Highlevel Data Engineering Exercise: Build a Data Pipeline

Objective:

To design, implement, and optimize a data pipeline that ingests data from multiple sources, performs necessary transformations, and loads the data into a target data warehouse.

Requirements:

1. **Data Sources:**
   - Source 1: A JSON data of restaurants. Static data - consider that this updates every week once.
   - Source 2: A sqlite data source of restaurant review. Dump it in SQL database, and build a pipeline according to that. Consider that there can be continuous inflow of data.
2. **Data Transformation:**
   - Normalize the data (e.g., consistent date formats, removing duplicates).
   - Perform necessary data cleaning (e.g., handling missing values, standardizing text fields).
3. **Data Loading:**
   - Load the transformed data into a target data warehouse (e.g., Amazon Redshift, Google BigQuery, Snowflake).
4. **Data Integrity:**
   - The order of data inflow has to be maintained. Race conditions should be handled accordingly.
5. **Performance Optimization:**
   - Optimize the pipeline for performance and scalability.
   - Implement logging and monitoring to track the pipeline's performance and errors.
6. **[Optional] Automation and Scheduling:**
   - Automate the pipeline to run at scheduled intervals (e.g., using Apache Airflow or cron jobs).
7. **Documentation:**
   - Provide clear documentation of the pipeline design, implementation, and any assumptions made.
   - Include a README file with instructions on how to run the pipeline.

Deliverables:

1. **Source Code:** Well-structured and documented code repository on Github. Please create a public/private repository on github for this test and add this user as collaborator (dev@gohighlevel.com / dev-highlevel)once the task is done
2. **Documentation:** Detailed documentation including pipeline design, setup instructions, and usage guidelines. Please include a video walkthrough of the project.
3. **Performance Metrics:** Report on the performance of the pipeline, including any optimizations made.
4. **Error Handling:** Description of error handling strategies implemented in the pipeline.

Data Set references:

JSON -

https://www.kaggle.com/datasets/ashishjangra27/swiggy-restaurants-dataset

SQL -

https://www.kaggle.com/datasets/ajaysh/amazon-fine-food-reviews