Catherine Schuster

DS 3002: Data Project 2
Solution Methodology

For my final project, I designed and populated a dimensional data mart containing the relationship between **publishers, authors, titles, stores, and sales data of books found in the open-sourced Pubs SQL database**. The original database contained 11 tables (employees, jobs, sales, stores, discounts, titles, authors, publishers, publisher information, titleauthor, and roysched).

The main goal of my solution sought to perform an ETL process almost entirely on the cloud. This included **(a)** creating and populating the database on the cloud using Azure and Azure Data Studio, **(b)** streaming and ingesting the database into Azure Databricks via the Azure SQL database server connection where cleaning and integration could be performed, and **(c)** loading the finalized product into MongoDB as a final destination for business analysis and storage.

Specific requirements for this project are met as follows:

create a fact table and a date dimension table aligning with the database in Azure Studio, (c) stream and ingest the database into Databriucks through SQL server connectiuon date

| Requirement | Completion |
|---|---|
| 1. Your solution must include a **Date dimension** to enable the analysis of the business process over various intervals of time | Created a date dimension table in Azure Data Studio via the pubs SQL Database Server Connection. Dates started and ended in alignment with the dates within the database, easily allowing for join statements with other data tables or performing time series analysis. (found in pubsDim_Date.sql) |
| 2. Your solution must include at least 3 additional dimension tables (e.g., buyers, sellers, products)<br><br>3. Your solution must include at least 1 fact table that models the business process | Final product involves 8 dimension tables: titles_dim, authors_dim, publishers_dim, employee_dim, sales_dim, stores_dim, date_dim, **fact_table**<br><br>These finalized dimension tables were sourced from the Azure SQL connection, and cleaned, simplified, and exported in Azure Databricks. |

| | |
|---|---|
| | The fact table was created in Azure Data Studio by joining various keys and fact metrics across all tables into one table. (found in pubsFact_Table.sql) |
| 4. Your solution must populate its dimensions using data originating from multiple sources:<br>- A relational database like MySQL, Oracle or SQL Server<br>- A NoSQL database like MongoDB, Redis, Cassandra or HBase<br>- An API that returns a message payload (e.g., JSON, CSV, text) | Dimensions are created, populated, modified, and imported through<br><br>- SQL relational database (Azure)<br>- Cloud based SQL and python queries in Databricks (Azure Databricks)<br>- API (DBFS in Databricks)<br>- MongoDB NoSQL Database |
| 5. Your solution must demonstrate accumulating data that originates from a real-time (streaming) data source for a predetermined interval (mini-batch) | As seen in StreamingData.sql file, my solution demonstrates how I pulled mini batches of real-time sales data based on time series conditionals. |
| 6. Your solution must include one or more visualizations that demonstrate the business value of your solution. | Visualization screenshot attached in github folder, demonstrates how the solution can be used to track and measure sales performance overtime for business value. |