Big Data Aplicado

UD 6 - Apache Hadoop

Práctica 4 - Spark Streaming

1. Usando cualquiera de las opciones disponibles Spark (cluster propio, docker o Databricks), realiza la siguiente práctica

Objetivo: Desarrolla una aplicación de Spark Streaming que procese un flujo de eventos simulados de transacciones financieras. Cada evento contiene múltiples piezas de información, incluyendo timestamps de generación del evento. La aplicación debe filtrar y procesar los eventos respetando una lógica de watermark para manejar datos tardíos y realizar agregaciones útiles para análisis financieros.

2. Cada evento tiene la siguiente información:

- **transaction_id**: Identificador único de la transacción.
- **amount**: Cantidad de la transacción
- **event_timestamp**: Timestamp que indica cuándo se hizo la transacción.
- **transaction_amount**: Cantidad de la transacción.
- **transaction_type**: Tipo de transacción (por ejemplo, depósito, retirada, transferencia).
- **currency**: Tipo de moneda de la transacción.

3. Los datos los tienes en formato csv en el propio repositorio.

Práctica. Realiza los siguientes apartados

Ejercicio 1.

1. Leer los eventos de transacciones desde una fuente de datos simulada. Puedes realizarla por **consola** o desde un **directorio** donde se encuentren los archivos
2. Calcula la suma, el número y media de las transacciones por tipo de moneda y tipo de transacción
3. Escribir los resultados agregados a la consola para su análisis.
4. Justifica el tipo de `output mode` usado.

Ejercicio 2.

1. Leer los eventos de transacciones desde una fuente de datos simulada. Puedes realizarla por **consola** o desde un **directorio** donde se encuentren los archivos
2. Aplica watermark y window para los datos tardíos utilizando el `timestamp` de la transacción
    i. Utiliza una ventana de 5 minutos
    ii. Configura una marca de agua de 10 minutos
    iii. Calcula en cada ventana, teniendo en cuenta la marca de agua, la suma y número de transacciones, agrupadas por tipo de transacción y moneda
3. Escribir los resultados agregados a la consola para su análisis.
4. Justifica el tipo de `output mode` usado.

Entrega:

La práctica debe ser entregada como un notebook de Jupyter o un script de Python que incluya comentarios explicativos sobre cada paso del análisis, asegurando que el código sea comprensible y bien organizado.

Datos:

En el propio repositorio