MODULO 07 - EJERCICIO 01-B

ALEXIS YURI M.

Diseña una arquitectura básica de Big Data para una app de movilidad urbana. Debe incluir las capas de adquisición, almacenamiento, procesamiento, análisis y visualización.

1. Fuentes de Datos

Una aplicación de movilidad urbana, como Uber o Didi, genera datos de diversas fuentes:

- GPS: Datos de geolocalización de vehículos y usuarios en tiempo real.

- Interacciones de usuario: Solicitudes de viaje, calificaciones, comentarios y pagos.

- Operaciones: Información de tarifas, asignación de vehículos y tiempos de espera.

2. Tecnologías por Capa

Para cada capa se eligen las siguientes herramientas:

- Adquisición: Para la ingesta de datos en tiempo real (GPS, sensores) se propone usar Apache Kafka. Es ideal para manejar grandes volúmenes de datos.

- Almacenamiento: Los datos se pueden guardar en HDFS (Hadoop Distributed File System) para su almacenamiento escalable y tolerante a fallos. Es la base del ecosistema Hadoop para el almacenamiento de datos masivos.

- Procesamiento: Se propone Apache Spark para procesar los datos. Spark puede manejar tanto el procesamiento en tiempo real (Spark Streaming) para analizar datos de GPS como el procesamiento por lotes (Spark SQL) para realizar análisis históricos.

- Análisis y Visualización: Para analizar los datos estructurados, se puede usar Apache Hive, que permite consultar datos en HDFS usando un lenguaje similar a SQL. Para la visualización de datos de negocio, como la eficiencia de las rutas o la demanda por zonas, se pueden usar herramientas de BI como Power BI o Tableau.

3. Justificación de las Elecciones

- Apache Kafka: Su capacidad para manejar un alto flujo de eventos lo hace perfecto para capturar datos de GPS y sensores en tiempo real.

- HDFS (Hadoop Distributed File System): Es la solución estándar para almacenar Big Data de manera fiable y económica, permitiendo escalabilidad horizontal.

- Apache Spark: Es versátil y potente. Puede realizar procesamiento de streaming para análisis en tiempo real (por ejemplo, para calcular el tiempo estimado de llegada) y también procesar grandes lotes de datos históricos para análisis predictivo o de optimización de rutas.

- Hive: Facilita el análisis de los datos almacenados en HDFS sin necesidad de un conocimiento profundo de programación, lo que permite a los analistas de datos usar SQL.

- Power BI: Es una herramienta de visualización robusta que puede conectarse a Hive para crear paneles de control interactivos que muestren métricas clave del negocio.

4. Esquema de la Arquitectura

A continuación se presenta un diagrama simple de la arquitectura diseñada.

