Final Project Overview

Estimated Effort: 5 mins

Scenario

You are a data engineer who has been hired by a European online retail company to design a data workflow for their operations. The company requires that you perform all of the following tasks:

1. Propose a detailed data architecture for the whole data process.

    The client does not want cloud-based processing resources. The company wants an SQL-based central data repository that their employees from multiple countries can access for their use.

2. Propose a detailed data warehouse schema and design its entity requirements document (ERD).

    a. The client wants customer information, seller information, inventory information, and transaction invoice information, to be recorded.
    b. The client wants the final data prepared such that the final record of sales invoices contains the headers `InvoiceNo`, `StockCode`, `Description`, `Quantity`, `InvoiceDate`, `UnitPrice`, `CustomerID`, and `Country`.

3. Recommend the infrastructure requirements for the proposed data architecture.
4. Create an ETL pipeline to clean, process, and load the data to an SQL server for analysis. Test the pipeline on a sample database.

    a. The recorded data is available at a provided URL.
    b. The InvoiceNo starting with the character `C` is a credit entry, and you should remove these entries from the records before starting your analysis.
    c. The StockCode values of `C2`, `D`, `M`, and `POST` correspond to Carraige, Discount, Manual, and Postage entries, none of which are required for analysis.
    d. The CustomerID is missing from a few entries. Remove these entries from the data before analysis.
    e. Load the final transaction record to an SQLite3 database `Invoice_Records` under the table `Purchase_transactions`.

5. Query the SQL database to access data from the server.

    You are required to extract the data of a selected country from the database table created in the previous step.

6. Implement data analysis and data mining strategies on the final data.

    Then, implement Apriori algorithm and perform association rule mining on the data for the specified country.

Author(s)

Abhishek Gagneja