Data Engineering Hackathon: Real-Time Ad Engagement Analysis

Background

In the digital advertising realm, understanding user engagement is pivotal. Tracking when users click on ads and then analysing the resultant actions provides businesses with the insights needed to optimize their ad campaigns, thus maximizing ROI.

Objective

Develop a real-time data processing pipeline to ingest, process, and visualize ad data from the click and conversion streams. Integrate this data with existing dimensional data to derive actionable insights.

Data Sources

Real-time streams:

- **Stream 1:** Ad Click Stream *(Refer the sample data and write script to generate mock data with real time frequency)*
    - Fields: ad_id, user_id, click_timestamp, platform, location

```
{
   "ad_id": 1001,
   "user_id": "U1234",
   "click_timestamp": "2023-09-09 10:00:00",
   "platform": "Web",
   "location": "New York"
}

{
   "ad_id": 1002,
   "user_id": "U5678",
   "click_timestamp": "2023-09-09 10:05:23",
   "platform": "Mobile",
   "location": "Los Angeles"
}

{
   "ad_id": 1003,
   "user_id": "U9101",
   "click_timestamp": "2023-09-09 10:10:34",
   "platform": "Tablet",
   "location": "San Francisco"
}
```

- **Stream 2**: Ad Conversions Stream *(Refer the sample data and write script to generate mock data with real time frequency)*
    - Fields: ad_id, user_id, conversion_timestamp, product_id, revenue

```
{
  "ad_id": 1001,
  "user_id": "U1234",
  "conversion_timestamp": "2023-09-09 10:20:00",
  "product_id": "P501",
  "revenue": 50.00
}

{
  "ad_id": 1002,
  "user_id": "U5678",
  "conversion_timestamp": "2023-09-09 10:30:23",
  "product_id": "P502",
  "revenue": 75.00
}
```


- **Campaigns** *(Slowly changing dimension table which should be loaded beforehand in downstream system, refer sample data):*
    - Fields: ad_id, campaign, product, target_start_date, target_end_date

      ad_id, campaign, product, target_start_date, target_end_date
      1001, Back-to-School, Laptop, 2023-08-01, 2023-10-31
      1002, Summer Sale, Smartphone, 2023-07-15, 2023-09-15
      1003, Fall Fashion, Jacket, 2023-09-01, 2023-11-30

- **User Demographics** *(Slowly changing dimension table which should be loaded beforehand in downstream system, refer sample data):*
    - Fields: user_id, age, gender, interests

      user_id, age, gender, interests
      U1234, 22, Male, Music;Travel
      U5678, 30, Female, Reading;Photography
      U9101, 27, Male, Gaming;Cooking

Requirements

- Real-time Data Ingestion

Data Processing:

- Join the real-time streams with the dimensional data to offer a comprehensive view of each event.
- Calculate, in real-time, metrics like:
    - Click-Through Rate (CTR) using click counts.
    - Conversion Rate based on clicks leading to conversions.

    - Average Revenue Per Conversion.
    - Conversion Latency (time between a click and its resultant conversion).
    - Trending Campaign

Data Storage:

Store processed real-time data in a NoSQL database for speedy querying and retrieval. Periodically offload data to a data warehouse for in-depth batch analysis and historical reporting.

Tech Stack:

Anything - Any open source, Any paid tool, Any cloud solution, Any programming language, free to use anything

Bonus:

- Present the real-time metrics on a dashboard.
- Features like:
    - Filtering by ad campaign, date range, or platform.
    - Real-time trends of CTR, Conversion Rate, etc.
    - Insights into top-performing ads based on revenue and conversion rates.

Evaluation Criteria

- Completeness of the Pipeline: End-to-end functionality from source to destination data ingestion
- Data Processing Efficiency: Speed and precision of calculations.
- Dashboard Usability: Clarity, real-time capabilities, and user experience.
- Scalability: Ability to handle vast data volumes and scale with increased data rates.
- Documentation: Clear documentation of architecture, data flows, and tools/components used.

Must do things for project submission

- Complete code should be checked in into Github along with readme file, architecture diagram, installation steps and step by step process to execute the project
- Teams need to create a 4-5 minutes of working Video demo of their project, need to explain the architecture and working of complete project with real time data
- No project submission will be accepted if above 2 points are not taken care