Departamento de Informática y Automática

Facultad de Ciencias

Big Data na gestão eficiente das Smart Grids

HDS: Uma Plataforma Híbrida, Dinâmica e Inteligente

Ph.D. Thesis

Candidate

Eugénia Margarida Pinto Vinagre Moreira

Supervisors

Doutor Carlos Fernando da Silva Ramos

Doutor Juan Manuel Corchado Rodríguez

Septiembre 2019

Agradecimentos

Aos meus orientadores, Professor Doutor Carlos Fernando da Silva Ramos e Professor Doutor Juan Manuel Corchado Rodríguez pela orientação e apoio na concretização do presente trabalho.

Agradeço à Professora Doutora Zita Maria Almeida do Vale por me ter acolhido no Grupo de Investigação em Engenharia e Computação Inteligente para a Inovação e Desenvolvimento (GECAD) e por ter aceitado sem hesitação a realização deste trabalho.

À Professora Doutora María Navelonga Moreno García pela sua sempre prontidão no esclarecimento de dúvidas relacionadas com as questões formais do presente PhD, e pela sua compreensão em momentos mais complicados e menos bons, o meu muito obrigada.

Agradeço a todos os meus colegas do GECAD pelo apoio facultado durante a execução do presente trabalho.

Aos colegas do *Bioinformatics Intelligent Systems and Educational Technology Research Group* (BISITE) que tive o privilégio de conhecer e trabalhar durante as diversas deslocações e estadias realizadas em Salamanca no âmbito deste projeto, o meu obrigada pela forma exemplar que sempre me receberam e pela partilha de conhecimentos.

À Fundação para a Ciência e Tecnologia (FCT), agradeço pelo financiamento concedido para o desenvolvimento do presente trabalho, sem o qual não seria impossível a sua realização.

Finalmente, agradeço a todos os meus amigos e família pelo apoio incondicional durante esta longa jornada.

This work is supported by FEDER Funds through COMPETE program and by National Funds through FCT under the project UID/EEA/00760/2013 and by FCT under the project SFRH/BD/103089/2014 (Eugénia Vinagre PhD)

Resumo

Nos últimos anos tem-se verificado um acréscimo exponencial de informação gerada e disponibilizada a cada dia. Devido ao rápido avanço tecnológico (dispositivos móveis; sensores; comunicação wireless; etc.) biliões e biliões de bytes são criados todos os dias. Este fenómeno, denominado por Big Data, é caracterizado por 5 Vs (i.e. Volume, Velocidade, Variedade, Veracidade, Valor) e cada um deles representa verdadeiros desafios (e.g. como recolher e transportar um grande volume de informação; como armazenar essa informação; como minerá-la, como analisá-la e extrair conhecimento, como garantir a sua segurança e privacidade, como processá-la em tempo real, etc.). É unanime na comunidade científica que o valor a extrair de toda esta informação constituirá um fator de extrema importância para a tomada de decisão, determinante no sucesso das mais variadíssimas áreas económicas, bem como na resolução de inúmeros problemas. Nestas áreas inclui-se o ecossistema energético que por razões ecológicas, económicas e políticas conduziu ao repensar da forma como consumimos e produzimos energia. Devido ao aumento das necessidades energéticas provocado pelo avanço tecnológico, ao previsto esgotamento dos recursos energéticos não renováveis e devido às diretivas para a eficiência energética impostas pela União Europeia, muitos têm sido os estudos feitos na área da gestão de recursos energéticos. O termo Smart Grids surgiu nas últimas décadas com o objetivo de definir um ecossistema energético inteligente, que visa não só a integração de inteligência, mas também de automação na operabilidade extremamente complexa de todos os seus processos. As Smart Grids têm sido alvo de grandes estudos e investimentos dos quais têm resultado avanços significativos. No entanto, alguns desafios estão ainda por concretizar nomeadamente na gestão do seu complexo fluxo de dados. É neste contexto que se enquadra a presente dissertação cujo principal objetivo se centra na obtenção de soluções para alguns dos problemas identificados no domínio de Smart Grids com recurso às novas técnicas e metodologias propostas na área de Big Data.

Este trabalho apresenta um estudo sobre os recentes e crescentes avanços tecnológicos realizados na área de Big Data, onde são identificados os seus grandes desafios. Destes destacam-se a complexidade na gestão de fluxos contínuos e desordenados, a necessidade de reduzir o tempo despendido na pré-preparação dos dados e o desafio de explorar soluções que proporcionem a automatização analítica. Por outro lado, o estudo analisa o impacto da aplicação nas novas tecnologias no desenvolvimento das Smart Grids, no qual se conclui que apesar de embrionária, a sua aplicação é imprescindível para a evolução do ecossistema energético. Deste estudo resultou ainda a identificação dos principais desafios na área das Smart Grids, dos quais se destacam a complexidade na gestão do seu fluxo de dados em tempo real e a necessidade de melhorar a precisão das previsões de consumo e produção de energia.

Face aos desafios identificados foi proposto um modelo conceptual, baseado na arquitetura *Docker Container*, para o desenvolvimento de uma plataforma. Este modelo objetiva a flexibilidade e agilidade de forma a permitir a integração e validação das novas e crescentes abordagens tecnológicas propostas

na área de Big Data, necessárias ao desenvolvimento das Smart Grids. A fim de validar o modelo proposto, foi desenvolvida uma *stack* onde foram implementados vários serviços que visaram contribuir para os desafios identificados na área de Big Data e Smart Grids, nomeadamente: visualização e monitorização dos dados recolhidos em tempo real; preparação dos dados recolhidos em tempo real; previsão em tempo real de várias séries temporais simultaniamente; deteção de anomalias; avaliação da precisão das previsões e geração de novos modelos para a previsão de consumo e produção de energia segundo determinados critérios.

Finalmente foram desenvolvidos vários casos de estudo cujos resultados obtidos permitiram concluir sobre a importância da pré-preparação dos dados na fase analítica, sobre a eficiência na automatização analítica e sobre as vantagens da análise de ponta (*Edge Analytics*). Ao contrário de abordagens mais tradicionais que visam a execução centralizada do processo analítico, o *edge analytics* explora a possibilidade de executar a analise de dados de forma descentralizada a partir de um ponto não central do sistema. Os resultados permitiram concluir que o *edge analytics* traz vantagens acrescidas para a precisão das previsões. Permitiram ainda, inferir sobre como recolher os resultados a fim de se obter uma melhor precisão nas previsões, i.e., quanto mais específica e ajustada ao contexto forem executadas as previsões maior será a sua precisão.

Palavras-chave: Big Data, Smart Grid, Docker Containers, Processamento em Tempo Real, Big Data Analytics, Auto-ML, Edge Analytics, Big Data Visualization

Resumen

En los últimos años se ha verificado un aumento exponencial de información generada y disponible cada día. Debido al rápido avance tecnológico (dispositivos móviles, sensores, comunicación inalámbrica, etc.) billones y billones de bytes se crean todos los días. Este fenómeno, denominado Big Data, se caracteriza por 5 Vs (es decir, Volumen, Velocidad, Variedad, Veracidad, Valor) y cada uno de ellos representa verdaderos desafíos (por ejemplo, cómo recoger y transportar un gran volumen de información, cómo almacenar esa información, minarla, cómo analizarla y extraer conocimiento, cómo garantizar su seguridad y privacidad, cómo procesarla en tiempo real, etc.). Es unánime en la comunidad científica que el valor a extraer de toda esta información constituirá un factor de extrema importancia para la toma de decisión, determinante el éxito de las variadísimas áreas económicas, así como en la resolución de innumerables problemas. En estas áreas se incluye el ecosistema energético que por razones ecológicas, económicas y políticas condujo a repensar la forma en que consumimos y producimos energía. Debido al aumento de las necesidades energéticas provocado por el avance tecnológico, al previsto agotamiento de los recursos energéticos no renovables y debido a las directivas para la eficiencia energética impuestas por la Unión Europea, muchos han sido los estudios realizados en el ámbito de la gestión de recursos energéticos. El término Smart Grid surgió en las últimas décadas con el objetivo de definir un ecosistema energético inteligente, que apunta no sólo a la integración de inteligencia, sino también de automatización en la operatividad extremadamente compleja de todos sus procesos. Las Smart Grids han sido objeto de grandes estudios e inversiones de los cuales han resultado avances significativos. Sin embargo, algunos desafíos aún no se concretan en la gestión de su complejo flujo de datos. Es en este contexto que se encuadra la presente disertación cuyo principal objetivo se centra en la obtención de soluciones para algunos de los problemas identificados en el dominio de Smart Grids utilizando las nuevas técnicas y metodologías propuestas en el área de Big Data.

Este trabajo presenta un estudio sobre los recientes y crecientes avances tecnológicos realizados en el área de Big Data, donde se identifican sus grandes desafíos. De ellos se destacan la complejidad en la gestión de flujos continuos y desordenados, la necesidad de reducir el tiempo empleado en la pre-preparación de los datos y el desafío de explorar soluciones que proporcionen la automatización analítica. Por otro lado, el estudio analiza el impacto de la aplicación de nuevas tecnologías en el desarrollo de las Smart Grids, en el que se concluye que, a pesar de embrionaria, su aplicación es imprescindible para la evolución del ecosistema energético. De este estudio resultó también la identificación de los principales desafíos en el área de las Smart Grids, de los cuales se destacan la complejidad en la gestión de su flujo de datos en tiempo real y la necesidad de mejorar la precisión de las previsiones de consumo y producción de energía.

En cuanto a los desafíos identificados, se propuso un modelo conceptual, basado en la arquitectura *Docker Container*, para el desarrollo de una plataforma. Este modelo tiene como objetivo la flexibilidad

y agilidad para permitir la integración y validación de los nuevos y crecientes enfoques tecnológicos propuestos en el área de Big Data, necesarios para el desarrollo de las Smart Grids. Con el fin de validar el modelo propuesto, se desarrolló una *stack* donde se implementaron varios servicios que pretendían contribuir a los desafíos identificados en el área de Big Data y Smart Grids, en particular: visualización y seguimiento de los datos recogidos en tiempo real; preparación de los datos recogidos en tiempo real; previsión en tiempo real de multillas series temporales simultáneamente; detección de anomalías; evaluación de la precisión del predicción y generación de nuevos modelos para la previsión de consumo y producción de energía según ciertos criterios.

Finalmente, se desarrollaron una serie de casos de estudo cuyos resultados nos permitieron concluir sobre la importancia de la preparación previa de los datos en la fase analítica, la eficiencia en la automatización analítica y las ventajas del análisis de borde (*Edge Analytics*). A diferencia de los enfoques más tradicionales para la ejecución centralizada del proceso analítico, el análisis de borde explora la posibilidad de realizar análisis de datos de forma descentralizada desde un punto no central del sistema. Los resultados permitieron concluir que el análisis de borde aporta ventajas añadidas a la precisión de los pronósticos. También nos permitieron inferir cómo recopilar los resultados para obtener una mejor precisión en las predicciones, por ejemplo, cuanto más precisos y ajustados al contexto se ejecuten los pronósticos, mayor será su precisión.

Palabras clave: Big Data, Smart Grid, Docker Containers, Real Time Processing, Big Data Analytics, Auto-ML, Edge Analytics, Big Data Visualization

Abstract

In recent years, there has been an exponential increase of information generated and made available every day. Due to rapid technological advancement (e.g., mobile devices, sensors, wireless communication, etc.) billions and billions of bytes are created every day. This phenomenon, called Big Data, is characterized by 5 Vs (i.e., Volume, Velocity, Variety, Veracity, Value) and each represents real challenges (e.g., how to collect and carry a large amount of information; how to store this information; how mining it, analyzing it and extracting knowledge; how to ensure its security and privacy; how to process it in real time, etc.). It is unanimous in the scientific community that the value to be extracted from all this information will be a factor of extreme importance for the decision making, determining the success of the most varied economic areas, as well as the resolution of numerous problems. These areas include the energy ecosystem that, for ecological, economic and political reasons, led us to rethink the way we consume and produce energy. Due to the increase in energy needs caused by technological advances, the expected depletion of non-renewable energy resources and due to the energy efficiency directives imposed by the European Union, many studies have been carried out in the area of energy resources management. The term Smart Grid has emerged in the last decades with the objective of defining an intelligent energy ecosystem, which aims not only to integrate intelligence but also to automate the extremely complex operability of all its processes. Smart grids have been the subject of major studies and investments which have resulted in significant advances. However, some challenges have to be addressed in the management of its complex data flow. It is in this context that the present dissertation falls, with the main objective on obtaining solutions to some of the problems identified in the field of Smart Grids using new techniques and methodologies proposed in the area of Big Data.

This paper presents a study on the recent and growing technological advances in the area of Big Data, where its major challenges are identified. These include complexity in the management of continuous and disordered flows, the need to reduce the time spent in pre-preparation of data and the challenge of exploring solutions that provide analytical automation. On the other hand, the study analyzes the impact of the application in the new technologies in the development of the Smart Grids, in which it is concluded that, although embryonic, its application is essential for the evolution of the energy ecosystem. This study also resulted in the identification of the main challenges in the area of Smart Grids, which highlight the complexity in managing its data flow in real time and the need to improve the accuracy of energy consumption and production forecasts.

Given the identified challenges, a conceptual model, based on the Docker Container architecture, was proposed for the development of a platform. This model aims at flexibility and agility in order to allow the integration and validation of the new and growing technological approaches proposed in the area of Big Data, necessary for the development of Smart Grids. In order to validate the proposed model, a *stack*

was developed where several services were implemented that aimed to contribute to the challenges identified in the area of Big Data and Smart Grids, namely: visualization and monitoring of data collected in real time; preparation of data collected in real time; real-time forecasting of multiple time series simultaneously; detection of anomalies; evaluation of the accuracy of forecasting and generation of new models for the forecast of consumption and production of energy according to certain criteria.

Finally, a number of case studies were developed whose results allowed us to conclude on the importance of the pre-preparation of the data in the analytical phase, on the efficiency in the analytical automation and on the advantages of the Edge Analytics. Unlike more traditional approaches to the centralized execution of the analytic process, edge analytics explores the possibility of performing data analysis in a decentralized way from a non-central point of the system. The results allowed to conclude that edge analytics brings added advantages to the precision of the forecasts. Results allowed us to infer how to collect the data in order to obtain a better precision in the predictions, i.e., the more precise and context-adjusted the forecasts are executed the greater their accuracy.

Keywords: Big Data, Smart Grid, Docker Containers, Real Time Processing, Big Data Analytics, Auto-ML, Edge Analytics, Big Data Visualization

Índice

Índice de Figuras

Índice de Tabelas

Notação e Glossário

AB	Apache Beam
ACID	Atomicidade, Consistência, Isolamento, Durabilidade
AGI	Artificial General Inteligente
AI	Artificial Inteligente
AMQP	Advanced Message Queuing Protocol
API	Application Programming Interface
AWS	Amazon Web Services
BASE	Basic Availability, Soft-state, Eventual Consistency
BD	Big Data
BDA	Big Data Analítica
BDR	Bases de Dados Relacionais
BDVA	Big Data Value Association
BI	Business Intelligence
BISITE	Bioinformatics Intelligent Systems and Educational Technology Research Group
BSON	Binary JSON
CAP	Consistency, Availability, Partition tolerance
CEN	Comité Europeu de Normalização
CENELEC	Comité Europeu de Normalização Eletrotécnica
CEO	Chief Executive Officer
CLI	Comand-line interface
CNTK	Microsoft Cognitive Toolkit
CPP	Critical Peak Pricing
CPU	Central Process Unit
DC	Docker Containers
DER	Distributed Energy Resource
DFS	Distributed file systems
DL	Deep Learning
DR	Demand Response
ETSI	Instituto Europeu de Normas de Telecomunicações
EV	Electric Vehicle
FCT	Fundação para a Ciência e Tecnologia
GECAD	Grupo de Investigação em Engenharia e Computação Inteligente para a Inovação e Desenvolvimento
HDFS	Hadoop Distributed File System
HDS	Hybrid, Dynamics and Smart
HIPAA	Health Insurance Portability and Accountability Act
HPC	High Performance Computing
HVAC	Heating, ventilation, and air conditioning
I&D	Investigação e Desenvovimento
IIOT	Industrial Internet of Things
IoT	Internet of Things

IPP	Instituto Politécnico do Porto
ISEP	Instituto Superior de Engenharia do Porto
iSOC	Integrated Smart Operations Center
JOSM	Java OpenStreetMap
JSM	Java Message Service
JSON	JavaScript Object Notation
M2M	Machine-to-Machine
MAE	Mean absolute error
MAPE	Mean Absolute Percentage Error
MG	Microgrid
ML	Machine Learning
MOM	Message-oriented Middleware
MPI	Message Passing Interface
MQTT	Message Queue Telemetry Transport
MSE	Mean Squared Error
MTC	Many-Task Computing
NIST	National Institute of Standards and Technology
NoSQL	Not Only SQL
NXD	Native XML DBMS
NYPA	New York Power Authority
OCI	Open Container Initiative
P2P	Peer to Peer
PaaS	Platform as a Service
PCI-DSS	Payment Card Industry Data Security Standard
PLC	Power Line Communication
PTR	Peak Time Rebate
PUM	Phasor Measurement Units
RDD	Resilient Distributed Dataset
RDF	Resource Description Framework
REST	Representational State Transfer
RFID	Radio-Frequency Identification
RGPD	Regulamento Geral de Proteção de Dados
RMSE	Root Mean Squared Error
S3	Amazon Simple Storage Service
S4	Simple Scalable Streaming System
SCADA	Supervisory Control and Data Acquisition
SDK	Software Development Kit
SG	Smart Grids
SGBD	Sistema de Gestão e Base de Dados
SQL	Structured Query Language
STOMP	Simple Text Oriented Messaging Protocol
SVM	Support Vector Machine
TOU	Time-of-use

TSDB	Time Series Database
UDBMS	Unified Database Management System
UE	União Europeia
UPS	Uninterruptible Power Supply
VM	Virtual Machine
VPP	Virtua Power Player
XML	Extensible Markup Language
XMPP	Extensible Messaging and Presence Protocol
XSLT	eXtensible Stylesheet Language Transformations
YAML	YAML Ain't markup language
YCSB	Yahoo! Cloud Serving Benchmark

1 Introdução

Este capítulo pretende fornecer uma visão global do tema subjacente na presente dissertação, i.e., a utilização e desenvolvimento das tecnologias Big Data na gestão do complexo fluxo de dados inerente ao ecossistema energético. Após o enquadramento do tema, é descrita sumariamente a principal motivação do presente trabalho, seguindo-se a apresentação dos principais objetivos a serem concretizados, bem como os seus principais contributos. Finalmente é apresentada a estrutura da dissertação.

1.1 Enquadramento

Esta nova era digital caracterizada por um crescimento exponencial do sector de tecnologias impulsionou e deu origem a um novo fenómeno denominado Big Data. Big Data é caracterizada por 5V's (i.e., volume, velocidade, variedade, veracidade e valor). O termo não se refere apenas à grande quantidade de dados gerados a cada segundo (i.e., *petabytes* ou mais), mas também ao tipo de dados (i.e., estruturados, semiestruturados e destruturados). Para além disso, caracteriza-se pela necessidade de validar a veracidade da informação e dela extrair valor.

Esta nova era proporciona às organizações a possibilidade de desenvolver soluções para melhorar o desempenho das suas atividades. Os dados são um dos seus principais ativos, a matéria prima para a descoberta de novo conhecimento e o principal suporte no apoio à tomada de decisões. No entanto, a gestão e tratamento do fluxo de dados gerado em redor das variadíssimas atividades organizacionais, normalmente caracterizado como sendo continuo e desordenado, consiste num processo cada vez mais complexo. Muitos esforços têm sido feitos pela comunidade científica e organizacional no sentido de solucionar muitos dos desafios inerentes à gestão da informação. Integração e correlação de informação correspondem a um dos pontos principais para a extração de conhecimento. É complexo o desenvolvimento de arquiteturas flexíveis com capacidade de: agilizar a validação e integração de um grande volume de dados originários de variadíssimas fontes; gerir de forma segura os complexos fluxos de dados contínuos e desordenados; e proporcionar analise e visualização de dados em tempo real.

Por outro lado, por razões ecológicas, económicas e políticas, temos vindo a assistir, a grandes esforços no sentido de repensar e implementar a forma como consumimos e produzimos energia. O termo Smart Grids surgiu nas últimas décadas como definição de ecossistemas energéticos inteligentes, que visam não só a integração de inteligência, mas também de automação na operabilidade extremamente complexa de todos os seus processos. Muitos desafios estão ainda por conquistar na área que se dedica ao desenvolvimento das Smart Grids, e.g., descentralização do mercado de energias; comunicações bidirecionais seguras entre todos os intervenientes do ecossistema energético; autonomização analítica nos processos de previsão de consumo/produção de energia; rentabilização dos ativos energéticos de

forma a dar precedência às energias renováveis; automação da análise preditiva de todos os equipamentos do ecossistema; desenvolvimento sustentável de casas e edifícios inteligentes (*smart homes* e *smart buildings*) de forma a responderem positivamente aos requisitos da demanda, sem porem em risco o conforto mínimo dos seu ocupantes; gestão dos veículos elétricos de forma a contribuírem positivamente para o balanceamento e estabilidade da rede elétrica; e sensibilização dos consumidores finais para os impactos do desperdício e incorreto consumo de energia. O ecossistema energético é um dos maiores pilares no funcionamento das sociedades. Por essa razão não é simples a implementação de novas abordagens tecnológicas que aparentemente possam parecer benéficas para a conquista de muitos dos seus desafios. Todas as novas abordagens devem ser devidamente testadas e experimentadas antes de serem aplicadas no contexto real do ecossistema energético, por forma a não o colocarem em risco.

É neste contexto que se enquadra a presente tese, i.e., propor uma solução suficientemente ágil de forma a permitir a experimentação das novas tecnologias desenvolvidas no âmbito de Big Data no contexto das Smart Grids. De entre muitos desafios inerentes à gestão complexa que caracteriza o fluxo de dados do ecossistema energético, o trabalho desenvolvido e apresentado nesta tese objetiva contribuir para o problema do processamento em real time da previsão das principais variáveis equilibradoras do ecossistema energético, i.e., previsão do consumo e produção de energia. Não obstante, ambiciona a proposta de uma solução capaz de abranger a experimentação das mais diversas operações que envolvem a gestão de dados no ecossistema energético, de forma a contribuir positivamente para a tomada de decisão em tempo real, fundamental à sua sustentabilidade. Por outro lado, objetiva que os contributos feitos no contexto das Smart Grids contribuam de igual modo para os grandes desafios que caracterizam Big Data.

1.2 Motivação

A principal motivação para o desenvolvimento do presente trabalho emergiu pela grande curiosidade sobre o principal assunto nele focado, i.e., *Big Data*. De facto, a descoberta de conhecimento é algo fascinante e esta nova era de informação constitui uma excelente oportunidade para a sua concretização. São inúmeras as atividades económicas que podem beneficiar do crescimento exponencial de informação que se tem vindo a verificar nos últimos tempos, por forma a se transformarem em algo mais sustentável. Da mesma forma se torna difícil enumerar as áreas científicas que podem recolher os benefícios desta nova era de informação, visto que a informação é a matéria prima para a descoberta de conhecimento. Desta forma torna-se bastante desafiante explorar e contribuir positivamente para os desafios inerentes a área de *Big Data*, como forma indireta de contribuir para os desafios de outras áreas. Juntam-se a esta motivação os desafios impostos ao ecossistema energético, com vista a garantir a sua sustentabilidade.

Com base nestas principais motivações foi proposta e desenvolvida a presente tese que objetiva um contributo positivo para os desafios na área de *Big Data* no contexto das *Smart Grids*.

1.3 Objetivos da dissertação

A presente dissertação tem como principal objetivo obter soluções para alguns dos problemas identificados no domínio de Smart Grids com recurso às novas técnicas e metodologias de Big Data. Pretende-se melhorar a previsão da demanda de energia de forma a agilizar a sua produção, evitando o desperdício e garantindo a sua sustentabilidade. Para que tal seja possível é necessário dotar as Smart Grids de processamento em tempo real. O processamento em tempo real torna-se elemento fundamental na tomada de decisões quer na gestão da produção quer na gestão do consumo. A seguir descrevem-se sumariamente os objetivos delineados:

Objetivos na área de Smart Grids:

- Melhorar o resultado das previsões (i.e., produção e consumo), com a agregação de informação de várias fontes. Para a execução deste objetivo propôs-se um estudo inicial das técnicas de Big Data. Em resultado desse estudo visou-se a aplicação, adaptação ou desenvolvimento de novas metodologias que permitam a concretização do objetivo;
- Processamento em tempo real nos sistemas de gestão das Smart Grids com vista a melhorar a execução de modelos analíticos. Para atingir este objetivo propôs-se um estudo das metodologias de processamento em tempo real existentes no domínio de Big Data e sua posterior aplicação, adaptação, ou caso necessário o desenvolvimento de novas metodologias;
- Melhorar a interação e comunicação entre os intervenientes da rede. Para a concretização deste objetivo propôs-se um estudo inicial visando as tecnologias *Big Data Visualization*, para a sua posterior aplicação nos sistemas de Smart Grids.

Objetivos na área de Big Data:

- Estudo sobre as evoluções tecnológicas operada na área de Big Data bem como a sua aplicabilidade na área de Smart Grids, contemplando soluções abertas (*open source*) e comerciais;
- Aplicação, adaptação ou desenvolvimento de novas metodologias para o melhoramento do processo analítico relacionado com as previsões;
- Aplicação, adaptação ou desenvolvimento de novas metodologias para o melhoramento relacionado com o processamento e gestão de um grande volume de dados em tempo real;
- Conceção e desenvolvimento de aplicações que utilizem as metodologias desenvolvidas no contexto das Smart Grids.

1.4 Contribuições

Ao longo do desenvolvimento deste trabalho verificou-se uma progressiva evolução das tecnologias projetadas com o objetivo de colmatar os grandes desafios inerentes a esta nova era de informação. Destes desafios pode-se, por exemplo, referir a evolução na área de processamento. Esta, aquando da submissão da preposta do presente trabalho, encontrava-se praticamente confinada à abordagem *Hadoop*. Esta abordagem foi projetada unicamente para resolver o problema do processamento de grandes volumes de dados. Só mais tarde é que se verificou o aparecimento de abordagens que visaram o processamento em tempo real, nomeadamente o *Apache Spark*, *Apache Flink*, *Apache Aplex*, etc. Outro bom exemplo a referir foca-se na área de armazenamento cujas abordagens disponibilizadas cresceram exponencialmente. Face aos sucessivos avanços operados na área de Big Data e aos seus grandes desafios que se encontram ainda por solucionar, o principal contributo do presente trabalho pode ser resumido da seguinte forma:

- Apresentação de um estudo sobre as evoluções tecnológicas operadas na área de Big Data bem como a sua aplicabilidade na área de Smart Grids. O estudo identifica os grandes desafios inerentes à área de Big Data dos quais se salientam: o processamento em tempo real de fluxo de dados contínuos e desordenados; o tempo despendido na pré-preparação de dados; e a automatização analítica. O estudo destaca ainda os grandes desafios na área das Smart Grids, nomeadamente a complexidade inerente ao seu fluxo de dados e a necessidade imprescindível de o tratar e processar em tempo real;
- Proposta de um modelo conceptual para a implementação de uma plataforma flexível, com capacidade para permitir a integração, experimentação e avaliação das novas tecnologias desenvolvidas no âmbito de Big Data, aplicadas ao contexto das Smart Grids;
- Desenvolvimento de uma *stack* objetivando testar o modelo proposto e a implementação de serviços capazes de responder positivamente aos desafios identificados em Big Data (i.e., processamento em tempo real visando a pré-preparação de dados e a automatização analítica) e nas Smart Grids (i.e., deteção de anomalia e processamento em tempo real da previsão de consumo e produção de energia com a maior precisão possível);
- Desenvolvimento de vários casos de estudo visando a experimentação e validação dos serviços implementados e disponíveis na *stack*, i.e., visualização e monitorização dos dados recolhidos em tempo real; preparação dos dados recolhidos em tempo real; previsão em tempo real; deteção de anomalias; avaliação da precisão da previsão; geração de novos modelos para a previsão de consumo/produção de energia segundo determinados critérios.

Ao longo do desenvolvimento deste trabalho resultaram ainda algumas publicações que visaram sobretudo a exploração das metodologias analíticas para a previsão do consumo e de produção de

energia. Foram também feitas publicações com o objetivo de propor modelos conceptuais para o desenvolvimento de plataformas Big Data no contexto das Smart Grids. Por fim foram ainda feitas outras publicações resultantes da divulgação de trabalhos (i.e., trabalhos que visaram a integração de serviços relacionados com a operabilidade das Smart Grids), executado no âmbito de alguns projetos. As publicações referidas são a seguir enumeradas:

- E. Vinagre, L. Gomes and Z. Vale, "*Electrical Energy Consumption Forecast Using External Facility Data*", Computational Intelligence, 2015 IEEE Symposium Series on, Cape Town, 2015, pp. 659-664.; doi: 10.1109/SSCI.2015.101;
- E. Vinagre, T. Pinto, S. Ramos, Z. Vale and J. M. Corchado, "*Electrical Energy Consumption Forecast Using Support Vector Machines*", 2016 27th International Workshop on Database and Expert Systems Applications (DEXA), Porto, 2016, pp. 171-175. doi: 10.1109/DEXA.2016.046;
- E. Vinagre, J. F. De Paz, T. Pinto, Z. Vale, J. M. Corchado and O. Garcia, "*Intelligent energy forecasting based on the correlation between solar radiation and consumption patterns*", 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, 2016, pp. 1-7. doi: 10.1109/SSCI.2016.7849853;
- E. Vinagre, T. Pinto, I. Praça, L.Gomes, J. Soares, Z. Vale "*Shared Intelligence Platform for Collaborative Simulations using Sequences of Algorithms: An Electricity Market participation case study*", 12th IEEE Power and Energy Society, PowerTech Conference, Manchester, UK 18-22 june 2017;
- I. Praça, J. Soares, L. Gomes, E. Vinagre, Z. Vale and L. A. Belhaj, "*Shared Intelligence for smart grids management*", 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, 2016, pp. 1-8. doi: 10.1109/SSCI.2016.7849855;
- E. Vinagre, T. Pinto, G. Pinheiro, Z. Vale, J. M. Corchado and C. Ramos, "*Knowledge management system for big data in a smart electricity grid context*" 2017 IFKAD Knowledge Management in the 21st Century: Resilience, Creativity and Co-creation, St. Petersburg, Russia 7-9 June 2017;
- E. Vinagre, T. Pinto, Z. Vale, C. Ramos, "*Big Data in efficient Smart Grids management*", PAAMS'17 Doctoral Consortium, 15th International Conference on Practical Applications of Agents and Multi-Agent Systems, Porto, Portugal 21-23 June 2017;
- G. Pinheiro, E. Vinagre, I. Praça, Z. Vale, C. Ramos, "*Smart Grids Data Management: A Case for Cassandra*", DCAI'17 14th International Conference on Distributed Computing and Artificial Intelligence, Porto, Portugal 21-23 June 2017.

Os projetos, conforme referido, são os seguintes:

- DREAM-GO - *Enabling demand response for short and real-time efficient and market-based smart grid operation. An intelligent and real-time simulation approach*. Este projeto recebeu financiamento do programa de investigação e inovação Horizonte 2020 da União Europeia no âmbito do acordo de financiamento Marie Skłodowska-Curie n.º 641794;
- GID-Microgrid - *Intelligent management of private decentralized microgrids*. Este projeto foi financiado pelo QREN, (Ref. 34086);
- SEAS - *Smart Energy Aware Systems*. Este projeto foi financiado no âmbito do programa europeu ITEA2, com a referência 12004.

1.5 Estrutura da dissertação

A presente dissertação está estruturada em sete capítulos. Este primeiro capítulo introdutório tem como principal objetivo descrever o enquadramento do tema subjacente à presente dissertação, assim como apresentar a principal motivação e os objetivos delineados para o seu desenvolvimento. Os restantes capítulos estão organizados como a seguir se descreve.

2º Capítulo - Big Data

Este capítulo apresenta um estudo pormenorizado sobre as abordagens que progressivamente têm sido propostas na área de Big Data com o objetivo de minimizar a complexidade inerente ao tratamento de dados. O estudo está organizado de acordo com as suas principais subáreas, i.e., recolha e transporte de dados, processamento, análise, armazenamento, visualização e governação de dados. São identificados os seus grandes desafios dos quais se salientam o processamento em tempo real de fluxo de dados contínuos e desordenados, o tempo consumido na pré-preparação de dados e ainda a automatização analítica. Apresenta ainda uma revisão sobre a aplicabilidade das novas abordagens.

3º Capítulo - Smart Grids

Este capítulo descreve os principais conceitos que definem as Smart Grids destacando a sua evolução no contexto Europeu. Revisa ainda a importância e o impacto da evolução tecnológica operada na área de Big Data no contexto das Smart Grids em três perspetivas diferentes, i.e., desenvolvimentos realizados na comunidade científica, impacto sectorial de Big Data na área das Smart Grids e finalmente, a posição da União Europeia relativamente ao investimento para o desenvolvimento da área de Big Data no contexto das Smart Grids.

4º Capítulo - Soluções propostas e pesquisa experimental

Este capítulo apresenta as várias soluções propostas e exploradas no sentido de darem uma resposta positiva aos problemas identificados como desafios nas áreas de Big Data e Smart Grids. São identificadas as suas principais limitações, as quais conduziram à proposta de um modelo conceptual. É ainda introduzido o conceito da arquitetura adotada na solução selecionada, i.e. *Docker Container*. O

capítulo termina com uma breve conclusão sobre os aspectos principais que conduziram à solução adotada.

5º Capítulo - Implementação da Stack HDS

Neste capítulo é apresentada a *stack* desenvolvida com o objetivo de testar o modelo proposto no capítulo anterior. Ao longo dos seus subcapítulos são descritos alguns detalhes da implementação dos vários serviços desta *stack*, que objetivaram contribuir para a minimização de alguns problemas identificados nas áreas de Big Data e Smart Grids.

6º Capítulo - Análise de resultados

Este capítulo apresenta a análise de resultados na perspectiva da eficiência dos serviços implementados e configuradas na *stack* desenvolvida. Apresenta os vários casos de estudo desenvolvidos para esse fim, nomeadamente, validação da eficiência do serviço de pré-prepararão dos dados, validação dos critérios implementados para a remodelagem de novos modelos preditivos, validação da automatização analítica, avaliação do melhor cenário de forma a ser garantida a precisão das previsões e finalmente a exploração do conceito subjacente a *edge analytic*.

7º Capítulo - Conclusões

Neste capítulo são apresentadas as conclusões do trabalho realizado e descrito ao longo desta dissertação. São ainda sumariadas as principais contribuições e limitações do corrente trabalho, bem como a proposta para futuros trabalhos.

2 Big Data

O presente capítulo apresenta um estado da arte na área Big Data explorando os seus principais conceitos, desafios e avanços. O capítulo encontra-se subdividido pelos vários ramos que integram Big Data, i.e., recolha e transporte de dados, processamento, análise, armazenamento, visualização e governação de dados.

2.1 Introdução

Big Data (BD) foi o termo designado para definir esta nova era de informação, caracterizada por 5V's (i.e., volume, velocidade, variedade, veracidade e valor). O Big Data não se refere apenas à grande quantidade de dados gerados a cada segundo (i.e., *petabytes* ou mais), mas também ao tipo de dados (i.e., estruturados, semiestruturados e destruturados). Para alem disso, caracteriza-se pela necessidade de validar a veracidade da informação e dela extrair valor. Com o impulsionamento da Internet das Coisas (IoT), Big Data transformou-se num desafio ainda maior e de extrema complexidade. Os dados representam uma matéria prima de enorme valor. Muitos têm sido os esforços feitos no sentido de desenvolver soluções capazes de os gerir. A evolução dos *landspaces* referentes às soluções Big Data desenvolvidas e disponibilizadas entre 2012 e 2018, conforme Figura 2.1, é bem ilustrativo deste fenómeno. Escolher uma solução, ou mesmo um conjunto de várias soluções complementares, não é tarefa simples.

Figura 2.1- Landspaces de soluções Big Data

No planeamento da implementação de uma plataforma baseada nas soluções disponíveis, requer uma definição clara dos principais objetivos que se pretende atingir (e.g., que tipo de dados são necessários recolher, que operações farão uso desses dados, que modelos analíticos serão usados, qual o tempo de resposta esperado no seu processamento, etc.). Por outro lado, é igualmente necessário efetuar uma pesquisa sobre os requisitos, funcionalidades e desempenho das várias soluções disponíveis a fim de selecionar as que melhor se enquadram nos objetivos delineados. Com este propósito, o presente capítulo apresenta um estado da arte na área Big Data e encontra-se subdividido pelos vários ramos que a integram, i.e., recolha e transporte de dados, processamento, processamento, análise, armazenamento, visualização e governação de dados.

2.2 Recolha e transporte de dados

Recolha de dados, (denominado em inglês pelo termo *"Data collect"*), refere-se à capacidade de um sistema recolher dados das mais variadíssimas fontes. Os dados podem ter origem em fontes externas ao sistema ou serem oriundos da própria operacionalidade do sistema. Para o transporte de dados entre sistemas ou entre os componentes que integram o sistema é comum o uso de sistemas de mensagens. Estes sistemas podem ser classificados segundo o seu tipo de sincronismo/assincronismo no processo de envio e receção de mensagens[1]:

- *Message Passing Interface* (MPI) - é uma abordagem que se baseia na troca de mensagens de forma síncrona (i.e., o emissor e o recetor das mensagens têm de estar ativos e disponíveis simultaneamente para que a transmissão das mensagens se processe);
- *Message-oriented Middleware* (MOM) – é uma abordagem que se baseia no conceito de comunicação desacoplada. Surgiu como uma solução para colmatar a complexidade das plataformas mais modernas (i.e., plataformas caracterizadas por um conjunto de componentes distribuídos) e a complexidade inerente dos fluxos de dados (i.e., fluxos de dados contínuos e fora de ordem).

MOM é uma infraestrutura desenvolvida para suportar o envio e a receção de mensagens de forma assíncrona, evitando deste modo congestionamentos na operabilidade dos sistemas. Assim, MOM pode ser visto como um componente fundamental nos sistemas modernos, não só pela sua faculdade de recolher dados externos, mas também pela sua capacidade de estabelecer um canal de transmissão de dados entre os componentes da sua arquitetura. O componente principal do MOM é o denominado *Broker de mensagens*, conforme ilustrado na Figura 2.2.

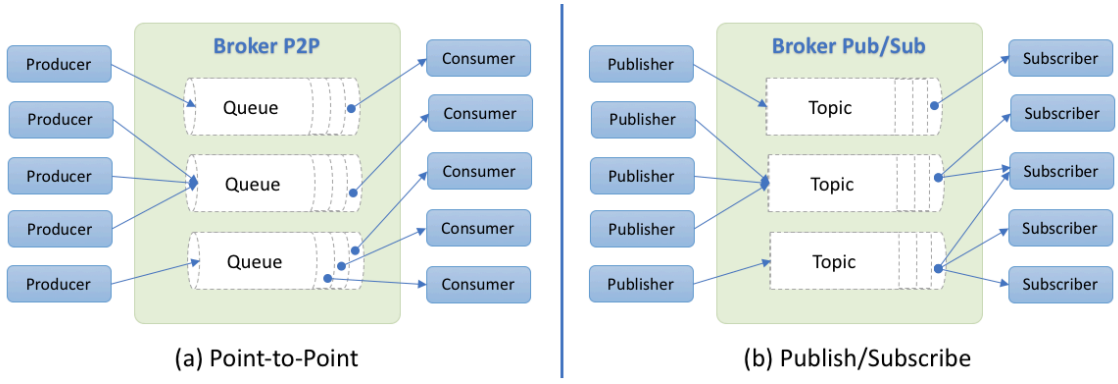


Figura 2.2 - MOM Messaging models

As mensagens enviadas para o *broker* são organizadas em filas e/ou tópicos. Segundo a metodologia utilizada na organização das mensagens, o MOM pode ser classificado de acordo com um dos seguintes modelos [2]:

- Ponto a Ponto (P2P) – Neste modelo as mensagens são enviadas e entregues de forma ordenada uma determinada fila especifica. Cada fila pode ter um ou vários recetores e consumidores, no entanto, cada mensagem é entregue uma e uma só vez a um dos recetores que se encontre em escuta na fila. Após a sua entrega, a mensagem é eliminada da fila.
- Publicar-subscrever – Neste modelo as mensagens são organizadas em tópicos. Os tópicos são devidamente identificados a fim de facilitar a sua subscrição. Todas as mensagens enviadas para um determinado tópico são transmitidas a todos os subscritores desse tópico.
- Modelo híbrido – Neste modelo as mensagens são organizadas em filas e em tópicos.

Os sistemas MOM mais recentes suportam a sua implementação em *clustering*, de forma a permitir escalonamento, tolerância a falhas e alta performance. De seguida descrevem-se algumas características e propriedades possíveis de serem implementadas nos sistemas MOM:

- Protocolos de mensagens – existe uma grande variedade de protocolos possíveis de serem implementados em sistemas MOM, dos quais enumeramos alguns dos mais utilizados:
    - AMQP (*Advanced Message Queuing Protocol*) [3] – Este protocolo é definido pela norma ISO/IEC 19464:2014. Foi desenvolvido com o principal objetivo de garantir interoperabilidade entre sistemas MOM. É compatível com ambos os modelos, P2P e pub/sub. Outro ponto forte é a garantia de confiabilidade. É composto por uma variadíssima gama de funcionalidades e configurações (e.g., definições de garantia de entrega, restrição a filas e/ou tópicos, definição de prioridades e durabilidade das mensagens, definição de transações, etc.);
    - MQTT (*Message Queue Telemetry Transport*) [4] – Este protocolo é definido pela norma ISO/IEC 20922:2016. Foi desenvolvido com o principal objetivo de facilitar a comunicação entre dispositivos com recursos limitados e baixa largura de banda (*network bandwidth*), i.e., sistemas desenvolvidos com o compromisso de garantirem a redução da latência de rede. É compatível com o modelo pub/sub e disponibiliza funcionalidades de forma a garantir alguma confiabilidade e garantia de entrega. Os seus pontos fortes são: simplicidade, rapidez na comunicação e performance. É amplamente utilizado nas comunicações máquina a máquina (M2M - *machine-to-machine*), na comunicação entre dispositivos conectados na Internet das coisas (IoT-*Internet of Things*) e em aplicações moveis (*mobile applications*);
    - STOMP (*Simple Text Oriented Messaging Protocol*) [5] - Este protocolo é orientado a texto e foi desenvolvido com o objetivo de conciliar simplicidade e interoperabilidade. Os pontos fortes deste protocolo são: simples na sua implementação, leve, amplamente interoperável e compatível com várias linguagens de programação;

- XMPP (*Extensible Messaging and Presence Protocol*) [6] – O protocolo XMPP é uma tecnologia XML aberta para comunicação em tempo real, compatível com uma vasta gama de aplicativos (e.g. mensagens instantâneas, presença, colaboração, etc.).

- Garantia de entrega – Esta característica representa a capacidade que o sistema detem para evitar perda de informação transmitida em rede e evitar sempre que possível a receber informação duplicada. Em termos de garantia de entrega os sistemas MOM podem operar segundo três modalidades distintas:

  - Em-mais-uma vez - neste modelo, também conhecido como *Fire-and-Forget*, não é garantida a entrega da mensagem ao recetor. A mensagem é enviada uma só vez e o recetor não confirma a mensagem lhe foi entregue, i.e., uma mensagem é enviada apenas uma vez, e pode ser entregue ou perdida;

  - Pelo menos uma vez - este é o modelo mais usado e consiste no reenvio da mensagem até que o destinatário confirme a sua receção. Neste caso, se a confirmação da receção for perdida, o destinatário poderá receber várias vezes a mesma mensagem;

  - Exatamente-uma vez - este modelo é o mais desejado, no entanto devido a erros de transmissão de dados na rede, garantir que uma mensagem é sempre entregue e apenas uma só vez é extremamente complexo.

- Propriedades da entrega – Os sistemas MOM podem ter a capacidade de implantar as seguintes modalidades na entrega de mensagens:

  - Entregas *Push/Pull* - numa entrega *Push* o sistema entrega ao destinatário as mensagens o mais rapidamente possível, enquanto que numa entrega *Pull* as mensagens são transmitidas para o destinatário quando este as solicita explicitamente;

  - Entrega ordenada - representa a capacidade de o sistema entregar as mensagens pela ordem em que as mesmas foram geradas. Têm em conta que o momento em que as mensagens são criadas pode ser diferente do momento em que estas são entregues;

  - Entrega transacional - representa a capacidade que um sistema tem em agrupar mensagens como uma unidade atômica. Esta característica garante que esse conjunto seja entregue na totalidade com êxito, caso contrário a entrega deverá ser invalidada;

  - Entrega filtrada - representa a capacidade que o sistema tem em facultar a distribuição e entrega de mensagens de acordo com regras impostas pelos seus clientes;

  - Entrega com prioridade – representa a capacidade do sistema poder atribuir prioridade na entrega das mensagens, independentemente da sua ordem de chegada.

- Persistência e Durabilidade - Persistência significa que o sistema tem capacidade de proceder ao armazenamento das mensagens durante o período que decorre entre a sua receção e o seu envio. Durabilidade reside na capacidade que o sistema detém para não perder as mensagens, i.e., a sua capacidade em reter as mensagens até que seja possível efetuar a sua entrega.
- Latência e Taxa de transferência - Latência representa o tempo que decorre entre receber e enviar uma mensagem, enquanto que a taxa de transferência está diretamente relacionada com a largura de banda, ou seja, o volume de dados que o sistema tem capacidade para despachar em determinado tempo. Quanto maior for a taxa de transferência no despacho das mensagens, menor será a sua latência.
- Robustez - Alta disponibilidade e tolerância a falhas representam a capacidade de o sistema minimizar o tempo de inatividade face a uma falha inesperada.

Existe uma vasta gama de sistemas MOM. Nos subcapítulos seguintes são descritos com mais detalhe alguns dos mais populares sistemas MOM, classificados na categoria de *open source*.

2.2.1 ActiveMQ

ActiveMQ [7] é um dos mais antigos sistemas MOM disponibilizados como *open source*. Foi desenvolvido com o intuito de implementar as especificações do Serviço de Mensagens Java (JMS - *Java Message Service*). A API JMS facilita a reescrita de clientes para sistema MOM do tipo P2P e *Pub/Sub*. ActiveMQ é uma solução compacta e altamente configurável. Facilita a transmissão de mensagens entre sistemas MOM, garante tolerância a falhas, balanceamento de cargas, escalabilidade, suporte transacional, configuração avançada de *clustering*, durabilidade, persistência, etc. O seu ponto forte é sem dúvida a vasta gama de recursos que disponibiliza, a interoperabilidade entre sistemas, a compatibilidade com vários protocolos e linguagens de programação e ainda a facilidade na configuração avançada do sistema.

2.2.2 ZeroMQ

Tal como indica o seu nome, o sistema ZeroMQ [8] foi desenvolvido com o objetivo de proporcionar latência zero na transmissão de mensagens entre os seus clientes. Destaca-se por ser extremamente rápido, leve e de fácil implementação. Outra característica que o faz destacar das demais soluções é o facto de disponibilizar uma extensão da tecnologia *sockets*. Para garantir uma comunicação assíncrona, um *socket* adicional é implementado nas filas de mensagens. Outra das suas particularidades é a incorporação de um elemento de segurança, denominado CURVE, disponibilizado na sua última versão. Segundo os recentes testes desenvolvidos em [9], no contexto de segurança de edifícios inteligentes, esta nova abordagem parece bastante prometedora. O sistema apresenta uma boa performance no transporte de um grande volume de dados, conforme testes desenvolvidos em [10]. ZeroMQ é compatível com vários protocolos, plataformas e linguagens (e.g., C, C++, C#, JAVA, Python, etc.).

Funciona nos modelos P2P e *Pub/Sub*, e é possível de ser implementado de forma independente à existência de um broker. Os seus pontos menos positivos residem no facto de não implementar durabilidade e exigir algum trabalho adicional na implementação de roteamentos mais complexos.

2.2.3 RabbitMQ

O sistema RabbitMQ [11] sobressai pala conjugação da sua performance com o seu alto nível de configuração, confiabilidade e interoperabilidade com plataformas, protocolos e linguagens de programação. Toda a sua arquitetura é sustentada pela implementação do Exchange. O Exchange pode ser configurado como *Fanout* (i.e., as mensagens são entregues a todas as filas que a ele estão ligadas), *Direct* (i.e., apenas entregará dados às filas que contiverem a mesma *routing key* que a mensagem) ou como *Topic* (i.e., apenas entregará dados às filas cuja expressão regular corresponda à expressão definida na *routing key* da mensagem). Assim, é possível refinar os critérios de filtragem de mensagens bem como implementar encaminhamentos de *broadcast*, *unicast* e *multicast*. Outros pontos fortes são: um *Load Balancer* implementado de raiz e uma interface gráfica que facilita a gestão e monitorização do sistema. É compatível com vários protocolos, plataformas e linguagens de programação (i.e., C, C++, C#, JAVA, *JavaScript*, *Python*, *Node.js*).

2.2.4 Apache Kafka

Kafka [12] – Kafka foi projetado para resolver o problema de performance na transmissão de um grande volume de dados, garantindo uma elevada taxa de transferência sem comprometer a persistência e a tolerância a falhas. Com um design exclusivo, conforme mostra Figura 2.3, Kafka é um sistema distribuído de mensagens robusto, particionado e replicado. Este sistema introduz conceitos inovadores como:

- Particionamento por tópicos, i.e., cada tópico é subdividido por várias partições para onde são enviadas as mensagens recebidas pelos subscritores.;
- Deslocamento de mensagens, i.e., cada mensagem armazenada numa partição recebe um identificador incremental e único, o que possibilita aos subscritores se deslocarem nas partições e acederem a mensagens mais antigas;
- Partições líderes e replicadas, i.e., as partições são distribuídas de forma balanceada por vários *brokers*. Cada *broker* é repassável por uma ou mais partições e pela sua replicação para outros *brokers*. Quando um *broker* falha as partições que liderava passam para a responsabilidade de outro *broker* que detenha uma cópia sincronizada das mesmas. O envio e receção de mensagens é um processo exclusivamente da responsabilidade dos líderes das partições;
- Grupo de consumidores, i.e., um conjunto de consumidores que desempenham a mesma tarefa com as mensagens recebidas de um tópico/partição. Assim, neste caso, apenas um dos

consumidores do grupo irá receber a mensagem para dar seguimento à referida tarefa. Vários grupos e vários subscritores independentes podem subscrever o mesmo tópico/partição.

Kafka tem o seu próprio protocolo de mensagens, mas disponibiliza várias APIs para a interoperabilidade com outros protocolos. É compatível com várias linguagens de programação o que lhe concede a flexibilidade de se redesenhar e se ajustar às necessidades dos seus clientes. Os seus pontos menos positivos estão relacionados com a durabilidade das mensagens nos tópicos, apenas configurável por período de tempo. Para alem disso, Kafka não é disponibilizado como uma solução totalmente empacotada, i.e., necessita do Apache *Zookeeper* para a instanciação e coordenação de todo o serviço. No entanto, *Zookeeper* faculta a descoberta de serviços para aplicações distribuídas de alta performance e é responsável pela gestão do estado e balanceamento dos nós da arquitetura kafka. Outro ponto menos positivo reside na ausência de uma consola com boa usabilidade de forma a facilitar a gestão e monitorização de todo o sistema.

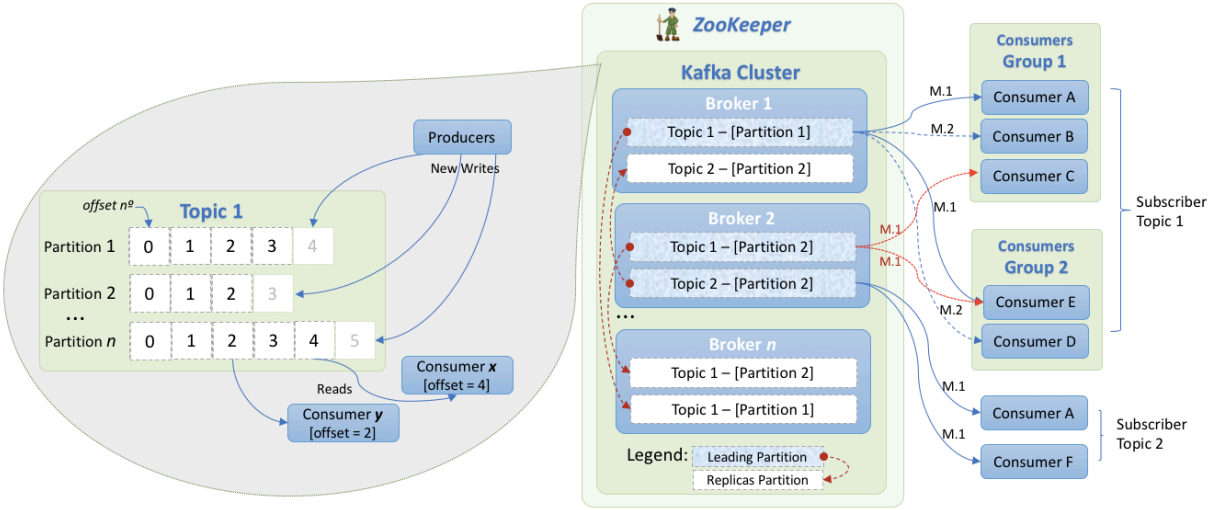


Figura 2.3 - Ecossistema do sistema Kafka

2.2.5 Sistemas MOM - Conclusão

Todos os sistemas descritos nas subsecções anteriores apresentam compatibilidade com os protocolos AMQP, MQTT, STOMP e XMPP. São compatíveis com os modelos P2P e Pub/Sub e apresentam interoperabilidade com outras plataformas. Permitem de uma forma mais ou menos indireta (i.e., através da disponibilidade de APIs) a implementação adicional de muitas funcionalidades (e.g., escalabilidade, elasticidade, resiliência, garantias de entrega, garantias de ordenação, durabilidade, persistência, transações, etc.). Por outro lado, todos os sistemas descritos fundamentam-se nos mesmos objetivos e propósitos, i.e., distribuição de mensagens de forma desacoplada. No entanto, diferem relativamente aos princípios sobre os quais as suas arquiteturas se apoiam.

Benchmarks recentes sobre sistemas MOM são raros. Contudo, *Benchmarks* recentes e contínuos são de extrema importância para o processo de seleção e adoção de soluções deste tipo, visto que estes sistemas

estão continuamente em evolução. Por outro lado, novas soluções vão surgindo, como, por exemplo, o sistema AmazonMQ disponibilizado recentemente pela empresa Amazon [13], com o propósito de facilitar a interoperabilidade com a ActiveMQ. Em [14] o autor compara o desempenho do sistema RabbitMQ com o Apache Kafka, e no trabalho apresentado em [15] são executados vários testes que comparam Kafka com o RabbitMQ configurado com o protocolo AMQP. Ambos os autores chegam à mesma conclusão, i.e., Kafka é mais indicado quando é necessária a transferência dum grande volume de dados, que implica uma alta taxa de transferência, e quando a persistência e durabilidade das mensagens são um fator critico. Por outro lado, quando é necessária a implementação de transações, filtragens e roteamentos complexos, o RabbitMQ será a opção mais adequada. Num trabalho realizado para a execução de tarefas MTC (*many-task computing)* configurado num cluster Hadoop, os autores comparam Kafka a ActiveMQ e concluem que ActiveMQ superou Kafka no balanceamento de cargas [16]. Em [2] o autor faz uma pesquisa acerca dos sistemas ActiveMQ e Kafka, concluindo que não existem sistemas capazes de resolver todos os problemas inerentes aos sistemas operacionais. A escolha por um sistema MOM deve ser feita tendo em conta quer a especificidade de cada sistema, quer analisando os casos de uso que se pretende solucionar com a sua implementação. Partilha-se da mesma opinião sugerida pelo autor em [14] que propõe o uso combinado dos sistemas MOM, de forma a resolver um maior número de casos de uso. Outros exemplos de uso compartilhado destes sistemas podem ser observados em [17] onde o autor propõe um sistema de comunicação de mensagens baseadas na implementação dos protocolos AMQP (via RabbitMQ) e MTTQ (via Mosquitto [18]), e em [19] onde o autor apresenta uma solução robusta baseada no sistema RabbitMQ com a implementação combinada dos protocolos AMQP e MTTQ.

2.3 Processamento de dados

Tradicionalmente o processamento de dados era única e exclusivamente suportado pelo denominado processamento em lote (denominado em inglês por "*batch processing*"). No processamento em lote os dados são recolhidos, agrupados e armazenados de forma integrada para posteriormente serem processados em conjunto, numa única operação. Este tipo de processamento é recomendado para previsões a médio e longo prazo, análise e exploração de novas metodologias, processamento operacional de suporte empresarial a curto prazo, etc. No entanto, esta nova era digital é caracterizada pela implementação crescente de uma vastíssima gama de hardware e software, pela presença de milhões e milhões de sensores e dispositivos digitais que continuamente geram dados a uma cadencia continua, infinita e muitas vezes de forma desordenada. Recolher e processar estes fluxos de dados (denominado em inglês por "*streaming data*") é hoje um desafio extremamente complexo. As necessidades impostas pelas atividades de negócio, que apelam à tomada de decisão em tempo real, impulsionaram e motivaram o desenvolvimento de novas abordagens na área de processamento, conforme se descreve:

- *Processamento micro-lote* (denominado em inglês por *"micro-batch processing"*) – nesta abordagem os dados são processados em lotes muito pequenos de forma frequente em reduzidos intervalos de tempo;
- *Processamento de fluxos* (denominado em inglês por *"stream processing"*) – esta abordagem os dados são processados imediatamente após chegarem ao sistema. Apesar de visar o processamento individual dos dados, a maioria destes sistemas suportam operações denominadas de "Janelas" (conforme conceito descrito detalhadamente na secção 2.3.1), para facilitarem o processamento de fluxos contínuos e desordenados.

O processamento de fluxo, conforme já referido, é apropriado e necessário para casos de uso que exigem dados constantemente atualizados e respostas em tempo real (e.g. monitorização de equipamentos, deteção de fraudes, análises preditivas em tempo real, etc.). Estes sistemas implementem uma arquitetura orientada a eventos, uma arquitetura na qual o fluxo de trabalho é continuamente monitorizado. Estes fluxos são muitas vezes denominados por fluxos de eventos [20].

Para uma melhore perceção sobre os principais conceitos que envolvem os sistemas de processamento, a seguir descreve-se a sua evolução, onde se realça o sistema Apache Beam, justificada pala adoção dos seus conceitos na implementação da plataforma desenvolvida no presente projeto (que se encontra descrita mais detalhadamente no capitulo 5.3.2). Adicionam-se ainda, referências a trabalhos efetuados nesta área de investigação, bem como as mais recentes soluções disponibilizadas quer *open source* quer comercialmente.

2.3.1 Evolução dos sistemas de processamento de dados

De forma a sintetizar os principais conceitos que caracterizam os sistemas de fluxos de dados, a seguir são descritas algumas das abordagens mais populares e que mais têm contribuído para a evolução dos sistemas de processamento.

MapReduce

MapReduce [21] foi a primeira abordagem proposta, pela Google, com o objetivo de resolver o problema do processamento de um grande volume de dados. Baseada no processamento em lote, destacou-se pela sua simplicidade e escalabilidade. Conforme mostra a Figura 2.4, a metodologia desta abordagem consiste em dividir e distribuir a tarefa de processamento por múltiplos nós (i.e., *Split*). Cada nó executará parte da tarefa (i.e., *Map*). Os resultados obtidos em cada nó são ordenados e conjugados (i.e., *Shuffle*) e resumidos (i.e., *Reduce*). Finalmente, a partir deste resumo será gerado o resultado.

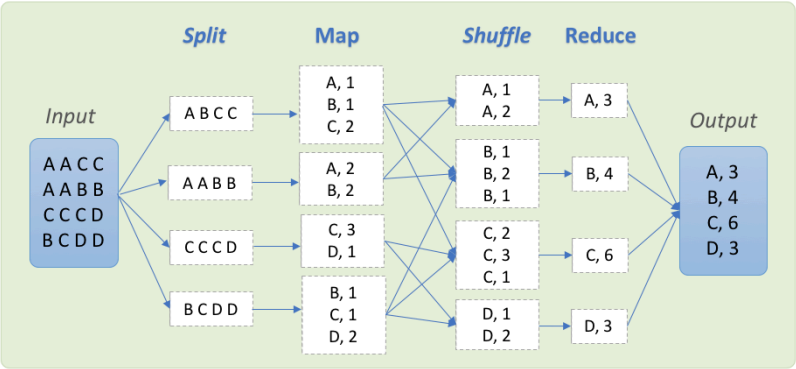


Figura 2.4 - Processamento MapReduce: exemplo na contagem de caracteres

Mais tarde, Doug Cutting and Mike Cafarella, adicionaram uma camada de processamento, baseada em *MapReduce*, sobre um sistema distribuído de ficheiros, denominado por *Hadoop Distributed File System* (HDFS). Assim nasceu o sistema *Hadoop*, designado como sendo a primeira Framework disponibilizada com código aberto na área de Big Data.

FlumeJava

Esta abordagem [21], [22], igualmente denominada por Flume (e que não deve ser confundido com o Apache Flume), foi desenvolvida pela Google com o objetivo de superar a falta de otimização em processamentos mais complexos executados pelo *MapReduce*. Flume é designado como sendo um pipeline de alto nível que permite a otimização automática de tarefas complexas.

Apache Storm

Apache Storm [21], [23] foi desenvolvido com o principal objetivo de proporcionar baixa latência no processamento de dados. Apesar da sua fraca consistência, é considerada como sendo a primeira abordagem puramente baseada em streaming. Storm está na origem do surgimento e ascensão da arquitetura Lambda (i.e., arquitetura híbrida baseada no processamento em lote e em memória). Tendo em conta as limitações do Storm (i.e., fraca consistência), a arquitetura Lambda consistia em conjugar o processamento do Storm com o processamento fortemente consistente do MapReduce.

Spark Streaming

Spark Streaming [21], [24] é um sub-projeto do Apache Spark. Esta abordagem foi desenvolvida com o objetivo de garantir baixa latência e forte consistência no processamento de fluxos, sem depender da adição de outras abordagens de processamento em lote. O Spark consegue executar pequenas tarefas totalmente em memória. No entanto, quando os trabalhos são muito pesados para a capacidade de memória do sistema, o Spark recorre à escrita de dados em disco, o que provoca perda de performance em termos de latência. O Spark baseia-se no conceito de micro-batch para o processamento de fluxos de dados contínuos e ilimitados. O micro-batch consiste no agrupamento de pequenas partes de um fluxo repartido por unidades de tempo. Estes micro-batch, denominados *Resilient Distributed Dataset* (RDD),

são distribuídos e processados pelos vários nós do sistema. Spark é considerado mais adequado para o processamento de fluxo de dados ordenados. Para alem disso, é completamente dependente da adição de um sistema de ficheiros distribuídos;

MillWheel

MillWheel [21] foi desenvolvido pela Google com o objetivo de garantir baixa latência, forte consistência e processamento de fluxos fora de ordem. Com esta abordagem foi introduzido o conceito importantíssimo para o processamento de fluxo de dados ilimitados e fora de ordem, denominado por Marca de Água (do termo em inglês, *Watermark*). Conforme mostra a Figura 2.5, *watermark* tenta definir o tempo que decorre entre o momento em que os dados foram gerados e o momento em que chegam ao sistema de processamento. A *watermark* pode ser determinada manualmente ou através de heurísticas.

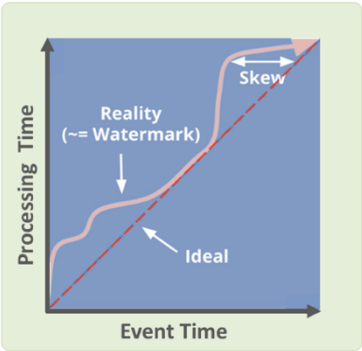


Figura 2.5- Watermark [21]

Cloud Dataflow

Cloud Dataflow [21], [25] foi desenvolvido pela Google com o objetivo de disponibilizar em nuvem uma solução unificada para o processamento de fluxo de dados, com as melhores funcionalidades dos seus projetos desenvolvidos até então (i.e., MapReduce, FlumeJava e MillWheel).

Apache Flink

Apache Flink [21], [26] foi desenvolvido pela Apache com o objetivo de proporcionar o processamento de fluxos fora de ordem. Destaca-se principalmente pela sua robustez ao adicionar o conceito de pontos de salvaguarda (i.e., *snapshotting*) que permite a reconstrução completa de um pipeline. É ainda caracterizado por permitir um processamento altamente escalável, com alto desempenho e tolerante a falhas. No entanto, tal como o Spark, depende de um sistema distribuído de ficheiros.

Apache Apex

Apache Apex [27] tal como o Flink, foi desenvolvido para proporcionar alta escalabilidade e alto desempenho. Unifica o processamento em lote com o processamento de fluxos de dados contínuos. Destaca-se por ser tolerante a falhas, seguro e facilmente operável. É compatível com várias linguagens

de programação (e.g. Java, Ruby, Python, R, etc.), no entanto, a reprogramação do pipeline só é possível com recurso a Java. Disponibiliza ainda uma biblioteca de operadores, denominada *Malhar* [28], que facilita a criação de aplicações mais complexas. *Malhar* também disponibiliza muitos conectores o que facilita a interoperabilidade do Apex com sistemas de ficheiros, bases de dados, sistemas de mensagens, etc. Tal como o Spark e o Flink, o Apex é dependente do suporte de uma plataforma de armazenamento.

Apache Beam

Apache Beam (AB) [21], [29] nasceu recentemente do esforço conjunto da Google e da Apache Foundation. O sistema AB contém as melhores funcionalidades desenvolvidas por estas duas organizações, no que respeita ao processamento de fluxo de dados. A grande premissa desta abordagem é: "*escrever uma vez e executar em qualquer lado*". É compatível com várias linguagens de programação (e.g., Python, Java, Scala, Go, etc.). Outro ponto forte desta abordagem é a sua interoperabilidade com outros sistemas de streaming (e.g., Apache Spark, Apache Flink, Apache Apex, DataFlow, etc.). Por se tratar de uma solução que poderá contribuir de forma significativa para o desenvolvimento da área de processamento de fluxos de dados, a seguir aprofunda-se os principais conceitos em que se fundamenta este sistema:

- *Estruturas – PColection* e *PTransform* são as estruturas principais do AB. O AB paraleliza automaticamente um fluxo de dado, denominado de *PColection* e executa sobre ele transformações possíveis de serem configuradas com recurso ao *PTransform* (que pode ser do tipo: *Element-wise*, *Aggregating* ou *Composite*).
- *Windowing* – permite definir a subdivisão dos fluxos de dados contínuos e infinitos em janelas temporais. Existe 3 tipos de abordagem para esta divisão, conforme mostra Figura 2.6.

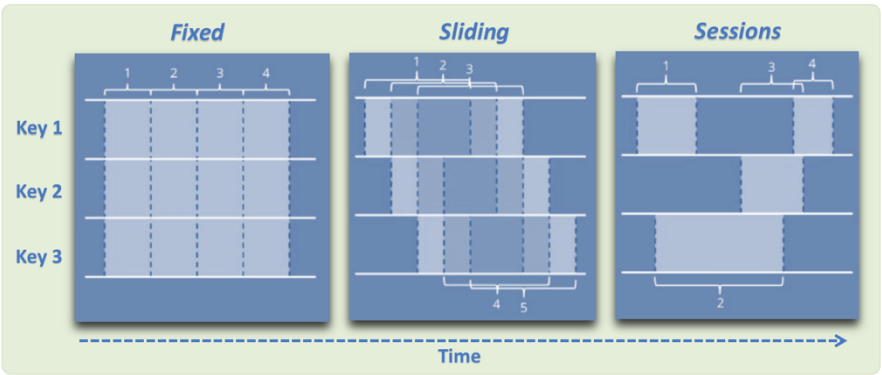


Figura 2.6 - Estratégias de Windowing [21]

Os exemplos mais simples desta estratégia são as janelas fixas (i.e., janelas de tamanho fixo ou variável) e as janelas deslizantes (i.e., janelas que se sobrepõem). As estratégias mais sofisticadas consistem na definição de sessões onde os dados são agrupados de acordo com determinado contexto;

- *Transformations* - Permite definir relações entre os elementos de uma janela temporal. As Transformações são definidas pela estrutura *PTransform.*

- *Watermarks* – conforme mostrado na (Figura 2.5), *watermarks* consiste numa linha temporal que tenta determinar o tempo que decorre entre a geração de um evento e o momento em que este chega ao sistema para ser processado. No caso do AB esta métrica é calculada por heurísticas proprietárias do próprio sistema e não é possível alterá-la. O risco de processar um fluxo de acordo com uma marca de água heurística é o facto de alguns elementos não serem considerados no processamento por chegarem tardiamente ao sistema (i.e., conforme mostra a Figura 2.7, o elemento '9' não irá ser processado na janela temporal 12:00-12:06);

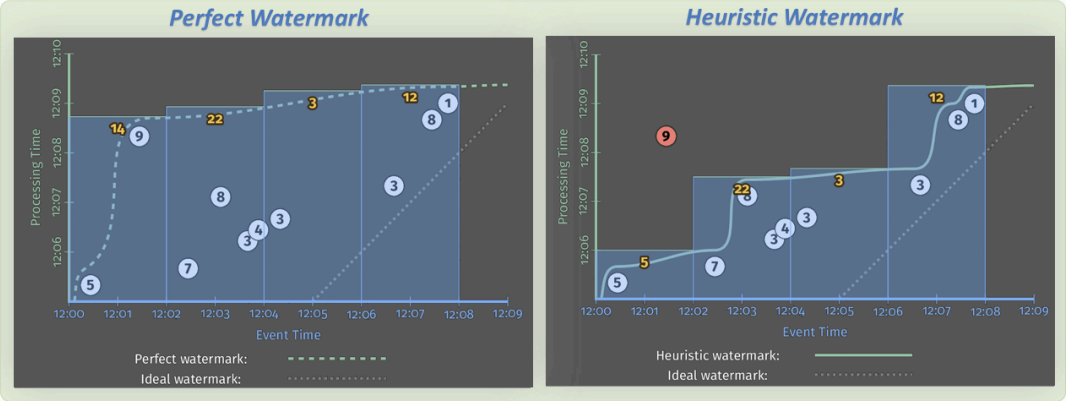


Figura 2.7- Diferença watermark perfeita (à esquerda) e watermark heurística (à direita) [21]

- *Triggers* – Os triggers permitem redefinir o momento em que uma janela temporal deve ser processada. É um método que permite efetuar correções de forma a ser possível processar dados que chegam tarde em relação a uma marca de água heurística. Supondo que o fluxo de dados apresentado na Figura 2.7 requer um processamento consistente e/ou não é exigente em termos de baixa latência, então é possível configurar um *trigger* para que a janela dos dados gerados entre as 12:00 e 12:02 só seja processada após a chegada de todos os elementos. Assim, desta forma o elemento '9' seria processado. Os *triggers* permitem várias configurações, e.g., definir quanto tempo uma janela pode ficar à espera de elementos atrasados para gerir de forma mais eficiente os recursos de memória do sistema, definir que uma janela deve ser reprocessada de x em x tempo para captar valores em atraso, etc.;

- *Accumulation* – Permite definir como os resultados obtidos parcelarmente por cada janela devem ser tratados (i.e., acumulados ou descartados). Os valores parcelares obtidos pelas janelas são armazenados parcialmente nos denominados *Panes*. Uma janela pode ter um ou vários *Panes*. Uma janela tem várias *Panes* quando por exemplo é configurado um *trigger* que define o reprocessamento de uma janela para evitar a perda de valores atrasados.

De acordo com conceitos descritos, a configuração de um pipeline no AB consiste em responder às seguintes perguntas:

- *O quê?* – Consiste em definir quais os fluxos a ser processados e quais transformações devem ser operadas sobre eles;
- *Onde?* – Consiste em definir as janelas temporais para o processamento dos fluxos;
- *Quando?* – Consiste em configurar as regras que definem quando uma janela contém todos os dados esperados para ser processada, i.e., pode ser utilizada a marca de água ou refinar esta definição com a utilização dos *Triggers*;
- *Como?* – Permite definir se os valores obtidos parcelarmente por cada janela devem ser acumulados ou descartados.

Alguns exemplos do uso desta abordagem podem ser analisados em [30] onde é proposto um *pipeline* para a monitorização de tráfego de veículos, e em [31] para a gestão de uma frota de táxis. Uma das grandes vantagens do AB, é permitir de forma simples, programar e ajustar um *pipeline* de acordo com as necessidades operacionais e de acordo com os recursos computacionais disponíveis. Outra grande vantagem, é facilitar o desenvolvimento de excelentes *benchnmarks* na área de *streaming*, devido à sua interoperabilidade com vários sistemas e à sua premissa "escrever uma vez e corre em qualquer lado".

2.3.2 Soluções e trabalhos de investigação dos sistemas de processamento

Conforme já referido neste subcapítulo destacamos os trabalhos mais recentes na área de processamento, ao qual se adiciona referência a soluções comerciais e *open source*.

O aumento significativo de soluções propostas na área de processamento de fluxos de dados desencadeou na comunidade científica o interesse pelo desenvolvimento de ferramentas capazes de as avaliar. Trabalhos recentes realizados na área de *benchmarks* para a avaliação destes sistemas, podem ser analisados em [32]–[35]. No entanto, *Yahoo Streaming Benchmark, SreamBench* e *Linear Road,* continuam a ser algumas das mais populares soluções propostas para o *Streaming Benchmark* [20], [35]. Em [36], [37] os autores desenvolveram e apresentaram outras abordagens a fim de avaliar a performance destes sistemas, chegando às mesmas conclusões, i.e., não existe um vencedor, cada sistema tem vantagens e desvantagens. Os sistemas de *streaming* têm características particulares que os tornam mais adequados em determinados contextos. Ou seja, o desempenho não é a única característica a considerar na escolha de um sistema. O mais importante é ter em conta todas as suas características (e.g., consistência, tolerância a falhas, escalabilidade, interoperabilidade, qual o seu comportamento mediante determinado contexto, etc.). Este e outros trabalhos desenvolvidos no âmbito de sistemas de *streaming* encontram-se sumariados na Tabela 2.1.

Tabela 2.1 - Trabalhos recentes realizados na área de sistemas de *streaming*

Ano	***Trabalho de investigação***
2018	*"Recent Advancements in Event Processing" [20]* Compara as características e funcionalidades das plataformas de computação de fluxos distribuídos (Storm, Spark Streaming, Samza, Flink, Apex).
2018	*"Benchmarking Distributed Stream Processing Engines" [36]* É proposta uma framework para avaliação dos sistemas Apache Storm, Spark e Flink. A avaliação é feita em relação à latência e ao rendimento dos sistemas. O autor conclui que não há um único vencedor.
2018	*"Exactly-Once Semantics with Real-Time Data Pipelines" [38]* É proposta uma solução para execução de fluxos complexos e fora de ordem, garantindo baixa latência. Esta arquitetura é baseada no Spark Streaming, mas adaptável a outros sistemas, e.g. Flink.
2018	*"Lambda-Based Data Processing Architecture for Two-Level Load Forecasting in Residential Buildings" [39]* É proposta uma arquitetura híbrida, baseada em *Spark Streaming*, para as previsões, a curto prazo e em tempo real, do consumo de energia em edifícios.
2018	*"A Distributed Online Learning Approach for Pattern Prediction over Movement Event Streams with Apache Flink" [40]* Proposta de um modelo de previsão para a distribuição de múltiplos fluxos, baseado em Apache Flink.
2018	*"A hybrid approach for alarm verification using stream processing, machine learning and text analytics" [41]* Solução proposta para a monitorização e validação de alarmes em tempo real. A solução proposta tem por base o sistema Streaming Spark.
2018	*"Scotty: Efficient Window Aggregation for out-of-order Stream Processing" [42]* É proposto o uso do operador *Scotty* para a gestão de janelas e processamento eficiente de fluxos fora de ordem.
2017	*"RIoTBench: An IoT benchmark for distributed stream processing systems" [32]* *Benchmark* "RIoTBench", para avaliação de sistemas de processamento de fluxo distribuído em aplicações IoT. A solução foi testada com o Apache Storm e com quatro fluxos de dados reais e distintos (redes inteligentes, transporte inteligente, deteção urbana e domínios de aptidão pessoal na IoT).
2017	*"Pilot-Streaming: A Stream Processing Framework for High-Performance Computing" [33]* Propõem um sistema combinado com as frameworks *Pilot-Streaming* [39] *e Streaming Mini-Apps* [40] para a avaliação do processamento de fluxo em *High Performance Computing* (HPC).
2017	*"Extremely Fast Decision Tree Mining for Evolving Data Streams" [34]* Apresenta um sistema, denominado "STREAMDM-C ++", desenvolvido em C ++ e baseado na metodologia de árvore de decisão, para a avaliação de sistemas de streaming.
2017	*"A New Application Benchmark for Data Stream Processing Architectures in an Enterprise Context: Doctoral Symposium" [35]* Proposta de um *benchmark* para análise de aplicativos de processamento de fluxo de dados no contexto da *Industrial Internet of Things* (IIOT)
2017	*"An IoT ecosystem for the implementation of scalable wireless home automation systems at smart city level" [43]* Proposta de um sistema baseado em Spark Streaming para auditoria de energia em tempo real no contexto da casa inteligente.
2017	*"Low Latency Stream Processing: Apache Heron with Infiniband & Intel Omni-Path" [44]* Proposta de um sistema baseado em *Spark Streaming* para auditoria de energia em tempo real no contexto da casa inteligente.
2017	*"Large-Scale Data Stream Processing Systems" [45]* Explora os recursos do sistema Apache Flink, seus desafios e oportunidades.
2017	*"Kafka interfaces for composable streaming genomics pipelines" [46]* Proposta de um sistema baseado em kafka Streaming, no contexto da bioinformática.

Ano	*Trabalho de investigação*
2016	*"A survey of systems for massive stream analytics" [47]* Comparação de recursos e apresentação de casos de uso (Storm, Spark Streaming e *Simple Scalable Streaming System* (S4))
2016	*"Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming" [37]* Benchmark, onde são comparados o desempenho de sistemas de streaming (Storm, Flink e Spark) em relação à latência e à taxa de transferência.

Como conclusão desta pesquisa sobre sistemas de processamento sumariza-se na Tabela 2.2 algumas das mais populares e recentes soluções nesta área.

Tabela 2.2- Frameworks para o processamento de dados

	ORGANIZAÇÃO	*PRODUTO*	*DESCRIÇÃO*	*WEBSITE*
Open Source	Apache Foundation	Apex	Stream and batch processing on YARN	http://apex.apache.org
	Apache Foundation	Beam	Programming model for batch and streaming data processing	https://beam.apache.org
	Apache Foundation	Flink	Streaming dataflow engine for Java	http://flink.apache.org
	Apache Foundation	Flume	Streaming data ingestion for Hadoop	http://flume.apache.org
	Apache Foundation	Samza	Distributed stream processing framework	http://samza.apache.org
	Apache Foundation	Spark Streaming	Discretized stream processing with Spark's RDDs	http://spark.apache.org/streaming/
	Apache Foundation	Storm	Distributed realtime streaming	http://storm.apache.org
	Google	Cloud Dataflow	Streaming dataflow	https://cloud.google.com/dataflow/
Commercial	Amazon Web Services	Amazon Kinesis	Stream data ingestion, storage, query, and analytics PaaS	https://aws.amazon.com/pt/kinesis/
	Confluent	Confluent Platform	Data integration and streaming data platform based on Kafka Streaming	https://www.confluent.io/product/confluent-platform/
	DataTorrent	DataTorrent RTS	Stream and batch application development platform, based on Apache Apex	http://docs.datatorrent.com/rts/
	Hortonworks	Hortonworks DataFlow (HDF)	Provides the only end-to-end platform that collects, curates, analyzes and acts on data in real-time, on-premises or in the cloud	https://hortonworks.com/products/data-platforms/hdf/
	IBM	IBM Streams	Software platform for enables continuous and fast analysis of massive volumes and moving data.	https://console.bluemix.net/catalog/services/ streaming-analytics
	IBM	IBM Streaming Analytics	Streaming data application development and analytics platform in Cloud	https://console.bluemix.net/catalog/services/ streaming-analytics
	Informatica	Big Data Streaming	Enable real-time analytics with fast, reliable, codeless stream processing	https://www.informatica.com/products/big-data/big-data-streaming.html#fbid=tQBWUBE7ipm
	MapR	MapR Event Streams	Global publish-subscribe event streaming system	https://mapr.com/products/mapr-streams/
	Microsoft	StreamInsight	Platform for develop and deploy complex event processing (CEP) applications	https://msdn.microsoft.com/en-us/library/ee362541(v=sql.111).aspx
	Pivotal	Spring Cloud Data Flow	A toolkit for building data integration and real-time data processing pipelines	https://cloud.spring.io/ spring-cloud-dataflow/
	SAP	SAP HANA	In-memory data platform, real-time analyzing of a wide variety of data, including live transactions, text, spatial and streaming data all with high-speed performance	https://www.sap.com/products/hana.html

2.4 Armazenamento de dados

O Armazenamento de Dados é uma área que se dedica à exploração de abordagens capazes de solucionar os grandes desafios inerentes à persistência dos dados nas plataformas Big Data. Como já abordamos nas subsecções anteriores, a persistência é uma característica muito importante nos sistemas de comunicação e processamento dos dados, mas não só. As primeiras abordagens na área de armazenamento eram em grande parte sustentadas pelos sistemas de base de dados relacionais. Devido à sua "rigidez" em termos de modelação, que de certa forma impediam a obtenção de altas taxas de performance, característica exigida nas plataformas Big Data, outras abordagens foram sendo propostas conforme se descreve nas subsecções seguintes.

2.4.1 NoSQL

NoSQL significa "Not Only SQL", ou seja, é um termo para descrever base de dados não relacionais caracterizadas pelo seu alto desempenho, escalabilidade horizontal, alta disponibilidade e resiliência. São soluções baseadas em arquiteturas de memória distribuídas com replicação e segmentação dos dados por vários clusters. Estas características permitem-lhe dar suporte a um grande número de operações I/O (i.e., leitura, escrita) executadas ao segundo [48], [49]. Ao contrário das Bases de Dados Relacionais (DBRs), a maioria das soluções NoSQL não garantem as propriedades ACID (Atomicidade, Consistência, Isolamento, Durabilidade). Em vez das propriedades ACID esta nova abordagem é caracterizada pelo teorema CAP (i.e., Consistência, Disponibilidade, Tolerância ao particionamento) em conformidade com o princípio BASE (i.e. Basicamente Disponível - *Basic Availability*, Estado Leve - *Soft-state* e Eventualmente Consistente - *Eventual Consistency*). O teorema CAP foi proposto por Eric Brewer [50] que define as suas propriedades como:

- (C) Consistência (*Consistency*): um sistema diz-se consistente sempre que após determinada atualização de um registo, todas as operações que acedam a esse registo terão como resultado a mesma versão;
- (A) Disponibilidade (*Availability*): refere-se à capacidade do sistema se manter disponível para a execução das operações durante determinado período;
- (P) Tolerância ao particionamento (*Partition tolerance*): refere-se à capacidade de um sistema continuar operacional mesmo na presença de uma falha de rede (i.e., uma falha na ligação entre os seus nós).

Segundo este teorema é muito difícil garantir todas as 3 propriedades em simultâneo. Ou seja, para serem garantidas 2 das suas propriedades é obrigatoriamente necessário relaxar uma terceira (i.e., o princípio BASE). As combinações possíveis são CP, AP e CA. Apesar desta abordagem ser fortemente caracterizada pelo teorema CAP e pelo princípio BASE, não impede que sejam implementadas as

propriedades ACID. Com algum esforço de programação estas propriedades poderão ser garantidas pela camada operacional que faz uso do sistema de armazenamento.

Os modelos mais comuns que caracterizam este tipo de armazenamento são: *Document, Key-Value, Column e Graph,* conforme mostra Figura 2.8 e a seguir se descreve:

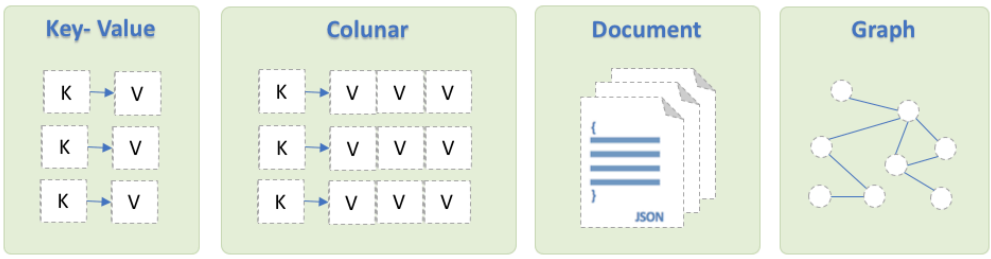


Figura 2.8 - Os quatro modelos mais comuns em armazenamento NoSQL

- *Document data model* - este modelo permite o armazenamento dos dados em coleções de documentos geralmente em formatos como JSON, JOSM, XML, YAML, JSON, BSON. Cada documento possui uma chave única e uma estrutura flexível (i.e., pode possuir diferentes campos de dados sem dependerem de estruturas pré-definidas). Em analogia com as bases de dados relacionais pode-se dizer que uma coleção representa uma tabela, um documento representa os registos da tabela (i.e., as linhas) e os campos contidos no documento (como referido, definidos de forma flexível) representam as colunas da tabela. As consultas a qualquer campo contido nos documentos podem ser executadas facilmente. Este modelo prima pela sua alta flexibilidade.
- *Key-value data model* - é provavelmente o modelo NoSQL mais simples e mais fácil de implementar. Baseiam-se no armazenamento de valores indexados para posteriormente serem acedidos através de uma chave. São caracterizadas pelo seu elevado desempenho e escalabilidade. No entanto, não são adequadas para aplicações demasiado complexas. Geralmente fornecem mecanismo de persistência e funcionalidades adicionais (e.g., replicação, controle de versão, bloqueio, transações, etc.)
- *Column data model* – neste modelo o armazenamento dos dados é orientado às colunas. As principais técnicas utilizadas consistem no particionamento (horizontal e vertical entre vários nós), indexação, compressão (i.e., capacidade de organizar informações semelhantes de forma contigua). Operações de I/O têm um excelente desempenho. São geralmente mais eficientes para grande volume de registos com poucas colunas.
- *Graph data model* – este modelo emprega conceitos da teoria de grafos do tipo multigrafo (também denominados por pseudografo) para a representação do relacionamento entre diferentes conjuntos de dados. Os vértices (i.e., os nós) e as arestas contêm as propriedades (i.e., atributos) dos dados a armazenar. Dois vértices podem ser interligados por várias arestas. Esse

modelo é vantajoso para atender a aplicações que requerem consultas complexas entre os vários níveis de relacionamento dos dados.

Apesar de a classificação atrás referida ser a mais comum para agrupar os modelos baseados em BDs NoSQL, devido à especificidade de determinados tipos de dados, recentemente têm sido desenvolvidas e propostas novas abordagens como [51]–[53]:

- *Time series data base* (TSDB) – estes modelos são otimizados para a escrita de um grande volume de registos organizados pelo tempo. São indicadas para o armazenamento de dados recolhidos em real time de variadíssimas fontes como e.g. de sensores, RFIDs, Smart Meters, etc. Apesar de outros modelos serem igualmente recomendados para este tipo de dados, (e.g. relacionais, baseados em Key-Valeu, colunas, etc.) estas bases de dados são especializadas no tratamento de dados temporais de forma a obterem grande performance;
- *Native XML DBMS* (NXD) – este modelo pode ser considerado um subconjunto do modelo orientado a Documentos. Neste caso específico, os documentos são nativamente escritos em XML. Apesar da sua especificidade, podem eventualmente suportar outros formatos. Utiliza linguagens de consulta específicas como XPath, XQuery ou XSLT;
- *RDF Storage* – este modelo foi concebido para dar suporte ao armazenamento de dados do tipo *Resource Description Framework* (RDF). RDF é uma metodologia para modelar e descrever informação no formato de triplos (i.e., um facto é expresso sob a forma de 3 componentes: sujeito, predicado e objeto). A informação do tipo RDF é frequentemente utilizada em web semântica. Este modelo baseia-se na teoria de Grafos em que o nó origem contém informação sobre o sujeito, o nó destino informação sobre o objeto e a propriedade contida na aresta representa o predicado;
- *Object storage* – estes modelos visam o armazenamento de dados completamente destruturados (e.g. imagens, vídeos, *streams* de áudio, vários tipos de ficheiros, etc.). Os objetos são identificados por IDs a partir dos quais é possível efetuar consultas com grande desempenho.

Existem mais de 300 soluções para o armazenamento de dados, das quais mais 200 estão classificadas como NoSQL. O grande crescimento que se tem verificado no desenvolvimento desta abordagem e na sua adoção deve-se em grande parte ao seu bom desempenho na resposta a algumas operações organizacionais e ao seu custo reduzido, ainda que muitas vezes puramente ilusório. A escolha por um modelo e por uma solução é uma tarefa extremamente difícil devido à especificidade de cada modelo e à arquitetura de cada solução [59]. Muitas questões devem ser consideradas, como por exemplo, qual o tipo de dados a armazenar e o fim a que se destinam, qual das propriedades do teorema CAP está a ser relaxada, etc. Segundo Stefan Edlich [60] mais de 50 questões devem ser respondidas antes de ser tomada uma decisão. Geralmente, uma só solução não é suficiente para responder a todas as necessidades operacionais de uma organização. Esta situação poderá conduzir à proliferação de silos de

dados. Como consequência, a organização terá um custo acrescido na administração, gestão e governação do armazenamento de dados. Por outro lado, poderão ocorrer situações em que as soluções NoSQL sejam consideradas inviáveis, face à criticidade de determinadas operações que exigem consistência e integridade transacional.

2.4.2 NewSQL

As bases de dados NewSQL são caracterizadas por oferecer o melhor dos dois mundos, i.e., a integridade, consistência e controlo de transações que caracteriza as tradicionais DBRs e a escalabilidade que caracteriza as abordagens NoSQL. Rick Cattell [49] numa comparação entre as abordagens NoSQL e NewSQL, conclui que na teoria é possível a implementação de escalabilidade nas DBRs, o que na prática lhes proporcionará vantagens sobre as soluções NoSQL, devido à simplicidade das transações SQL e às propriedades ACID. Segundo Matt Aslett [56], as soluções NewSQL podem ser classificadas do seguinte modo:

- *New Architectures* – representam o conjunto de sistemas NewSOL desenvolvidos de raiz (i.e., não são uma extensão de soluções RDBs já existentes). São baseadas em arquiteturas distribuídas com total suporte ao controlo de concorrência entre clusters, tolerância a falhas e alto desempenho. Alguns exemplos: VoltDB; NuoDB; MemSQL; Clustrix; CockroachDB; H-Store; HyPer; Google Spanner; SAP Hana.
- *Transparent Sharding Middleware* – Reporesentam o conjunto de soluções que emergiam de soluções já existentes (e.g., MariaDB que foi desenvolvida a a partir da BDR MySQL, ScaleArc; ScaleBase3, etc.). A arquitetura neste tipo de abordagem é composta por um Middleware central que comunica com as várias instâncias das RDBs implementadas em diversos clusters;
- *Database-as-a-Service* – Representa o grupo de soluções NewSQL disponibilizadas como um serviço, por provedor de computação em nuvem. Alguns exemplos: Amazon Aurora; ClearDB.

No entanto, estas soluções podem ainda ser avaliadas de acordo com as seguintes propriedades: *Main Memory Storage* (i.e., o armazenamento é feito na memória principal, o que lhes confere a característica de elevada performance); *Partitioning / Sharding* (i.e., a capacidade de subdividir a SGBD NewSQL em vários subconjuntos e os distribuir por várias partições); *Concurrency Control* (i.e., fazem uso de um protocolo para a coordenação de transações, podendo estas ser centralizadas ou descentralizadas); *Secondary indexes* (i.e., capacidade de particionar índices secundários, o que lhes garante uma maior performance, em detrimento do aumento de complexidade); *Replication* (i.e., consiste na capacidade de replicação e sincronização dos dados entre as várias partições, garantindo a consistência e integridade transacional); *Crash Recovery* (i.e., funcionalidade que lhes garante a tolerância a falhas).

2.4.3 Multi-Model

Multi-Model é uma recente abordagem que, tem como principal foco a unificação da diversidade de metodologias propostas na área de armazenamento num único sistema. Emergiu pela necessidade de retificar as grandes desvantagens apontadas aos denominados "data silos". Pete Aven e Diane Burley, visando uma solução capaz de integrar modelos NoSQL e modelos relacionais, definem Multi-Model como *"A database that supports multiple data models in their natural form within a single, integrated backend, and uses data standards and query standards appropriate to each model. Queries are extended or combined to provide seamless query across all the supported data models. Indexing, parsing, and processing standards appropriate to the data model are included in the core database product."* [57].

No entanto, conceber um sistema capaz de gerir modelos relacionais e modelos NoSQL é um problema bastante complexo e desafiante conforme referido em [58]. Os autores neste artigo propõem uma plataforma a que chamaram UDBMS (*Unified Database Management System*) e identificam os principais desafios a vencer neste tipo de abordagem: diversidade (i.e., refere-se à capacidade de integrar modelos e dados diversos); extensibilidade (i.e., capacidade de integrar todos os tipos de dados); flexibilidade (i.e., suportar a integração de todo o tipo de esquemas). No seguimento desta proposta foram executados vários trabalhos [59], [60] no sentido de validar a plataforma UDBMS. Os autores concluem que é necessário executar diversos testes, para e.g., selecionar a melhor sequência de consultas a ser executada nos diversos modelos. Relativamente às ferramentas para comparar soluções de armazenamento, estas continuam a representar um enorme desafio. As ferramentas para *benchmarks* disponíveis (e.g., YCSB, BigBench, TPCx-BB, Bigframe, etc.), apenas permitem testar sistemas específicos, e nenhum se adequa às Multi-Model.

Na análise de alguns *benchmarks* executados para a avaliação de variadas soluções de armazenamento [61]–[64], concluiu-se que os resultados apurados pelos autores são inconclusivos, e extremamente dependentes do caso de uso. Por outro lado, também se concluiu que ainda não existem soluções Multi-Model capazes de abranger por completo a diversidade de modelos de armazenamento.

2.4.4 Data Lake

Data Lake (DL) é um conceito e não uma tecnologia. O seu conceito refere-se ao armazenamento centralizado de todos os dados, recolhidos por uma organização, no seu formato original. O termo DL foi definido como, *"Data Lake, it can be defined as a vast repository of a variety of enterprise-wide, raw information that can be acquired, processed, analyzed and delivered"* [65]. DL é um conceito extremamente genuíno para a comunidade denominada "cientistas de dados". Esta centralização de dados é muito mais apelativa para a comunidade científica que se dedica à análise de dados, pelo fato dos dados se encontrarem no seu estado original. Em outras abordagens fundamentadas na centralização de dados, como por exemplo no *Data Werehouse*, os dados são transformados para serem integrados numa determinada estrutura pré-modelada. Como consequência, a informação que poderá ser importante

na exploração de futuros processos analíticos, é simplesmente perdida. Eventualmente, os dados poderão conter valor impossível de ser extraído, face ao estado presente das metodologias de análise. No entanto, num futuro próximo, com o desenvolvimento de novas metodologias de análise, estes dados armazenados no seu estado original, poderão vir a dar resposta a muitas questões e à descoberta de novos conhecimentos, até então longe de serem supostos.

Um dos grandes problemas do DL ocorre quando mal gerido, i.e., pode facilmente ser transformado num "pântano" quando armazena um grande volume de dados. Para evitar este tipo de situações é fundamental a aplicação de boas técnicas de catalogação [66].

Outra questão importante a ter em conta na implementação de um DL é garantir a privacidade dos dados. Face à imposição de leis que visam a proteção de dados pessoais (e.g., Regulamento Geral de Proteção de Dados (RGPD), *Health Insurance Portability and Accountability Act* (HIPAA), *Payment Card Industry Data Security Standard* (PCI – DSS), etc.), nem sempre será possível armazenar os dados no seu estado original. No entanto, é possível executar técnicas (e.g., Anonymization, Pseudonymization Trusted Computation, Encrypted Computation, Perturbation, Zero Knowledge Proofs, etc.) sobre eles por forma a serem garantidas as questões relativas à privacidade [67]. Desta forma será sempre possível armazenar e disponibilizar estes dados num formato original ou semi-original, que serão de grande valia para o desenvolvimento de trabalhos científicos.

2.4.5 Sistemas de ficheiros distribuídos

Sistemas de ficheiros distribuído *(Distributed File System* (DFS)) tal como o nome sugere, é uma abordagem que consiste no armazenamento dos dados num sistema de ficheiros distribuídos. As arquiteturas que suportam este tipo de armazenamento, geralmente são caracterizadas como sendo altamente tolerantes a falhas, tendo alto rendimento e de baixo custo. O exemplo mais popular deste tipo de abordagem é o *Hadoop Distributed File System* (HDFS)[1] integrado na framework *Apache Hadoop*. No entanto, existem soluções alternativas, e.g., HopsFS[68]; GlusterFS[74]; GPFS[75]; MapR-FS[2]; CephFS[3]; etc.

Alluixio [71] é outro exemplo neste tipo de armazenamento. O principal objetivo desta solução é garantir a interoperabilidade com outros sistemas de ficheiros distribuídos (e.g. Amazon S3, OpenStack Swift, GlusterFS, HDFS, Ceph, OSS, etc.) e sistemas de streaming (e.g., Spark Streaming, Apache MapReduce, Apache Flink, etc.).

A interoperabilidade é uma característica fundamental no desenvolvimento de plataformas distribuídas, e deve ser tomada em conta na seleção de componentes a integrar numa plataforma Big Data.

[1] http://hadoop.apache.org
[2] https://mapr.com/products/mapr-fs/
[3] http://docs.ceph.com/docs/master/cephfs/

2.5 Analise de dados

A Analítica é uma peça fundamental nas plataformas Big Data. São necessárias metodologias de análise para a validação da qualidade dos dados na fase de recolha, análise para facilitar a integração de dados, metodologias para a governação e segurança dos dados, análise para monitorização de equipamentos e operações, análise para deteção de anomalias e fraudes, análise para previsões e otimização de recursos. A Analítica deve estar presente em todo o ciclo da gestão de dados. O sucesso das plataformas que visam o tratamento de fluxos complexos depende invariavelmente da sua capacidade de análise. Conforme foca o autor em [72], estas plataformas têm evoluído por forma a fornecerem capacidades não só reativas (com a analítica descritiva e de diagnóstico), mas também capacidades proactivas (com suporte à análise preditiva e prescritiva). Estes quatro tipos de análise são definidos como:

- *Análise de Diagnóstico:* este tipo de análise foca-se em estabelecer relações de causa-efeito sobre dados recolhidos em determinado contexto. Esta abordagem é extremamente utilizada para entender acontecimentos passados e traçar contramedidas a fim de serem atingidos resultados e objetivos futuros (i.e., tenta responder à questão "porque aconteceu?");
- *Análise Descritiva:* baseia-se no cruzamento de dados históricos com o objetivo de obter informações claras que permitam traçar cenários precisos sobre temas relevantes no momento presente (i.e., tenta responder à questão "o que está acontecendo?");
- *Análise Preditiva:* esta abordagem tentar prever um acontecimento em determinados contextos (tendências de consumo, flutuações de mercado, etc.), com o objetivo de antever ameaças e oportunidades, (i.e., tenta responder à questão "O que é espectável acontecer?"). A análise preditiva permite a execução de ações não só reativas, mas sobretudo proactivas (manutenção preditiva de equipamentos, previsão de demanda, etc.);
- *Análise Prescritiva:* esta abordagem tenta avaliar as consequências de possíveis acontecimentos futuros. É frequentemente aplicada a decisões estratégicas (e.g. redução de custos, redução de riscos, maximização de lucros, otimização de recursos, antever o impacto e a reação do consumidor face a um novo produto, etc.). Ou seja, tenta responder à questão "E se ... o que é espectável acontecer?"). A análise prescritiva quando associada à abordagem preditiva, poderá conduzir à automatização de algumas decisões operacionais próximas do tempo real.

Na área de Big Data as várias abordagens analíticas são baseadas em metodologias de Machine Learning (ML). ML é uma área da Inteligência Artificia (AI), conforme mostra Figura 2.9, focada no desenvolvimento de tecnologias e metodologias de análise para a extração de conhecimento com base nos dados.

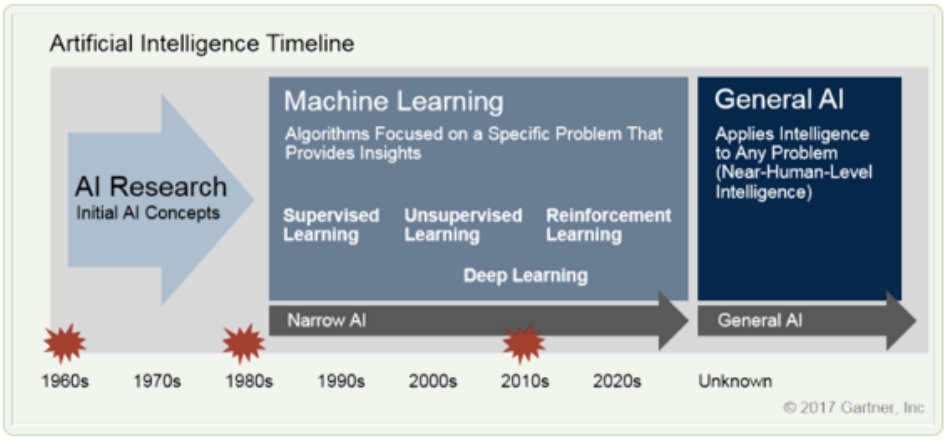


Figura 2.9 - Evolução histórica da AI [73]

As metodologias de aprendizagem em ML podem ser classificadas conforme a seguir se descreve:

- *Aprendizagem supervisionada:* Esta metodologia é baseada em várias fases de processamento. Numa primeira fase são treinados grandes volumes de dados históricos rotulados pelo ser humano. Numa segunda fase é gerado o modelo, i.e., são criadas regras gerais para serem aplicadas em problemas semelhantes que ocorram no futuro. Quanto melhores forem os dados do conjunto de treinamento, melhor será a previsão.
- *Aprendizagem sem supervisão:* consiste em treinar dados não-estruturados. Esses algoritmos exploram os dados não rotulados e tentam encontrar uma estrutura.
- *Aprendizagem de reforço:* nesta abordagem não é criado um modelo com base no treinamento de um conjunto de dados. Em vez disso, é aplicado algo semelhante à psicologia comportamental, i.e., os algoritmos são avaliados e recompensados de acordo com o resultado obtido.

Grandes avanços têm sido feitos na área de ML no sentido de solucionar inúmeros problemas complexos. Muitos dos esforços residem na transformação e conjugação de metodologias mais tradicionais, de forma a serem integradas nas novas abordagens computacionais. Um bom exemplo é a solução apresentada em [74]. Neste trabalho é proposta uma nova metodologia baseada na execução distribuída do algoritmo SVM para a previsão de preços do mercado elétrico no contexto das Smart Grids, e cujos resultados obtidos demostraram uma precisão superior relativamente a abordagens mais tradicionais. Outro excelente exemplo refere-se a uma área especifica de ML denominada Deep Learning (DL). DL é uma abordagem recente que visa a exploração de metodologias baseadas no princípio das redes neuronais. São metodologias com capacidade para serem executadas de forma distribuída, o que lhes confere altos níveis de desempenho. Têm capacidade para processar grandes volumes de dados complexos (e.g., texto, imagens, vídeo, áudio, etc.) o que lhes permite obter uma precisão superior relativamente a outros algoritmos mais tradicionais. Apesar de permitirem a resolução de inúmeros problemas altamente complexos como a tradução de idiomas, reconhecimento facial,

deteção de fraudes e anomalias ou previsões, foram otimizadas para problemas de classificação. Os problemas de classificação, baseados em redes neuronais, são bastante exigentes relativamente ao volume de dados para treino, i.e., quanto mais dados obtiver na fase de treino, maior será a sua capacidade de previsão. Mais recentemente, esta área tem se dedicado à exploração de soluções para problemas de previsão baseados em regressão linear. O desenvolvimento de DL está em plena ascensão e com grande aceitação no mercado [75]. Inúmeras soluções estão disponíveis, quer *open source* (Caffe, TensorFlow, Theano, Deeplearning4j, etc.), quer soluções comercias (Arimo, deepsense.Io, H2O.ai, Kaggle, Psiori, Skymind, etc.). Estas e outras soluções podem ser consultadas na (Tabela 2.3), no final deste subcapítulo.

Atualmente os desenvolvimentos na área de ML estão focados na resolução de problemas específicos. É espectável que a AI evolua para algo denominado *Artificial General Inteligente* (AGI) [76]. AGI tenta explorar soluções que detenham capacidade para resolver qualquer tipo de problema, i.e., resolver problemas de uma forma mais genérica. A ideia chave é desenvolver mecanismos capazes de imitar cognitivamente o comportamento humano. No entanto AGI é apenas um conceito embrionário, cujos avanços vão pouco para além de meras discussões sobre o assunto.

Um dos grandes desafios na área Analítica é o desenvolvimento de mecanismos que permitam a sua própria automatização, i.e., automatização dos processos que envolvam analítica. Para além da simplificação dos próprios processos, o grande objetivo é tornar a análise de dados acessível a todos. Tendo em conta estes objetivos, as plataformas Big Data tendem a disponibilizar mecanismos analíticos para diferente tipo de utilizadores, i.e., cientistas de dados que pretendem explorar novos modelos, utilizadores que necessitam de executar experimentação analítica sobre os modelos existentes, e utilizadores meramente operacionais que necessitam de tomar decisões com base na execução de modelos sofisticados.

Com o objetivo de se obter uma melhor compreensão acerca dos grandes desafios propostos na área Analítica de Big Data, a seguir, e com base no relatório disponibilizado pela empresa Gartner [76], são descritas as maiores tendências do mercado tecnológico:

- *Human-in-the-Loop Crowdsourcing*: Consiste numa nova abordagem híbrida para conciliar a experiência humana ao desempenho de algoritmos inteligentes. O objetivo é permitir a intervenção do ser humano de forma a melhorar os resultados dos algoritmos. Alguns exemplos de uso: rastreamento de ações humanas para identificar padrões a fim de consegui-lo em futuras operações, automação de tarefas em que é difícil descrever regras (e.g., rotulação e validação da qualidade dos dados na fase de recolha), tarefas relacionadas com a integração de dados [77], tarefas que requerem habilidades raras (e.g., conhecimento de nichos de mercado), etc.;

- *Conversational Analytics:* Esta abordagem visa a criação de metodologias para facilitar a interação do ser humano com os dispositivos móveis. É o caso por exemplo dos assistentes pessoais;
- *Algorithm Marketplaces:* Esta abordagem foca-se no desenvolvimento de infraestruturas capazes de facilitar a reutilização de algoritmos desenvolvidos e testados pelos cientistas de dados. Estas plataformas podem ser usadas para disponibilizar o uso de algoritmos como um serviço ou como suporte operacional de uma organização;
- *Guided Analytics:* Explora o desenvolvimento de ferramentas analíticas interativas com o objetivo de orientar e facilitar as etapas de preparação e exploração de dados, para utilizadores menos qualificados na área da ciência dos dados;
- *AutoML:* Tentar encontrar o algoritmo e a configuração ideal dos seus parâmetros, para a solução de determinado problema, é na maioria das vezes uma tarefa rotineira de tentativa-e-erro que exige muito tempo. O objetivo desta abordagem consiste na automação destas tarefas. Os modelos recomendados são feitos com base no caso de uso e dados disponíveis. E o critério para a recomendação de configuração de parâmetros tem como base a avaliação da linguagem do modelo e da plataforma onde este será executado. A grande vantagem é reduzir o tempo despendido com estas tarefas, libertando os cientistas de dados para o foco em tarefas complexas de exploração de novas metodologias analíticas;
- *Embedded Analytics:* Tem como objetivo a incorporação da análise em aplicativos cooperativos e ferramentas de negócio que são consumidos pelos utilizadores finais de forma quase invisível;
- *IoT Edge Analytics:* Esta abordagem consiste em explorar a possibilidade de executar a análise de dados junto das fontes que os geram. Está estritamente relacionada com a *edge computing*. Surgiu devido à grande proliferação de dados gerados pelos dispositivos IoT e pela grande necessidade de obter analítica em tempo real, sobre esses mesmo dados. Este tipo de abordagem trará grandes vantagens para a obtenção de respostas mais rápidas, facilitará questões relacionadas com privacidade de dados pessoais, evitará congestionamento de dados na rede, aumentará a confiabilidade das operações, etc. No entanto, esta abordagem conduz a uma maior complexidade na sua implementação, questões relacionadas com ciber-segurança serão mais difíceis de ser garantidas, e por último, o facto dos dados serem filtrados e analisados na sua origem, poderá levar à perda de informação útil para a descoberta de novos conhecimentos;
- *Advanced Anomaly Detection:* Quando determinados eventos ou objetos se desviam do espectável ou da norma, estamos na presença de algo anómalo. As técnicas de deteção de anomalias visam identificar este tipo de acontecimentos e caracterizam-se como sistemas avançados quando incorporam técnicas analíticas avançadas que lhes conferem a capacidade de

antever com precisão futuras anomalias, i.e., capacidade de serem não apenas reativos, mas sobretudo proactivos;

- *Citizen Data Science:* Esta área tem como objetivo o desenvolvimento de plataformas analíticas que permitem a profissionais extrair conhecimento avançados a partir dos dados, sem que para tal tenham de ter conhecimentos avançados no âmbito da ciência dos dados. Estes profissionais, denominados *citizen data scientists*, têm geralmente grande conhecimento na área de negócio e habilidades analíticas para operar estes modelos. As ferramentas disponibilizadas por estas plataformas agilizam e automatizam tarefas relacionadas com a preparação de dados, disponibilizam orientação relativamente ao uso dos modelos, disponibilizam visualizações avançadas e interativas dos resultados, permitem a colaboração e partilha de conhecimento entre os seus utilizadores, etc.;
- *Augmented Data Discovery:* Esta área analítica dedica-se à exploração de metodologia para a descoberta de novos conhecimentos. Anteriormente denominada como "*Smart Data Discovery*", esta área é caracterizada pelas plataformas de Business Intelligence (BI) modernas, que permitem aos seus utilizadores, extrair novos conhecimentos sem terem de criar novos modelos e escrever novos algoritmos. Técnicas como correlação de dados, clusters, previsões, visualizações avançadas, estão normalmente disponíveis nestas plataformas;
- *Graph Analytics:* Esta abordagem é baseada em técnicas de grafos e surgiu pela necessidade de analisar grandes volumes de dados heterogéneos e complexos. Visa sobretudo explorar relações e influências entre diferentes entidades. Os resultados analíticos são representados por gráficos interativos, facto que facilita a sua visualização, interpretação e exploração. Esta abordagem é altamente eficiente, e.g., avaliação de riscos, otimização de rotas, análise de fraudes, balanceamento de cargas, análises de mercado, localização inteligente, etc.;
- *Optimization:* Esta abordagem consiste em encontrar a melhor solução num conjunto possível de soluções viáveis, usando algoritmos matemáticos que maximizam ou minimizam uma função objetivo sujeita a determinadas restrições;
- *Self-Service Data Preparation:* São abordagens analíticas para facilitar a preparação de dados em todas as tarefas a eles relacionadas, e.g., catalogação, transformação e modelação, integração, linhagem, manipulação dos dados para visualização ou para o processo de análises mais complexas, etc. Estas ferramentas analíticas têm assumido uma importância relevante para as tarefas relacionadas com a governação de dados. O seu desenvolvimento terá grande impacto uma vez que 80% das tarefas relacionadas com o processo analítico estão centralizadas na preparação dos dados;

- *Cognitive Computing:* Esta área dedica-se à exploração de tecnologias capazes de imitar e entender as habilidades cognitivas do ser humano de forma a melhor interagir com ele e melhorar o seu desempenho.

A tabela seguinte (Tabela 2.3), conforme já referido, resume algumas das mais populares *frameworks* disponibilizadas na área de Analítica, classificados como *Open Source.* Por fim, a Tabela 2.4, resume algumas das plataformas e *frameworks* comerciais que disponibilizam soluções de acordo com as tecnologias analíticas abordadas.

Tabela 2.3 - Big Data: *Frameworks* para a analítica

COMPANY	PRODUCT	DESCRIPTION	WEBSITE
Airbnb.io	Aerosolve	ML library designed from the ground up to be human friendly	http://airbnb.io/aerosolve/
Amazon	DSSTNE	Deep Scalable Sparse Tensor Neural Engine is a software for Deep Learning	https://github.com/amzn/amazon-dsstne
Apache Foundation	MADlib	Big data machine learning in SQL	http://madlib.apache.org/
Apache Foundation	Mahout	Machine learning and data mining on Hadoop	http://mahout.apache.org/
Apache Foundation	Singa	Machine learning library creation	http://singa.incubator.apache.org/en
Apache Foundation	Spark MLlib	Machine learning library for Apache Spark	http://spark.apache.org/mllib
Caffe2	Caffe2	Deep learning framework	http://caffe2.ai/
Chainer	Chainer	Neural network framework	http://chainer.org
Google	TensorFlow	Machine learning library	http://tensorflow.org/
Intel Nervana	neon	Deep learning framework	http://neon.nervanasys.com/
Keras	Keras	Deep learning library for Python	http://keras.io/
Microsoft	CNTK	Deep learning and Cognitive toolkit	http://github.com/Microsoft/CNTK
Microsoft	DMTK	Distributed Machine Learning Toolkit	http://dmtk.io/
MXNet	MXNet	Deep learning library	http://mxnet.io/
OpenAI	OpenAI (Gym; Baselines)	AI software for training, experimentation and benchmarking. The main goal is to the development of safe AGI	https://openai.com/

COMPANY	PRODUCT	DESCRIPTION	WEBSITE
PaddlePaddle	PaddlePaddle	"PArallel Distributed Deep Learning" is a platform base in DL, flexible, easy-to-use, efficient and scalable	htpp://www.paddlepaddle.org/
Preferred Networks, inc.	Chainer	A flexible framework of Neural Networks for Deep Learning (is a Python-based deep learning framework aiming at flexibility.)	https://chainer.org/
Samsung	Veles	Distributed machine learning platform	http://github.com/Samsung/veles
Scikit Learn	Scikit Learn	Machine learning libraries for Python	http://scikit-learn.org/stable
Skymind	DL4J	Deep learning software for Java and Scala	http://deeplearning4j.org/
Torch	Torch	Machine learning framework for use with GPUs	http://torch.ch/
University of Montreal	Theano	Deep learning library for Python	http://deeplearning.net/software/theano/
University of Waikato	Weka AutoWeka	Machine learning and data mining for Java; AutoML	http://cs.waikato.ac.nz/ml/weka

Tabela 2.4 - Soluções comerciais para a Data Science e ML

	Features															
Sample Vendors	Machine Learning	Deep Learning	Predictive Analytics	Prescriptive Analytics	Cognitive Computing	Augmented Data Discovery	Embedded Analytics	Citizen Data Science	Guided Analytics	AutoML	Graph Analytics	Optimization	Human-in-the-Loop Crowdsourcing	Conversational Analytics	IoT Edge Analytics	Advanced Anomaly Detection	Algorithm Marketplaces
IBM	x		x	x	x	x		x	x	x		x					x
SSAS	x		x	x		x		x	x			x					
Microsoft	x	x	x		x		x										x
DataRobot						x	x	x	x	x							
H2O.ai	x	x	x							x							
RapidMiner	x		x						x			x					
Alpine Data							x		x	x							

	Features															
Sample Vendors	Machine Learning	Deep Learning	Predictive Analytics	Prescriptive Analytics	Cognitive Computing	Augmented Data Discovery	Embedded Analytics	Citizen Data Science	Guided Analytics	AutoML	Graph Analytics	Optimization	Human-in-the-Loop Crowdsourcing	Conversational Analytics	IoT Edge Analytics	Advanced Anomaly Detection	Algorithm Marketplaces
Ayasdi									x		x					x	
KNIME	x		x						x								
Salesforce						x	x	x									
SAP	x						x	x									
Amazon		x											x	x			
Google		x			x					x							
AIMMS				x								x					
Alteryx	x																x
Dataiku	x		x														
Digital Reasoning					x						x						
Domino Data Lab	x		x														
FICO				x								x					
Gurobi Optimization				x								x					
Kaggle		x											x				
Maana											x		x				
MathWorks			x									x					
Palantir											x					x	
SparkBeyond						x		x									
Intel		x									x						
TIBCO															x		

2.6 Visualização de dados

A área de visualização de dados dedica-se ao estudo de formas de representar a informação. O principal objetivo é que esta seja apelativa e facilmente percetível pelo ser humano. O ser humano tem maior facilidade na interpretação de informação visual, ou seja, consegue facilmente reconhecer padrões, tendências e anomalias quando a informação está representada num gráfico. No entanto, não consegue obter a mesma perceção quando a mesma informação se encontra representada de forma tabelar. Os métodos de visualização para apresentar os dados de forma eficaz e interessante centram-se em tipos como: Gráficos; Diagramas; Mapas e Painéis.

Nesta nova era de informação, a complexidade inerente dos dados, induziram novos desafios na área de visualização. Novas questões foram colocadas. Como representar grande volume de dados? Como lidar com dados completamente destruturados? Como representá-los em tempo real? Muitos estudos têm sido feitos nesta área ([78]–[82]), onde se conclui que a área de visualização está muito interligada com a área analítica, i.e., as soluções analíticas tendem a incorporar a seu próprio painel (denominado em inglês por *dashboard*) para a visualização de resultados. No entanto, têm sido feitos esforços no sentido de desenvolver novas soluções completamente independentes e interoperáveis com outros componentes (*Grafana*, *Conograf*, *Kinbana*, etc.). Estas soluções tentam ainda acrescentar funcionalidades para facilitar a interatividade com os dados. A seguir, na Tabela 2.5, sumariza-se algumas das mais populares ferramentas na área de Visualização para Big Data.

Tabela 2.5 - Frameworks: Big Data Visualização

	ORGANIZAÇÃO	PRODUTO	DESCRIÇÃO	WEBSITE
Open Source	Apache Foundation	Zeppelin	Visualização interativa de dados	http://zeppelin.apache.org/
	D3.js	D3.js	Biblioteca de visualização em JavaScript	http://d3js.org/
	Eclipse Foundation	BIRT	Biblioteca de visualização e geração de relatório, em Java	http://eclipse.org/birt/
	Elasticsearch	Kibana	Dashboard integrado da solução Elasticsearch.	https://www.elastic.co/pt/products/kibana
	Grafana	Grafana	Plataforma visualização e analise de métricas.	https://grafana.com/
	Graphviz	Graphviz	Toolkit de visualização gráfica	http://graphviz.org/
	InfluxData	Conograf	Dashboards para visualizações em tempo real; interface para a plataforma InfluxDB.	https://www.influxdata.com/time-series-platform/chronograf/
	JUNG Framework	JUNG Framework	Framework Java para gráficos, modelagem, análise e visualização de dados	http://jung.sourceforge.net/
	Project Jupyter	Jupyter	Visualização interativa de dados; integração com Spark e Hadoop	http://jupyter.org/
	Sencha	InfoVis Toolkit	Biblioteca de visualização em JavaScript	philogb.github.io/jit
Commercial	1010data	Insights Platform	Gestão de dados, análise, modelagem, relatórios e visualização	http://1010data.com/products/insights-platform/analysis-modeling
	Alteryx	Alteryx Analytics	ETL, análise preditiva, análise espacial, fluxos de trabalho automatizados, relatórios e visualização.	https://www.alteryx.com/products/alteryx-platform/alteryx-designer
	Databricks	Databricks	Ciência de dados (ingestão, processamento, colaboração, exploração e visualização) para Spark	https://databricks.com/product/unified-analytics-platform

ORGANIZAÇÃO	PRODUTO	DESCRIÇÃO	WEBSITE
Datameer	Datameer	BI, integração de dados, ETL e visualização de dados no Hadoop	http://datameer.com/product/product-overview
DataWatch	DataWatch	Extração de dados, análise self-service, visualização	https://www.datawatch.com/in-action/monarch-desktop/
Domo	Domo	Integração, preparação e visualização de dados	http://domo.com/product
GoodData	GoodData	Distribuição de dados, visualização, análise (R, MAQL), BI e armazenamento	http://gooddata.com/platform
Informatica	Relate 360	Análise de Big Data, visualização, pesquisa e BI	https://www.informatica.com/
Qlik	Qlik Analytics Platform,	Plataforma de visualização de dados, Integração e pesquisa	https://www.qlik.com/us
Sisense	Sisense	BI, Analítica, Visualização e Relatórios	http://sisense.com/product/
Tableau	Tableau Desktop	Visualização, Análise	https://www.tableau.com/products/desktop
TIBCO	TIBCO Spotfire	Data mining e Visualização	http://spotfire.tibco.com/
Trifacta	Trifacta	Agrupamento, exploração e visualização de dados no Hadoop	https://www.trifacta.com/products/wrangler-editions/#wrangler
Yellowfin	Yellowfin	BI e Visualização de dados	http://yellowfinbi.com/platform
Zoomdata	Zoomdata	Analytics, visualization, and BI with self-service on Hadoop, Spark, many data stores	http://zoomdata.com/

2.7 Governação de Dados

A Governação de Dados é definida como sendo um conjunto de atividades que visa estabelecer regras para a criação, reutilização e consumo de dados. Segundo as boas práticas definidas em [83], estas atividades consistem em: definir a arquitetura, os metadados e os modelos de dados, garantir a qualidade dos dados, definir e garantir a integração e interoperabilidade dos dados, definir o armazenamento dos dados, definir processos e operações executadas nos dados, garantir a segurança dos dados (i.e., definir e implementar políticas de acesso e permissões, garantir a privacidade dos dados privados de acordos com os regulamentos impostos por lei).

Os dados representam um ativo cada vez mais importante na atividade das empresas, e são cada vez mais complexos de gerir. Na literatura científica são escassas as referências recentes relacionadas com o tema "Governação de Dados", analisado como um todo [84], [85]. Geralmente, os trabalhos realizados nesta área focam apenas um ou dois aspetos relacionados com o tema (segurança, qualidade dos dados, integração, etc.). No entanto, a inclusão da governação de dados em plataformas Big Data, continua a ser um assunto bastante desafiador.

Apache Atlas [86] e *Apache Range* [87] são dois componentes *open source* que se distinguem nesta área de governação de dados. O componente *Apache Range* permite definir permissões e implementar políticas de acesso a dados. Por sua vez, *Apache Atlas* disponibiliza funcionalidades para mesclarem os dados, definição de metadados, catalogação e linhagem dos dados, etc. Estes dois componentes são completamente integráveis no ecossistema *Hadoop* e estão disponíveis na plataforma *Hortonworks*.

Hortonworks é uma das primeiras e mais populares plataformas Big Data. Com Base no ecossistema *Hadoop*, outras plataformas foram desenvolvidas, e.g., *Cloudera*, *MapR*, etc.

Atualmente verifica-se uma proliferação de plataformas Big Data. Muitas delas disponibilizam as suas funcionalidades em nuvem como serviço, e.g., *Microsoft Azure*, *Amazon Web Services* (AWS), *Google Cloud*, *Clouder, BlueData*, etc. Por ouro lado, verifica-se o foco na especificidade das funcionalidades disponibilizadas (*Anaconda* e *BlueData*) que direcionam o seu foco para a análise de dados, *Informatica,* centraliza as suas funcionalidades numa visão 360º sobre dados de negócio, *Apache Nifi* tem o seu foco na recolha e integração dos dados, etc. Por fim, verifica-se que todas estas soluções são na sua maioria comerciais. No entanto, algumas estão disponíveis sem custo em *Sandbox* (i.e., ambiente virtual fechado que permite isolar a execução de determinados processos). *Sandbox*, ainda que não seja o ambiente esperado para o desenvolvimento de uma solução em produção, é um excelente recurso para experimentação de soluções. Como solução *open source,* destaca-se o projeto *Big Data Eurupe* (BDE), que visa disponibilizar uma solução Big Data baseada na conetividade de vários componentes via *Docker Container*. Estas e outras plataformas Big Data encontram-se sumariadas na tabela seguinte (Tabela 2.6).

Tabela 2.6 - Plataformas Big Data

PRODUCT	DESCRIPTION	WEBSITE
Anaconda	Plataforma para cientistas de dados	https://www.anaconda.com
Arcadia Data	Plataforma de análise e BI orientada AI; Interoperabilidade com *Data Lake*	https://www.arcadiadata.com
Attunity	Integração e Gestão de dado	https://www.attunity.com
AWS	*Amazon Web Services* - Todos os Serviços Big data disponíveis via Cloud	https://aws.amazon.com/pt/
Azure	Plataforma Big Data, proprietária da Microsoft; disponibiliza todos os serviços via Cloud	https://azure.microsoft.com/pt-pt/
BDE	Plataforma Big Data apoiada na interoperabilidade de componentes open source	https://www.big-data-europe.eu/pipeline-2/
BlueData	Plataforma de serviços de IA e ML e analítica, baseados em Containers e disponibilizados em Cloud	https://www.bluedata.com
CDAP	*Open source framework* para o desenvolvimento de aplicações analíticas	https://cdap.io
Cloudera	Serviço disponibilizado via Cloud para analítica - IA e Edge	https://www.cloudera.com/about.html
Dataiku	Plataforma colaborativa de análise e operacionalização de modelos de ML	https://www.dataiku.com
DataRPM	Solução disponibilizada pela empresa *Progress* - Deteção de anomalias e Previsão	https://www.progress.com/datarpm
Datawatch	Unificação da recolha, análise, visualização e previsão de dados.	https://www.datawatch.com
Hortonworks	Plataformas e serviços de gestão de dados globais	https://br.hortonworks.com
Informatica	Análise de Big Data, visualização, pesquisa e BI	https://www.informatica.com/#fbid=rlL6vTJV4lx
MapR	Data Platform for AI and Analytics	https://mapr.com

PRODUCT	*DESCRIPTION*	*WEBSITE*
Apache NiFi	*Framework* para automatização de fluxo de dados entre sistemas de software	https://nifi.apache.org
OperaSulution	*Edge data science* com desempenho em escala	https://www.operasolutions.com
QLIK	Plataforma de visualização de dados, Integração e pesquisa	https://www.qlik.com/us
SAP HANA	Business Data Platform	https://www.sap.com/products/hana.html
Seematix	Plataforma Big Data, IA, IoT e Análise de dados	https://www.semantix.com.br/en/
StreamSets	Plataforma para Integração e análise de fluxo de dados	https://streamsets.com/products/dataops-platform
Talend	Visualização interativa de dados para BI	https://www.tableau.com

2.8 Conclusão

Neste capítulo foi abordada a complexidade inerente aos fluxos de dados que caracterizam esta nova era de informação. Os principais conceitos relacionados com as diversas subáreas que caracterizam Big Data foram igualmente abordados. Foram descritas pesquisas sobre a disponibilidade de componentes para a implementação de plataformas Big Data, focadas nos seus principais conceitos, maturidade, funcionalidades e performance. Salienta-se ainda a importância da interoperabilidade entre componentes.

Na área de análise de dados conclui-se que os grandes desafios estão centrados no desenvolvimento de soluções que permitam a sua automatização. Por outro lado, é necessário desenvolver soluções adaptadas ao perfil dos utilizadores, (i.e., de acordo com o seu nível de conhecimento analítico) de forma a todos poderem usufruir dos benefícios da análise de dados. Verifica-se ainda, a necessidade de desenvolver soluções direcionadas para casos específicos de uso, e.g., *Anomaly Detection, IoT Edge Analytics, Data Preparation,* etc.

Relativamente às plataformas Big Data conclui-se que estas são na sua maioria soluções comerciais, disponibilizadas em nuvem como um serviço. Para além disso, são focadas na resolução de problemas específicos, e.g., análise de dados, recolha e integração de dados, etc.

Por fim, conclui-se que a área de Big Data está repleta de desafios e um longo caminho está ainda por percorrer, apesar dos notórios avanços tecnológicos feitos nesta área.

3 Smart Grids

Este capítulo explora o contexto sobre o qual se pretende a aplicabilidade da evolução tecnológica verificada em Big Data. Assim, neste capítulo serão abordados os conceitos e desafios inerentes ao ecossistema energético inteligente, com destaque para o panorama Europeu. É ainda apresentada uma revisão da literatura sobre a integração de Big Data no contexto das Smart Grids. No mesmo âmbito acrescenta-se uma revisão sobre o setor energético, e os investimentos em investigação e desenvolvimento promovidos pela União Europeia. Finalmente são apresentadas as conclusões sobre o estudo desenvolvido.

3.1 Introdução

Por razões ecológicas, económicas e políticas surgiu há mais de uma década o conceito de Smart Grid (SG) para definir o desenvolvimento do ecossistema energético. Ambiciona-se prover este ecossistema com sistemas inteligentes, seguros e descentralizados de forma a facilitar a incorporação de energias renováveis em grande escala. Conforme mostra Figura 3.1, o setor energético é caracterizado por dois fluxos: fluxo energético e fluxo de dados. O fluxo energético refere-se ao produto propriamente dito, cuja gestão é cada vez mais complexa face aos novos desafios e objetivos definidos. Por sua vez, o fluxo de dados representa um dos maiores ativos do setor. Conforme se depreende, é um fluxo extremamente complexo composto por um conjunto avultoso de comunicações bidirecional entre os vários domínios que compõem o ecossistema. É expectável que a gestão eficiente deste ativo resulte em sucesso para os complexos desafios inerentes ao setor energético. Espera-se que a gestão do fluxo de dados seja eficiente na resposta a questões relacionadas com a tomada de decisão, monitorização e controlo de equipamentos, fiabilidade da rede, segurança e otimização de recursos, etc.

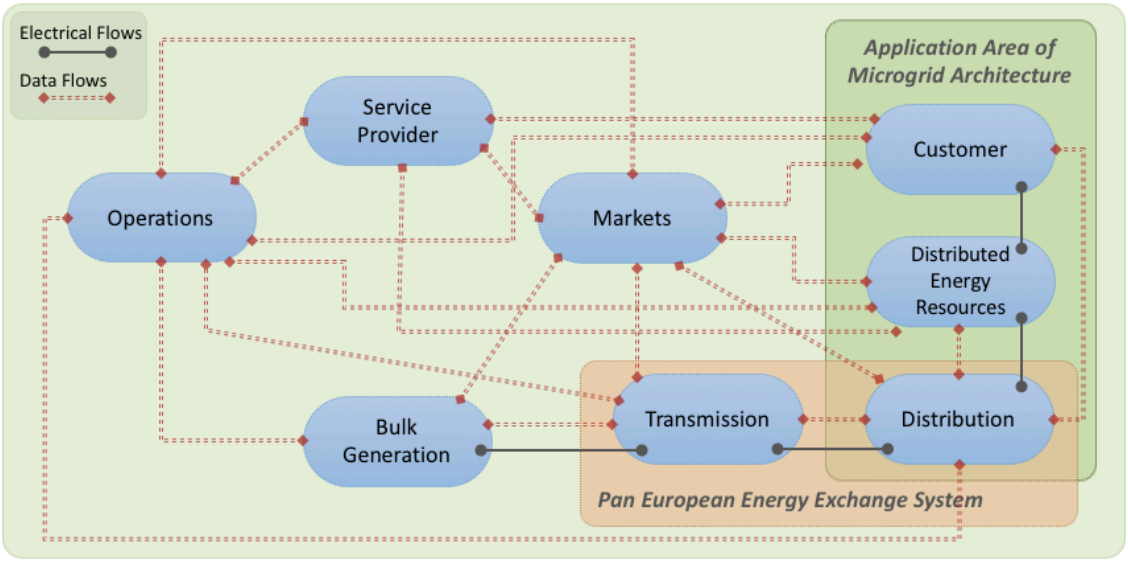


Figura 3.1- Extensão do modelo NIST para Smart Grid proposto pela EU (adaptado de [88]*)*

Esta nova era digital representa uma excelente oportunidade para vencer os complexos desafios impostos ao sector energético, conforme mencionado por Gil Quiniones, presidente e CEO da empresa NYPA (*New York Power Authority*), aquando da recente inauguração do centro digital para monitorização operacional iSOC (*Integrated Smart Operations Center)* [89]. Na mesma altura e partilhando da mesma opinião, a União Europeia (UE) emite um comunicado onde afirma que esta nova era digital é uma grande oportunidade para o desenvolvimento europeu, no qual destaca o setor energético [90]. Acrescenta ainda que, os sistemas inteligentes de energia, seguros e sustentáveis, são da responsabilidade de todos. As tecnologias digitais têm vindo a desempenhar um papel muito importante no desenvolvimento das Smart Grids (e.g., sistemas inteligentes de medição, casas e edifícios inteligentes, eletrodomésticos inteligentes, soluções inteligentes para carregamento de veículos elétricos, etc.). No entanto, apesar do setor energético estar mais descentralizado e mais descarbonizado, carecem da cooperação continua desta nova era digital. Esta era digital representa um sem número de oportunidades para as Smart Grids e deve ser olhada como o caminho a seguir, para que as Smart Grids se transformem em algo verdadeiramente sustentável e real. Ressalva ainda a importância desta nova era de dados como parte integrante no desenvolvimento das Smart Grids e apela para a mudança cultural das empresas energéticas no sentido de compartilharem dados. No entanto, admite que este assunto necessita de uma maior regulamentação e como tal será revisto no corrente programa Horizonte 2020 (i.e., 2018-2020).

Por outro lado, verifica-se que a implementação de novas tecnologias no setor energético é mais lenta em relação a outras áreas de negócio [91]. O que de certa forma se justifica pelo facto do sistema energético ser um dos maiores pilares na sustentabilidade de qualquer sociedade [92]. Ou seja, antes da implementação de qualquer nova tecnologia no seu ecossistema, esta tem de ser devidamente testada e validada, nos mais variadíssimos contextos energéticos, para que possa ser considerada uma solução viável. Caso contrário os riscos e prejuízos podem ser avultados e incalculáveis. O desenvolvimento e implementação de ambientes de simulação e emulação constituem um grande desafio, contudo, são essenciais para o sucesso e evolução das Smart Grids.

É neste contexto que se enquadra um dos principais objetivos desta dissertação, i.e., propor uma arquitetura para a gestão do fluxo de dados num ecossistema energético simulado e emulado. Pretende-se o desenvolvimento de uma solução flexível, ágil e escalável de forma a facilitar a implementação e experimentação das mais diversas tecnologias emergentes na área de Big Data. Com o propósito de atingir o objetivo proposto, a seguir serão abordados os principais conceitos das Smart Grids e sua evolução. Por fim, será apresentada uma revisão sobre a forma como as novas abordagens tecnológicas propostas na área de Big Data (descritas no capítulo 2) estão a ser implementadas no contexto das Smart Grids.

3.2 Smart Grids – Visão geral

Apesar da noção de redes energéticas inteligentes estar associada ao conceito de Smart Grids, esta não é unanime devido às diferentes estratégias delineadas pelos diferentes países relativamente ao seu desenvolvimento. O *National Institute of Standards and Technology* (NIST), organização responsável por coordenar o desenvolvimento das Smart Grids nos EUA, defini uma Smart Grids como:

> "*Smart Grid is a modernized grid that enables bidirectional flows of energy and uses two-way communication and control capabilities that will lead to an array of new functionalities and applications*" [93]

No entanto, no contexto da União Europeia SG é definida como:

> "*A Smart Grid is an electricity network that can cost efficiently integrate the behaviour and actions of all users connected to it – generators, consumers and those that do both – in order to ensure economically efficient, sustainable power system with low losses and high levels of quality and security of supply and safety*" [94].

Um recente estudo[95] feito nesta área é bem elucidativo em relação a estas divergências. Este estudo, conforme se resume na Tabela 3.1, compara a evolução das Smart Grids nos países da UE, EUA, Japão e China e é relevante para o entendimento dos trabalhos de investigação propostos por estes países.

Tabela 3.1- Estudo comparativo da evolução das Smart Grids (SGs) na UE, EUA, Japão e China

ÂMBITO	***DESCRIÇÃO***
Visão geral da evolução das SGs	- O desenvolvimento das SGs na Europa e na América apresentam um estado mais maduro relativamente ao Japão e à China. - Na UE, EUA e Japão a gestão das SGs é descentralizada ao contrário da China. O sector energético na UE e EUA é composto por vários intervenientes o que dificulta a definição de padrões. No Japão o sector é composto por apenas 10 empresas o que facilita todo o processo de fixação de padrões. - Distribuição uniforme de energia na EUA e UE. A China apresenta desequilíbrio na distribuição de energia e carga. O japão é altamente dependente das importações, justificação pela qual aposta claramente em energias renováveis.
Mercados e tarifas	- EUA e UE: RTP (preços em tempo real); CPP (preço de pico critico); PTR (preço de desconto de ponta); TOU (preço de tempo de uso). ainda desenvolveram processos para o DR (Resposta à Demanda). - Japão: vários tipos de TOU, e simulação em projetos pilotos do CPP. - China: Aplica preços independentemente do período de uso tendo, no entanto, iniciado a simulação do TOU em alguns projetos piloto.
Políticas	- EUA: políticas para incentivar o investimento privado. - UE: políticas que visam a redução de emissões de carbono. - Japão: políticas direcionadas para a implementação de energias renováveis a fim de reduzir a sua dependência energética. - China: políticas direcionadas par o desenvolvimento de energias renováveis e para a poupança de energia.
Projetos	- EUA e UE: foco no estudo de gestão da rede elétrica; agregadores de DR (*Demand Response*) e VPPs (*Virtua Power Player*); consumidores inteligentes e tecnologias de casa inteligentes.

	- UE: projetos maioritariamente direcionados para questões de DER (*Distributed Energy Resource*) e EVs (*Electric Vehicle*). - EUA: projetos maioritariamente centralizados nas questões relacionadas com os *Smart Meters*. - Japão: projetos com foco na monitorização do uso de energia, DR, EVs e otimização de sistemas de armazenamento. Maior destaque para projetos relacionados com a utilização de EMS (sistemas de gestão de energia). - China: vários projetos piloto com o objetivo de otimizar a distribuição de energia.
Desafios	- EUA: consciencialização do consumidor final para o real valor das SGs, melhoria nas comunicações entre consumidores e restantes atores do ecossistema energético; incentivar o investimento de empresas privadas nas SGs; sistemas para a tomada de decisão relativamente ao investimento em SGs (i.e., melhorar as leis e regulamentos e fortalecer a comunicação entre todos os intervenientes do ecossistema energético). - UE: maior apoio financeiro para o investimento em SGs e implementação de mecanismos que permitam a clara visão entre custos e benefícios. - Japão: participação ativa no desenvolvimento de padrões a nível internacional (i.e., concentrar-se na cooperação internacional, tirando proveito das vantagens tecnológicas que é possuidora). - China: definição de políticas nacionais claras relativamente ao desenvolvimento das SGs; transição entre a tradicional rede energética para o novo paradigma das SGs (i.e., concentrar-se na implementação de reformas da rede elétrica).

Apesar da divergência no seu conceito, tem-se assistido à colaboração entre os vários países no sentido de estabelecer padrões universais para o ecossistema energético, uma vez que o principal objetivo é comum, i.e., um ecossistema energético sustentável, económico e limpo. Por outo lado, conclui-se que apesar de metas e objetivos diferentes, o desenvolvimento das Smart Grids depende inevitavelmente dos avanços tecnológicos, definições de padrões e da sensibilização de todos para o benefício das redes elétricas inteligentes.

A seguir é descrito com mais pormenor a evolução das Smart Grids na UE a fim de introduzir o conceito de microgrids, visto ser este o cenário sobre o qual foi desenvolvido o presente trabalho. Por outo lado, é fundamental o entendimento do modelo conceptual adotado pela UE, uma vez que este está na origem do avultado volume de dados gerados no ecossistema energético.

3.2.1 Smart Grids na UE

Os principais objetivos da UE relativamente à estratégia energética consistem em garantir o abastecimento confiável de energia num ambiente competitivo e justo para todos; reduzir a emissão de gases com efeito de estufa, a poluição e a dependência de combustíveis fosseis; reduzir a dependência das importações de energia. Para atingir os objetivos propostos estabeleceu as seguintes metas a serem atingidas a médio e longo prazo:

- 2020: 20% na redução das emissões de gases com efeito de estufa; 20% na melhoria da eficiência energética; pelo menos 20% do consumo energético deve ser garantido por energias renováveis.

- 2030: aumento para 40% na redução das emissões de gases com efeito estufa, 27% na eficiência energética e 27% na geração e utilização de energias renováveis. Para além destes aumentos, visa ainda a interconexão energética entres os países da UE em pelo menos 15%. De acordo com os resultados que estão a ser obtidos relativamente ao cumprimento das metas estabelecidas para 2020, a comissão europeia propôs, no final de 2016, um conjunto de objetivos adicionais para 2030 que consistem em: assumir a liderança mundial nas energias renováveis; dar prioridade à eficiência energética; estabelecer condições equitativas para os consumidores [3]
- 2050: aumento para 80% a 95% na redução de gases com efeito estufa.

No seguimento das metas a serem atingidas para 2020 a UE aprovou o mandato M490 em março de 2011[4] com o objetivo de revisar um conjunto de padrões de forma a facilitar o desenvolvimento e implementação das Smart Grids no quadro Europeu. De acordo com o referido mandato, foi criado um grupo formado por três organismos do setor privado que se dedica à normalização, (i.e., Comité Europeu de Normalização - CEN, Comité Europeu de Normalização Eletrotécnica - CENELEC e o Instituto Europeu de Normas de Telecomunicações - ETSI), que ficou denominado por Grupo CEN-CENELEC-ETSI Smart Grid de Coordenação (SG-CG). O grupo SG-CG adotou o modelo conceptual proposto pelo NIST, tendo, no entanto, necessidade de adicionar um novo domínio para contemplar alguns requisitos específicos do contexto energético da UE, conforme mostra (Figura 3.1). O modelo conceptual proposto pelo NIST fornece uma visão de alto nível para expor a complexidade do ecossistema energético inteligente, auxiliando no entendimento das inter-relações entre todos os seus intervenientes. O modelo define os fluxos bidirecionais, de energia e de dados, entre os seus sete domínios (i.e., Geração, Transmissão, Distribuição, Consumidores, Operações, Provedor de Serviços e Mercados). Por sua vez, cada domínio é composto por elementos da rede inteligente (i.e., atores e aplicações) [1].

Tendo por base este modelo, o grupo SG-CG adicionou-lhe um novo domínio, denominado DER (i.e., *Distributed Energy Resource*), como forma de representar o conceito de flexibilidade na geração e consumo de energia, fundamental para atingir os objetivos propostos pela UE. Ou seja, o DER permite a representação dos pequenos produtores de energias renováveis, no ecossistema energético. Com a inclusão deste domínio é ainda possível agrupar o consumo, a produção e o armazenamento, que por sua vez, permite a representação de um novo conceito, i.e., microgrid (MG). Uma MG é definida como sendo uma rede de baixa e/ou media tensão equipada com recursos capazes de produzir, armazenar e distribuir energia, de forma autónoma e de acordo com a procura.

Pela análise do modelo verifica-se um acentuado fluxo de dados entre os vários domínios. Com a inclusão do domínio DER e consequentemente com a implementação massiva de variadíssimos dispositivos (e.g., smart meters, sensores, atuadores, PLC´s, sistemas SCADA, etc.), necessários à operabilidade inteligente do ecossistema energético, prevê-se um significativo e constante acréscimo no volume de dados gerado a cada instante. A este volume de dados, gerado pelo próprio ecossistema,

acrescem dados externos necessários para o contexto de muitas operações (e.g., os dados atmosféricos são fundamentais para se obter uma melhor acurácia na previsão de energias geradas por fontes renováveis).

Está-se na presença de um grande volume de dados, cuja tendência será para um crescimento acentuado ao logo do tempo, quer intensificado pela complexidade do próprio ecossistema, quer pelo aumento substancial que se prevê [96] no consumo de energia. Por outro lado, o equilíbrio que caracteriza o ecossistema, exige que estes dados sejam processados em tempo real de forma a oferecerem suporte à tomada de decisão. Neste sentido, as questões relacionadas com a gestão e governação dos dados passaram a ser um assunto de grande relevância neste domínio. Assim, a seguir revisa-se como estão as tecnologias Big Dada a ser aplicadas no contexto das Smart Grids.

3.3 Big Data no contexto das Smart Grids

A fim de atingir o objetivo proposto nesta tese, foi feita uma revisão no sentido de se entender de que forma os avanços tecnológicos realizados na área de Big Data estão a ser implementados no ecossistema energético, no que diz respeito à gestão do seu fluxo de dados.

No entanto, verificou-se que na literatura científica as palavras chave "Big Data" e "Smart Grid" raramente se cruzam numa mesma publicação. Quando tal acontece, normalmente, o objetivo é justificar a presença de Big Data no ecossistema energético, face às características intrínsecas do seu fluxo de dados. Na avaliação de trabalhos de âmbito mais genérico, i.e., revisões biográficas (2016 – 02/2018), sumariados na Tabela 3.2, conclui-se que maioritariamente se focam no problema de análise de grandes volumes de dados. As plataformas Big Data nem sempre são identificadas, e em caso afirmativo, as referências ficam restritas à *framework* Hadoop. No entanto, todos são unanimes ao reconhecer os grandes benefícios que as tecnologias Big Data provocariam no desenvolvimento das Smart Grids.

Tabela 3.2 – Revisão biográfica: *Smart Grids* vs *Big Data*

Ano	*Título*	*Âmbito*
2018	*Power systems big data analytics: An assessment of paradigm shift barriers and prospects*	Revisão biográfica, identificando o grande volume de dados no sector de energias e o grande benefício de se recorrer a BDA para inferir conhecimento [97]
2017	*Big Data management in smart grid: concepts, requirements and implementation*	Revisão biográfica, apresenta uma visão geral das oportunidades, conceitos e desafios na gestão das SGs resumindo as tecnologias Big Data que podem ser usados para lidar com os requisitos das SGs. [98]
2017	*Big data issues in smart grid – A review*	Revisão literária identificando as principais fontes de dados das SGs, bem como, os benefícios da aplicação de técnologias BDA nas SGs. [99]
2017	*Energy Big Data Security Threats in IoT-Based Smart Grid Communications*	Neste artigo são revisados os principais desafios relacionados com a segurança da comunicação de dados no contexto das SGs [100]

2017	*Systematic Review of Smart Grid Analytics*	Revisão sistemática com foco em BDA, concluindo que as SGs exigem recursos de processamento em tempo real para atender às necessidades do consumidor, previsões de curto prazo para determinar os picos de geração de energia renovável. [101]
2017	*Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges*	Examina a necessidade de BDA na eficiência energética de edifícios [102]
2017	*Survey of Security Advances in Smart Grid: A Data Driven Approach*	Revisão biográfica sobre os avanços feitos na segurança das SGs caracterizando as vulnerabilidades e soluções, de acordo com o ciclo de vida dos dados (i.e., geração, recolha, armazenamento e processamento). [103]
2017	*Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges*	Revisão biográfica sobre BDA aplicada à área IoT, no qual estão incluídos os sensores e SM das SGs; Proposta de uma arquitetura para a análise de dados IoT [104]
2016	*Energy Big Data Analytics and Security: Challenges and Opportunities*	Revisão biográfica focada nos desafios inerentes à segurança das SGs. Apresenta uma taxonomia para um melhor entendimento das relações entre componentes, dados e operações existentes na segurança das SGs. Por fim enumera as questões em aberto, i.e., necessidade de arquiteturas escaláveis e interoperáveis, processamento e descoberta de conhecimento em tempo real, segurança e privacidade. [105]
2016	*Energy big data: A survey*	Nesta revisão biográfica são identificadas as principais fontes de dados nas SGs, trabalhos relacionados e desafios. Explora as questões relacionadas com a segurança da rede. [106]

A fim de se obter trabalhos mais específicos, de acordo com os desafios das Smar Grids, refinou-se a pesquisa com termos relacionados à área de Big Data e à área de Smart Grids. Dos trabalhos analisados, resumidos na Tabela 3.3, conclui-se que estes estão maioritariamente centralizados no domínio DER. Os problemas referidos estão geralmente relacionados com metodologias de análise de grande volume de dados. Visam na sua maioria explorar questões relacionadas com previsão de consumo e produção de energia, monotorização da rede e deteção de anomalias, segurança nas comunicações e otimização de recursos.

Quanto às tecnologias, i.e., *frameworks* Big Data Analítica (BDA), raramente são mencionadas e utilizadas. Em vez disso, são exploradas e propostas novas metodologias, geralmente híbridas e baseadas em abordagens tradicionais, implementadas de forma distribuída para fazer face ao problema do processamento em tempo real de grande volume de dados. As fontes de dados exploradas são na sua maioria os dados dos *Smart Meters*, Sensores e *Phasor Measurement Units* (PMUs). Outro aspeto amplamente focado consiste na necessidade de analisar e extrair conhecimento destes dados. As *frameworks* Big Data mais experimentadas são o *Hadoop* e *MapReduce*.

Tabela 3.3 - I&D na área de SGs & BD

Ano	*Título*	*Descrição*	*DA*	*P*	*O*	*SM*	*Ot*	*Tecnologia Big Data*
2018	*A Big Data Scale Algorithm for Optimal Scheduling of Integrated Microgrids*	Movo modelo de otimização de recursos, baseado em técnicas de decomposição. O Modelo é executado num cluster configurado com *Hadoop MapReduce.* [107]	x		x			Hadoop MapReduce
2017	*Graph Signal Processing in Applications to Sensor Networks, Smart Grids, and Smart Cities*	Proposta de uma metodologia baseada em grafos para a representação de dados recolhidos por sensores [108]				x		
2017	*Hadoop-based framework for big data analysis of synchronised harmonics in active distribution network*	Proposta de uma solução baseada no processamento MapReduce para tratamento de dados Harmonicos em redes energéticas [109]	x			x		Hadoop MapReduce
2017	*Big Data Analytics for Electric Vehicle Integration in Green Smart Cities*	Pesquisa sobre ferramentas de análise e dados no domínio de veículos elétricos. [110]					x	
2017	*Big data framework for analytics in smart grids*	Plataforma - *Cloud computer* - para a gestão dos dados dos *Smart Meters* [111]					x	Hadoop
2017	*Energy Big Data Security Threats in IoT-Based Smart Grid Communications*	Revisão literária sobre a segurança na rede de comunicações; Simulação de um ataque à rede elétrica. [100]				x		
2017	*Data quality of electricity consumption data in a smart grid environment*	Apresenta uma revisão sobre os métodos de deteção de dados atípicos de consumo de eletricidade; analisa e classifica a qualidade dos dados de consumo de energia em: dados de ruído, dados incompletos e dados de outliers; descreve as causas dos dados atípicos do consumo de eletricidade [112]	x				x	
2017	*A Two-Way Street: Green Big Data Processing for a Greener Smart Grid*	Orquestrador de planos cruzados para a eficiência energética em *Data Centers* [113]	x	x			x	
2017	*A Distributed Computing Platform Supporting Power System Security Knowledge Discovery Based on Online Simulation*	Plataforma de computação distribuída, baseada em *Hadoop*, para apoio à descoberta de conhecimento na área de segurança dos sistemas de energia. [114]	x			x		Hadoop
2017	*Wireless Big Data Computing in Smart Grid*	Proposta de uma arquitetura hierárquica WiFi para a comunicação dos dados nas SGs e uma abordagem híbrida constituída por: otimização externa baseada na teoria dos jogos e um algoritmo de otimização interna para o agendamento de energia. [115]	x		x			
2017	*Data Compression in Smart Distribution Systems via Singular Value Decomposition*	Proposta de um método de compressão de dados para lidar eficientemente com o problema de transporte e muitos dados. [116]	x				x	
2017	*Big Data Analysis based Security Situational Awareness for Smart Grid*	Método ML de reforço baseado em *Fuzzy cluster* e na teoria dos jogos [117]	x			x		

DA= Big Data Analítica; Pr=Previsões; Ot=Otimização; SM=Segurança, Monotorização e deteção de anomalias; Ou=Outros

Ano	*Título*	*Descrição*	*DA*	*P*	*O*	*SM*	*Ot*	*Tecnologia Big Data*
2017	*Game-Theoretical Energy Management for Energy Internet with Big Data-Based Renewable Power Forecasting*	Nova abordagem, baseada na teoria de jogos e num algoritmo combinado (*backpropagation* e genético) para a previsão de energia eólica a curto prazo [118]	x	x				
2017	*Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid*	É proposto um método baseado em SVM aplicado a um sistema distribuído, para a previsão de preços da eletricidade [74]	x	x				
2017	*Mining Energy Consumption Behavior Patterns for Households in Smart Grid*	Nova abordagem baseada na mineração incremental dos dados, com o objetivo de eliminar dados redundantes. (dados dos SMs) [119]	x				x	
2017	*Deep Learning for Household Load Forecasting - A Novel Pooling Deep RNN*	Proposta de um novo modelo: *pooling-based deep recurrent neural network* (PDRNN); consiste no agrupamento de perfis antes da aplicação de *deep lerning* visando desta forma resolver os problemas de overfiting [120]	x	x				Tensorflow
2017	*A Spatio-temporal Data Summarization Paradigm for Real-time Operation of Smart Grid*	Desenvolvimento de um novo paradigma de sumarização para dados de consumo de energia [121]	x				x	
2017	*A Hierarchical Framework for Smart Grid Anomaly Detection Using Large-Scale Smart Meter Data*	Método baseado em *data mining* e regras de associação para a determinação dos parâmetros de estado dos transformadores [122]	x			X		
2017	*Data-Driven Charging Strategy of PEVs Under Transformer Aging Risk*	Framework para tratamento e análise dos dados no contexto dos EVs [123]	x				x	
2017	*Efficient customer selection process for various DR objectives*	É proposta uma nova abordagem baseada em modelos estocásticos, com o objetivo de selecionar os melhores clientes para a participação em programas de DR [124]	x				x	
2017	*Temporal, Functional and Spatial Big Data Computing Framework for Large-Scale Smart Grid*	É proposta uma nova metodologia para a distribuição e processamento de dados das SG., a fim de obter melhor eficiência computacional. A metodologia é baseada na subdivisão dos dados por grupos funcionais. [125]	x				x	
2017	*Big Data Acquisition under Failures in FiWi Enhanced Smart Grid*	Modelo de otimização com base em algoritmos de roteamento, para melhorar a eficiência de comunicações WiFi [126]	x		x			
2017	*Massive Streaming PMU Data Modeling and Analytics in Smart Grid State Evaluation Based on Multiple High-Dimensional Covariance Tests*	Proposta de um novo algoritmo para avaliação da qualidade de energia [127]	x			x		
2017	*Advanced and Adaptive Dispatch for Smart Grids by means of Predictive Models*	Redes neurais aplicadas em *cloud* [128]	x	x				
2016	*Distributed Data Analytics Platform for Wide-Area Synchrophasor Measurement Systems*	Proposta de uma plataforma distribuída para a análise de dados dos sincrofasores [129]	x			x		

DA= Big Data Analítica; Pr=Previsões; Ot=Otimização; SM=Segurança, Monotorização e deteção de anomalias; Ou=Outros

Conclui-se ainda que, existe uma consciência evidente, por parte da comunidade científica da área das Smart Grids, da existência dos novos desafios trazidos por esta nova era de Big Data, e um claro reconhecimento dos benefícios que ela pode acrescentar no desenvolvimento das Smart Grids. No entanto, apesar de existir esta consciência, verifica-se que as tecnologias e abordagens propostas na área de Big Data (conforme abordadas no capítulo 2), raramente são mencionadas e alvo de experimentação no contexto das Smart Grids. Este fenómeno certamente se explica, por um lado, devido ao facto dos assuntos relacionados com Big Data serem relativamente recentes e estarem ainda numa faze de imaturidade, e por outro, devido à criticidade subjacente dos sistemas energéticos.

Outra questão que se avaliou diz respeito à utilização de plataformas e ferramentas Big Data por empresas do sector energético. Em [99][130] os autores disponibilizam essa informação. No entanto, verificou-se que o mercado empresarial na área de energias está em constante mutação, o que inviabiliza a informação obtida. A título de exemplo podem-se citar as recentes aquisições feitas pela empresa Enel [131]. Esta empresa, para reforçar a sua área de *Enery Storage,* adquiriu as empresas *Demand Energy Networks* e a Tyn*emouth Energy Storage*, como reforço dos seus serviços de DER adquiriu a *EnerNOC* e finalmente para reforçar a sua infraestrutura de *Electric Vehicle* adquiriu a *eMotorWerks*. Outros bons exemplos são as recentes aquisições da *Opower* pela *Oracle* [132], da *EnergySavvy* pela *Tendril* [133] e da *Utopus Insights* pela *Vestas* [134]. Apesar de existirem ferramentas disponíveis para a análise de tendências e mutações no sector energético [135], estas não são de acesso livre. Porém a mutação do mercado empresarial não se restringe ao ecossistema energético. O mesmo sucede na área das técnologias direcionadas para o assunto de Big Data. O recente relatório emitido pala *Gartner* acerca da avaliação de plataformas ML para a ciência dos dados [136] , onde refere que "só nos últimos quatro anos a plataforma *Statistica* passou da *Statsoft* para a *Dell*, desta para a *Quest* e finalmente para a *TIBCO*", é um bom exemplo deste cenário.

Assim, e a fim de superar os obstáculos atrás referidos, revisou-se o uso de ferramentas e plataformas Big Data no sector energético, com recurso não só a publicações científicas sobre estes assuntos, mas igualmente, com recurso aos media e à informação disponibilizada nos sites oficias de organizações do sector energético e tecnológico (e.g., *Energia Central*[4], *GTM Research*[5], *Solar Plaza*[6], *Gartner Research*[7], *Data Science Central*[8], etc.). Da análise desenvolvida, da qual se selecionou algumas empresas que se encontram sumariadas na (Tabela 3.4), constatou-se que as tecnologias Big Data estão a ser absorvidas pelo sector energético, ainda que de forma embrionária. Este tema, constituído por enormes desafios e oportunidades, tem ainda um longo caminho a percorrer.

[4] https://www.energycentral.com
[5] https://www.greentechmedia.com/
[6] https://www.solarplaza.com/channels/
[7] https://www.gartner.com/en
[8] https://www.datasciencecentral.com

Tabela 3.4 - Big Data no Sector Energético

Atividade Principal		EMPRESA	DESCRIPCÃO
	Energia	Agder Energi	É uma das maiores empresas de energia hidroelétrica da Noruega, produzindo cerca de 8,1 TWh de energia renovável anualmente. Para a gestão do fluxo de dados, utiliza os serviços disponibilizados pela plataforma *Microsoft Azure* [137]
		Allego	É uma empresa que fornece soluções para carregamento de carros elétricos. A sua plataforma de serviços em nuvem é baseada na tecnologia *Azure.* [138]
		C&J Energy Services	*C&J Energy Services* implementou uma nova plataforma para a gestão dos seus dados com base na plataforma MapR. Integra ainda o Apache Drill para acelerar o acesso aos a dados históricos do seu ERP (*Enterprise Resource Planning*) [139]
		Centerica	É uma empresa distribuidora de energia. Para melhor gerir os seus dados e servir os seus clientes, a empresa apostou claramente na plataforma *Hortonworks.*[140]
		Chevron	É uma empresa multinacional do ramo energético, com investimentos em todas as áreas das indústrias do petróleo, gás natural e energia geotérmica. Para uma melhor eficiência e agilidade na gestão dos seus dados decidiu migrá-los para *cloud*. (e tirar partido dos serviços da plataforma Azure) [141]
		e.on	É uma empresa de distribuição de energia. Para melhor servir os seus clientes desenvolveu a solução *E.ON Home*, em parceria com a Microsoft [142]
		Exelon	Exelon é uma grande empresa de energia responsável por abastecer milhões de clientes em 48 estados do Canadá. Produz energia eólica, solar, hidroelétrica e nuclear. Para a gestão eficiente do seu ativo de dados selecionou como solução a plataforma *Predix* [143]. Em parceria com a *General Electric Digital,* detentora desta plataforma, desenvolveu a suas própria solução que lhe permite analisar de forma reativa e proactiva toda a sua atividade, e.g., previsão de produção de energia renovável, monitorização e prevenção do equipamento, etc. [144]
		Podo	Podo é uma empresa espanhola distribuidora de energia. A análise avançada dos dados e a disponibilidade dos seus serviços é operada com recurso a uma plataforma da empresa *Cloudera.*[145]
		Schneider Electric	A *Schneider Electric* desenvolve tecnologias e soluções para a gestão e automação da energia de uma forma segura, fiável, eficiente e sustentável. A sua solução *EcoStruxure™ Platform,* desenvolvida em parceria com a *Microsoft* fornece um conjunto de serviços em *cloud* [146]. Integra ainda uma plataforma mobile para a previsão e deteção de anomalias em equipamento desenvolvido com as tecnologias disponibilizadas pela *DataRPM* da *Progress* [147].
		Southwest Power Pool	SPP é uma organização de transmissão de energia. A fim de garantir o equilíbrio entre a oferta e a demanda de energia recorreu à plataforma da Informatica. Desta forma consegui a centralização e replicação quase em tempo real de dados adquiridos de mais de 400 sistemas, reduzindo o tempo de análise destes dados de um dia para 20 minutos.[148]
		Vector	*Vector* é uma das principais empresas de fornecimento de serviços de energia e comunicação da Nova Zelândia. *Vector* selecionou a plataforma *mPrest* para a gestão e análise dos seus dados.[149].
		Vestas	É uma das maiores empresas mundiais da área de produção de energia eólica. Face à necessidade de melhorar a gestão e análise dos seus dados, adquiriu a empresa *Utopus Insights*, e desenvolveu a sua própria plataforma. *Scipher*, é uma plataforma de análise de energia escalável, segura e flexível, que permite a ingestão e análise em tempo real de um grande volume de dados [150].
	Software	AutoGrid EPI	*Energy Internet Platform* (EIP), Plataforma unificada que fornece serviços em *cloud* para o controlo, previsão, e otimização de recursos em tempo real
		Enel X.	*Enel X* faz parte do grupo *Enel*, uma das maiores empresas do sector energético, e desenvolve soluções para empresas do sector energético. As principais áreas de atuação são: resposta à demanda; energias renováveis; *microgrids*; armazenamento de energia; eficiência energética; carregamento de EVs; gestão do fornecimento de energia. [151]
		eSmart Systems	Desenvolve soluções de software, com foco em Inteligência Artificial. A plataforma *eSmart* disponibiliza serviços na área de operação e prevenção da rede elétrica, e ainda para os mercados de energia. Os serviços são disponibilizados em cloud através da plataforma Azure. [152]
		General Electric Digital	*GE Digital* é uma subsidiária da multinacional *General Electric*, e tem como missão o desenvolvimento de software e a prestação de serviços de consultoria na área de tecnologia operacional. A empresa desenvolveu a plataforma *Predix* [143] com foco na indústria da internet das coisas (IIOT). A plataforma permite a recolha, centralização e análise de grandes volumes de dados provenientes de uma ampla variedade de fontes, e.g., máquinas, sensores, sistemas de controle, dispositivos, etc. Foi concebida para respostas eficientes ao tratamento de dados de diversas atividades, e.g., geração, transporte e distribuição de energia, aviação, saúde, etc. Para além dos serviços disponíveis, permite ainda, o desenvolvimento de soluções personalizadas. Os Serviços providos pela Plataforma podem ser acedidos via *cloud* da *Microsoft Azure* [153].
		mPrest	*mPrest*, empresa sediada em Israel, desenvolve software de monitoramento, controle e análise. A sua plataforma *mPrest*, fornece serviços de análise, integração e otimização de dados em tempo real, para várias ares de atividade, e.g., segurança, cidades inteligentes, energias distribuídas, etc. A plataforma pode ser implementada localmente ou em nuvem a partir de provedores como *AWS* ou *Microsoft Azure* [154].
		Oracle	Oracle disponibiliza uma série de servições na área de eficiência energética. Em parceria com *Cloudera*, disponibiliza os serviços da sua plataforma, i.e., *Oracle Utilities Opower* Utilities em *cloud.*[155]
		Southern Company	É uma das maiores empresas produtoras de energia dos EUA. Para melhorar a gestão e integração centralizada dos seus dados recorreu à solução disponibilizada pela Informatica. [156]

O sector energético, face aos seus grandes desafios na gestão e análise de dados, representa uma grande oportunidade de negócio para o sector tecnológico. As empresas de desenvolvimento de software estão atentas a este fenómeno. Novas *startups* nascem a cada dia com a missão de se especializarem no desenvolvimento de soluções para o sector energético e ambicionando que estas possam contribuir para a sustentabilidade e eficiência energética. Estas empresas inovadoras estão a ser absorvidas pelas grandes empresas do sector energético (e.g. *Utopus Insights* pela *Vestas*) ou pelas grandes empresas do sector de tecnologias (e.g., *Opower* pela *Oracle*). Assim, conforme se pode constatar pela análise da Tabela 3.4, a absorção das tecnologias Big Data está a ser implementada no sector energético de forma direta (i.e., contratando produtos e/ou serviços, ou adquirindo empresas cuja missão é o desenvolvimento de soluções ramo Big Data), ou indiretamente (i.e., contratando produtos e/ou serviços, ou adquirindo empresas cuja missão é o desenvolvimento de soluções para o sector energético e que já implementam tecnologias Big Data).

Verifica-se ainda que apesar dos desafios inerentes ao sector energético ser transversal a todo o sector, (i.e. geração, transporte e distribuição), a interação entre as duas áreas ocorre com maior frequência no domínio *Distributed Energy Resource*. Por outro lado, verifica-se que quando a solução recai sobre a aquisição de serviços, *cloud* é geralmente a opção selecionada. Finalmente, apesar da complexidade inerente à gestão de dados, o foco principal continua a ser a procura de plataformas avançadas de análise de dados, para deles extrair o precioso valor imprescindível à tomada de decisão em tempo real.

Por fim, para consolidar este estudo foi feita uma análise sobre a posição da UE relativamente ao envolvimento destas duas áreas, e de que forma a área de investigação e desenvolvimento (I&D) pode constituir uma valência no assunto. Da pesquisa efetuada na página oficial da UE [157], com o termo 'Big Data', verificou-se que desde 2010 até à data corrente foram aprovados 676, sendo 603 do programa H2020 e 73 do programa FP7. Pela análise do gráfico representado na Figura 3.2 (a) comprova-se uma tendência crescente na adoção do termo 'Big Data'.

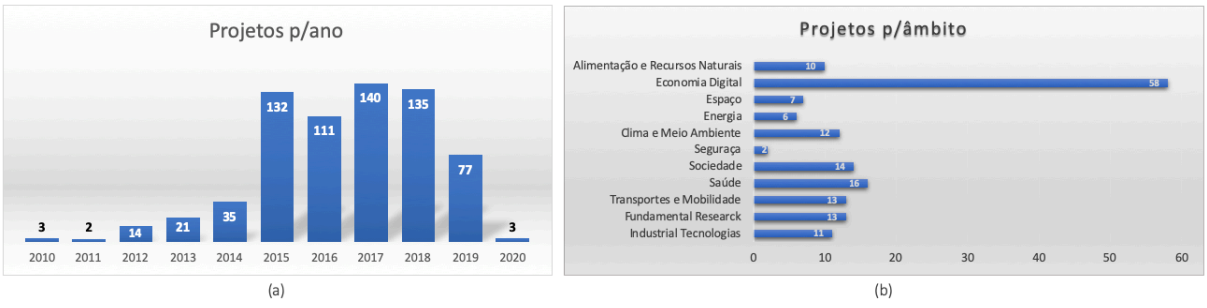


Figura 3.2- Projetos financiados pela UE com referência a Big Data (valores atualizados em 05/2019)

Afim de se analisar o impacto sectorial do termo pesquisado, foram aplicados os filtros disponíveis no site, cujos resultados obtidos se apresentam na Figura 3.2 (b). Do total de 676 projetos apenas foram obtidos 162 resultados, representando um valor inferior relativamente ao número de projetos classificados quanto ao seu âmbito. Constata-se assim que a maioria dos projetos não está classificada

quanto ao seu âmbito de aplicabilidade. Por outro lado, dos projetos classificados verificou-se que o seu âmbito incide sobre áreas distintas. Por exemplo o projeto 'MATRIX CHARGING', que visou o desenvolvimento de uma nova tecnologia para carregamento de EV, está classificado em três áreas (i.e., *energy, trasport and mobility, digital economy*) [158]. Desta forma os resultados apresentados no gráfico Figura 3.2 (b) são irrealistas, visto que a maioria dos projetos classificados se cruzam na área de *digital economy*. Não sendo possível obter por esta via a resposta à questão formulada, foi feita nova pesquisa fundamentada na informação da UE, onde identifica os projetos financiados no âmbito de Big Data [159]. Da análise desta informação foi possível identificar 68 projetos financiados desde 2015, conforme representado na figura (Figura 3.3) (a). Estes projetos resultaram num investimento superior a 329 Milhões de EUR, dos quais 290 (i.e., 88%) foram suportados pela UE e o restante pelo sector empresarial, conforme mostra o gráfico Figura 3.3 (b).

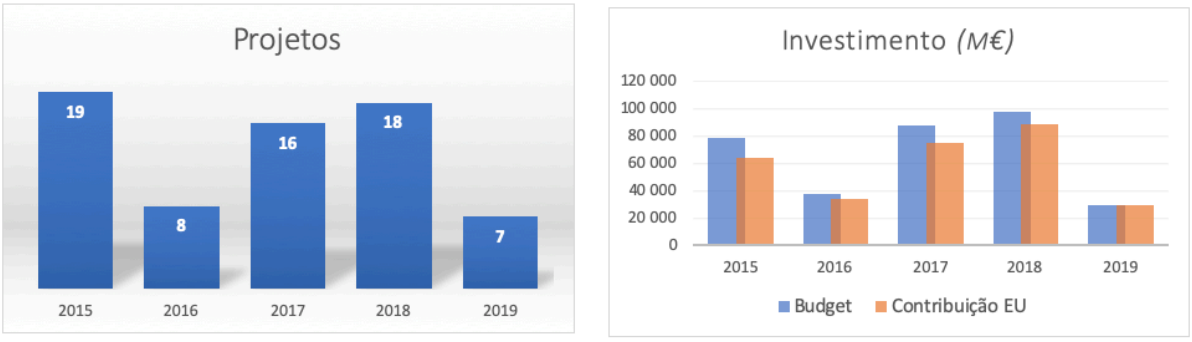


Figura 3.3 - Projetos financiados pela UE no âmbito de Big Data (valores atualizados em 05/2019)

Quanto ao seu impacto sectorial, chegou-se à mesma constatação. i.e., a maioria dos projetos não estão classificados quanto ao seu âmbito de aplicabilidade. Informação mais detalha sobre estes projetos pode ser consultada na tabela do Anexo 1.

A informação obtida até então não é conclusiva quanto à posição da UE sobre a envolvência de Big Data no contexto das Smart Grids. No entanto, nos subsecivos relatórios emitido pela organização *Big Data Value Association* BDVA[9] [160]–[162], é possível constatar que o sector energético é um dos visados para o financiamento no âmbito de Big Data.

Assim, conclui-se que a Europa reconhece claramente o valor acrescido pelos investimentos na área de Big Data para impulsionar desenvolvimento económico, transversal a todos os sectores no qual se inclui o sector energético.

[9] *Big Data Value Association* (BDVA) é uma organização internacional sem fins lucrativos, constituída por mais de 180 membros de toda a Europa, que representa de forma equilibrada as grandes, pequenas e médias indústrias, bem como organizações de investigação e desenvolvimento. BDVA é a contrapartida privada da Comissão da UE para implementar o programa *Public-Private Partnership* (PPP) *Big Data Value*. A missão da BDVA é desenvolver o Ecossistema de Inovação que permitirá a transformação digital orientada por dados na Europa, proporcionando o máximo benefício económico e social, e alcançando e sustentando a liderança da Europa na criação de Valor a partir de Big Data.

3.4 Conclusão

Este capítulo revisou os principais conceitos que definem as Smart Grids, destacando a sua evolução no cenário UE. O fluxo de dados do ecossistema energético foi caracterizado e identificado como um ativo fundamental para o desenvolvimento das Smart Grids. Revisou-se a importância e o impacto da evolução tecnológica operada na área de Big Data no contexto das Smart Grids em três vertentes: desenvolvimentos realizados na comunidade científica, impacto sectorial de Big Data na área de Smart Grids e finamente a posição da UE sobre investimentos na área de Big Data no contexto da Smart Grids.

Da análise efetuada ao longo deste capítulo conclui-se que a maturidade na aplicação de tecnologias Big Data no ecossistema energético está ainda num estágio inicial. É unanime a posição de todos, i.e., comunidade científica e setorial, de que esta nova era digital é bastante promissora para o desenvolvimento das Smart Grids. No entanto, é imprescindível incorporar nas Smart Grids tecnologias de Big Data para a gestão e análise do seu complexo fluxo de dados. Só desta forma será possível extrair o valor essencial para a tomada de decisão em tempo real, a fim de garantir a eficiência e sustentabilidade do ecossistema energético.

4 Soluções propostas e pesquisa experimental

Neste capítulo são descritas as várias soluções propostas e desenvolvidas no âmbito do presente trabalho. É descrito o cenário experimental sobre o qual foram planeadas e avaliadas as várias soluções. É ainda introduzido o conceito da arquitetura adotada na solução selecionada, i.e. *Docker Container*, face ao seu grande impacto positivo no sucesso da mesma. O capítulo termina com uma breve conclusão sobre os aspectos principais que conduziram à solução adotada.

4.1 Introdução

O presente trabalho tem como objetivo a experimentação e a validação das novas tecnologias desenvolvidas no âmbito de Big Data no contexto das Smart Grids. Um dos principais focos da experimentação visa o processamento em tempo real de um grande volume de dados, de forma a contribuir para desafios como a previsão de consumo/produção de energia. Estas duas variáveis são apontadas como determinantes para a eficiência do ecossistema energético. Muitas das operações executadas no ecossistema energético dependem invariavelmente da precisão destas duas variáveis.

Por outro lado, verifica-se que existem ainda muitos desafios apontados na área de Big Data de forma a extrair o valor dos dados em tempo real. Um dos grandes desafios identificado está relacionado com a área de análise. Estima-se que aproximadamente 80% do esforço na análise de dados seja despendido na preparação de dados e que apenas 20% do tempo seja consumido no processamento da metodologia de análise. Assim, as várias propostas descritas neste capítulo, visam dar uma resposta positiva face aos problemas identificados. Para além deste, outros desafios relacionados com a complexa gestão do fluxo de dados, como a monitorização e detecção de anomalias, a otimização de recursos e a eficiência em programas de resposta à demanda são tidos em conta no planeamento das soluções propostas.

Importa ainda, identificar o cenário de experimentação, conforme mostra Figura 4.1 e que a seguir se descreve.

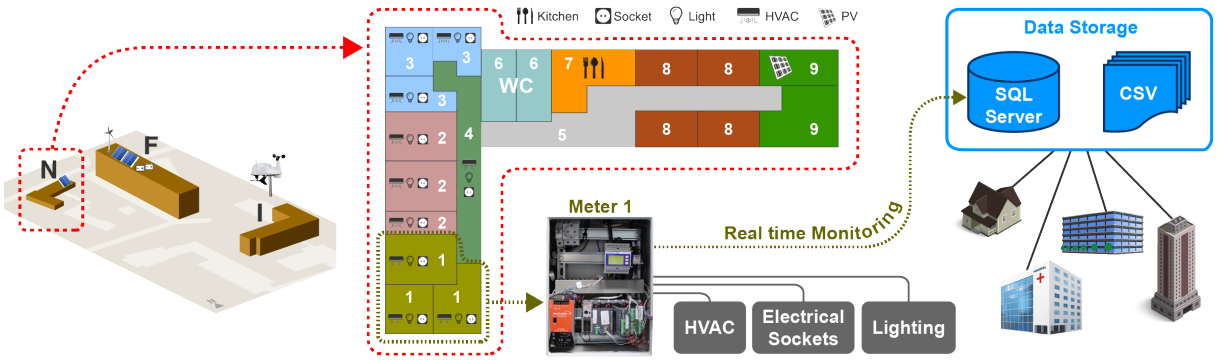


Figura 4.1- Fontes de dados disponíveis na MicroGrid do GECAD

O cenário para a experimentação do presente trabalho é caracterizado por uma infraestrutura de micro-rede existente no GECAD [163]. O laboratório de micro-rede do GECAD inclui vários edifícios no campus do Instituto de Engenharia do Politécnico do Porto (ISEP / IPP), com geração de energia (fotovoltaica - PV e eólica). O edifício N, conforme representado na (Figura 4.1) está dividido por zonas. Cada zona é caracterizada pelo agrupamento de várias salas. O edifício está equipado com analisadores que medem o consumo/produção de energia a cada 10 segundos, em cada uma das zonas. Por sua vez, cada analisador está subdividido em três grupos de carga: Ar Condicionado (HVAC); iluminação; e tomadas. Mais detalhes sobre a micro-rede GECAD podem ser encontrados em [164]. Outros dados externos ao ecossistema energético estão a ser capturados por sensores, tais como a temperatura, a humidade e a presença, ainda que de forma experimental. Finalmente, todos estes dados são armazenados numa base de dados SQL Server. Para além destes, o histórico de dados é ainda composto por um vasto conjunto de dados disponibilizados para as áreas de investigação relacionadas com as Smart Grids [165].

De acordo com o cenário disponível para experimentação e os grandes desafios inerentes ao ecossistema energético, a seguir são descritas as várias abordagens identificadas como possíveis soluções.

4.2 Extensão da SMACK Stack

No início do presente trabalho, os avanços tecnológicos no âmbito da área Big Data, encontravam-se ainda num estágio muito embrionário. Os avanços e experimentações na área de processamento pouco mais tinham avançado para além do ecossistema Hadoop. Hadoop, conforme descrito no capítulo 2, visou colmatar o grande desafio no processamento de um grande volume de dados. No entanto, nesta altura a framework *Apache Spark*, projetada para dar resposta ao processamento em tempo real, começava a ganhar algum ênfase. Assim, tendo como principal objetivo a procura de uma solução que permitisse o processamento de um grande volume de dados em tempo real, requisito fundamental para a gestão de fluxo de dados no ecossistema energético, a primeira proposta foi feita tendo como base a arquitetura SMACK Stack [166]. Esta abordagem é composta por vários componentes (i.e. *Apache Spark*, Mesos, Akka, Cassandra, Kafka) do qual se destaca a biblioteca *Spark Streming,* da *framework Apache Spark,* pela sua superior capacidade de processamento, i.e., mais rápido relativamente ao *MapReduce* do *Hadoop*. Para além desta biblioteca, a *framework Apache Spark* é composta por um conjunto de bibliotecas que lhe confere um elevado grau de interoperabilidade com outros componentes. Assim, tirando proveito desta interoperabilidade, propôs-se a integração de alguns componentes, de forma a enriquecer as camadas de análise e visualização de dados, conforme mostra Figura 4.2.

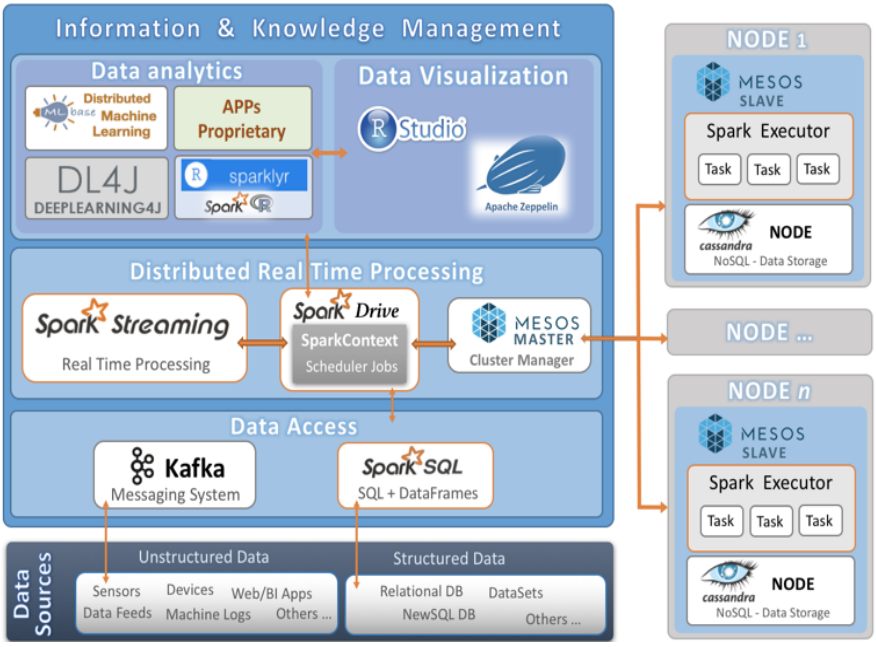


Figura 4.2- Extensão da SMACK Stack

A arquitetura proposta é composta por várias camadas como a seguir se descreve:

- Camada de análise, que inclui os seguintes componentes:
  - *MLbase* - É um projeto de código aberto desenvolvido para otimizar e simplificar a construção de algoritmos de *Machine Learning* em ambientes distribuídos [167];
  - *DL4J* - Deeplearning4j é uma *framework* projetada para a execução de algoritmos de *Deep Learning*, desenvolvida em Java e *Scala*, e de acesso livre [168];
  - *Sparklyr* – Tal como a API SparkR [169], este projeto disponibiliza uma interface para a conexão com o RStudio [170]. No entanto, disponibiliza funcionalidades adicionais, em relação ao SparkR, das quais se destinge o tratamento mais avançado na manipulação de fontes de dados, através de operações como a agregação, a transformação e a filtragem [171]. Para além de facilitar a conexão com o RStudio, é completamente compatível com a biblioteca de análise disponibilizada na *framework* Spark, i.e., MLlib [172];
  - *Apps Proprietary* – Aplicações e algoritmos proprietários do GECAD.
- Camada de visualização: Para esta camada propôs-se o componente Apache Zeppelin [173]. O Apache Zeppelin é um notebook de código aberto, baseado no serviço web e compatível com a *framework* Apache Spark. Na versão 0.5, último lançamento oficial aquando desta proposta, Zeppelin ainda não suportava a interoperabilidade com a linguagem de programação R. Para colmatar este problema, propôs-se a integração do RStudio que disponibiliza uma interface para visualização;

- Camada de processamento: propôs-se a execução distribuída com suporte ao Spark Streaming;
- Camada de acesso a dados: Esta camada é responsável por ingerir com os dados na plataforma, cuja tarefa é desempenhada pelos componentes SparkSQL [174] (i.e., para acesso a dados históricos armazenados na base de dados SQL Server) e Apache Kafka [12] (i.e., para a recolha de dados não estruturados, tais como, dados de sensores e dados de analisadores);

Finalmente, todos os dados recolhidos em tempo real são armazenados em Cassandra, i.e., uma base de dados NoSQL. Cassandra é caracterizada pelo armazenamento do tipo coluna. É totalmente compatível com Apache Spark e, de acordo com a literatura [175], esta base de dados é uma excelente solução para armazenar séries temporais, fundamentada pelo seu bom desempenho.

A solução proposta foi implementada num cluster de três máquinas, 1 nó físico e 2 nós virtualizados, gerido pela ferramenta Apache Mesos [176]. As experiências iniciais desenvolvidas sobre a solução resultaram num bom desempenho da mesma, relativamente a arquiteturas mais tradicionais [177].

Prosseguiu-se com a avaliação do desempenho da solução relativamente à performance do armazenamento executado em Cassandra versos SQLServer, com auxílio da ferramenta YCSB. Os testes de desempenho efetuados não foram conclusivos relativamente à superioridade da Base Dados Cassandra. Por outro lado, Cassandra demostrou ser pouco flexível quanto à modelação de casos mais complexos, conforme descrito em [178].

Verificou-se ainda que em arquiteturas de várias camadas, como é o caso do SMACK Stack, os problemas de desempenho de latência acontecem com frequência quando ocorrem falhas nos seus componentes. Outro dos problemas desta abordagem é a sua limitada flexibilidade e complexidade acrescida na sua manutenção. Sempre que é necessário atualizar os componentes existentes para novas versões ou implementar novos componentes, torna-se necessário efetuar a paragem do sistema e muitas vezes a atualização e adaptação do código que estabelece a conetividade entre os vários componentes [179].

Detetou-se ainda que inúmeros avanços têm sido feitos nas várias áreas de Big Data (e.g., sistemas de striming, sistemas MOM, armazenamento, etc.). As novas abordagens propostas nesta área podem conter soluções interessantes para responder positivamente à complexidade crescente da gestão do fluxo de dados no contexto das Smart Grids. Como tal, merecem ser experimentadas e validadas neste contexto. Para que isso seja possível, é necessário desenvolver uma solução mais flexível. Com esse objetivo revisou-se a atual arquitetura com o propósito de se propor uma plataforma mais flexível, escalável e tolerante a falhas, conforme se descreve no subcapítulo seguinte.

4.3 Plataforma HDS

Para a proposta de uma solução mais flexível foram primeiramente revisadas e atualizadas as novas abordagens tecnológicas operadas na área de Big Data, e a forma como estas estão a ser implementadas no contexto das Smart Grids, conforme descrito nos capítulos 2 e 3. Foi ainda necessário revisar e experimentar os avanços tecnológicos feitos na área de arquiteturas de sistemas. Este último assunto acabou por ter grande importância e sucesso na nova solução proposta, ao ponto de esta ser a selecionada para o desenvolvimento do trabalho proposto nesta tese. Importa, antes de descrever a nova solução, referir, ainda que sumariamente, os principais conceitos inerentes à arquitetura que a sustenta, i.e., *Docker Containers* [180].

4.3.1 Docker Containers

Docker Containers (DC) é uma nova abordagem tecnológica na área de arquiteturas de sistemas que proporciona o isolamento e a interoperabilidade entre todos os componentes da arquitetura. Foi concebida com o objetivo de facilitar o desenvolvimento, a implantação e a execução de aplicações em ambientes isolados. Possibilita a disponibilização de aplicações de forma mais rápida, uma vez que agiliza todo o processo de criação, manutenção e modificação das mesmas. É escalável e flexível, facilitando a criação de cenários de acordo com o que se pretende simular.

Esta arquitetura distribuída é mais leve e apresenta melhor desempenho quando comparada a arquiteturas implementadas com Máquinas Virtuais. Isto porque o conceito de virtualização dos DC se baseia no isolamento de processos e não no isolamento do sistema operativo como acontece com as máquinas virtuais. DC distinguem-se ainda por outras vantagens, como: portabilidade; reutilização; resiliência; confiabilidade, tolerância a falhas, automatização no balanceamento de cargas, etc. [181].

Breve história do Docker

A tecnologia DC teve a sua origem numa organização, *platform as a service* (PaaS), denominada *dotCloud*. O projeto desenvolvido surgiu pela necessidade de atender positivamente à necessidade crescente dos seus clientes. Com o crescente aumento da sua infraestrutura composta por servidores e máquinas virtuais era necessário criar soluções para agilizar o seu desempenho e ao mesmo tempo facilitar a complexidade da sua manutenção. Em Março de 2013 o projeto Docker foi disponibilizado como *open source* e a empresa *dotCloud* passou a ser denominada por Docker. A partir de então esta tecnologia tem ganho grandes adeptos e tem sido alvo de grandes avanços [182]. Em 2015 foi criada a *Open Container Initiative* (OCI) cujo objetivo é desenvolver e especificar os padrões do formato *container* de forma a garantir a sua compatibilidade em todas as plataformas [183].

Em março de 2017 a empresa subdividiu o projeto Docker em duas versões: versão *Docker Edition Enterprise* (Docker EE); versão *Docker Edition Community* (Docker EC), conforme Figura 4.3. A versão EE exige a assinatura da plataforma, disponibiliza suporte técnico, tem integrada a orquestração

de contentores, segurança e ambiente gráfico para gestão e manutenção dos *containers*. Por sua vez, a versão CE tem como lema "Faça você mesmo", mas é de livre acesso. Esta alteração não inviabiliza a possibilidade de se instalar e usar o Docker CE em ambientes de produção, mas exige muito mais trabalho e complexidade especialmente para quem se inicia nesta nova tecnologia [184]–[186].

Figura 4.3 - Docker Edition Community vs Docker Edition Enterprise

Nesta mesma data a empresa anunciou o início de um projeto denominado *Moby* para promover o avanço e desenvolvimento de *containers* de software, em parceria com várias organizações como Microsoft, IBM, Google, etc. *Moby* é um projeto open source e devido aos esforços realizados no seu âmbito, hoje já é possível executar *containers* nativamente em vários sistemas operativos (i.e., sistemas baseados em Linux, Windows server 2016 e Windows 10) localmente e/ou em *cloud* publica e/ou privada [187].

Docker Container e Virtualização

O conceito de virtualização em DC baseia-se no isolamento de processos enquanto que a tradicional virtualização (i.e., *Virtual Machine* (VM)) se baseia no isolamento de sistemas operativos, conforme Figura 4.4. Neste último caso, a virtualização pode ser caracterizada por dois tipos distintos:

- *Bera-metal* – Conforme mostra Figura 4.4 (a), o software de virtualização, geralmente denominado por *hypervisor*, é instalado diretamente sobre o hardware. Este tipo de software cria dispositivos de hardware artificial com tudo o que é necessário para executar outros sistemas operativos (e.g., Windows, MacOS, Linux, Unix, etc.) no mesmo *host,* de forma completamente isolada. Hyper-V, Xen, e VMware ESXi, são alguns exemplos de software deste tipo de virtualização;
- *Hosted Virtualization* – Neste caso o software de virtualização é instalado sobre o sistema operativo. A camada de virtualização, conforme mostra Figura 4.4 (b), contém todo o software

necessário para hospedar e gerir as máquinas virtuais. No entanto, neste tipo de abordagem os sistemas virtualizados dependem do Sistema Operativo do *host* para aceder ao hardware. Alguns exemplos deste tipo de software são: VirtualBox, Microsoft Virtual PC e VMware Workstation.

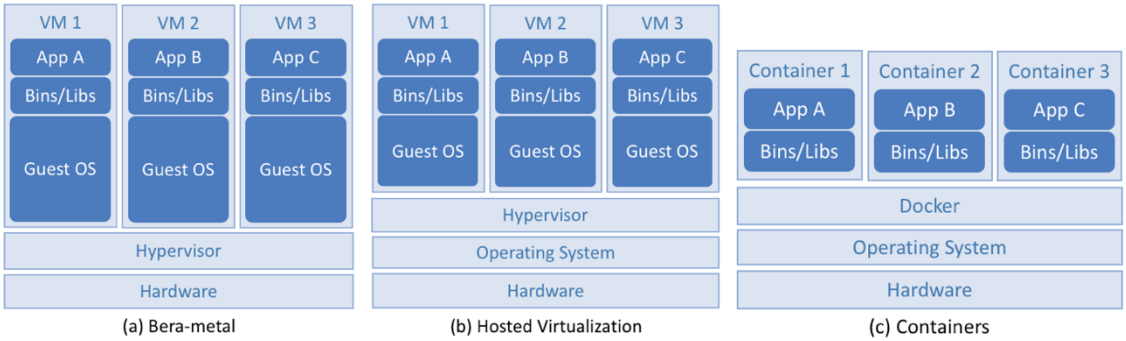


Figura 4.4 - Virtualização vs Containers

No caso dos *containers*, conforme mostra Figura 4.4 (c), as aplicações para correrem de forma isolada não necessitam da instalação de um sistema operativo dedicado à sua execução. Para criar o isolamento necessário das aplicações, a plataforma Docker utiliza funcionalidades do *Kernel* do sistema operativo do *host*. O processo de isolamento é denominado por *Containers*. Os *containers* "empacotam" a aplicação e tudo que é necessário para a sua execução (i.e., código, ferramentas, bibliotecas, etc.). São isolados a nível de disco, memoria, CPU e rede. Os processos de uma aplicação em execução num determinado *container* não têm acesso aos processos a serem executados em outros *containers*, a menos que seja explicitamente configurado. Depois de configuradas e devidamente testadas as aplicações podem ser transformadas em *Imagens*. No processo de construção de uma *imagem* é-lhe adicionado um *template* a partir do qual é possível instanciar aplicações isoladas em *containers* segundo determinadas características. A ideia é construir uma vez e executar onde se quiser e sempre que for necessário.

Docker Engine vs Docker-Machine

A Figura 4.4 (c) apresenta o cenário DC a correr nativamente sobre o sistema operativo do *host*. No entanto, no caso de ser necessário instalar DC num servidor com um SO não contemplado na Figura 4.3, é ainda possível criar um ambiente DC a partir de uma máquina virtual (de preferência do tipo *Bera-metal*, uma vez que tem um desempenho superior relativamente ao tipo *Hosted Virtualization*). É igualmente possível desenvolver um cenário misto, ou seja, num mesmo servidor: DC a ser executado nativamente no *host* na máquina física [188] (i.e. *Docker Engine*), e o DC a ser executado a partir de máquinas virtuais (i.e. *Docker-Machine*) [189], conforme mostra Figura 4.5. Este último, é um cenário particularmente interessante no caso de ser necessário executar *containers* Windows e Linux no mesmo *host* [190]. No entanto, é necessário referir que o *Hyper-V* nos sistemas Windows é um requisito essencial para a instalação do *Docker Engine* [188]. Algumas versões Windows 10 não disponibilizam esta funcionalidade (Windows 10 Home) e outras não trazem esta funcionalidade ativa, o que não invalida a possibilidade de executar a sua ativação conforme explanado em [191]. Ainda de referir que

em sistemas Windows poderá ser necessário ativar as opções de virtualização na *bios* da máquina. No caso de sistemas operativos Mac o hipervisor é garantido pela *framework* HyperKit. No entanto, é indispensável compreender a interoperabilidade entre o *Docker Engine* e os *hosts* criados com o *Docker-Machine*, bem como os vários cenários em que esta arquitetura poderá ser implementada, como definido em [192].

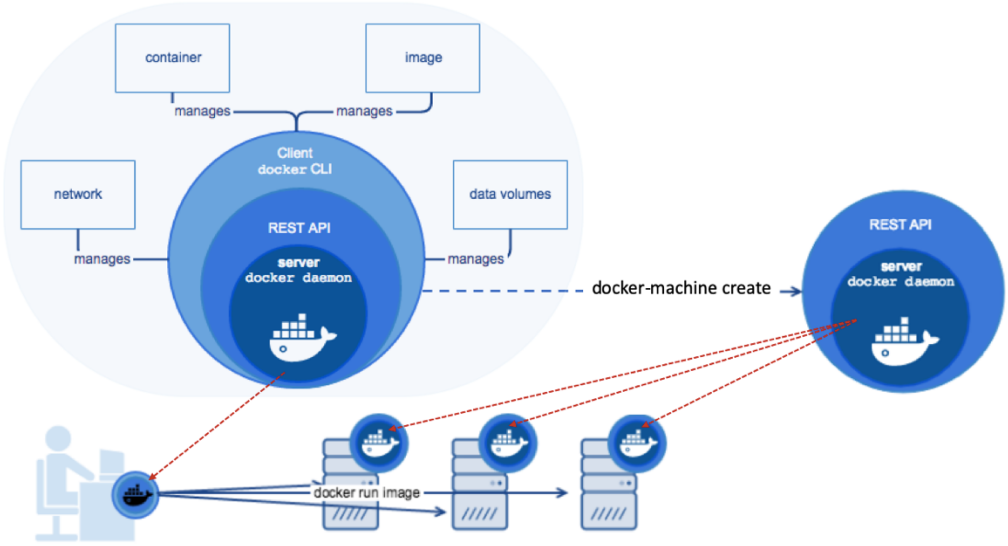


Figura 4.5 - Docker Engine vs Docker-Machine (adaptado de *[189]*)

O *Docker Engine*, conforme representado na Figura 4.5, é uma aplicação cliente servidor para a execução de *containers*. É composta pelo *Docker Deamon* (i.e., o servidor propriamente dito) e pelo Docker CLI (i.e., aplicação de linha comandos a partir da qual é possível solicitar pedidos ao *Docker Deamon*). REST API é o protocolo através do qual cliente e servidor se comunicam. *Docker Engine* permite a gestão de *imagens*, *containers*, *redes* e *volumes de dados*.

Docker-Machine é uma interface a partir da qual é possível instalar e gerir o *Docker Engine* em *hosts* virtuais, que podem residir, ou não, na mesma máquina física, ou em provedores de serviços em *cloud*, e.g., Microsoft Azure, Amazon Web Services, Digital Ocean, etc. Por defeito, em sistemas Windows o *docker-machine* instala os *hosts* virtuais no hyper-v. Em sistemas Mac e Linux os *hosts* virtuais são instalados na VirtualBox.

Orquestração de Docker Containers

A orquestração de Dokers permite a gestão de *containers* a serem executados em múltiplos *hosts* físicos e/ou virtuais. O *Docker Engine*, a partir da versão 12.1, disponibiliza e incorpora um orquestrador, i.e. Docker Swarm [193]. Existem, no entanto, outros orquestradores, como, por exemplo, Kubernetes[10],

[10] https://kubernetes.io

DC/OS[11], Cattle[12]. O Docker Swarm permite definir cada nó do cluster como gestor ou trabalhador. Os nós gestores por norma não executam *containers*. A sua principal função é gerir o cluster. Por sua vez, os nós trabalhadores têm como função executar as tarefas enviadas pelos gestores. Ao iniciar-se o Swarm, este cria automaticamente um gestor. No entanto, são necessários pelo menos três gestores para que o cluster seja tolerante a falhas. O Swarm adotou o algoritmo Ralf [194] para a definição do número de nós possíveis de estarem em falha sem inviabilizarem o seu funcionamento, i.e., $(n+1)/2$, em que n representa o número de nós do cluster. Para além da resiliência, o Swarm é caracterizado por: escalabilidade dos serviços, segurança, balanceamento automático da carga de trabalhos e descoberta de serviços.

4.3.2 Modelo conceptual – Plataforma HDS

A fim de validar a arquitetura DC, foram feitos alguns testes inicias. Desenvolveu-se um cluster constituído por um desktop com Windows 10 e um Laptop com MacOS, orquestrado pelo Swarm. Foram executados alguns exemplos disponibilizados na página oficial do DC. Esta experimentação inicial permitiu verificar o balanceamento de cargas e a comunicação entre os *containers* dos diferentes *hosts*. Face ao sucesso obtido, DC foi adotado como a arquitetura base da solução a propor conforme mostra figura seguinte.

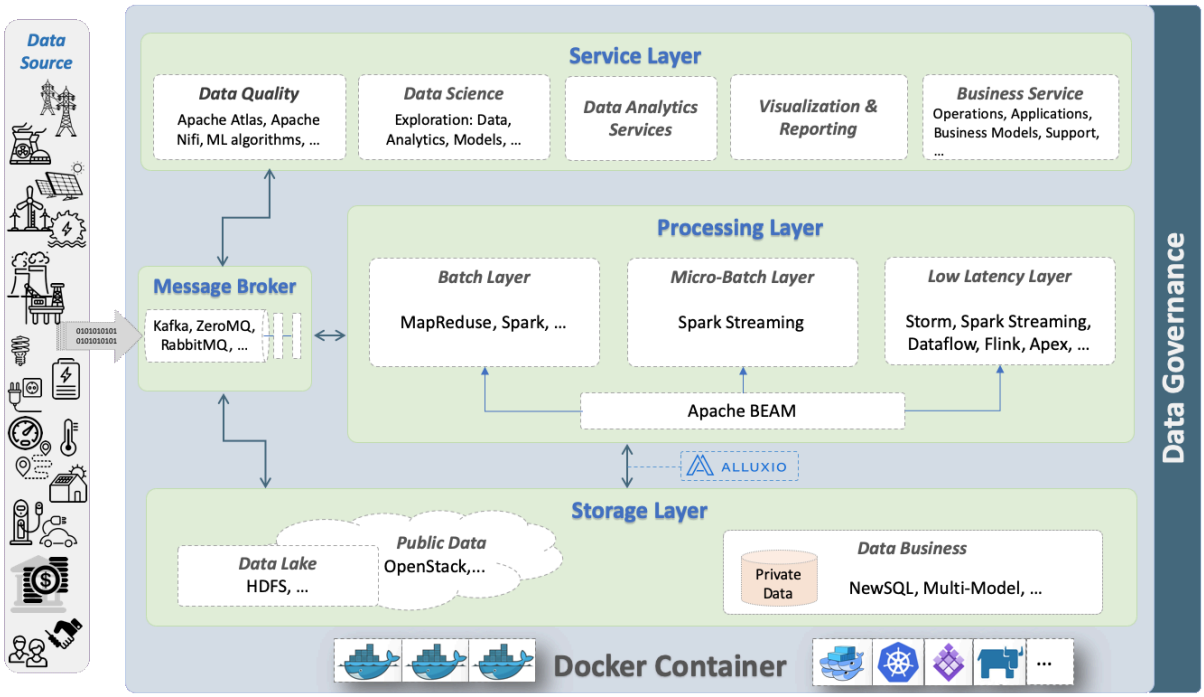


Figura 4.6- Modelo conceptual da plataforma HDS

[11] https://dcos.io
[12] https://rancher.com/rancher-glossary-1-6-to-2-0-terms-concepts/

O modelo proposto é composto por várias camadas, i.e., serviços, processamento e armazenamento. Os pontos chaves da arquitetura são:

- Simulação dos vários sistemas de transporte de mensagens, fundamental para a recolha e comunicação entre componentes e operações. Propõe-se a integração de soluções como por exemplo: Apache Kafkas, RabbitMQ, ZeroMQ, atendendo ao facto de que, conforme referido na literatura ([14] e [15]), os desempenhos dependerem de cada caso de uso;
- Simulação de vários sistemas de processamento onde se realça a inclusão do componente Apache Beam [21], [29]. Este componente é uma recente abordagem que visa a interoperabilidade entre vários sistemas de streaming;
- Na área de armazenamento pretende-se a simulação de várias abordagens com o objetivo de: garantir a centralização e correlação de dados de diversas fontes (fundamental para a descoberta de novos conhecimentos); e garantir a partilha de dados entre a comunidade de I&D sem pôr em causa a privacidade de dados pessoais imposta pelo novo regulamento (i.e., RPGD). Para facilitar a partilha de dados propõe-se o desenvolvimento de um Data Lake [66], disponibilizado numa cloud a desenvolver com base em abordagens open source, por exemplo, OpenStack [195]. Destaca-se ainda a possível inclusão do componente Alluixio [71], uma nova abordagem que visa a interoperabilidade entre diversos sistemas de ficheiros distribuídos e sistemas de streaming;
- Propõe-se ainda uma camada de serviços com o objetivo de assegurar as várias operações diretamente e indiretamente relacionadas com a operabilidade do ecossistema energético;
- Relativamente à segurança da arquitetura propõe-se a simulação de vários componentes que têm vindo a ser propostos para este fim, e.g. Apache Alfa [86], Apache Nifi [196] e Apache Range [87].

4.4 Conclusão

Neste capítulo foram apresentadas as várias soluções exploradas no sentido de darem uma resposta positiva aos problemas identificados como desafios nas áreas de Big Data e Smart Grids. A proposta inicial teve como principal foco a procura de uma solução que contribuísse para a minimização do problema de processamento em tempo real de grandes volumes de dados inerentes ao ecossistema energético. Nesse contexto, e extraindo o melhor que oferece a *framework* Apache Spark relativamente à performance, foi proposta uma solução baseada na SMACK Stack, à qual se propôs a adição de componentes para agilizar os processos relacionados com a análise e visualização dos dados.

No entanto, verificou-se que muitos avanços têm sido feitos na área de Big Data e que a arquitetura anteriormente proposta apresentava pouca flexibilidade para a experimentação dessas novas abordagens.

Por outro lado, verificou-se ainda um crescimento avultado na complexidade que envolve a gestão de dados no ecossistema energético, devido aos desenvolvimentos tecnológicos desta nova era digital que pouco a pouco vão impulsionando e tornando real a implementação das Smart Grids. Assim, face ao aumento da complexidade que envolve a gestão de dados no ecossistema energético, foi revisada a arquitetura inicial e propõe-se uma nova solução com o objetivo de facilitar a implementação e experimentação das novas abordagens que têm sido feitas na área de Big Data. Para tal, foi feita a revisão dos avanços tecnológicos operados na área de Big Data, bem como a forma como estes estão a ser implementados no ecossistema energético (conforme descrito nos capítulos 2 e 3). Por outro lado, foi ainda necessário revisar os avanços efetuados na área de arquiteturas de sistemas, a fim de encontrar uma solução mais flexível e ágil que permitisse a experimentação das novas abordagens propostas na área de Big Data no contexto das Smart Grids. Desta revisão conclui-se que a arquitetura *Docker Container* possui as características ideais para o objetivo pretendido. Assim, foi proposta uma nova solução, tendo como base para a sua implementação a arquitetura *Docker Container*. A nova solução, i.e., Plataforma HDS, visa a resolução dos desafios relacionados com a tomada de decisão em tempo real, sem, no entanto, deixar caminho aberto para a resolução de outros assuntos relacionados com as Smart Grids (e.g., Governação e segurança dos dados; partilha de dados para a comunidade I&D; processamento próximo do tempo real para a tomada de decisões de médio e longo prazo, de forma a melhor rentabilizar recursos de hardware).

5 Implementação da Stack HDS

Neste capítulo é apresentada a *stack* desenvolvida com o objetivo de testar o modelo proposto no capítulo anterior, bem como, contribuir para a minimização de alguns problemas identificados nas áreas de Big Data e Smart Grids. Ao longo dos subcapítulos seguintes serão ainda descritos alguns dos detalhes da implementação desta *stack*. Por fim serão resumidas as principais conclusões sobre este assunto.

5.1 Introdução

Conforme descrito no capítulo anterior, foram exploradas e propostas várias soluções com o propósito de validar e experimentar os avanços tecnológicos operados na área de Big Data no contexto das Smart Grids. Esse estudo culminou com a proposta de um modelo para a implementação de uma plataforma cujas principais características são a sua flexibilidade e agilidade, garantidas pela incorporação da arquitetura *Docker Container*. A fim de testar o modelo proposto pretende-se desenvolver uma *stack*, a *stack* HDS. Esta *stack* tem como principal objetivo contribuir para alguns desafios identificados na área de Big Data, tais como a redução no tempo de preparação de dados na área analítica e a promoção da automação do processo analítico, e na área das Smart Grids (processamento em tempo real da previsão do consumo/produção de energia, melhoria na precisão das previsões e deteção de anomalias).

Face aos avanços operados e disponibilizados pelo projeto *Big Data Europe* (BDE) [197], a primeira estratégia para a implementação da *stack* HDS consistiu na utilização da plataforma BDE à qual se adicionariam outros componentes a fim de se atingir os objetivos propostos. O projeto BDE, financiado pela UE tem como principal objetivo o desenvolvimento de uma plataforma Big Data que unifique o tratamento de fluxo de dados de vários sectores de atividade, i.e., agricultura, saúde, transportes, energia, segurança, clima e ciências sociais. A plataforma desenvolvida por este projeto é baseada em *docker containers*. A plataforma disponibiliza alguns componentes Big Data em *containers*: Kafka para o transporte de dados; Hadoop, Spark e Flink para o processamento; Elasticsearch para o armazenamento e Kibana para a visualização.

Foram feitas várias experiências a fim de implementar a plataforma BDE. A primeira experiência consistiu em instalar a plataforma num cluster com cinco nós, dois físicos e três virtualizados, orquestrados pelo *Swarm* e instalados em duas máquinas físicas. Uma das máquinas, com 32 Gb de memória e com o sistema operativo Windows 10, e outra com 8 Gb de memória e com o sistema MacOS. Na primeira máquina foi instalado o *Docker Engine*, que ficou definido como gestor, e criado dois nós virtuais definidos como trabalhadores. O *Docker Engine* foi configurado para executar os *hosts* em Linux. Por sua vez, na máquina com o sistema MacOs, foram criados dois nós, um gestor e um trabalhador. A experiência falhou devido à definição de partilha de volumes declarada no ficheiro

docker-compose da aplicação BDE (i.e., "*/var/run/docker.sock:/var/run/docker.sock*"). Muitos assuntos sobre esta matéria estão abertos no fórum dedicado ao desenvolvimento da tecnologia *Docker Containers* [198]–[200].

A segunda experiência consistiu na alteração do nó Windows por um nó com o sistema Ubuntu. Nesta experiência foi possível chegar à interface da plataforma a partir da qual se pode configurar um *workflow*. Após esta etapa, construída de acordo com os exemplos facultados pelo projeto BDE, a execução da mesma resultou numa espera interminável, sem nunca se poder verificar o resultado da execução do *workflow*. Durante a execução do *workflow* as máquinas indicaram um excessivo consumo de memória e CPU. Os requisitos de hardware não estão especificados, no entanto existe uma grande probabilidade de a experiência ter resultado em insucesso devido a recursos insuficientes.

Verificou-se ainda que para tirar o maior partido da gestão dos *workflows*, a utilização desta plataforma, como base para o desenvolvimento da *stack* HDS, implicava a dependência do componente kibana. Por outro lado, o *dashboard* Kibans, é um *plugin* projetado para a visualização de conteúdos indexados e armazenados no Elasticsearch. A dependência de determinados componentes não corresponde ao objetivo definido e proposto no modelo projetado para a plataforma HDS.

Assim, face aos resultados obtidos, decidiu-se implementar a *stack* HDS de raiz, conforme se descreve nos subcapítulos seguintes.

5.2 Stack HDS – Componentes

A *Stack* HDS foi inicialmente implementada em duas máquinas físicas, ao qual se adicionou uma terceira máquina para garantir a tolerância a falhas. As especificações das máquinas físicas bem como a configuração do cluster implementado podem ser analisados na figura seguinte.

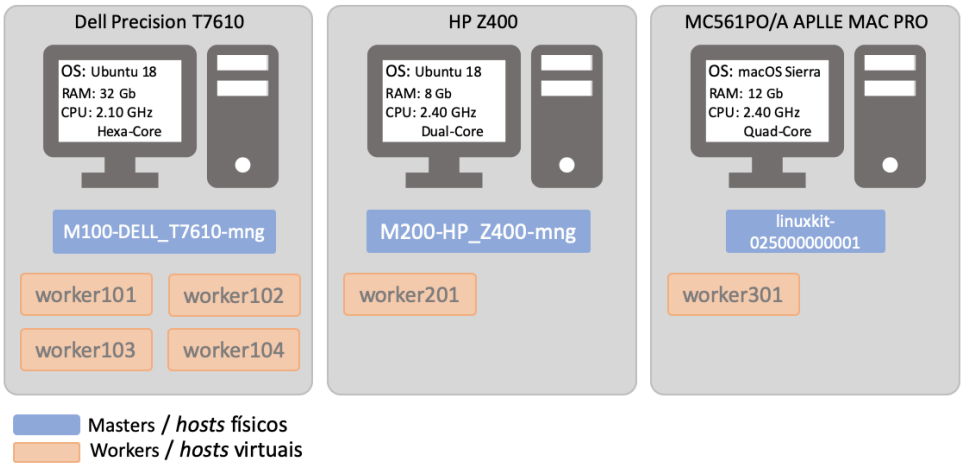


Figura 5.1 - Especificações do cluster Docker Container

Sobre este cluster de *Docker Containers*, orquestrado pelo *Docker Swarm*, foram instalados e desenvolvidos vários *containers* conforme mostra Figura 5.2.

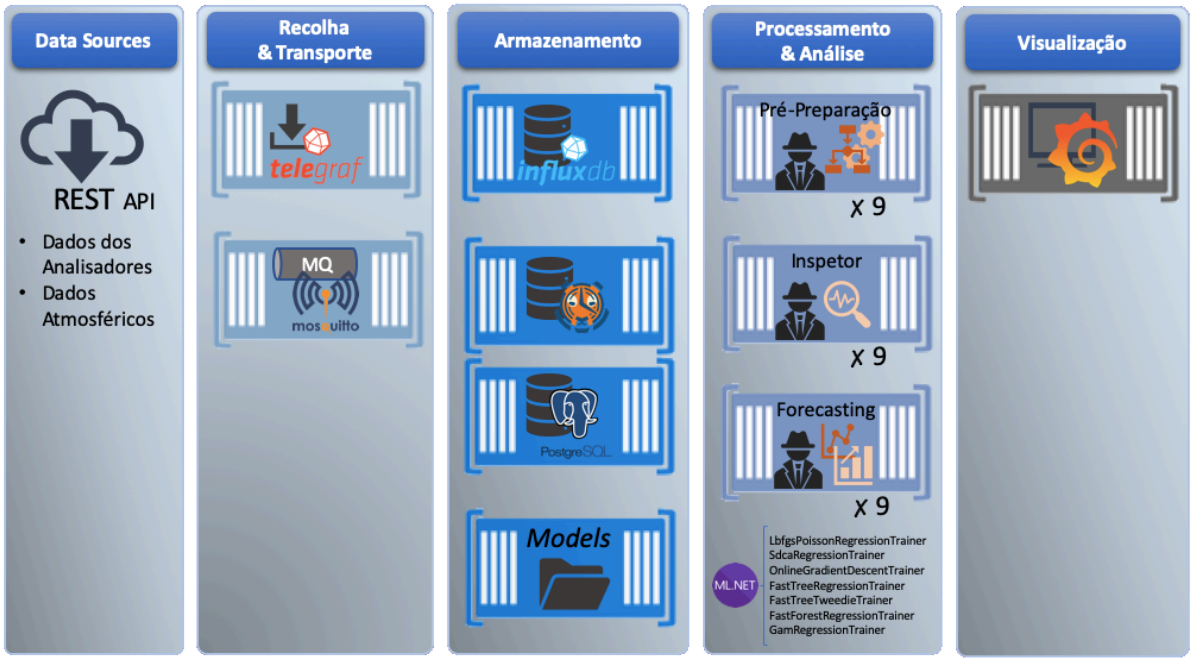


Figura 5.2 - Componentes da Stack HDS

Os componentes da *stack* HDS foram classificados de acordo com os serviços que disponibilizam, e que se descrevem nos subcapítulos seguintes.

5.2.1 Data Sources

Os fluxos de dados recolhidos e analisados pela *stack* desenvolvida são referentes aos dados de consumo e produção de energia do edifício N do campus do ISEP, no GECAD, conforme descrito no subcapítulo 4.1. Para além destes, são recolhidos e analisados dados atmosféricos, capturados por um sensor de temperatura instalado no exterior do edifício. Todos estes dados são obtidos através de uma aplicação *Representational State Transfer* (REST) desenvolvida para esse fim. Os pedidos têm de ser feitos por analisador e por sensor. Cada pedido retorna vários campos e.g., potência ativa, potência reativa, intensidade de corrente, tensão, etc. Um exemplo detalhado de todos os dados que estão a ser capturados através desta fonte pode ser analisado no Anexo 2. A Tabela 5.1 especifica os dados referentes às séries temporais a serem tratadas no âmbito da *stack* HDS.

Tabela 5.1 - Especificação de dados coletados através da API REST GECAD

Zona	Localização	Analisador	Campo	Métrica
1	Salas:101, 102, 103	Analyzer1_V4	Ph1_P	HVAC
			Ph2_P	Sockets
			Ph3_P	Lights
			P_Total	Total
2	Salas:104, 105, 106	Analyzer2_V3	Ph1_P	Sockets
			Ph2_P	HVAC

Zona	Localização	Analisador	Campo	Métrica
			Ph3_P	Lights
			P_Total	Total
3	Salas:107, 108, 109	Analyzer3_V3	Ph1_P	Lights
			Ph2_P	HVAC
			Ph3_P	Sockets
			P_Total	Total
4	common left, WC, hall	Analyzer4_V4	P1	Sockets
			P2	Lights
			P3	HVAC
			Ptotal	Total
7	Sala: 110	AnalyzerKitHall_V2	kitchen_ac_activePower	HVAC
			hallway_lights_activePower	Lights
			dishwasher_active	Sockets
			microwave_active	Sockets
			water_active	Sockets
			kettle_active	Sockets
8a	Salas: 111, 116	Analyzer116_V1	P1_W	Lights
			P2_W	HVAC
			P3_W	Sockets
			P_Total_W	Total
8b	Salas: 112, 115	Analyzer115_V1	P1_W	Sockets
			P2_W	HVAC
			P3_W	Lights
			P_Total_W	Total
9	Sobre as Salas: 113, 114	Inverter6_V3	N6_P	PV - Generação
bld	Building	∑ de todos analisadores	DR_shift	DR_shift
			DR_reduce	DR_reduce
			hvac	HVAC
			lights	Lights
			sockets	Sockets
			consumption	Consumption
			generation	Generation
-	Outdoor	Sensors_V1	Outdoor_temperature_x10	Temperatura /10

Para além dos dados referidos, estão ainda a ser coletados e analisados outros dados referentes às condições atmosféricas (i.e., dados correntes e dados de previsão), disponibilizados pelas empresas AccuWeather e Open Weather Map, através de RESTful APIs [201], [202].

5.2.2 Recolha & transporte

Nesta subsecção descrevem-se os componentes selecionados para os serviços de recolha e transporte de dados da *stack* HDS.

Telegraf

O telegraf é um componente desenvolvido e disponibilizado como *open source* pela empresa influxData, com o objetivo de facilitar a recolha e envio de métricas obtidas pelos dispositivos da IoT, pelos sistemas e por eventos de base de dados. É bastante flexível e fácil de configurar. Prima pela sua interoperabilidade com variadíssimos componentes, i.e., filas de mensagens, base de dados, etc. Por ouro lado, é compatível com vários protocolos de comunicação e formatos de input/output dos dados. Disponibiliza ainda funções para agregação e filtragem dos dados [203]. Este componente foi selecionado para o desenvolvimento da stack HDS pelas características que lhe são inerentes. A sua imagem oficial está disponível no repositório Docker Hub [204].

Mosquitto

O Mosquitto é um *message broker*, *open source*, que implementa o protocolo MQTT. É extremamente leve e adequado para projetos que dispõem de recursos limitados de hardware [205]. Esta foi a principal rezão pela qual Mosquitto foi selecionado como fila de mensagens a ser implementado na *stack* HDS. Kafka foi outra das soluções equacionadas, devido ao facto de ser uma solução distribuída e tolerante a falhas. No entanto, no caso da *stack* HDS, essas características são por si só garantidas pela arquitetura que a sustenta, i.e. *Docker Containers*. A imagem oficial do Mosquitto está disponível no repositório Docker Hub [206].

5.2.3 Armazenamento

Nesta subsecção são descritos os componentes selecionados e desenvolvidos para a execuções de todos os serviços relacionados com a persistência e partilha de dados da *stack* HDS.

influxdb

influxdb é uma base de dados projetada e desenvolvida pela empresa influxData, para o armazenamento de séries temporais (i.e., TSDB conforme descrito em 2.4.1). É uma solução *open source*, com exceção do seu componente para a implementação de clusters [3]. É baseada na linguagem SQL, no entanto, em novembro de 2018, a empresa influxData anunciou uma nova linguagem funcional para consulta e processamento de dados das séries temporais (i.e., linguagem Flux) [207].

influxdb é caracterizada pela sua alta performance e otimização, conseguida pala indexação de *Tags* e não dos *Fields* como é habitual em outras bases de dados. Os *Fields* armazenam medições num determinado instante temporal, e devem ser vistos como dados que não precisam de critérios para serem selecionados. Aplicar critérios de seleção a *Fields* resulta em consultas mais lentas uma vez que implica a leitura de todos os valores após o qual serão aplicados os filtros de seleção. Por outro lado, os *Tags* podem ser vistos como "metadados" que caracterizam as medições armazenadas nos *Fields*. Os *tags* são indexados e devem ser usados como critérios de seleção dos dados [208]. Num recente *benchmarking*

[209], o autor compara a performance da infuxdb com outras bases de dados e conclui que esta é 20 vezes mais rápida que SQL Server, 16 vezes mais rápida que Cassandra, 10 vezes mais rápida que Elasticsearch e 4 vezes mais rápida que a base de dados OpenTSDB.

Outra característica bastante interessante da influxdb é o facto de facultar a possibilidade de se configurar e aplicar políticas de retenção dos dados [208]. Por outro lado, a influxdb carece de uma interface mais amigável, i.e., disponibiliza apenas uma interface de linha de comandos (CLI) para a sua gestão.

influxdb foi selecionado como um dos componentes de armazenamento para a *stack* HDS (i.e., para o armazenamento temporário dos dados recolhidos), pela alta performance, pela disponibilidade de políticas de retenção e ainda pela interoperabilidade com o componente telegraf. A imagem oficial do influxdb está disponível no Docker Hub [210].

PostgreSQL & TimescaleDB

PostgreSQL é uma base de dados relacional, *open source*, e disponibiliza as funções que normalmente caracterizam este tipo de armazenamento [14]. No entanto, salienta-se duas das suas funcionalidades, por terem sido determinantes na escolha deste componente. Essas funcionalidades são:

- *Copy*: esta funcionalidade permite a importação e exportação eficiente de dados em massa (i.e., bulk), para e a partir de uma tabela. É muito mais rápido transacionar grandes volumes de dados desta forma quando comparado com os processos de *Insert* e *Select*. Existem três modos distintos para efetuar esta operação, i.e., *binary*, *text* e *raw binary* [211];
- *Listen and Notify*: é um processo através do qual um cliente pode ser notificado de determinados eventos que ocorrem na base de dados[212].

Para além destas características, PostgreSQL disponibiliza uma API [213] para a interoperabilidade com as imagens "agentes", desenvolvidas de raíz com suporte à *framework* .Net, descrito nos subcapítulos 5.2.4 e 5.3. PostgreSQL disponibiliza ainda uma extensão para o tratamento e armazenamento de séries temporais, i.e., TimescaleDB.

O armazenamento de séries temporais implica um rápido aumento no tamanho das bases de dados. O grande problema na performance das bases de dados advém da latência que ocorre na operação de escrita em disco. O armazenamento de séries temporais implica o armazenamento de um grande volume de dados. Manter um grande volume de dados em memória é extremamente dispendioso e muitas vezes impraticável. Por esta razão, a maioria das soluções TSDB se fundamenta na abordagem NoSQL, a fim de garantir a escalabilidade do armazenamento e consequentemente assegurar a sua performance. No entanto, TimescaleDB apoia a sua arquitetura nos princípios que caracterizam as bases de dados relacionais, apesar da dificuldade no escalonamento de algumas das suas funções, e.g. indexação de chave secundárias, *joins* entre tabelas, etc. Pode dizer-se que esta solução oferece o melhor dos dois

mundos, i.e., das bases de dados TSDB e relacionais. Esta nova abordagem fundamenta-se nas diferenças existentes no armazenamento de séries temporais relativamente ao armazenamento de operações mais complexas que exigem garantia de transações, modelações complexas, constante atualização de dados, etc. De facto, o armazenamento de séries temporais implica principalmente *inserts* de dados referenciados no tempo, que correspondem maioritariamente a um intervalo de tempo recente. Para garantir alta performance e resolver o problema de escalabilidade, TimescaleDB baseia-se no conceito de Hypertables, i.e., tabelas que contêm todos os dados referenciados no tempo e na divisão destas em pedaços (i.e., *chunks*) mais pequenos e de tamanho dinâmico. Os *chunks* que correspondem aos intervalos de tempo mais recente mantêm-se em memória, a fim de se garantir uma melhor performance nas operações de escrita e leitura de dados. Por outro lado, o conceito de *chunks* facilita a aplicação de políticas de retenção de dados [214].

A versão atual de TimesacleDB não suporta a sua instalação em *cluster*. A sua implementação em cluster encontra-se ainda em desenvolvimento. No entanto, os testes realizados para avaliar a performance desta abordagem implementada num único nó provaram a sua superioridade quando comparada a outras soluções disponibilizadas de forma distribuída [215].

TimesacaleDB foi selecionada como um componente para armazenamento dos dados históricos, (i.e., dados tratados após recolha e dados das previsões), pela sua alta performance. Para além disso, verificou-se que outras soluções facultam performance com recurso ao escalonamento distribuído do armazenamento. No entanto, a funcionalidade de implementar essas soluções em cluster não está disponível em versões *open source*, e.g., influxdb, Mariadb, etc. Por outro lado, TimescaleDB atinge uma performance superior, apesar de ser executada num único nó. Assim, TimescaleDB é a solução que melhor se adapta ao cenário da *stack* HDS, face aos recursos limitados de hardware em dois dos seus nós, que compõem o cluster.

Por sua vez, PosgreSQL, foi adotado para o armazenamento de regras a serem aplicadas pelos agentes "Pré-Processamento" e "Inspector", como se explica nos subcapítulos seguintes. A imagem oficial do PostgreSQL com a extensão TimescaleBD, está disponível no Docker Hub [216] .

Models

Models é um *container* que armazena um sistema de ficheiros. Este componente *Models* foi desenvolvido de forma a ser possível persistir e partilhar os modelos de previsão de consumo/produção de energia, entre os agentes "Forecasting" e "Inspetor", descritos nas secções seguintes. O seu desenvolvimento consistiu na utilização dos mecanismos nativos do *Docker Container*, disponibilizados para a criação e gestão de volumes [217].

5.2.4 Processamento e Análise

Para a execução das funções de processamento e análise de dados, foram desenvolvidos os componentes "agentes". Existem três tipos de agentes, i.e. *Agentes Pré-preparação*, *Agentes de Previsão* e *Agentes*

Inspetores. Para cada um deles foi desenvolvida uma imagem (i.e., *agente:agente, agente:forecast* e *agente:inspetor*), com recurso à *framework* .Net. As principais funções de cada tipo de agente são descritas a seguir:

- *Agentes Pré-preparação* - estes agentes são responsáveis por: detetar e corrigir anomalias nos registos (i.e., dados obtidos em tempo real das fontes identificadas em 5.2.1) de acordo com regras pré-estabelecidas, reconstruir registos em falha, normalizar e agregar os dados;
- *Agentes de Previsão* (*Forecasting*)– de acordo com o modelo armazenado para a sua zona, estes agentes são responsáveis por: gerar a previsão para a sua zona, de minuto a minuto e para a hora seguinte (parâmetros configuráveis), armazenar e partilhar os resultados das previsões;
- *Agentes Inspetor* - Estes agentes são responsáveis por: inspecionar anomalias no consumo/produção de energia da sua zona de acordo com determinadas regras pré-estabelecidas, lançar alertas, inspecionar os resultados das previsões, caso necessário deve pesquisar, criar e partilhar novo modelo

Cada uma destas imagens permite a configuração e a criação de novos agentes. Para isso basta configurar um ficheiro do tipo *json*, com os parâmetros necessários à execução das suas tarefas. Na criação do *container* que instanciará o serviço para o novo agente deverá ser identificado o parâmetro "hostname", cujo nome deverá ser igual ao ficheiro de configuração (i.e., o ficheiro *json* atrás referido). Desta forma a imagem contida no *containeir* irá carregar automaticamente o ficheiro de configuração a fim de obter todos os seus valores. O exemplo representado na Figura 5.3 mostra a criação de um novo *Agente de Previsão* para a zona99.

Figura 5.3 - Criação e configuração de novos agentes

As imagens *agente:forecast* e *agente:inspetor* integram um componente, i.e. ML.NET, para a análise de dados relacionada com o processo de previsões, que a seguir se descreve.

ML.NET

ML.NET é uma *framework open source* desenvolvida e disponibilizada pela Microsoft, com o objetivo de facilitar a integração de *Machine Learning* em aplicações desenvolvidas em .NET. É compatível com as plataformas de desenvolvimento Visual Studio .Net e NET Core, assim como, com as linguagens de programação C#, F#, VB.Net e Python [218]. Esta *framework* disponibiliza vários algoritmos para a resolução de diversos problemas [219], i.e.:

- problemas de classificação binária: os algoritmos desta categoria visam prever e classificar a qual de duas classes os novos dados pertencem, e.g., sim ou não, 0 ou 1, etc.;
- problemas de classificação multi-classe: os algoritmos desta categoria visam prever e classificar a qual dos conjuntos pertencem os novos dados, e.g., raça de cães, categoria de plantas, etc.;
- problemas de regressão: os algoritmos desta categoria visam prever novos dados que se caracterizam por serem valores contínuos, e.g. previsão de temperatura, previsão de consumo de energia, etc.

A tabela seguinte (Tabela 5.2), exibe os vários algoritmos disponibilizados pala *framework*, dos quais se destacam os algoritmos para a resolução de problemas de regressão, uma vez que foram estes os implementados no *Agente Inspetor*.

Tabela 5.2 - ML.NET: Algoritmos

CLASSIFICAÇÃO BINÁRIA	*CLASSIFICAÇÃO MULTICLASSE*	*REGRESSÃO*
AveragedPerceptron	LightGbmMulticlass	FastTree
SdcaLogisticRegressionBinary	SdcaMaximumEntropyMulticlass	FastTreeTweedie
SdcaNonCalibratedBinary	SdcaNonCalibratedMulticlass	Stochastic Dual Coordinate Ascent (SDCA)
SymbolicSgdLogisticRegressionBinary	LbfgsMaximumEntropyMulticlass	Poisson
LightGbmBinary	NaiveBayesMulticlass	FastForest
FastTreeBinary	OneVersusAll	GeneralizedAdditiveModel
FastForestBinary	PairwiseCoupling	OnlineGradientDescent
GamBinary		
FieldAwareFactorizationMachine		
LinearSvm		

A *framework* ML.NET disponibiliza ainda um conjunto de funções para a normalização dos dados de entrada (i.e., *Binning*, *LogMeanVariance*, *MeanVariance*, *MinMax*, *SupervisedBinning*) [220], para a avaliação dos modelos (i.e., *RSquared*, *RootMeanSquaredError, MeanAbsoluteError*, *MeanSquaredError* e *LossFunction*) [221] e salvaguarda dos modelos gerados. Para além disso, permite a integração de outras bibliotecas para o processamento de algoritmos *deep learning*, como é o caso do

TensorFlow. A interoperabilidade com outras *frameworks* de análise de dados está atualmente em desenvolvimento, como é o caso da *framework* Microsoft Cognitive Toolkit (CNTK).

Por fim, importa referir que o processamento e análise do fluxo de dados da *stack* HDS ficou confinado aos componentes atrás referidos. Era espectável, conforme o modelo proposto para a plataforma HDS (Figura 4.6), a utilização do Apache Beam como principal componente para o processamento do fluxo de dados. No entanto, tornou-se inviável a sua utilização. A *framework* Apache Beam apenas disponibiliza um *Docker Container* para o SDK Python [222]. No entanto, este SDK só disponibiliza conectores para a fila de mensagem *Google Cloud Pus/Sub* e para as bases de dados *Google BigQuery*, *Google Cloud Datastore* e *Google Cloud Bigtable* (i.e., componentes disponíveis apenas no serviço da *cloud* da Google que não é de acesso livre). O SDK Java da *framework* disponibiliza os conectores necessários para a *stack* HDS (i.e., conector para a fila de mensagem MQTT), no entanto, não foi possível criar uma imagem *Docker Container* a partir do código fonte deste SDK [223]. Por outro lado, apesar dos avanços operados nesta *framework,* que permitem a execução de várias transformações e agregações dos dados, não disponibiliza mecanismos para a reconstrução de dados em falta. Assim, optou-se pela implementação de alguns conceitos inerentes à *framework* Apache Beam no *Agente Pré-preparação*, nomeadamente o conceito referente às janelas deslizantes, a fim de contemplar o problema dos fluxos de dados infinitos e fora de ordem. Espera-se, no entanto, o amadurecimento da *framework* Apache Beam a fim de ser possível a sua integração e experimentação no contexto do modelo proposto. As imagens dos agentes estão disponíveis num repositório privado do Docker Hub, conforme mostra Figura 5.4.

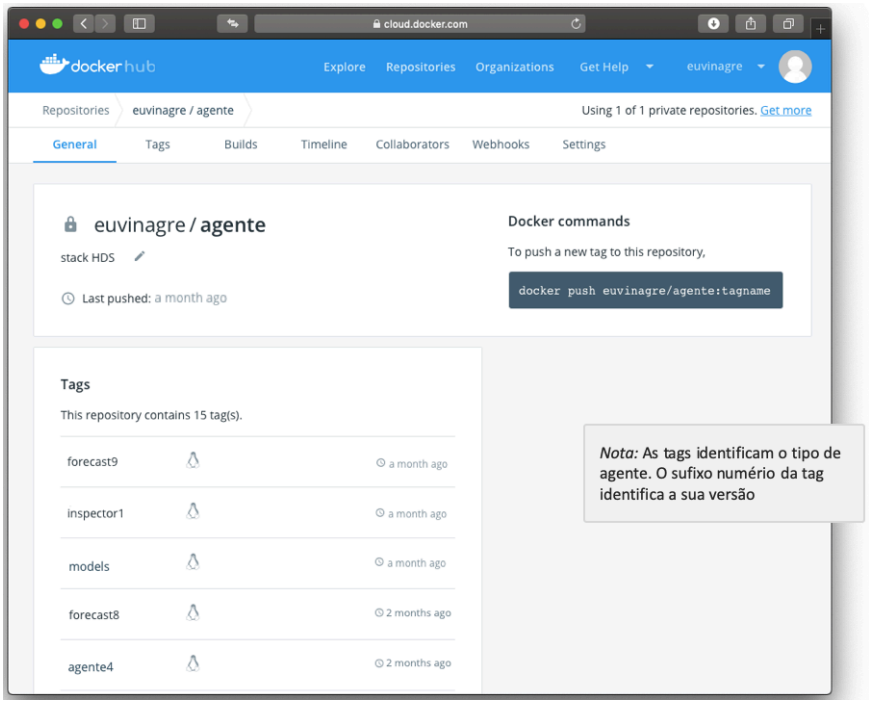


Figura 5.4 - Repositorio Docker Hub: imagens agentes

5.2.5 Visualização

Para a execução dos serviços relacionados com a visualização foi selecionado o componente Grafana. Contudo, Chronograf (i.e., *dashboard* disponibilizado pela empresa influxData) foi a primeira opção pela sua interoperabilidade com os componentes telegraf e influxdb [224]. No entanto, aquando da implementação do primeiro serviço na *stack* HDS (i.e., representação de dados recolhidos em tempo real), o Chronograf demostrou algum atraso no refrescamento dos ecrãs. Na tentativa de se obter uma melhor solução, foram feitas experimentações sobre o comportamento do *dashboard* Grafana. Grafana apresentou um desempenho superior. Para além disso Grafana disponibiliza variadíssimas funcionalidades (vários tipos de gráficos e tabelas, integração com variadíssimas fontes de dados, exportação e importação dos *dashboards* em formato Json e configuração de alertas). É extremamente interativo e disponibiliza um grande número de *plugins*, o que lhe confere um grau elevado de interoperabilidade com outros componentes [225]. A imagem oficial do Grafana está disponível no Docker Hub [226].

5.3 Stack HDS - Serviços

Nesta secção são descritos os vários serviços implementados na *stack* HDS, nomeadamente a visualização e monitorização dos dados recolhidos em tempo real, a preparação dos dados recolhidos em tempo real, a previsões em tempo real, a deteção de anomalias, a avaliação da precisão das previsões e geração de novos modelos para a previsão de consumo/produção de energia.

5.3.1 Visualização e monitorização de dados recolhidos em tempo real

Para a implementação deste serviço foram utilizados os *containers* telegraf, influxdb, Mosquitto e o Grafana, conforme mostra Figura 5.5.

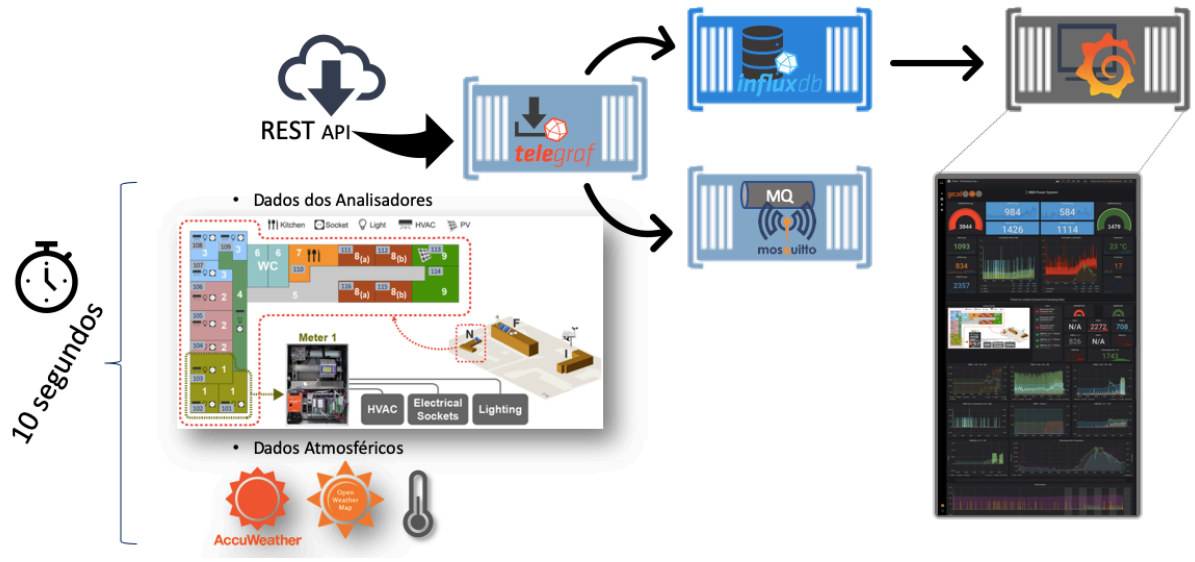


Figura 5.5 - Serviço: Visualização e monitorização em tempo real dos dados recolhidos

O ficheiro *telegraf.conf* do componente telegraf, que pode ser consultado no Anexo 3, foi configurado de forma a identificar as entradas (i.e., identificação dos urls a partir dos quais o telegraf irá executar pedidos GET às APIs REST), e saídas (i.e., para onde o telegraf deverá enviar as métricas coletadas). Como saídas foram definidos o influxdb e o MQTT. Por fim, foram configuradas no Grafana as fontes de dados (i.e., base de dados do influxdb) bem como a configuração de todos os painéis dos *dashboars* de forma a representar graficamente todas as métricas coletadas.

Pela observação contínua dos gráficos configurados no Grafana, foi possível verificar a existência de inúmeros erros nas métricas obtidas a partir dos analisadores referentes às zonas de consumo de energia, i.e., valores negativos e valores excessivamente altos, conforme mostra a Figura 5.6 e a Figura 5.7 respetivamente.

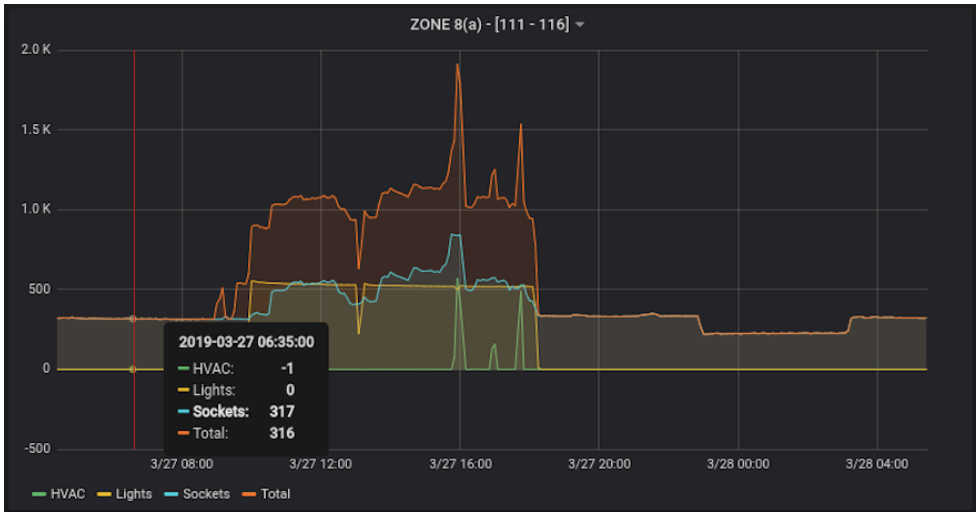


Figura 5.6 - Métricas com valores negativos

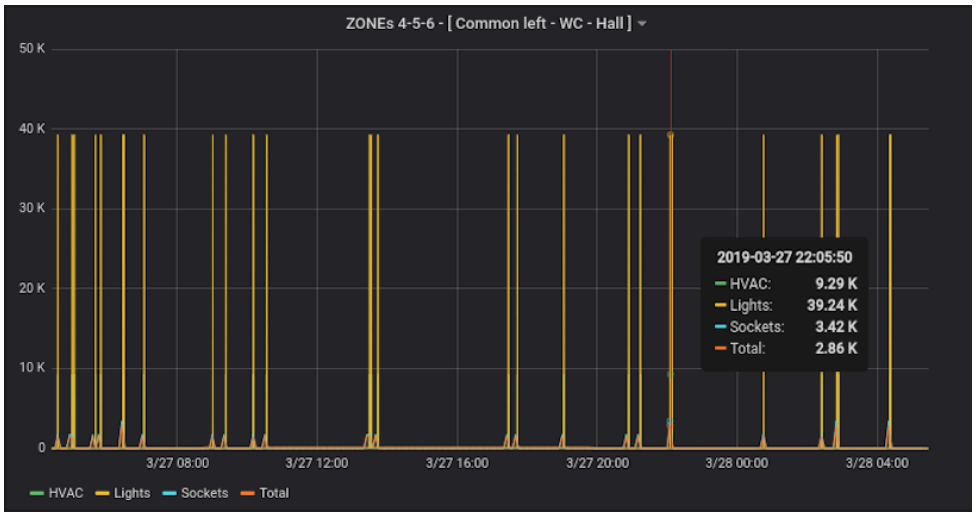


Figura 5.7 - Métricas com valores elevados

Pela análise da Figura 5.7 pode ainda concluir-se que o valor da métrica "Total" não corresponde ao somatório das restantes métricas, (i.e., HVAC, Lights e Sockets).

Após a análise destes registos, foi possível elaborar um conjunto de regras de forma a ser possível filtrar estes erros no serviço de pré-preparação, conforme descrito no subcapítulo seguinte. Estas regras denominadas por *Regras de Primeiro Nível*, estabelecem limites de máximos e mínimos para o consumo e produção de energia em cada uma das zonas. Algumas destas regras foram ainda configuradas como alertas nos *dashboards* do Grafana. Para além disso, estas regras foram armazenadas na base de dados PostgreSQL, sendo possível a sua alteração de forma a se ajustarem a possíveis alterações de contexto.

A representação dos dados foi ainda importante para uma avaliação provisória de possíveis características das séries temporais. Desta análise conclui-se que a maioria das séries são aleatórias, não apresentam tendências, sendo que a sazonalidade fica confinada à zona de produção de energia. Verificou-se ainda uma fraca corelação entre as variáveis das séries temporais.

5.3.2 Pré-preparação de dados em tempo real

Como mostra Figura 5.8, este serviço visa a pré-preparação dos dados coletados em tempo real. Estes dados, conforme referido no subcapítulo anterior, estão a ser enviados pelo telegraf para o *Broker* Mosquitto. O serviço é ainda composto pelos componentes *Agentes Pré-preparação*, um por cada zona, e pelas Bases de Dados PostgreSQL e TimesacleDB.

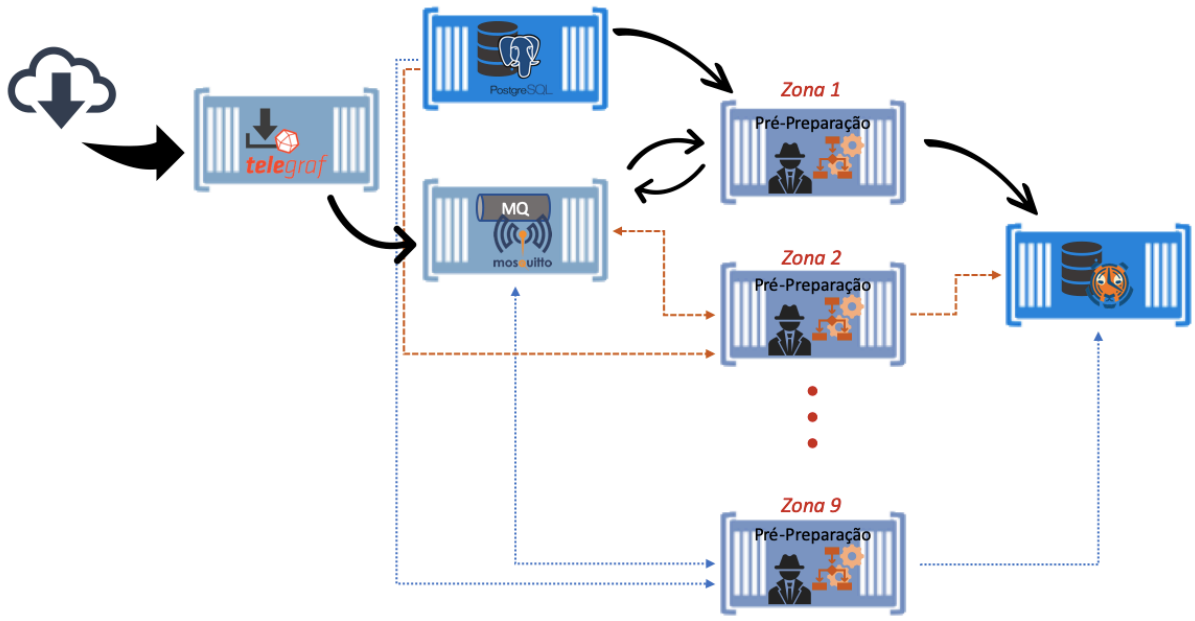


Figura 5.8 - Serviço: Pré-preparação de dados em tempo real

O *Agente Pré-Preparação* conecta-se ao servidor PostgreSQL para obter o conjunto de regras referente à sua zona, (i.e., regras para a validação dos dados). Conecta-se ao Broker Mosquitto para subscrever os tópicos referentes aos dados atmosféricos e às métricas do analisador da sua zona. Sempre que recebe uma nova mensagem procede à sua normalização, validação e agregação. Finalmente, envia os dados agregados para o Mosquitto e para a base de dados TimescaleDB. A seguir descreve-se detalhadamente cada uma das tarefas executadas pelo agente.

Normalização dos registos

As mensagens subscritas pelo agente e referentes aos dados dos analisadores das zonas são compostas por vários campos conforme ilustra a Tabela 5.1 e o Anexo 2. A identificação dos campos das mensagens não está normalizada, por exemplo, o campo P2 numa zona pode significar Potência Ativa na Iluminação e noutra pode significar Potência Ativa no circuito de Tomadas, a Potência Ativa HVAC numa zona pode ser identificada por P1, noutra por P2 e noutra por Ph2_P. Face a este cenário é necessário converter e normalizar as mensagens num formato unificado e entendido por todas as operações da *stack* HDS. Assim, o agente quando recebe uma mensagem procede à sua conversão para um objeto unificado (ver Figura 5.9), de acordo com as regras especificadas para a sua zona.

```
public class MyObjConsumer
{
   public double HVAC { get; set; }
   public double Lights { get; set; }
   public double Sockets { get; set; }
   public double Total { get; set; }
   public int WorkDay { get; set; }
   public int Minute { get; set; }
   public int Hour { get; set; }
   public int Month { get; set; }
   public double temperature { get; set; }
   public long timestamp { get; set; }
}
```

```
public class MyObjGeneration
{
   public double Power { get; set; }
   public int WorkDay { get; set; }
   public int Minute { get; set; }
   public int Hour { get; set; }
   public int Month { get; set; }
   public double temperature { get; set; }
   public long timestamp { get; set; }
}
```

```
public class MyObjBuilding
{
   public double DR_reduce { get; set; }
   public double DR_shift { get; set; }
   public double consumption { get; set; }
   public double generation { get; set; }
   public double hvac { get; set; }
   public double lights { get; set; }
   public double sockets { get; set; }
   public int WorkDay { get; set; }
   public int Minute { get; set; }
   public int Hour { get; set; }
   public int Month { get; set; }
   public double temperature { get; set; }
   public long timestamp { get; set; }
}
```

```
MyObjConsumer: zona1, zona2, zona3, zona4,
               zona7, zona8a, zona8b

MyObjGeneration: zona9

MyObjBuilding: zonabld
```

Figura 5.9 - Objetos para a Normalização da informação

Nesta conversão está ainda contemplada a adição de dados de contexto relacionado com o momento em que o evento ocorreu, i.e., mês, dia, hora, minuto e WorkDay.

WorkDay consite em identificar o dia com 0 ou 1, i.e., se o dia é um feriado e/ou fim de semana, ou se é um dia de trabalho. Para isso foi desenvolvida uma função que calcula a data da Páscoa para um determinado ano (i.e., primeiro domingo após a primeira lua cheia). A partir desta data é possível determinar todos os feriados móveis e adicionar os feriados fixos. A classe IsWorkDay desenvolvida para retornar o valor WorkDay contempla apenas o cenário da cidade do Porto, podendo, no entanto, ser alterada para outros contextos.

Após conversão do registo, este passa para a fase de validação.

Validação dos registos

A validação destes registos consiste em verificar se estes obedecem a determinadas regras pré-estabelecidas, i.e., *Regras Básicas* e *Regras de Primeiro Nível*, conforme referido no subcapítulo 5.3.1. As *Regras Básicas* foram implementadas em todos os agentes responsáveis pelas zonas relacionadas com consumo de energia, e consistem em:

- Se alguma métrica referente à potência ativa for negativa, então eliminar registo;
- Se a soma das potências ativas das várias fases do analisador (i.e., *Hvac*, *Lihgts* e *Sockets*) for diferente do valor da Potência ativa expressa no campo *Total*, então o registo é eliminado.

Por sua vez, conforme já referido, as *Regras de Primeiro Nível* consistem em validar os valores dos campos referentes à potência ativa de acordo com determinados limites. Estas regras estão armazenadas em PostgresSQL e estão sujeitas a possíveis alterações. Assim, o agente quando inicia a sua atividade conecta-se à Base de Dados a fim de obter estas regras, que ficam armazenadas na sua memória. Por outro lado, solicita à Base de Dados para ser notificado sempre que ocorrerem alterações no conjunto das regras referentes à sua zona. Desta forma só voltará a executar a tarefa relacionada com a obtenção das regras quando for notificado. Caso contrário teria de o fazer sempre que validasse um novo registo, i.e., de 10 em 10 segundos. A figura seguinte mostra a aplicabilidade da funcionalidade *Listen and Notify* do PostgresSQL, para a notificação de alterações efetuadas na tabela *Regras de Primeiro Nível* da zona1.

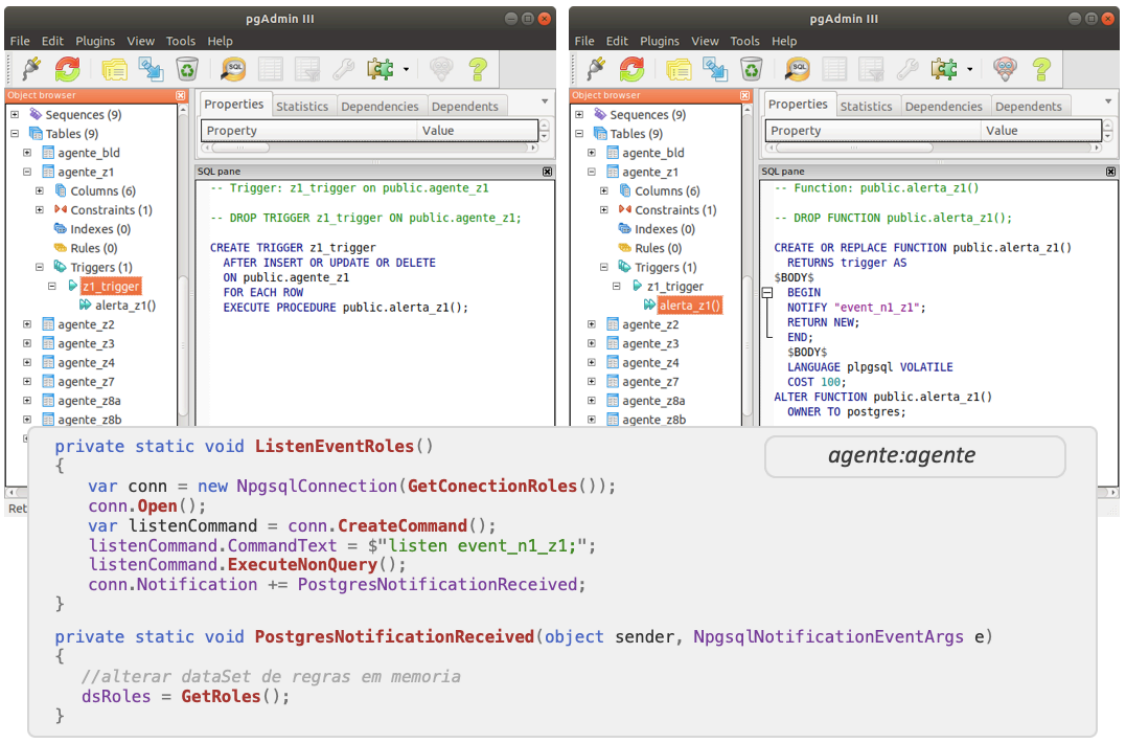


Figura 5.10 - Listen and Notify - Regras de Primeiro Nivel da zona1

Agregação dos registos

Os registos chegam a uma cadência de 10 segundos. A tarefa de agregação consiste em os agrupar em intervalos de um minuto. Para a execução desta tarefa, aplicaram-se alguns dos conceitos do Apache Beam de forma a contemplar a chegada de registos fora de ordem. Para isso foi criada uma *stack*, i.e., *stack agregação*, com três posições. Cada uma das suas posições representa um intervalo de tempo, i.e., minuto passado, minuto corrente e minuto futuro. Assim, o registo após validação é agregado numa destas posições conforme mostra Figura 5.11.

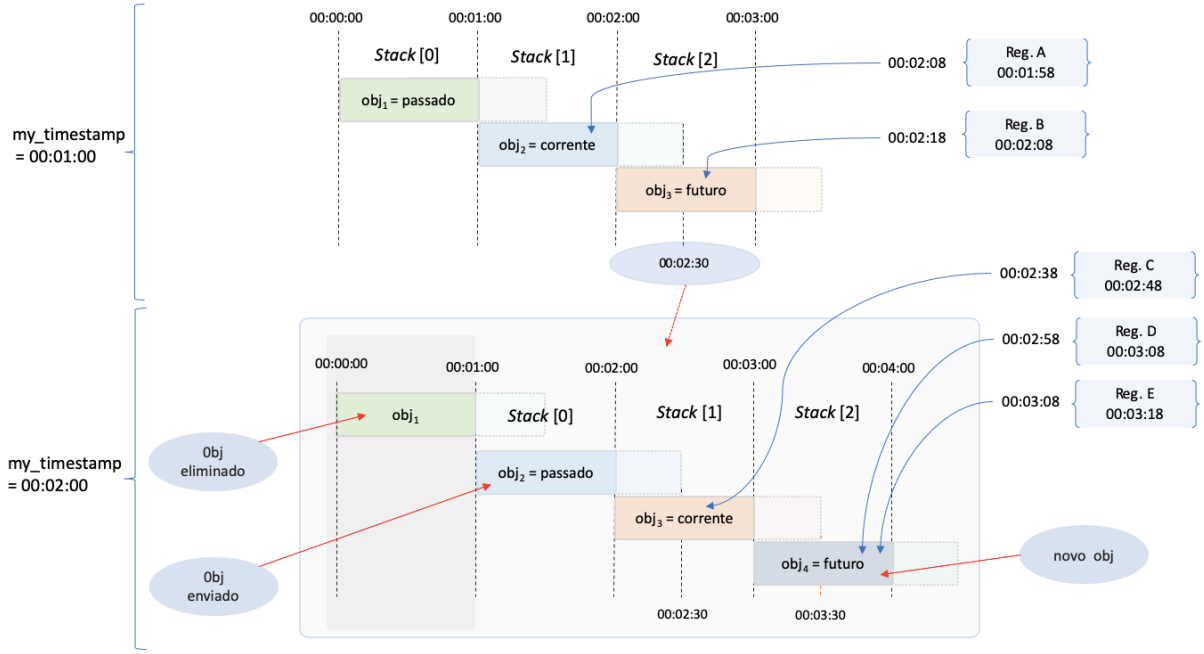


Figura 5.11 - Stack agregação

Conforme se pode observar pela figura, cada intervalo de tempo permanece ativo 30 segundos para além do tempo que representa. Desta forma os registos atrasados podem ser agrupados corretamente. O apontador *my_timestamp* indica o início do objeto corrente e a partir dele é possível determinar a que posição da *stack agregação* devem ser associados os novos registos. Após validação/agregação de um novo registo, o agente verifica a hora corrente e caso tenham decorridos mais de 90 segundos para além do valor indicado pelo apontador *my_timestamp*, o objeto é enviado para o um tópico do Broker Mosquitto, criado para os registos da sua zona. Para além disso, armazena o registo na Base de Dados TimescaleDB. No entanto, antes de proceder ao envio, o agente verifica se o objeto na posição corrente da *stack agregação* não é nulo, (i.e., durante todo o período que esta posição se manteve ativa, todos os registos foram eliminados na fase de validação). Caso o objeto seja nulo o agente procede à recriação do objeto em falha, com suporte aos valores contidos nos objetos armazenados nos extremos da *stack agregação*. Finalmente, o agente procede à atualização das posições da *stack agregação* e do apontador *my_timestamp*, conforme mostrado na Figura 5.11.

As figuras seguintes, apresentam o resultado da tarefa desempenhada pelos *Agentes Inspetores*.

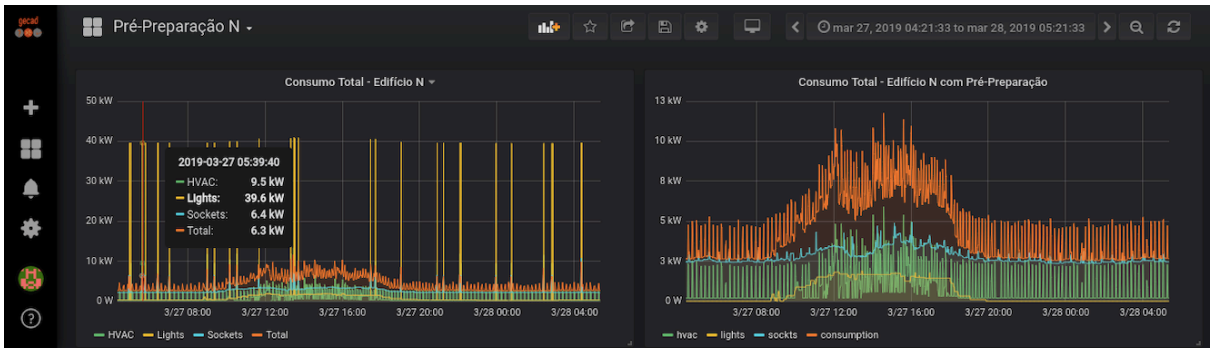


Figura 5.12 - Impacto da Pré-preparação de dados na zonaBld

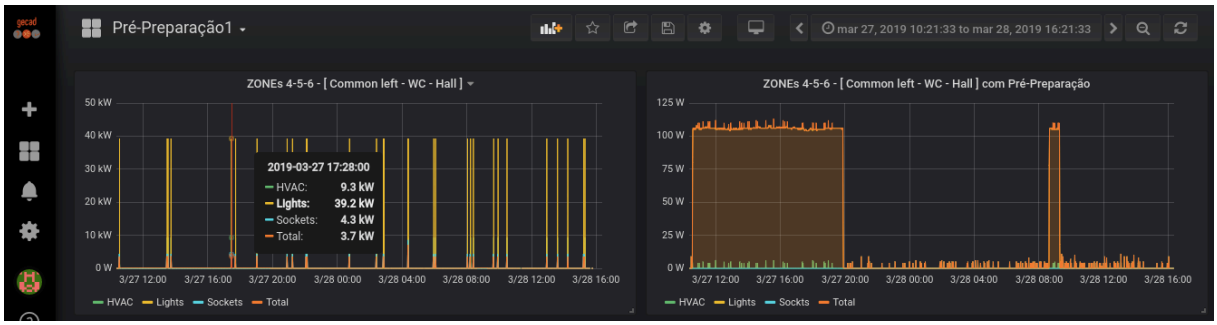


Figura 5.13 - Impacto da Pré-Preparação de dados na zona4

Figura 5.14 – Impacto da Pré-Preparação de dados na zona8a

Como é visível pela observação das imagens, os *Agentes Inspetores* estão a executar eficientemente a tarefa de pré-preparação dos dados.

5.3.3 Previsão em tempo real

Para a implementação deste serviço, conforme mostra Figura 5.15, foram utilizados os componentes *Agentes Forecasting*, um por cada zona, e ainda os componentes Modelos, Mosquitto, TimescalaDB e Grafana.

O *Agente Forecasting* é responsável por executar a previsão de consumo/produção de energia de acordo com as séries temporais da sua zona (i.e., 4 séries no caso das zonas de consumo (Hvac, Lights, Sockets e Total) e uma no caso da zona de produção). Para além disso, é responsável por armazenar os resultados

das previsões na Base de Dados TimescaleDB, assim como, por enviá-los para um tópico do Broker Mosquitto.

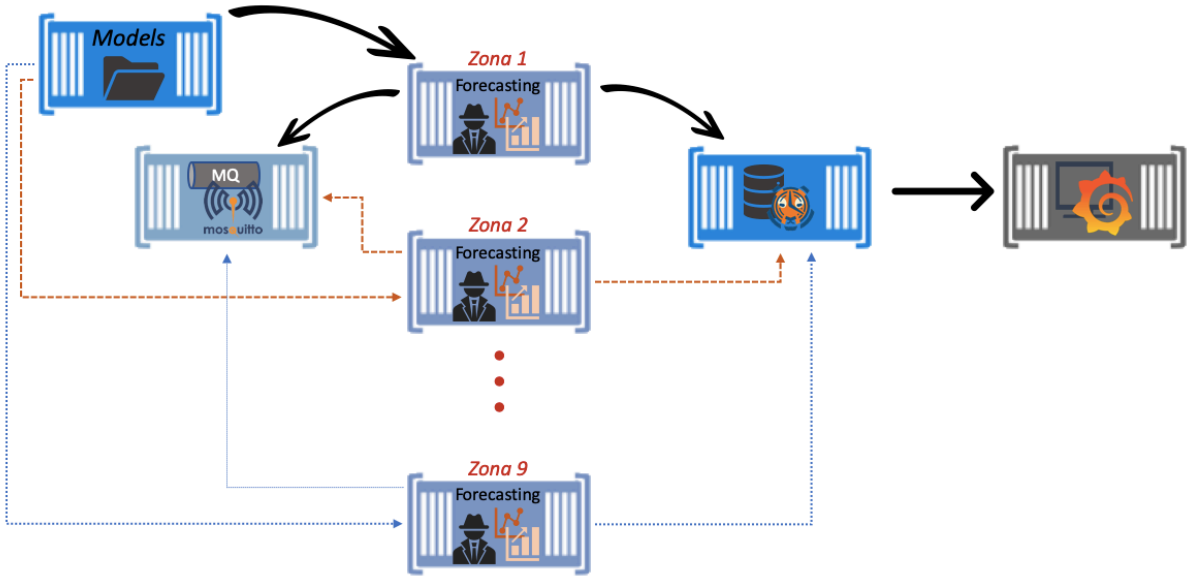


Figura 5.15 - Serviço: Previsões em tempo real

Para executar esta tarefa o agente faz uso do modelo gerado pelo *Agente Inspetor*, modelo esse que se encontra armazenado no contentor *Models* partilhado por todos os agentes. O agente está orientado aos eventos despoletados por um temporizador, de forma executar a previsão no momento exato, i.e., de minuto a minuto para a hora seguinte. Apesar das previsões estarem a ser executadas de minuto a minuto e para a hora seguinte, estes parâmetros são configuráveis.

A definição dos dados de input / output para a execução dos algoritmos de previsão resultaram da análise e dos testes operados pelo *Agente Inspetor*, após elaboração dos primeiros algoritmos, conforme descrito no subcapítulo seguinte (5.3.4). Destes testes resultou a seguinte definição implementada nos *Agentes Forecasting*:

- *Previsão Hvac*: input – timestamp, WorkDay, Month, Hour, Minute, temperature; output – hvac;
- *Previsão Lights*: input – timestamp, WorkDay, Month, Hour, Minute, temperature; output – lights;
- *Previsão Sockets*: input – timestamp, WorkDay, Month, Hour, Minute, temperature; output – sockets;
- *Previsão Total*: input – timestamp, WorkDay, Month, Hour, Minute, temperature; output – total;
- *Previsão Generação*: input – timestamp, Month, Hour, Minute, temperature; output – power.

Por fim, os resultados das previsões foram configurados no *dashboard* Grafana de forma a poderem ser visualizados e analisados.

5.3.4 Deteção de anomalias & automação analítica

Este serviço está a ser coordenado pelos *Agentes Inspetores*. Conforme mostra figura seguinte, para a implementação deste serviço foram utilizados os componentes, PostegreSQL, Mosquitto, TimescaleDB e Models.

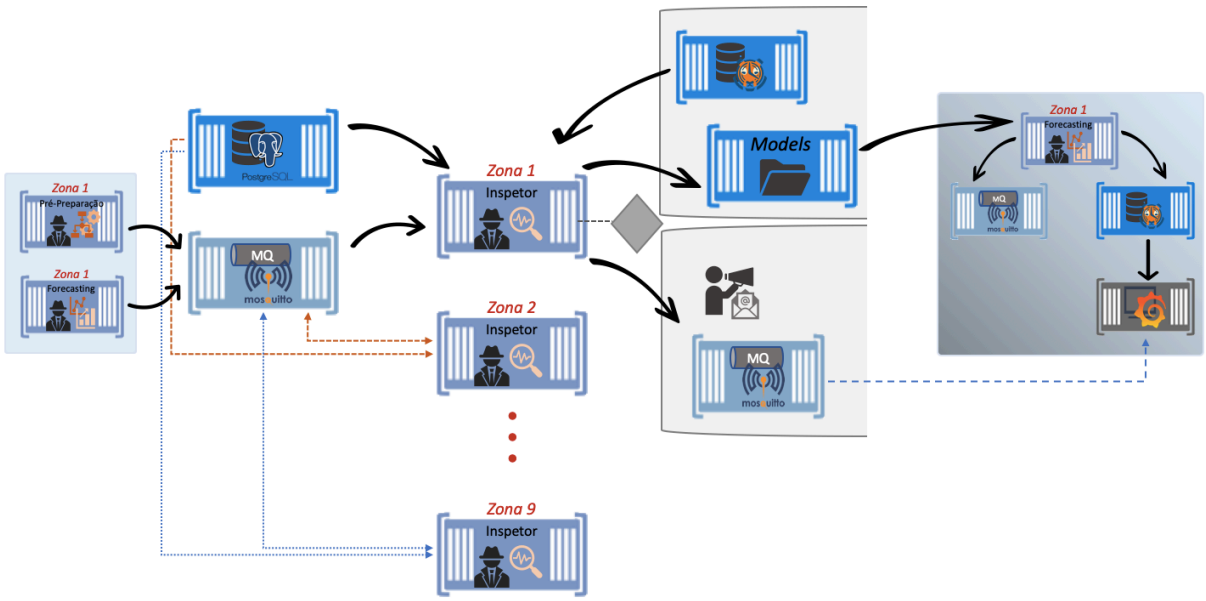


Figura 5.16 - Serviço: Deteção de anomalias & Automação analítica

O *Agente Inspetor* subescreve os tópicos referentes aos registos de pré-preparação e previsão, referentes à sua zona, no Broker Mosquitto. Estes registos, conforme já referido, dizem respeito aos registos enviados para o Broker pelos *Agentes Pré-Preparação* e *Agentes Previsão*.

Este agente é responsável por validar se os valores dos consumos/produção que constam nos registos pré-preparados estão de acordo com o expectável, i.e., se estão de acordo com um conjunto de regras, denominadas *Regras de Segundo Nível*, a fim de detetar possíveis anomalias, e.g., um aparelho de ar-condicionado ou uma luz que por descuido ficou ligada durante a noite, o ar-condicionado da sala de servidores está desligado por uma possível avaria, etc. Em caso de deteção de uma anomalia o agente envia uma notificação por e-mail, e envia essa mesma notificação para um tópico do *Broker*, de forma a poder ser representada no *dashboard* Grafana.

Para além disso, é ainda responsável por verificar os resultados das previsões e em caso de estas não estarem de acordo com o espectável, irá procurar um modelo que melhor se ajuste ao contexto atual. Para isso irá treinar vários algoritmos, tendo como base datasets com gamas diferentes, solicitados à Base de Dados TimescaleDB. A seguir descreve-se com maior detalhe a aplicação das *Regras de*

Segundo Nível e os critérios para a geração de novos modelos de previsão, bem como, a análise dos primeiros resultados obtidos com a implementação dos algoritmos da biblioteca ML.NET.

Regras de Segundo Nível

Para a elaboração das Regras de segundo nível, bem como para a execução e testes dos algoritmos da biblioteca ML.NET, foi necessário converter e migrar o histórico armazenado em SQL Server. Foi desenvolvido um programa a fim de facilitar a execução desta tarefa. Este programa leu os registos armazenados em SQLServer, aplicou sobre eles as *Regras Básicas*, agregou e recriou os registos em falha. Após esta fase de conversão, os dados foram enviados em bloco e armazenados na base de dados TimescaleDB, com recurso à função *Copy*.

Após a migração do histórico foi possível elaborar o conjunto de regras de segundo nível para cada zona. No caso das zonas de consumo, as regras consistiram em determinar o valor máximo, valor mínimo e valor médio da potência ativa, por tipo de dia (i.e., workday), por mês e por hora, para cada tipo de série temporal (i.e., hvac, lights, sockets e total). No caso da zona de produção de energia, os valores determinados foram os mesmos com a exceção da sua subdivisão por tipo de dia, visto que a produção não é influenciada por esta variável.

Tal com acontece com as *Regras de Primeiro nível*, estas regras são passiveis de serem alteradas. Assim, aplicou-se o mesmo processo de escuta e notificação entre a base de dados e os *Agentes Pré-Preparação*, conforme exemplificado em 5.3.2 para o caso das *Regras de Primeiro nível*.

Implementação e testes da biblioteca ML.NET

Antes da implementação da biblioteca ML.NET, os dados das séries temporais foram configurados e representados no *dashboard* Grafana a fim de detetar tendências, sazonalidades, aleatoriedade e correlação entre a suas variáveis. Foram ainda, elaborados testes com suporte à Framework R, de forma a auxiliar a observação feita. Após esta análise, conclui-se que a aleatoriedade é a característica mais comum a todas as séries temporais, da qual se exclui a série referente à produção de energia. Verificou-se novamente a fraca corelação entre as variáveis das séries temporais. Face a este cenário, e não sendo possível determinar quais seriam os modelos mais adequados à previsão destas séries temporais, foram implementados todos os algoritmos disponibilizados pela biblioteca ML.NET para a previsão de problemas de regressão.

Para além disso, todos os algoritmos foram treinados com vários intervalos de dados, i.e., um ano de histórico, últimos seis e três meses de histórico e ainda o último mês de histórico. Foram igualmente treinados com intervalos específicos de dados, i.e., conjuntos de dados filtrados por dia útil, por hora, por dia útil e hora, e por mês. A seleção e conjugação de variáveis de entrada foi outro aspeto testado, i.e., timestamp, workday, mês, hora, minuto e temperatura; timestamp, workday, mês, hora e minuto; timestamp e temperatura; timestamp, etc. Por fim, foi ainda testada a influência dos algoritmos

disponibilizados para a normalização dos dados de treino, i.e., *Binning*, *LogMeanVariance*, *MeanVariance*, *MinMax* e *SupervisedBinning*.

Após a análise dos resultados, verificou-se que os algoritmos frequentemente candidatos a melhor modelo foram o *FastTree* e o *FastTreeTweedie*, i.e., 25 e 8 respetivamente, para as 33 séries temporais em tratamento. Para a avaliação dos modelos gerados, foram utilizadas as funções disponibilizadas pela própria biblioteca ML.NET (i.e., *RSquared*, *RootMeanSquaredError, MeanAbsoluteError*, *MeanSquaredError* e *LossFunction*). O tempo de execução para a geração e avaliação do modelo é em média 10 segundos para cada um dos algoritmos, com exceção feita ao algoritmo *Poisson* (com um minuto de tempo médio), e ao algoritmo *GeneralizedAdditiveModel*. Neste último caso o tempo de execução é bastante variável podendo atingir valores superiores a seis minutos. Devido ao seu tempo de execução excessivamente elevado em comparação com os restantes algoritmos, este foi testado em várias máquinas com recursos diferentes de hardware, onde se concluiu que quanto mais memória e processador a máquina detém, maior é o tempo de execução do algoritmo. Relativamente ao intervalo de dados para treino do modelo, verificou-se uma clara tendência para os dados referentes ao último mês, i.e., 20 vezes candidato como melhor intervalo de dados. No entanto, o intervalo com dados dos últimos dois meses foi candidato 9 vezes, o conjunto de dados referentes aos últimos três meses foi candidato 3 vezes, enquanto que os dados referentes a todo o histórico apenas foi candidato uma vez. Por outro lado, os restantes intervalos de dados filtrados por dia útil, hora, etc., nunca obtiveram resultados satisfatórios. Da análise feita às funções de normalização, conclui-se que as funções *Binning* e *MinMax* são as que permitem a geração de melhores modelos. No entanto a função *MinMax* degrada consideravelmente a performance na excussão do treino e geração dos modelos, podendo atingir um valor total superior a 20 minutos (i.e., tempo total para a execução de todos os algoritmos). Assim, optou-se por utilizar a função *Binning* para a normalização dos dados. Por fim, da análise feita relativamente à seleção e conjugação de variáveis de entrada, conclui-se que os melhores resultados são obtidos quando se utiliza todas as variáveis disponíveis (i.e., as variáveis de contexto, dia útil, mês, hora, minuto e temperatura). Uma análise mais detalhada das experiências e resultados obtidos, pode ser consultada no Anexo 4.

Para além da implementação e experimentação dos algoritmos atrás mencionados, foi testada a integração da biblioteca TensorFlow. No entanto este teste resultou em insucesso porque as bibliotecas do TensorFlow integradas com a ML.NET dependem da disponibilidade de uma *unidade de processamento gráfico* (GPU), recurso este, não disponível no hardware da *stack* HDS. Assim, a previsão das séries temporais da *stack*, ficou confinada aos algoritmos nativos da biblioteca ML.NET.

Critérios para a geração de alertas e de novos modelos

O critério para a geração de alertas consiste em validar de 5 em 5 minutos se os registos gerados e pré-preparados em tempo real estão de acordo com o espectável e delineado pelas *Regras de Segundo Nível*.

Se todos os registos neste período de tempo estiverem em desacordo com o estipulado o agente envia um alerta por e-mail e uma mensagem para o *Broker* Mosquitto. Caso contrário, o agente procede à verificação da eficiência das previsões. Esta verificação é avaliada com o cálculo do *Mean Absolute Percentage Error* (MAPE). Se o erro obtido for superior a 30% é registada a ocorrência. Ao fim de três ocorrências sucessivas o agente toma a decisão de procurar um novo modelo. O agente aplica os critérios atrás referidos a cada uma das séries temporais da zona pela qual é responsável.

A aplicação dos critérios permite estabelecer consistência na deteção de anomalias e no envio de alertas. Por outro lado, a validação da eficiência das previsões com base nos resultados obtidos pela deteção de anomalias, permite evitar que os modelos de previsão se ajustem demasiado a *outlines*. Estes *outlines* apesar de reais são consequência de possíveis anomalias, ou seja, não deviam ocorrer. Face à sua ocorrência é espectável que se proceda imediatamente à sua correção. A aplicação de critérios na geração de novos modelos permite ainda gerir de forma mais adequada os recursos da *stack* HDS, garantindo desta forma a performance de todos os seus processos.

Por fim, a configuração dos *containers* criados para a execução de todos os serviços atrás referidos pode ser consultada e analisado no Anexo 5.

5.4 Conclusão

Neste capítulo foi descrita a implementação da *stack* HDS, tendo como principais objetivos testar o modelo conceptual proposto para a plataforma HDS e ao mesmo tempo contribuir para colmatar alguns dos desafios propostos na área de Big Data e Smart Grids, nomeadamente a pré-preparação de dados e a previsão do consumo/produção de energia em tempo real.

Todos os componentes que integram a *stack* foram descritos e caracterizados, justificando-se o porquê da sua seleção. Foram ainda mencionadas as contramedidas adotadas face ao insucesso obtido na implementação de alguns componentes, dos quais se destacam a plataforma BDE e a *framework* Apache Beam.

Por fim, foram descritos os vários serviços implementados na *stack* HDS, i.e., visualização e monitorização dos dados recolhidos em tempo real; preparação dos dados recolhidos em tempo real; previsões em tempo real; deteção de anomalias; avaliação da precisão das previsões e geração de novos modelos para a previsão de consumo/produção de energia. Alguns pormenores na implementação destes serviços foram destacados por terem contribuído significativamente para a performance da *stack* HDS, conforme o objetivo proposto, tais como a implementação da funcionalidade *Listen and Notify* do PostgresSQL e o escalonamento dos serviços dos agentes por cada uma das zonas, etc. Outros ainda, foram aprofundados pelo seu contributo para a eficácia da *stack*, nomeadamente o tratamento de fluxos contínuos e desordenados, a definição de regras básicas e regras de primeiro e segundo nível, a imposição de critérios para a deteção de anomalias e para a geração de novos modelos de previsão.

Foram ainda descritos e analisados os resultados que se foram obtendo ao longo da implementação da *stack* HDS por forma a garantir a sua eficiência.

Conclui-se que os objetivos propostos foram concretizados. A flexibilidade no modelo proposto para a plataforma HDS permitiu a implementação de contramedidas face ao insucesso obtido na integração de alguns componentes. Por outro lado, a *stack* HDS, apesar dos seus recursos limitados de hardware, executa com eficiência a pré-prepararão dos dados em tempo real. Executa ainda a detecção de anomalias e a previsão de energia para 33 séries temporais em tempo real e de forma autónoma, podendo desta forma contribuir significativamente para a tomada de decisão em tempo real no contexto das Smart Grids.

6 Análise de resultados

A análise de resultados foi um processo constante ao longo de todo o trabalho desenvolvido na presente tese. No entanto, este capítulo visa a exploração e análise de questões relacionadas com a avaliação das previsões executadas pela *stack* HDS.

Pela observação dos resultados obtidos pelas previsões que estão a ser executadas sobre as 33 séries temporais, muitas questões podem ser colocadas, planeadas e analisadas no sentido de os melhorar. As regras implementadas na regeneração dos modelos de previsão são as mais adequadas? Faz sentido executar a previsão para quatro séries temporais nas zonas de consumo (i.e., o somatório das previsões de *hvac*, *lights* e *sockets* supera os resultados obtidos pela previsão do *total*)? Extrapolando o cenário para *edge computer*, que dados devem ser enviados para uma central (dados pré-preparados ou os resultados das previsões)? Os casos de estudo descritos nos subcapítulos seguintes exploram algumas destas questões.

Foram utilizadas várias métricas para a avaliação da precisão das previsões apresentadas nos dos casos de estudo. A *Mean Absolute Percentage Error* (MAPE) é uma métrica de fácil interpretação, pelo facto de ser expressa em termos percentuais. No entanto a MAPE, é muito sensível à escala e a valores perto de zero, podendo o seu resultado ser mal interpretado. Por outro lado, não é possível de ser calculada quando o valor real é zero. Assim, para uma melhor interpretação dos resultados foram adotadas as seguintes métricas:

- *Mean Absolute Percentage Error (MAPE)* [227]: consiste no cálculo do erro médio expresso em percentagem e é definida pela fórmula,

$$MAPE = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{|R_i - P_i|}{|R_i|} \right) * 100 \quad (1)$$

  onde R_i representa o valor real e P_i representa o valor previsto;

- *Mean absolute error (MAE)* [228]: mede a magnitude média dos erros num conjunto de previsões, sem considerar a sua direção, i.e., mede a direção absoluta entre a previsão e a observação real, onde a diferença individual tem o mesmo peso. Quanto menor, melhor será a previsão. É expressa pela seguinte equação,

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |R_i - P_i| \quad (2)$$

  onde R_i representa o valor real e P_i representa o valor previsto;

- *Mean Squared Error (MSE)* [229]: mede o erro quadrado médio da previsão, i.e., para cada ponto calcula a diferença quadrática entre a previsão e o real, seguida do cálculo da média desses

valores. Todas as diferenças são ponderadas de igual forma. Quanto maior esse valor, pior é a previsão. É expressa pela seguinte equação,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(R_i - P_i)^2 \quad (3)$$

onde R_i representa o valor real e P_i representa o valor previsto;

- *Root Mean Squared Error (RMSE)* [230]: mede a magnitude média do erro com base no cálculo da raiz quadrada da diferença entre a previsão e a observação real, i.e., a raiz quadrada da métrica MSE. RMSE atribui maior peso aos erros de maior amplitude, penalizando assim grandes erros. Quanto menor, melhor será a previsão. É expressa pela seguinte fórmula,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - P_i)^2} = \sqrt{MSE} \quad (4)$$

onde R_i representa o valor real e P_i representa o valor previsto;

- *R-squared (R^2)* [231]: *R-squared*, também denominado coeficiente de determinação, mede a variância inexplicada (i.e., variância do erro de previsão) em relação à variância total (i.e., variância dos dados observados). O resultado apresentado por R^2 varia entre zero e um, indicando o grau de precisão do modelo. Quanto maior o seu valor, menor é o erro. Esta métrica é expressa pela seguinte formulação,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(R_i - P_i)^2}{\sum_{i=1}^{N}\left(R_i - \left(\frac{1}{n}\sum_{j=1}^{n} r_j\right)\right)^2} \quad (5)$$

onde N e n representam o número de observações, R_i e r_j valores reais, e P_i o valor previsto.

Os dados reais e previstos relativos às experiências realizadas, nos casos de estudo apresentados a seguir, estão disponíveis em *cloud* no endereço CS_HDS.

6.1 Caso de estudo: Auto-ML

Este caso de estudo visou testar a eficiência do *Agente Inspetor* na sua tomada de decisão relativamente a geração de novos modelos de previsão.

Para se atingir o objetivo definido, foram configurados dois dos novos *Agentes Forecasting*, um responsável pela execução de previsões para as séries temporais da zona1 e outro para a previsão de energia da zona9. Para a execução das previsões foram-lhes atribuídos os modelos ativos da zona1 e zona9, mas sem o suporte de *Agentes Inspetores*. Desta forma estes novos agentes são obrigados a executar as previsões sempre com os mesmos modelos. Os resultados das duas experiências são detalhados a seguir.

Eficiência da automação analítica na zona de produção

Os resultados obtidos para a zona9 podem ser analisados nas figuras seguintes.

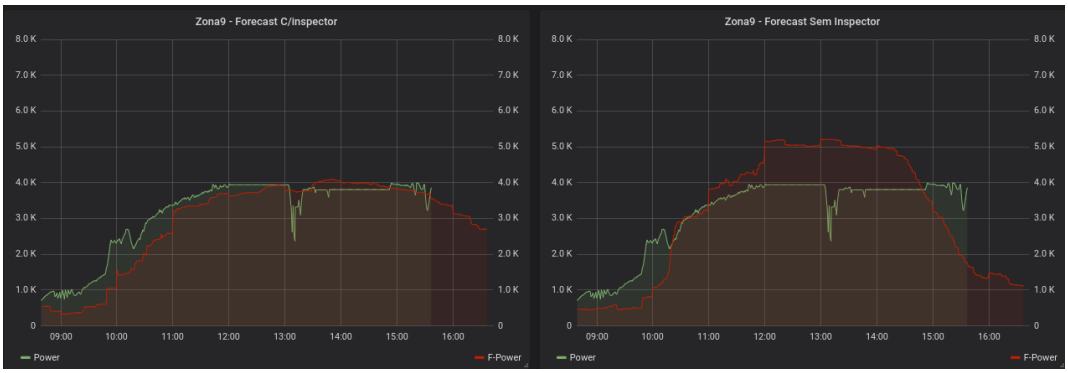


Figura 6.1 – Previsão na zona9 com e sem Agente Inspetor (2019-05-27 08:30 – 15:40)

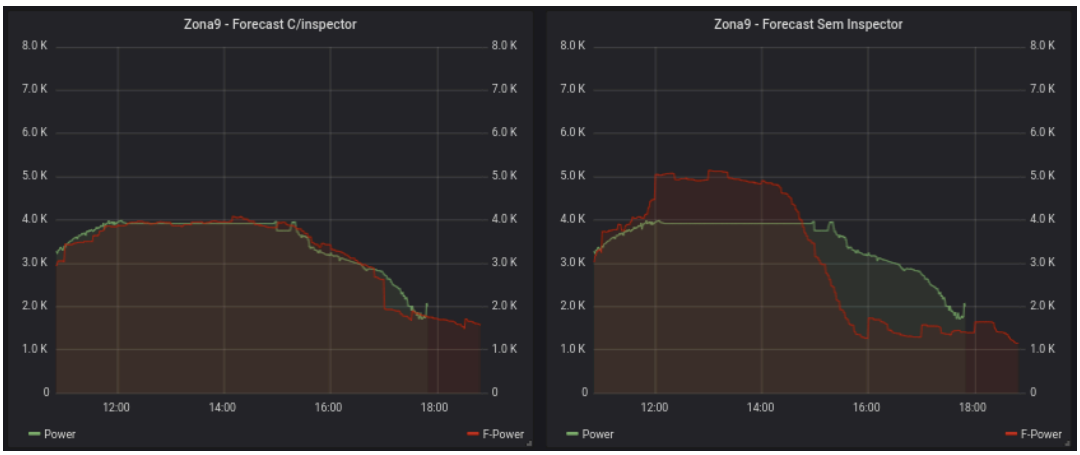


Figura 6.2 – Previsão na zona9 com e sem Agente Inspetor (2019-05-30 11:10 – 17:50)

Para um melhor entendimento dos resultados representados nos gráficos da *Figura 6.1* e *Figura 6.2* foram extraídos da Base de Dados os dados referentes à experimentação realizada, sobre os quais foram calculadas as métricas para a avaliação da precisão das previsões, conforme sumariadas na tabela seguinte.

Tabela 6.1- Resultados das Previsões: zona9 com e sem Agente Inspetor

		R^2	MAPE	MAE	RMSE	MSE
Figura 6.1	*Auto-ML*	0,85	13,13 %	355	471	221820
	Modelo fixo	0,49	32,41 %	1053	1218	1483848
Figura 6.2	*Auto-ML*	0,96	6,11 %	155	228	52093
	Modelo fixo	0,68	31,69 %	996	1089	1186741

Pela análise dos resultados rapidamente se conclui que a aplicabilidade do Agente Inspetor é muito benéfica. Por ouro lado, comprova-se a eficiência da metodologia implementada para a geração de novos modelos, conforme descrito em 5.3.4.

Eficiência da automação analítica na zona de consumo

Os resultados obtidos para as séries temporais da zona1 podem ser analisados nas figuras seguintes.

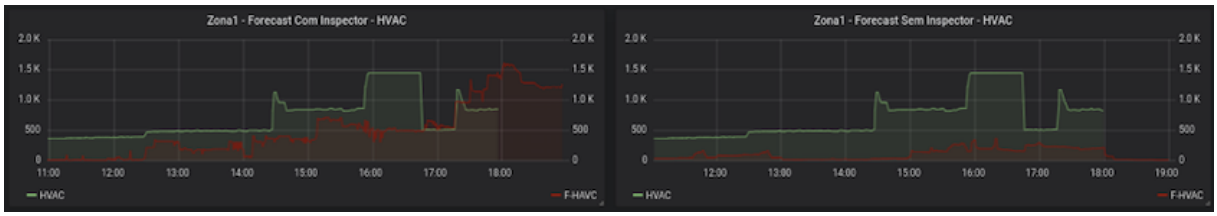


Figura 6.3 – Previsão na zona1 [Hvac] com e sem Agente Inspetor

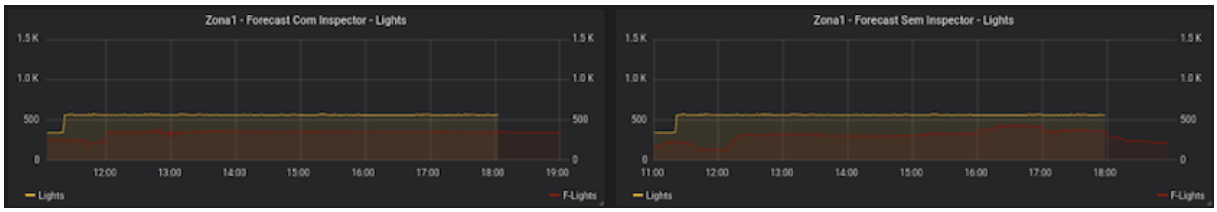


Figura 6.4 – Previsão na zona1 [Lights] com e sem Agente Inspetor

Figura 6.5 – Previsão na zona1 [Sockets] com e sem Agente Inspetor

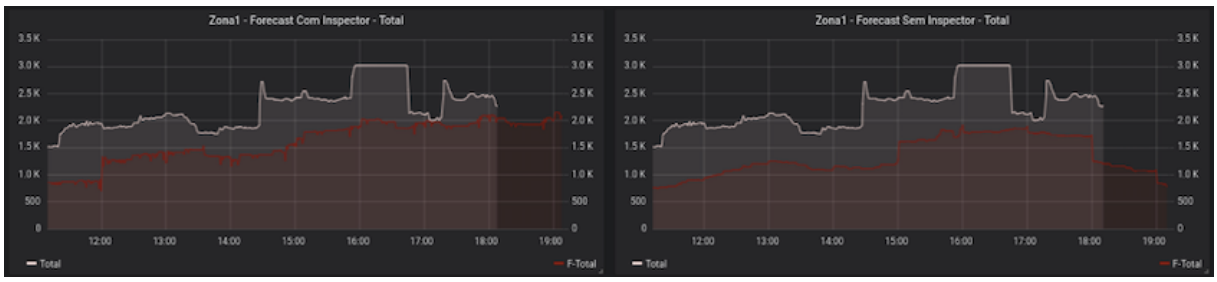


Figura 6.6 – Previsão na zona1 [Total] com e sem Agente Inspetor

Tal como na experiência anterior, foram extraídos da Base de Dados os dados referentes à experimentação realizada (i.e., 30/05/2019 11:00:00 – 18:00:00), sobre os quais foram calculadas as métricas para a avaliação da precisão das previsões, conforme sumariadas na tabela seguinte.

Tabela 6.2 – Resultados das Previsões: Séries temporais da zona1, com e sem Agente Inspetor

Séries Temporais		*R^2*	*MAPE*	*MAE*	*RMSE*	*MSE*
Hvac (Figura 6.3)	*Auto-ML*	0,19	53,82 %	412	490	240562
	Modelo fixo	0,07	85,50 %	652	713	508620

Séries Temporais		R^2	MAPE	MAE	RMSE	MSE
Lights (Figura 6.4)	Auto-ML	0,0019	37,94 %	213	213	45467
	Modelo fixo	0,0013	43,48 %	244	251	63171
Sockets (Figura 6.5)	Auto-ML	0,43	10,10 %	98	122	14965
	Modelo fixo	0,43	10,10 %	98	122	14965
Total (Figura 6.6)	Auto-ML	0,46	25,74 %	597	663	440178
	Modelo fixo	0,39	38,47 %	878	919	844848

Pela análise dos resultados verifica-se que a precisão das previsões nesta área de consumo de energia está longe de atingir o sucesso obtido na zona de produção, com a exceção da série temporal *Sockets* onde se obteve resultados satisfatórios. No entanto, a analítica implementada com recurso apoiado pelo *Agente Inspetor* foi substancialmente superior aos resultados obtidos com um modelo fixo. Verificou-se ainda que ao longo do período experimentado o *Agente Inspetor* procurou novos modelos na tentativa de melhorar os resultados das previsões, ainda que com eles não se tenham obtido resultados satisfatórios. Por outro lado, face aos bons resultados obtidos pelo modelo implementado para a previsão da série temporal *sockets*, o agente decidiu de forma acertada, visto não ter alterado o modelo.

Apesar dos resultados menos satisfatórios obtidos nas previsões *Hvac*, *Lights* e *Total*, verificou-se na experimentação feita no caso de estudo descrito no subcapítulo seguinte uma acentuada melhoria nos resultados obtidos na previsão *Total*, conforme mostra Figura 6.8. Assim, explorou-se o histórico das previsões obtidas na zona1, desde a corrente experiência (i.e., 30/05/2019 11:00:00) até à experiência apresentada no caso de estudo seguinte (i.e., 17/06/2019 16:30), conforme mostra figura seguinte.

Figura 6.7 – Previsão na Zona1 (30/05/2019 11:00 - 17/06/2019 16:30)

Pela análise da Figura 6.7 depreende-se a eficiência do *Agente Inspetor* na aplicabilidade da automação analítica na zona1.

Neste caso de estudo conclui-se que a aplicação da automação analítica foi implementada com sucesso.

6.2 Caso de estudo: Soma das partes

A previsão do consumo de energia individualizado por tipo é bastante interessante em alguns cenários, tais como a gestão de energia em casas inteligentes ou a implementação de programas de resposta à demanda sem pôr em causa atividades críticas. No entanto, há outras situações em que poderá ser interessante a previsão de energia de forma mais global, por exemplo, para uma melhor gestão de recursos ou para apoio à tomada de decisões mais genéricas. Independentemente dos benefícios de cada um dos cenários, este caso de estudo pretendeu responder à questão: o somatório das previsões de *hvac*, *lihgts* e *sockets* supera os resultados obtidos pela previsão do *total?*

Para a execução deste caso de estudo, procedeu-se à configuração do *dashboard* Grafana de forma a representar as previsões do *total* e o somatório das previsões *hvac*, *lihgts* e *sockets*, cujo resultado pode ser observado nas figuras seguintes.

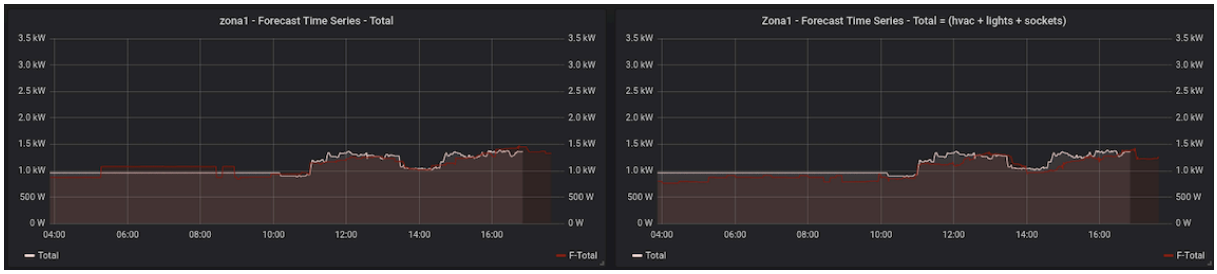


Figura 6.8 – Previsão na zona1: soma das partes

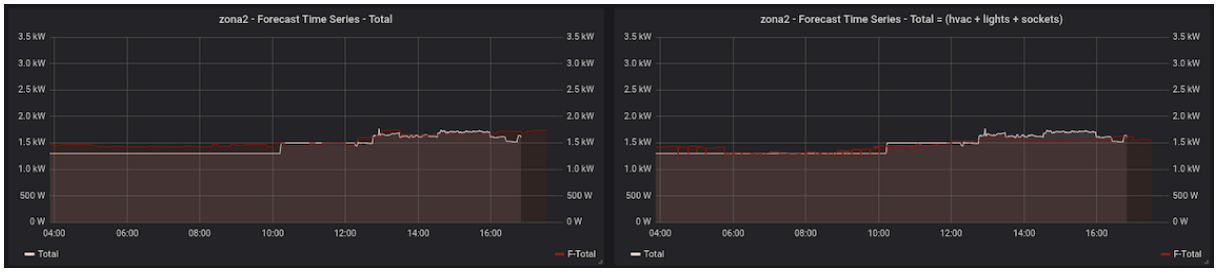


Figura 6.9 – Previsão na zona2: soma das partes

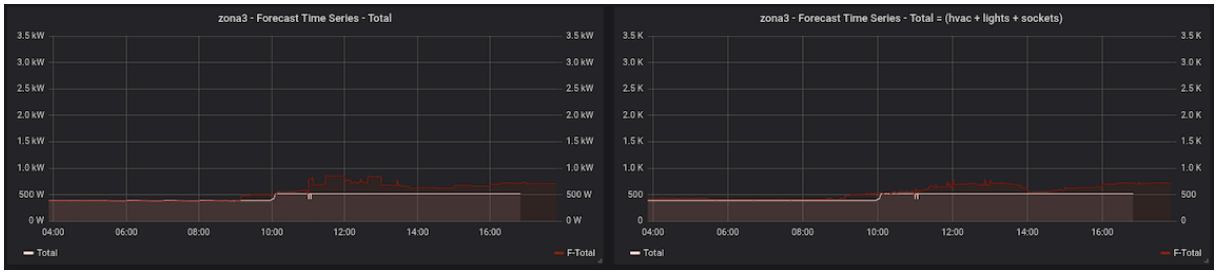


Figura 6.10 – Previsão na zona3: soma das partes

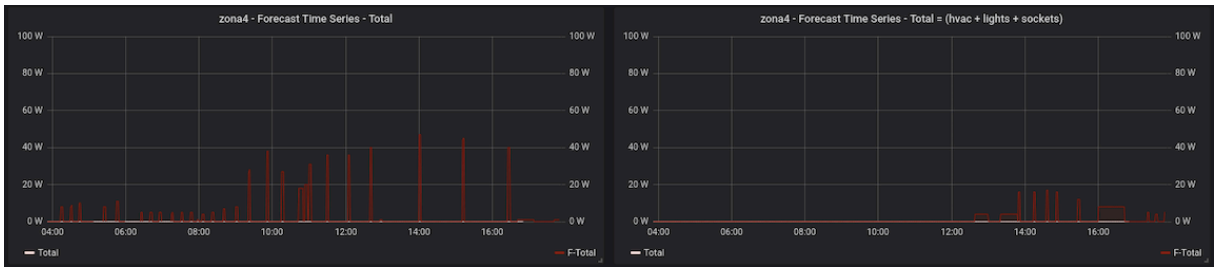


Figura 6.11 – Previsão na zona4: soma das partes

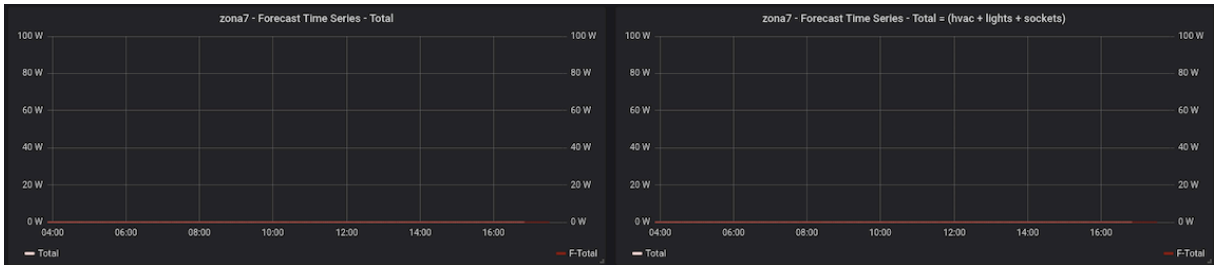


Figura 6.12 – Previsão na zona7: soma das partes

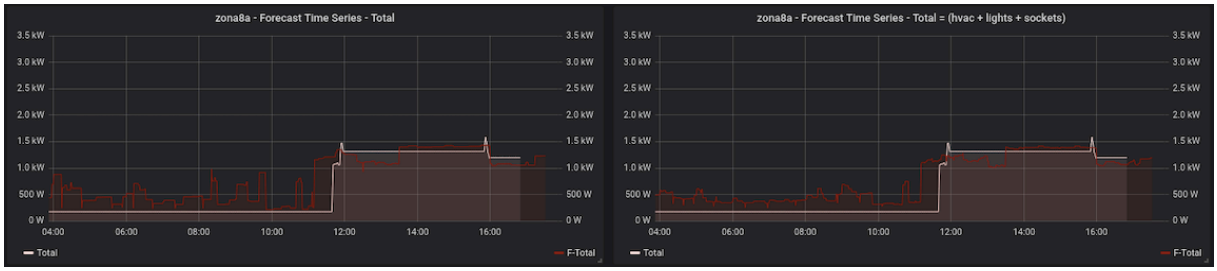


Figura 6.13 – Previsão na zona8(a): soma das partes

Figura 6.14 – Previsão na zona8(b): soma das partes

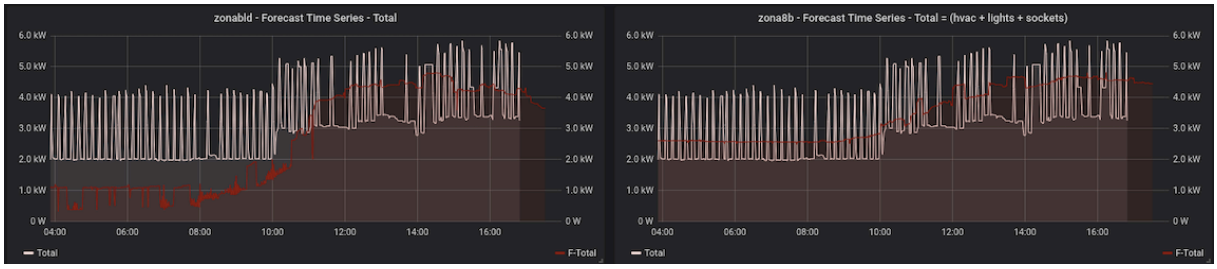


Figura 6.15 – Previsão na zonaBld: soma das partes

Os resultados observados nas Figura 6.8 – Figura 6.15, foram quantificados conforme sumariados na tabela seguinte.

Tabela 6.3 – Resultados das Previsões: Soma das partes p/zona

		R^2	MAPE	MAE	RMSE	MSE
Zona1 (Figura 6.8)	F-Total	0,72	7,44 %	79	8304	91
	F- (H + L+ S)	0,81	8,52 %	93	12405	111
Zona2 (Figura 6.9)	F-Total	0,75	6,71 %	94	109	11919
	F- (H + L+ S)	0,80	4,00 %	60	77	5929
Zona3 (Figura 6.10)	F-Total	0,79	21,78 %	110	148	21857
	F- (H + L+ S)	0,76	18,70 %	92	115	13307
Zona4 (Figura 6.11)	F-Total	—	238,93 %	2	8,35	70
	F- (H + L+ S)	—	106,93 %	1	3,14	10
Zona7 (Figura 6.12)	F-Total	—	—	0	0	0
	F- (H + L+ S)	—	—	0	0	0
Zona8a (Figura 6.13)	F-Total	0,76	97,66	228	313	98119
	F- (H + L+ S)	0,82	93,07	215	283	80186
Zona8b (Figura 6.14)	F-Total	—	44,73	214	230	52783
	F- (H + L+ S)	—	44,06	211	228	52033
ZonaBld (Figura 6.15)	F-Total	0,29	38,89	1211	1436	2061500
	F- (H + L+ S)	0,29	28.09	897	997	994261

Pela análise da tabela verifica-se que os melhores resultados foram obtidos pela soma das partes, exceção feita à zona1. Desta forma, e caso seja necessário replanear a otimização dos recursos de hardware na *stack* HDS, a execução da previsão para as séries temporais *Total*, deve ser tomada em conta.

Verificou-se ainda a obtenção de resultados pouco satisfatórios na zonaBld. Os resultados menos conseguidos para esta zona são de certa forma justificados por uma menor eficiência na pré-preparação dos dados. As regras básicas, conforme referidas no capítulo 5.3.2, permitem detetar e filtrar os registos errados obtidos na fase de recolha de dados. No entanto, a regra referente aos registos que apresentam potência ativa negativa é impossível de ser aplicada nesta zona, visto que os registos desta zona representam o somatório dos valores obtidos em todas as outras zonas. Desta forma, os valores negativos ficam completamente camuflados sendo impossível a sua deteção e consequente tratamento. Por outro lado, a aplicação das regras de primeiro nível nesta zona é feita num âmbito mais amplo (i.e., não podem ser tão refinadas e ajustadas ao contexto como acontece nas restantes zonas).

Neste caso de estudo conclui-se que quanto mais especifica for a implementação das previsões, maior será a probabilidade de se obter bons resultados. Por outro lado, conclui-se que a execução de uma boa pré-preparação dos dados é determinante para o sucesso das previsões.

6.3 Caso de estudo: Edge Computing & Edge Analytic

Este caso de estudo tem como principal objetivo fazer prova das abordagens *Edge Computing* e *Edge Analytic*. Para a experimentação deste caso de estudo supõe-se que cada zona representa um ponto de computação e valida-se qual das hipóteses é mais válida, i.e., pré-preparar os dados recolhidos em cada ponto e enviá-los para um ponto central onde será executada a previsão como um todo, ou executar a previsão em cada ponto e enviar os resultados obtidos para o ponto central. No entanto, este caso de estudo não contempla a latência da rede no transporte de dados, visto que o cenário exploratório está implementado em três máquinas físicas conectadas no mesmo domínio. Ou seja, o caso de estudo tem como foco a avaliação da precisão das previsões.

Para a sua implementação, os dados pré-preparados de cada zona foram agrupados por tipo de série temporal (i.e., hvac, lights, sockets e total). Foi gerado um modelo preditivo para cada série temporal de acordo com os mesmos critérios aplicados em cada zona. Finalmente, a previsão foi executada no período compreendido entre as 04:00 e as 17:30 do dia 17-06-2019. Os resultados comparativos entre a previsão executada no ponto central e a previsão executada nos restantes pontos pode ser observado nas figuras seguintes.

Figura 6.16 – Simulação Edge Analytic

A Figura 6.16 representa e compara os quatro cenários explorados, i.e.:

(1) Previsão da série temporal *Total* executada no ponto central;

(2) Somatório das previsões das séries temporais *Hvac*, *Lights* e *Sockets* executadas no ponto central;

(3) Somatório das previsões das séries temporais *Total*, de todas as zonas;

(4) Somatório das previsões das séries temporais *Hvac*, *Lights* e *Sockets*, de todas as zonas.

Os resultados obtidos nos quatro cenários experimentados encontram-se sumariados na tabela seguinte.

Tabela 6.4 - Resultados da simulação: Edge Analytic

		R^2	MAPE	MAE	RMSE	MSE
Central	(1) Total	0,81	8,89 %	383	462	213188
	(2) Σ(H+L+S)	0,84	8,45 %	104	135	18107
Zonas	(3) Σ Total	0,93	8,75 %	515	454	206215
	(4) Σ(H+L+S)	0,94	5,10 %	397	459	210609

Pela análise dos resultados, conclui-se que o melhor cenário é a execução da previsão por zona e por tipo. Neste cenário a previsão do consumo total do Edifício N resultou num erro de apenas 5,1%.

Finalmente, foi verificada a eficiência da previsão por tipo de consumo, i.e., *Hvac*, *Lights* e *Sockets*, conforme mostra figura seguinte.

Figura 6.17 - Previsão p/tipo: Central vs Zonas

A tabela seguinte sumariza os resultados obtidos.

Tabela 6.5 - Resultados das Previsões p/tipo: Central vs Zonas

		R^2	MAPE	MAE	RMSE	MSE
Hvac	Central	0,46	25,94 %	88	118	14007
	Zonas	0,66	19,2 %	71	84	7013
Lights	Central	0,81	26,90 %	224	316	99619
	Zonas	0,93	19,50 %	206	172	29602
Sockets	Central	0,89	13,85 %	229	252	63661
	Zonas	0,87	6,42 %	236	229	52531

Pela análise dos resultados obtidos depreende-se mais uma vez que quanto mais especifica for a previsão, i.e., quanto mais esta se puder ajustar ao cenário que se pretende prever, melhor serão os resultados obtidos.

Este caso de estudo pretendeu avaliar de que forma se pode beneficiar da abordagem *Edge Analytic*. Conclui-se que, para além de todas as vantagens inerentes à computação distribuída que sustenta o conceito, *Edge Anatytic* acrescenta vantagens significativas no que diz respeito à precisão das previsões.

6.4 Conclusão

Este capítulo visou a análise de resultados na perspectiva da eficiência da previsão em execução nas várias zonas configuradas na *stack* HDS. Foram desenvolvidos vários casos de estudo a fim de validar os critérios definidos e configurados na *stack* para a execução das previsões de consumo e produção de energia. Foram ainda explorados e experimentados vários cenários com o objetivo de se coletarem os melhores resultados obtidos nas previsões.

Os resultados obtidos permitiram concluir que os critérios configurados para a geração de novos modelos preditivos estão de acordo com o espectável, tendo contribuído para o sucesso da automação analítica implementada na *stack* HDS. Por outro lado, conclui-se que quanto mais específica e particularizada for a configuração de regras na pré-preparação de dados, melhores serão os resultados obtidos na precisão das previsões. De igual modo, quanto mais específica for a implementação das previsões (i.e., por zona e por tipo), maior será a probabilidade de se obter bons resultados. Por fim, foi analisado o impacto da aplicação da abordagem *Edge Analytic* na precisão das previsões. Conclui-se que esta abordagem é não só viável, como contribui significativamente e positivamente para a obtenção de melhores resultados na precisão das previsões.

7 Conclusões

Este capítulo aborda, ainda que de forma resumida, o trabalho efetuado ao longo desta dissertação. Apresenta os objetivos propostos e concretizados e os principais contributos desta tese, conforme já referido no capítulo introdutório, ainda que de uma forma mais pormenorizada. Prossegue com a descrição das principais delimitações enfrentadas no desenvolvimento do presente trabalho e com a enumeração de algumas sugestões para a realização de trabalhos futuros, no sentido de aperfeiçoar e melhorar o trabalho já realizado, de forma a poder culminar num maior contributo para o desenvolvimento quer da área de Big Data quer para a área das Smart Grids. Por fim é feita uma apreciação de todo o trabalho realizado nesta dissertação.

7.1 Síntese da dissertação

Esta dissertação iniciou-se com um estudo aprofundado dos principais conceitos, desafios e avanços operados na área de Big Data. Identificaram-se as principais abordagens que têm vindo a ser propostas nesta área, bem como a sua evolução e aplicabilidade. Analisou-se a disponibilidade de componentes para a conceção de plataformas Big Data, onde se realçou a sua maturidade, funcionalidades e performance. Relativamente às plataformas Big Data conclui-se que estas são na sua maioria soluções comerciais, disponibilizadas em nuvem como um serviço. Para além disso, são focadas na resolução de problemas específicos, como, por exemplo, análise de dados, recolha e integração de dados. Por fim, conclui-se que a área de Big Data está repleta de desafios e um logo caminho está ainda por percorrer. Dos desafios identificados realçou-se a complexidade inerente ao processamento em tempo real de fluxos de dados contínuos e desordenados. Na área de análise de dados destacou-se o problema relacionado com o tempo despendido na pré-preparação de dados, a necessidade de desenvolver soluções que permitam a automatização analítica, soluções adaptadas ao nível de conhecimento de cada utilizador e soluções direcionadas para casos específicos de uso, tais como, deteção de anomalias (*Anomaly Detection*)*, IoT Edge Analytics,* e preparação de dados (*Data Preparation*). Ainda de salientar que, face ao crescimento exponencial de soluções propostas nesta área, houve a necessidade de atualizar este estudo durante todo o período em que decorreu o desenvolvimento da presente dissertação.

Prosseguiu-se com um estudo centrado nas Smart Grids a fim de identificar os seus principais desafios. Deste estudo foi possível concluir-se que o fluxo de dados do ecossistema energético representa um dos seus principais ativos, cuja gestão é de extrema complexidade. Por outro lado, identificou-se a grande necessidade de solucionar os problemas direcionados para o processamento em tempo real e para a melhoria da precisão das previsões de consumo e produção de energia. A fim de avaliar o possível impacto da aplicabilidade das abordagens propostas na área de Big Data no desenvolvimento das Smart Grids, revisou-se a sua importância em três vertentes distintas, i.e., desenvolvimentos realizados na comunidade científica, impacto sectorial de Big Data na área de Smart Grids, posição da UE

relativamente ao investimento na área de Big Data no contexto das Smart Grids. O estudo realizado permitiu concluir que a maturidade da aplicação de tecnologias Big Data no ecossistema energético está ainda num estágio inicial. No entanto, é unanime a posição de todos (i.e., comunidade científica e setorial), de que é imprescindível incorporar nas Smart Grids tecnologias de Big Data para a gestão e análise do seu complexo fluxo de dados. Só desta forma será possível extrair o valor essencial para a tomada de decisão em tempo real, a fim de garantir a eficiência e sustentabilidade do ecossistema energético.

Face aos estudos atrás referidos, foram propostas e exploradas várias soluções no sentido de darem uma resposta positiva aos problemas identificados como desafios nas áreas de Big Data e Smart Grids. A proposta inicial teve como principal foco a procura de uma solução que contribuísse para a minimização do problema de processamento em tempo real de grandes volumes de dados inerentes ao ecossistema energético. Nesse contexto, e extraindo o melhor que oferece a *framework Apache Spark* relativamente ao desempenho, foi proposta uma solução baseada na *SMACK Stack*, à qual se propôs a adição de componentes para agilizar os processos relacionados com a análise e visualização dos dados. No entanto, verificou-se que muitos avanços têm sido feitos na área de Big Data e que a arquitetura anteriormente proposta apresentava pouca flexibilidade para a experimentação dessas novas abordagens. Por outro lado, verificando-se um crescimento avultado na complexidade que envolve a gestão de dados no ecossistema energético, foi revisada a arquitetura inicial culminando numa nova solução com o objetivo de facilitar a implementação e experimentação das novas abordagens que têm sido feitas na área de Big Data. Pretendia-se uma solução mais flexível e ágil. Ao se identificar que a arquitetura *Docker Container* possui as características ideais para a concretização desse objetivo, a nova solução foi desenvolvida tendo como base esta arquitetura. O modelo conceptual proposto na nova solução visa a resolução dos desafios relacionados com a tomada de decisão em tempo real, sem, no entanto, deixar caminho aberto para a resolução de outros assuntos relacionados com as Smart Grids (e.g., Governação e segurança dos dados; partilha de dados para a comunidade I&D; processamento próximo do tempo real para a tomada de decisões de médio e longo prazo, etc.).

A fim de validar o modelo proposto procedeu-se ao desenvolvimento de uma *stack* (i.e., a *stack* HDS). A *stack* visa contribuir para a minimização de alguns problemas identificados nas áreas de Big Data e Smart Grids. Assim, foram desenvolvidos e implementados os seguintes serviços:

- Visualização e monitorização dos dados recolhidos em tempo real, que permitem a deteção de anomalias contidas nos dados recolhidos;
- Preparação dos dados recolhidos em tempo real, i.e., normalização, validação e agregação dos dados;
- Previsão em tempo real de 33 séries temporais;

- Deteção de anomalias relacionadas com o comportamento incorreto no consumo de energia e com possíveis falhas do sistema;
- Avaliação da precisão da previsão e geração de novos modelos para a previsão de consumo e produção de energia de acordo com determinados critérios, de forma a garantir a automatização do processo analítico.

Destacam-se alguns pormenores na implementação destes serviços por terem contribuído significativamente para a performance e eficácia da *stack,* nomeadamente o tratamento de fluxos contínuos e desordenados, definição de regras (i.e., regras básicas e regras de primeiro e segundo nível), imposição de critérios para a deteção de anomalias e para a geração de novos modelos de previsão, implementação da funcionalidade *Listen and Notify* do PostgresSQL e o escalonamento dos serviços.

De salientar ainda que, foi necessário adotar contramedidas para fazer face ao insucesso obtido na implementação de alguns componentes, dos quais se destacam a plataforma BDE e a *framework Apache Beam*. Destas contramedidas resultou o desenvolvimento de três imagens *Docker Container*, i.e., *agente:agente*, *agente:Inspetor* e *agente:forecasting*.

Finalizada a implementação da *stack*, conclui-se que os objetivos propostos foram concretizados. A flexibilidade no modelo proposto para a plataforma HDS permitiu a implementação de contramedidas face ao insucesso obtido na integração de alguns componentes. Por outro lado, a *stack* HDS, apesar dos seus recursos limitados de hardware, executa com eficiência a pré-prepararão dos dados em tempo real. Executa ainda a detecção de anomalias e a previsão de energia para 33 séries temporais em tempo real e de forma autónoma, podendo desta forma contribuir significativamente para a tomada de decisão em tempo real, no contexto das Smart Grids.

Finalmente, a fim de se avaliar a eficiência dos serviços a serem executados pela *stack* HDS, foram elaborados vários casos de estudo dos quais se concluiu que os critérios configurados para a geração de novos modelos preditivos estão de acordo com o espectável, tendo contribuído para o sucesso da automação analítica implementada na *stack*. Por outro lado, conclui-se que quanto mais específica e particularizada for a configuração de regras na pré-prepararão de dados, melhores serão os resultados obtidos na precisão das previsões. De igual modo, quanto mais específica for a implementação das previsões (i.e., por zona e por tipo), maior será a probabilidade de se obter bons resultados. Por fim, analisando-se o impacto da aplicação da abordagem *Edge Analytic*, conclui-se que esta pode contribuir significativamente e positivamente para a obtenção de melhores resultados na precisão das previsões.

7.2 Objetivos realizados

O principal objetivo desta tese pretende a aplicação das novas abordagens operadas na área de Big Data no contexto da Smart Grids. Assim, focando este objetivo nos grandes desafios inerentes quer à área de Big Data quer à área das Smart Grids, foram delineados os seguintes objetivos:

Objetivos na área de Smart Grids:

- Melhorar o resultado das previsões (i.e., produção e consumo), com a agregação de informação de várias fontes. Para a execução deste objetivo propôs-se um estudo inicial das técnicas de Big Data. Em resultado desse estudo visou-se a aplicação, adaptação ou desenvolvimento de novas metodologias que permitam a concretização do objetivo;
- Processamento em tempo real nos sistemas de gestão das Smart Grids com vista a melhorar a execução de modelos analíticos. Para atingir este objetivo propôs-se um estudo das metodologias de processamento em tempo real existentes no domínio de Big Data e sua posterior aplicação, adaptação, ou caso necessário o desenvolvimento de novas metodologias;
- Melhorar a interação e comunicação entre os intervenientes da rede. Para a concretização deste objetivo propôs-se um estudo inicial visando as tecnologias *Big Data Visualization*, para uma posterior aplicação nos sistemas de Smart Grid.

Objetivos na área de Big Data:

- Estudo sobre as evoluções tecnológicas operadas na área de Big Data bem como a sua aplicabilidade na área de Smart Grids, contemplando soluções *open source* e comerciais;
- Aplicação, adaptação ou desenvolvimento de novas metodologias para o melhoramento do processo analítico relacionado com as previsões;
- Aplicação, adaptação ou desenvolvimento de novas metodologias para o melhoramento do processamento e gestão de grandes volumes de dados em tempo real;
- Conceção e desenvolvimento de aplicações que utilizem as metodologias desenvolvidas no contexto das Smart Grids.

Dos objetivos delineados pode-se afirmar que todos foram possíveis de ser concretizados e cujos principais contributos podem ser analisados no subcapítulo seguinte.

7.3 Principais contributos

Os principais contributos e originalidades do presente trabalho, conforme já referido no capítulo introdutório, podem ser resumidos da seguinte forma:

- Apresentação de um estudo sobre as evoluções tecnológicas operadas na área de *Big Data* bem como a sua aplicabilidade na área de *Smart Grids*. O estudo identifica os grandes desafios inerentes à área de *Big Data* dos quais se salientam o processamento em tempo real de fluxo de dados contínuos e desordenados, o tempo despendido na pré-preparação de dados e ainda a automatização analítica. O estudo destaca ainda os grandes desafios na área das *Smart Grids*, nomeadamente a complexidade inerente ao seu fluxo de dados e a necessidade imprescindível

de o tratar e processar em tempo real. Por outro lado, realça o grande desafio inerente à precisão das previsões do consumo e produção de energia, sendo que estas duas variáveis assumem uma importância relevante no equilíbrio e sustentabilidade de todo o ecossistema energético. Finalmente, avalia o impacto positivo na evolução das Smart Grids, que se obteria com a integração das abordagens tecnológicas operadas na área de Big Data;

- Proposta de um modelo conceptual para a implementação de uma plataforma flexível, com capacidade para permitir a integração, experimentação e avaliação das novas tecnologias desenvolvidas no âmbito de Big Data, aplicadas ao contexto das Smart Grids;
- Desenvolvimento de uma *stack* objetivando testar o modelo proposto e a implementação de serviços capazes de responder positivamente aos desafios identificados em Big Data (i.e., processamento em tempo real visando a pré-preparação de dados e a automatização analítica) e nas Smart Grids (i.e., deteção de anomalia e processamento em tempo real da previsão de consumo e produção de energia com a maior precisão possível). Os serviços implementados e disponíveis na *stack* são: visualização e monitorização dos dados recolhidos em tempo real; preparação dos dados recolhidos em tempo real; previsão em tempo real; deteção de anomalias; avaliação da precisão da previsão e geração de novos modelos para a previsão de consumo/produção de energia segundo determinados critérios;
- Desenvolvimento de vários casos de estudo visando a experimentação e validação dos serviços implementados e disponíveis na *stack*, dos quais foi possível extrair conclusões sobre a precisão das previsões, a importância da pré-preparação de dados em tempo real, a eficácia da automatização analítica e a viabilidade da implementação da abordagem *Edge Analytic*.

Relativamente aos contributos originais do presente trabalho realça-se o desenvolvimento da *stack* HDS, que conforme mostra Figura 7.1, visou a resolução de grandes desafios, quer na área de Big Data quer na área das Smart Grids.

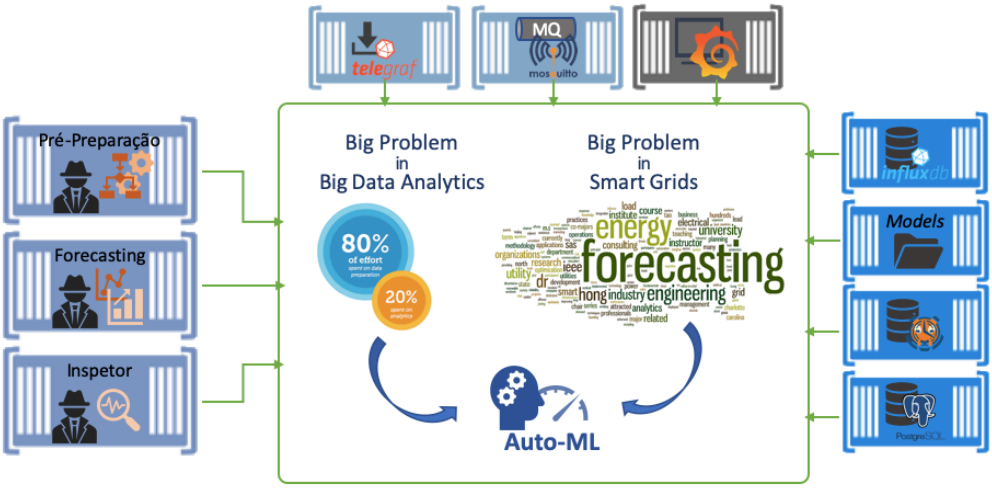


Figura 7.1 - Stack HDS: Solução para os grandes desafios nas áreas de Big Data e Smart Grids

As imagens desenvolvidas e encapsuladas nos *containers Agente Pré-Preparação, Agente Forcasting* e *Agente Inspetor* constituem uma proposta original, assim como as metodologias nelas integradas como a seguir se descreve:

- *Agente Pré-Preparação*: As funcionalidades de agregação e transformação de dados começam a transformar-se numa realidade graças aos desenvolvimentos realizados na área de Big Data. O mesmo se começa a verificar no processamento de fluxos de dados infinitos e desordenados. No entanto, a existência de componentes que a par destas funcionalidades tenham a capacidade de validar e refazer dados não é propriamente uma realidade. A metodologia implementada neste agente constitui uma proposta original, não tendo sido usada anteriormente na área de Big Data ou na área das Smart Grids;
- *Agente Forecasting*: Este agente desenvolvido apenas para executar previsões de acordo com o modelo que lhe foi facultado, tem a particularidade de poder ser configurado de forma a se determinar sobre quais séries temporais deve executar a previsão, de quanto em quanto tempo e para que espaço temporal. Estas funcionalidades constituem uma proposta inovadora no cenário de previsões em tempo real no contexto das Smar Grids.
- *Agente Inpetor*: Este agente tem implementada uma metodologia para a deteção de anomalias baseada em regras. Serve-se do resultado desta validação para a execução de outra metodologia que visa a verificação dos resultados da precisão das previsões. A conjugação destas duas metodologias constitui a base para a tomada de decisão por parte do agente relativamente à geração de novos modelos preditivos. A metodologia no seu todo, constitui uma proposta não tendo sido usada anteriormente nas áreas de Big Data e Smart Grids.

Para além do contributo original dos agentes desenvolvidos, a Stack HDS proposta constitui uma contribuição original na área de Big Data aplicada às Smart Grids relativamente à pré-preparação de dados em tempo real e consequentemente à analítica automatizada das previsões com precisões muito satisfatórias. Ainda relativamente à automatização analítica, a implementação distribuída da Framework ML.NET bem como a sua experimentação no âmbito das previsões de energia é uma proposta original do presente trabalho não tendo sido usada anteriormente na área das Smart Grids. Por outro lado, a flexibilidade e agilidade, que caracteriza a *stack* HDS baseada em *Docker Containers*, é uma proposta original na área de Big Data aplicada a Smart Grids. O projeto Big Data Europe é a única plataforma passível de ser instalada localmente, contemplando à área das Smart Grids e tendo como base a arquitetura *Docker Container*. No entanto, a sua funcionalidade depende de um *workflow* proprietário e de componentes Big Data cujo funcionamento foi ajustado à própria plataforma. Ao contrário da BDE, a plataforma HDS foi desenvolvida com o objetivo de ser completamente flexível à incorporação de novos serviços, dependendo apernas da compatibilidade entre os componentes disponibilizados pela

área de Big Data e dos desenvolvimentos disponibilizados pela arquitetura *Docker Container*, sem, no entanto, indisponibilizar a integração e implementação de novos componentes.

Para além do contributo original do modelo conceptual da plataforma HDS sobre o qual foi implementada a *stack* HDS, é ainda de destacar o contributo original relacionado com a prova das mais valias trazidas pela aplicação da *Edge Analytics* para a precisão das previsões, uma vez que nunca antes foi provado se de facto a sua implementação implicaria prejuízos ou benefícios na precisão das previsões.

7.4 Limitações & trabalho futuro

Foram muitas as limitações no desenvolvimento do presente trabalho, as quais se podem sintetizar nos recursos de hardware para a implementação da *stack* e na imaturidade de muitas abordagens disponibilizadas na área de Big Data. Relativamente à limitação dos recursos de hardware a limitação foi transformada num desafio. De facto, a necessidade de implementar uma plataforma para a gestão do fluxo de dados com as características de Big Data com recursos de hardware tão limitados levou a uma análise muito pormenorizada sobre que componentes usar, como os implementar, como os distribuir, como os conectar, etc. Só desta forma foi possível desenvolver uma *stack* com a capacidade de pré-preparar dados em tempo real e executar previsões de consumo e produção para 33 series temporais em tempo real. À *stack* implementada foi ainda adicionada a capacidade de detetar anomalias e o poder de decisão no processo de automatização analítica, de forma a garantir com a maior exatidão possível a precisão das previsões.

Relativamente à segunda limitação referida, i.e., imaturidade nas abordagens disponibilizadas na área de Big Data, concretamente abordagens *open source*, é extremamente compreensível visto que esta área é relativamente recente. No mundo *open source* relacionado com novas e recentes abordagens tecnológicas, o termo bug é a palavra de ordem. No entanto, este foi encarado no presente trabalho como um desafio. A procura constante de soluções que permitissem ultrapassar esta limitação traduziu-se numa excelente oportunidade de aquisição de conhecimentos. Por outro lado, permitiu comprovar o quanto ágil é o modelo conceptual proposto na presente tese, caso contrário seria impossível concretizar os objetivos propostos.

Como trabalho futuro, propõem-se melhorias nos serviços já implementados na *stack*, o desenvolvimento e a integração de novos serviços de forma a se concretizar na totalidade o proposto no modelo conceptual, como a seguir se descrevem:

- Melhorias nos serviços já implementados:
    - Integrar e validar os novos algoritmos disponibilizados recentemente na *framework* ML.NET (i.e., *Ordinary Least Squares* (OLS) [232] e *LightGbm* [233]);

- Extrair e armazenar os *logs* dos *containers Agentes Inspetores*, relativamente à geração de novos modelos, para facilitar a sua análise. É necessário aceder a cada um dos *containers* para se obter os *logs,* o que torna a sua análise um processo difícil. No entanto, esta análise pode ser muito útil para melhor se entender o comportamento das previsões, e.g., com que frequência os modelos são gerados, quais os modelos que frequentemente são candidatados em determinadas zonas, etc.;
- Desenvolver um agente para agregar todos os resultados obtidos nas previsões. Conforme se concluiu da análise feita às previsões, no capítulo 6, o melhor resultado para a previsão do edifício, resultou deste somatório. Assim, seria interessante ter este valor representado no *dashboard*;
- Melhorar a configuração dos serviços no *dashboard*, e.g., configurar os alertas que estão a ser enviados para o Broker MQTT, etc.;
- Desenvolver uma interface mais amigável para a manutenção das regras de primeiro e segundo nível, que no presente momento apenas é possível efetuar com acesso direto à Base de Dados;
- Implementação de um plano de *disaster recovery*. Apesar de *stack* ser tolerante a falhas, a ocorrência de falhas prolongadas de rede ou da energia traz como consequência a perda na pré-preparação de dados, isto porque as três máquinas físicas estão instaladas no mesmo edifício e no mesmo domínio de rede. Se a folha for inferior a uma hora, o *Agente Pré-preparação* está dotado para fazer a recuperação e sincronização automática dos dados. No entanto, se a falha se verificar por um período superior, esta sincronização tem de ser feita de forma manual para evitar inconsequências no histórico de dados pré-preparados. Assim, propõem-se o desenvolvimento de um componente para a sincronização dos dados em coso de falhas prolongadas de rede ou energia, e ainda a instalação de *Uninterruptible Power Supply* (UPS) bem como, a distribuição das máquinas pelos diferentes edifícios do campus do ISEP.

▪ Desenvolvimento e integração de novos serviços:

- Desenvolver a camada de serviços com o objetivo de integrar outras operações diretamente ou indiretamente relacionadas com a operabilidade do ecossistema energético;
- Ainda na camada de serviço, propõem-se como trabalho futuro a experimentação e integração de algumas bibliotecas para a área de Data Science, que já se encontram disponíveis em *Docker Container* e.g., Anaconda [234], Flask [235], FloydHub [236], etc.;

- Desenvolvimento de uma *cloud*, e.g. com recurso à *framework OpenStack*;
- Implementação de um Data Lake na *cloud* desenvolvida de forma a facilitar a partilha de dados entre a comunidade de I&D sem pôr em causa a privacidade de dados pessoais;
- Implementação, integração e simulação de componentes de forma a garantirem todas as atividades relacionadas com a governação dos dados, das quais se destacam todas as que estão diretamente relacionadas com a privacidade de dados privados, de forma a garantir que a utilização dos dados se encontra de acordo com os regulamentos impostos por lei. Destas atividades destacam-se a segurança dos dados (i.e., implementação de políticas de acesso, definição de permissões, aplicação de algoritmos de anonimização ou outras técnicas que garantam a privacidade dos dados) e a linhagem dos dados (i.e., definição dos processos e operações executadas sobre os dados). Numa primeira fase propõe-se a experimentação dos componentes Apache Range, Apache Alfa e Apache Nifi;
- Verificação dos avanços realizados e disponibilizados na *framework* Apache Beam a fim de concretizar a sua integração;
- Dar continuidade à integração e experimentação de outros componentes Big Data, de forma a validar as vantagens e desvantagens dos componentes já implementados;
- Apesar de ser possível obter informação sobre os recursos de hardware consumidos pelos *containers* em execução através dos comandos do *Docker Container*, esta é de pouca usabilidade. Assim, para facilitar a monitorização dos recursos de hardware da *stack*, propõe-se como trabalho futuro a recolha das métricas dos *hosts* e dos *containers* e a sua representação no *dashboard*.

Por fim, como trabalho futuro de âmbito mais específico propõe-se um estudo mais aprofundado das metodologias analíticas relacionadas com as previsões. Propõe-se o estudo das que já estão integradas de forma a se perceber que relação existe entre elas e as eventuais características das séries temporais (i.e., quais as possíveis condições que determinam a adoção de determinada metodologia em detrimento de outra). No mesmo âmbito propõe-se o estudo, a integração e adaptação de novas metodologias.

7.5 Apreciação final

Como apreciação final do trabalho realizado na presente dissertação cabe referir que, face à extrema complexidade que envolve as duas áreas em estudo, i.e., Big Data e Smart Grids, a concretização dos objetivos propostos resultou em enorme satisfação. Por ouro lado, conforme explanado no subcapítulo anterior, o trabalho aqui realizado abre horizontes para a exploração de soluções para muitos dos desafios inerentes a ambas as áreas.

Por fim, cabe ainda referir que o presente trabalho culminou numa enorme aprendizagem. Espero no futuro continuar a ter a oportunidade de juntar ao conhecimento adquirido, outro tanto, e com ele continuar a contribuir de forma positiva para novos desafios.

Bibliografia

[1] S. Celar, E. Mudnic, and Z. Seremet, “State-Of-The-Art of Messaging for Distributed Computing Systems,” in *DAAAM Proceedings*, 1st ed., vol. 1, B. Katalinic, Ed. DAAAM International Vienna, 2016, pp. 0298–0307.

[2] J. Korab, *Understanding Message Brokers*. OReilly, 2017.

[3] “AMQP.” [Online]. Available: http://www.amqp.org/. [Accessed: 05-Mar-2018].

[4] “MQTT.” [Online]. Available: http://mqtt.org/. [Accessed: 05-Mar-2018].

[5] “STOMP.” [Online]. Available: http://stomp.github.io/. [Accessed: 05-Mar-2018].

[6] “XMPP.” [Online]. Available: https://xmpp.org/. [Accessed: 05-Mar-2018].

[7] “Apache ActiveMQ ™ -- Index.” [Online]. Available: http://activemq.apache.org/. [Accessed: 30-Mar-2018].

[8] “ZeroMQ Feature List - zeromq.” [Online]. Available: http://zeromq.org/docs:features. [Accessed: 30-Mar-2018].

[9] G. Lilis and M. Kayal, “A secure and distributed message oriented middleware for smart building applications,” *Autom. Constr.*, vol. 86, pp. 163–175, Feb. 2018.

[10] M. Araya *et al.*, “A New ACS Bulk Data Transfer Service for CTA,” p. THBPL03, Jan. 2018.

[11] “RabbitMQ - Messaging that just works.” [Online]. Available: https://www.rabbitmq.com/. [Accessed: 30-Mar-2018].

[12] “Apache Kafka,” *Apache Kafka*. [Online]. Available: https://kafka.apache.org/intro.html.

[13] “Amazon MQ - Developer Guide,” p. 78.

[14] P. Dobbelaere and K. S. Esmaili, “Kafka versus RabbitMQ,” *ArXiv170900333 Cs*, Sep. 2017.

[15] V. John and X. Liu, “A Survey of Distributed Message Broker Queues,” *ArXiv170400411 Cs*, Apr. 2017.

[16] C. N. Nguyen, J. Lee, S. Hwang, and J.-S. Kim, “On the role of message broker middleware for many-task computing on a big-data platform,” *Clust. Comput.*, pp. 1–14, Mar. 2018.

[17] P. Y. Lai, C. R. Dow, and Y. Y. Chang, “Rapid-Response Framework for Defensive Driving based on Internet of Vehicles Using Message-Oriented Middleware,” *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2018.

[18] “Eclipse Mosquitto,” *Eclipse Mosquitto*, 08-Jan-2018. [Online]. Available: https://mosquitto.org/. [Accessed: 30-Mar-2018].

[19] O. Andreassen, F. Marazita, and M. Miskowiec, “Upgrade of the CERN RADE framework architecture using RabbitMQ and MQTT,” p. THPHA038, Jan. 2018.

[20] M. Dayarathna and S. Perera, “Recent Advancements in Event Processing,” *ACM Comput Surv*, vol. 51, no. 2, pp. 33:1–33:36, Feb. 2018.

[21] Tyler Akidau, Slava Chernyak, and Reuven Lax, *Streaming Systems:The What, Where, When, & How of Large-Scale Data Processing*, Book Preview. O’Reilly Media, 2017.

[22] C. Chambers *et al.*, “FlumeJava: Easy, Efficient Data-Parallel Pipelines,” 2010.

[23] “Apache Storm.” [Online]. Available: http://storm.apache.org/. [Accessed: 03-Apr-2018].

[24] “Spark Streaming | Apache Spark.” [Online]. Available: https://spark.apache.org/streaming/. [Accessed: 02-Apr-2018].

[25] “Cloud Dataflow — Processamento de dados de stream e em lote,” *Google Cloud*. [Online]. Available: https://cloud.google.com/dataflow/?hl=pt-br. [Accessed: 03-Apr-2018].

[26] “Apache Flink: Scalable Stream and Batch Data Processing.” [Online]. Available: https://flink.apache.org/. [Accessed: 03-Apr-2018].

[27] “Apache Apex.” [Online]. Available: https://apex.apache.org/. [Accessed: 03-Apr-2018].

[28] “Apache Apex Malhar Documentation.” [Online]. Available: http://apex.apache.org/docs/malhar/#operator-library-overview. [Accessed: 03-Apr-2018].

[29] “Apache Beam.” [Online]. Available: https://beam.apache.org/. [Accessed: 03-Apr-2018].

[30] F. Ultremare, *traffic-flow: Traffic Flow Demo*. 2017.

[31] *cloud-dataflow-nyc-taxi-tycoon: This is the support code and solutions for the NYC Taxi Tycoon Dataflow Codelab*. Google Codelabs, 2018.

[32] A. Shukla, S. Chaturvedi, and Y. Simmhan, “RIoTBench: An IoT benchmark for distributed stream processing systems,” *Concurr. Comput. Pract. Exp.*, vol. 29, no. 21, p. e4257, Nov. 2017.

[33] A. Luckow, G. Chantzialexiou, and S. Jha, “Pilot-Streaming: A Stream Processing Framework for High-Performance Computing,” *ArXiv180108648 Cs*, Jan. 2018.

[34] A. Bifet *et al.*, “Extremely Fast Decision Tree Mining for Evolving Data Streams,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2017, pp. 1733–1742.

[35] G. Hesse, C. Matthies, B. Reissaus, and M. Uflacker, “A New Application Benchmark for Data Stream Processing Architectures in an Enterprise Context: Doctoral Symposium,” in *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, New York, NY, USA, 2017, pp. 359–362.

[36] J. Karimov, T. Rabl, A. Katsifodimos, R. Samarev, H. Heiskanen, and V. Markl, “Benchmarking Distributed Stream Processing Engines,” *ArXiv180208496 Cs*, Feb. 2018.

[37] S. Chintapalli *et al.*, “Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming,” 2016, pp. 1789–1792.

[38] A. K. Rastogi, N. Malik, and S. Hooda, “Exactly-Once Semantics with Real-Time Data Pipelines,” in *Ambient Communications and Computer Systems*, Springer, Singapore, 2018, pp. 293–303.

[39] G. D. Nugraha, A. Musa, J. Cho, K. Park, and D. Choi, “Lambda-Based Data Processing Architecture for Two-Level Load Forecasting in Residential Buildings,” *Energies*, vol. 11, no. 4, p. 772, Mar. 2018.

[40] E. Qadah, E. Alevizos, M. Mock, and G. Fuchs, “A Distributed Online Learning Approach for Pattern Prediction over Movement Event Streams with Apache Flink,” p. 8, Mar. 2018.

[41] A.-C. Sima, K. Stockinger, K. Affolter, M. Braschler, P. Monte, and L. Kaiser, “A hybrid approach for alarm verification using stream processing, machine learning and text analytics,” presented at the International Conference on Extending Database Technology (EDBT), March 26-29, 2018, 2018.

[42] J. Traub *et al.*, “Scotty: Efficient Window Aggregation for out-of-order Stream Processing,” p. 4, 2018.

[43] J. Lohokare, R. Dani, A. Rajurkar, and A. Apte, “An IoT ecosystem for the implementation of scalable wireless home automation systems at smart city level,” 2017, pp. 1503–1508.

[44] S. Kamburugamuve, K. Ramasamy, M. Swany, and G. Fox, “Low Latency Stream Processing: Apache Heron with Infiniband & Intel Omni-Path,” 2017, pp. 101–110.

[45] P. Carbone *et al.*, “Large-Scale Data Stream Processing Systems,” in *Handbook of Big Data Technologies*, Springer, Cham, 2017, pp. 219–260.

[46] F. Versaci, L. Pireddu, and G. Zanetti, “Kafka interfaces for composable streaming genomics pipelines,” *bioRxiv*, p. 182030, Aug. 2017.

[47] M. P. Singh, M. A. Hoque, and S. Tarkoma, “A survey of systems for massive stream analytics,” *ArXiv160509021 Cs*, May 2016.

[48] M. Stonebraker, “SQL databases v. NoSQL databases,” *Commun. ACM*, vol. 53, no. 4, p. 10, Apr. 2010.

[49] R. Cattell, “Scalable SQL and NoSQL data stores,” *ACM SIGMOD Rec.*, vol. 39, no. 4, p. 12, May 2011.

[50] E. Brewer, “CAP twelve years later: How the ‘rules’ have changed,” *Computer*, vol. 45, no. 2, pp. 23–29, Feb. 2012.

[51] MikeWasson, “Choose the right data store.” [Online]. Available: https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/data-store-overview. [Accessed: 21-Nov-2017].

[52] “DB-Engines Ranking - popularity ranking of database management systems.” [Online]. Available: https://db-engines.com/en/ranking. [Accessed: 21-Nov-2017].

[53] “NOSQL Databases.” [Online]. Available: http://nosql-database.org/. [Accessed: 21-Sep-2017].

[54] MikeWasson, “Criteria for choosing a data store.” [Online]. Available: https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/data-store-comparison. [Accessed: 11-Oct-2017].

[55] Stefan Edlich, “Databases & Psychology,” *GOTO Magazine*, vol. 2, no. 2, pp. 28–30, Aug-2012.

[56] A. Pavlo and M. Aslett, “What’s Really New with NewSQL?,” *SIGMOD Rec*, vol. 45, no. 2, pp. 45–55, Sep. 2016.

[57] Pete Aven and Diane Burley, *Building on Multi-Mode Databases - How to Manage Multiple Schemas Using a Single Platform*, First Edition. Gravenstein Highway North, Sebastopol, CA: O’Reilly Media, Inc., 2017.

[58] J. Lu, Z. H. Liu, P. Xu, and C. Zhang, “UDBMS: Road to Unification for Multi-model Data Management,” *ArXiv161208050 Cs*, Dec. 2016.

[59] Jiaheng Lu, “Towards Benchmarking Multi-Model Databases,” presented at the CIDR 2017. 8th Biennial Conference on Innovative Data Systems Research, California, USA, 2017.

[60] “UniBench.” [Online]. Available: http://udbms.cs.helsinki.fi/?projects/ubench. [Accessed: 21-Nov-2017].

[61] V. Reniers, D. Van Landuyt, A. Rafique, and W. Joosen, “On the State of NoSQL Benchmarks,” 2017, pp. 107–112.

[62] F. R. Oliveira and L. del Val Cura, “Performance Evaluation of NoSQL Multi-Model Data Stores in Polyglot Persistence Applications,” in *Proceedings of the 20th International Database Engineering & Applications Symposium*, New York, NY, USA, 2016, pp. 230–235.

[63] A. C. Weinberger, “Benchmark: PostgreSQL, MongoDB, Neo4j, OrientDB and ArangoDB,” *ArangoDB*, 13-Oct-2015. .

[64] A. C. Weinberger, “Performance comparison between ArangoDB, MongoDB, Neo4j and OrientDB,” *ArangoDB*, 11-Jun-2015. .

[65] Tomcy John and Pankaj Misra, *Data Lake for Enterprises*. Packt Book, 2017.

[66] B. Inmon, *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications, 2016.

[67] CEN-CENELEC-ETSI, "SG-CG/M490/H_ Smart Grid Information Security," Dec. 2014.

[68] M. Ismail, S. Niazi, M. Ronstrom, S. Haridi, and J. Dowling, "Scaling HDFS to More Than 1 Million Operations Per Second with HopsFS," 2017, pp. 683–688.

[69] G. Donvito, G. Marzulli, and D. Diacono, "Testing of several distributed file-systems (HDFS, Ceph and GlusterFS) for supporting the HEP experiments analysis," *J. Phys. Conf. Ser.*, vol. 513, no. 4, p. 042014, 2014.

[70] D. Quintero *et al.*, *IBM Spectrum Scale (formerly GPFS)*. IBM Redbooks, 2017.

[71] "Alluxio - Open Source Memory Speed Virtual Distributed Storage," *Alluxio*. [Online]. Available: http://www.alluxio.org. [Accessed: 08-Apr-2018].

[72] Sean Patrick Murphy, *Data and Electric Power - From Deterministic Machines to Probabilistic Systems in Traditional Engineering*, First Edition. Gravenstein Highway North, Sebastopol, CA: O'Reilly Media, Inc., 2016.

[73] "Top 10 Strategic Technology Trends for 2018," Oct-2017. [Online]. Available: https://www.gartner.com/doc/3811368/top--strategic-technology-trends. [Accessed: 28-Feb-2018].

[74] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Zomaya, "Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid," *IEEE Trans. Big Data*, vol. PP, no. 99, pp. 1–1, 2017.

[75] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *ArXiv180304311 Cs*, Mar. 2018.

[76] Peter Krensky and Jim Hare, "Hype Cycle for Data Science and Machine Learning, 2017," Gartner, ID: G00325005, Jul. 2017.

[77] G. Li, "Human-in-the-loop Data Integration," *Proc. VLDB Endow.*, vol. 10, p. 12, 2017.

[78] N. Bikakis, "Big Data Visualization Tools," *ArXiv180108336 Cs*, Jan. 2018.

[79] A. S. S. Fiaz, N. Asha, D. Sumathi, and A. S. S. Navaz, "Data Visualization: Enhancing Big Data More Adaptable and Valuable," vol. 11, no. 4, p. 4, 2016.

[80] S. M. Ali, N. Gupta, G. K. Nayak, and R. K. Lenka, "Big data visualization: Tools and challenges," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016, pp. 656–660.

[81] L. Wang, G. Wang, and C. A. Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress," 2015.

[82] R. Agrawal, A. Kadadi, X. Dai, and F. Andres, “Challenges and Opportunities with Big Data Visualization,” in *Proceedings of the 7th International Conference on Management of Computational and Collective intElligence in Digital EcoSystems*, New York, NY, USA, 2015, pp. 169–173.

[83] D. International, *DAMA-DMBOK: Data Management Body of Knowledge*, Edição: Second. Basking Ridge, New Jersey: Technics Publications, 2017.

[84] A. Al-Badi, A. Tarhini, and A. I. Khan, “Exploring Big Data Governance Frameworks,” *Procedia Comput. Sci.*, vol. 141, pp. 271–277, Jan. 2018.

[85] R. S. Moghadam and R. Colomo-Palacios, “Information security governance in big data environments: A systematic mapping,” *Procedia Comput. Sci.*, vol. 138, pp. 401–408, Jan. 2018.

[86] “Apache Atlas – Data Governance and Metadata framework for Hadoop.” [Online]. Available: http://atlas.apache.org/. [Accessed: 11-Oct-2017].

[87] “Apache Ranger – Introduction.” [Online]. Available: https://ranger.apache.org/. [Accessed: 11-Oct-2017].

[88] CEN-CENELEC-ETSI, “Smart Grid Reference Architecture,” Nov. 2012.

[89] “Governor Cuomo Announces Grand Opening of Digital Command Center to Monitor NYPA Power Plant Operations, Encourage Energy Innovation, and Improve Operational Efficiencies,” *Governor Andrew M. Cuomo*, 11-Dec-2017. [Online]. Available: https://www.governor.ny.gov/news/governor-cuomo-announces-grand-opening-digital-command-center-monitor-nypa-power-plant. [Accessed: 20-Dec-2017].

[90] “Digitising the energy sector: an opportunity for Europe,” *Digital Single Market*. [Online]. Available: https://ec.europa.eu/digital-single-market/en/blog/digitising-energy-sector-opportunity-europe. [Accessed: 31-Dec-2017].

[91] Department of Energy, “Modern Distribution Grid: Volume-II,” Mar. 2017.

[92] F. Gangale, J. Vasiljevska, C. F. Covrig, A. Mengolini, and G. Fulli, “Smart grid projects outlook 2017,” 2017.

[93] S. Grid *et al.*, “NIST Special Publication 1108r3 NIST Framework and Roadmap for Smart Grid Interoperability NIST Special Publication 1108r3 NIST Framework and Roadmap for Smart Grid Interoperability,” 2014. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.1108r3.

[94] EUROPEAN COMMISSION, “Smart Grids: from innovation to deployment,” *Brussels, 12.4.2011 COM(2011) 202 final*, 2011. [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0202:FIN:EN:PDF.

[95] Y. Zhang, W. Chen, and W. Gao, “A survey on the development status and challenges of smart

grids in main driver countries," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 137–147, Nov. 2017.

[96] "WEO 2017." [Online]. Available: https://www.iea.org/weo2017/. [Accessed: 01-May-2018].

[97] H. Akhavan-Hejazi and H. Mohsenian-Rad, "Power systems big data analytics: An assessment of paradigm shift barriers and prospects," *Energy Rep.*, vol. 4, pp. 91–100, Nov. 2018.

[98] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big Data management in smart grid: concepts, requirements and implementation," *J. Big Data*, vol. 4, no. 1, Dec. 2017.

[99] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid – A review," *Renew. Sustain. Energy Rev.*, vol. 79, pp. 1099–1107, Nov. 2017.

[100] W.-L. Chin, W. Li, and H.-H. Chen, "Energy Big Data Security Threats in IoT-Based Smart Grid Communications," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 70–75, Oct. 2017.

[101] A. Bhardwaj and W. Singh, "Systematic Review of Smart Grid Analytics," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 2333–2338, Jun. 2017.

[102] N. Koseleva and G. Ropaite, "Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges," *Procedia Eng.*, vol. 172, pp. 544–549, Jan. 2017.

[103] S. Tan, D. De, W.-Z. Song, J. Yang, and S. K. Das, "Survey of Security Advances in Smart Grid: A Data Driven Approach," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 1, pp. 397–422, 2017.

[104] Mohsen Marjani, Fariza Nasaruddin, and Abdullah Gani, "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.

[105] J. Hu and A. V. Vasilakos, "Energy Big Data Analytics and Security: Challenges and Opportunities," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2423–2436, Sep. 2016.

[106] H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. Zhang, "Energy big data: A survey," *IEEE Access*, vol. 4, pp. 3844–3861, 2016.

[107] H. K. Nguyen, A. Khodaei, and Z. Han, "A Big Data Scale Algorithm for Optimal Scheduling of Integrated Microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 1, pp. 274–282, Jan. 2018.

[108] I. Jabłoński, "Graph Signal Processing in Applications to Sensor Networks, Smart Grids, and Smart Cities," *IEEE Sens. J.*, vol. 17, no. 23, pp. 7659–7666, Dec. 2017.

[109] Z. Cao, J. Lin, C. Wan, Y. Song, G. Taylor, and M. Li, "Hadoop-based framework for big data analysis of synchronised harmonics in active distribution network," *IET Gener. Transm. Distrib.*, vol. 11, no. 16, pp. 3930–3937, Nov. 2017.

[110] B. Li, M. C. Kisacikoglu, C. Liu, N. Singh, and M. Erol-Kantarci, "Big Data Analytics for Electric Vehicle Integration in Green Smart Cities," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 19–25, Nov. 2017.

[111] A. A. Munshi and Y. A.-R. I. Mohamed, “Big data framework for analytics in smart grids,” *Electr. Power Syst. Res.*, vol. 151, pp. 369–380, Oct. 2017.

[112] W. Chen, K. Zhou, S. Yang, and C. Wu, “Data quality of electricity consumption data in a smart grid environment,” *Renew. Sustain. Energy Rev.*, vol. 75, pp. 98–105, Aug. 2017.

[113] Z. Asad and M. A. R. Chaudhry, “A Two-Way Street: Green Big Data Processing for a Greener Smart Grid,” *IEEE Syst. J.*, vol. 11, no. 2, pp. 784–795, Jun. 2017.

[114] T. e Huang, Q. Guo, and H. Sun, “A Distributed Computing Platform Supporting Power System Security Knowledge Discovery Based on Online Simulation,” *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1513–1524, May 2017.

[115] K. Wang *et al.*, “Wireless Big Data Computing in Smart Grid,” *IEEE Wirel. Commun.*, vol. 24, no. 2, pp. 58–64, Apr. 2017.

[116] J. C. S. de Souza, T. M. L. Assis, and B. C. Pal, “Data Compression in Smart Distribution Systems via Singular Value Decomposition,” *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 275–284, Jan. 2017.

[117] J. Wu, K. Ota, M. Dong, J. Li, and H. Wang, “Big Data Analysis based Security Situational Awareness for Smart Grid,” *IEEE Trans. Big Data*, vol. PP, no. 99, pp. 1–1, 2017.

[118] Z. Zhou *et al.*, “Game-Theoretical Energy Management for Energy Internet With Big Data-Based Renewable Power Forecasting,” *IEEE Access*, vol. 5, pp. 5731–5746, 2017.

[119] S. Singh and A. Yassine, “Mining Energy Consumption Behavior Patterns for Households in Smart Grid,” *IEEE Trans. Emerg. Top. Comput.*, vol. PP, no. 99, pp. 1–1, 2017.

[120] H. Shi, M. Xu, and R. Li, “Deep Learning for Household Load Forecasting - A Novel Pooling Deep RNN,” *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[121] Z. Shah, A. Anwar, A. N. Mahmood, Z. Tari, and A. Y. Zomaya, “A Spatio-temporal Data Summarization Paradigm for Real-time Operation of Smart Grid,” *IEEE Trans. Big Data*, vol. PP, no. 99, pp. 1–1, 2017.

[122] R. Moghaddass and J. Wang, “A Hierarchical Framework for Smart Grid Anomaly Detection Using Large-Scale Smart Meter Data,” *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[123] C. Li, C. Liu, K. Deng, X. Yu, and T. Huang, “Data-Driven Charging Strategy of PEVs Under Transformer Aging Risk,” *IEEE Trans. Control Syst. Technol.*, vol. PP, no. 99, pp. 1–14, 2017.

[124] J. Kwac, J. I. Kim, and R. Rajagopal, “Efficient customer selection process for various DR objectives,” *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[125] W. Hou, Z. Ning, L. Guo, and X. Zhang, “Temporal, Functional and Spatial Big Data Computing

Framework for Large-Scale Smart Grid," *IEEE Trans. Emerg. Top. Comput.*, vol. PP, no. 99, pp. 1–1, 2017.

[126] H. Guo, J. Liu, and L. Zhao, "Big Data Acquisition under Failures in FiWi Enhanced Smart Grid," *IEEE Trans. Emerg. Top. Comput.*, vol. PP, no. 99, pp. 1–1, 2017.

[127] L. Chu, R. C. Qiu, X. He, Z. Ling, and Y. Liu, "Massive Streaming PMU Data Modeling and Analytics in Smart Grid State Evaluation Based on Multiple High-Dimensional Covariance Tests," *IEEE Trans. Big Data*, vol. PP, no. 99, pp. 1–1, 2017.

[128] G. Capizzi, G. L. Sciuto, C. Napoli, and E. Tramontana, "Advanced and Adaptive Dispatch for Smart Grids by means of Predictive Models," *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[129] D. Zhou *et al.*, "Distributed Data Analytics Platform for Wide-Area Synchrophasor Measurement Systems," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2397–2405, Sep. 2016.

[130] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renew. Sustain. Energy Rev.*, vol. 56, pp. 215–225, Apr. 2016.

[131] "European Utility Giants Are on a Grid Edge Shopping Spree in 2017 | Greentech Media." [Online]. Available: https://www.greentechmedia.com/research/report/european-utility-giants-are-on-a-grid-edge-shopping-spree-in-2017#gs.z39TFYo. [Accessed: 11-Jan-2018].

[132] "After Two Years as an Oracle Company, What's Next for Opower? | Greentech Media." [Online]. Available: https://www.greentechmedia.com/articles/read/what-is-next-for-opower#gs.cxx566. [Accessed: 30-Apr-2018].

[133] J. S. John, "Tendril Acquires EnergySavvy to Enhance Utility Customer Data Analytics," 15-May-2019. [Online]. Available: https://www.greentechmedia.com/articles/read/tendril-acquires-energysavvy-adds-utility-customer-data-analytics-to-offeri. [Accessed: 20-May-2019].

[134] "Utopus Insights - about." [Online]. Available: https://www.utopusinsights.com/about. [Accessed: 22-Feb-2019].

[135] "» Data Services Navigant Research." [Online]. Available: https://www.navigantresearch.com/data-services. [Accessed: 26-Jan-2018].

[136] "2018 Magic Quadrant for Analytics and Business Intelligence Platforms," *Not all business intelligence platforms are created equal.*, Feb-2018. [Online]. Available: http://go.qlik.com/GMQ-2018.html. [Accessed: 28-Feb-2018].

[137] "Agder Energi - About us." [Online]. Available: http://www.ae.no/konsernet/om/english/. [Accessed: 21-May-2018].

[138] Microsoft, *Microsoft, Agder Energi partner to show how smarter grid can help Norway become*

more sustainable. .

[139] "C&J Energy | MapR." [Online]. Available: https://mapr.com/customers/candj-energy/. [Accessed: 21-May-2019].

[140] © 2019 Cloudera, I. A. rights reserved Terms, C. | P. Policy, D. P. A. Hadoop, associated open source project names are trademarks of the A. S. F. F. a complete list of trademarks, and C. Here, "Centrica | Customer Success," *Cloudera*. [Online]. Available: https://www.cloudera.com/content/www/en-us/about/customers/centrica.html. [Accessed: 22-May-2019].

[141] C. P. Affairs Government and Public, "Chevron Partners with Microsoft to Fuel Digital Transformation," *chevron.com*. [Online]. Available: https://www.chevron.com/stories/chevron-partners-with-microsoft. [Accessed: 21-Feb-2018].

[142] "E.ON develops highly-secure smart and efficient connected home solution in collaboration with Microsoft." [Online]. Available: https://www.eon.com/en/about-us/media/press-release/2018/eon-develops-highly-secure-smart-and-efficient-connected-home-solution-in-collaboration-with-microsoft.html. [Accessed: 21-Dec-2018].

[143] "Predix Platform | GE Digital." [Online]. Available: https://www.ge.com/digital/iiot-platform. [Accessed: 23-Jan-2019].

[144] "Cloud-Mining for Big Data With GE's Predix." [Online]. Available: https://www.exeloncorp.com/sustainability/interactive-csr/2017/innovation/predix-cloud-mining. [Accessed: 12-Jun-2018].

[145] © 2019 Cloudera, I. A. rights reserved Terms, C. | P. Policy, D. P. A. Hadoop, associated open source project names are trademarks of the A. S. F. F. a complete list of trademarks, and C. Here, "Energy & Utilities Data Solutions," *Cloudera*. [Online]. Available: https://www.cloudera.com/solutions/energy-and-utilities.html. [Accessed: 22-May-2019].

[146] "Microsoft and Schneider Electric Co-innovation." [Online]. Available: //www.schneider-electric.com/en/partners/alliances/microsoft.jsp. [Accessed: 11-Nov-2018].

[147] "Schneider Electric Success Story - Progress." [Online]. Available: https://www.progress.com/customers/schneider-electric-success. [Accessed: 28-Nov-2018].

[148] "Southwest Power Pool - Informatica Customer Success Story | Informatica US." [Online]. Available: https://www.informatica.com/about-us/customers/customer-success-stories/southwest-power-pool.html#fbid=7r2HjSELD6E. [Accessed: 22-May-2018].

[149] "Vector Invests in World Leading Internet of Energy Software." [Online]. Available: https://www.vector.co.nz/news/media-release-vector-invests-in-world-leading-int. [Accessed:

23-May-2019].

[150] V.- www.vestas.com, "Vestas - Scipher." [Online]. Available: https://www.vestas.com/en/services/utopus insights. [Accessed: 22-Feb-2019].

[151] "Enel X: technologies and innovation for smart electricity solutions | Enel X." [Online]. Available: /n-a/en.html. [Accessed: 21-May-2018].

[152] "eSmart Systems Platform." [Online]. Available: https://www.esmartsystems.com/products/platform/. [Accessed: 21-Feb-2018].

[153] "GE and Microsoft partner to bring Predix to Azure, accelerating digital transformation for industrial customers," *Stories*, 11-Jul-2016. [Online]. Available: https://news.microsoft.com/2016/07/11/ge-and-microsoft-partner-to-bring-predix-to-azure-accelerating-digital-transformation-for-industrial-customers/. [Accessed: 23-Jan-2019].

[154] "mPrest -Technology." [Online]. Available: https://www.mprest.com/technology. [Accessed: 23-Feb-2019].

[155] "Cloudera Enterprise Data Hub on Oracle Cloud Infrastructure." [Online]. Available: https://cloud.oracle.com/en_US/iaas/cloudera. [Accessed: 22-May-2019].

[156] "Southern Company – Informatica Customer Success Story | Informatica US." [Online]. Available: https://www.informatica.com/about-us/customers/customer-success-stories/southern-company.html#fbid=7r2HjSELD6E. [Accessed: 22-May-2018].

[157] "European Commission : CORDIS : Projects and Results : Home." [Online]. Available: https://cordis.europa.eu/projects. [Accessed: 25-May-2019].

[158] "Novel, automated charging infrastructure for electric vehicles | Projects | H2020," *CORDIS | European Commission*. [Online]. Available: https://cordis.europa.eu/project/rcn/211244/brief/en. [Accessed: 26-Jan-2019].

[159] eydnema, "EU-funded projects on data," *Digital Single Market - European Commission*, 22-Sep-2014. [Online]. Available: https://ec.europa.eu/digital-single-market/en/programme-and-projects/project-factsheets-data. [Accessed: 25-May-2019].

[160] BDVA, "Big Data Value cPPP Monitoring Report 2017," 2018.

[161] BDVA, "BDVA Strategic Research and Innovation Agenda v4 (BDVA SRIA v4)," Jan. 2017.

[162] BDVA, "Future Challenges for European Leadership in the Global Data Economy and Data-Driven Society: Input to Framework Programme 9," Mar. 2018.

[163] E. Vinagre, L. Gomes, and Z. A. Vale, "Electrical Energy Consumption Forecast Using External Facility Data," in *IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town,*

South Africa, December 7-10, 2015, 2015, pp. 659–664.

[164] L. Gomes, M. Lefrançois, P. Faria, and Z. Vale, "Publishing real-time microgrid consumption data on the web of Linked Data," in *2016 Clemson University Power Systems Conference (PSC)*, 2016, pp. 1–8.

[165] "IEEE PES Intelligent Systems Subcommittee." [Online]. Available: http://sites.ieee.org/pes-iss/. [Accessed: 02-Jun-2019].

[166] R. Estrada and I. Ruiz, *Big Data SMACK*. Berkeley, CA: Apress, 2016.

[167] T. Kraska, A. Talwalkar, and J. Duchi, "Mlbase: A distributed machine-learning system," in *In CIDR*, 2013.

[168] "Deeplearning4j." [Online]. Available: https://deeplearning4j.org/. [Accessed: 02-Feb-2018].

[169] "SparkR (R on Spark) - Spark 2.4.3 Documentation." [Online]. Available: https://spark.apache.org/docs/latest/sparkr.html. [Accessed: 02-Feb-2018].

[170] "RStudio," *RStudio*. [Online]. Available: https://www.rstudio.com/. [Accessed: 02-Feb-2018].

[171] "SparkR vs sparklyr today," *RStudio Community*, 20-Nov-2017. [Online]. Available: https://community.rstudio.com/t/sparkr-vs-sparklyr-today/18532. [Accessed: 21-Nov-2017].

[172] "MLlib | Apache Spark." [Online]. Available: http://spark.apache.org/mllib/. [Accessed: 02-Feb-2018].

[173] "Zeppelin." [Online]. Available: https://zeppelin.apache.org/. [Accessed: 02-Feb-2018].

[174] "Spark SQL & DataFrames | Apache Spark." [Online]. Available: http://spark.apache.org/sql/. [Accessed: 02-Feb-2018].

[175] A. Chebotko, A. Kashlev, and S. Lu, "A Big Data Modeling Methodology for Apache Cassandra," in *2015 IEEE International Congress on Big Data*, 2015, pp. 238–245.

[176] "Apache Mesos," *Apache Mesos*. [Online]. Available: http://mesos.apache.org/. [Accessed: 02-Feb-2018].

[177] E. Vinagre, T. Pinto, G. Pinheiro, Z. Vale, C. Ramos, and J. M. Corchado, "Knowledge Management System For Big Data In A Smart Electricity Grid Context," 2018.

[178] G. Pinheiro, E. Vinagre, I. Praça, Z. Vale, and C. Ramos, "Smart Grids Data Management: A Case for Cassandra," in *International Symposium on Distributed Computing and Artificial Intelligence*, 2017, pp. 87–95.

[179] Ciara Byrne, *Fast Data Use Cases for Telecommunications: How Fast Data Can Help Telcos Virtualize, Monetize, and Deal with the Data Deluge*, First Edition. Gravenstein Highway North, Sebastopol, CA: O'Reilly Media, Inc., 2017.

[180] “Docker,” *Docker*. [Online]. Available: https://www.docker.com/. [Accessed: 28-Jul-2017].

[181] V. Kohli, J. Wooten, and R. Dua, *Troubleshooting docker: strategically design, troubleshoot, and automate Docker containers from development to deployment*. Birmingham Mumbai: Packt, 2017.

[182] “dotCloud, Inc. is Becoming Docker, Inc.,” *Docker Blog*, 29-Oct-2013. [Online]. Available: https://blog.docker.com/2013/10/dotcloud-is-becoming-docker-inc/. [Accessed: 24-Jun-2017].

[183] “Open Containers Initiative: About,” *Open Containers Initiative*. .

[184] “Announcing Docker Enterprise Edition,” *Docker Blog*, 02-Mar-2017. [Online]. Available: https://blog.docker.com/2017/03/docker-enterprise-edition/. [Accessed: 04-Jul-2017].

[185] “About Docker CE,” *Docker Documentation*, 31-May-2019. [Online]. Available: https://docs.docker.com/install/. [Accessed: 04-Jul-2017].

[186] “Overview of Docker editions,” *Docker Documentation*, 31-May-2017. [Online]. Available: https://docs.docker.com/install/overview/. [Accessed: 04-Jul-2017].

[187] “Moby.” [Online]. Available: https://mobyproject.org/. [Accessed: 08-Jul-2017].

[188] “About Docker Engine,” *Docker Documentation*, 04-May-2017. [Online]. Available: https://docs.docker.com/engine/docker-overview/. [Accessed: 24-Jul-2017].

[189] “Docker Machine Overview,” *Docker Documentation*, 31-May-2019. [Online]. Available: https://docs.docker.com/machine/overview/. [Accessed: 28-Jul-2017].

[190] “What is a Container?,” *Docker*. [Online]. Available: https://www.docker.com/resources/what-container. [Accessed: 04-Jul-2017].

[191] “Get started with Docker for Windows,” *Docker Documentation*, 31-May-2019. [Online]. Available: https://docs.docker.com/docker-for-windows/. [Accessed: 04-Jun-2019].

[192] “Docker Desktop for Mac vs. Docker Toolbox,” *Docker Documentation*, 31-May-2017. [Online]. Available: https://docs.docker.com/docker-for-mac/docker-toolbox/. [Accessed: 28-Jul-2017].

[193] “Swarm mode overview,” *Docker Documentation*, 05-Jun-2017. [Online]. Available: https://docs.docker.com/engine/swarm/. [Accessed: 25-Jul-2017].

[194] D. Ongaro and J. Ousterhout, “In Search of an Understandable Consensus Algorithm,” p. 18.

[195] “Open source software for creating private and public clouds.,” *OpenStack*. [Online]. Available: https://www.openstack.org/. [Accessed: 28-Nov-2017].

[196] “Apache NiFi.” [Online]. Available: https://nifi.apache.org/. [Accessed: 26-Nov-2017].

[197] “Big Data Europe » Empowering Communities with Data Technologies.” .

[198] “docker-compose bind mount docker.sock not a valid Windows path · Issue #1829 · docker/for-win,” *GitHub*. [Online]. Available: https://github.com/docker/for-win/issues/1829. [Accessed: 07-May-2018].

[199] “Docker build does not work on Windows · Issue #1341 · containous/traefik,” *GitHub*. [Online]. Available: https://github.com/containous/traefik/issues/1341. [Accessed: 07-May-2018].

[200] “Can’t set Docker Volume for Container in Windows Docker CE,” *Docker Forums*, 04-May-2017. [Online]. Available: https://forums.docker.com/t/cant-set-docker-volume-for-container-in-windows-docker-ce/31841. [Accessed: 07-May-2018].

[201] “AccuWeather APIs | home.” [Online]. Available: https://developer.accuweather.com/. [Accessed: 22-Sep-2018].

[202] “Weather API - OpenWeatherMap.” [Online]. Available: https://openweathermap.org/api. [Accessed: 22-Sep-2018].

[203] “Telegraf 1.9 documentation | InfluxData Documentation.” [Online]. Available: https://docs.influxdata.com/. [Accessed: 11-Sep-2018].

[204] “telegraf - Docker Hub.” [Online]. Available: https://hub.docker.com/_/telegraf. [Accessed: 12-Nov-2018].

[205] “Eclipse Mosquitto,” *Eclipse Mosquitto*, 08-Jan-2018. [Online]. Available: https://mosquitto.org/. [Accessed: 11-Sep-2018].

[206] “eclipse-mosquitto - Docker Hub.” [Online]. Available: https://hub.docker.com/_/eclipse-mosquitto. [Accessed: 12-Nov-2018].

[207] “Flux 0.7 Technical Preview | Blog | InfluxData.” [Online]. Available: https://www.influxdata.com/blog/flux-0-7-technical-preview/. [Accessed: 11-Dec-2018].

[208] “InfluxDB key concepts | InfluxData Documentation.” [Online]. Available: https://docs.influxdata.com/. [Accessed: 11-Jun-2018].

[209] S. N. Zehra, “Time Series Databases and InfluxDB,” p. 45, 2018.

[210] “influxdb - Docker Hub.” [Online]. Available: https://hub.docker.com/_/influxdb. [Accessed: 12-Nov-2018].

[211] “COPY | Npgsql Documentation.” [Online]. Available: https://www.npgsql.org/doc/copy.html. [Accessed: 28-Oct-2018].

[212] “PostgreSQL: Documentation: 9.1: NOTIFY.” [Online]. Available: https://www.postgresql.org/docs/9.1/sql-notify.html. [Accessed: 21-Oct-2018].

[213] “Documentation | Npgsql Documentation.” [Online]. Available:

https://www.npgsql.org/doc/index.html. [Accessed: 12-Oct-2018].

[214] “Time-series data: Why (and how) to use a relational database instead of NoSQL.” [Online]. Available: https://blog.timescale.com/time-series-data-why-and-how-to-use-a-relational-database-instead-of-nosql-d0cd6975e87c/#2362. [Accessed: 11-Nov-2018].

[215] “TimescaleDB Docs | Architecture.” [Online]. Available: https://docs.timescale.com/v1.3/introduction/architecture. [Accessed: 10-Nov-2018].

[216] “timescale/timescaledb - Docker Hub.” [Online]. Available: https://hub.docker.com/r/timescale/timescaledb. [Accessed: 05-Jan-2019].

[217] “Use volumes,” *Docker Documentation*. [Online]. Available: https://docs.docker.com/storage/volumes/. [Accessed: 12-Feb-2019].

[218] “Microsoft.ML 0.11.0.” [Online]. Available: https://www.nuget.org/packages/Microsoft.ML/. [Accessed: 20-Feb-2019].

[219] natke, “Machine learning tasks - ML.NET.” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/tasks. [Accessed: 20-Feb-2019].

[220] natke, “Data transformations - ML.NET.” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/transforms. [Accessed: 21-Feb-2019].

[221] sfilipi, “RegressionMetrics Class (Microsoft.ML.Data).” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.data.regressionmetrics. [Accessed: 22-Feb-2019].

[222] “Contribute - docker-images.” [Online]. Available: https://beam.apache.org/contribute/docker-images/. [Accessed: 04-Mar-2019].

[223] “Built-in I/O Transforms.” [Online]. Available: https://beam.apache.org/documentation/io/built-in/. [Accessed: 04-Mar-2019].

[224] “Chronograf 1.7 documentation | InfluxData Documentation.” [Online]. Available: https://docs.influxdata.com/. [Accessed: 13-Sep-2018].

[225] “Graph Panel,” *Grafana Labs Blog*. [Online]. Available: https://grafana.com/docs/features/panels/graph/. [Accessed: 18-Dec-2018].

[226] “grafana/grafana - Docker Hub.” [Online]. Available: https://hub.docker.com/r/grafana/grafana/. [Accessed: 18-Dec-2018].

[227] “Mean absolute percentage error,” *Wikipedia*. 12-Jun-2019.

[228] “Mean absolute error,” *Wikipedia*. 23-May-2019.

[229] “Mean squared error,” *Wikipedia*. 04-May-2019.

[230] “Root-mean-square deviation,” *Wikipedia*. 19-Apr-2019.

[231] “Coefficient of determination,” *Wikipedia*. 12-Jun-2019.

[232] sfilipi, “OlsTrainer Class (Microsoft.ML.Trainers).” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.trainers.olstrainer. [Accessed: 01-Jul-2019].

[233] sfilipi, “LightGbmRegressionTrainer(IDataView, IDataView) Class (Microsoft.ML.Trainers.LightGbm).” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/api/microsoft.ml.trainers.lightgbm.lightgbmregressiontrainer. [Accessed: 01-Jul-2019].

[234] “continuumio/anaconda - Docker Hub.” [Online]. Available: https://hub.docker.com/r/continuumio/anaconda/. [Accessed: 01-Jul-2019].

[235] “Welcome | Flask (A Python Microframework).” [Online]. Available: http://flask.pocoo.org/. [Accessed: 01-Jul-2019].

[236] “FloydHub - Deep Learning Platform - Cloud GPU.” [Online]. Available: https://www.floydhub.com/. [Accessed: 01-Jul-2019].

Anexo 1 Projetos Big Data financiados pela União Europeia

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
643937	01/02/2015 31/01/2018	EDSA	European Data Science Academy	Data explosion on the web, fuelled by social networking, micro-blogging, as well as crowdsourcing, has led to the Big Data phenomenon. This is characterized by increasing volumes of structured, semi-structured and unstructured data, originating from sources that generate them...	29 76 189,75 29 76 189,75
644497	01/03/2015 31/08/2017	proDataMarket	Enabling the property Data Marketplace for Novel Data-driven Business Models	Property data are one of the most valuable datasets managed by governments worldwide and extensively used in various domains by private and public organizations. Unfortunately these data are not always easy to access. House and property data is used in a variety of ways to...	4 482 957,00 3 499 144,50
644564	01/01/2015 31/12/2017	BigDataEurope	Integrating Big Data, Software and Communities for Addressing Europe's Societal Challenges	BigDataEurope will provide support mechanisms for all the major aspects of a data value chain, in terms of the employed data and technology assets, the participating roles and the established or evolving processes. The effectiveness of the provided support mechanisms will be...	4 984 238,75 4 984 238,75
644632	01/04/2015 31/03/2017	MixedEmotions	Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets	MixedEmotions will develop innovative multilingual multi-modal Big Data analytics applications that will analyze a more complete emotional profile of user behavior using data from mixed input channels: multilingual text data sources, A/V signal input (multilingual speech...	3 529 617,14 3 036 910,00
644657	01/04/2015 31/03/2018	AutoMat	Automotive Big Data Marketplace for Innovative Cross-sectorial Vehicle Data Services	Inside today's vehicles ~4000 CAN-Bus signals/sec are processed in comparison to very few signals in smart phones and alike. This large amount of continuously gathered vehicle data represents major big data business potentials, not only for the automotive industry but in...	5 751 785,50 4 460 269,00
644683	01/02/2015 31/07/2017	ODINE	Open Data INcubator for Europe	The Open Data INcubator for Europe (ODINE) project will set up an environment to support and advice SMEs and start-ups in creating commercial added value from open data. Drawing on the experience from key players in the consortium including Wayra (an incubator/accelerator) ...	8 136 115,44 882 980,50
644715	02/02/2015 01/02/2017	AquaSmart	Aquaculture Smart and Open Data Analytics as a Service	AQUASMART's objective is to enhance innovation capacity to the aquaculture sector, by addressing the problem of global knowledge access and data exchanges between aquaculture companies and its related stakeholders. Offering aquaculture production companies the tools to...	3 109 077,50 2 717 432,37

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
644753	01/02/2015 31/07/2017	KConnect	Khresmoi Multilingual Medical Text Analysis, Search and Machine Translation Connected in a Thriving Data-Value Chain	The overall objective of the KConnect project is to create a medical text Data-Value Chain with a critical mass of participating companies using cutting-edge commercial cloud-based services for multilingual Semantic Annotation, Semantic Search and Machine Translation of...	3 889 842,50 3 083 083,00
644771	01/02/201531/01/2017	FREME	Open Framework of E-Services for Multilingual and Semantic Enrichment of Digital Content	The aim of the FREME innovative action is to establish an “Open Framework of E-Services for Multilingual and Semantic Enrichment of Digital Content”. Six enrichment services will be designed, piloted, and validated during the action. Their innovation, usability, and...	3 606 751,44 3 212 626,00
644906	01/03/2015 31/08/2018	AEGLE	AEGLE (Ancient Greek: Αἴγλη) – An analytics framework for integrated and personalized healthcare services in Europe	The data generated in the health domain is coming from heterogeneous, multi-modal, multi-lingual, dynamic and fast evolving medical technologies. Today we are found in a big health landscape characterized by large volume, versatility and velocity (3Vs) which has led to the...	6 079 642,50 5 230 698,75
645012	01/03/2015 28/02/2018	KRISTINA	Knowledge-Based Information Agent with Social Competence and Human Interaction Capabilities	In Europe, migration is tradition – and not only since the European legislation changed towards free migration of European citizens. This is not free of challenges. Especially in the case of care, migrants, often face a double challenge: (i) not to speak the language and not...	3 633 801,25 3 633 801,25
645244	01/02/2015 31/01/2018	EuDEco	Modelling the European data economy	The Modelling the European Data Economy (EuDEco) project will assist European sci-ence and industry in understanding and exploiting the potentials of data reuse in the context of Big and Open Data big data. The aim isto establish a self sustaining data market and thereby...	2 276 625,00 2 276 625,00
645323	01/01/2015 31/12/2017	BISON	BIg Speech data analytics for cONtact centres	Contact centers (CC) are an important business for Europe: 35,000 contact centers generate 3.2 Million jobs (~1% of Europe’s active population). A typical CC produces a wealth of multilingual spoken data that is nowadays mined by humans (CC agents and supervisors) or by...	4 097 952,50 3 090 824,50
645378	01/01/2015 31/12/2017	ARIA-VALUSPA	Artificial Retrieval of Information Assistants - Virtual Agents with Linguistic Understanding, Social skills, and Personalised Aspects	The ARIA-VALUSPA project will create a ground-breaking new framework that will allow easy creation of Artificial Retrieval of Information Assistants (ARIAs) that are capable of holding multi-modal social interactions in challenging and unexpected situations. The system can ...	2 949 318,76 2 949 317,78

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
645425	01/03/2015 28/02/2018	SSIX	Social Sentiment analysis financial IndeXes	Social Sentiment Indices powered by X-Scores (SSIX) aims to provide European SMEs with a collection of easy to interpret tools to analyse and understand social media users attitudes for any given subject; these sentiment characteristics can be exploited to help SMEs to operate...	4 259 125,50 3 315 963,00
687591	01/01/2016 31/12/2018	datACRON	Big Data Analytics for Time Critical Mobility Forecasting	datACRON is a research and innovation collaborative project introducing novel methods for threat and abnormalactivity detection in very large fleets of moving entities spread across large geographical areas.Specifically, datACRON aims to develop novel methods for real-time...	3 993 835,00 3 993 835,00
687691	01/12/2015 30/11/2018	PROTEUS	Scalable online machine learning for predictive analytics and real-time interactive visualization	PROTEUS mission is to investigate and develop ready-to-use scalable online machine learning algorithms and interactive visualization techniques for real-time predictive analytics to deal with extremely large data sets and data streams. The developed algorithms and techniques...	3 156 525,00 3 156 525,00
688082	01/02/201631/01/2019	SETA	SETA: An open, sustainable, ubiquitous data and service ecosystem for efficient, effective, safe, resilient mobility in metropolitan areas	Around 50% of the global population lives in metropolitan areas, and this is expected to grow to 75% by 2050. Mobility within these areas is complex as it involves multiple modalities of transport, multiple managing authorities, as well as several millions of citizens. The...	5 565 247,50 5 565 247,50
688099	01/01/2016 31/12/2018	Cloud-LSVA	Cloud Large Scale Video Analysis	Cloud-LSVA will create Big Data Technologies to address the open problem of a lack of software tools, and hardware platforms, to annotate petabyte scale video datasets. The problem is of particular importance to the automotive industry. CMOS Image Sensors for Vehicles are the...	4 604 431,25 4 604 431,25
688139	01/02/2016 31/01/2019	SUMMA	Scalable Understanding of Multilingual Media	Media monitoring enables the global news media to be viewed in terms of emerging trends, people in the news, and the evolution of story-lines. The massive growth in the number of broadcast and Internet media channels means that current approaches can no longer cope with the...	7 963 951,25 6 193 361,25
688191	01/12/2015 30/11/2018	STREAMLINE	STREAMLINE	STREAMLINE will address the competitive advantage needs of European online media businesses (EOMB) by delivering fast reactive analytics suitable in solving a wide array of problems, including addressing customer retention, personalised recommendation, and more broadly...	3 291 294,99 3 291 294,99
688227	01/12/2015 30/11/2018	HOBBIT	Holistic Benchmarking of Big Linked Data	Linked Data has gained significant momentum over the last years. It is now used at industrial scale in many sectors in which an increasingly large amount of rapidly changing data needs to be processed. HOBBIT is an ambitious project that aims to push the development of Big...	3 798 453,75 3 718 250,00

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
688356	01/01/2016 31/03/2018	SEE.4C	SpatiotEmporal ForEcasting: Coopetition to meet Current Cross-modal Challenges	Fast, accurate forecasting of spatiotemporal data is needed in critical industrial domains such as energy (prediction of spatiotemporal patterns in renewable generation, usage and traffic) as well as in public policy. The task is so challenging in scale and scope however as...	760 806,25 760 806,25
688380	01/12/2015 30/11/2018	VaVeL	Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors	Urban environments are awash with data from fixed and mobile sensors and monitoring infrastructures from public, private, or industry sources. Making such data useful would enable developing novel big data applications to benefit the citizens of Europe in areas such as...	3 999 668,75 3 999 668,75
688797	01/01/2016 30/11/2018	TOREADOR	TrustwOrthy model-awaRE Analytics Data platfORm	The TOREADOR project is aimed at overcoming some major hurdles that until now have prevented many European companies from reaping the full benefits of Big Data analytics (BDA). Companies and organisations in Europe have become aware of the potential competitive advantage they...	6 311 218,75 6 311 218,75
731581	01/01/2017 31/12/2019	SLIPO	Scalable Linking and Integration of Big POI data	POIs are the content of any application, service, and product even remotely related to our physical surroundings. From navigation applications, to social networks, to tourism, and logistics, we use POIs to search, communicate, decide, and plan our actions. The Big Data assets...	3 087 000,00 2 635 500,00
731583	01/01/2017 31/12/2019	SODA	Scalable Oblivious Data Analytics	More and more data is being generated, and analyzing this data drives knowledge and value creation across society. Unlocking this potential requires sharing of (often personal) data between organizations, but this meets unwillingness from data subjects and data controllers...	2 980 610,00 2 980 609,75
731601	01/01/2017 31/12/2019	SPECIAL	Scalable Policy-awarE linked data arChitecture for prIvacy, trAnsparency and compLiance	The SPECIAL project will address the contradiction between Big Data innovation and privacy-aware data protection by proposing a technical solution that makes both of these goals realistic. We will develop technology that: (i) supports the acquisition of user consent at...	3 991 388,75 3 991 388,75
731873	01/01/2017 31/12/2019	e-Sides	Ethical and Societal Implications of Data Sciences	Data-driven innovation is deeply transforming society and the economy. Although there are potentially enormous economic and social benefits this innovation also brings new challenges for individual and collective privacy, security, as well as democracy and participation. The...	999 940,00 999 940,00

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
731932	01/01/2017 31/07/2019	TT	Transforming Transport	Big Data will have a profound economic and societal impact in the mobility and logistics sector, which is one of the most-used industries in the world contributing to approximately 15% of GDP. Big Data is expected to lead to 500 billion USD in value worldwide in the form of...	18 703 369,39 14 631 935,45
732003	01/01/2017 30/06/2019	euBusinessGraph	Enabling the European Business Graph for Innovative Data Products and Services	Corporate information, including basic company firmographics (e.g., name(s), incorporation data, registered addresses, ownership and related entities), financials (e.g., balance sheets, ratings) as well as contextual data (e.g., cadastral data on corporate properties, geo...	3 601 500,00 3 099 900,00
732064	01/01/2017 31/12/2019	DataBio	Data-Driven Bioeconomy	The data intensive target sector selected for the DataBio project is the Data-Driven Bioeconomy, focusing in production of best possible raw materials from agriculture, forestry and fishery/aquaculture for the bioeconomy industry to produce food, energy and biomaterials taking...	16 164 596,28 12 580 486,16
732189	01/01/2017 30/06/2019	AEGIS	Advanced Big Data Value Chain for Public Safety and Personal Security	AEGIS, brings together the data, the network & the technologies to create a curated, semantically enhanced, interlinked & multilingual repository for public & personal safety-related big data. It delivers a data-driven innovation that expands over multiple business sectors &...	3 927 599,95 2 999 200,00
732194	01/12/2016 30/11/2019	QROWD	QROWD - Because Big Data Integration is Humanly Possible	Big Data integration in European cities is of utmost importance for municipalities and companies to offer effective information services, enable efficient data-driven transportation and mobility, reduce CO2 emissions, assess the efficiency of infrastructure, as well as enhance...	3 993 505,00 2 969 367,50
732310	01/01/2017 30/06/2019	BigDataOcean	BigDataOcean - Exploiting Ocean's of Data for Maritime Applications	The main objective of BigDataOcean is to enable maritime big data scenarios for EU-based companies, organisations and scientists, through a multi-segment platform that will combine data of different velocity, variety and volume under an inter-linked, trusted, multilingual...	3 566 172,50 2 998 569,50
732328	01/01/2017 31/12/2019	FashionBrain	Understanding Europe’s Fashion Data Universe	The primary goal of each retailer is to “understand your customers”. Our interviews with retailers show a primary demand from the retail industry for predicting a customer's next demand. Surprisingly , even a complete record of past purchases (and returns) is not...	3 087 000,00 2 635 500,00
732340	02/01/17	K-PLEX	Knowledge Complexity	One of the major terminological forces driving ICT development today is that of ‘big data.’ While the phrase may sound inclusive and integrative, in fact, ‘big data’ approaches are highly selective, excluding any input that cannot be effectively structured, represented, or, indeed, digitised.	499 978,75 499 977,00

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
732506	01/01/2017 31/12/2019	Data Pitch	Accelerating data to market	Information technology has driven, directly or indirectly, much of Europe's economic growth during the last decades as the role of data transitioned from the support of business decisions to becoming a good in itself. An open approach towards data value creation has become...	7 067 230,00 6 994 105,00
732590	01/01/2017 31/12/2019	EW-Shopp	EW-Shopp - Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey	In this project we aim at supporting companies operating in the fragmented European ecosystem of the eCommerce, Retail and Marketing industries to increase their efficiency and competitiveness by leveraging deep customer insights that are too challenging for them to obtain...	3 566 250,00 2 902 875,00
732630	01/01/2017 31/12/2020	BDVe	Big Data Value ecosystem	The mission of BDVe is to support the Big Data Value PPP in realizing a vibrant data-driven EU economy or said in other words, BDVe will support the implementation of the PPP to be a SUCCESS.Behind that mission, there are multiple goals to achieve, which should be taken into...	4 940 286,25 4 940 286,25
732907	01/11/2016 31/10/2019	MH-MD	My Health - My Data	Issues of data subjects' privacy and data security represent a crucial challenge in the biomedical sector more than in other industries. The current IT landscape in this field shows a myriad of isolated, locally hosted patient data repositories, managed by clinical centres...	3 982 440,00 3 456 188,50
779747	01/01/2018 31/12/2020	BigDataStack	High-performance data-centric stack for big data applications and operations	The new data-driven industrial revolution highlights the need for big data technologies to unlock the potential in various application domains. To this end, BigDataStack delivers a complete high-performant stack of technologies addressing the emerging needs of data operations...	2 240 988,75 2 240 988,75
779780	01/01/2018 31/12/2020	BodyPass	API-ecosystem for cross-sectorial exchange of 3D personal data	BodyPass aims to break barriers between health sector and consumer goods sector and eliminate the current data silos. The main objective of BodyPass is to foster exchange, linking and re-use, as well as to integrate 3D data assets from the two sectors. For this, BodyPass has...	3 110 581,25 2 552 018,38
779790	01/01/2018 30/06/2021	EDI	European Data Incubator	Only 2 out of the top 20 companies changing lives and making money out of Big Data are European. Europe is not playing the role it should in the global market. European Data Incubator project aims to revert this situation. Every company is born as a startup, and Big Data has...	2 894 448,75 1 699 323,75

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
780167	01/12/2017 30/11/2020	Cross-CPP	Ecosystem for Services based on integrated Cross-sectorial Data Streams from multiple Cyber Physical Products and Open Data Sources	The objective is to establish an IT environment for the integration and analytics of data streams coming from high volume (mass) products with cyber physical features, as well from Open Data Sources, aiming to offer new cross sectorial services and focusing on the commercial...	7 708 441,51 7 142 856,76
780245	01/01/2018 31/12/2020	E2DATA	European Extreme Performing Big Data Stacks	Imagine a Big Data application with the following characteristics: (i) it has to process large amounts of complex streaming data, (ii) the application logic that processes the incoming data must execute and complete within a strict time limit, and (iii) there is a limited...	4 676 250,00 4 676 250,00
780247	01/01/2018 31/12/2020	TheyBuyForYou	Enabling procurement data value chains for economic development, demand management, competitive markets and vendor intelligence	Procurement affects virtually every organisation. Public spending alone will soon exceed €2trn per annum in the EU. In times of slow economic recovery and enhanced transparency, there is a pressing need for better management of government finances. The interaction between...	3 274 440,00 2 925 693,00
780251	01/01/2018 31/12/2020	TYPHON	Polyglot and Hybrid Persistence Architectures for Big Data Analytics	The need for levels of availability and scalability beyond those supported by relational databases has led to the emergence of a new generation of purpose-specific databases grouped under the term NoSQL. In general, NoSQL databases are designed with horizontal scalability as a...	4 499 447,50 4 499 447,50
780355	01/01/2018 31/12/2020	FANDANGO	FAke News discovery and propagation from big Data ANalysis and artificial intelliGence Operations	Fake News are now a hot issue in Europe as well as worldwide, particularly referred to Political and Social Challenges that reflect in business as well as in industry. Europe is lacking of a systematic knowledge and data transfer across organizations to address the aggressive...	3 583 125,00 2 879 250,00
780495	01/01/2018 28/02/2021	BigMedilytics	Big Data for Medical Analytics	There are three main reasons for an immediate innovation action to apply big data technologies in Healthcare. Firstly, a Healthy nation is a Wealthy nation! An improvement in health leads to economic growth through long-term gains in human and physical capital, which...	16 940 837,50 14 997 306,25
780602	01/12/2017 30/11/2020	Lynx	Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe	European small and medium-sized enterprises (SMEs) and companies operating internationally face multiple difficulties to trade abroad and to localise their products and services in other countries. Complying with regulatory and legal aspects is crucial for these companies, who...	3 638 065,00 2 959 247,52
780622	01/01/2018 31/12/2020	CLASS	Edge and CLoud Computation: A Highly Distributed Software Architecture for Big Data AnalyticS	Big data applications processing extreme amounts of complex data are nowadays being integrated with even more challenging requirements such as the need of continuously processing vast amount of information in real-time.Current data analytics systems are usually designed...	3 900 802,50 3 900 802,50

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
780732	01/01/2018 31/12/2020	BOOST 4.0	Big Data Value Spaces for COmpetitiveness of European COnnected Smart FacTories 4.0	EFFRA recommendations on Factories 4.0 and Beyond (Sept 2016) clearly stated the need for development of large scale experimentation and demonstration of data-driven “connected smart” Factories 4.0, to retain European manufacturing competitiveness. BOOST 4.0 will address...	18 843 440,00 14 983 516,26
780751	01/01/2018 31/12/2020	BigDataGrapes	Big Data to Enable Global Disruption of the Grapevine-powered Industries	Big data is becoming a hype that is going to completely redefine industries within very traditional sectors like agriculture, food and beauty. The emergence of niche big data companies like Enolytics (“bringing big data insights to the wine industry”) is threatening to...	4 441 500,00 4 441 500,00
780754	01/01/2018 31/12/2020	Track and Know	Big Data for Mobility Tracking Knowledge Extraction in Urban Areas	Track&Know will research, develop and exploit a new software framework that aims at increasing the efficiency of Big Data applications in the transport, mobility, motor insurance and health sectors. Stemming from industrial cases, Track&Know will develop user friendly...	4 848 013,75 4 848 013,75
780787	01/01/2018 31/12/2020	I-BiDaaS	Industrial-Driven Big Data as a Self-Service Solution	Organizations leverage data pools to drive value, while it is variety, not volume or velocity, which drives big-data investments. The convergence of IoT, cloud, and big data, create new opportunities for self-service analytics towards a completely paradigm towards big data...	4 997 035,00 4 997 035,00
780792	01/01/2018 31/12/2020	ICARUS	Aviation-driven Data Value Chain for Diversified Global and Local Operations	The European aviation industry needs to leverage the surge of multi-source and multi-lingual data streams to gain augmented intelligence on its status quo and open up a wide spectrum of unprecedented services for the whole ecosystem (airlines, airports, passengers, service...	3 505 978,75 2 836 490,52
780966	01/01/2018 31/12/2020	DataBench	Evidence Based Big Data Benchmarking to Improve Business Performance	Organisations rely on evidence from the Benchmarking domain to provide answers to how their processes are performing. There is extensive information on how and why to perform technical benchmarks for the specific management and analytics processes, but there is a lack of...	2 240 988,75 2 240 988,75
824988	01/12/2018 30/11/2021	MUSKETEER	Machine learning to augment shared knowledge in federated privacy-preserving scenarios	The massive increase in data collected and stored worldwide calls for new ways to preserve privacy while still allowing data sharing among multiple data owners. Today, the lack of trusted and secure environments for data sharing inhibits data economy while legality, privacy...	4 380 346,25 4 380 335,00

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
825014	01/01/2019 31/12/2021	Data Market Services	Supporting the European data market providing free support services to data-centric SMEs and start-ups	The EU data market has been analysed in the past years by several studies and reports. The estimate of the overall value of the data market in EU28 had a growth rate of a 9.5% between 2015 and 2016. Despite of this growing bottom-line market, there are some barriers: •Europe...	2 993 961,25 2 993 961,25
825041	01/01/2019 31/12/2021	SmartDataLake	Sustainable Data Lakes for Extreme-Scale Analytics	Data lakes are raw data ecosystems, where large amounts of diverse data are retained and coexist. They facilitate self-service analytics for flexible, fast, ad hoc decision making. SmartDataLake enables extreme-scale analytics over sustainable big data lakes. It provides an...	3 945 450,00 3 945 450,00
825070	01/01/2019 31/12/2021	INFORE	Interactive Extreme-Scale Analytics and Forecasting	At an increasing rate, industrial and scientific institutions need to deal with massive data flows streaming in from a multitude of sources. For instance, maritime surveillance applications combine high-velocity data streams, including vessel position signals emitted from...	4 435 586,25 4 435 586,25
825184	01/01/2019 31/12/2021	CloudButton	Serverless Data Analytics Platform	This project is inspired by the following sentence from a professor of computer graphics at UC Berkeley : “Why is there no cloud button?” He outlined how his students simply wish they could easily “push a button” and have their code – existing, optimized...	4 277 507,50 4 277 507,50
825225	01/12/2018 30/11/2021	Safe-DEED	Safe Data Enabled Economic Development	As privacy and trust remain key in the data sharing debate, Privacy enhancing technologies (PET) will play a prominent role by 2025. Safe-DEED takes a highly interdisciplinary approach, bringing together partners from cryptography, data science, business innovation, and legal...	2 996 400,00 2 996 400,00
825258	01/01/2019 31/12/2021	ExtremeEarth	From Copernicus Big Data to Extreme Earth Analytics	Copernicus is the European program for monitoring the Earth. The geospatial data produced by the Sentinel satellites puts Copernicus at the forefront of the Big Data paradigm, giving rise to all the relevant challenges: volume, velocity, variety, veracity and value...	5 988 301,25 5 988 301,25
825292	01/01/2019 31/12/2022	EXA MODE	EXtreme-scale Analytics via Multimodal Ontology Discovery & Enhancement	Exascale volumes of diverse data from distributed sources are continuously produced. Healthcare data stand out in the size produced (production 2020 >2000 exabytes), heterogeneity (many media, acquisition methods), included knowledge (e.g. diagnostic reports) and commercial...	4 333 281,25 4 333 281,25

ID/ URL	Start date / End date	Acronym	Title	Teaser	Budget (€) / Contribuição UE (€)
825333	01/01/2019 31/12/2021	MOSAICrOWN	Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and Owner control	The application of data analysis techniques over large data collections provides great benefits, from the personal, to the business, research, and social domain. The availability of large data collections recording actions and choices of individuals and organisations can lead...	3 203 750,00 3 203 750,00
825473	01/12/2018 30/11/2021	ELASTIC	A Software Architecture for Extreme-ScaLe Big-Data AnalyticS in Fog CompuTIng ECosystems	Big data is nowadays being integrated in systems requiring to process a vast amount of information from (geographically) distributed data sources, while fulfilling the non-functional properties (real-time, energy-efficiency, communication quality and security) inherited from...	5 920 581,25 5 920 581,25

Anexo 2 Data Sources: REST API GECAD

 # Zona1
 # GET: http://192.168.2.5:8520/resource/Analyzer1_V4


```
{
        "Analyzer1_V4": {
        "Ph3_U": 238,
        "Ph3_Cos": 100,
        "Ph2_U": 231,
        "Ph1_Cos": 14,
        "Ph3_I": 14,
        "Ph2_I": 44,
        "AC101_P": 10,
        "Ph1_I": 3,
        "Ph2_Cos": 85,
        "P_Total": 1200,
        "Ph1_U": 231,
        "Ph3_Q": 0,
        "Ph2_Q": -489,
        "Ph3_P": 350,
        "Ph1_Q": -59,
        "Ph2_P": 870,
        "Ph1_P": 10,
        "Q_Total": -549
      },
      "timestamp": "2018-09-13 17:54:19.302"
}
```


 # Zona2
 # GET: http://192.168.2.5:8520/resource/Analyzer2_V3


```
{
        "Analyzer2_V3": {
        "Ph3_U": 233,
        "Ph3_Cos": 100,
        "Ph2_U": 230,
        "Ph1_Cos": 91,
        "Ph3_I": 9,
        "Ph2_I": 7,
        "Ph1_I": 57,
        "Ph2_Cos": 100,
        "P_Total": 1620,
        "Ph1_U": 224,
        "Ph3_Q": 0,
        "Ph2_Q": 30,
        "Ph3_P": 220,
        "Ph1_Q": -519,
        "Ph2_P": 170,
        "Ph1_P": 1230,
        "Q_Total": -489
      },
      "timestamp": "2018-09-13 17:55:26.120"
}
```


 # Zona3
 # GET: http://192.168.2.5:8520/resource/Analyzer3_V3


```
{
```

```
        "Analyzer3_V3": {
        "Ph3_U": 228,
        "Ph3_Cos": 84,
        "Ph2_U": 232,
        "Ph1_Cos": 100,
        "Ph3_I": 26,
        "Ph2_I": 2,
        "Ph1_I": 0,
        "Ph2_Cos": 0,
        "P_Total": 560,
        "Ph1_U": 228,
        "Ph3_Q": -259,
        "Ph2_Q": -29,
        "Ph3_P": 560,
        "Ph1_Q": 0,
        "Ph2_P": 0,
        "Ph1_P": 0,
        "Q_Total": -289
      },
      "timestamp": "2018-09-13 17:56:16.467"
}
```


 # Zona4
 # GET: http://192.168.2.5:8520/resource/Analyzer4_V4


```
{
        "Analyzer4_V4": {
        "P1": 0,
        "P2": 0,
        "P3": 0,
        "PTotal": 0,
        "I1": 0,
        "I2": 0,
        "I3": 0,
        "U1": 235,
        "U2": 226,
        "U3": 229
      },
      "timestamp": "2018-09-13 17:57:49.873"
}
```


 # Zona7
 # GET: http://192.168.2.5:8520/resource/AnalyzerKitHall_V2


```
{
        "AnalyzerKitHall_V2": {
        "hallway_lights_reactivePower": 0,
        "dishwasher_control": 0,
        "hallway_lights_powerQuality": 0,
        "hallway_ac_current_x10": 0,
        "water_active": 0,
        "hallway_lights_current_x10": 0,
        "dishwasher_reactive": 0,
        "hallway_lights_activePower": 0,
        "hallway_ac_voltage": 0,
        "hallway_ac_activePower": 0,
        "kitchen_ac_current_x10": 0,
        "kettle_reactive": 0,
        "kitchen_ac_voltage": 0,
```

```
        "kitchen_ac_reactivePower": 0,
        "water_current_x10": 0,
        "kitchen_ac_activePower": 0,
        "dishwasher_current_x10": 0,
        "kettle_active": 0,
        "microwave_reactive": 0,
        "kitchen_ac_powerQuality": 0,
        "water_control": 0,
        "microwave_current_x10": 0,
        "kettle_current_x10": 0,
        "microwave_control": 0,
        "microwave_active": 0,
        "hallway_ac_powerQuality": 0,
        "kettle_voltage": 0,
        "kettle_quality": 0,
        "kettle_control": 0,
        "microwave_quality": 0,
        "hallway_lights_voltage": 0,
        "microwave_voltage": 0,
        "dishwasher_active": 0,
        "water_voltage": 0,
        "water_quality": 0,
        "hallway_ac_reactivePower": 0,
        "water_reactive": 0,
        "dishwasher_voltage": 0,
        "dishwasher_quality": 0
      },
      "timestamp": "2018-09-13 18:01:39.972"
}

# Zona8a
# GET: http://192.168.2.5:8520/resource/Analyzer115_V1
{
        "Analyzer116_V1": {
        "Curr3_A": 2,
        "PQuality3": 0,
        "Curr2_A": 0,
        "U3N_V": 224,
        "Curr1_A": 2,
        "PQuality1": 0,
        "PQuality2": 0,
        "U1N_V": 231,
        "U2N_V": 229,
        "P_Total_W": 921,
        "Reactive3_VA": -194,
        "Reactive1_VA": -121,
        "Reactive2_VA": -31,
        "Reactive_Total_VA": -349,
        "P1_W": 569,
        "P3_W": 352,
        "P2_W": 0
      },
      "timestamp": "2018-09-13 18:03:55.760"
}

# Zona8b
# GET: http://192.168.2.5:8520/resource/Analyzer116_V1
```

```
{
        "Analyzer115_V1": {
        "P_Total_W": 919,
        "U2N_Vx10": 2368,
        "Curr1_mA": 1575,
        "U1N_Vx10": 2238,
        "P1_W": 277,
        "Curr2_mA": 151,
        "Curr3_mA": 2810,
        "P3_W": 630,
        "P2_W": 12,
        "U3N_Vx10": 2243
       },
       "timestamp": "2018-09-13 18:04:41.142"
}
```


 # Zona9
 # GET: http://192.168.2.5:8520/resource/Inverter6_V3


```
{
       "Inverter6_V3": {
       "N6_U1N": 223,
       "N6_LifeEnerProd": 12062,
       "N6_U2N": 227,
       "N6_U3N": 234,
       "N6_Freq": 49,
       "N6_PF": 100,
       "N6_P": 1656
      },
     "timestamp": "2018-09-13 18:05:38.655"
}
```


 # ZonaBld
 # GET: http://192.168.2.5:8520/building/energy


```
{
     "generation": 1623,
     "hvac": 182,
     "DR_shift": 163.8,
     "DR_reduce": 830.1999999999999,
     "consumption": 3737,
     "sockets": 2348,
     "lights": 1186,
     "timestamp": "2018-09-13 18:07:36.106"
}
```


 # Sensor Temperatura (Outdoor)
 # GET: http://192.168.2.5:8520/resource/Sensors_V1/Outdoor_temperature_x10


```
{
       "Sensors_V1": {
        "Outdoor_temperature_x10": 189
       },
       "timestamp": "2018-09-13 18:09:37.534"
}
```


 # Hvac Sala 102
 # GET: http://192.168.2.5:8520/resource/Analyzer102ac_V1


```
{
       "timestamp": "2018-09-13 18:15:04.934",
```

```
        "Analyzer102ac_V1": {
         "AC102_P": 0,
         "AC102_Q": -29,
         "AC102_U": 222,
         "AC102_I": 2,
         "AC102_Cos": 0
        }
}
```


 # Hvac Sala 103

 # GET: http://192.168.2.5:8520/resource/Analyzer103ac_V1


```
{
       "Analyzer103ac_V1": {
       "AC103_P": 0,
       "AC103_Q": -29,
       "AC103_Cos": 0,
       "AC103_U": 226,
       "AC103_I": 2
      },
     "timestamp": "2018-09-13 18:16:05.372"
}
```


 # Hvac Sala 105

 # GET: http://192.168.2.5:8520/resource/Analyzer105ac_V1


```
{
        "Analyzer105ac_V1": {
         "AC105_Cos": 100,
         "AC105_Q": 0,
         "AC105_P": 0,
         "AC105_U": 235,
         "AC105_I": 0
        },
        "timestamp": "2018-09-13 18:16:49.907"
}
```


 # Hvac Salas 107, 108 e 109

 # GET: http://192.168.2.5:8520/resource/Analyzer107_108_109ac_V1


```
{
        "Analyzer107_108_109ac_V1": {
         "AC107_Cos": 100,
         "AC107_108_109_P_Total": 0,
         "AC109_I": 0,
         "AC107_I": 0,
         "AC108_I": 0,
         "AC107_P": 0,
         "AC108_Cos": 100,
         "AC107_108_109_Q_Total": 30,
         "AC109_Q": 0,
         "AC109_Cos": 100,
         "AC107_Q": 0,
         "AC108_P": 0,
         "AC108_Q": -29,
         "AC109_P": 0,
         "AC109_U": 234,
         "AC107_U": 236,
         "AC108_U": 233
        },
        "timestamp": "2018-09-13 18:17:40.625"
}
```

Anexo 3 Ficheiro de configuração: telegraf

```
## ############################################
## Definir Inputs
## ############################################
# Ler metrics - HTTP endpoints - Weather -teste
[[inputs.http]]
  urls = [
"http://api.openweathermap.org/data/2.5/weather?q=Porto&appid=f5674da8fa4622d3e691c10dea1f7fb1
&units=metric"
  ]
  name_suffix = "_weather1"
  interval = "1h"
  data_format = "json"

## ############################################ (500 call - 6 meses) -> 144
# Ler metrics - HTTP endpoints - Weather current (96)
[[inputs.http]]
  urls = [
"http://dataservice.accuweather.com/currentconditions/v1/275317?apikey=CmIGUQGhBcj7qEBhz58u1zi
AqP5qKdeD&language=pt&details=true&metric=true"
  ]
  name_suffix = "_weatherc"
  interval = "15m"
  data_format = "json"

## ############################################
# Ler metrics - HTTP endpoints - Weather Forecasting (48)
[[inputs.http]]
  urls = [
"http://dataservice.accuweather.com/forecasts/v1/hourly/12hour/275317?apikey=CmIGUQGhBcj7qEBhz
58u1ziAqP5qKdeD&language=pt&details=true&metric=true"
  ]
  name_suffix = "_weather"
  interval = "30m"
  data_format = "json"

## ############################################
# Ler metrics - HTTP endpoints - Gecad_SensorV1_Temperatura_outdoor
[[inputs.http]]
  urls = [
    "http://192.168.2.5:8520/resource/Sensors_V1/Outdoor_temperature_x10"
  ]
  name_suffix = "_wtemp"
  interval = "10m"
  data_format = "json"

## ############################################
# Ler metrics - HTTP endpoints - Gecad_power
[[inputs.http]]
  urls = [
```

```
    "http://192.168.2.5:8520/resource/Analyzer1_V4",
    "http://192.168.2.5:8520/resource/Analyzer2_V3",
    "http://192.168.2.5:8520/resource/Analyzer3_V3",
    "http://192.168.2.5:8520/resource/Analyzer4_V4",
    "http://192.168.2.5:8520/resource/Analyzer5_V3",
    "http://192.168.2.5:8520/resource/Analyzer115_V1",
    "http://192.168.2.5:8520/resource/Analyzer116_V1",
    "http://192.168.2.5:8520/resource/Analyzer102ac_V1",
    "http://192.168.2.5:8520/resource/Analyzer103ac_V1",
    "http://192.168.2.5:8520/resource/Analyzer105ac_V1",
    "http://192.168.2.5:8520/resource/Analyzer107_108_109ac_V1",
    "http://192.168.2.5:8520/resource/Inverter6_V3"
  ]
  name_suffix = "_power"
  data_format = "json"

## ###############################################
# Ler metrics - HTTP endpoints - Gecad Global
[[inputs.http]]
  urls = [
    "http://192.168.2.5:8520/building/energy"
  ]
  name_suffix = "_global"
  data_format = "json"

## ###############################################
# Ler metrics - HTTP endpoints - Gecad Kit
[[inputs.http]]
  urls = [
    "http://192.168.2.5:8520/resource/AnalyzerKitHall_V2"
  ]
  name_suffix = "_kitpz7"
  data_format = "json"

## ###############################################
# receber alertas
[[inputs.mqtt_consumer]]
servers = ["192.168.2.62:1883"]
qos = 0
topics = [
  "alertas"
]
name_suffix = "_z"
data_format = "influx"

## ###############################################
## Definir Outputs
## ###############################################
# DB - visualizar dados em real-time
#
[[outputs.influxdb]]
```

```
  database = "weather_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_weather"]

[[outputs.influxdb]]
  database = "power_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_power"]

[[outputs.influxdb]]
  database = "global_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_global"]

[[outputs.influxdb]]
  database = "z7_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_kitpz7"]

[[outputs.influxdb]]
  database = "weather1_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_weather1"]

[[outputs.influxdb]]
  database = "wtemp_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_wtemp"]

[[outputs.influxdb]]
  database = "alertas_metrics"
  urls = ["http://influxdb:8086"]
  namepass = ["*_z"]

## ###############################################
# MQ - tratamento dados em real-time
## = Power Por zona
[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz1"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
     url = ["http://192.168.2.5:8520/resource/Analyzer1_V4"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz2"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
```

```
    url = ["http://192.168.2.5:8520/resource/Analyzer2_V3"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz3"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
    url = ["http://192.168.2.5:8520/resource/Analyzer3_V3"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz4"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
    url = ["http://192.168.2.5:8520/resource/Analyzer4_V4"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz7"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
    url = ["http://192.168.2.5:8520/resource/Analyzer5_V3"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz8a"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
    url = ["http://192.168.2.5:8520/resource/Analyzer116_V1"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz8b"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
    url = ["http://192.168.2.5:8520/resource/Analyzer115_V1"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz9"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
    url = ["http://192.168.2.5:8520/resource/Inverter6_V3"]

## = Gecad HAVC
```

```
[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "power102ac"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
     url = ["http://192.168.2.5:8520/resource/Analyzer102ac_V1"]

  [[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "power102ac"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
     url = ["http://192.168.2.5:8520/resource/Analyzer103ac_V1"]

  [[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "power103ac"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
     url = ["http://192.168.2.5:8520/resource/Analyzer105ac_V1"]

  [[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "powerz3ac"
  qos = 2
  data_format = "json"
  [outputs.mqtt.tagpass]
     url = ["http://192.168.2.5:8520/resource/Analyzer107_108_109ac_V1"]

## = Gecad Global
[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "global"
  qos = 2
  data_format = "json"
  namepass = ["*_global"]

## = Gecad z7
[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "kitpz7"
  qos = 2
  data_format = "json"
  namepass = ["*_kitpz7"]

## = Wether
[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
```

```
  topic_prefix = "weather"
  qos = 2
  data_format = "json"
  namepass = ["*_weather"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "weatherc"
  qos = 2
  data_format = "json"
  namepass = ["*_weatherc"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "wtemp"
  qos = 2
  data_format = "json"
  namepass = ["*_wtemp"]

[[outputs.mqtt]]
  servers = ["192.168.2.62:1883"]
  topic_prefix = "weather1"
  qos = 2
  data_format = "json"
  namepass = ["*_weather1"]

## ############################################
```

Anexo 4 Analise e avaliação dos resultados obtidos pelos modelos ML.NET

Este anexo apresenta os resultados das experiências feitas sobre as funcionalizas disponibilizadas pela *framework* ML.NET, i.e., normalização e datasets para treino, algoritmos e performance na geração de modelos preditivos. O anexo está assim subdividido em: Funções de Normalização; Precisão das Previsões: DataSets & Algoritmos;

Funções de Normalização

Foram feitas experimentações sobre as funcionalidades disponibilizada pela Framework ML.NET, o qual permitiu concluir que estas têm influência sobre a previsão das previsões, bem como no tempo de execução na geração do modelo preditivo. As experimentações foram feitas sobre as quatro séries temporais da zona1, conforme mostra figuras seguintes.

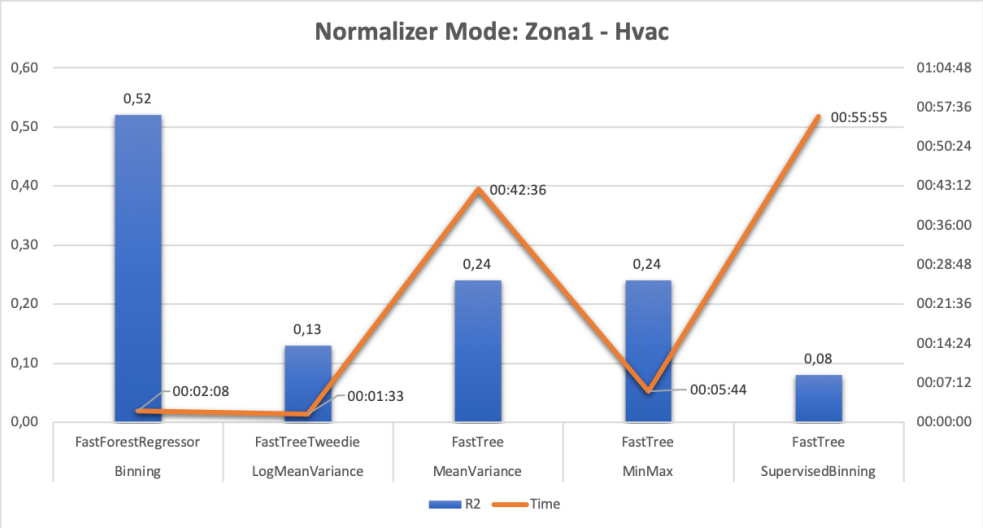


Figura A4.1 - Função de Normalização: Zona1 - Hvac

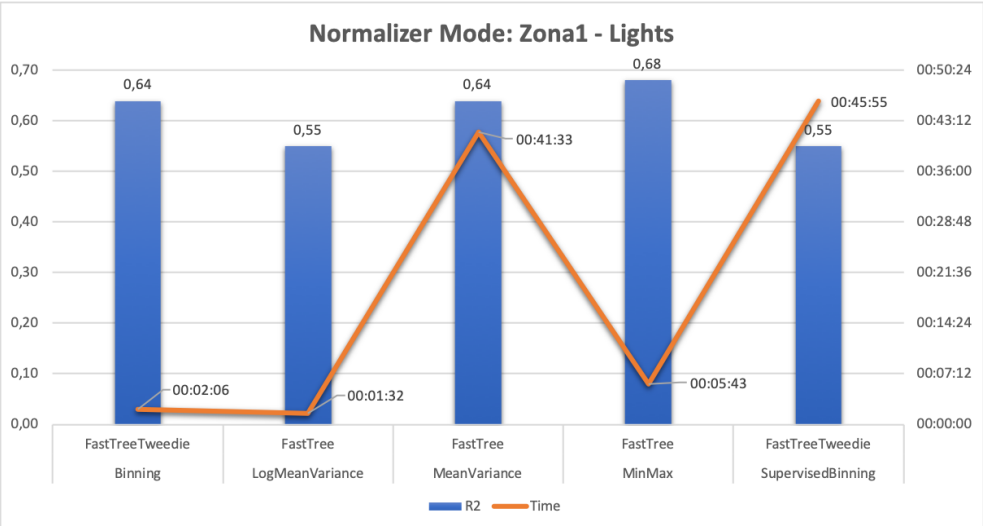


Figura A4.2 - Função de Normalização: Zona1 - Lights

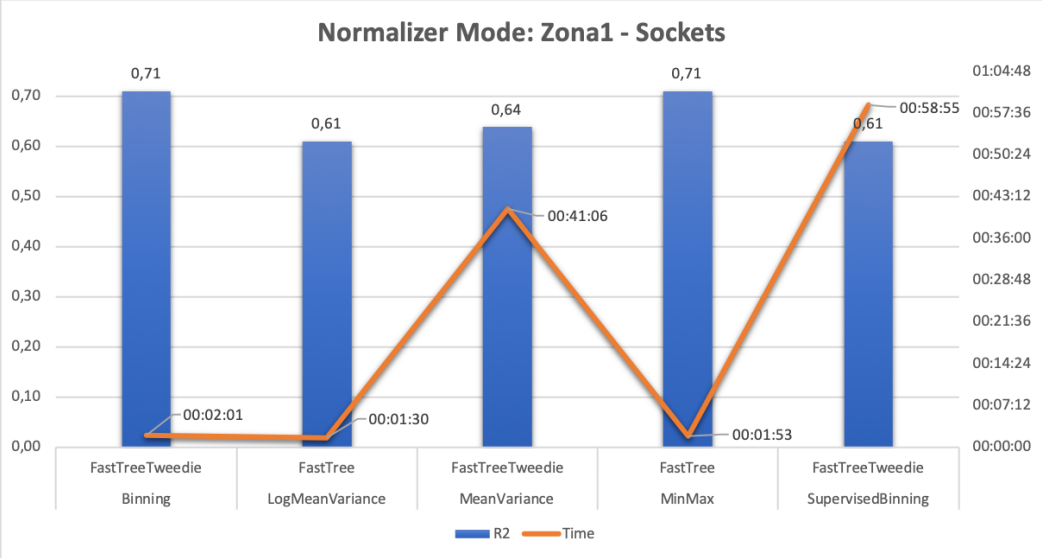


Figura A4.3 - Função de Normalização: Zona1 - Sockets

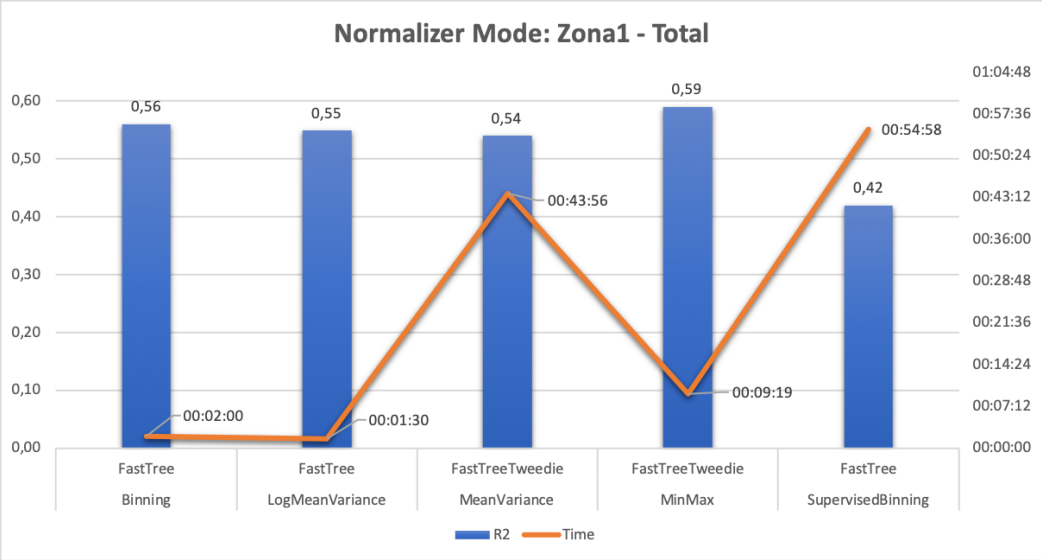


Figura A4.4 - Função de Normalização: Zona1 - Total

As experiências foram feitas tendo em conta as variáveis de input, no qual se concluiu que estas variáveis são determinantes para a precisão das previsões, mas não tem uma influência direta com os resultados obtidos pela função de normalização.

Como resultado da experimentação concluiu-se que as funções Binning e MinMax são as que conseguem obter melhores resultados na previsão das previsões. No entanto, A função Binning consegue obter uma melhor performance na execução dos modelos preditivos. Face aos resultados obtidos, foi adotada a função Binning para a normalização dos dados de treino no processo da geração de modelos preditivos. Os dados respeitantes às experimentações estão disponíveis nas tabelas seguinte.

Tabela A4.1 - Normalizer Mode: Resultados obtidos com a série temporal Hvac da Zona1

	contexto:	Without Temperature			Withe Temperature		
	time:	start	end	total	start	end	total
		11:42:36	11:45:10	0:02:34	11:48:41	11:50:49	0:02:08
Binning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,17	173,50	72,40	0,33	156,36	61,05
	FastTreeTweedie	0,17	173,33	71,43	0,32	157,18	57,36
	SDCA	0,04	186,98	89,75	0,05	185,59	90,38
	Poisson	-0,06	196,55	52,54	-0,06	196,20	52,83
	FastForestRegressor	0,15	175,51	78,53	0,52	73,64	45,26
	GeneralizedAdditiveModel	0,08	182,62	89,84	0,24	166,57	71,68
	OnlineGradientDescent	0,04	186,98	91,50	0,16	174,44	82,45
		11:54:17	11:55:28	0:01:11	12:01:12	12:02:45	0:01:33
LogMeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,10	200,54	118,72	0,13	154,85	104,15
	FastTreeTweedie	0,10	200,57	118,76	0,13	156,24	104,78
	SDCA	0,02	269,59	163,94	0,05	220,02	151,83
	Poisson	0,01	269,74	163,54	0,05	219,62	151,49
	FastForestRegressor	0,09	209,87	124,88	0,10	179,66	121,20
	GeneralizedAdditiveModel	0,08	212,94	126,84	0,10	184,88	124,51
	OnlineGradientDescent	0,01	269,72	163,58	0,04	220,73	151,89
		12:10:20	12:52:03	0:41:43	13:07:21	13:49:57	0:42:36
MeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,11	188,12	76,92	0,24	174,11	67,99
	FastTreeTweedie	0,11	188,07	76,11	0,23	175,21	64,79
	SDCA	0,01	198,28	96,97	0,02	196,91	98,45
	Poisson	-0,06	205,06	54,08	-0,05	204,06	54,59
	FastForestRegressor	0,09	190,29	81,85	0,14	184,19	77,17
	GeneralizedAdditiveModel	0,05	193,79	87,85	0,10	188,74	84,40
	OnlineGradientDescent	0,02	197,36	89,87	0,03	196,04	89,74
		14:56:05	15:01:27	0:05:22	15:17:08	15:22:52	0:05:44
MinMax	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,11	188,12	76,92	0,24	174,11	67,99
	FastTreeTweedie	0,11	188,07	76,11	0,23	175,21	64,79
	SDCA	0,01	198,28	96,97	0,02	196,91	98,45
	Poisson	-0,06	205,06	54,08	-0,05	204,06	54,59
	FastForestRegressor	0,09	190,29	81,85	0,14	184,19	77,17
	GeneralizedAdditiveModel	0,05	193,79	87,85	0,10	188,74	84,40
	OnlineGradientDescent	0,02	197,36	89,87	0,03	196,04	89,74
	--> erro	15:27:11	0:00:00	##########	15:44:33	0:00:00	##########
SupervisedBinning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,04	167,67	53,71	0,08	164,51	51,73
	FastTreeTweedie	0,04	167,60	53,49	0,01	169,81	43,16
	SDCA	0,03	168,60	57,47	0,03	168,46	56,20
	Poisson	-0,04	174,10	35,43	-0,04	174,26	35,48
	FastForestRegressor	0,05	166,68	53,15	0,06	166,12	51,73
	GeneralizedAdditiveModel	0,07	165,32	53,13	0,07	164,85	53,65
	OnlineGradientDescent	0,03	168,75	56,49	0,03	168,62	55,60

Tabela A4.2 - Normalizer Mode: Resultados obtidos com a série temporal Lihgt da Zona1

	contexto	Without Temperature			Withe Temperature		
	time	start	end	total	start	end	total
		16:12:38	16:14:43	0:02:05	16:26:23	16:28:29	0:02:06
Binning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,55	62,37	47,07	0,64	53,58	39,72
	FastTreeTweedie	0,55	62,33	47,03	0,64	53,71	39,84
	SDCA	0,11	94,07	70,10	0,20	88,61	66,31
	Poisson	0,08	95,45	69,18	0,18	89,61	65,26
	FastForestRegressor	0,46	70,08	51,62	0,49	68,21	50,18
	GeneralizedAdditiveModel	0,42	72,93	55,31	0,43	71,96	54,33
	OnlineGradientDescent	0,11	94,07	70,10	0,20	88,67	66,30
		16:33:41	16:35:15	0:01:34	16:41:32	16:43:04	0:01:32
LogMeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,44	71,62	52,89	0,55	62,78	46,36
	FastTreeTweedie	0,44	71,63	52,91	0,54	63,34	46,64
	SDCA	0,07	96,28	73,04	0,19	89,20	67,58
	Poisson	0,06	96,34	72,86	0,19	89,04	67,43
	FastForestRegressor	0,40	74,95	55,64	0,42	72,84	53,95
	GeneralizedAdditiveModel	0,38	76,05	56,51	0,40	74,95	55,42
	OnlineGradientDescent	0,06	96,33	72,88	0,18	89,49	67,61
		17:00:20	17:41:04	0:40:44	17:53:12	18:34:45	0:41:33
MeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,55	62,37	47,07	0,64	53,58	39,72
	FastTreeTweedie	0,55	62,31	47,03	0,64	53,71	39,84
	SDCA	0,05	97,02	70,14	0,06	96,46	69,44
	Poisson	-14,95	420,16	387,07	-14,48	413,86	381,20
	FastForestRegressor	0,46	70,08	51,62	0,49	68,21	50,18
	GeneralizedAdditiveModel	0,42	72,93	55,31	0,43	71,96	54,33
	OnlineGradientDescent	0,11	94,07	70,55	0,23	86,68	65,23
		18:38:48	18:44:33	0:05:45	18:58:02	19:03:45	0:05:43
MinMax	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,57	69,44	34,42	0,68	59,43	28,97
	FastTreeTweedie	0,57	69,37	34,17	0,67	60,56	27,51
	SDCA	0,13	98,92	68,33	0,21	94,31	66,34
	Poisson	-0,13	112,58	53,84	-0,01	106,45	50,09
	FastForestRegressor	0,48	76,32	57,36	0,52	73,64	45,26
	GeneralizedAdditiveModel	0,40	82,22	60,13	0,41	81,28	59,32
	OnlineGradientDescent	0,13	98,89	68,73	0,20	94,45	66,94
	--> erro	19:10:57	0:00:00	#########	19:27:51	0:00:00	########
SupervisedBinning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,55	62,37	47,07	0,55	62,37	47,07
	FastTreeTweedie	0,55	62,31	47,03	0,55	62,31	47,03
	SDCA	0,05	97,02	70,14	0,05	97,02	70,14
	Poisson	-14,95	420,16	387,07	-14,95	420,16	387,07
	FastForestRegressor	0,46	70,08	51,62	0,46	70,08	51,62
	GeneralizedAdditiveModel	0,42	72,93	55,31	0,42	72,93	55,31
	OnlineGradientDescent	0,11	94,07	70,55	0,11	94,07	70,55

Tabela A4.3 Normalizer Mode: Resultados obtidos com a série temporal Sockets da Zona1

	contexto:	Without Temperature			Withe Temperature		
	time:	start	end	total	start	end	total
		22:42:36	22:44:40	0:02:04	22:46:43	22:48:44	0:02:01
Binning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,30	0,71	59,53	44,13
	FastTreeTweedie	0,61	69,26	52,26	0,71	59,68	44,27
	SDCA	0,12	104,52	77,89	0,22	98,45	73,68
	Poisson	0,09	106,06	76,87	0,20	99,57	72,51
	FastForestRegressor	0,51	77,87	57,36	0,54	75,79	55,76
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,48	79,96	60,37
	OnlineGradientDescent	0,12	104,52	77,89	0,22	98,52	73,67
		23:33:47	23:35:08	0:01:21	0:01:12	0:02:42	0:01:30
LogMeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,49	79,58	58,77	0,61	69,75	51,51
	FastTreeTweedie	0,49	79,59	58,79	0,60	70,38	51,82
	SDCA	0,08	106,98	81,16	0,21	99,11	75,09
	Poisson	0,07	107,04	80,96	0,21	98,93	74,92
	FastForestRegressor	0,44	83,28	61,82	0,47	80,93	59,94
	GeneralizedAdditiveModel	0,42	84,50	62,79	0,44	83,28	61,58
	OnlineGradientDescent	0,07	107,03	80,98	0,20	99,43	75,12
		0:10:20	0:51:02	0:40:42	1:07:41	1:48:47	0:41:06
MeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,30	0,71	59,53	44,13
	FastTreeTweedie	0,61	69,23	52,26	0,71	59,68	44,27
	SDCA	0,06	107,80	77,93	0,07	107,18	77,16
	Poisson	-16,61	466,84	430,08	-16,09	459,84	423,55
	FastForestRegressor	0,51	77,87	57,36	0,54	75,79	55,76
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,48	79,96	60,37
	OnlineGradientDescent	0,12	104,52	78,39	0,25	96,31	72,48
		1:48:53	1:50:31	0:01:38	1:55:14	1:57:07	0:01:53
MinMax	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,03	0,71	59,53	44,13
	FastTreeTweedie	0,61	69,26	52,26	0,71	59,68	44,27
	SDCA	0,12	104,51	78,30	0,25	96,18	72,61
	Poisson	0,09	106,30	77,31	0,23	97,43	71,32
	FastForestRegressor	0,51	77,87	57,36	0,54	75,79	55,76
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,48	79,96	60,37
	OnlineGradientDescent	0,12	104,55	78,47	0,25	96,37	72,47
	--> erro	2:02:18	0:00:00	##########	2:14:31	0:00:00	##########
SupervisedBinning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,30	0,61	69,30	52,30
	FastTreeTweedie	0,61	69,23	52,26	0,61	69,23	52,26
	SDCA	0,06	107,80	77,93	0,06	107,80	77,93
	Poisson	-16,61	466,84	430,08	-16,61	466,84	430,08
	FastForestRegressor	0,51	77,87	57,36	0,51	77,87	57,36
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,47	81,03	61,46
	OnlineGradientDescent	0,12	104,52	78,39	0,12	104,52	78,39

Tabela A4.4 - Normalizer Mode: Resultados obtidos com a série temporal Total da Zona1

	contexto:	Without Temperature			Withe Temperature		
	time:	start	end	total	start	end	total
		22:42:36	22:44:40	0:02:04	22:46:43	22:48:44	0:02:01
Binning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,30	0,71	59,53	44,13
	FastTreeTweedie	0,61	69,26	52,26	0,71	59,68	44,27
	SDCA	0,12	104,52	77,89	0,22	98,45	73,68
	Poisson	0,09	106,06	76,87	0,20	99,57	72,51
	FastForestRegressor	0,51	77,87	57,36	0,54	75,79	55,76
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,48	79,96	60,37
	OnlineGradientDescent	0,12	104,52	77,89	0,22	98,52	73,67
		23:33:47	23:35:08	0:01:21	0:01:12	0:02:42	0:01:30
LogMeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,49	79,58	58,77	0,61	69,75	51,51
	FastTreeTweedie	0,49	79,59	58,79	0,60	70,38	51,82
	SDCA	0,08	106,98	81,16	0,21	99,11	75,09
	Poisson	0,07	107,04	80,96	0,21	98,93	74,92
	FastForestRegressor	0,44	83,28	61,82	0,47	80,93	59,94
	GeneralizedAdditiveModel	0,42	84,50	62,79	0,44	83,28	61,58
	OnlineGradientDescent	0,07	107,03	80,98	0,20	99,43	75,12
		0:10:20	0:51:02	0:40:42	1:07:41	1:48:47	0:41:06
MeanVariance	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,30	0,71	59,53	44,13
	FastTreeTweedie	0,61	69,23	52,26	0,71	59,68	44,27
	SDCA	0,06	107,80	77,93	0,07	107,18	77,16
	Poisson	-16,61	466,84	430,08	-16,09	459,84	423,55
	FastForestRegressor	0,51	77,87	57,36	0,54	75,79	55,76
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,48	79,96	60,37
	OnlineGradientDescent	0,12	104,52	78,39	0,25	96,31	72,48
		1:48:53	1:50:31	0:01:38	1:55:14	1:57:07	0:01:53
MinMax	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,03	0,71	59,53	44,13
	FastTreeTweedie	0,61	69,26	52,26	0,71	59,68	44,27
	SDCA	0,12	104,51	78,30	0,25	96,18	72,61
	Poisson	0,09	106,30	77,31	0,23	97,43	71,32
	FastForestRegressor	0,51	77,87	57,36	0,54	75,79	55,76
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,48	79,96	60,37
	OnlineGradientDescent	0,12	104,55	78,47	0,25	96,37	72,47
	--> erro	2:02:18	0:00:00	##########	2:14:31	0:00:00	##########
SupervisedBinning	Treiners	R2 Score	RMS loss	L1	R2 Score	RMS loss	L1
	FastTree	0,61	69,30	52,30	0,61	69,30	52,30
	FastTreeTweedie	0,61	69,23	52,26	0,61	69,23	52,26
	SDCA	0,06	107,80	77,93	0,06	107,80	77,93
	Poisson	-16,61	466,84	430,08	-16,61	466,84	430,08
	FastForestRegressor	0,51	77,87	57,36	0,51	77,87	57,36
	GeneralizedAdditiveModel	0,47	81,03	61,46	0,47	81,03	61,46
	OnlineGradientDescent	0,12	104,52	78,39	0,12	104,52	78,39

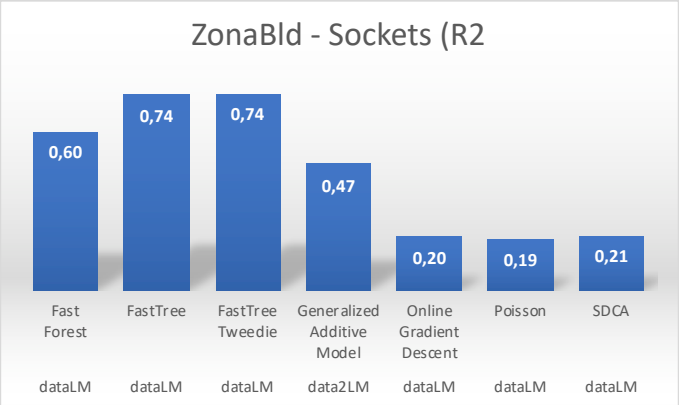
Precisão das Previsões: DataSets & Algoritmos

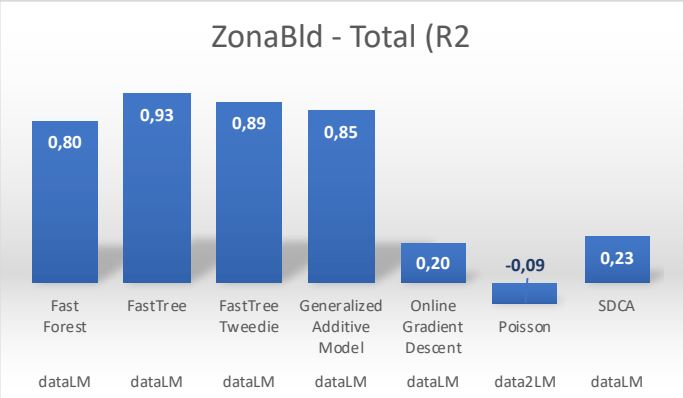
Estas experiências visaram testar a eficiência dos algoritmos de acordo com o *datasete* disponibilizado para treino do modelo. As experimentações foram feitas sobre as quatro séries temporais da zona1 e zonaBld e ainda sobre a série de produção da zona9. Os resultados obtidos podem ser observados nas figuras seguintes.

Figura A4.5 - Precisão das Previsões: DataSets & Algoritmos - zona

Dos resultados obtidos na avaliação dos modelos preditivos é possível concluir que os melhores algoritmos são: FastTree e o FastTreeTweedie. O intervalo de dados mais favorável para o treino dos algoritmos corresponde na maioria ao último mês de dados históricos.

Figura A4.6 - Precisão das Previsões: DataSets & Algoritmos - zonaBld

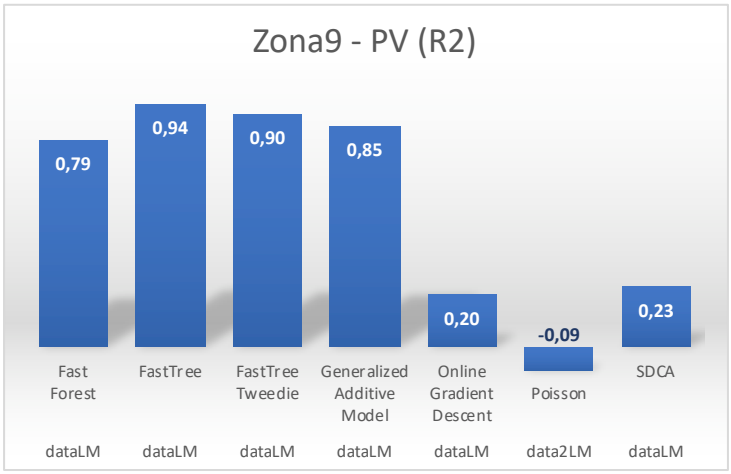




Dos resultados obtidos na avaliação dos modelos preditivos para a zonaBld é possível concluir novamente que os melhores algoritmos são: FastTree e o FastTreeTweedie. Da mesma forma, é possível concluir que o intervalo de dados mais favorável para o treino dos algoritmos corresponde na sua maioria ao último mês de dados históricos.

Finalmente para a zona de produção, os dados obtidos permitiram concluir que o melhor resultado foi obtido pelo algoritmo FastTree e o melhor dataset para treino dos algoritmos corresponde aquele que contem como histórico o último mês de dados.

Figura A4.7 - Precisão das Previsões: DataSets & Algoritmos – zona9



Os dados respeitantes às experimentações estão disponíveis nas tabelas seguinte.

Previsões: Tempo de execução

Os tempos de execução resultantes das experiências realizadas podem ser analisadas nas figuras e tabelas seguintes

Figura A4.8 - Tempo total na execução por algoritmo e dataset para treino: zona1 [Hvac]

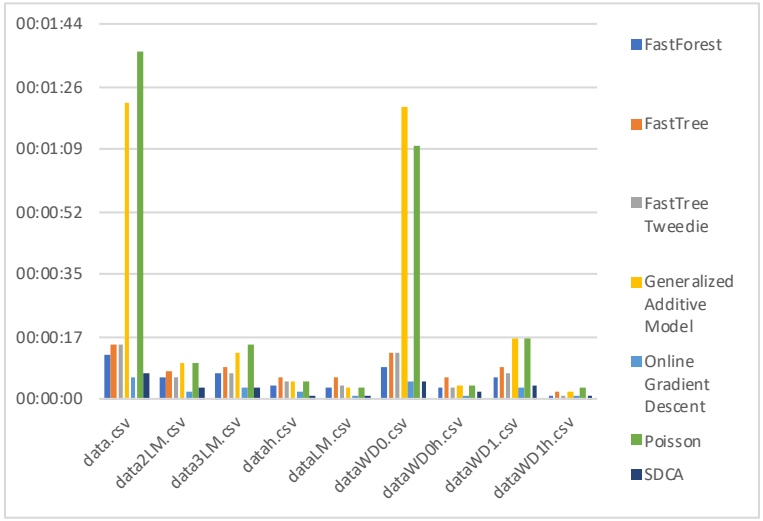


Tabela A4.5 - Tempo de execução do processo de geração de novos modelos preditivos - Zona1 [Hvac]

DataSets	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data.csv	00:00:12	00:00:15	00:00:15	00:01:22	00:00:06	00:01:36	00:00:07	00:01:36
data2LM.csv	00:00:06	00:00:08	00:00:06	00:00:10	00:00:02	00:00:10	00:00:03	00:00:10
data3LM.csv	00:00:07	00:00:09	00:00:07	00:00:13	00:00:03	00:00:15	00:00:03	00:00:15
datah.csv	00:00:04	00:00:06	00:00:05	00:00:05	00:00:02	00:00:05	00:00:01	00:00:06
dataLM.csv	00:00:03	00:00:06	00:00:04	00:00:03	00:00:01	00:00:03	00:00:01	00:00:06
dataWD0.csv	00:00:09	00:00:13	00:00:13	00:01:21	00:00:05	00:01:10	00:00:05	00:01:21
dataWD0h.csv	00:00:03	00:00:06	00:00:03	00:00:04	00:00:01	00:00:04	00:00:02	00:00:06
dataWD1.csv	00:00:06	00:00:09	00:00:07	00:00:17	00:00:03	00:00:17	00:00:04	00:00:17
dataWD1h.csv	00:00:01	00:00:02	00:00:01	00:00:02	00:00:01	00:00:03	00:00:01	00:00:03
Maximo	**00:00:12**	**00:00:15**	**00:00:15**	**00:01:22**	**00:00:06**	**00:01:36**	**00:00:07**	**00:01:36**
Total	**00:00:51**	**00:01:14**	**00:01:01**	**00:03:37**	**00:00:24**	**00:03:43**	**00:00:27**	**00:04:00**

Figura A4.9 - Tempo total na execução por algoritmo e dataset para treino: zona1 [Lights]

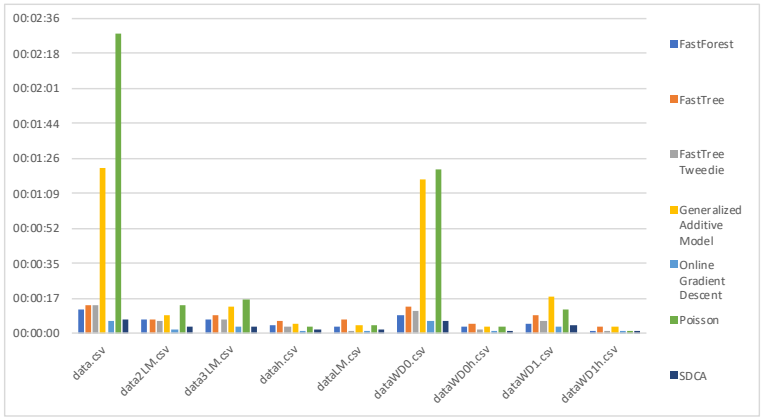


Tabela A4.6 - Tempo de execução do processo de geração de novos modelos preditivos - Zona1 [Lights]

DataSets	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data.csv	00:00:12	00:00:14	00:00:14	00:01:22	00:00:06	00:02:28	00:00:07	00:02:28
data2LM.csv	00:00:07	00:00:07	00:00:06	00:00:09	00:00:02	00:00:14	00:00:03	00:00:14
data3LM.csv	00:00:07	00:00:09	00:00:07	00:00:13	00:00:03	00:00:17	00:00:03	00:00:17
datah.csv	00:00:04	00:00:06	00:00:03	00:00:05	00:00:01	00:00:03	00:00:02	00:00:06
dataLM.csv	00:00:03	00:00:07	00:00:01	00:00:04	00:00:01	00:00:04	00:00:02	00:00:07
dataWD0.csv	00:00:09	00:00:13	00:00:11	00:01:16	00:00:06	00:01:21	00:00:06	00:01:21
dataWD0h.csv	00:00:03	00:00:05	00:00:02	00:00:03	00:00:01	00:00:03	00:00:01	00:00:05
dataWD1.csv	00:00:05	00:00:09	00:00:06	00:00:18	00:00:03	00:00:12	00:00:04	00:00:18
dataWD1h.csv	00:00:01	00:00:03	00:00:01	00:00:03	00:00:01	00:00:01	00:00:01	00:00:03
Maximo	**00:00:12**	**00:00:14**	**00:00:14**	**00:01:22**	**00:00:06**	**00:02:28**	**00:00:07**	**00:02:28**
Total	**00:00:51**	**00:01:13**	**00:00:51**	**00:03:33**	**00:00:24**	**00:04:43**	**00:00:29**	**00:04:59**

Figura A4.10 - Tempo total na execução por algoritmo e dataset para treino: zona1 [Sockets]

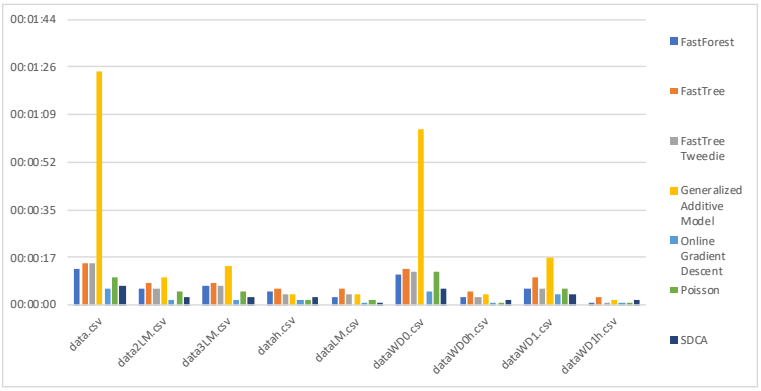


Tabela A4.7 - Tempo de execução do processo de geração de novos modelos preditivos - Zona1 [Sockets]

DataSets	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data.csv	00:00:13	00:00:15	00:00:15	00:01:25	00:00:06	00:00:10	00:00:07	00:01:25
data2LM.csv	00:00:06	00:00:08	00:00:06	00:00:10	00:00:02	00:00:05	00:00:03	00:00:10
data3LM.csv	00:00:07	00:00:08	00:00:07	00:00:14	00:00:02	00:00:05	00:00:03	00:00:14
datah.csv	00:00:05	00:00:06	00:00:04	00:00:04	00:00:02	00:00:02	00:00:03	00:00:06
dataLM.csv	00:00:03	00:00:06	00:00:04	00:00:04	00:00:01	00:00:02	00:00:01	00:00:06
dataWD0.csv	00:00:11	00:00:13	00:00:12	00:01:04	00:00:05	00:00:12	00:00:06	00:01:04
dataWD0h.csv	00:00:03	00:00:05	00:00:03	00:00:04	00:00:01	00:00:01	00:00:02	00:00:05
dataWD1.csv	00:00:06	00:00:10	00:00:06	00:00:17	00:00:04	00:00:06	00:00:04	00:00:17
dataWD1h.csv	00:00:01	00:00:03	00:00:01	00:00:02	00:00:01	00:00:01	00:00:02	00:00:03
Maximo	**00:00:13**	**00:00:15**	**00:00:15**	**00:01:25**	**00:00:06**	**00:00:12**	**00:00:07**	**00:01:25**
Total	**00:00:55**	**00:01:14**	**00:00:58**	**00:03:24**	**00:00:24**	**00:00:44**	**00:00:31**	**00:03:30**

Figura A4.11 - Tempo total na execução por algoritmo e dataset para treino: zona1 [Total]

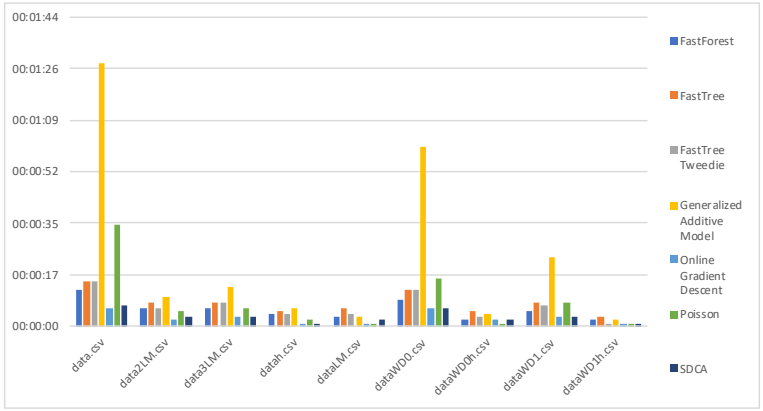


Tabela A4.8 - Tempo de execução do processo de geração de novos modelos preditivos - Zona1 [Total]

DataSets	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data.csv	00:00:12	00:00:15	00:00:15	00:01:28	00:00:06	00:00:34	00:00:07	00:01:28
data2LM.csv	00:00:06	00:00:08	00:00:06	00:00:10	00:00:02	00:00:05	00:00:03	00:00:10
data3LM.csv	00:00:06	00:00:08	00:00:08	00:00:13	00:00:03	00:00:06	00:00:03	00:00:13
datah.csv	00:00:04	00:00:05	00:00:04	00:00:06	00:00:01	00:00:02	00:00:01	00:00:06
dataLM.csv	00:00:03	00:00:06	00:00:04	00:00:03	00:00:01	00:00:01	00:00:02	00:00:06
dataWD0.csv	00:00:09	00:00:12	00:00:12	00:01:00	00:00:06	00:00:16	00:00:06	00:01:00
dataWD0h.csv	00:00:02	00:00:05	00:00:03	00:00:04	00:00:02	00:00:01	00:00:02	00:00:05
dataWD1.csv	00:00:05	00:00:08	00:00:07	00:00:23	00:00:03	00:00:08	00:00:03	00:00:23
dataWD1h.csv	00:00:02	00:00:03	00:00:01	00:00:02	00:00:01	00:00:01	00:00:01	00:00:03
Maximo	**00:00:12**	**00:00:15**	**00:00:15**	**00:01:28**	**00:00:06**	**00:00:34**	**00:00:07**	**00:01:28**
Total	**00:00:49**	**00:01:10**	**00:01:00**	**00:03:29**	**00:00:25**	**00:01:14**	**00:00:28**	**00:03:34**

Figura A4.12 - Tempo total na execução por algoritmo e dataset para treino: zonaBld [Hvac]

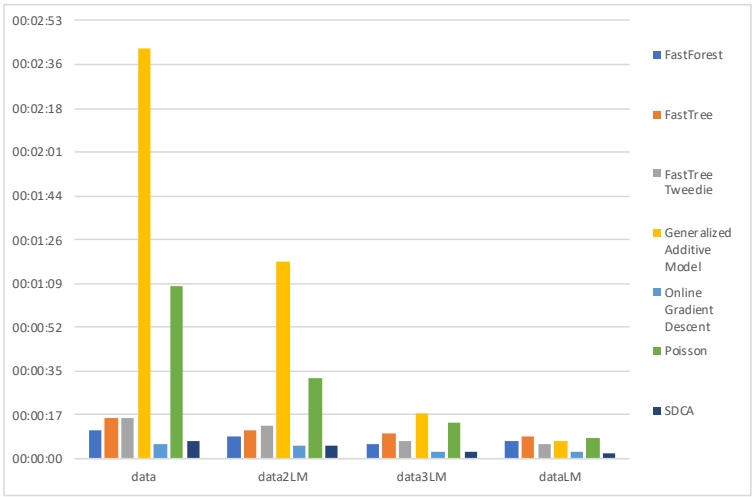


Tabela A4.9 - Tempo de execução do processo de geração de novos modelos preditivos - ZonaBld [Hvac]

DataSet	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data	00:00:11	00:00:16	00:00:16	00:02:42	00:00:06	00:01:08	00:00:07	00:02:42
data2LM	00:00:09	00:00:11	00:00:13	00:01:18	00:00:05	00:00:32	00:00:05	00:01:18
data3LM	00:00:06	00:00:10	00:00:07	00:00:18	00:00:03	00:00:14	00:00:03	00:00:18
dataLM	00:00:07	00:00:09	00:00:06	00:00:07	00:00:03	00:00:08	00:00:02	00:00:09
Maximo	**00:00:11**	**00:00:16**	**00:00:16**	**00:02:42**	**00:00:06**	**00:01:08**	**00:00:07**	**00:02:42**
Total	**00:00:33**	**00:00:46**	**00:00:42**	**00:04:25**	**00:00:17**	**00:02:02**	**00:00:17**	**00:04:27**

Figura A4.13 - Tempo total na execução por algoritmo e dataset para treino: zonaBld [Lights]

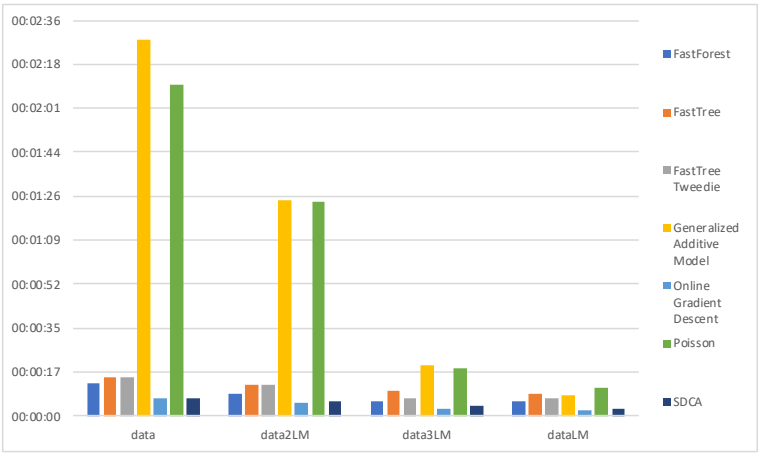


Tabela A4.10 - Tempo de execução do processo de geração de novos modelos preditivos - ZonaBld [Lights]

DataSet	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data	00:00:13	00:00:15	00:00:15	00:02:28	00:00:07	00:02:10	00:00:07	00:02:28
data2LM	00:00:09	00:00:12	00:00:12	00:01:25	00:00:05	00:01:24	00:00:06	00:01:25
data3LM	00:00:06	00:00:10	00:00:07	00:00:20	00:00:03	00:00:19	00:00:04	00:00:20
dataLM	00:00:06	00:00:09	00:00:07	00:00:08	00:00:02	00:00:11	00:00:03	00:00:11
Maximo	**00:00:13**	**00:00:15**	**00:00:15**	**00:02:28**	**00:00:07**	**00:02:10**	**00:00:07**	**00:02:28**
Total	**00:00:34**	**00:00:46**	**00:00:41**	**00:04:21**	**00:00:17**	**00:04:04**	**00:00:20**	**00:04:24**

Figura A4.14 - Tempo total na execução por algoritmo e dataset para treino: zonaBld [Sockets]

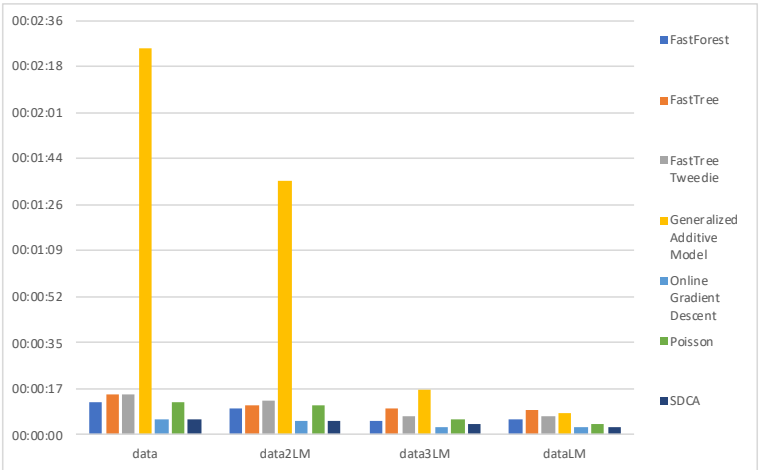


Tabela A4.11 - Tempo de execução do processo de geração de novos modelos preditivos - Zona1 [Sockets]

DataSet	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data	00:00:12	00:00:15	00:00:15	00:02:25	00:00:06	00:00:12	00:00:06	00:02:25
data2LM	00:00:10	00:00:11	00:00:13	00:01:35	00:00:05	00:00:11	00:00:05	00:01:35
data3LM	00:00:05	00:00:10	00:00:07	00:00:17	00:00:03	00:00:06	00:00:04	00:00:17
dataLM	00:00:06	00:00:09	00:00:07	00:00:08	00:00:03	00:00:04	00:00:03	00:00:09
Maximo	**00:00:12**	**00:00:15**	**00:00:15**	**00:02:25**	**00:00:06**	**00:00:12**	**00:00:06**	**00:02:25**
Total	**00:00:33**	**00:00:45**	**00:00:42**	**00:04:25**	**00:00:17**	**00:00:33**	**00:00:18**	**00:04:26**

Figura A4.15 - Tempo total na execução por algoritmo e dataset para treino: zonaBld [Total]

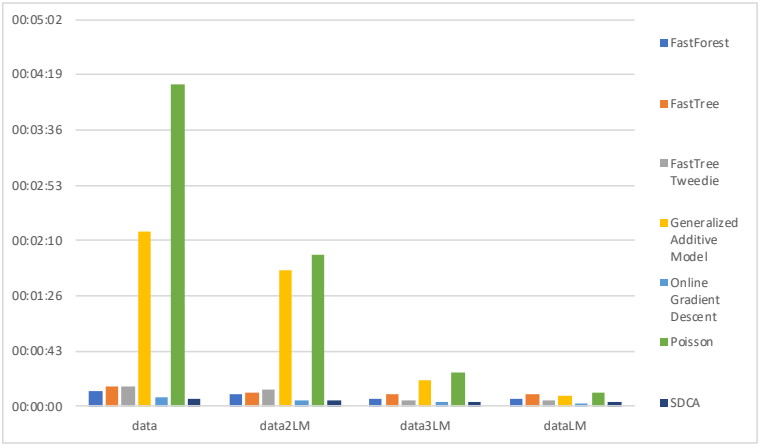


Tabela A4.12 - Tempo de execução do processo de geração de novos modelos preditivos - Zona1 [Total]

DataSet	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data	00:00:12	00:00:16	00:00:16	00:02:16	00:00:07	00:04:11	00:00:06	00:04:11
data2LM	00:00:10	00:00:11	00:00:13	00:01:46	00:00:05	00:01:58	00:00:05	00:01:58
data3LM	00:00:06	00:00:10	00:00:05	00:00:20	00:00:03	00:00:27	00:00:04	00:00:27
dataLM	00:00:06	00:00:09	00:00:05	00:00:08	00:00:02	00:00:11	00:00:03	00:00:11
Maximo	**00:00:12**	**00:00:16**	**00:00:16**	**00:02:16**	**00:00:07**	**00:04:11**	**00:00:06**	**00:04:11**
Total	**00:00:34**	**00:00:46**	**00:00:39**	**00:04:30**	**00:00:17**	**00:06:47**	**00:00:18**	**00:06:47**

Figura A4.16 - Tempo total na execução por algoritmo e dataset para treino: zona9 [PV]

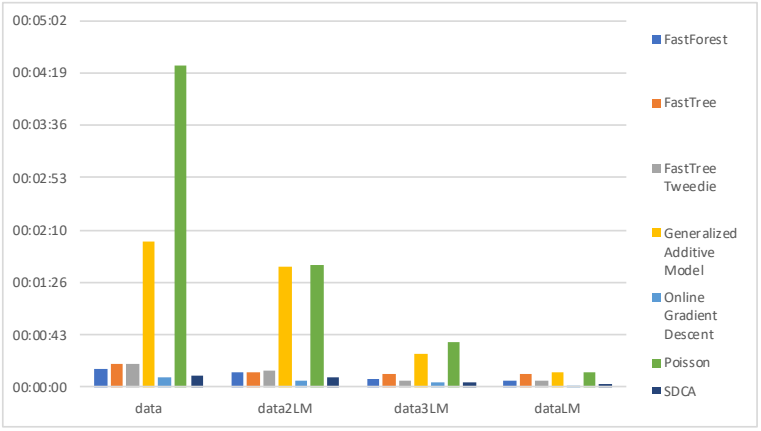


Tabela A4.13 - Tempo de execução do processo de geração de novos modelos preditivos - Zona9 [PV]

DataSets	FastForest	FastTree	FastTree Tweedie	Generalized Additive Model	Online Gradient Descent	Poisson	SDCA	Maximo
data	00:00:15	00:00:19	00:00:19	00:02:00	00:00:08	00:04:26	00:00:09	00:04:26
data2LM	00:00:12	00:00:13	00:00:14	00:01:39	00:00:06	00:01:41	00:00:08	00:01:41
data3LM	00:00:07	00:00:11	00:00:06	00:00:27	00:00:04	00:00:37	00:00:04	00:00:37
dataLM	00:00:06	00:00:11	00:00:05	00:00:12	00:00:02	00:00:12	00:00:03	00:00:12
Maximo	**00:00:15**	**00:00:19**	**00:00:19**	**00:02:00**	**00:00:08**	**00:04:26**	**00:00:09**	**00:04:26**
Total	**00:00:40**	**00:00:54**	**00:00:44**	**00:04:18**	**00:00:20**	**00:06:56**	**00:00:24**	**00:06:56**

Da análise dos ressoltados obtidos pode concluir-se que o tempo total para a execução do processo de geração de novos modelos varia entre 3 a 6 minutos. O treino e a geração de um novo modelo são normalmente inferiores a 10 segundos para a maioria dos algoritmos, dos quais se excluem o Poisson e o GeneralizedAdditiveModel.

Nas tabelas seguintes estão disponíveis todos os dados relativos às análises efetuadas.

Tabela A4.14 - Dados da avaliação dos modelos gerados para a Zona1 [Hvac]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data.csv	FastTree	0,24	174	68	23/03/19	15:06:41	00:00:15
	FastTreeTweedie	0,23	175	65	23/03/19	15:06:56	00:00:15
	SDCA	0,03	196	91	23/03/19	15:07:03	00:00:07
	Poisson	-0,05	204	55	23/03/19	15:08:39	00:01:36
	FastForest	0,14	184	77	23/03/19	15:08:51	00:00:12
	GeneralizedAdditiveModel	0,10	189	84	23/03/19	15:10:13	00:01:22
	OnlineGradientDescent	0,03	196	90	23/03/19	15:10:19	00:00:06
dataLM.csv	FastTree	0,49	29	6	23/03/19	15:10:25	00:00:06
	FastTreeTweedie	0,54	28	4	23/03/19	15:10:29	00:00:04
	SDCA	0,01	41	7	23/03/19	15:10:30	00:00:01
	Poisson	0,00	41	5	23/03/19	15:10:33	00:00:03
	FastForest	0,22	36	6	23/03/19	15:10:36	00:00:03
	GeneralizedAdditiveModel	0,05	40	8	23/03/19	15:10:39	00:00:03
	OnlineGradientDescent	0,01	41	7	23/03/19	15:10:40	00:00:01
data2LM.csv	FastTree	0,23	224	104	23/03/19	15:10:48	00:00:08
	FastTreeTweedie	0,24	222	96	23/03/19	15:10:54	00:00:06
	SDCA	0,08	245	139	23/03/19	15:10:57	00:00:03
	Poisson	-0,08	265	82	23/03/19	15:11:07	00:00:10
	FastForest	0,17	231	120	23/03/19	15:11:13	00:00:06
	GeneralizedAdditiveModel	0,14	237	128	23/03/19	15:11:23	00:00:10
	OnlineGradientDescent	0,07	246	133	23/03/19	15:11:25	00:00:02
data3LM.csv	FastTree	0,24	222	110	23/03/19	15:11:34	00:00:09
	FastTreeTweedie	0,23	223	104	23/03/19	15:11:41	00:00:07
	SDCA	0,04	249	139	23/03/19	15:11:44	00:00:03
	Poisson	-0,08	265	81	23/03/19	15:11:59	00:00:15
	FastForest	0,14	236	126	23/03/19	15:12:06	00:00:07
	GeneralizedAdditiveModel	0,07	245	136	23/03/19	15:12:19	00:00:13
	OnlineGradientDescent	0,04	249	138	23/03/19	15:12:22	00:00:03
dataWD0.csv	FastTree	0,25	180	77	23/03/19	15:50:10	00:00:13
	FastTreeTweedie	0,24	181	74	23/03/19	15:50:23	00:00:13
	SDCA	0,04	203	103	23/03/19	15:50:28	00:00:05
	Poisson	-0,05	213	62	23/03/19	15:51:38	00:01:10
	FastForest	0,17	189	86	23/03/19	15:51:47	00:00:09
	GeneralizedAdditiveModel	0,15	191	89	23/03/19	15:53:08	00:01:21
	OnlineGradientDescent	0,04	203	102	23/03/19	15:53:13	00:00:05
dataWD1.csv	FastTree	0,17	131	35	23/03/19	15:53:22	00:00:09
	FastTreeTweedie	0,17	131	31	23/03/19	15:53:29	00:00:07
	SDCA	0,01	143	41	23/03/19	15:53:33	00:00:04
	Poisson	-0,02	146	24	23/03/19	15:53:50	00:00:17
	FastForest	0,08	138	38	23/03/19	15:53:56	00:00:06
	GeneralizedAdditiveModel	0,04	141	40	23/03/19	15:54:13	00:00:17
	OnlineGradientDescent	0,01	143	42	23/03/19	15:54:16	00:00:03
dataWD0h.csv	FastTree	0,08	165	52	23/03/19	15:54:22	00:00:06
	FastTreeTweedie	0,01	170	43	23/03/19	15:54:25	00:00:03
	SDCA	0,03	168	56	23/03/19	15:54:27	00:00:02
	Poisson	-0,04	174	35	23/03/19	15:54:31	00:00:04
	FastForest	0,06	166	52	23/03/19	15:54:34	00:00:03
	GeneralizedAdditiveModel	0,07	165	54	23/03/19	15:54:38	00:00:04
	OnlineGradientDescent	0,03	169	56	23/03/19	15:54:39	00:00:01
dataWD1h.csv	FastTree	0,22	170	52	23/03/19	15:54:41	00:00:02
	FastTreeTweedie	0,17	175	42	23/03/19	15:54:42	00:00:01
	SDCA	0,01	192	54	23/03/19	15:54:43	00:00:01
	Poisson	-0,04	196	42	23/03/19	15:54:46	00:00:03
	FastForest	0,19	173	47	23/03/19	15:54:47	00:00:01
	GeneralizedAdditiveModel	0,12	180	50	23/03/19	15:54:49	00:00:02
	OnlineGradientDescent	0,00	192	54	23/03/19	15:54:50	00:00:01
datah.csv	FastTree	-0,01	153	47	23/03/19	15:54:56	00:00:06
	FastTreeTweedie	-0,10	160	39	23/03/19	15:55:01	00:00:05
	SDCA	0,02	151	49	23/03/19	15:55:02	00:00:01
	Poisson	-0,03	154	28	23/03/19	15:55:07	00:00:05
	FastForest	0,02	150	46	23/03/19	15:55:11	00:00:04
	GeneralizedAdditiveModel	0,02	151	48	23/03/19	15:55:16	00:00:05
	OnlineGradientDescent	0,02	151	49	23/03/19	15:55:18	00:00:02

Tabela A4.15 - Dados da avaliação dos modelos gerados para a Zona1 [Lights]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data.csv	FastTree	0,70	63	31	23/03/19	13:54:37	00:00:14
	FastTreeTweedie	0,69	64	30	23/03/19	13:54:51	00:00:14
	SDCA	0,19	104	78	23/03/19	13:54:58	00:00:07
	Poisson	-0,08	119	60	23/03/19	13:57:26	00:02:28
	FastForest	0,56	76	50	23/03/19	13:57:38	00:00:12
	GeneralizedAdditiveModel	0,47	84	65	23/03/19	13:59:00	00:01:22
	OnlineGradientDescent	0,19	103	79	23/03/19	13:59:06	00:00:06
dataLM.csv	FastTree	0,95	25	7	23/03/19	13:59:13	00:00:07
	FastTreeTweedie	0,89	37	13	23/03/19	13:59:14	00:00:01
	SDCA	0,47	83	60	23/03/19	13:59:16	00:00:02
	Poisson	-7,33	330	95	23/03/19	13:59:20	00:00:04
	FastForest	0,82	49	29	23/03/19	13:59:23	00:00:03
	GeneralizedAdditiveModel	0,63	69	50	23/03/19	13:59:27	00:00:04
	OnlineGradientDescent	0,39	89	67	23/03/19	13:59:28	00:00:01
data2LM.csv	FastTree	0,78	53	27	23/03/19	13:59:35	00:00:07
	FastTreeTweedie	0,76	56	26	23/03/19	13:59:41	00:00:06
	SDCA	0,31	95	73	23/03/19	13:59:44	00:00:03
	Poisson	-1,32	174	68	23/03/19	13:59:58	00:00:14
	FastForest	0,62	71	48	23/03/19	14:00:05	00:00:07
	GeneralizedAdditiveModel	0,47	83	65	23/03/19	14:00:14	00:00:09
	OnlineGradientDescent	0,29	96	74	23/03/19	14:00:16	00:00:02
data3LM.csv	FastTree	0,74	55	27	23/03/19	14:00:25	00:00:09
	FastTreeTweedie	0,73	56	26	23/03/19	14:00:32	00:00:07
	SDCA	0,33	88	66	23/03/19	14:00:35	00:00:03
	Poisson	-2,06	189	60	23/03/19	14:00:52	00:00:17
	FastForest	0,58	70	45	23/03/19	14:00:59	00:00:07
	GeneralizedAdditiveModel	0,45	80	61	23/03/19	14:01:12	00:00:13
	OnlineGradientDescent	0,31	90	68	23/03/19	14:01:15	00:00:03
dataWD0.csv	FastTree	0,68	72	40	23/03/19	15:27:21	00:00:13
	FastTreeTweedie	0,67	73	40	23/03/19	15:27:32	00:00:11
	SDCA	0,14	118	95	23/03/19	15:27:38	00:00:06
	Poisson	-0,19	139	81	23/03/19	15:28:59	00:01:21
	FastForest	0,57	83	57	23/03/19	15:29:08	00:00:09
	GeneralizedAdditiveModel	0,55	85	57	23/03/19	15:30:24	00:01:16
	OnlineGradientDescent	0,14	118	95	23/03/19	15:30:30	00:00:06
dataWD1.csv	FastTree	0,63	17	4	23/03/19	15:30:39	00:00:09
	FastTreeTweedie	0,61	18	3	23/03/19	15:30:45	00:00:06
	SDCA	0,03	28	8	23/03/19	15:30:49	00:00:04
	Poisson	-0,01	28	5	23/03/19	15:31:01	00:00:12
	FastForest	0,36	23	6	23/03/19	15:31:06	00:00:05
	GeneralizedAdditiveModel	0,10	27	9	23/03/19	15:31:24	00:00:18
	OnlineGradientDescent	0,04	28	9	23/03/19	15:31:27	00:00:03
dataWD0h.csv	FastTree	0,11	15	3	23/03/19	15:31:32	00:00:05
	FastTreeTweedie	-0,05	17	2	23/03/19	15:31:34	00:00:02
	SDCA	0,01	16	5	23/03/19	15:31:35	00:00:01
	Poisson	-0,01	16	2	23/03/19	15:31:38	00:00:03
	FastForest	0,14	15	4	23/03/19	15:31:41	00:00:03
	GeneralizedAdditiveModel	0,02	16	5	23/03/19	15:31:44	00:00:03
	OnlineGradientDescent	0,01	16	5	23/03/19	15:31:45	00:00:01
dataWD1h.csv	FastTree	0,46	3	0	23/03/19	15:31:48	00:00:03
	FastTreeTweedie	0,43	3	0	23/03/19	15:31:49	00:00:01
	SDCA	0,02	3	1	23/03/19	15:31:50	00:00:01
	Poisson	0,00	3	0	23/03/19	15:31:51	00:00:01
	FastForest	0,40	3	1	23/03/19	15:31:52	00:00:01
	GeneralizedAdditiveModel	0,20	3	1	23/03/19	15:31:55	00:00:03
	OnlineGradientDescent	0,01	3	1	23/03/19	15:31:56	00:00:01
datah.csv	FastTree	0,25	22	4	23/03/19	15:32:02	00:00:06
	FastTreeTweedie	0,32	21	3	23/03/19	15:32:05	00:00:03
	SDCA	0,01	25	5	23/03/19	15:32:07	00:00:02
	Poisson	-0,01	25	3	23/03/19	15:32:10	00:00:03
	FastForest	0,24	22	5	23/03/19	15:32:14	00:00:04
	GeneralizedAdditiveModel	0,05	25	6	23/03/19	15:32:19	00:00:05
	OnlineGradientDescent	0,01	25	5	23/03/19	15:32:20	00:00:01

Tabela A4.16 - Dados da avaliação dos modelos gerados para a Zona1 [Sockets]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data.csv	FastTree	0,64	83	58	23/03/19	15:29:51	00:00:15
	FastTreeTweedie	0,64	83	58	23/03/19	15:30:06	00:00:15
	SDCA	0,11	131	94	23/03/19	15:30:13	00:00:07
	Poisson	0,13	129	90	23/03/19	15:30:23	00:00:10
	FastForest	0,47	100	70	23/03/19	15:30:36	00:00:13
	GeneralizedAdditiveModel	0,40	107	78	23/03/19	15:32:01	00:01:25
	OnlineGradientDescent	0,15	128	92	23/03/19	15:32:07	00:00:06
dataLM.csv	FastTree	0,89	24	15	23/03/19	15:32:13	00:00:06
	FastTreeTweedie	0,90	24	15	23/03/19	15:32:17	00:00:04
	SDCA	0,22	65	43	23/03/19	15:32:18	00:00:01
	Poisson	0,22	66	42	23/03/19	15:32:20	00:00:02
	FastForest	0,64	45	28	23/03/19	15:32:23	00:00:03
	GeneralizedAdditiveModel	0,45	55	38	23/03/19	15:32:27	00:00:04
	OnlineGradientDescent	0,09	71	53	23/03/19	15:32:28	00:00:01
data2LM.csv	FastTree	0,79	56	39	23/03/19	15:32:36	00:00:08
	FastTreeTweedie	0,79	56	39	23/03/19	15:32:42	00:00:06
	SDCA	0,55	83	65	23/03/19	15:32:45	00:00:03
	Poisson	0,54	83	64	23/03/19	15:32:50	00:00:05
	FastForest	0,64	74	59	23/03/19	15:32:56	00:00:06
	GeneralizedAdditiveModel	0,61	77	59	23/03/19	15:33:06	00:00:10
	OnlineGradientDescent	0,54	84	65	23/03/19	15:33:08	00:00:02
data3LM.csv	FastTree	0,81	54	35	23/03/19	15:33:16	00:00:08
	FastTreeTweedie	0,81	54	36	23/03/19	15:33:23	00:00:07
	SDCA	0,27	106	87	23/03/19	15:33:26	00:00:03
	Poisson	0,26	107	84	23/03/19	15:33:31	00:00:05
	FastForest	0,67	71	54	23/03/19	15:33:38	00:00:07
	GeneralizedAdditiveModel	0,64	74	56	23/03/19	15:33:52	00:00:14
	OnlineGradientDescent	0,25	107	88	23/03/19	15:33:54	00:00:02
dataWD0.csv	FastTree	0,63	92	64	23/03/19	14:59:39	00:00:13
	FastTreeTweedie	0,63	92	64	23/03/19	14:59:51	00:00:12
	SDCA	0,11	143	105	23/03/19	14:59:57	00:00:06
	Poisson	0,08	145	102	23/03/19	15:00:09	00:00:12
	FastForest	0,49	108	76	23/03/19	15:00:20	00:00:11
	GeneralizedAdditiveModel	0,46	111	78	23/03/19	15:01:24	00:01:04
	OnlineGradientDescent	0,11	143	105	23/03/19	15:01:29	00:00:05
dataWD1.csv	FastTree	0,61	48	34	23/03/19	15:01:39	00:00:10
	FastTreeTweedie	0,62	48	34	23/03/19	15:01:45	00:00:06
	SDCA	0,01	77	57	23/03/19	15:01:49	00:00:04
	Poisson	0,00	77	58	23/03/19	15:01:55	00:00:06
	FastForest	0,32	64	46	23/03/19	15:02:01	00:00:06
	GeneralizedAdditiveModel	0,22	68	49	23/03/19	15:02:18	00:00:17
	OnlineGradientDescent	0,00	77	58	23/03/19	15:02:22	00:00:04
dataWD0h.csv	FastTree	0,63	44	32	23/03/19	15:02:27	00:00:05
	FastTreeTweedie	0,61	45	33	23/03/19	15:02:30	00:00:03
	SDCA	0,03	71	54	23/03/19	15:02:32	00:00:02
	Poisson	-0,01	73	55	23/03/19	15:02:33	00:00:01
	FastForest	0,37	57	42	23/03/19	15:02:36	00:00:03
	GeneralizedAdditiveModel	0,26	63	46	23/03/19	15:02:40	00:00:04
	OnlineGradientDescent	-0,14	77	58	23/03/19	15:02:41	00:00:01
dataWD1h.csv	FastTree	0,83	30	20	23/03/19	15:02:44	00:00:03
	FastTreeTweedie	0,83	30	20	23/03/19	15:02:45	00:00:01
	SDCA	0,03	72	54	23/03/19	15:02:47	00:00:02
	Poisson	-0,01	74	55	23/03/19	15:02:48	00:00:01
	FastForest	0,56	49	38	23/03/19	15:02:49	00:00:01
	GeneralizedAdditiveModel	0,31	61	45	23/03/19	15:02:51	00:00:02
	OnlineGradientDescent	-0,18	80	60	23/03/19	15:02:52	00:00:01
datah.csv	FastTree	0,62	48	34	23/03/19	15:02:58	00:00:06
	FastTreeTweedie	0,62	48	34	23/03/19	15:03:02	00:00:04
	SDCA	0,03	77	57	23/03/19	15:03:05	00:00:03
	Poisson	0,00	78	59	23/03/19	15:03:07	00:00:02
	FastForest	0,34	63	46	23/03/19	15:03:12	00:00:05
	GeneralizedAdditiveModel	0,25	68	50	23/03/19	15:03:16	00:00:04
	OnlineGradientDescent	-0,04	79	60	23/03/19	15:03:18	00:00:02

Tabela A4.17 - Dados da avaliação dos modelos gerados para a Zona1 [Total]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data.csv	FastTree	0,56	224	124	23/03/19	15:48:27	00:00:15
	FastTreeTweedie	0,56	225	123	23/03/19	15:48:42	00:00:15
	SDCA	0,16	310	215	23/03/19	15:48:49	00:00:07
	Poisson	0,14	314	196	23/03/19	15:49:23	00:00:34
	FastForest	0,44	253	153	23/03/19	15:49:35	00:00:12
	GeneralizedAdditiveModel	0,36	270	178	23/03/19	15:51:03	00:01:28
	OnlineGradientDescent	0,16	309	213	23/03/19	15:51:09	00:00:06
dataLM.csv	FastTree	0,94	43	21	23/03/19	15:51:15	00:00:06
	FastTreeTweedie	0,94	43	22	23/03/19	15:51:19	00:00:04
	SDCA	0,37	136	97	23/03/19	15:51:21	00:00:02
	Poisson	0,36	137	91	23/03/19	15:51:22	00:00:01
	FastForest	0,79	78	48	23/03/19	15:51:25	00:00:03
	GeneralizedAdditiveModel	0,58	111	78	23/03/19	15:51:28	00:00:03
	OnlineGradientDescent	0,32	141	106	23/03/19	15:51:29	00:00:01
data2LM.csv	FastTree	0,46	240	147	23/03/19	15:51:37	00:00:08
	FastTreeTweedie	0,46	241	147	23/03/19	15:51:43	00:00:06
	SDCA	0,07	316	217	23/03/19	15:51:46	00:00:03
	Poisson	0,03	322	203	23/03/19	15:51:51	00:00:05
	FastForest	0,33	268	181	23/03/19	15:51:57	00:00:06
	GeneralizedAdditiveModel	0,24	285	196	23/03/19	15:52:07	00:00:10
	OnlineGradientDescent	0,06	316	219	23/03/19	15:52:09	00:00:02
data3LM.csv	FastTree	0,45	255	150	23/03/19	15:52:17	00:00:08
	FastTreeTweedie	0,44	257	149	23/03/19	15:52:25	00:00:08
	SDCA	0,14	317	217	23/03/19	15:52:28	00:00:03
	Poisson	0,12	320	200	23/03/19	15:52:34	00:00:06
	FastForest	0,32	282	181	23/03/19	15:52:40	00:00:06
	GeneralizedAdditiveModel	0,25	297	198	23/03/19	15:52:53	00:00:13
	OnlineGradientDescent	0,14	318	220	23/03/19	15:52:56	00:00:03
dataWD0.csv	FastTree	0,56	241	143	22/03/19	03:55:28	00:00:12
	FastTreeTweedie	0,56	241	142	22/03/19	03:55:40	00:00:12
	SDCA	0,14	336	244	22/03/19	03:55:46	00:00:06
	Poisson	0,12	339	229	22/03/19	03:56:02	00:00:16
	FastForest	0,46	266	166	22/03/19	03:56:11	00:00:09
	GeneralizedAdditiveModel	0,45	270	169	22/03/19	03:57:11	00:01:00
	OnlineGradientDescent	0,14	336	242	22/03/19	03:57:17	00:00:06
dataWD1.csv	FastTree	0,28	146	66	22/03/19	03:57:25	00:00:08
	FastTreeTweedie	0,28	146	65	22/03/19	03:57:32	00:00:07
	SDCA	0,02	169	85	22/03/19	03:57:35	00:00:03
	Poisson	0,00	171	81	22/03/19	03:57:43	00:00:08
	FastForest	0,11	161	76	22/03/19	03:57:48	00:00:05
	GeneralizedAdditiveModel	0,06	165	80	22/03/19	03:58:11	00:00:23
	OnlineGradientDescent	0,02	169	85	22/03/19	03:58:14	00:00:03
dataWD1h.csv	FastTree	0,33	172	67	22/03/19	03:58:17	00:00:03
	FastTreeTweedie	0,33	172	65	22/03/19	03:58:18	00:00:01
	SDCA	0,03	208	92	22/03/19	03:58:19	00:00:01
	Poisson	-0,03	213	92	22/03/19	03:58:20	00:00:01
	FastForest	0,23	185	76	22/03/19	03:58:22	00:00:02
	GeneralizedAdditiveModel	0,15	194	83	22/03/19	03:58:24	00:00:02
	OnlineGradientDescent	-0,07	217	97	22/03/19	03:58:25	00:00:01
dataWD0h.csv	FastTree	0,13	173	78	22/03/19	03:58:30	00:00:05
	FastTreeTweedie	0,12	174	77	22/03/19	03:58:33	00:00:03
	SDCA	0,06	180	92	22/03/19	03:58:35	00:00:02
	Poisson	0,01	184	86	22/03/19	03:58:36	00:00:01
	FastForest	0,11	175	82	22/03/19	03:58:38	00:00:02
	GeneralizedAdditiveModel	0,10	176	85	22/03/19	03:58:42	00:00:04
	OnlineGradientDescent	0,01	185	90	22/03/19	03:58:44	00:00:02
datah.csv	FastTree	0,13	162	76	22/03/19	03:58:49	00:00:05
	FastTreeTweedie	0,15	161	74	22/03/19	03:58:53	00:00:04
	SDCA	0,04	171	89	22/03/19	03:58:54	00:00:01
	Poisson	0,01	174	83	22/03/19	03:58:56	00:00:02
	FastForest	0,10	165	81	22/03/19	03:59:00	00:00:04
	GeneralizedAdditiveModel	0,08	168	83	22/03/19	03:59:06	00:00:06
	OnlineGradientDescent	0,01	172	86	22/03/19	03:59:07	00:00:01

Tabela A4.18 - Dados da avaliação dos modelos gerados para a ZonaBld [Hvac]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data	FastTree	0,26	937	755	23/03/19	18:57:11	00:00:16
	FastTreeTweedie	0,26	937	753	23/03/19	18:57:27	00:00:16
	SDCA	0,06	1054	851	23/03/19	18:57:34	00:00:07
	Poisson	-0,10	1142	779	23/03/19	18:58:42	00:01:08
	FastForest	0,19	975	800	23/03/19	18:58:53	00:00:11
	GeneralizedAdditiveModel	0,16	998	796	23/03/19	19:01:35	00:02:42
	OnlineGradientDescent	0,06	1053	850	23/03/19	19:01:41	00:00:06
dataLM	FastTree	0,34	836	644	23/03/19	19:01:50	00:00:09
	FastTreeTweedie	0,33	842	638	23/03/19	19:01:56	00:00:06
	SDCA	0,06	997	785	23/03/19	19:01:58	00:00:02
	Poisson	-0,12	1089	731	23/03/19	19:02:06	00:00:08
	FastForest	0,24	896	716	23/03/19	19:02:13	00:00:07
	GeneralizedAdditiveModel	0,19	929	726	23/03/19	19:02:20	00:00:07
	OnlineGradientDescent	0,06	998	784	23/03/19	19:02:23	00:00:03
data2LM	FastTree	0,29	960	777	23/03/19	19:02:34	00:00:11
	FastTreeTweedie	0,29	960	775	23/03/19	19:02:47	00:00:13
	SDCA	0,05	1107	898	23/03/19	19:02:52	00:00:05
	Poisson	-0,10	1190	829	23/03/19	19:03:24	00:00:32
	FastForest	0,23	998	822	23/03/19	19:03:33	00:00:09
	GeneralizedAdditiveModel	0,21	1013	813	23/03/19	19:04:51	00:01:18
	OnlineGradientDescent	0,05	1107	897	23/03/19	19:04:56	00:00:05
data3LM	FastTree	0,09	828	666	23/03/19	19:05:06	00:00:10
	FastTreeTweedie	0,09	829	664	23/03/19	19:05:13	00:00:07
	SDCA	0,04	853	710	23/03/19	19:05:16	00:00:03
	Poisson	-0,13	923	615	23/03/19	19:05:30	00:00:14
	FastForest	0,07	838	684	23/03/19	19:05:36	00:00:06
	GeneralizedAdditiveModel	0,06	843	687	23/03/19	19:05:54	00:00:18
	OnlineGradientDescent	0,03	854	711	23/03/19	19:05:57	00:00:03

Tabela A4.19 - Dados da avaliação dos modelos gerados para a ZonaBld [Lights]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data	FastTree	0,84	198	102	23/03/19	21:56:54	00:00:15
	FastTreeTweedie	0,82	205	102	23/03/19	21:57:09	00:00:15
	SDCA	0,23	430	331	23/03/19	21:57:16	00:00:07
	Poisson	-0,10	514	276	23/03/19	21:59:26	00:02:10
	FastForest	0,71	266	182	23/03/19	21:59:39	00:00:13
	GeneralizedAdditiveModel	0,60	310	244	23/03/19	22:02:07	00:02:28
	OnlineGradientDescent	0,23	429	328	23/03/19	22:02:14	00:00:07
dataLM	FastTree	0,93	150	74	23/03/19	22:02:23	00:00:09
	FastTreeTweedie	0,93	155	69	23/03/19	22:02:30	00:00:07
	SDCA	0,41	448	349	23/03/19	22:02:33	00:00:03
	Poisson	-1,09	844	371	23/03/19	22:02:44	00:00:11
	FastForest	0,80	259	181	23/03/19	22:02:50	00:00:06
	GeneralizedAdditiveModel	0,64	349	281	23/03/19	22:02:58	00:00:08
	OnlineGradientDescent	0,39	457	360	23/03/19	22:03:00	00:00:02
data2LM	FastTree	0,82	229	131	23/03/19	22:03:12	00:00:12
	FastTreeTweedie	0,81	236	132	23/03/19	22:03:24	00:00:12
	SDCA	0,17	494	393	23/03/19	22:03:30	00:00:06
	Poisson	-0,24	603	380	23/03/19	22:04:54	00:01:24
	FastForest	0,73	279	197	23/03/19	22:05:03	00:00:09
	GeneralizedAdditiveModel	0,74	274	180	23/03/19	22:06:28	00:01:25
	OnlineGradientDescent	0,17	494	395	23/03/19	22:06:33	00:00:05
data3LM	FastTree	0,55	49	17	23/03/19	22:06:43	00:00:10
	FastTreeTweedie	0,52	51	15	23/03/19	22:06:50	00:00:07
	SDCA	0,04	72	31	23/03/19	22:06:54	00:00:04
	Poisson	-0,04	75	18	23/03/19	22:07:13	00:00:19
	FastForest	0,24	64	27	23/03/19	22:07:19	00:00:06
	GeneralizedAdditiveModel	0,13	69	31	23/03/19	22:07:39	00:00:20
	OnlineGradientDescent	0,04	72	31	23/03/19	22:07:42	00:00:03

Tabela A4.20 - Dados da avaliação dos modelos gerados para a ZonaBld [Sockets]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data	FastTree	0,60	430	250	23/03/19	21:44:16	00:00:15
	FastTreeTweedie	0,59	432	250	23/03/19	21:44:31	00:00:15
	SDCA	0,13	634	423	23/03/19	21:44:37	00:00:06
	Poisson	0,11	638	390	23/03/19	21:44:49	00:00:12
	FastForest	0,47	492	293	23/03/19	21:45:01	00:00:12
	GeneralizedAdditiveModel	0,39	528	343	23/03/19	21:47:26	00:02:25
	OnlineGradientDescent	0,14	630	411	23/03/19	21:47:32	00:00:06
dataLM	FastTree	0,74	274	156	23/03/19	21:47:41	00:00:09
	FastTreeTweedie	0,74	272	154	23/03/19	21:47:48	00:00:07
	SDCA	0,21	477	328	23/03/19	21:47:51	00:00:03
	Poisson	0,19	481	314	23/03/19	21:47:55	00:00:04
	FastForest	0,60	340	204	23/03/19	21:48:01	00:00:06
	GeneralizedAdditiveModel	0,47	390	269	23/03/19	21:48:09	00:00:08
	OnlineGradientDescent	0,20	478	325	23/03/19	21:48:12	00:00:03
data2LM	FastTree	0,57	488	291	23/03/19	21:48:23	00:00:11
	FastTreeTweedie	0,57	489	291	23/03/19	21:48:36	00:00:13
	SDCA	0,08	717	479	23/03/19	21:48:41	00:00:05
	Poisson	0,05	729	457	23/03/19	21:48:52	00:00:11
	FastForest	0,48	538	330	23/03/19	21:49:02	00:00:10
	GeneralizedAdditiveModel	0,47	541	334	23/03/19	21:50:37	00:01:35
	OnlineGradientDescent	0,08	717	479	23/03/19	21:50:42	00:00:05
data3LM	FastTree	0,53	187	127	23/03/19	21:50:52	00:00:10
	FastTreeTweedie	0,52	189	128	23/03/19	21:50:59	00:00:07
	SDCA	0,19	246	188	23/03/19	21:51:03	00:00:04
	Poisson	0,18	248	187	23/03/19	21:51:09	00:00:06
	FastForest	0,35	221	157	23/03/19	21:51:14	00:00:05
	GeneralizedAdditiveModel	0,36	220	156	23/03/19	21:51:31	00:00:17
	OnlineGradientDescent	0,19	247	188	23/03/19	21:51:34	00:00:03

Tabela A4.21 - Dados da avaliação dos modelos gerados para a ZonaBld [Total]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data	FastTree	0,85	464	216	23/03/19	22:09:22	00:00:16
	FastTreeTweedie	0,85	471	211	23/03/19	22:09:38	00:00:16
	SDCA	0,19	1089	774	23/03/19	22:09:44	00:00:06
	Poisson	-0,11	1272	647	23/03/19	22:13:55	00:04:11
	FastForest	0,70	663	418	23/03/19	22:14:07	00:00:12
	GeneralizedAdditiveModel	0,68	679	458	23/03/19	22:16:23	00:02:16
	OnlineGradientDescent	0,19	1086	760	23/03/19	22:16:30	00:00:07
dataLM	FastTree	0,93	240	109	23/03/19	22:16:39	00:00:09
	FastTreeTweedie	0,89	308	143	23/03/19	22:16:44	00:00:05
	SDCA	0,23	814	615	23/03/19	22:16:47	00:00:03
	Poisson	-0,09	972	513	23/03/19	22:16:58	00:00:11
	FastForest	0,80	415	276	23/03/19	22:17:04	00:00:06
	GeneralizedAdditiveModel	0,85	360	224	23/03/19	22:17:12	00:00:08
	OnlineGradientDescent	0,20	831	624	23/03/19	22:17:14	00:00:02
data2LM	FastTree	0,84	477	219	23/03/19	22:17:25	00:00:11
	FastTreeTweedie	0,83	482	213	23/03/19	22:17:38	00:00:13
	SDCA	0,19	1063	744	23/03/19	22:17:43	00:00:05
	Poisson	-0,09	1239	629	23/03/19	22:19:41	00:01:58
	FastForest	0,72	629	384	23/03/19	22:19:51	00:00:10
	GeneralizedAdditiveModel	0,68	673	447	23/03/19	22:21:37	00:01:46
	OnlineGradientDescent	0,19	1065	740	23/03/19	22:21:42	00:00:05
data3LM	FastTree	0,89	402	185	23/03/19	22:21:52	00:00:10
	FastTreeTweedie	0,86	464	216	23/03/19	22:21:57	00:00:05
	SDCA	0,20	1111	782	23/03/19	22:22:01	00:00:04
	Poisson	-0,16	1333	686	23/03/19	22:22:28	00:00:27
	FastForest	0,75	617	389	23/03/19	22:22:34	00:00:06
	GeneralizedAdditiveModel	0,71	662	459	23/03/19	22:22:54	00:00:20
	OnlineGradientDescent	0,18	1119	793	23/03/19	22:22:57	00:00:03

Tabela A4.22 - Dados da avaliação dos modelos gerados para a Zona9 [PV]

DataSet	Algiritmo	R2	RMS	L1	Data	Hora	Tempo execução
data	FastTree	0,85	462	217	23/03/19	23:14:08	00:00:19
data	FastTreeTweedie	0,85	471	211	23/03/19	23:14:27	00:00:19
data	SDCA	0,18	1093	740	23/03/19	23:14:36	00:00:09
data	Poisson	-0,11	1271	647	23/03/19	23:19:02	00:04:26
data	FastForest	0,71	655	413	23/03/19	23:19:17	00:00:15
data	GeneralizedAdditiveModel	0,68	679	458	23/03/19	23:21:17	00:02:00
data	OnlineGradientDescent	0,19	1086	761	23/03/19	23:21:25	00:00:08
dataLM	FastTree	0,94	236	109	23/03/19	23:21:36	00:00:11
dataLM	FastTreeTweedie	0,90	300	139	23/03/19	23:21:41	00:00:05
dataLM	SDCA	0,23	814	615	23/03/19	23:21:44	00:00:03
dataLM	Poisson	-0,09	970	513	23/03/19	23:21:56	00:00:12
dataLM	FastForest	0,79	422	283	23/03/19	23:22:02	00:00:06
dataLM	GeneralizedAdditiveModel	0,85	360	224	23/03/19	23:22:14	00:00:12
dataLM	OnlineGradientDescent	0,20	831	624	23/03/19	23:22:16	00:00:02
data2LM	FastTree	0,84	475	218	23/03/19	23:22:29	00:00:13
data2LM	FastTreeTweedie	0,83	482	214	23/03/19	23:22:43	00:00:14
data2LM	SDCA	0,19	1063	742	23/03/19	23:22:51	00:00:08
data2LM	Poisson	-0,09	1237	626	23/03/19	23:24:32	00:01:41
data2LM	FastForest	0,73	609	369	23/03/19	23:24:44	00:00:12
data2LM	GeneralizedAdditiveModel	0,68	673	447	23/03/19	23:26:23	00:01:39
data2LM	OnlineGradientDescent	0,19	1065	740	23/03/19	23:26:29	00:00:06
data3LM	FastTree	0,90	400	185	23/03/19	23:26:40	00:00:11
data3LM	FastTreeTweedie	0,86	458	212	23/03/19	23:26:46	00:00:06
data3LM	SDCA	0,20	1111	786	23/03/19	23:26:50	00:00:04
data3LM	Poisson	-0,17	1338	687	23/03/19	23:27:27	00:00:37
data3LM	FastForest	0,77	590	366	23/03/19	23:27:34	00:00:07
data3LM	GeneralizedAdditiveModel	0,71	662	459	23/03/19	23:28:01	00:00:27
data3LM	OnlineGradientDescent	0,18	1119	793	23/03/19	23:28:05	00:00:04

Anexo 5 Ficheiros docker-compose da stack HDS

Ficheiro docker-compose: componetes bese e *agentes:agentes*

```
version: "3.3"

services:
  mosquito:
    image: hds_mosquitto
    networks:
      - hds-net
    ports:
      - 1883:1883
      - 9001:9001
    environment:
      - TZ=UTC-1
    hostname: mosquitto
    deploy:
      restart_policy:
        condition: on-failure
      placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

  telegraf:
    image: telegraf
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    hostname: mytelegraf
    depends_on:
      - mosquito
      - influxdb
    volumes:
      - /var/run/docker.sock:/var/run/docker.sock
    configs:
      - source: telegraf-config
        target: /etc/telegraf/telegraf.conf
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == M100-DELL_T7610-wrk #HP_manager

  influxdb:
    image: influxdb
```

```
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - data:/var/lib/influxdb
    deploy:
      restart_policy:
        condition: on-failure
      placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

  grafana:
    image: ev/grafana
    networks:
      - hds-net
    ports:
      - 3000:3000
    environment:
      - GF_SECURITY_ADMIN_USER=${ADMIN_USER:-admin}
      - GF_SECURITY_ADMIN_PASSWORD=${ADMIN_PASSWORD:-admin}
      - GF_USERS_ALLOW_SIGN_UP=false
      - TZ=Europe/Lisbon
    volumes:
      - grafana:/var/lib/grafana
    deploy:
      restart_policy:
       condition: on-failure
      placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

  agente_z1:
    image: euvinagre/agente:agente4
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    hostname: agente_z1
    depends_on:
      - telegraf
    deploy:
      restart_policy:
       condition: on-failure
      placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

  agente_z2:
    image: euvinagre/agente:agente4
```

```
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    hostname: agente_z2
    depends_on:
      - telegraf
    deploy:
      restart_policy:
       condition: on-failure
      placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

  agente_z3:
    image: euvinagre/agente:agente4
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    hostname: agente_z3
    depends_on:
      - telegraf
    deploy:
      restart_policy:
        condition: on-failure

  agente_z4:
    image: euvinagre/agente:agente4
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    hostname: agente_z4
    depends_on:
      - telegraf
    deploy:
      restart_policy:
        condition: on-failure

  agente_z7:
    image: euvinagre/agente:agente4
    networks:
      - hds-net
    environment:
      - TZ=Europe/Lisbon
    hostname: agente_z7
    depends_on:
      - telegraf
    deploy:
      restart_policy:
```

```
      condition: on-failure

agente_z8a:
  image: euvinagre/agente:agente4
  networks:
    - hds-net
  environment:
    - TZ=Europe/Lisbon
  hostname: agente_z8a
  depends_on:
    - telegraf
  deploy:
    restart_policy:
      condition: on-failure

agente_z8b:
  image: euvinagre/agente:agente4
  networks:
    - hds-net
  environment:
    - TZ=Europe/Lisbon
  hostname: agente_z8b
  depends_on:
    - telegraf
  deploy:
    restart_policy:
      condition: on-failure

agente_z9:
  image: euvinagre/agente:agente4
  networks:
    - hds-net
  environment:
    - TZ=Europe/Lisbon
  hostname: agente_z9
  depends_on:
    - telegraf
  deploy:
    restart_policy:
      condition: on-failure

agente_bld:
  image: euvinagre/agente:agente4
  networks:
    - hds-net
  environment:
    - TZ=Europe/Lisbon
  hostname: agente_bld
  depends_on:
    - telegraf
  deploy:
```

```
      restart_policy:
        condition: on-failure


configs:
  telegraf-config:
    file: $PWD/conf/telegraf/telegraf.conf

networks:
  hds-net:
    driver: overlay

volumes:
    grafana: {}
    data: {}
```

Ficheiro docker-compose: *agentes:forecast*

```
version: "3.3"

services:

  agente_forecast_z1:
    image: euvinagre/agente:forecast8
    networks:
      - hds1-net
    hostname: AG_Forecast_Z1
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == worker101

  agente_forecast_z2:
    image: euvinagre/agente:forecast8
    networks:
      - hds1-net
    hostname: AG_Forecast_Z2
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
```

```
    placement:
        constraints:
          - node.hostname == worker102

agente_forecast_z3:
  image: euvinagre/agente:forecast8
  networks:
    - hds1-net
  hostname: AG_Forecast_Z3
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == worker103

agente_forecast_z4:
  image: euvinagre/agente:forecast8
  networks:
    - hds1-net
  hostname: AG_Forecast_Z4
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == worker104

agente_forecast_z7:
  image: euvinagre/agente:forecast8
  networks:
    - hds1-net
  hostname: AG_Forecast_Z7
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk
```

```
agente_forecast_z8a:
  image: euvinagre/agente:forecast8
  networks:
    - hds1-net
  hostname: AG_Forecast_Z8a
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == M200-HP_Z400-mng

agente_forecast_z8b:
  image: euvinagre/agente:forecast8
  networks:
    - hds1-net
  hostname: AG_Forecast_Z8b
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == worker201

agente_forecast_z9:
  image: euvinagre/agente:forecast8
  networks:
    - hds1-net
  hostname: AG_Forecast_Z9
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

agente_forecast_bld:
  image: euvinagre/agente:forecast8
```

```
    networks:
      - hds1-net
    hostname: AG_Forecast_Bld
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == M100-DELL_T7610-wrk #HP_manager

networks:
  hds1-net:
    driver: overlay

volumes:
  Models:
    external: true
```

Ficheiro docker-compose: *agente:inspetores*

```
version: "3.3"

services:

  agente_inspect_z1:
    image: agente:inspector
    networks:
      - hds1-net
    hostname: AG_Inspct_Z1
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == worker101

  agente_inspect_z2:
    image: agente:inspector
    networks:
      - hds1-net
    hostname: AG_Inspct_Z2
    environment:
      - TZ=Europe/Lisbon
```

```
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == worker102

  agente_inspect_z3:
    image: agente:inspector
    networks:
      - hds1-net
    hostname: AG_Inspct_Z3
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == worker103

  agente_inspect_z4:
    image: agente:inspector
    networks:
      - hds1-net
    hostname: AG_Inspct_Z4
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == worker104

  agente_inspect_z7:
    image: agente:inspector
    networks:
      - hds1-net
    hostname: AG_Inspct_Z7
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
```

```
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == M100-DELL_T7610-wrk #HP_manager

agente_inspect_z8a:
  image: agente:inspector
  networks:
    - hds1-net
  hostname: AG_Inspct_Z8a
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models/:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == M200-HP_Z400-mng

agente_inspect_z8b:
  image: agente:inspector
  networks:
    - hds1-net
  hostname: AG_Inspct_Z8b
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models/:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
        constraints:
          - node.hostname == worker201

agente_inspect_z9:
  image: agente:inspector
  networks:
    - hds1-net
  hostname: AG_Inspct_Z9
  environment:
    - TZ=Europe/Lisbon
  volumes:
    - Models/:/app/Models
  deploy:
    restart_policy:
      condition: on-failure
    placement:
```

```
          constraints:
            - node.hostname == M100-DELL_T7610-wrk #HP_manager

  agente_inspect_bld:
    image: agente:inspector
    networks:
      - hds1-net
    hostname: AG_Inspct_Bld
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == M100-DELL_T7610-wrk #HP_manager

networks:
  hds1-net:
    driver: overlay

volumes:
  Models:
    external: true
```

Ficheiro docker-compose: ***agentes:forecast (simulação e validação da geração de modelos)***

```
version: "3.3"

services:

  agente_forecast_z11:
    image: euvinagre/agente:forecast9
    hostname: AG_Forecast_Z11
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == M100-DELL_T7610-wrk           #HP_manager

  agente_forecast_z91:
    image: euvinagre/agente:forecast9
    hostname: AG_Forecast_Z91
    environment:
```

```
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == M100-DELL_T7610-wrk #HP_manager

  agente_forecast_bld1:
    image: euvinagre/agente:forecast9
    hostname: AG_Forecast_Bld1
    environment:
      - TZ=Europe/Lisbon
    volumes:
      - Models/:/app/Models
    deploy:
      restart_policy:
        condition: on-failure
      placement:
          constraints:
            - node.hostname == M100-DELL_T7610-wrk #HP_manager

volumes:
  Models:
    external: true
```