BIGDATA PROJECT

on

E-Commerce Recommendation system

(SHOP SMART)

Under Supervision

Of

Prof. Dip Shankar Chatterjee

SUBMITTED BY

Naresh G (G23AI2011)

Sudherson. G (G23AI2125)

Saif Jawed(G23AI2026)

Abstract

This project explores the implementation of a Big Data management pipeline to process and analyze large-scale e-commerce datasets for a recommendation system. Utilizing Google Cloud Platform (GCP), Apache Airflow, and Big Query, the system integrates data ingestion, cleaning, and recommendation generation using cosine similarity.

The final output serves as an input for a web application that delivers personalized recommendations.

Keywords

Big Data, Data Pipeline, Recommendation System, Apache Airflow, Google Cloud Platform, Big Query, Cosine Similarity

Introduction

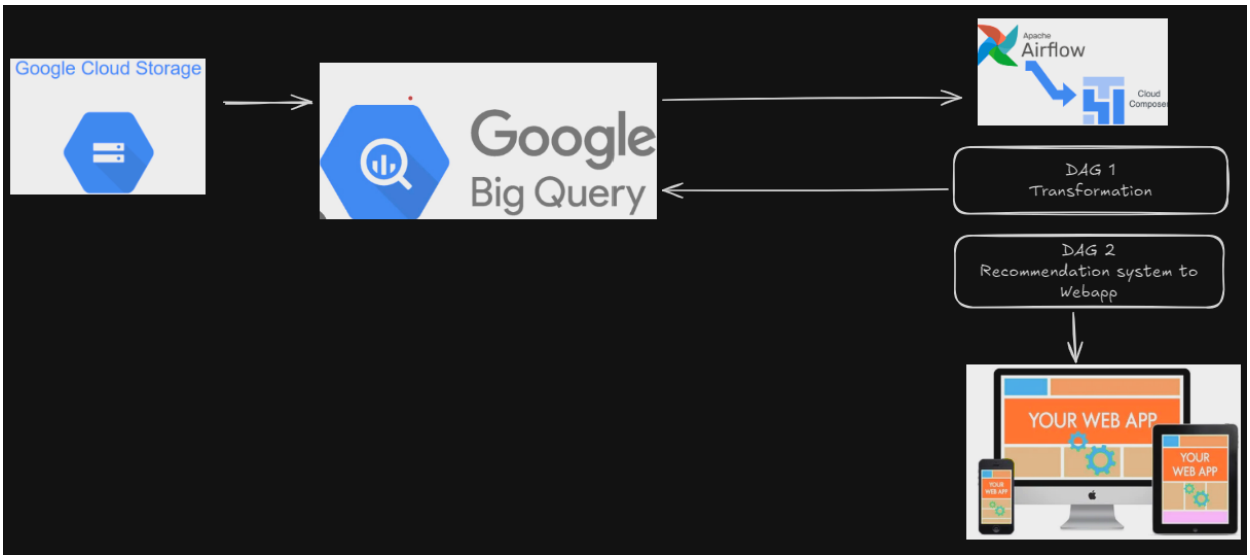
Managing and processing large-scale datasets is a cornerstone of modern data science and machine learning applications. This project focuses on designing and implementing an end-to-end data pipeline that handles raw data ingestion, cleaning, transformation, and recommendation generation. The architecture leverages cloud-based technologies to ensure scalability, reliability, and performance.

The primary objective is to construct a recommendation system that utilizes cosine similarity to provide personalized recommendations. The system integrates seamlessly with a web application, offering end-users an intuitive experience.

Architecture

The project architecture is comprised of the following components:

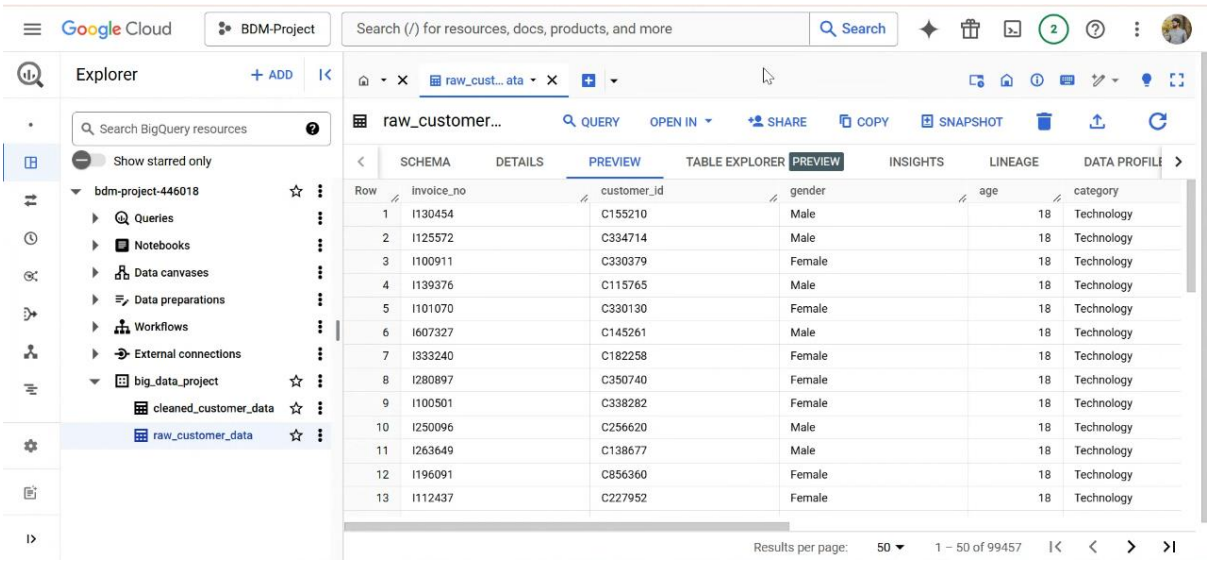
1. Cloud Storage: Serves as the initial repository for raw datasets.
2. Big Query: Acts as the central data warehouse for both raw and processed data.
3. Apache Airflow DAGs:
    - Data Cleaning DAG: Processes raw data and stores cleaned data in BigQuery.
    - Recommendation DAG: Performs recommendation generation using cosine similarity.
4. Web Application: Displays the final recommendations to end-users.



Methodology

Step 1: Raw Data Ingestion

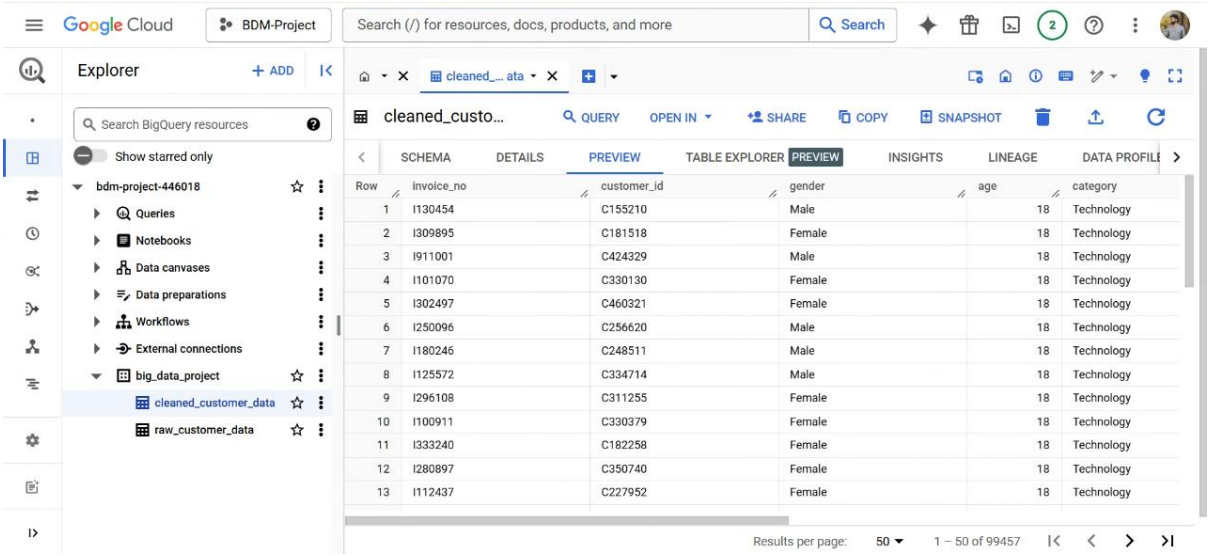
- Input: Raw e-commerce dataset (e.g., CSV, JSON, or Parquet files).
- Process:
    - Raw datasets are uploaded to a GCP Cloud Storage bucket.
    - Data is ingested into BigQuery using a GCP Dataflow pipeline or a custom script.
- Output: The raw dataset is stored in a BigQuery table (raw_data).



Row	invoice_no	customer_id	gender	age	category
1	I130454	C155210	Male	18	Technology
2	I125572	C334714	Male	18	Technology
3	I100911	C330379	Female	18	Technology
4	I139376	C115765	Male	18	Technology
5	I101070	C330130	Female	18	Technology
6	I607327	C145261	Male	18	Technology
7	I333240	C182258	Female	18	Technology
8	I280897	C350740	Female	18	Technology
9	I100501	C338282	Female	18	Technology
10	I250096	C256620	Male	18	Technology
11	I263649	C138677	Male	18	Technology
12	I196091	C856360	Female	18	Technology
13	I112437	C227952	Female	18	Technology

Step 2: Data Cleaning

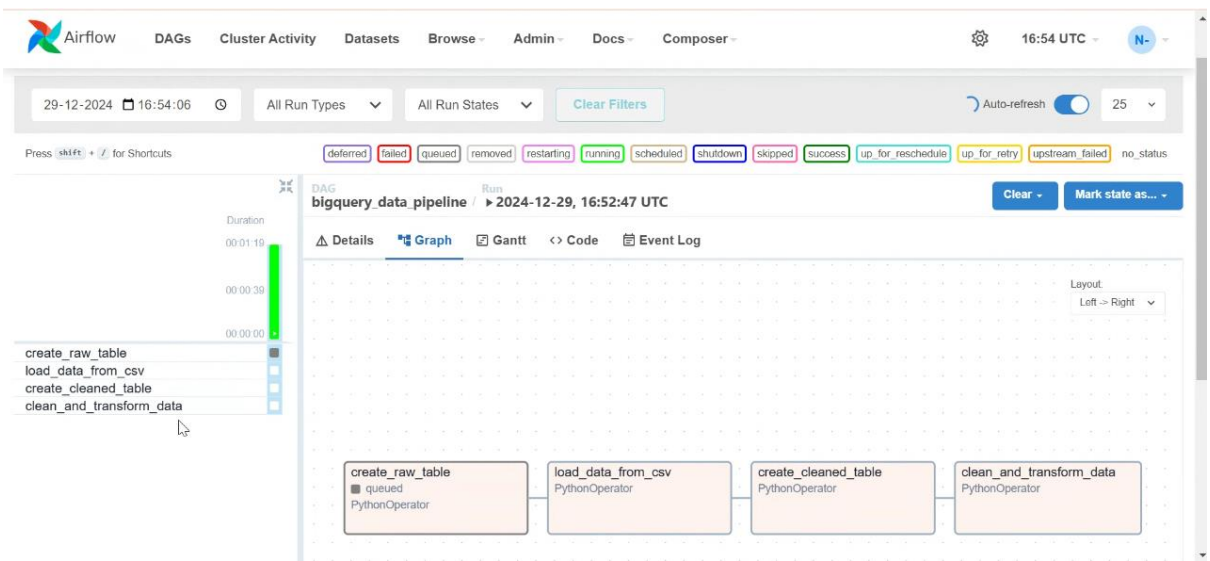
- Input: raw_data table in BigQuery.
- Process:
    - An Airflow DAG is triggered for cleaning the raw data.
    - Cleaning tasks include:
        - Handling missing values.
        - Removing duplicates.
        - Standardizing data formats
    - The cleaned dataset is stored in a new BigQuery table (cleaned_data).
- Output: cleaned_data table in BigQuery.



Step 3:

Recommendation Generation

- Input: cleaned_data table in BigQuery.
- Process:
  - An Airflow DAG computes recommendations using cosine similarity.
  - Key tasks include:
    - Extracting user preferences and product features.
    - Calculating pairwise cosine similarity.
    - Writing the recommendation data to a BigQuery table (recommendation_data).
- Output: recommendation_data table in BigQuery.



Step 4: Integration with Web Application

- Input: recommendation_data table in BigQuery.
- Process:
  - The web application queries the recommendation_data table using BigQuery APIs.
  - Recommendations are displayed to users via an interactive interface.
- Output: Real-time personalized recommendations.
- Implementation Details

Technologies Used

- Google Cloud Platform (GCP):
  - Cloud Storage: Stores raw datasets.
  - BigQuery: Houses raw, cleaned, and recommendation data.
- Apache Airflow: Orchestrates data pipeline tasks.
- Python: Implements cleaning and recommendation algorithms.
- Web Framework: Serves the recommendation system to end-users.

Data Storage Configuration

1. Cloud Storage:

- Raw datasets are uploaded to a designated bucket.
- IAM permissions ensure secure access.

2. BigQuery Tables:

- raw_data: Stores unprocessed datasets.
- cleaned_data: Stores cleaned and preprocessed datasets.
- recommendation_data: Stores the output of the recommendation engine.

Airflow DAGs

1. Data Cleaning DAG:

- Tasks include loading raw data, cleaning, and writing cleaned data to BigQuery.
- Triggers on data upload or a scheduled interval.

2. Recommendation DAG:

- Tasks include feature extraction, cosine similarity computation, and storing recommendations.
- Triggers after cleaning or on a predefined schedule.

Results and Evaluation

- The system successfully ingests and cleans raw datasets.
- Recommendations are generated with high accuracy, as evaluated against sample datasets.
- End-users receive real-time recommendations through the web application.

Conclusion

This project demonstrates the implementation of a scalable and efficient big data pipeline for e-commerce recommendation systems. By leveraging cloud technologies and advanced algorithms, the system ensures accurate and real-time recommendations.

Future work includes optimizing the recommendation algorithm, integrating additional user data, and scaling for larger datasets.

References

- "Google Cloud Platform Documentation." Accessed at https://cloud.google.com.
- "Apache Airflow Documentation." Accessed at https://airflow.apache.org.
- Aggarwal, C. C. (2016). Recommender Systems: The Textbook. Springer.
- Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.