Detailed Documentation

Data Model

The data model integrates and standardizes data from multiple sources into a unified format. The key components include:

- **Sales Data**: Contains transaction details such as Transaction_ID, Product_ID, Quantity, Price, and Transaction_Date.
- **Exchange Rates**: Provides currency exchange rates with fields like Currency_Code, Exchange_Rate, and Date.
- **Customer Demographics**: Includes customer information such as Customer_ID, Customer_Name, Age, Gender, Location, and Date_Joined.
- **Products**: Contains product details including Product_ID, Product_Name, Category, Price, and Stock_Available.
- **Transactions**: Records transaction details with fields like Transaction_ID, Customer_ID, Product_ID, Quantity, Transaction_Date, and Total_Amount.

The standardized data model ensures consistency in data types, naming conventions, and structures across all sources.

Pipeline Architecture

The pipeline architecture is designed to be robust, scalable, and modular. It includes the following stages:

1. **Data Ingestion**:
    - **Flat Files**: Read and parse CSV files.
    - **APIs**: Fetch data from external and internal APIs with proper authentication and error handling.
    - **Database**: Query data from PostgreSQL tables.
2. **Data Standardization**:
    - Merge data from different sources.
    - Standardize column names and data types.
    - Save the standardized data for further processing.
3. **Data Preprocessing**:
    - Handle missing data, duplicate records, and inconsistent entries.
    - Perform feature engineering, such as converting categorical variables and normalizing data.
    - Ensure the pipeline is modular and easily extendable.

Cloud Architecture (Step 4)

The cloud-based architecture uses Google Cloud Platform (GCP) services for scalability, resilience, and cost-efficiency. Here's a high-level overview:

- **Data Ingestion**: Cloud Functions for API data, Cloud Dataflow for CSV files.
- **Data Storage**: Google Cloud Storage for raw data, Cloud SQL for relational data.
- **Data Processing**: Cloud Dataflow for ETL jobs.
- **Data Serving**: Cloud Endpoints to expose APIs, Cloud Functions for serverless functions.

Diagram of Cloud Architecture: Cloud Architecture Diagram

Documentation and Presentation (Step 5)

- **Detailed Documentation**:
    - **Data Model**: Integrates and standardizes data from multiple sources into a unified format.
    - **Pipeline Architecture**: Robust, scalable, and modular, including data ingestion, standardization, preprocessing, and cloud deployment.
    - **Steps Taken**:
        - **Data Ingestion**: Read CSV files, fetch data from APIs, query PostgreSQL.
        - **Data Standardization**: Merge and standardize data.
        - **Data Preprocessing**: Handle missing data, remove duplicates, filter abnormal values, and perform feature engineering.
        - **Cloud Deployment**: Design and deploy a cloud-based architecture using GCP services.
- **Challenges and Solutions**:
    - **Challenges**: Handling missing data, ensuring data consistency, managing API rate limits, and designing a scalable cloud architecture.
    - **Solutions**: Implementing robust error handling, using forward filling for missing data, standardizing data formats, and leveraging GCP services for scalability and resilience.