Integración de Nextflow y AWS para el Análisis Genómico a Gran Escala: Un Estudio de Caso hipotético

Ángel Canal-Alonso[1], Pedro Jiménez[1] and Noelia Egido[1], Javier Prieto[1], Juan Manuel Corchado[1]

[1] Departamento de Bioinformática y Biología Computacional, AIR Institute, Carbajosa de la Sagrada, España

E-mail: acanal@air-institute.com

Resumen

El presente artículo explora la combinación innovadora de Nextflow y Amazon Web Services (AWS) para enfrentar los desafíos inherentes al análisis genómico a gran escala. Centrándose en un caso hipotético denominado "El Atlas Genómico del Pacífico", se ilustra cómo una organización de investigación podría abordar la secuenciación y análisis de 10,000 genomas. Aunque el "Atlas Genómico del Pacífico" es un ejemplo ficticio utilizado únicamente con fines ilustrativos, destaca los desafíos reales asociados con proyectos genómicos de gran envergadura, como el manejo de enormes volúmenes de datos y la necesidad de análisis computacional intensivo. A través de la integración de Nextflow, una herramienta de gestión de workflows, con la infraestructura en la nube de AWS, se demuestra cómo se pueden superar estos desafíos, ofreciendo soluciones escalables, flexibles y coste-efectivas para la investigación genómica. La adopción de tecnologías modernas, como las descritas en este artículo, es esencial para avanzar en el campo de la genómica y acelerar descubrimientos científicos.

Palabras Clave: Next-Generation sequencing, Cloud computing, Distributed Computing

---

Introducción

Secuenciación de ADN

La secuenciación del ADN es un proceso que determina el orden exacto de los nucleótidos (adenina, timina, guanina y citosina) en una molécula de ADN. Desde su invención en la década de 1970, la secuenciación del ADN ha revolucionado el campo de la genómica y ha proporcionado información inestimable sobre la genética de numerosos organismos, incluidos los seres humanos.

Historia y Desarrollo

El desarrollo inicial de la secuenciación del ADN se basó en técnicas manuales que eran laboriosas y requerían mucho tiempo. El método más famoso de esta era fue el método de terminación de cadena desarrollado por Frederick Sanger en 1975, que le valió un Premio Nobel. Este método, conocido como secuenciación de Sanger, fue la técnica estándar durante muchos años y se utilizó, por ejemplo, en el Proyecto Genoma Humano, que tenía como objetivo secuenciar todo el genoma humano.

Con el tiempo, las técnicas de secuenciación del ADN han evolucionado, dando paso a lo que se conoce como Secuenciación de Nueva Generación (NGS, por sus siglas en inglés). La NGS es capaz de secuenciar millones de fragmentos de ADN simultáneamente, lo que la hace mucho más rápida y eficiente que las técnicas anteriores.

Importancia de la Secuenciación del ADN

La secuenciación del ADN ha tenido un impacto profundo en la biología y la medicina. Ha permitido a los científicos descubrir miles de genes humanos, identificar variantes genéticas asociadas con enfermedades y trastornos, y

comprender mejor la evolución y diversidad de la vida en la Tierra.

En medicina, la secuenciación del ADN ha llevado al desarrollo de pruebas genéticas para enfermedades hereditarias, la identificación de nuevos objetivos terapéuticos y la personalización de tratamientos basados en el perfil genético de un individuo.

La secuenciación del ADN también tiene aplicaciones en ecología, forense, agricultura, y muchos otros campos, lo que demuestra su versatilidad y relevancia en la ciencia contemporánea.

Desafíos de la Secuenciación

A pesar de los avances, la secuenciación del ADN no está exenta de desafíos. El volumen masivo de datos generados por técnicas como la NGS requiere soluciones de almacenamiento y análisis de datos sofisticadas. Aquí es donde las técnicas de computación distribuida entran en juego, proporcionando herramientas y metodologías para manejar, procesar y analizar grandes conjuntos de datos de secuenciación de manera eficiente.

Secuenciación de nueva generacion

La Secuenciación de Nueva Generación (NGS), también conocida como secuenciación masiva en paralelo, ha revolucionado el campo de la genómica al permitir la secuenciación simultánea de millones de fragmentos de ADN. A diferencia de la secuenciación de Sanger, que secuencia un fragmento de ADN a la vez, la NGS puede analizar todo un genoma en un solo experimento. A continuación, se describe cómo funcionan estas técnicas:

Antes de la secuenciación, el ADN de interés (que puede ser el genoma completo o regiones específicas) se fragmenta en piezas más pequeñas. Estos fragmentos se ligan con adaptadores específicos que permiten la amplificación y la secuenciación posterior.

Una característica común en las técnicas de NGS es la amplificación de los fragmentos de ADN. Una de las técnicas más utilizadas es la amplificación en puente, donde los fragmentos de ADN se unen a una superficie sólida y se copian en múltiples iteraciones, generando un "cluster" de fragmentos idénticos.

A continuación, se describen algunas de las técnicas de secuenciación más populares:

- Secuenciación por Síntesis (SBS): Esta técnica, utilizada en las plataformas de Illumina, implica la incorporación secuencial de nucleótidos fluorescentemente marcados. Cada vez que se incorpora un nucleótido, se emite una señal fluorescente que se detecta y registra. La intensidad y el color de la fluorescencia determinan el nucleótido específico que se ha incorporado.

- Secuenciación Iónica: Utilizada en las plataformas de Ion Torrent, esta técnica detecta los iones liberados durante la incorporación de un nucleótido en la cadena de ADN en crecimiento. No se requieren marcadores fluorescentes; en su lugar, un sensor detecta el cambio en el pH causado por la liberación de un ion hidrógeno cada vez que se añade un nucleótido.

- Nanoporos: En esta técnica, utilizada en las plataformas de Oxford Nanopore, las moléculas de ADN pasan a través de nanoporos en una membrana. A medida que cada nucleótido pasa por el nanoporo, provoca cambios en la corriente eléctrica que se pueden detectar y traducir en una secuencia.

Dada la gran cantidad de datos generados en un solo experimento de NGS, el análisis de datos es un paso crítico. El proceso generalmente incluye:

- Alineación/mapeo: Los fragmentos de secuencia, conocidos como "lecturas", se alinean o mapean a una secuencia de referencia, como el genoma humano de referencia.
- Identificación de variantes: Una vez mapeadas las lecturas, se pueden identificar variantes genéticas, como polimorfismos de nucleótido único (SNPs) o inserciones y deleciones (indels).
- Análisis funcional: Se pueden utilizar bases de datos y herramientas bioinformáticas para interpretar el impacto potencial de las variantes identificadas.

Las técnicas de NGS han abierto nuevas posibilidades en la investigación genómica, permitiendo desde el estudio de la diversidad genética hasta la identificación de mutaciones asociadas a enfermedades. Sin embargo, el volumen masivo de datos generados requiere soluciones informáticas avanzadas, y aquí es donde la computación distribuida juega un papel crucial.

Computación distribuida

La computación distribuida es un campo de la informática que estudia los sistemas distribuidos y cómo se pueden diseñar y coordinar para alcanzar un objetivo común. Un sistema distribuido consiste en múltiples computadoras autónomas conectadas a través de una red, donde cada computadora tiene

sus propios recursos locales, como CPU, memoria y almacenamiento. A diferencia de un sistema centralizado donde todas las operaciones y recursos están concentrados en una única máquina o ubicación, en un sistema distribuido, las operaciones se reparten entre las diferentes máquinas, y estas trabajan juntas como un solo sistema cohesivo.

Uno de los principales beneficios de la computación distribuida es su capacidad para manejar grandes cantidades de datos y realizar cálculos complejos de manera eficiente. Al dividir una tarea en subprocesos más pequeños y distribuirlos entre varias máquinas, es posible procesar grandes volúmenes de datos en paralelo, acelerando significativamente el tiempo de procesamiento. Esta característica es especialmente valiosa en campos como la genómica, donde los conjuntos de datos son enormes y el análisis puede ser computacionalmente intensivo.

Otra ventaja es la escalabilidad. A medida que aumentan las demandas de cálculo o almacenamiento, se pueden agregar más máquinas al sistema distribuido, permitiendo que el sistema crezca y se adapte a las necesidades cambiantes. Esto contrasta con los sistemas centralizados, donde la escalabilidad puede ser limitada por la capacidad de la máquina central.

La tolerancia a fallos es otro beneficio clave. En un sistema distribuido, si una máquina falla, las otras pueden continuar trabajando. Los datos y las tareas se pueden replicar en múltiples máquinas, por lo que si una máquina se desconecta o falla, otra máquina puede asumir su trabajo, garantizando así la continuidad y la disponibilidad del sistema.

Sin embargo, la computación distribuida también presenta desafíos. La coordinación entre las máquinas, la garantía de consistencia de datos y la gestión de fallos son aspectos complejos que deben ser manejados cuidadosamente. Además, la comunicación entre las máquinas en una red puede introducir latencia, lo que puede afectar el rendimiento del sistema.

En el contexto de la secuenciación de nueva generación (NGS), la computación distribuida ha demostrado ser una herramienta valiosa para manejar y analizar grandes conjuntos de datos genómicos. Al distribuir el análisis a través de múltiples máquinas, es posible acelerar significativamente el tiempo de procesamiento, lo que permite a los investigadores obtener resultados más rápidamente y realizar análisis más complejos que antes no eran posibles.

En el mundo de la computación distribuida, existen diversas arquitecturas y modelos que se han desarrollado para abordar distintos tipos de problemas y necesidades. A continuación, se presentan algunas de las arquitecturas más comunes:

1. Sistemas Cliente-Servidor:
   - En esta arquitectura, hay una clara distinción entre los proveedores de servicios (servidores) y los solicitantes de servicios (clientes).
   - Los servidores mantienen y ofrecen servicios y recursos a los clientes, que a su vez solicitan y consumen estos servicios.
   - Ejemplo clásico: servidores web y navegadores.

2. Sistemas Peer-to-Peer (P2P):
   - En un sistema P2P, todos los nodos tienen capacidades similares y pueden actuar tanto como clientes como servidores.
   - Está diseñado para ser descentralizado, eliminando la necesidad de un servidor central.
   - Son ampliamente utilizados para compartir archivos, como en el caso de BitTorrent.

3. Arquitecturas basadas en Grillas (Grid Computing):
   - Diseñado para resolver problemas que requieren grandes cantidades de recursos computacionales, combinando recursos de múltiples ubicaciones y organizaciones.
   - Los nodos en una grilla no necesariamente tienen una relación constante o de largo plazo, y a menudo se agregan y retiran dinámicamente según sea necesario.
   - Ejemplo: Proyecto SETI@home para la búsqueda de inteligencia extraterrestre.

4. Clusters de Computadoras:
   - Un cluster es un grupo de computadoras conectadas entre sí, trabajando juntas para ofrecer alta disponibilidad, alta capacidad de cálculo o ambos.
   - A menudo, las máquinas en un cluster son similares y están ubicadas físicamente cerca unas de otras, conectadas por una red local de alta velocidad.
   - Ejemplo: Clusters de Hadoop para procesamiento de big data.

5. Computación en la Nube (Cloud Computing):
   - Proporciona recursos computacionales (CPU, memoria, almacenamiento) como un servicio a través de internet.
   - Los recursos pueden ser escalados dinámicamente según las necesidades.
   - Ejemplos: Servicios como Amazon Web Services (AWS), Google Cloud Platform y Microsoft Azure.

6. Sistemas basados en Volúmenes de Datos (Data-intensive Computing):

- Diseñado específicamente para problemas que requieren el procesamiento de grandes cantidades de datos.
- A menudo utiliza sistemas distribuidos de almacenamiento y procesamiento, como Hadoop y su sistema de archivos distribuidos (HDFS).

7. Arquitecturas basadas en Microservicios:
- Divide una aplicación en pequeños servicios independientes que se ejecutan como procesos separados y se comunican entre sí a través de mecanismos ligeros, generalmente HTTP.
- Cada microservicio se encarga de una función específica y puede desarrollarse, desplegarse y escalarse de manera independiente.

Estas arquitecturas pueden ser aplicadas en diferentes contextos según las necesidades específicas del problema a resolver. En el caso de la secuenciación de nueva generación y el análisis genómico, la capacidad de procesar y almacenar grandes cantidades de datos de manera eficiente es esencial, lo que hace que ciertas arquitecturas, como la computación en la nube y los sistemas basados en volúmenes de datos, sean particularmente relevantes.

La confluencia entre NGS y Computación Distribuida

La incorporación de la computación distribuida en la Secuenciación de Nueva Generación (NGS) es esencial para abordar los desafíos computacionales y de almacenamiento que presenta esta tecnología. Los datos generados por las plataformas de NGS son masivos, y el análisis de estos datos es computacionalmente intensivo. La computación distribuida ofrece soluciones para superar estos desafíos. A continuación, se describen las formas en que la computación distribuida puede ser aplicada en NGS:

1. Almacenamiento Distribuido:
- Las plataformas de NGS generan terabytes de datos en un solo experimento. Almacenar estos datos en una sola máquina o servidor no es práctico ni eficiente.
- Los sistemas de archivos distribuidos, como el Sistema de Archivos Distribuidos de Hadoop (HDFS), permiten almacenar grandes conjuntos de datos distribuyendo los datos en múltiples máquinas, proporcionando redundancia y alta disponibilidad.

2. Procesamiento Paralelo:
- Los algoritmos de análisis de NGS, como la alineación de lecturas y la identificación de variantes, son computacionalmente intensivos.
- Los frameworks de procesamiento distribuido, como Hadoop y Spark, permiten dividir estas tareas en subprocesos más pequeños que se ejecutan en paralelo en múltiples nodos de un cluster, acelerando significativamente el proceso.

3. Escalabilidad:
- A medida que aumenta el volumen de datos o las demandas de cálculo, es posible agregar más nodos al sistema distribuido.
- Esta escalabilidad es esencial para mantenerse al día con el crecimiento exponencial de los datos genómicos.

4. Tolerancia a Fallos:
- En un sistema distribuido, si un nodo falla, el trabajo puede ser redistribuido a otros nodos, garantizando que no se pierda el progreso y que el análisis pueda continuar sin interrupciones.

5. Análisis en la Nube:
- Las soluciones de computación en la nube, como AWS, Google Cloud y Microsoft Azure, ofrecen capacidades de computación distribuida "bajo demanda".
- Los investigadores pueden alquilar recursos según sea necesario, evitando la inversión en infraestructura costosa y permitiendo el acceso a herramientas y algoritmos de vanguardia.

6. Optimización de Workflows:
- Las plataformas de gestión de workflows, como Apache Airflow o Nextflow, permiten diseñar, coordinar y optimizar flujos de trabajo complejos en entornos distribuidos, garantizando que las tareas se ejecuten en el orden correcto y aprovechando al máximo los recursos disponibles.

7. Colaboración y Compartición de Datos:
- Los sistemas distribuidos facilitan la compartición y el acceso a datos entre investigadores y organizaciones, promoviendo la colaboración y permitiendo análisis conjuntos de grandes conjuntos de datos.

En resumen, la computación distribuida es esencial para aprovechar al máximo las capacidades de la Secuenciación de Nueva Generación. Al proporcionar soluciones para el almacenamiento y procesamiento de grandes conjuntos de datos, así como para la optimización y colaboración, la computación distribuida ha ampliado las posibilidades de la investigación genómica y ha acelerado los descubrimientos en este campo.

Caso de Uso: Implementación de Análisis Genómico con Nextflow y AWS

La genómica ha experimentado una revolución sin precedentes en términos de generación de datos y complejidad de análisis. Los investigadores y científicos de datos se

enfrentan al desafío de procesar y analizar rápidamente terabytes de datos de secuenciación, y esto requiere infraestructuras computacionales robustas, flexibles y escalables. En este contexto, las herramientas y servicios basados en la nube, como AWS, y los sistemas de gestión de workflows, como Nextflow, han emergido como soluciones líderes para abordar estos desafíos.

En este caso de uso, exploraremos cómo una organización de investigación genómica hipotética implementa un flujo de trabajo de análisis de NGS utilizando Nextflow para coordinar y optimizar el proceso, y AWS para proporcionar la infraestructura computacional y de almacenamiento necesaria. La combinación de estas dos tecnologías permite a la organización escalar dinámicamente sus recursos, optimizar costes y acelerar el tiempo de descubrimiento.

Este escenario ilustra un enfoque práctico y moderno para la investigación genómica, demostrando cómo las soluciones de computación distribuida y en la nube pueden ser implementadas de manera efectiva para superar los desafíos computacionales inherentes al campo de la genómica.

Contexto y Necesidad

La Organización Genómica del Pacífico (OGP) es un prominente centro de investigación situado en la costa oeste de los Estados Unidos. Durante más de una década, la OGP ha liderado investigaciones en áreas de genómica humana, microbioma y genómica evolutiva. Con un equipo de más de 200 científicos, la organización ha producido algunos de los estudios más citados en el campo de la genómica.

En los últimos años, la OGP ha iniciado un ambicioso proyecto: "El Atlas Genómico del Pacífico". Este proyecto busca secuenciar y analizar los genomas de 10,000 individuos de la región del Pacífico para estudiar la diversidad genética, identificar variantes raras y comprender mejor las adaptaciones evolutivas de las poblaciones en esta área geográfica única.

Dada la magnitud del proyecto "El Atlas Genómico del Pacífico", la OGP se enfrentó a varios desafíos:

1. Volumen de Datos: Cada genoma completo secuenciado genera alrededor de 100-150 GB de datos en formato raw. Con 10,000 genomas, se anticipa la generación de más de un petabyte de datos.

2. Análisis Computacional Intensivo: El análisis de los datos de secuenciación no es una tarea trivial. Incluye pasos como alineación, identificación de variantes y anotación, cada uno de los cuales requiere recursos computacionales significativos.

3. Tiempo: Con las infraestructuras tradicionales, procesar un solo genoma puede llevar días. Con 10,000 genomas, esto se traduce en años de tiempo de procesamiento, lo cual es inaceptable para la organización.

4. Costes: El almacenamiento y análisis de grandes volúmenes de datos pueden resultar en costes prohibitivos si no se gestionan y optimizan adecuadamente.

Por lo tanto, la OGP necesitaba una solución que no solo pudiera manejar el volumen masivo de datos y el análisis computacional, sino que también fuera coste-efectiva y redujera significativamente el tiempo total de procesamiento. La adopción de técnicas de computación distribuida y soluciones en la nube emergió como la solución más viable para abordar estos desafíos.

Selección de Herramientas

Dada la necesidad de la Organización Genómica del Pacífico (OGP) de procesar y analizar grandes volúmenes de datos genómicos de manera eficiente, se inició un proceso de evaluación y selección de herramientas adecuadas. Varias soluciones y plataformas se consideraron, pero las herramientas seleccionadas debían satisfacer criterios específicos, como escalabilidad, flexibilidad, facilidad de uso, capacidad de integración y coste-efectividad.

1. Nextflow:
Razones para la elección:
- Flexibilidad: Nextflow permite escribir flujos de trabajo complejos utilizando un lenguaje de script sencillo y fácil de leer. Esto facilita la adaptación y modificación de flujos de trabajo según las necesidades del proyecto.
- Portabilidad: Nextflow es agnóstico en cuanto a la infraestructura, lo que significa que los flujos de trabajo pueden ejecutarse en diferentes plataformas sin necesidad de modificaciones, desde una laptop local hasta clusters de alta performance o la nube.
- Paralelización y Optimización: Nextflow maneja automáticamente la paralelización de tareas y optimiza la utilización de recursos.
- Reproducibilidad: Permite una ejecución consistente y reproducible de análisis, asegurando que los resultados sean confiables y repetibles en diferentes entornos.

2. Amazon Web Services (AWS):
Razones para la elección:

- Escalabilidad: AWS ofrece una amplia gama de servicios que se pueden escalar dinámicamente según las necesidades. Esto es crucial para manejar los picos de demanda durante las fases intensivas de procesamiento.
- Variedad de Servicios: Desde EC2 para computación, S3 para almacenamiento de datos, hasta servicios más especializados como AWS Batch para la ejecución de trabajos por lotes, AWS cubre todas las necesidades de la OGP.
- Coste-Efectividad: Con su modelo de pago por uso, AWS permite a la OGP optimizar costes, ya que solo paga por los recursos que utiliza. Además, AWS ofrece opciones de instancias spot y reservadas que pueden reducir aún más los costes.
- Integración con Nextflow: AWS se integra perfectamente con Nextflow, permitiendo que los flujos de trabajo se desplieguen y ejecuten directamente en la nube sin esfuerzo adicional.

Tras una cuidadosa consideración y pruebas preliminares, la combinación de Nextflow y AWS emergió como la solución más adecuada para las necesidades del proyecto "El Atlas Genómico del Pacífico". Esta combinación no solo aborda los desafíos de escalabilidad y rendimiento, sino que también ofrece una solución coste-efectiva para el análisis genómico a gran escala.

Diseño del Workflow con Nextflow

El diseño del workflow en Nextflow para el proyecto "El Atlas Genómico del Pacífico" de la OGP se centró en abordar los principales pasos del análisis de secuenciación de NGS, desde la recepción de datos en formato raw hasta la identificación y anotación de variantes genéticas. A continuación, se presenta una descripción detallada del flujo de trabajo diseñado:

1. Entrada de Datos:
- Lecturas Raw: Las lecturas de secuenciación raw, generalmente en formato FASTQ, son la entrada principal. Estas lecturas pueden estar almacenadas localmente o en un bucket de S3 en AWS.

2. Control de Calidad (QC):
- Herramienta: FastQC.
- Se realiza una evaluación de calidad de las lecturas raw para identificar problemas potenciales como contaminación o degradación del ADN.
- Los informes generados por FastQC se consolidan para una revisión detallada.

3. Preprocesamiento de Lecturas:
- Herramientas: Trimmomatic y Cutadapt.
- Eliminación de adaptadores y bases de baja calidad de las lecturas.
- Las lecturas limpias resultantes se utilizan para los siguientes pasos del análisis.

4. Alineación de Lecturas:
- Herramienta: BWA-MEM.
- Las lecturas se alinean contra un genoma de referencia (por ejemplo, el genoma humano GRCh38) para obtener un archivo BAM alineado.
- La alineación permite mapear las lecturas a su ubicación genómica correspondiente.

5. Procesamiento Post-Alineación:
- Herramientas: SAMtools y GATK.
- Se realizan varias operaciones, como la eliminación de duplicados, la recalibración de calidad de bases y la realineación local para optimizar la calidad del archivo BAM.

6. Identificación de Variantes:
- Herramienta: GATK HaplotypeCaller.
- Se identifican variantes genéticas, como SNPs y indels, generando un archivo VCF.
- Este paso es crucial para el objetivo principal del proyecto: descubrir variantes genéticas en la población del Pacífico.

7. Anotación de Variantes:
- Herramienta: ANNOVAR o SnpEff.
- Las variantes identificadas se anotan para proporcionar información sobre su impacto potencial, su ubicación genómica, cambios en aminoácidos (si es aplicable) y otras características relevantes.

8. Salida y Reporte:
- Se generan informes detallados y resúmenes del análisis.
- Los datos procesados y los resultados se pueden almacenar de nuevo en S3 o en otro sistema de almacenamiento adecuado para su posterior análisis o compartición.

El flujo de trabajo diseñado con Nextflow se estructura en procesos individuales, donde cada proceso representa un paso específico del análisis (por ejemplo, control de calidad, alineación, identificación de variantes). Nextflow maneja automáticamente la paralelización de tareas, la gestión de dependencias y el control de errores, garantizando un flujo de trabajo robusto y eficiente. Además, gracias a su integración con AWS, los recursos computacionales se escalan dinámicamente según las necesidades de cada paso del análisis.

Infraestructura en AWS

Para el proyecto "El Atlas Genómico del Pacífico" llevado a cabo por la Organización Genómica del Pacífico (OGP), la infraestructura en Amazon Web Services (AWS) se diseñó para maximizar la eficiencia, la escalabilidad y el coste-efectividad. A continuación, se detalla la infraestructura seleccionada y configurada en AWS:

1. Amazon S3 (Simple Storage Service):
- Uso: Almacenamiento de datos.
- Las lecturas raw de secuenciación en formato FASTQ, así como todos los datos intermedios y finales, se almacenan en buckets de S3.
- S3 ofrece durabilidad, alta disponibilidad y escalabilidad para grandes conjuntos de datos.

2. Amazon EC2 (Elastic Compute Cloud):
- Uso: Procesamiento y análisis.
- Las instancias EC2 proporcionan la potencia computacional necesaria para ejecutar el flujo de trabajo de Nextflow.
- Se utilizan diferentes tipos y tamaños de instancias según las necesidades específicas de cada paso del análisis. Por ejemplo, se pueden usar instancias con alta memoria para la identificación de variantes y instancias con alto rendimiento de CPU para la alineación.

3. AWS Batch:
- Uso: Gestión y ejecución de trabajos.
- AWS Batch gestiona la ejecución de trabajos en contenedores, escalando automáticamente el número de instancias EC2 según las necesidades.
- Está integrado con Nextflow, permitiendo que los trabajos del flujo de trabajo se ejecuten de manera eficiente en la nube.

4. Amazon RDS (Relational Database Service):
- Uso: Almacenamiento de metadatos y resultados de análisis.
- RDS proporciona bases de datos relacionales que se utilizan para almacenar metadatos sobre las muestras, así como resultados resumidos de los análisis.

5. Amazon EFS (Elastic File System):
- Uso: Almacenamiento compartido.
- EFS proporciona un sistema de archivos escalable y de alta disponibilidad que se puede montar en varias instancias EC2. Es útil para datos que deben ser accesibles desde múltiples instancias simultáneamente.

6. AWS Lambda y Step Functions:
- Uso: Automatización y coordinación.
- AWS Lambda permite la ejecución de código en respuesta a eventos específicos, como la finalización de un trabajo o la llegada de nuevos datos a S3.
- Step Functions se utiliza para coordinar microservicios y automatizar flujos de trabajo.

7. Amazon CloudWatch:
- Uso: Monitoreo y alertas.
- CloudWatch monitoriza el rendimiento de los recursos de AWS, proporcionando métricas en tiempo real, registros y alertas. Esto es esencial para garantizar que el sistema funcione de manera óptima y para detectar y abordar rápidamente cualquier problema.

8. AWS Identity and Access Management (IAM):
- Uso: Seguridad y control de acceso.
- IAM se utiliza para definir permisos y políticas, garantizando que solo los usuarios y servicios autorizados puedan acceder a los recursos de AWS.

La combinación de estos servicios en AWS proporciona una infraestructura robusta y escalable para el proyecto. Gracias a la flexibilidad y la variedad de servicios disponibles en AWS, la OGP pudo diseñar una solución que se adapta perfectamente a sus necesidades, garantizando la eficiencia en el procesamiento y análisis de datos genómicos a gran escala.

Integración de Nextflow con AWS

La integración de Nextflow con AWS permite que los flujos de trabajo genómicos se ejecuten de manera eficiente en la infraestructura de la nube, aprovechando la escalabilidad, la flexibilidad y el poder computacional de AWS. A continuación, se detalla cómo se llevó a cabo esta integración:

1. Configuración de Credenciales:
- Para permitir que Nextflow interactúe con los servicios de AWS, es necesario configurar las credenciales de AWS. Esto se hace utilizando roles y políticas de AWS Identity and Access Management (IAM) para garantizar un acceso seguro.
- Las credenciales se almacenan de manera segura y se proporcionan a Nextflow, ya sea a través de variables de entorno o de archivos de configuración.

2. Ejecución en AWS Batch:
- Nextflow tiene un ejecutor específico para AWS Batch, lo que facilita la ejecución de trabajos en la infraestructura de AWS.
- Al usar el ejecutor de AWS Batch, Nextflow automáticamente envía trabajos para su ejecución en AWS, maneja la paralelización de tareas y gestiona las dependencias entre tareas.

3. Almacenamiento en S3:
- Los datos de entrada, intermedios y de salida se almacenan en buckets de Amazon S3.
- Nextflow puede leer directamente los datos desde S3 y también escribir resultados en S3, eliminando la necesidad de transferencias manuales de datos.

4. EFS para Almacenamiento Compartido:
- Para ciertos pasos del flujo de trabajo que requieren acceso a datos desde múltiples instancias, se utiliza Amazon EFS.
- Nextflow se configura para montar volúmenes EFS en las instancias EC2, permitiendo un acceso compartido a los datos.

5. Monitoreo con CloudWatch:
- Se integra Nextflow con Amazon CloudWatch para monitorizar el rendimiento y estado de los trabajos en tiempo real.
- Las métricas y registros generados por Nextflow se envían a CloudWatch, lo que permite alertas y supervisión en caso de fallos o cuellos de botella.

6. Automatización con Lambda:
- Se utilizan funciones AWS Lambda para automatizar ciertos aspectos post-procesamiento, como la notificación de la finalización de un flujo de trabajo o la limpieza de recursos temporales.
- Nextflow puede activar estas funciones Lambda al finalizar ciertas etapas del análisis.

7. Optimización de Costes:
- Gracias a la integración con AWS, se pueden utilizar instancias EC2 spot para ejecutar trabajos, lo que puede reducir significativamente los costes.
- Nextflow se configura para solicitar y utilizar estas instancias spot cuando estén disponibles.

La integración de Nextflow con AWS ofrece una solución potente y cohesiva para la ejecución de análisis genómicos a gran escala en la nube. Al combinar la capacidad de gestión de workflows de Nextflow con la infraestructura de AWS, la OGP pudo crear un sistema que maximiza la eficiencia, reduce el tiempo de análisis y proporciona una excelente relación calidad-precio.

Conclusiones

El proyecto "El Atlas Genómico del Pacífico", llevado a cabo por la Organización Genómica del Pacífico (OGP), representa una de las iniciativas más ambiciosas en el campo de la genómica en la región del Pacífico. La magnitud de los datos generados y la complejidad del análisis requerían una infraestructura y herramientas que fueran escalables, eficientes y coste-efectivas. La solución encontrada en la combinación de Nextflow y Amazon Web Services (AWS) demostró ser la elección acertada para enfrentar estos desafíos.

Las principales conclusiones del proyecto son:

1. Escalabilidad y Rendimiento: La integración de Nextflow con AWS permitió a la OGP procesar grandes volúmenes de datos genómicos en un tiempo significativamente reducido en comparación con infraestructuras tradicionales. La capacidad de AWS para escalar dinámicamente recursos garantizó que el análisis se realizara de manera óptima, independientemente del volumen de datos.

2. Flexibilidad y Adaptabilidad: La naturaleza modular y adaptable de los flujos de trabajo en Nextflow, combinada con la amplia gama de servicios ofrecidos por AWS, permitió a la OGP adaptar y modificar el análisis según las necesidades emergentes del proyecto.

3. Coste-Efectividad: Al aprovechar servicios específicos de AWS, como las instancias EC2 spot y el almacenamiento optimizado en S3, la OGP pudo mantener los costes bajo control, obteniendo un excelente retorno de la inversión.

4. Reproducibilidad y Consistencia: Una de las principales ventajas de utilizar Nextflow es la garantía de reproducibilidad. Cada análisis realizado puede ser replicado con precisión, asegurando la validez y confiabilidad de los resultados obtenidos.

5. Colaboración y Compartición: La infraestructura basada en la nube facilitó la compartición de datos y resultados con colaboradores y partes interesadas, promoviendo una mayor colaboración y transparencia en el proyecto.

6. Visión de Futuro: Con la infraestructura y flujos de trabajo ya establecidos en AWS y Nextflow, la OGP está bien posicionada para futuros proyectos y expansiones. La adaptabilidad de la solución garantiza que la organización estará preparada para enfrentar desafíos futuros en el campo de la genómica.

En resumen, la implementación de análisis genómicos utilizando Nextflow y AWS ha demostrado ser una combinación poderosa y efectiva. Este caso de uso sirve como un excelente ejemplo de cómo las tecnologías modernas pueden ser utilizadas para avanzar en la investigación genómica, proporcionando soluciones innovadoras a desafíos tradicionales en el campo.

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6

Acknowlegments

The present study has been funded by the AIR Genomics project (file number CCTT3/20/SA/0003) through the 2020 call for R&D Projects Oriented towards Excellence and Competitive Improvement of CCTT by the Institute of Business Competitiveness of Castilla y León and FEDER funds.