Universidad de La Habana

Facultad de Matemática y Computación
Procesamiento de Grandes Volúmenes de Datos

Informe Técnico del Proyecto

Sistema Distribuido de Generación y Procesamiento de Datos Genómicos Sintéticos

Integrantes:

Alberto E. Marichal Fonseca – Ciencia de Datos
Jabel Resendiz Aguirre – Ciencia de la Computación

20 de octubre de 2025

Descripción del proyecto

El presente proyecto implementa un sistema distribuido de generación , transmisión, almacenamineot y análisis de datos genómicos sintéticos utilizando un enfoque de procesamiento en streaming. El objetivo es simular un entorno de Procesamiento de gran volumen de datos que permita la generación continua de secuencias genéticas sintéticas (SNPs) de familias humanas y su posterior procesamineto de una infraestructura distribuida basada en Kafka y HDFS.

0.1. Objetivo Central

Diseñar e implementar un pipeline distribuido que genere y transmita información genómica sintética en tiempo real, garantizando la escalabilidad y persistencia del flujo mediante Apache Kafka y Hadoop HDFS.

0.2. Dataset seleccionado

El proyecto utiliza el **dataset "1000 Genome Data"** disponible públicamente en la plataforma **Kaggle**,accesible : kaggle.com. Este conjunto de datos forma parte del **1000 Genemes Project**, una de las iniciativas más reconocidas internacionalmente en genómica poblacional, orientada a caracterizar la diversidad genética humana mediante el análisis de **Single Nucleotide Polymorphism (SNPs)**

Este conjunto de datos no contiene las secuencias genéticas completas, sino una colección de metadatos experimentales que describen el proceso de secuenciación, la calidad de las muestras y la cobertura obtenida para cada individuo o cohorte. Entre sus columnas principales se encuentran:

- **Sample:** identificador único de la muestra genómica
- **Population:** población o grupo étnico que pertenece la muestra
- **Center:** laboratorio o institución que realizó la secuenciación.
- **Platform:** tecnología de secuenciación utilizada
- **Total Sequence:** cantidad total de bases genéticas secuenciadas por muestras
- **Aligned Non Duplicated Coverage:** porcentaje de cobertura efectiva tras el alineamiento y eliminación de duplicados.
- **Passed QC:**indicador binario del control de calidad de la muestra.

0.3. Justificación del Dataset

El dataset **"1000 Genome Data"** es adecuado para nuestro proyecto por varias razones:

- **Volumen:** Contiene miles de muestras genómicas, cada una con información detallada de SNPs y metadatos experimentales, lo que permite simular escenarios de **Grandes Volúmenes de Datos**. Este volumen es suficiente para probar la escalabilidad del sistema distribuido y la capacidad de procesamiento en streaming de Kafka y HDFS.
- **Características:** El conjunto de datos incluye atributos clave como identificación de muestra, población, laboratorio, plataforma de secuenciación, cobertura de secuenciación y control de calidad. Estos atributos permiten generar perfiles genómicos sintéticos realistas y diversificados para cada miembro de una familia simulada.
- **Pertinencia:** Dado que el proyecto se centra en la **generación y transmisión de SNPs sintéticos**, este dataset proporciona una base estadística y poblacional confiable para crear genomas sintéticos con distribuciones realistas. La información sobre cobertura y calidad asegura que los datos generados reflejen variaciones biológicas y técnicas similares a las observadas en secuenciación real.

Arquitectura Propuesta

El sistema implementa un pipeline de procesamiento de datos genómicos sintéticos en **streaming**, que permite generar, transmitir y analizar SNPs en tiempo real. La arquitectura general se describe en la

Descripción de los bloques

- **Producer:** Genera SNPs sintéticos para cada miembro de la familia, simulando datos genómicos reales y enviándolos a Kafka.Esto permite probar y validar el pipeline sin depender de datos reales, manteniendo la privacidad y consistencia de la información.
- **Kafka:** Actúa como intermediario de mensajes, transmitiendo los datos en tiempo real desde el producer hacia Spark Streaming y la consola.Su uso garantiza baja latencia y escalabilidad, asegurando que los datos se entreguen de manera confiable a múltiples consumidores.
- **Spark Streaming:** Procesa los datos en tiempo real, calcula estadísticas por familia y por miembro, y publica los resultados nuevamente en Kafka.Esto permite análisis

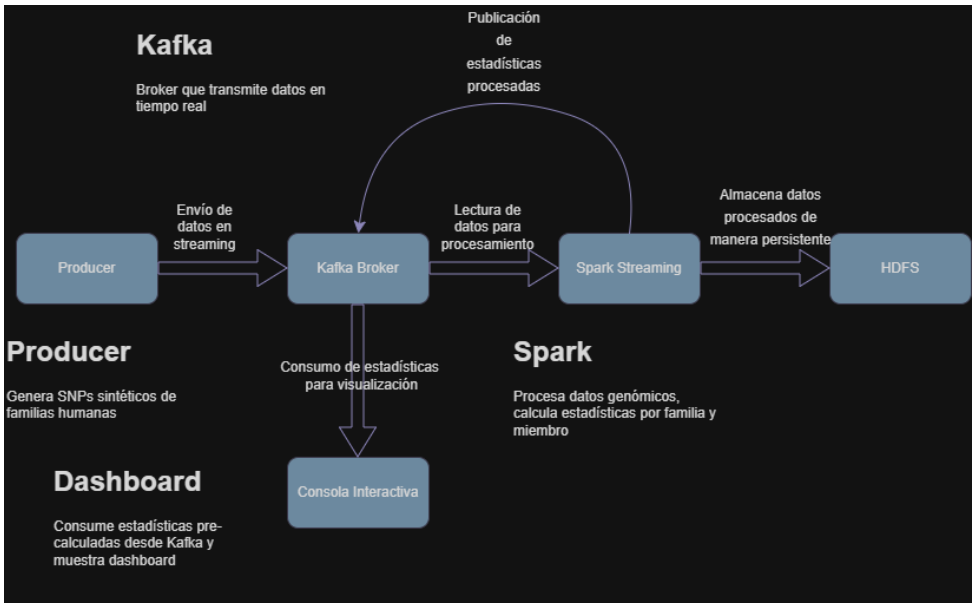

Figura 1: Arquitectura general del pipeline de procesamiento de datos genómicos sintéticos en streaming.

continuos y actualización inmediata de la información, crítico para la monitorización en tiempo real.

- **HDFS:** Almacena las estadísticas procesadas de forma persistente. Cada micro-batch se guarda en formato Parquet, permitiendo análisis históricos, reconstrucción de métricas y escalabilidad del sistema.
- **Consola Interactiva:** Consume las estadísticas pre-calculadas desde Kafka y presenta un dashboard interactivo que permite monitorear en tiempo real el estado de las familias y sus SNPs.Su función es facilitar la interpretación y toma de decisiones rápida, visualizando los resultados del pipeline de manera clara.

Enfoque

El enfoque del proyecto es **streaming**, dado que se requiere la generación y análisis continuo de datos genómicos, garantizando baja latencia y escalabilidad en tiempo real.

Avances

Hasta la fecha, el proyecto ha alcanzado los siguientes avances, reflejando la implementación de un prototipo funcional del pipeline de procesamiento de datos genómicos sintéticos en streaming:

- **Infraestructura de streaming y cluster:** Instalación y configuración de Kafka, Spark Streaming, HDFS, incluyendo contenedores Docker para facilitar la reproduci-

bilidad y pruebas locales. HDFS está disponible y configurado para almacenamiento batch, aunque no se utiliza en el flujo principal de streaming.

- **Tópicos y comunicación:** Creación de los tópicos en Kafka y verificación de la comunicación entre el *producer* y el *consumer*.
- **Generador de familias sintéticas:** Desarrollo del módulo que genera familias completas con SNPs simulando datos genómicos reales y envío de los mismos a Kafka en tiempo real.
- **Procesamiento en tiempo real:** Implementación de procesos en Spark Streaming para calcular estadísticas por familia y miembro, sin necesidad de procesamiento batch.
- **Visualización inicial:** Desarrollo de una consola interactiva que consume las estadísticas pre-calculadas desde Kafka y presenta un dashboard en tiempo real.
- **Métricas y monitoreo:** Registro de estadísticas de generación y envío de SNPs, incluyendo número de familias procesadas, SNPs enviados y latencia aproximada.
- **Pruebas de escalabilidad:** Ejecución de pruebas con múltiples hilos de generación para evaluar el comportamiento del pipeline bajo cargas incrementales.
- **Manejo de errores y logs:** Implementación de control de errores y logging en los hilos de streaming para garantizar la estabilidad del prototipo.

Estos avances constituyen un prototipo funcional que valida el enfoque de streaming y permite continuar con la expansión y optimización del pipeline.