Kindle-1 Training Project

Problem Statement

Customer Segmentation: how to focus on growth

Kathy's manager has sent her an Excel file with endless rows of company customers and their purchase history. The goal is simple - to find **20%** of customers that drive **80%** of the company revenue. One solution is to work in spreadsheets , sore eyes and complete customer segmentation which is a tedious solution. She is looking for an optimal solution. Her colleague in Data Analytics has introduced her to the idea of reporting in Data Studio. However, data needs to be prepared in order to be used in Data Studio. She is looking forward to you to help her achieve the goal with a performant solution.

Solution Approach

You have a source file 'data.csv' which contains raw ecommerce data. Write a simple python script to load data in raw format into a table in a SQL Server. As part of data migration (ETL), create a data pipeline to extract data into a staging table in Big query in raw format. Note, no changes are made to the data as part of the extraction layer. In the transformation layer, perform below mentioned data cleaning and transformation steps. Feel free to use intermediate tables/ temporary tables as per your requirement. In the final step, load layer, create tables as mentioned and add audit columns (created_datetime, modified_datetime) to the table. In the end, create multiple views to be used for reporting to answer required business questions.

Data cleaning and preparing

1. **Filtering our negative (canceled) orders.** If you look at the Quantity column, you notice some negative values there - canceled or returned orders. Remove those orders.
2. **Handling NA values in the CustomerID** field. Customer ID is crucial for this analysis, so drop NA values there.
3. **Changing the InvoiceDate column to DateTime format.** In order to prevent any issues with dates, make sure all the data is in the appropriate format.
4. **Removing incomplete data.** Eg: December 2011 - not a full month in this data set.
5. **Calculating the sales column** by multiplying the Quantity and UnitPrice columns.
6. **Transforming data -** each record represents the purchase history of individual customers.
    a. Create a final table ONLINE_RETAIL to store cleansed data in Big Query
    b. Create a table CUSTOMER_SUMMARY to store summary for each customer purchase.
    c. Create a table SALES_SUMMARY to store summary of sales for each country"

Data Model

Table - ONLINE_RETAIL

Invoice No	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	ItemTotal
<string>	<string>	<string>	<integer>	<timestamp>	<numeric>	<numeric>	<string>	<numeric>

Table - CUSTOMER_SUMMARY

CustomerID	TotalSales	OrderCount	AvgOrderValue
<integer>	<double>	<integer>	<double>

Table - SALES_SUMMARY

Country	TotalSales	PercentofCountrySales
<string>	<integer>	<double>

Resources to be provided

1. Source file
2. Repository - MW
3. GCP Project - MW

Technology Stack

- Source

  - Csv file
  - SQL Server
- Data Pipeline
  - Python
  - Apache Beam
  - Dataflow
- DataWarehouse
  - BigQuery
- Orchestration
  - Cloud Composer
  - Airflow
- Reporting
  - Data Studio

Timeline

2 weeks

Acceptance Criteria

- Data modelling for target tables based on source
- Orchestration of job to execute pipeline, SQLs
- Documented flow of the entire project including architecture diagram, details of the tech stack, Challenges faced during implementation
- Answers the following business questions based on this dataset:
  - What are the sales figures for each country?
  - What is the overall sales trend?
  - How many new customers are there each month?
  - When do customers make the most purchases?
  - Which is the best selling product in each country?
  - When were the largest orders made?
  - Which customers made the largest orders?

References

- Dataset E-Commerce Data | Kaggle