Project Brief

Created by Mr. Polakorn Anantapakorn

Title: TMDB Recommendation Pipeline System (TMDB-RecoFlow)

Introduction

In 2025, **watching movies remains a widely accessible and popular activity**. Streaming platforms greatly benefit from this popularity and offer various services, such as providing online movie streaming services, managing subscription payments, and offering comprehensive customer support. One crucial service for retaining customers is **movie recommendation**. This process is driven by **data analysis** and powered by a **recommendation system** that utilizes high-quality data derived from these analytical processes.

The fundamental process that ensures high-quality data for data analytics and building models is **data engineering**. It focuses on efficient data processing and **data pipeline system maintenance**, ultimately leading to actionable, high-quality data.

This project utilizes the **TMDB (The Movie Database) dataset**, a free, community-driven platform providing detailed information about movies, TV shows, and cast members. It includes metadata such as genres, ratings, production details, and images. Developers can access this rich dataset via a public API, making it ideal for building recommendation systems and movie-related projects.

This project presents the architecture and data pipeline system utilizing tools within the **data engineering field**. For example, **Apache Airflow** for orchestrating and managing data pipelines, **Apache Spark** for processes and transforms the large TMDB dataset efficiently, and **BigQuery**, a data warehouse service from Google Cloud Platform (GCP), for storing high-quality data. This project demonstrates **the end-to-end data engineering process**. The outcome of this project is high-quality data prepared for data analytics and recommendation systems.

Objective

1.**To design and implement a robust and scalable data architecture** that integrates all components (data ingestion, processing, storage, and serving) to support the end-to-end data pipeline.

2.**To build a data processing pipeline** that ensures high-quality data is extracted, transformed, and loaded into the data warehouse.

3.**To establish a scalable data warehouse** to house high-quality movie data, optimized to efficiently support a movie recommendation system.

Data Source

Name: Full TMDB Movies Dataset 2024 (1M Movies)

Source: https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies/data

Characteristic:

The TMDB Movies Dataset (2023) used in this project is a comprehensive and regularly updated collection of **film information**. It contains a vast number of movies, totaling **1,000,000 entries from the TMDB database**, with daily updates. Each entry includes essential details such as the movie's ID, Title, Average Vote, Vote Count, Status, Release Date, Revenue, and Runtime. Additionally**, the dataset features various other attributes that contribute to effective analysis and the development of robust movie recommendation systems.**

Concern:

- **Secondary Source**: The data is obtained from Kaggle, a secondary source, rather than directly from the TMDB API. This means the project relies on someone else's extraction and aggregation process, which might introduce unforeseen biases or limitations from their collection methodology.
- **Snapshot Nature**: The dataset is a static snapshot, containing data only up to (02-06-2025)

Output

The successful execution of this project, **TMDB-RecoFlow**, will yield the following key outputs:

- **A robust and scalable data pipeline**: This end-to-end pipeline, built with Apache Airflow for orchestration and Apache Spark for processing, ensures continuous data flow from ingestion to serving.
- **An optimized BigQuery data warehouse**: A well-structured central repository for high-quality, transformed movie metadata, designed for efficient data retrieval to support analytical queries and recommendation system training.
- **A high-quality dataset for recommendation**: The project will provide a clean, consistent, and readily accessible dataset specifically prepared to feed into a movie recommendation system for suggesting similar titles.
- **Demonstration of end-to-end data engineering**: This project will serve as a practical showcase of the complete data engineering lifecycle, from raw data acquisition to delivering actionable data for advanced analytics.

Benefits

- **Enhanced Operational Efficiency**: The automated data pipeline, powered by Apache Airflow and Spark, streamlines data processing, reducing manual effort and ensuring timely data availability.
- **Actionable Business Insights**: The project delivers clean, structured data in BigQuery, enabling precise analytics and valuable insights into movie trends and user behavior.
- **Practical Skill Development**: For the developer, it provides invaluable hands-on experience in designing, implementing, and managing an end-to-end data engineering pipeline using industry-standard tools.