GROW DATA SKILLS

Spark Streaming Assignment: Real-Time Advertisement Data Aggregation

Objective: Process real-time advertisement data using Spark Streaming to gain business insights and store the aggregated data into Cassandra.

Background:
You have been provided with a Kafka topic named ads_data that contains advertisement data in the following format:

```
{
   "ad_id": "12345",
   "timestamp": "2023-08-23T12:01:05Z",
   "clicks": 5,
   "views": 10,
   "cost": 50.75
}
```

The goal is to process this real-time data, compute business insights using window-based aggregation, and write the aggregated results into a Cassandra table. The aggregation key is ad_id, and aggregated values should update previous values in the Cassandra table.

Tasks:

- **Kafka setup and Mock data producer:**
    - Set up Confluent Kafka on cloud or local
    - Create topic named as **ads_data**
    - Write a python script which will use above mentioned data format and keep on publishing random mock data in avro serialized form into Kafka topic
- **Reading Data from Kafka:**
    - Set up a Spark Streaming application.
    - Use the Kafka connector to read data from the ads_data topic.
    - Parse & desearialize the incoming data into the appropriate structure.

- **Windowing Based Aggregation:**

- Perform a window-based aggregation over a window duration (e.g., 1 minute) and sliding interval (e.g., 30 seconds).
    - Aggregate the following:
        - Total clicks per ad_id.
        - Total views per ad_id.
        - Average cost per view for each ad_id.

- **Write Aggregated Data to Cassandra:**
    - For each ad_id, check if an entry already exists in the Cassandra table.
    - If an entry exists, update the values:
        - Add new clicks/views to the existing counts.
        - Update the average cost per view.
    - If an entry doesn't exist, create a new row with the aggregated values.

Submission:

Submit your Spark Streaming application code, along with a brief report detailing the results and any challenges faced during the assignment.