**TV VIEWERSHIP**

**Objective:**

Design and implement an ETL pipeline to process data generated by TV set-top boxes. The processed data should be stored in a well-designed data warehouse, enabling efficient access for business analysis.

**Scenario:**

Collects data from TV set-top boxes to analyze viewership patterns, track channel popularity, and understand user preferences. The raw data is semi-structured and arrives as log files in real-time from thousands of set-top boxes.

Your task is to design an end-to-end solution for processing this data and making it available for business analysts to query and generate insights.

**Requirements:**

1. **Data Sources:**
   * Raw logs from set-top boxes include the following details:
     + Mac ID (md5)
     + Event Date
     + Event Time
     + Channel Name
     + Program ID
     + Geo Location
     + Event Code
       1. SSTART: Event logged when box is turned on
       2. ELWS: Event logged when channel is Changed
       3. SALIVE: Event received every 10 min in case the user is watching and inactive on remote.
       4. SSTAND: Event logged when box is turned off
     + Satellite Name
   * Program logs include the following details:
     + Program ID
     + Program Name
     + Program Genre
     + Program Dialect
2. **Pipeline Design:**
   * Extract data from incoming logs (e.g., via SFTP, Kafka, or cloud storage like AWS S3).
   * Transform the data by cleaning, validating, and enriching it. Example transformations include:
     + Mapping Program IDs to Program Names using reference data.
     + Categorizing users based on location or program genre.
   * Load the processed data into a data warehouse which includes:
     + **Raw Data Table:** A table to store processed raw data for auditing and troubleshooting purposes.
     + **Downstream Table:** A table to store minute-by-minute user viewership data as shown in the table below. If a user switches between multiple channels within the same minute, the record should retain only the latest channel viewed during that minute. This table will support detailed analysis of user behaviour and channel popularity.

|  |  |  |  |
| --- | --- | --- | --- |
| **MAC ID** | **datetime** | **channel** | **Program name** |
| MAC1 | 01/01/2024 3:01 | CHANNEL1 | Program 1 |
| MAC1 | 01/01/2024 3:02 | CHANNEL1 | Program 1 |
| MAC1 | 01/01/2024 3:03 | CHANNEL2 | Program 01 |
| MAC1 | 01/01/2024 3:04 | CHANNEL2 | Program 01 |
| MAC1 | 01/01/2024 3:05 | CHANNEL1 | Program 1 |

1. **Data Warehouse Design:**
   * Design database views/SQL to support the following queries:
     + Total viewership hours by channel, and date.
     + Average viewing duration per user.
     + Top 10 most-watched channels daily.
   * Write a SQL statement that returns Reach and TRPs by Channel names.

Reach of a channel = No. of unique macs that watched the channel / viewers of that day.

Ratings = No. unique macs that watched the channel **on that minute**/ viewers of that day

TRP = Sum of Ratings.

1. **Tooling & Technologies:**
   * Use any tools (e.g., Python, SQL, Spark, Airflow, Redshift, etc.).
   * Implement automation for daily data processing.
2. **Performance Considerations:**
   * Ensure the pipeline can handle large-scale data (e.g., millions of records daily).
   * Optimize data warehouse queries for quick retrieval.
3. **Deliverables:**
   * Create a git repository and share it with us – keep pushing you commits so we monitor the progress and time
   * Please upload the following to the same git repo:
     + A high-level architecture diagram of the pipeline and data warehouse.
     + Code snippets for ETL implementation.
     + The processed raw data table, downstream table and other query result datasets in .txt format from the sample data shared.
     + Sample queries and Views for business analysis.
     + Documentation explaining design decisions, tools used, and future scalability considerations.
     + Create a PowerPoint presentation to present the findings to the executive team