Report: Building an End-to-End Data Pipeline

Project Guide

I built a complete data pipeline, from data extraction to visualization. I obtained a dataset from Kaggle, extracted and transformed it using PySpark, stored it in a DuckDB database, and then visualized it using Microsoft Power BI to extract meaningful insights.

1. Data Source Identification & Understanding

Data Source:

- **Large Dataset:** A suitable e-commerce-related dataset containing **1,051,784 rows** was identified from Kaggle. The dataset includes detailed information on e-commerce transactions, such as sales, customer data, product data, and more. The dataset is stored in `ShopSpectra Transaction Dataset.csv`.

Overview of Columns:

- **TransactionID:** Unique identifier for each transaction.
- **UserID:** Unique identifier for each user.
- **TransactionAmount:** Amount of the transaction.
- **TransactionDate1:** Date and time of the transaction.
- **PaymentMethod:** Method of payment used.
- **MerchantCategory:** Category of the merchant.
- **Quantity:** Quantity of items purchased.
- **CustomerAge:** Age of the customer.
- **Location:** Location of the customer.
- **DeviceType:** Type of device used for the transaction.
- **TransactionStatus:** Status of the transaction (e.g., Pending, Completed).
- **Is_Declined:** Indicator if the transaction was declined.
- **Is_Fraud:** Indicator if the transaction was fraudulent.
- **AccountAgeDays:** Age of the customer's account in days.
- **TransactionDate:** Date of the transaction.
- **Latitude:** Latitude of the transaction location.
- **Longitude:** Longitude of the transaction location.
- **TransactionHour:** Hour of the transaction.

2. Data Extraction

Tools Used:

- PySpark for handling large datasets efficiently.

Process:

- The e-commerce dataset was loaded from the specified path using PySpark.
- Displayed the first few rows of the loaded dataset to understand its structure.

3. Data Transformation

Cleaning Processes:

1. **Handling Missing Values:**
   - **UserID** and **TransactionID**: Removed rows where these values were missing, as they are critical identifiers.
   - **TransactionDate1**: Interpolated missing values using the closest valid date from the same `UserID`.
   - **Numerical Columns**: Filled missing values with the median of the column.
   - **Categorical Columns**: Filled missing values with the mode (most frequent value).
   - **Is_Declined** and **Is_Fraud**: Filled missing values with 0 (not declined, not fraudulent).

Rows Removed Due to Missing Values:

- Removed 34,506 rows where `UserID` or `TransactionID` was missing.

2. **Removing Duplicates:**
   - Removed 12,000 duplicate rows to ensure data integrity.
3. **Correct Formatting Errors:**
   - Standardized date formats.
   - Cleaned numerical columns of non-numeric characters.
4. **Column Renaming:**
   - Renamed columns for consistency and clarity.
5. **Feature Engineering:**
   - Created new features such as total transactions per user and average transaction amount per user.
   - Joined the new features with the original dataset.

Summary Statistics for Clean Data:

- The clean dataset contains **1,040,800 rows** and **18 columns**.
- Key columns and their statistics:
  - **TransactionAmount:** Mean = 7,506.12, Std = 4,333.60, Min = 1.02, Max = 14,999.97
  - **Quantity:** Mean = 5.50, Std = 2.87, Min = 1.00, Max = 10.00
  - **CustomerAge:** Mean = 43.01, Std = 24.55, Min = 1.00, Max = 85.00
  - **AccountAgeDays:** Mean = 182.88, Std = 105.37, Min = 1.00, Max = 365.00
  - **Latitude:** Mean = 36.09, Std = 4.64, Min = 29.42, Max = 47.61
  - **Longitude:** Mean = 95.85, Std = 15.83, Min = 71.06, Max = 122.42
  - **TransactionHour:** Mean = 11.26, Std = 6.93, Min = 0.00, Max = 23.00

4. Data Loading

Database Schema Design:

- Designed a relational database schema in PostgreSQL to effectively store the transformed data.

Database Tables:

1. **Customers Table:** Stores customer information.
    - Fields: `customer_id`, `user_id`, `customer_age`, `location`, `account_age_days`
2. **Orders Table:** Stores order information.
    - Fields: `order_id`, `transaction_id`, `customer_id`, `order_date`, `total_amount`
    - Relationship: Foreign key linking `customer_id` to the `customers` table
3. **Products Table:** Stores product information.
    - Fields: `product_id`, `product_category`
4. **Order Items Table:** Stores order item information.
    - Fields: `order_item_id`, `order_id`, `product_id`, `quantity`, `price`
    - Relationships: Foreign keys linking `order_id` to the `orders` table and `product_id` to the `products` table

Loading Data into PostgreSQL:

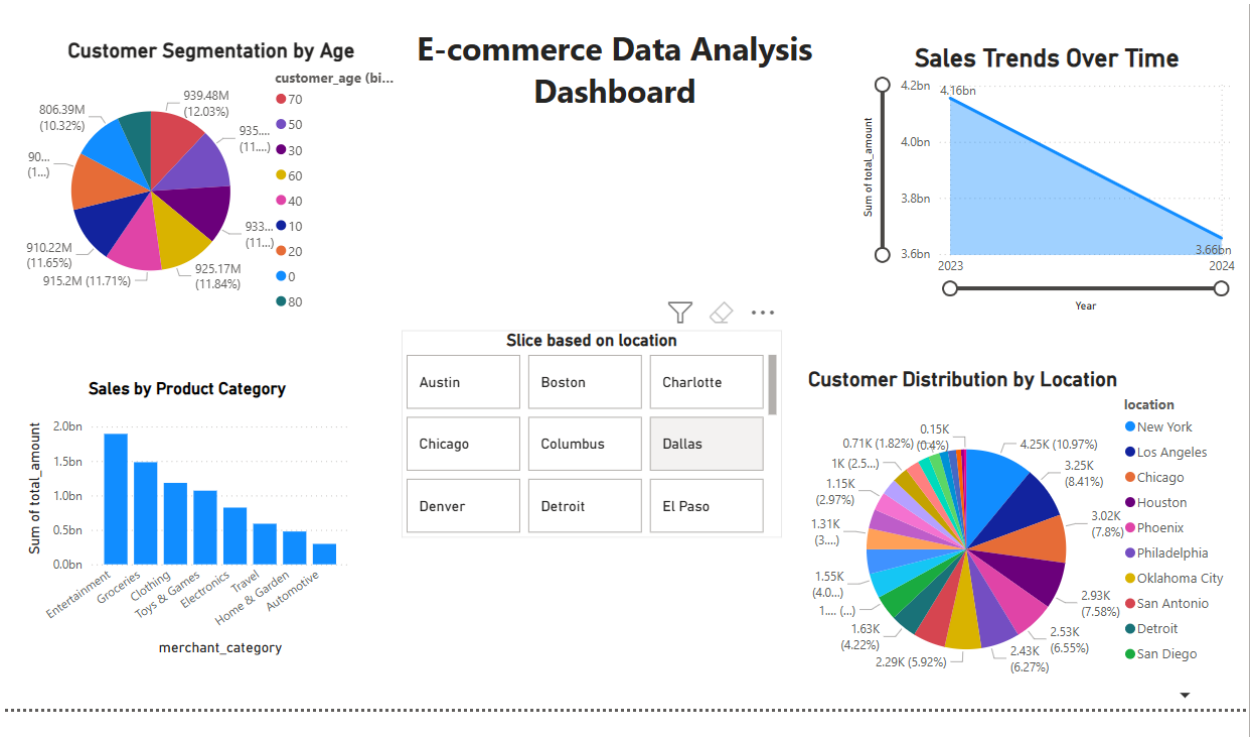
- Leveraged the powerful querying capabilities of DuckDB to efficiently process and transform the cleaned data.
- Created tables in DuckDB to facilitate data processing, ensuring optimal performance and ease of transformation.
- Employed SQLAlchemy to seamlessly load the cleaned and transformed data into the PostgreSQL database, maintaining data integrity and enabling efficient data management.

5. Data Visualization and Insights

Tools Used in Power BI:

- **Power BI Desktop** for creating and designing visualizations.
- **Connected Power BI with the SQL database** using the PostgreSQL connector to import cleaned and transformed data.

Visualizations on the Dashboard



1. **Top-Selling Products:**
   - Horizontal bar chart showing the count of product IDs for different merchant categories, with categories like Entertainment, Groceries, and Clothing leading the list.
2. **Sales by Product Category:**
   - Vertical bar chart showing the sum of total amounts for different merchant categories, reflecting the revenue contribution from each category.
3. **Sales Trends Over Time:**
   - Line chart tracking the sum of total amounts over the years 2023 to 2024, indicating overall sales performance and fluctuations.
4. **Customer Distribution by Location:**
   - Pie chart showing customer percentage distribution by location, highlighting top cities like New York, Los Angeles, and Chicago.
5. **Slice Based on Location:**
   - Slicer panel listing different cities for filtering data based on location.

Patterns and Trends

1. Sales Trends Over Time

- **Overall Sales Performance:**
  - Sales in 2023 reached approximately 4.2 billion, while 2024 sales are projected at 4.16 billion, indicating a slight decline of 10.32%.

    - This decline suggests a potential challenge in maintaining growth momentum, which could be due to market saturation, increased competition, or changing customer preferences.
- **Quarterly/Monthly Trends:**
    - Specific figures like 806.39 million and 910.22 million likely represent sales for specific periods (e.g., quarters or months).
    - The 11.65% and 11.71% changes indicate fluctuations in sales performance over time.
    - These fluctuations could reflect seasonal trends, promotional impacts, or external economic factors.

2. Customer Segmentation by Age

- The dashboard includes customer age groups, but exact ranges are unclear. However, age-based segmentation is a critical factor in understanding purchasing behavior.
    - **Potential Trend:** Younger age groups (e.g., millennials, Gen Z) may dominate online shopping, while older demographics might show slower adoption.

3. Sales by Product Category

- The total sales amount is 2.0 billion, distributed across various merchant categories.
    - **Pattern:** Certain product categories likely drive the majority of sales, while others underperform.

4. Customer Distribution by Location

- **Geographical Trends:**
    - Chicago stands out as the top city, contributing 32,000 customers (34.1%), followed by Houston (30,000, 7.8%), Phoenix (29,000, 7.58%), and San Antonio (25,000, 6.55%).
    - Other cities like Detroit, San Diego, and Philadelphia also show significant customer bases but with smaller shares.
    - **Pattern:** Urban areas with higher population densities tend to have larger customer bases, while smaller cities contribute less to overall sales.

Potential Business Insights

1. Declining Sales Growth

- The 10.32% decline in sales from 2023 to 2024 highlights a potential issue that needs further investigation. This could be due to factors such as customer churn, reduced marketing effectiveness, or external economic conditions.

2. High-Performing Regions

- Cities like Chicago, Houston, and Phoenix are key markets with large customer bases. These regions likely have strong brand awareness or effective local marketing strategies.

3. Underperforming Regions

- Smaller cities like Oklahoma City and El Paso have relatively low customer contributions. This could indicate untapped potential or barriers to customer acquisition in these areas.

4. Product Category Optimization

- The uneven distribution of sales across product categories suggests that some categories are more popular than others. Identifying these categories can provide insights into customer preferences and market demand.

5. Customer Age Segmentation

- Age-based segmentation reveals differences in purchasing behavior across demographics. Younger customers may prefer trendy or tech-savvy products, while older customers might prioritize reliability and convenience.

6. Seasonal or Periodic Fluctuations

- The 11.65% and 11.71% changes in sales suggest periodic fluctuations. These could be tied to seasonal trends, such as holiday shopping spikes or summer slumps.