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

8 de diciembre de 2025

Descripción del proyecto

El presente proyecto implementa un sistema distribuido para la generación, transmisión, almacenamiento y análisis de datos genómicos sintéticos en tiempo real. Su objetivo central es simular un flujo continuo de información genética de familias humanas, enfocándose en los **Single Nucleotide Polymorphisms (SNPs)**, para evaluar el rendimiento de una infraestructura distribuida basada en Apache Kafka y Hadoop HDFS.

0.1. Objetivo Central

Generar y transmitir datos genómicos sintéticos de manera escalable y persistente, simulando secuencias de SNPs de familias humanas, garantizando la continuidad del flujo y la capacidad de análisis en tiempo real mediante un pipeline distribuido.

0.2. Dataset seleccionado

El proyecto utiliza el **Family Genome Dataset** disponible públicamente en **Kaggle**: kaggle.com. Este dataset contiene cinco archivos CSV que representan los datos genómicos de una familia completa: `Father Genome.csv`, `Mother Genome.csv`, `Child1 Genome.csv`, `Child2 Genome.csv` y `Child3 Genome.csv`. Cada archivo incluye las siguientes columnas principales:

- **RSID:** identificador único de cada SNP.
- **Chromosome:** cromosoma donde se encuentra el SNP.
- **Position:** posición del SNP dentro del cromosoma.
- **Genotype:** variante genética observada para ese SNP.

Este formato permite asociar los SNPs a cada individuo de la familia y simular herencia genética en el pipeline distribuido.

0.3. Justificación del Dataset

El **Family Genome Dataset** es adecuado para nuestro proyecto por las siguientes razones:

- **Volumen:** Contiene miles de registros de SNPs por individuo y cinco archivos por familia, lo que permite simular **Grandes Volúmenes de Datos** y probar la escalabilidad de Kafka y HDFS.
- **Características:** Los atributos clave (`RSID`, `Chromosome`, `Position`, `Genotype`) permiten generar perfiles genómicos sintéticos realistas y diversificados, manteniendo coherencia familiar.

- **Pertinencia:** Al incluir datos de cada miembro de la familia, se puede simular herencia genética y transmitir SNPs sintéticos coherentes con distribuciones biológicas plausibles, alineándose con el objetivo central del proyecto.

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

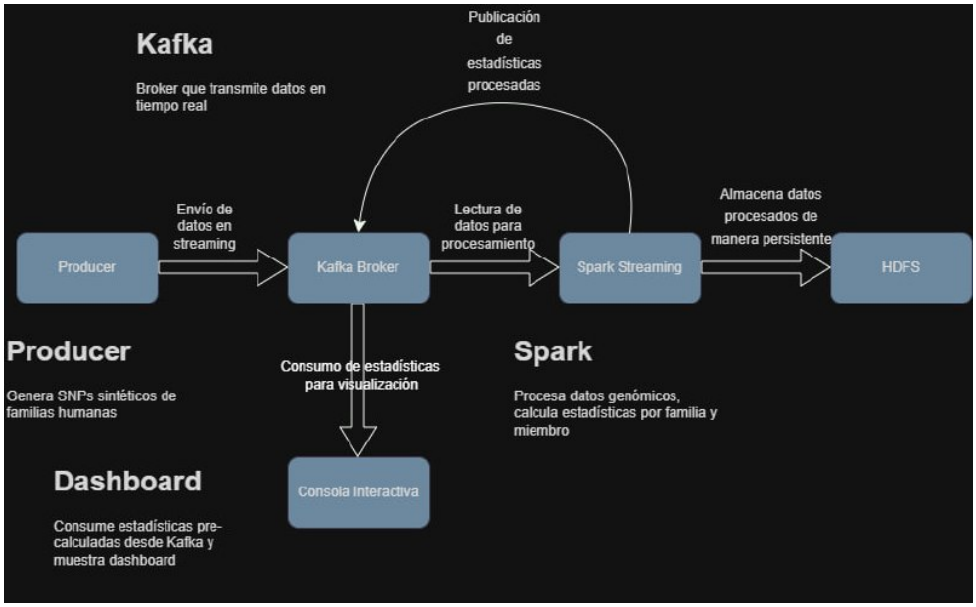

Figura 1: Arquitectura general del pipeline de procesamiento de datos genómicos sintéticos en streaming.

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

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 reproducibilidad 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.