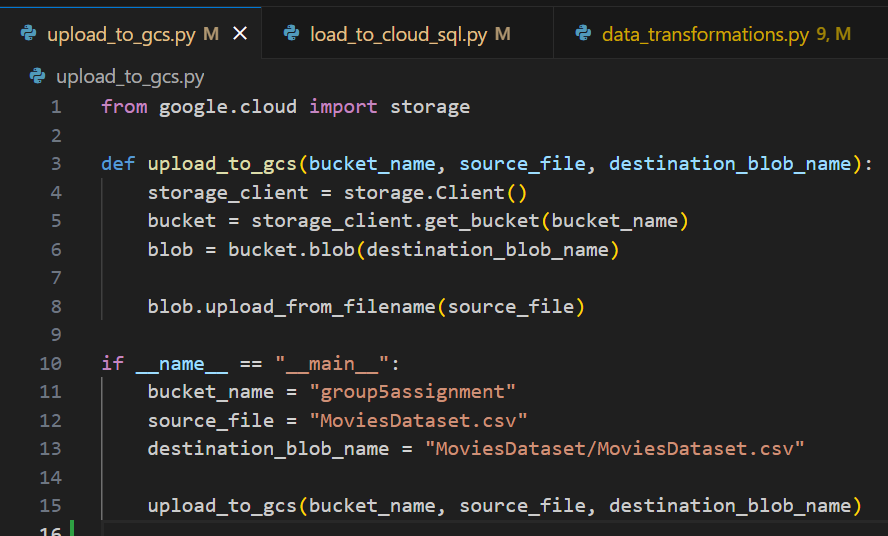
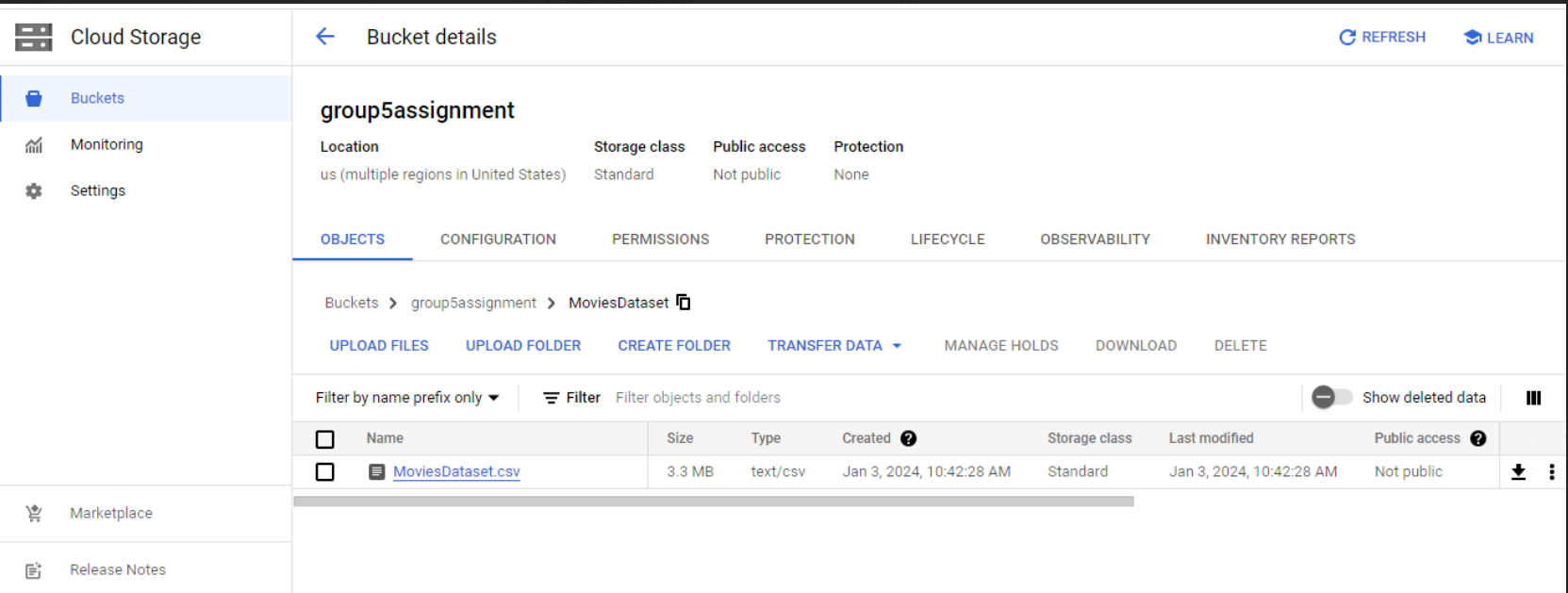
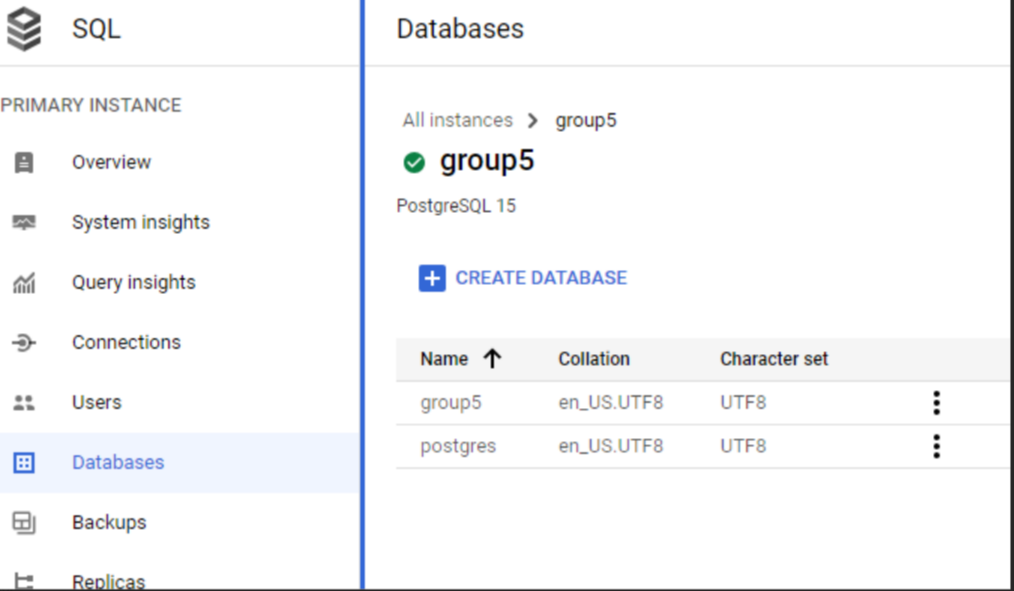
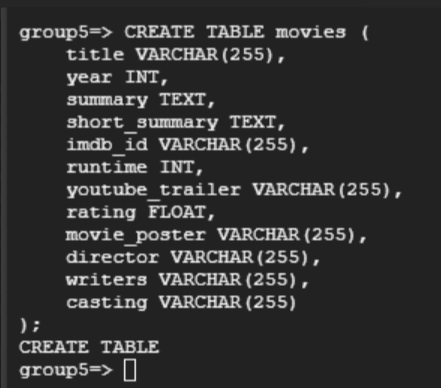
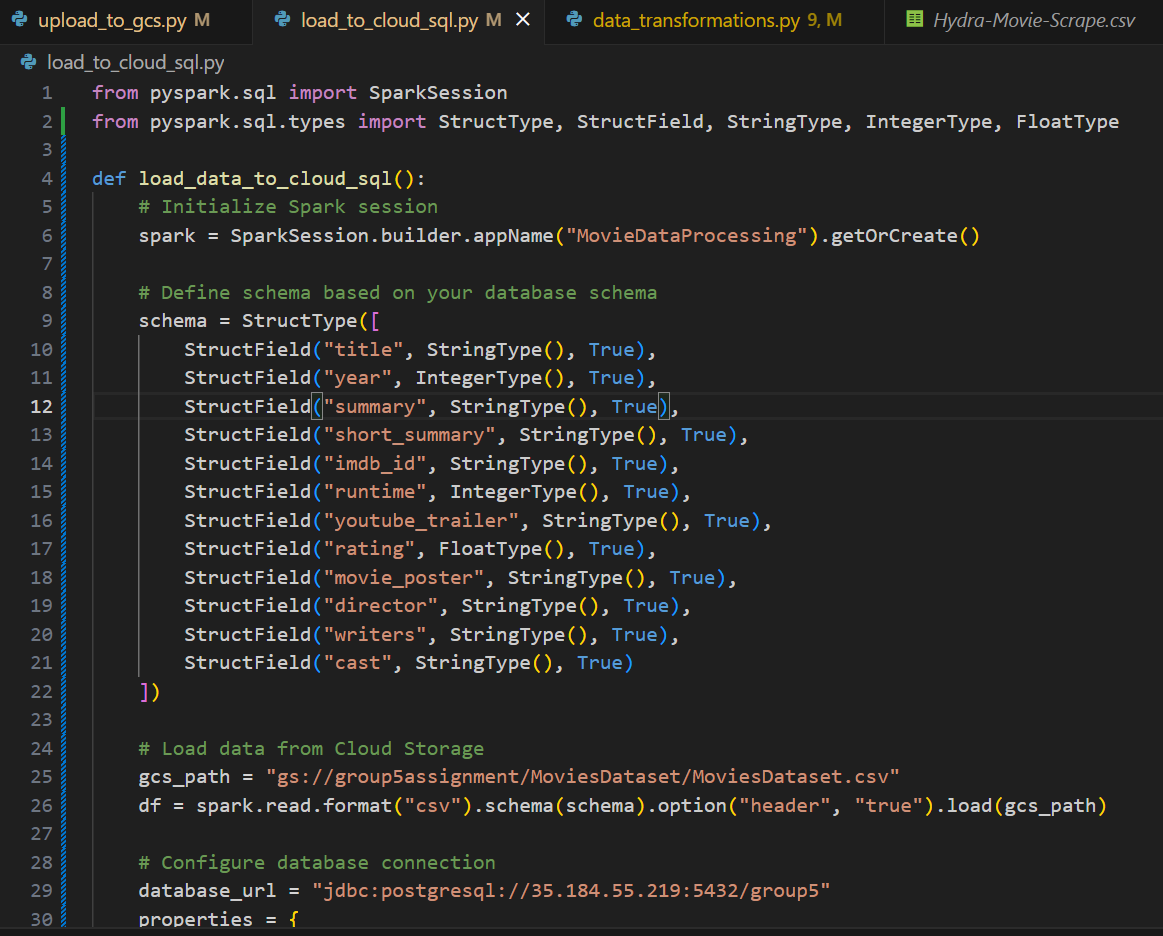
1. **Data Ingestion: Cloud Storage**
   1. **Write GCP Helper Python Script to perform Upload**
   2. **Upload the MoviesDataset CSV file to a Cloud Storage bucket.**

****

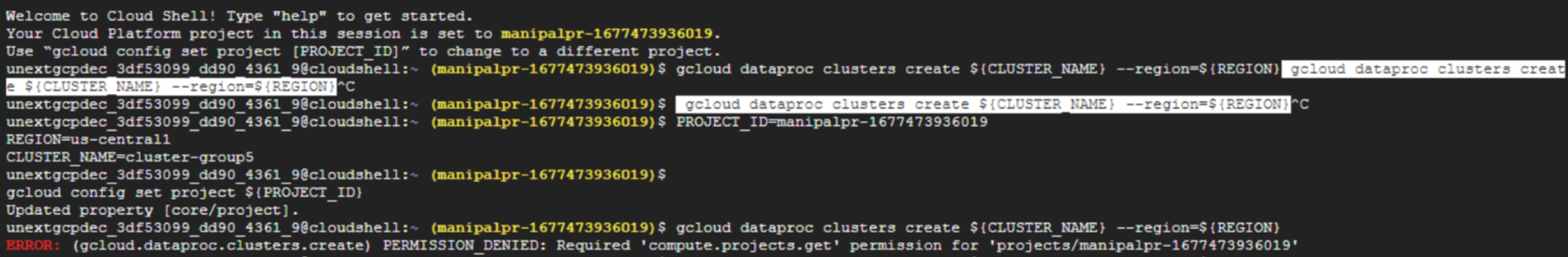
****

1. **Data Storage: Cloud SQL** 
   1. **Create a Cloud SQL instance (PostgreSQL/MySQL) to store normalized movie data using GCloud Utility.**
   2. **Design and implement a database schema that accommodates the dataset structure using SQL Client (CLI/GUI).**
   3. **Using PySpark Load the data from Cloud Storage into the Cloud SQL database.**

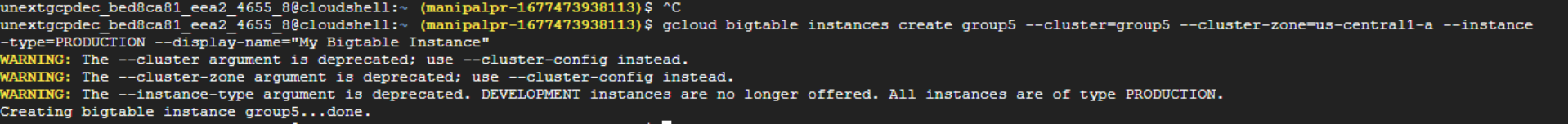
****

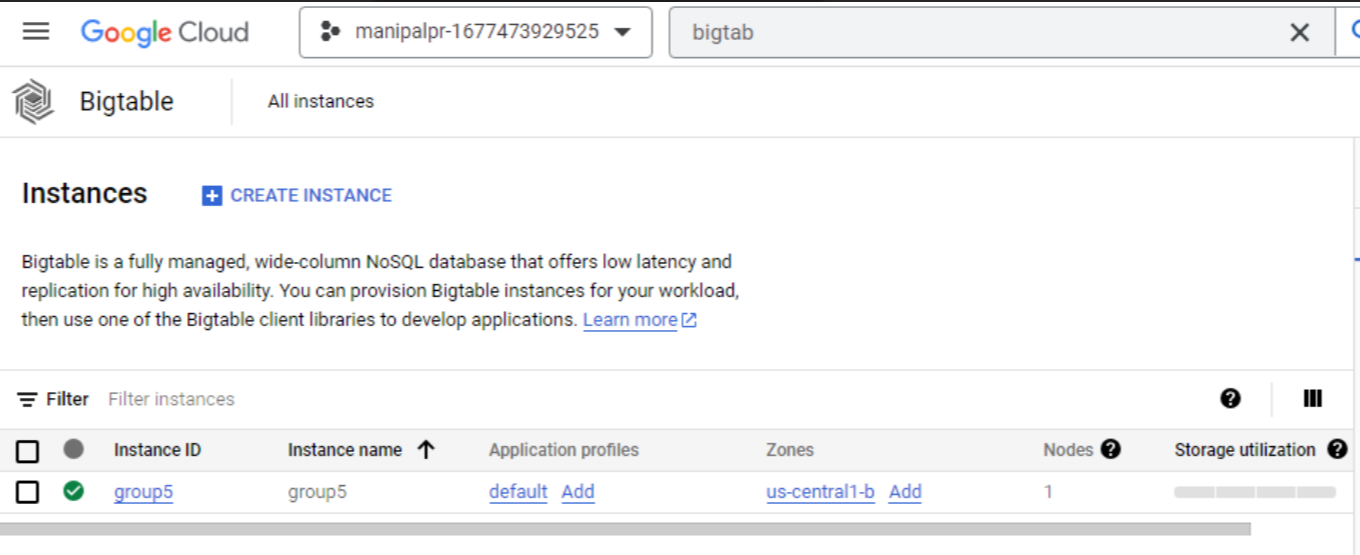
****

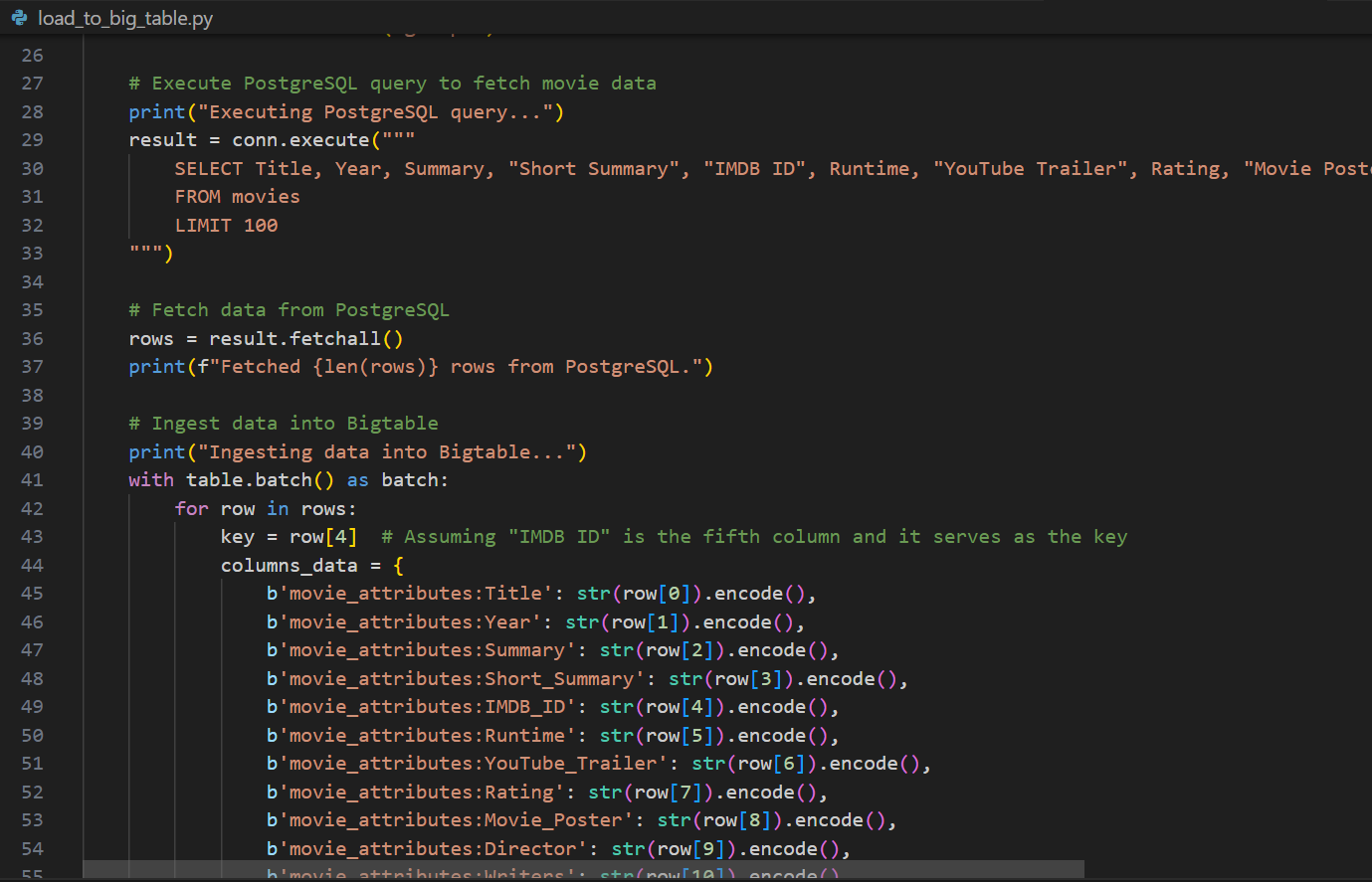
1. **Data Processing: DataProc/DataFlow**
   1. **Create a Dataproc cluster to process and transform the data using GCloud CLI/Portal.**
   2. **Explore and clean the dataset, handling missing values or outliers.**
   3. **Utilize Apache Spark on the Dataproc cluster to perform following data transformations:**
      1. **Which movies were released in the year 2020?**
      2. **What is the average IMDb rating of the movies in the dataset?**
      3. **Which movies have the longest and shortest runtimes?**
      4. **How many movies were directed by each director?**
      5. **Who are the unique writers in the dataset?**
      6. **Which movies have an IMDb rating greater than 8.0?**
      7. **Which movies do not have a YouTube trailer code?**
      8. **How many movies does each cast member appear in?**

****

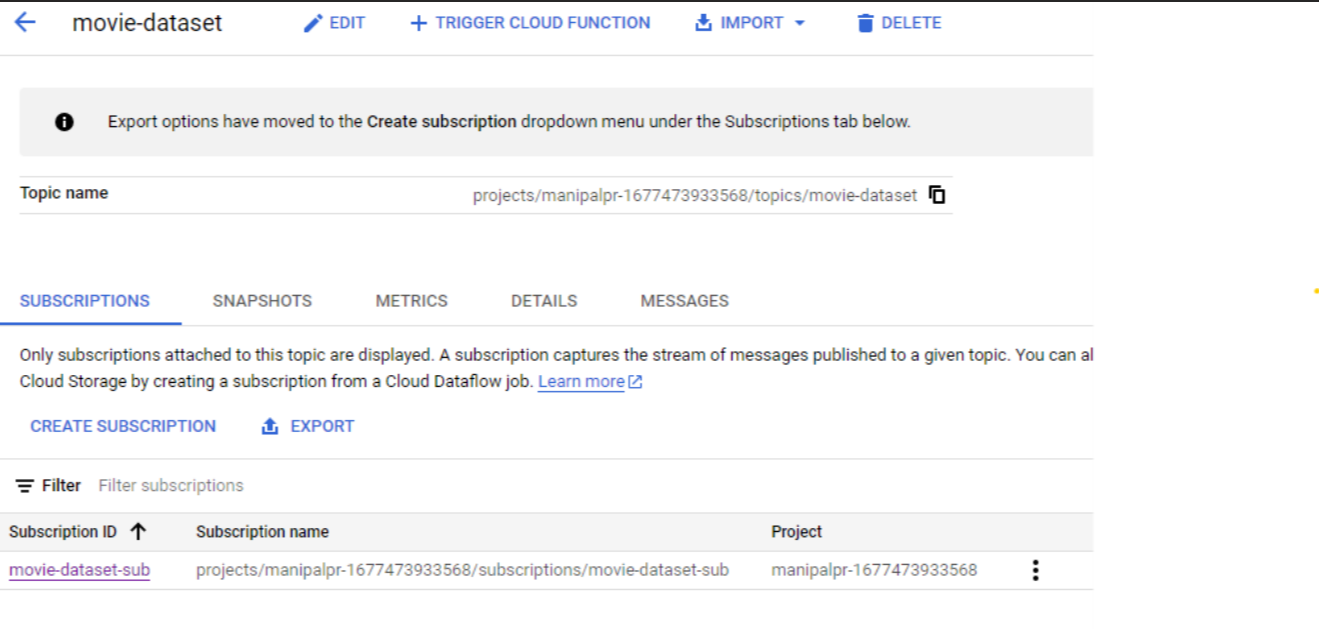
1. **Big Data Storage: BigTable**
   1. **Create a BigTable instance to store denormalized and indexed movie data.**
   2. **Design a schema that optimizes for the types of queries you might perform.**
   3. **Ingest a subset of the data into BigTable for analysis.**

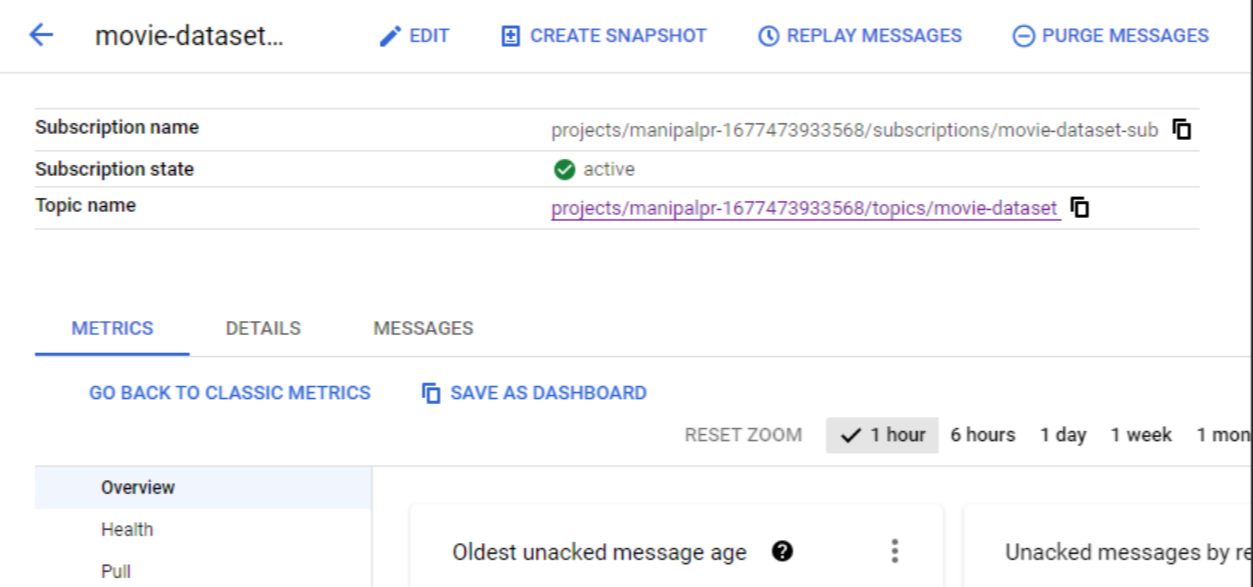
****

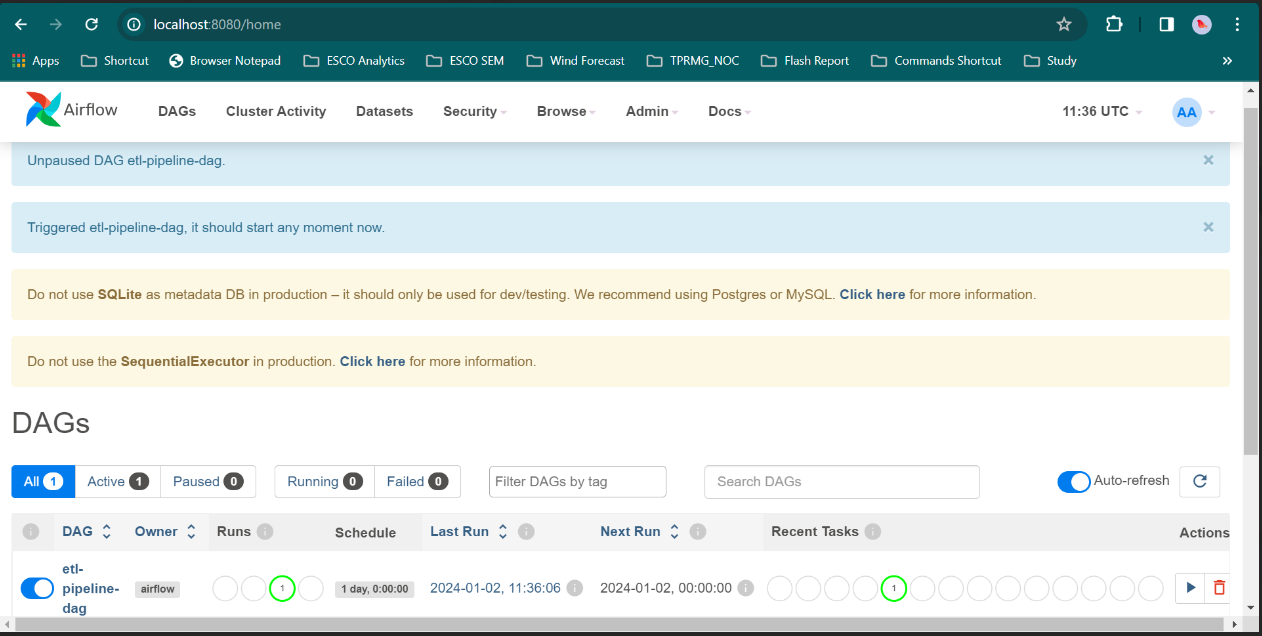
****

****

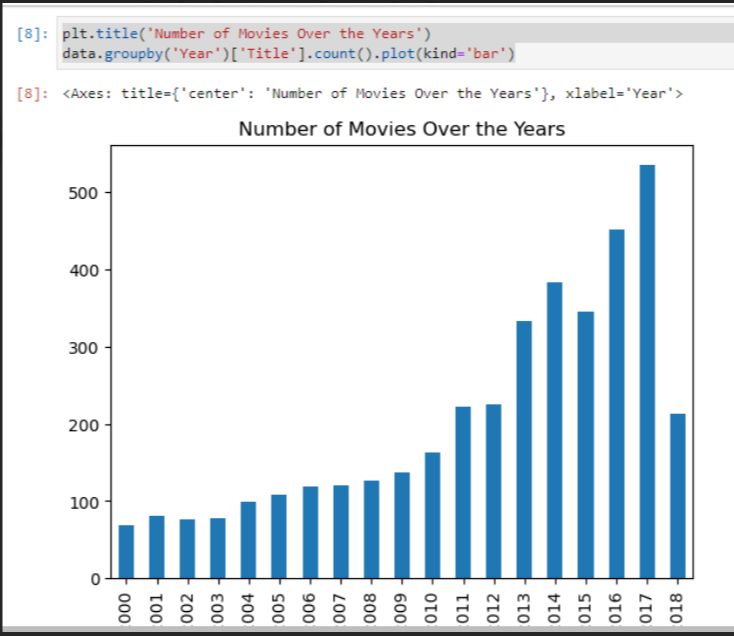
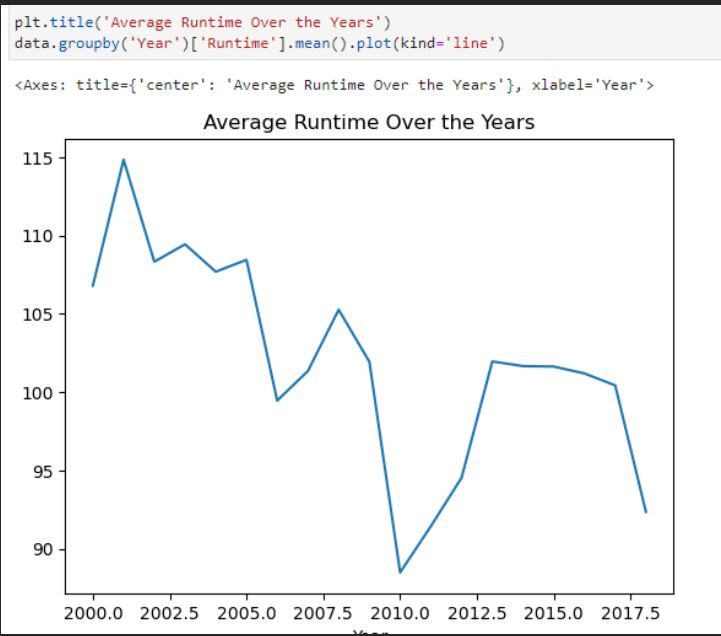
1. **Event-Driven Architecture: Pub/Sub**
   1. **Set up a Pub/Sub topic and subscription.**
   2. **Creating an event-driven pipeline using Airflow/GCP Scheduler, where new movie data is automatically ingested into Cloud Storage triggers a data processing job on Dataproc/DataFlow.**

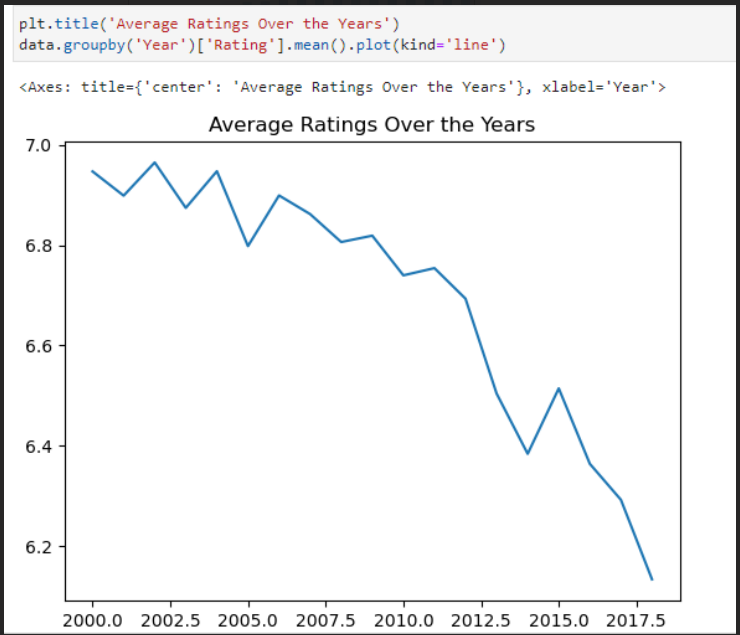
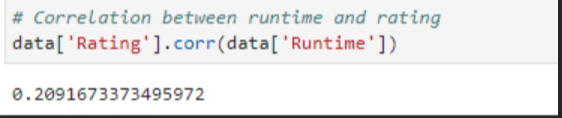
****

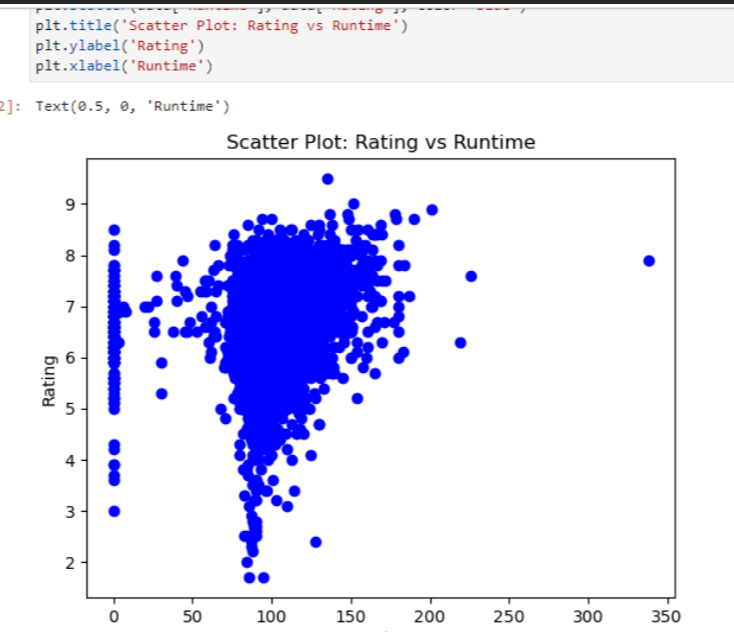
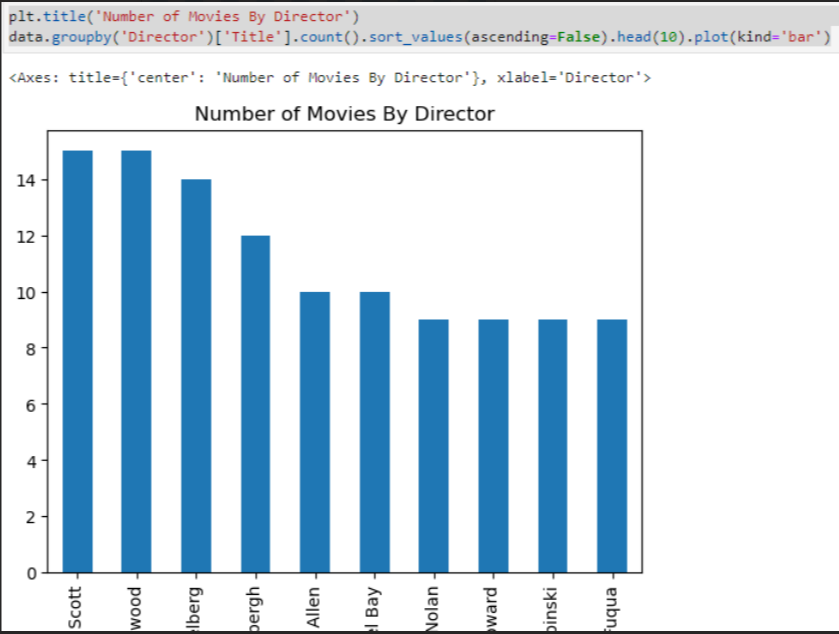
****

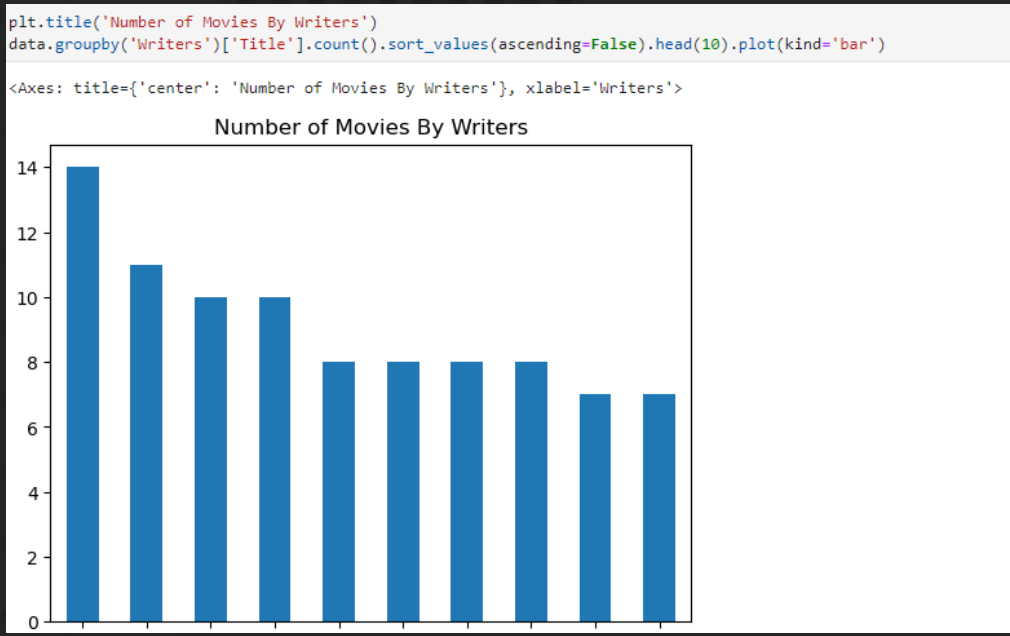
****

1. **Analysis and Reporting:**
   1. **Use Matplotlib or any preferred visualization tool to create reports and visualizations based on the processed data.**
   2. **Generate insights, trends, or interesting facts about the movies.**

** **

** **

** **

****