CSC1109 – Supermarket Sales dataset analysis using Apache Pig and Hive

Cathal Ryan – 20302111

Git repo here

1. Introduction

Apache Hadoop is a framework for processing large-scale data distributed across multiple computers. Built on MapReduce, a processing technique with a divide-and-conquer approach, it enables parallel processing of big data. Apache offers additional projects within the Hadoop ecosystem, like Apache Pig and Apache Hive, which utilize MapReduce to facilitate data processing and querying. Apache Pig is a high-level language for parallel computation suited for cleaning and processing large datasets. Apache Hive is a data warehousing tool, allowing users to perform ad-hoc querying.

This project utilized Apache Pig and Hive to analyse supermarket sales data. The files used and created in this assignment are provided along with the code used to generate them. The dataset was cleaned and pre-processed using Pig, and more advanced queries were conducted in Hive to derive insights. The following sections outline the steps taken at each stage and present an analysis of the query results within the supermarket sales dataset.

2. Data Cleaning

Before conducting any analysis, the supermarket sales dataset needed to be cleaned and transformed into a suitable format for querying. This process was performed locally using Apache Pig. The general steps and issues addressed are outlined below.

- CSVLoader was used to correctly load the data without splitting columns on quoted commas, ensuring data consistency.
- The header row containing column names was removed.
- Transaction dates were reformatted into DD-MM-YYYY format to facilitate chronological analysis.
- Rows with critical missing data were then filtered out.
- Once the data was cleaned and reformatted, the processed dataset was saved in outputs/clean_supermarket_sales. This output was structured to retain only essential attributes. After preparing the data in Pig, I used Hive for further in-depth analysis on this cleaned dataset.

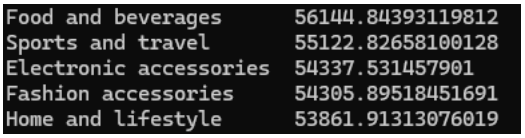
3. Data Querying

I first started with two simple queries in both Pig and Hive. I will show both results together with some analysis.

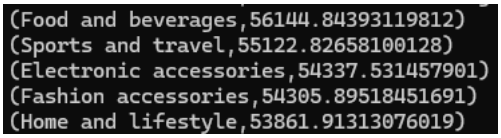
- *Total Sales By Product Line*

The first query calculates total sales by each product line, giving an overview of which product categories contribute the most to revenue. This query groups the data by ProductLine and sums up the Total sales in each category, then orders the results in descending order by TotalSales and limits the output to the top 5 product lines.

Hive output:



Pig output:

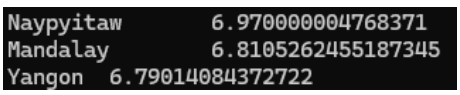


We can see in the two outputs that the results match. We can also see that Food and beverages contribute the most to revenue followed by Sport and travel, then Electronic accessories. Interestingly, there is very little difference between each of the categories.

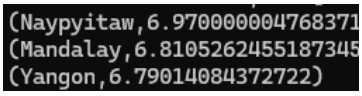
- *Average Rating By City for Orders Above 500*

The second query finds the average customer rating by city for orders over 500. This filters records to include only orders with Total greater than 500, groups them by City, calculates the average Rating, and sorts the results by the highest AverageRating values.

Hive output:



Pig output:



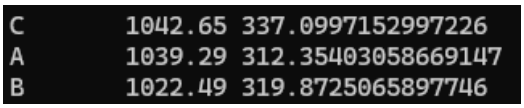
The Hive and Pig queries both reveal that the cities Naypyitaw and Mandalay have higher average ratings, which could imply stronger customer satisfaction or loyalty in these areas for high-value orders.

Now I will look at more advanced queries in hive.

- Maximum Total Sales Per Branch

For the first complex query I am using the MAX function and AVG function to find the highest transaction and average transaction for each branch. The results will then be ordered by the branches with the highest individual sales transaction.

Output:

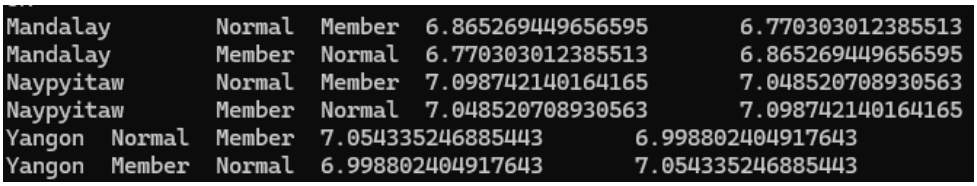


We can see from the analysis that branches C and A have slightly higher max transactions. However branch B has a higher average transaction than branch A while branch C also has the highest average transaction of the three.

- Average Rating Between Different Customer Types

For the second complex query I am performing a self-join to analyse and compare the average rating between different customer types (Member vs. Normal) in each city. This helps understand if there's a notable difference in customer satisfaction based on membership.

Output:

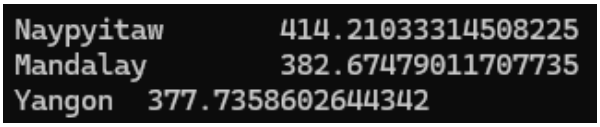


We can see from this output that the average ratings are very similar between normal customers and members. The small variations in ratings between customer types suggest a generally consistent customer experience.

- Average Total Over 100 By City

For the third complex query I will be using TABLESAMPLE to randomly sample the data to limit data processing. The query calculates the average total sales for orders greater than 100 in each city.

Output:



We can see that Naypyitaw has the highest average sales and Mandalay and Yangon with similar average sales. Since this result is from a sample (20% of the data), it provides a quick estimate of the average sales by city.

4. Conclusion

Through the use of Apache Pig and Hive, this analysis offers insights into the sales performance, customer satisfaction, and spending patterns within the supermarket dataset. Using Pig, data was pre-processed to remove inaccuracies, ensuring a clean, structured dataset that enabled reliable querying in Hive.

Some of the key findings from the Hive analysis include:

- The Food and Beverages category emerged as the top contributor to total sales, followed closely by Sports and Travel. This information is valuable for understanding which product lines drive revenue.

- High-value orders (over 500) show that cities like Naypyitaw and Mandalay have the highest average customer ratings. This trend could indicate higher customer satisfaction or loyalty in these regions, suggesting potential areas for targeted customer retention strategies.

- Complex queries further revealed that Branch C has the highest average and maximum transaction values, and a self-join analysis highlighted slight differences in customer satisfaction between Member and Normal customers. This indicates a generally consistent customer experience across membership types, affirming that the service quality does not significantly vary by customer type.

- Using a sample of the dataset to calculate average sales above 100 per city provided an efficient means of identifying spending trends without processing the full dataset. Naypyitaw showed the highest average sales, confirming it as a high-revenue area.

In summary, the use of Pig for data cleaning and Hive for analysis on this supermarket dataset has demonstrated the power of these tools for uncovering valuable business insights.