DS-2002 Final: Retail Sales Data Lakehouse with Streaming

For the DS-2022 Final, I constructed a dimensional data lakehouse for retail sales analysis using PySpark and the Bronze/Silver/Gold medallion architecture, using batch data and streaming data.

Data Lakehouse:

- The data lakehouse reflects customers purchases of products across various categories and regions
- The orders are tracked with pricing, quantity, and shipping values
- Transactions come in real-time and are enriched by reference data
- Allows for greater analysis of product sales, by category, time, and region

PySpark Data Lakehouse

- Created with estella_dw, a data warehouse created from Adventureworks data in MySQL. It includes the following:
    - 1 fact table:
        - "Fact_sales" - streaming order transactions with quantities, prices, and extended amounts
    - 6 Dimension tables:
        - "dim_date", calendar values for year/month/quarter analysis
        - "dim_product" product names, categories, and colors
        - "dim_customer" customer information
        - "dim_shipping_region" geographic shipping regions
        - "dim_currency" international currency
        - "dim_culture" localization and language

Medallion Architecture (Bronze/Silver/Gold) Pipeline

Using PySpark streaming in Jupyter Notebook, I implemented a 3-layer data lakehouse with Bronze, Silver, and Gold layers.

1. Bronze Layer – Raw Ingestion
    a. Data extraction: streaming ingestion from 3 JSON files containing order transactions
    b. Process: Process one file at a time to simulate data arriving in batches
    c. Load: Save as a parquet file w/ timestamps showing when each file was processed

2. Silver Layer - Cleaning and Connecting Data
    a. Read: Load the Bronze parquet files
    b. Join: Combine order data w/ dimension tables
    c. Transform: Add useful info such as product names, categories, ful dates
    d. Output: Cleaned files w/ all pieces connected together

3. Gold Layer - Final Business Reports
    a. Aggregate: Count products sold– grouped by month, year, category
    b. Metric: Total quantity of products sold
    c. Output: Temporary memory table for testing, permanent parquet files, SQL table for querying

Data Sources Used:

- estella_DW CSV files
- MongoDB Atlas
- MySQL table files
- JSON diles

Requirements Met:

- 6 dimension tables
- 1 fact table
- 3 sources (MySQL, MongoDB, CSV)
- Mixed batch and streaming data
- 3-Layer Architecture
- Connected facts and dimensions
- Business report created

Running the project:

- All code is located in Jupyter Notebook with PySpark
    - MySQL runs locally for some dimension tables
    - MongoDB Atlas for products and regions
    - CSV files stored locally for customers and dates
    - JSON files in a local folder for streaming

A query is provided at the end of the notebook to test the final table.