Coffee Sales Analysis Using Apache Hive

Busyra Binti Sukria[1], Nur Iriana Binti Muhamad Shukri[2], Mutmainnah Radiah Binti Jamal Abdul Hekim[3], Ros Sara Elysa Binti Ros Fauzi[4]

Computing Science Studies, College of Computing, Informatics and Mathematics
Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia

ABSTRACT

The rapid advancements in data analytics have revolutionized the way businesses make informed decisions. This project focuses on analyzing the coffee sales industry through the development of a comprehensive analytics system. The dataset, sourced from Kaggle, is processed using an Extract, Transform, Load (ETL) pipeline to ensure data accuracy and usability. The project integrates predictive models and advanced visualizations to identify key sales patterns, optimize inventory management, and enhance decision-making processes. Furthermore, a range of exploratory queries and analytics, such as determining the most profitable markets and evaluating marketing impact, provides actionable insights. This innovative approach seeks to assist stakeholders in improving operational efficiency, maximizing profitability, and addressing market demands effectively.

Keywords: *coffee sales; ETL; big data technologies; RDBMS;*

1.0 Introduction

This project focuses on leveraging data analytics to enhance decision-making processes within the coffee sales industry. Utilizing an Extract, Transform, Load (ETL) pipeline, it ensures accurate and clean data for analysis. The project explores various aspects such as sales trends, profit margins, and customer purchasing behaviors. Advanced analytics and predictive models provide actionable insights into optimizing inventory management, identifying profitable markets, and evaluating marketing effectiveness. By integrating exploratory queries and visualization techniques, this initiative empowers stakeholders to improve operational efficiency, maximize profitability, and address market demands effectively.

1.1. Background

The rapid evolution of data technologies has transformed industries, enabling organizations to leverage large-scale data for actionable insights. This project focuses on the coffee sales industry, exploring key performance indicators such as sales trends, profit margins, and customer purchasing behaviours. By utilizing modern tools and techniques, the project seeks to optimize decision-making processes through data analytics.

1.2. Objectives

1. To analyse sales trends, evaluate marketing effectiveness, and optimize operations for better decision-making in coffee sales.

2. To identify key sales patterns, improve profitability, and enhance marketing and inventory management.
3. To explore customer purchasing behaviors and preferences to design targeted marketing strategies and improve customer satisfaction.

2.0 Requirements Analysis

The requirements analysis phase establishes a thorough understanding of the tools, datasets, and processes essential for this project. By aligning technical capabilities with the project's goals, this phase creates a solid foundation for efficient implementation and analysis.

2.1 Source the dataset

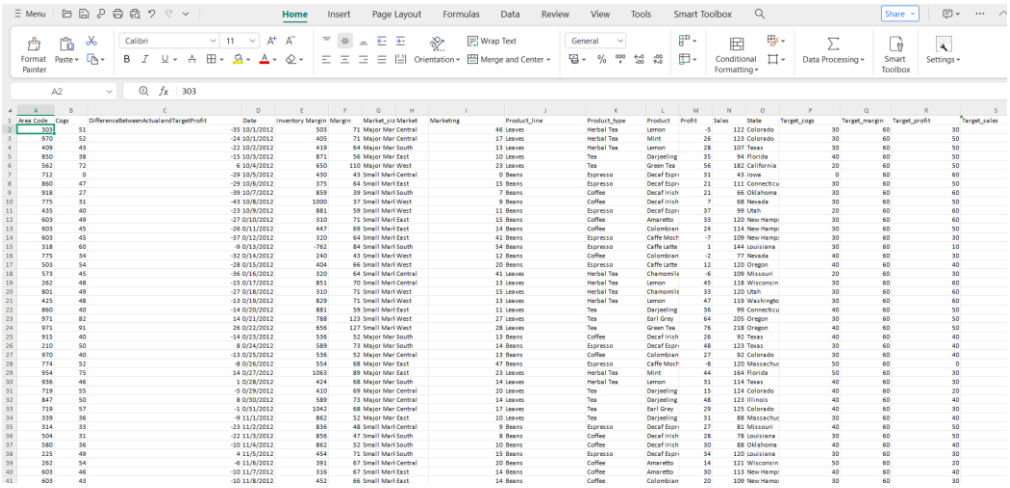


Figure 2.1 Coffee.csv dataset

Figure 2.1 shows the dataset for this project which is “coffee.csv,". The dataset is obtained from Kaggle. It contains 1,062 entries and 21 columns, providing detailed insights into the coffee sales industry. It includes product details such as Product, Product_line, and Product_type, along with market information like Market, Market_size, and State. Financial metrics, including Sales, Profit, Cogs (Cost of Goods Sold), Margin, and Total_expenses, are complemented by target metrics like Target_sales, Target_profit, and Target_margin. The dataset also features Marketing and Difference Between Actual and Target Profit to assess campaign effectiveness and profitability gaps. Additionally, Date records transaction timelines, and Area Code serves as a regional identifier, making the dataset a comprehensive resource for analyzing sales trends, market dynamics, and operational performance.

2.2 Data Analysis Requirement

To meet the project's objectives, the queries for the following will be implemented as shown below.

1. Determine which market has the highest sales.
2. Identify the top 5 products with the highest total sales.
3. Display the average profit margin for each product.
4. Display products with the highest marketing spend.
5. Determine which market has the highest sales (repeated).
6. Evaluate whether actual profit meets or exceeds the target profit for each market.
7. Examine the total expenses incurred by each market.
8. Calculate the total number of transactions for each market in the coffee new table.
9. Display distinct products sold in each market.
10. Identify the least profitable products in each market.

2.3 System Requirements

The system requirements for this project are tailored to support the efficient processing and analysis of large-scale datasets using big data tools like Apache Hive and Hadoop.

2.3.1 Big Data Tools and Platform

Apache Hive is a powerful tool for querying and analyzing big data, offering flexibility and efficiency for large-scale data processing. It extends traditional data warehousing capabilities by operating on top of Hadoop, enabling the handling of structured, semi-structured, and unstructured data (Małysiak-Mrozek et al., 2022). Hive's performance can be optimized through strategic data organization, with partitioning showing significant benefits in query response times (Costa et al., 2019). Hive also shows promise in handling complex XML schemas, utilizing techniques such as cataloging, deserialization, and positional explode for efficient processing (Chugh, 2021).

2.3.2 Hardware Requirements

A standard laptop is adequate for running the required tools and software for this project. The recommended hardware specifications are as stated below.

- **Processor**: At least an Intel i5 or an equivalent processor
- **Storage**: A 256GB SSD or larger to accommodate the dataset and essential tools
- **Operating System**: Windows.
- **RAM**: 16GB for optimal performance

2.3.3 Software Requirements

The software requirements for this project include VMware running Cloudera, which provides a robust virtualized environment for managing and processing big data. Cloudera offers a comprehensive ecosystem for data analytics, integrating tools like Apache Hadoop and Apache Hive for efficient data storage, processing, and querying. Hadoop serves as the backbone for distributed data management, enabling the handling of large datasets across multiple nodes. Hive, built on top of Hadoop, simplifies data analysis by providing a SQL-like interface, making it easier to query, summarize, and analyze structured data. Together, these tools ensure seamless data processing and analysis within a scalable and efficient framework.

3.0 Database Design

The database design focuses on creating a structured and efficient schema to manage and analyze the coffee sales data, ensuring seamless integration between entities and enabling accurate and meaningful insights.

3.1 Data Model

An Entity Relationship Diagram is developed for structured data

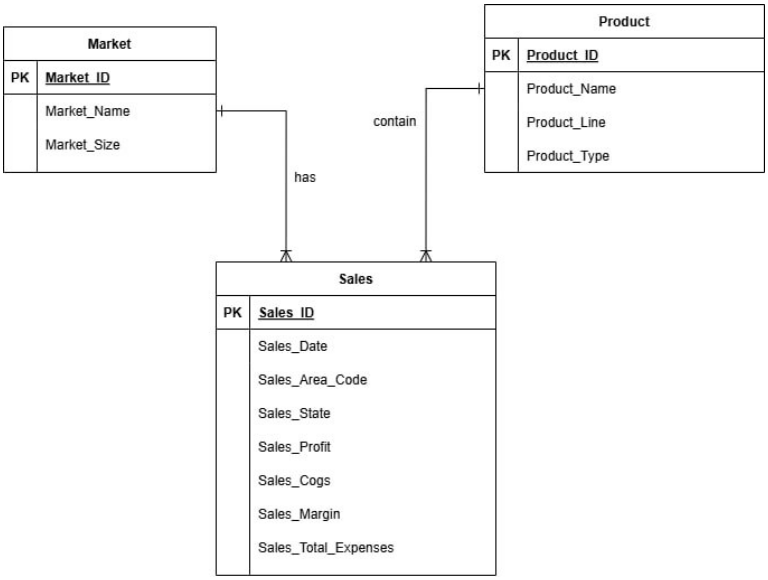


Figure 3.1 Entity Relationship Diagram

The Entity-Relationship Diagram (ERD) in Figure 3.1 illustrates the relationships between three main entities: Market, Product, and Sales. The Market entity, identified by Market_ID, contains attributes like Market_Name and Market_Size, representing market details. The Product entity, identified by Product_ID, includes attributes such as Product_Name, Product_Line, and Product_Type, defining product characteristics. The Sales entity, identified by Sales_ID, serves as the central table linking markets and products, with attributes like Sales_Date, Sales_Profit, Sales_Cogs, and Sales_Total_Expenses. A one-to-many relationship exists between Market and Sales, and between Product and Sales, allowing a market or product to be associated with multiple sales records.

This design enables comprehensive analysis of sales performance, market trends, and product profitability.

3.2 ETL Process

The ETL process is fundamental in ensuring that data is accurate, consistent, and organized, enabling efficient analytics and reporting.

3.2.1 Extraction

The extraction phase in data science involves retrieving and importing raw data from sources like datasets, ensuring accuracy and integrity (Mahalle et al., 2021). Proper data extraction is essential for systematic reviews and meta-analyses, as it impacts the quality of inputs and, consequently, the reliability of conclusions. Good practices in data extraction help reduce errors and bias, ensuring efficient processes and providing sufficient information for readers to assess the generalizability of findings (Taylor et al., 2021).

3.2.2 Transforming

The transformation phase focuses on cleaning, organizing, and converting the raw data into a structured format suitable for analysis. This involves cleaning, integrating, and restructuring data from various sources (Fernandes et al., 2023). The process includes handling missing or inconsistent data, standardizing formats, and deriving new metrics (Azeroual, 2020). HiveQL, an extension of Apache Hive, can be utilized for efficient querying and transformation of big data warehouses (Bożena Małysiak-Mrozek et al., 2022).

3.2.3 Load

In the loading phase, the processed and transformed data is stored in a Hadoop-based distributed storage system for scalability and reliability. The structured data is organized in tables within Apache Hive, allowing users to query and analyze the data efficiently. This phase ensures that the cleaned and structured data is readily available for running analytics, generating insights, and visualizing trends.

4.0 Implementation

During the implementation phase, the "Coffee.csv" dataset was loaded into Apache Hive, enabling efficient data processing and querying. The initial step involved creating a dedicated database and tables within Hive to organize the data. Using HiveQL, a series of queries were executed to transform and analyze the dataset, addressing project objectives such as identifying top-selling products, evaluating marketing effectiveness, and determining market profitability. The data was processed in batches using the Hadoop Distributed File System (HDFS) to ensure scalability and reliability. This phase also included generating insights through advanced queries, preparing the dataset for visualization and decision-making.

4.1 Data Preprocessing and Transformation

The data preprocessing and transformation phase focused on preparing the "Coffee.csv" dataset for efficient analysis in Apache Hive. This began with cleaning the dataset by addressing missing values, standardizing data formats, and removing inconsistencies to ensure accuracy. Key transformations included calculating derived metrics such as profit margins and the difference between actual and target sales, as well as categorizing products based on their market performance. HiveQL was utilized extensively to execute these transformations, including filtering, aggregating, and restructuring the data into meaningful formats. The transformed data was then organized into well-structured tables within Hive, ensuring it was ready for the querying and analysis phase.

Data loading

1. Create database coffee_sales

```
0: jdbc:hive2://localhost:10000> CREATE database coffee_sales
. . . . . . . . . . . . . . . .> ;
INFO  : Compiling command(queryId=hive_20250117021616_17d64dcd-f9c7-47bd-bb24-72
dead16e0a5): CREATE database coffee_sales
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250117021616_17d64dcd-f9c7-47
bd-bb24-72dead16e0a5); Time taken: 0.22 seconds
INFO  : Executing command(queryId=hive_20250117021616_17d64dcd-f9c7-47bd-bb24-72
dead16e0a5): CREATE database coffee_sales
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117021616_17d64dcd-f9c7-47
bd-bb24-72dead16e0a5); Time taken: 0.084 seconds
INFO  : OK
No rows affected (0.4 seconds)
```

Figure 4.1 Query to create database coffee_sales

2. Show the databases

```
f9c162a9f2): SHOW DATABASES
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117021919_6976b858-5091-45
89-893c-3cf9c162a9f2); Time taken: 0.01 seconds
INFO  : OK
+----------------+--+
| database_name  |
+----------------+--+
| 360dt          |
| coffee_sales   |
| default        |
| retail_db      |
| retail_db2     |
| us_death       |
+----------------+--+
6 rows selected (0.132 seconds)
```

Figure 4.2 Query to show the databases

3. Use the coffee_sales database

```
f6ed222410): use coffee_sales
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250117022121_417c183c-375a-4f
25-a1e2-86f6ed222410); Time taken: 0.011 seconds
INFO  : Executing command(queryId=hive_20250117022121_417c183c-375a-4f25-a1e2-86
f6ed222410): use coffee_sales
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117022121_417c183c-375a-4f
25-a1e2-86f6ed222410); Time taken: 0.01 seconds
INFO  : OK
No rows affected (0.04 seconds)
```

Figure 4.3 Query to use the coffee_sales database

4. Create table Sales

```
0: jdbc:hive2://localhost:10000> CREATE table sales(col_value string);
INFO  : Compiling command(queryId=hive_20250117031414_9b188c55-2908-4ac1-8dff-41
c3408a62f0): CREATE table sales(col_value string)
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250117031414_9b188c55-2908-4a
c1-8dff-41c3408a62f0); Time taken: 0.018 seconds
INFO  : Executing command(queryId=hive_20250117031414_9b188c55-2908-4ac1-8dff-41
c3408a62f0): CREATE table sales(col_value string)
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117031414_9b188c55-2908-4a
c1-8dff-41c3408a62f0); Time taken: 0.067 seconds
INFO  : OK
No rows affected (0.125 seconds)
```

Figure 4.4 Query to create table sales

5. Load the data

```
0: jdbc:hive2://localhost:10000> LOAD DATA local INPATH '/home/cloudera/coffee/c
offee.csv' OVERWRITE INTO TABLE sales;
INFO  : Compiling command(queryId=hive_20250117052525_469f406a-fc0d-454b-8547-70
72409e23ab): LOAD DATA local INPATH '/home/cloudera/coffee/coffee.csv' OVERWRITE
 INTO TABLE sales
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250117052525_469f406a-fc0d-45
4b-8547-7072409e23ab); Time taken: 0.014 seconds
INFO  : Executing command(queryId=hive_20250117052525_469f406a-fc0d-454b-8547-70
72409e23ab): LOAD DATA local INPATH '/home/cloudera/coffee/coffee.csv' OVERWRITE
 INTO TABLE sales
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table coffee_sales.sales from file:/home/cloudera/coffee
/coffee.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Table coffee_sales.sales stats: [numFiles=1, numRows=0, totalSize=122980
, rawDataSize=0]
INFO  : Completed executing command(queryId=hive_20250117052525_469f406a-fc0d-45
4b-8547-7072409e23ab); Time taken: 0.272 seconds
INFO  : OK
No rows affected (0.343 seconds)
```

Figure 4.5 Query to load the data

6. Create table coffee_new

```
0: jdbc:hive2://localhost:10000> CREATE TABLE coffee_new (
. . . . . .>     area_code INT,
. . . . . .>     cogs INT,
. . . . . .>     difference_between_actual_and_target_profit
 INT,
. . . . . .>     date STRING,
. . . . . .>     inventory_margin INT,
. . . . . .>     margin INT,
. . . . . .>     market_size STRING,
. . . . . .>     market STRING,
. . . . . .>     marketing INT,
. . . . . .>     product_line STRING,
. . . . . .>     product_type STRING,
. . . . . .>     product STRING,
. . . . . .>     profit INT,
. . . . . .>     sales INT,
. . . . . .>     state STRING,
. . . . . .>     target_cogs INT,
. . . . . .>     target_margin INT,
. . . . . .>     target_profit INT,
. . . . . .>     target_sales INT,
. . . . . .>     total_expenses INT,
. . . . . .>     type STRING
. . . . . .> );
INFO  : Compiling command(queryId=hive_20250117053737_cc517559-3c88-4ac0-9134-d3
eba1d7e3ac): CREATE TABLE coffee_new (
area_code INT,
cogs INT,
difference_between_actual_and_target_profit INT,
date STRING,
inventory_margin INT,
margin INT,
market_size STRING,
market STRING,
marketing INT,
product_line STRING,
product_type STRING,
product STRING,
profit INT,
sales INT,
state STRING,
target_cogs INT,
target_margin INT,
target_profit INT,
target_sales INT,
total_expenses INT,
type STRING
)
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117053434_3bd9d5b0-a81d-47
58-82b4-9b6cc8d4f9c6); Time taken: 0.04 seconds
INFO  : OK
No rows affected (0.121 seconds)
0: jdbc:hive2://localhost:10000> DROP TABLE coffee_master;
INFO  : Compiling command(queryId=hive_20250117053535_c5317ede-7aed-444d-a7ae-76
78d749cb1c): DROP TABLE coffee_master
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20250117053535_c5317ede-7aed-44
4d-a7ae-7678d749cb1c); Time taken: 0.123 seconds
INFO  : Executing command(queryId=hive_20250117053535_c5317ede-7aed-444d-a7ae-76
78d749cb1c): DROP TABLE coffee_master
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117053535_c5317ede-7aed-44
4d-a7ae-7678d749cb1c); Time taken: 0.081 seconds
INFO  : OK
No rows affected (0.222 seconds)
```

Figure 4.6 Query to create table coffee_new

7. Insert overwrite table

```
0: jdbc:hive2://localhost:10000> INSERT OVERWRITE TABLE coffee_new
. . . . . .> SELECT
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){1}', 1) AS INT) AS area_code,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){2}', 1) AS INT) AS cogs,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){3}', 1) AS INT) AS difference_between_actual_and_target_profit,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){4
}', 1) AS date,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){5}', 1) AS INT) AS inventory_margin,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){6}', 1) AS INT) AS margin,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){7
}', 1) AS market_size,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){8
}', 1) AS market,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){9}', 1) AS INT) AS marketing,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){1
0}', 1) AS product_line,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){1
1}', 1) AS product_type,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){1
2}', 1) AS product,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){13}', 1) AS INT) AS profit,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){14}', 1) AS INT) AS sales,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){1
5}', 1) AS state,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){16}', 1) AS INT) AS target_cogs,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){17}', 1) AS INT) AS target_margin,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){18}', 1) AS INT) AS target_profit,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){19}', 1) AS INT) AS target_sales,
. . . . . .>     CAST(regexp_extract(col_value, '^(?:([^,]*)
,?){20}', 1) AS INT) AS total_expenses,
. . . . . .>     regexp_extract(col_value, '^(?:([^,]*),?){2
1}', 1) AS type
. . . . . .> FROM sales
. . . . . .> WHERE col_value IS NOT NULL
. . . . . .>   AND size(split(col_value, ',')) >= 21 -- Ensu
re the row has at least 21 columns
. . . . . .>   AND NOT (col_value LIKE '%,,%' OR col_value R
LIKE '(^|,)(NULL|null)(,|$)');
INFO  : Compiling command(queryId=hive_20250117054343_7fba5cdb-659d-4f6d-b4af-16
5c4f631e9b): INSERT OVERWRITE TABLE coffee_new
SELECT
CAST(regexp_extract(col_value, '^(?:([^,]*),?){1}', 1) AS INT) AS area_code,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){2}', 1) AS INT) AS cogs,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){3}', 1) AS INT) AS difference_bet
ween_actual_and_target_profit,
regexp_extract(col_value, '^(?:([^,]*),?){4}', 1) AS date,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){5}', 1) AS INT) AS inventory_marg
in,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){6}', 1) AS INT) AS margin,
regexp_extract(col_value, '^(?:([^,]*),?){7}', 1) AS market_size,
regexp_extract(col_value, '^(?:([^,]*),?){8}', 1) AS market,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){9}', 1) AS INT) AS marketing,
regexp_extract(col_value, '^(?:([^,]*),?){10}', 1) AS product_line,
regexp_extract(col_value, '^(?:([^,]*),?){11}', 1) AS product_type,
regexp_extract(col_value, '^(?:([^,]*),?){12}', 1) AS product,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){13}', 1) AS INT) AS profit,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){14}', 1) AS INT) AS sales,
regexp_extract(col_value, '^(?:([^,]*),?){15}', 1) AS state,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){16}', 1) AS INT) AS target_cogs,
```

Figure 4.7 Query to insert overwrite table

```
CAST(regexp_extract(col_value, '^(?:([^,]*),?){17}', 1) AS INT) AS target_margin
,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){18}', 1) AS INT) AS target_profit
,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){19}', 1) AS INT) AS target_sales,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){20}', 1) AS INT) AS total_expense
s,
regexp_extract(col_value, '^(?:([^,]*),?){21}', 1) AS type
FROM sales
WHERE col_value IS NOT NULL
AND size(split(col_value, ',')) >= 21
AND NOT (col_value LIKE '%,,%' OR col_value RLIKE '(^|,)(NULL|null)(,|$)')
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:area_code,
type:int, comment:null), FieldSchema(name:cogs, type:int, comment:null), FieldSc
hema(name:difference_between_actual_and_target_profit, type:int, comment:null),
FieldSchema(name:date, type:string, comment:null), FieldSchema(name:inventory_ma
rgin, type:int, comment:null), FieldSchema(name:margin, type:int, comment:null),
 FieldSchema(name:market_size, type:string, comment:null), FieldSchema(name:mark
et, type:string, comment:null), FieldSchema(name:marketing, type:int, comment:nu
ll), FieldSchema(name:product_line, type:string, comment:null), FieldSchema(name
:product_type, type:string, comment:null), FieldSchema(name:product, type:string
, comment:null), FieldSchema(name:profit, type:int, comment:null), FieldSchema(n
ame:sales, type:int, comment:null), FieldSchema(name:state, type:string, comment
:null), FieldSchema(name:target_cogs, type:int, comment:null), FieldSchema(name:
target_margin, type:int, comment:null), FieldSchema(name:target_profit, type:int
, comment:null), FieldSchema(name:target_sales, type:int, comment:null), FieldSc
hema(name:total_expenses, type:int, comment:null), FieldSchema(name:type, type:s
tring, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117054343_7fba5cdb-659d-4f
6d-b4af-165c4f631e9b); Time taken: 0.226 seconds
INFO  : Executing command(queryId=hive_20250117054343_7fba5cdb-659d-4f6d-b4af-16
5c4f631e9b): INSERT OVERWRITE TABLE coffee_new
SELECT
CAST(regexp_extract(col_value, '^(?:([^,]*),?){1}', 1) AS INT) AS area_code,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){2}', 1) AS INT) AS cogs,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){3}', 1) AS INT) AS difference_bet
ween_actual_and_target_profit,
regexp_extract(col_value, '^(?:([^,]*),?){4}', 1) AS date,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){5}', 1) AS INT) AS inventory_marg
in,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){6}', 1) AS INT) AS margin,
regexp_extract(col_value, '^(?:([^,]*),?){7}', 1) AS market_size,
regexp_extract(col_value, '^(?:([^,]*),?){8}', 1) AS market,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){9}', 1) AS INT) AS marketing,
regexp_extract(col_value, '^(?:([^,]*),?){10}', 1) AS product_line,
regexp_extract(col_value, '^(?:([^,]*),?){11}', 1) AS product_type,
regexp_extract(col_value, '^(?:([^,]*),?){12}', 1) AS product,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){13}', 1) AS INT) AS profit,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){14}', 1) AS INT) AS sales,
regexp_extract(col_value, '^(?:([^,]*),?){15}', 1) AS state,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){16}', 1) AS INT) AS target_cogs,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){17}', 1) AS INT) AS target_margin
,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){18}', 1) AS INT) AS target_profit
,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){19}', 1) AS INT) AS target_sales,
CAST(regexp_extract(col_value, '^(?:([^,]*),?){20}', 1) AS INT) AS total_expense
s,
regexp_extract(col_value, '^(?:([^,]*),?){21}', 1) AS type
FROM sales
WHERE col_value IS NOT NULL
AND size(split(col_value, ',')) >= 21
AND NOT (col_value LIKE '%,,%' OR col_value RLIKE '(^|,)(NULL|null)(,|$)')
INFO  : Query ID = hive_20250117054343_7fba5cdb-659d-4f6d-b4af-165c4f631e9b
INFO  : Total jobs = 3
INFO  : Launching Job 1 out of 3
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks is set to 0 since there's no reduce operator
```

Figure 4.7 Query to insert overwrite table (continued)

```
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0009
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/applicat
ion_1737099421907_0009/
INFO  : Starting Job = job_1737099421907_0009, Tracking URL = http://quickstart.
cloudera:8088/proxy/application_1737099421907_0009/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0
009
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of redu
cers: 0
INFO  : 2025-01-17 05:43:46,690 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 05:43:51,953 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU
 3.12 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 120 msec
INFO  : Ended Job = job_1737099421907_0009
INFO  : Starting task [Stage-7:CONDITIONAL] in serial mode
INFO  : Stage-4 is selected by condition resolver.
INFO  : Stage-3 is filtered out by condition resolver.
INFO  : Stage-5 is filtered out by condition resolver.
INFO  : Starting task [Stage-4:MOVE] in serial mode
INFO  : Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/coff
ee_sales.db/coffee_new/.hive-staging_hive_2025-01-17_05-43-39_990_73820924050031
1421-8/-ext-10000 from hdfs://quickstart.cloudera:8020/user/hive/warehouse/coffe
e_sales.db/coffee_new/.hive-staging_hive_2025-01-17_05-43-39_990_738209240500311
421-8/-ext-10002
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table coffee_sales.coffee_new from hdfs://quickstart.clo
udera:8020/user/hive/warehouse/coffee_sales.db/coffee_new/.hive-staging_hive_202
5-01-17_05-43-39_990_73820924050031142l-8/-ext-10000
INFO  : Starting task [Stage-2:STATS] in serial mode
INFO  : Table coffee_sales.coffee_new stats: [numFiles=1, numRows=1061, totalSiz
e=121919, rawDataSize=120858]
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1   Cumulative CPU: 3.12 sec   HDFS Read: 131451 HDF
S Write: 122004 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 3 seconds 120 msec
INFO  : Completed executing command(queryId=hive_20250117054343_7fba5cdb-659d-4f
6d-b4af-165c4f631e9b); Time taken: 14.132 seconds
```

Figure 4.7 Query to insert overwrite table (continued)

8. Show table properties

```
4f4323c97b): show TBLPROPERTIES coffee_new
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:prpt_name,
type:string, comment:from deserializer), FieldSchema(name:prpt_value, type:strin
g, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117054747_28b0ad6f-06ce-4b
6b-9f14-8e4f4323c97b); Time taken: 0.034 seconds
INFO  : Executing command(queryId=hive_20250117054747_28b0ad6f-06ce-4b6b-9f14-8e
4f4323c97b): show TBLPROPERTIES coffee_new
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250117054747_28b0ad6f-06ce-4b
6b-9f14-8e4f4323c97b); Time taken: 0.017 seconds
INFO  : OK
+------------------------+-------------+--+
|       prpt_name        | prpt_value  |
+------------------------+-------------+--+
| COLUMN_STATS_ACCURATE  | true        |
| numFiles               | 1           |
| numRows                | 1061        |
| rawDataSize            | 120858      |
| totalSize              | 121919      |
| transient_lastDdlTime  | 1737121434  |
+------------------------+-------------+--+
6 rows selected (0.071 seconds)
```

Figure 4.8 Query to show table properties

4.2 Data Analytics

1. Query to determine which market has the highest sales

```
0: jdbc:hive2://localhost:10000> SELECT Market, SUM(Sales) AS Total_Sales
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Market
. . . . . .> ORDER BY Total_Sales DESC;
INFO  : Compiling command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4): SELECT Market, SUM(Sal
es) AS Total_Sales
FROM coffee_new
GROUP BY Market
ORDER BY Total_Sales DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchem
a(name:total_sales, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4); Time taken:
0.163 seconds
INFO  : Executing command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4): SELECT Market, SUM(Sal
es) AS Total_Sales
FROM coffee_new
GROUP BY Market
ORDER BY Total_Sales DESC
INFO  : Query ID = hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0016
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0016/
INFO  : Starting Job = job_1737099421907_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173
7099421907_0016/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0016
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 07:24:33,077 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 07:24:38,320 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.37 sec
INFO  : 2025-01-17 07:24:45,760 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.39 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 390 msec
INFO  : Ended Job = job_1737099421907_0017
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.31 sec   HDFS Read: 131360 HDFS Write: 201 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.39 sec   HDFS Read: 5156 HDFS Write: 48 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 6 seconds 700 msec
INFO  : Completed executing command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4); Time taken:
42.692 seconds
INFO  : OK
```

```
+----------+--------------+--+
|  market  | total_sales  |
+----------+--------------+--+
| West     | 67418        |
| Central  | 64859        |
| East     | 44108        |
| South    | 26388        |
+----------+--------------+--+
4 rows selected (42.922 seconds)
```

Figure 4.9 Query to determine which market has the highest sales

2. Query to identify the top 5 products with the highest total sales

```
0: jdbc:hive2://localhost:10000> SELECT Product, SUM(Sales) AS Total_Sales
. . . . . . . . . . . . . . . .> FROM coffee_new
. . . . . . . . . . . . . . . .> GROUP BY Product
. . . . . . . . . . . . . . . .> ORDER BY Total_Sales DESC
. . . . . . . . . . . . . . . .> LIMIT 5;
INFO  : Compiling command(queryId=hive_20250117064747_529d9704-65fe-4cb5-87ec-93eceb9b4d65): SELECT
Product, SUM(Sales) AS Total_Sales
FROM coffee_new
GROUP BY Product
ORDER BY Total_Sales DESC
LIMIT 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:product, type:string, comment:n
ull), FieldSchema(name:total_sales, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117064747_529d9704-65fe-4cb5-87ec-93eceb9b4d65
); Time taken: 0.109 seconds
INFO  : Executing command(queryId=hive_20250117064747_529d9704-65fe-4cb5-87ec-93eceb9b4d65): SELECT
Product, SUM(Sales) AS Total_Sales
FROM coffee_new
GROUP BY Product
ORDER BY Total_Sales DESC
LIMIT 5
INFO  : Query ID = hive_20250117064747_529d9704-65fe-4cb5-87ec-93eceb9b4d65
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0012
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application 1737099421907 00    _
12/
INFO  : Starting Job = job_1737099421907_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1737099421907_0012/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0012
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 06:48:04,274 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 06:48:09,674 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.72 sec
INFO  : 2025-01-17 06:48:17,066 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.67 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 670 msec
INFO  : Ended Job = job_1737099421907_0012
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0013
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_00
13/
INFO  : Starting Job = job_1737099421907_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1737099421907_0013/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0013
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 06:48:25,153 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 06:48:31,493 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.38 sec
INFO  : 2025-01-17 06:48:37,846 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.45 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 450 msec
INFO  : Ended Job = job_1737099421907_0013
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.67 sec   HDFS Read: 131353 HDFS Write:
501 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.45 sec   HDFS Read: 5581 HDFS Write: 83
 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 7 seconds 120 msec
INFO  : Completed executing command(queryId=hive_20250117064747_529d9704-65fe-4cb5-87ec-93eceb9b4d65
); Time taken: 41.596 seconds
INFO  : OK
+-----------------+--------------+--+
|     product     | total_sales  |
+-----------------+--------------+--+
| Colombian       | 30761        |
| Lemon           | 23926        |
| Caffe Mocha     | 21716        |
| Chamomile       | 19295        |
| Decaf Espresso  | 18888        |
+-----------------+--------------+--+
5 rows selected (41.765 seconds)
```

Figure 4.10 Query to identify the top 5 products with the highest total sales

3. Query to display average profit margin for each product.

```
7099421907_0052/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0052
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 09:06:45,347 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 09:06:51,663 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.59 sec
INFO  : 2025-01-17 09:06:59,079 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.66 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 660 msec
INFO  : Ended Job = job_1737099421907_0052
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0053
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0053/
INFO  : Starting Job = job_1737099421907_0053, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173
7099421907_0053/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0053
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 09:07:07,240 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 09:07:13,555 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.54 sec
INFO  : 2025-01-17 09:07:18,834 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.45 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 450 msec
INFO  : Ended Job = job_1737099421907_0053
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.66 sec   HDFS Read: 131814 HDFS Write: 566 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.45 sec   HDFS Read: 5543 HDFS Write: 341 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 7 seconds 110 msec
INFO  : Completed executing command(queryId=hive_20250117090606_0cb8be3e-e1aa-4a14-a206-3fb3053d1f1e); Time taken:
42.962 seconds
INFO  : OK
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> SELECT Product, AVG(Margin) AS Avg_Profit_Margin
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Product
. . . . . .> ORDER BY Avg_Profit_Margin DESC;
INFO  : Compiling command(queryId=hive_20250117090606_0cb8be3e-e1aa-4a14-a206-3fb3053d1f1e): SELECT Product, AVG(Ma
rgin) AS Avg_Profit_Margin
FROM coffee_new
GROUP BY Product
ORDER BY Avg_Profit_Margin DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:product, type:string, comment:null), FieldSche
ma(name:avg_profit_margin, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117090606_0cb8be3e-e1aa-4a14-a206-3fb3053d1f1e); Time taken:
0.095 seconds
INFO  : Executing command(queryId=hive_20250117090606_0cb8be3e-e1aa-4a14-a206-3fb3053d1f1e): SELECT Product, AVG(Ma
rgin) AS Avg_Profit_Margin
FROM coffee_new
GROUP BY Product
ORDER BY Avg_Profit_Margin DESC
INFO  : Query ID = hive_20250117090606_0cb8be3e-e1aa-4a14-a206-3fb3053d1f1e
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0052
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0052/
INFO  : Starting Job = job_1737099421907_0052, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173

+--------------------+---------------------+--+
|      product       |  avg_profit_margin  |
+--------------------+---------------------+--+
| Regular Espresso   | 221.22222222222223  |
| Colombian          | 151.25              |
| Earl Grey          | 125.38888888888889  |
| Chamomile          | 111.91666666666667  |
| Lemon              | 109.31932773109244  |
| Darjeeling         | 102.72916666666667  |
| Decaf Espresso     | 101.25490196078431  |
| Caffe Mocha        | 95.15               |
| Caffe Latte        | 88.14814814814815   |
| Decaf Irish Cream  | 77.1875             |
| Mint               | 71.54166666666667   |
| Amaretto           | 69.625              |
| Green Tea          | 44.94444444444444   |
+--------------------+---------------------+--+
13 rows selected (43.151 seconds)
```

Figure 4.11 Query to display average profit margin for each product

4. Query to display products with the highest marketing spend.

```
0: jdbc:hive2://localhost:10000> SELECT Product, SUM(Marketing) AS Total_Marketing
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Product
. . . . . .> ORDER BY Total_Marketing DESC
. . . . . .> LIMIT 5;
INFO  : Compiling command(queryId=hive_20250117101919_1d15e36a-2304-4cff-8a6f-c5e4ff450c59): SELECT Product, SUM(Market
ing) AS Total_Marketing
FROM coffee_new
GROUP BY Product
ORDER BY Total_Marketing DESC
LIMIT 5
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:product, type:string, comment:null), FieldSchema(n
ame:total_marketing, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117101919_1d15e36a-2304-4cff-8a6f-c5e4ff450c59); Time taken: 0.12
8 seconds
INFO  : Executing command(queryId=hive_20250117101919_1d15e36a-2304-4cff-8a6f-c5e4ff450c59): SELECT Product, SUM(Market
ing) AS Total_Marketing
FROM coffee_new
GROUP BY Product
ORDER BY Total_Marketing DESC
LIMIT 5
INFO  : Query ID = hive_20250117101919_1d15e36a-2304-4cff-8a6f-c5e4ff450c59
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0064
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0064/
INFO  : Starting Job = job_1737099421907_0064, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0064/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0064
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:19:53,419 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:19:59,698 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.85 sec
INFO  : 2025-01-17 10:20:05,961 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.41 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 410 msec
INFO  : Ended Job = job_1737099421907_0064
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0065
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0065/
INFO  : Starting Job = job_1737099421907_0065, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0065/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0065
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:20:15,598 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:20:21,956 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.07 sec
INFO  : 2025-01-17 10:20:30,386 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.76 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 760 msec
INFO  : Ended Job = job_1737099421907_0065
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.41 sec   HDFS Read: 131369 HDFS Write: 501 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.76 sec   HDFS Read: 5599 HDFS Write: 81 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 6 seconds 170 msec
INFO  : Completed executing command(queryId=hive_20250117101919_1d15e36a-2304-4cff-8a6f-c5e4ff450c59); Time taken: 43.9
85 seconds
INFO  : OK
+--------------------+-------------------+--+
|      product       | total_marketing   |
+--------------------+-------------------+--+
| Caffe Mocha        | 4900              |
| Colombian          | 4166              |
| Lemon              | 3858              |
| Chamomile          | 3048              |
| Decaf Irish Cream  | 2696              |
+--------------------+-------------------+--+
5 rows selected (44.168 seconds)
```

Figure 4.12 Query to display products with the highest marketing spend

5. Query to determine which market has the highest sales

```
0: jdbc:hive2://localhost:10000> SELECT Market, SUM(Sales) AS Total_Sales
. . . . . . . . . . . . . . . .> FROM coffee_new
. . . . . . . . . . . . . . . .> GROUP BY Market
. . . . . . . . . . . . . . . .> ORDER BY Total_Sales DESC;
INFO  : Compiling command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4): SELECT Market, SUM(Sal
es) AS Total_Sales
FROM coffee_new
GROUP BY Market
ORDER BY Total_Sales DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchem
a(name:total_sales, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4); Time taken:
0.163 seconds
INFO  : Executing command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4): SELECT Market, SUM(Sal
es) AS Total_Sales
FROM coffee_new
GROUP BY Market
ORDER BY Total_Sales DESC
INFO  : Query ID = hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0016
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0016/
INFO  : Starting Job = job_1737099421907_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173
7099421907_0016/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job 1737099421907 0016
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 07:24:33,077 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 07:24:38,320 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.37 sec
INFO  : 2025-01-17 07:24:45,760 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.39 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 390 msec
INFO  : Ended Job = job_1737099421907_0017
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.31 sec   HDFS Read: 131360 HDFS Write: 201 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.39 sec   HDFS Read: 5156 HDFS Write: 48 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 6 seconds 700 msec
INFO  : Completed executing command(queryId=hive_20250117072424_92e1b69e-a63b-47dc-a1c8-d763dfd68bc4); Time taken:
42.692 seconds
INFO  : OK
+----------+--------------+--+
|  market  | total_sales  |
+----------+--------------+--+
| West     | 67418        |
| Central  | 64859        |
| East     | 44108        |
| South    | 26388        |
+----------+--------------+--+
4 rows selected (42.922 seconds)
```

Figure 4.13 Query to determine which market has the highest sales

6. Query to evaluate whether actual profit meets or exceeds the target profit for each market.

```
0: jdbc:hive2://localhost:10000> SELECT Market, SUM(Profit) AS Actual_Profit, SUM(Target_profit) AS Target_Profit
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Market;
INFO  : Compiling command(queryId=hive_20250117073333_cd350265-48d2-4c5e-8e01-98f842d48aec): SELECT Market, SUM(Pro
fit) AS Actual_Profit, SUM(Target_profit) AS Target_Profit
FROM coffee_new
GROUP BY Market
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchem
a(name:actual_profit, type:bigint, comment:null), FieldSchema(name:target_profit, type:bigint, comment:null)], prop
erties:null)
INFO  : Completed compiling command(queryId=hive_20250117073333_cd350265-48d2-4c5e-8e01-98f842d48aec); Time taken:
0.096 seconds
INFO  : Executing command(queryId=hive_20250117073333_cd350265-48d2-4c5e-8e01-98f842d48aec): SELECT Market, SUM(Pro
fit) AS Actual_Profit, SUM(Target_profit) AS Target_Profit
FROM coffee_new
GROUP BY Market
INFO  : Query ID = hive_20250117073333_cd350265-48d2-4c5e-8e01-98f842d48aec
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0020
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0020/
INFO  : Starting Job = job_1737099421907_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173
7099421907_0020/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0020
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 07:33:28,569 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 07:33:33,868 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.64 sec
INFO  : 2025-01-17 07:33:41,262 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.32 sec
INFO  : MapReduce Total cumulative CPU time: 4 seconds 320 msec
INFO  : Ended Job = job_1737099421907_0020
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.32 sec   HDFS Read: 132728 HDFS Write: 70 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 4 seconds 320 msec
INFO  : Completed executing command(queryId=hive_20250117073333_cd350265-48d2-4c5e-8e01-98f842d48aec); Time taken:
20.71 seconds
INFO  : OK
+----------+----------------+----------------+--+
|  market  | actual_profit  | target_profit  |
+----------+----------------+----------------+--+
| Central  | 22906          | 23530          |
| East     | 14745          | 13580          |
| South    | 8396           | 8640           |
| West     | 18269          | 18120          |
+----------+----------------+----------------+--+
4 rows selected (20.876 seconds)
```

Figure 4.14 Query to evaluate whether actual profit meets or exceeds the target profit for each market

7. Query to examine the total expenses incurred by market

```
0: jdbc:hive2://localhost:10000> SELECT Market, SUM(Total_expenses) AS Total_Expenses
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Market
. . . . . .> ORDER BY Total_Expenses DESC;
INFO  : Compiling command(queryId=hive_20250117081717_c458753a-faea-40d4-8fe3-26eb12f6899e): SELECT Market, SUM(Tot
al_expenses) AS Total_Expenses
FROM coffee_new
GROUP BY Market
ORDER BY Total_Expenses DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchem
a(name:total_expenses, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117081717_c458753a-faea-40d4-8fe3-26eb12f6899e); Time taken:
0.094 seconds
INFO  : Executing command(queryId=hive_20250117081717_c458753a-faea-40d4-8fe3-26eb12f6899e): SELECT Market, SUM(Tot
al_expenses) AS Total_Expenses
FROM coffee_new
GROUP BY Market
ORDER BY Total_Expenses DESC
INFO  : Query ID = hive_20250117081717_c458753a-faea-40d4-8fe3-26eb12f6899e
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0046
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0046/
INFO  : Starting Job = job_1737099421907_0046, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173
7099421907_0046/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0046

INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 08:17:35,146 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 08:17:41,474 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.54 sec
INFO  : 2025-01-17 08:17:48,871 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.49 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 490 msec
INFO  : Ended Job = job_1737099421907_0046
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0047
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0047/
INFO  : Starting Job = job_1737099421907_0047, Tracking URL = http://quickstart.cloudera:8088/proxy/application_173
7099421907_0047/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0047
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 08:17:57,357 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 08:18:02,658 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.29 sec
INFO  : 2025-01-17 08:18:10,069 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.12 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 120 msec
INFO  : Ended Job = job_1737099421907_0047
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.49 sec   HDFS Read: 131378 HDFS Write: 200 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.12 sec   HDFS Read: 5179 HDFS Write: 47 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 6 seconds 610 msec
INFO  : Completed executing command(queryId=hive_20250117081717_c458753a-faea-40d4-8fe3-26eb12f6899e); Time taken:
43.289 seconds
INFO  : OK
+----------+-----------------+--+
|  market  | total_expenses  |
+----------+-----------------+--+
| West     | 19784           |
| Central  | 17048           |
| East     | 12436           |
| South    | 7830            |
+----------+-----------------+--+
4 rows selected (43.451 seconds)
```

Figure 4.15 Query to examine the total expenses incurred by market

8. Query to calculate the total number of transactions for each market in the coffee new table

```
0: jdbc:hive2://localhost:10000> SELECT Market, COUNT(*) AS Total_Transactions
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Market
. . . . . .> ORDER BY Total_Transactions DESC;
INFO  : Compiling command(queryId=hive_20250117100202_08ddf053-ceda-481f-b0d7-5479e373a477): SELECT Market, COUNT(*) AS
 Total_Transactions
FROM coffee_new
GROUP BY Market
ORDER BY Total_Transactions DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchema(na
me:total_transactions, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117100202_08ddf053-ceda-481f-b0d7-5479e373a477); Time taken: 0.08
9 seconds
INFO  : Executing command(queryId=hive_20250117100202_08ddf053-ceda-481f-b0d7-5479e373a477): SELECT Market, COUNT(*) AS
 Total_Transactions
FROM coffee_new
GROUP BY Market
ORDER BY Total_Transactions DESC
INFO  : Query ID = hive_20250117100202_08ddf053-ceda-481f-b0d7-5479e373a477
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0058
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0058/
INFO  : Starting Job = job_1737099421907_0058, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0058/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0058
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:02:54,691 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:02:59,968 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.0 sec
INFO  : 2025-01-17 10:03:06,229 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.52 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 520 msec
INFO  : Ended Job = job_1737099421907_0058
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0059
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0059/
INFO  : Starting Job = job_1737099421907_0059, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0059/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0059
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:03:13,209 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:03:19,481 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.25 sec
INFO  : 2025-01-17 10:03:26,819 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.78 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 780 msec
INFO  : Ended Job = job_1737099421907_0059
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.52 sec   HDFS Read: 131302 HDFS Write: 198 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.78 sec   HDFS Read: 5129 HDFS Write: 40 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 6 seconds 300 msec
INFO  : Completed executing command(queryId=hive_20250117100202_08ddf053-ceda-481f-b0d7-5479e373a477); Time taken: 39.3
76 seconds
INFO  : OK
+----------+---------------------+--+
|  market  | total_transactions  |
+----------+---------------------+--+
| West     | 336                 |
| Central  | 335                 |
| East     | 222                 |
| South    | 168                 |
+----------+---------------------+--+
4 rows selected (39.538 seconds)
```

Figure 4.16 Query to calculate the total number of transactions for each market in the coffee new table

9. Query to distinct products sold in each market

```
0: jdbc:hive2://localhost:10000> SELECT Market, COUNT(DISTINCT Product) AS Distinct_Products
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY Market
. . . . . .> ORDER BY Distinct_Products DESC;
INFO  : Compiling command(queryId=hive_20250117100707_47ec50be-d744-4b89-97d0-d87abc8349ce): SELECT Market, COUNT(DISTI
NCT Product) AS Distinct_Products
FROM coffee_new
GROUP BY Market
ORDER BY Distinct_Products DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchema(na
me:distinct_products, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117100707_47ec50be-d744-4b89-97d0-d87abc8349ce); Time taken: 0.08
5 seconds
INFO  : Executing command(queryId=hive_20250117100707_47ec50be-d744-4b89-97d0-d87abc8349ce): SELECT Market, COUNT(DISTI
NCT Product) AS Distinct_Products
FROM coffee_new
GROUP BY Market
ORDER BY Distinct_Products DESC
INFO  : Query ID = hive_20250117100707_47ec50be-d744-4b89-97d0-d87abc8349ce
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0060
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0060/
INFO  : Starting Job = job_1737099421907_0060, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0060/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0060
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:07:59,225 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:08:04,517 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.21 sec
INFO  : 2025-01-17 10:08:10,790 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.8 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 800 msec
INFO  : Ended Job = job_1737099421907_0060
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0061
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0061/
INFO  : Starting Job = job_1737099421907_0061, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0061/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0061
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:08:17,848 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:08:23,096 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.03 sec
INFO  : 2025-01-17 10:08:29,361 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.69 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 690 msec
INFO  : Ended Job = job_1737099421907_0061
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.8 sec   HDFS Read: 131657 HDFS Write: 192 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.69 sec   HDFS Read: 5171 HDFS Write: 35 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 5 seconds 490 msec
INFO  : Completed executing command(queryId=hive_20250117100707_47ec50be-d744-4b89-97d0-d87abc8349ce); Time taken: 37.0
5 seconds
INFO  : OK
+----------+--------------------+--+
|  market  | distinct_products  |
+----------+--------------------+--+
| West     | 12                 |
| East     | 12                 |
| Central  | 11                 |
| South    | 7                  |
+----------+--------------------+--+
4 rows selected (37.212 seconds)
```

Figure 4.17 Query to distinct products sold in each market

10. Query to identify the least profitable products in each market

```
0: jdbc:hive2://localhost:10000> SELECT Market, Product, MIN(Profit) AS Min_Profit
. . . . . . . . . . . . . . . .> FROM coffee_new
. . . . . . . . . . . . . . . .> GROUP BY Market, Product
. . . . . . . . . . . . . . . .> ORDER BY Market, Min_Profit ASC;
INFO  : Compiling command(queryId=hive_20250117103939_02c6493f-3914-4e85-b2a4-90fe8cd00dc7): SELECT Market, Product, MI
N(Profit) AS Min_Profit
FROM coffee_new
GROUP BY Market, Product
ORDER BY Market, Min_Profit ASC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:market, type:string, comment:null), FieldSchema(na
me:product, type:string, comment:null), FieldSchema(name:min_profit, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250117103939_02c6493f-3914-4e85-b2a4-90fe8cd00dc7); Time taken: 0.05
8 seconds
INFO  : Executing command(queryId=hive_20250117103939_02c6493f-3914-4e85-b2a4-90fe8cd00dc7): SELECT Market, Product, MI
N(Profit) AS Min_Profit
FROM coffee_new
GROUP BY Market, Product
ORDER BY Market, Min_Profit ASC
INFO  : Query ID = hive_20250117103939_02c6493f-3914-4e85-b2a4-90fe8cd00dc7
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0069
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0069/
INFO  : Starting Job = job_1737099421907_0069, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0069/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0069
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:39:23,198 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:39:28,404 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.94 sec
INFO  : 2025-01-17 10:39:32,545 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.07 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 70 msec
INFO  : Ended Job = job_1737099421907_0069
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737099421907_0070
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737099421907_0070/
INFO  : Starting Job = job_1737099421907_0070, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1737099
421907_0070/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737099421907_0070
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-17 10:39:40,448 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-17 10:39:45,682 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.76 sec
INFO  : 2025-01-17 10:39:50,843 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.08 sec
INFO  : MapReduce Total cumulative CPU time: 2 seconds 80 msec
INFO  : Ended Job = job_1737099421907_0070
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.07 sec   HDFS Read: 131685 HDFS Write: 1570 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.08 sec   HDFS Read: 6921 HDFS Write: 845 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 4 seconds 150 msec
INFO  : Completed executing command(queryId=hive_20250117103939_02c6493f-3914-4e85-b2a4-90fe8cd00dc7); Time taken: 33.2
75 seconds
INFO  : OK
```

Figure 4.18 Query to identify the least profitable products in each market

market	product	min_profit
Central	Lemon	-39
Central	Earl Grey	-24
Central	Chamomile	-12
Central	Decaf Irish Cream	-9
Central	Green Tea	-6
Central	Colombian	5
Central	Darjeeling	6
Central	Decaf Espresso	6
Central	Amaretto	8
Central	Caffe Mocha	16
Central	Mint	26
East	Caffe Mocha	-332
East	Mint	-280
East	Regular Espresso	-24
East	Lemon	-6
East	Darjeeling	8
East	Chamomile	10
East	Green Tea	15
East	Colombian	20
East	Decaf Espresso	21
East	Amaretto	30
East	Decaf Irish Cream	76
East	Earl Grey	127
South	Decaf Irish Cream	-39
South	Caffe Latte	-22
South	Caffe Mocha	-10
South	Lemon	5
South	Decaf Espresso	8
South	Chamomile	9
South	Colombian	20
West	Green Tea	-605
West	Decaf Irish Cream	-221
West	Amaretto	-131
West	Mint	-40
West	Darjeeling	-10
West	Colombian	-6
West	Caffe Latte	6
West	Caffe Mocha	9
West	Earl Grey	9
West	Chamomile	17
West	Decaf Espresso	17
West	Lemon	20

42 rows selected (33.379 seconds)

Figure 4.18 Query to identify the least profitable products in each market (continued)

11. Query to create table coffee_sales_summary

```
0: jdbc:hive2://localhost:10000> CREATE TABLE coffee_sales_summary
. . . . . .> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . .> LINES TERMINATED BY '\n'
. . . . . .> STORED AS textfile
. . . . . .> AS
. . . . . .> SELECT
. . . . . .>     State AS State,
. . . . . .>     Market AS Market,
. . . . . .>     Product AS Product,
. . . . . .>     COUNT(*) AS Total_Transactions,
. . . . . .>     SUM(Sales) AS Total_Sales,
. . . . . .>     AVG(Sales) AS Average_Sales,
. . . . . .>     MIN(Sales) AS Min_Sale,
. . . . . .>     MAX(Sales) AS Max_Sale,
. . . . . .>     SUM(Profit) AS Actual_Profit,
. . . . . .>     SUM(Target_profit) AS Target_Profit,
. . . . . .>     SUM(Profit) - SUM(Target_profit) AS Profit_Difference,
. . . . . .>     AVG(Profit / Sales) AS Avg_Profit_Margin,
. . . . . .>     SUM(Sales) - SUM(Target_sales) AS Sales_Difference,
. . . . . .>     SUM(Marketing) AS Total_Marketing,
. . . . . .>     MIN(Profit) AS Min_Profit,
. . . . . .>     SUM(Total_expenses) AS Total_Expenses,
. . . . . .>     MAX(Date) AS Last_Date_Sale,
. . . . . .>     COUNT(DISTINCT Product) AS Distinct_Products
. . . . . .> FROM coffee_new
. . . . . .> GROUP BY State, Market, Product
. . . . . .> ORDER BY State, Market, Total_Sales DESC;
INFO  : Compiling command(queryId=hive_20250120221515_164e0fd0-1068-452e-a47a-0452a70cd679): CREATE TABLE coffee_sale
s_summary
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS textfile
AS
SELECT
State AS State,
Market AS Market,
Product AS Product,
COUNT(*) AS Total_Transactions,
SUM(Sales) AS Total_Sales,
AVG(Sales) AS Average_Sales,
MIN(Sales) AS Min_Sale,
```

Figure 4.19 Query to create table coffee_sales_summary

```
SUM(Total_expenses) AS Total_Expenses,
MAX(Date) AS Last_Date_Sale,
COUNT(DISTINCT Product) AS Distinct_Products
FROM coffee_new
GROUP BY State, Market, Product
ORDER BY State, Market, Total_Sales DESC
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:state, type:string, comment:null), FieldSchema(n
ame:market, type:string, comment:null), FieldSchema(name:product, type:string, comment:null), FieldSchema(name:total_
transactions, type:bigint, comment:null), FieldSchema(name:total_sales, type:bigint, comment:null), FieldSchema(name:
average_sales, type:double, comment:null), FieldSchema(name:min_sale, type:int, comment:null), FieldSchema(name:max_s
ale, type:int, comment:null), FieldSchema(name:actual_profit, type:bigint, comment:null), FieldSchema(name:target_pro
fit, type:bigint, comment:null), FieldSchema(name:profit_difference, type:bigint, comment:null), FieldSchema(name:avg
_profit_margin, type:double, comment:null), FieldSchema(name:sales_difference, type:bigint, comment:null), FieldSchem
a(name:total_marketing, type:bigint, comment:null), FieldSchema(name:min_profit, type:int, comment:null), FieldSchema
(name:total_expenses, type:bigint, comment:null), FieldSchema(name:last_date_sale, type:string, comment:null), FieldS
chema(name:distinct_products, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250120221515_164e0fd0-1068-452e-a47a-0452a70cd679); Time taken: 0.
141 seconds
INFO  : Executing command(queryId=hive_20250120221515_164e0fd0-1068-452e-a47a-0452a70cd679): CREATE TABLE coffee_sale
s_summary
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS textfile
AS
SELECT
State AS State,
Market AS Market,
Product AS Product,
COUNT(*) AS Total_Transactions,
SUM(Sales) AS Total_Sales,
AVG(Sales) AS Average_Sales,
MIN(Sales) AS Min_Sale,
MAX(Sales) AS Max_Sale,
SUM(Profit) AS Actual_Profit,
SUM(Target_profit) AS Target_Profit,
SUM(Profit) - SUM(Target_profit) AS Profit_Difference,
AVG(Profit / Sales) AS Avg_Profit_Margin,
SUM(Sales) - SUM(Target_sales) AS Sales_Difference,
SUM(Marketing) AS Total_Marketing,
MIN(Profit) AS Min_Profit,
SUM(Total_expenses) AS Total_Expenses,
MAX(Date) AS Last_Date_Sale,
MAX(Date) AS Last_Date_Sale,
COUNT(DISTINCT Product) AS Distinct_Products
FROM coffee_new
GROUP BY State, Market, Product
ORDER BY State, Market, Total_Sales DESC
INFO  : Query ID = hive_20250120221515_164e0fd0-1068-452e-a47a-0452a70cd679
INFO  : Total jobs = 2
INFO  : Launching Job 1 out of 2
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737433801366_0005
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737433801366_0005/
INFO  : Starting Job = job_1737433801366_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_17374
33801366_0005/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737433801366_0005
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-20 22:15:45,888 Stage-1 map = 0%,  reduce = 0%
INFO  : 2025-01-20 22:15:53,250 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.95 sec
INFO  : 2025-01-20 22:16:00,670 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.63 sec
INFO  : MapReduce Total cumulative CPU time: 5 seconds 630 msec
INFO  : Ended Job = job_1737433801366_0005
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737433801366_0006
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737433801366_0006/
INFO  : Starting Job = job_1737433801366_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_17374
33801366_0006/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737433801366_0006
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-20 22:16:07,748 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-20 22:16:14,103 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
INFO  : 2025-01-20 22:16:20,433 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.42 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 420 msec
INFO  : Ended Job = job_1737433801366_0006
```

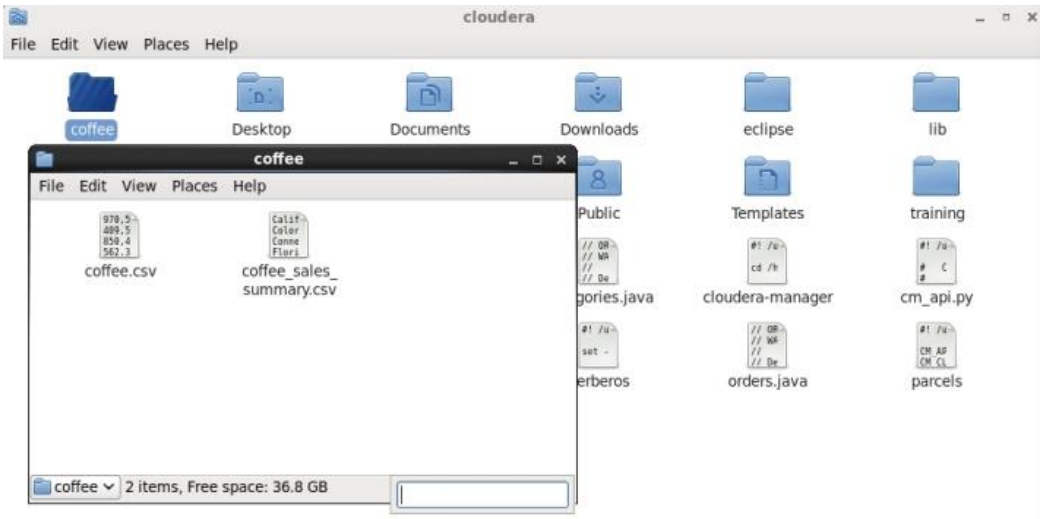
Figure 4.19 Query to create table coffee_sales_summary (continued)

```
INFO  : Launching Job 2 out of 2
INFO  : Starting task [Stage-2:MAPRED] in serial mode
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1737433801366_0006
INFO  : The url to track the job: http://quickstart.cloudera:8088/proxy/application_1737433801366_0006/
INFO  : Starting Job = job_1737433801366_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_17374
33801366_0006/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1737433801366_0006
INFO  : Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
INFO  : 2025-01-20 22:16:07,748 Stage-2 map = 0%,  reduce = 0%
INFO  : 2025-01-20 22:16:14,103 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
INFO  : 2025-01-20 22:16:20,433 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.42 sec
INFO  : MapReduce Total cumulative CPU time: 3 seconds 420 msec
INFO  : Ended Job = job_1737433801366_0006
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/coffee_sales.db/coffee_sales_summary from
 hdfs://quickstart.cloudera:8020/user/hive/warehouse/coffee_sales.db/.hive-staging_hive_2025-01-20_22-15-38_645_65189
72563455129723-3/-ext-10001
INFO  : Starting task [Stage-4:DDL] in serial mode
INFO  : Starting task [Stage-3:STATS] in serial mode
INFO  : Table coffee_sales.coffee_sales_summary stats: [numFiles=1, numRows=177, totalSize=19589, rawDataSize=19412]
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.63 sec   HDFS Read: 141043 HDFS Write: 16647 SUCCESS
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.42 sec   HDFS Read: 26879 HDFS Write: 19682 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 9 seconds 50 msec
INFO  : Completed executing command(queryId=hive_20250120221515_164e0fd0-1068-452e-a47a-0452a70cd679); Time taken: 43
.025 seconds
INFO  : OK
No rows affected (43.207 seconds)
0: jdbc:hive2://localhost:10000> DESCRIBE coffee_sales_summary;
INFO  : Compiling command(queryId=hive_20250120222121_ffb5a531-fa79-42d7-b937-63bf0a6904bc): DESCRIBE coffee_sales_su
mmary
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer
), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string, commen
t:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20250120222121_ffb5a531-fa79-42d7-b937-63bf0a6904bc); Time taken: 0.
089 seconds
INFO  : Executing command(queryId=hive_20250120222121_ffb5a531-fa79-42d7-b937-63bf0a6904bc): DESCRIBE coffee_sales_su
mmary
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250120222121_ffb5a531-fa79-42d7-b937-63bf0a6904bc); Time taken: 0.
021 seconds
INFO  : OK
+----------------------+------------+----------+--+
|      col_name        | data_type  | comment  |
+----------------------+------------+----------+--+
| state                | string     |          |
| market               | string     |          |
| product              | string     |          |
| total_transactions   | bigint     |          |
| total_sales          | bigint     |          |
| average_sales        | double     |          |
| min_sale             | int        |          |
| max_sale             | int        |          |
| actual_profit        | bigint     |          |
| target_profit        | bigint     |          |
| profit_difference    | bigint     |          |
| avg_profit_margin    | double     |          |
| sales_difference     | bigint     |          |
| total_marketing      | bigint     |          |
| min_profit           | int        |          |
| total_expenses       | bigint     |          |
| last_date_sale       | string     |          |
| distinct_products    | bigint     |          |
+----------------------+------------+----------+--+
18 rows selected (0.167 seconds)
```

Figure 4.19 Query to create table coffee_sales_summary (continued)

Export csv file

```
[cloudera@quickstart ~]$ hdfs dfs -cat /user/hive/warehouse/coffee_sales.db/coff
ee_sales_summary/* > ~/coffee_sales_summary.csv
```



4.20 Export csv file

5.0 Data Visualization using Power BI

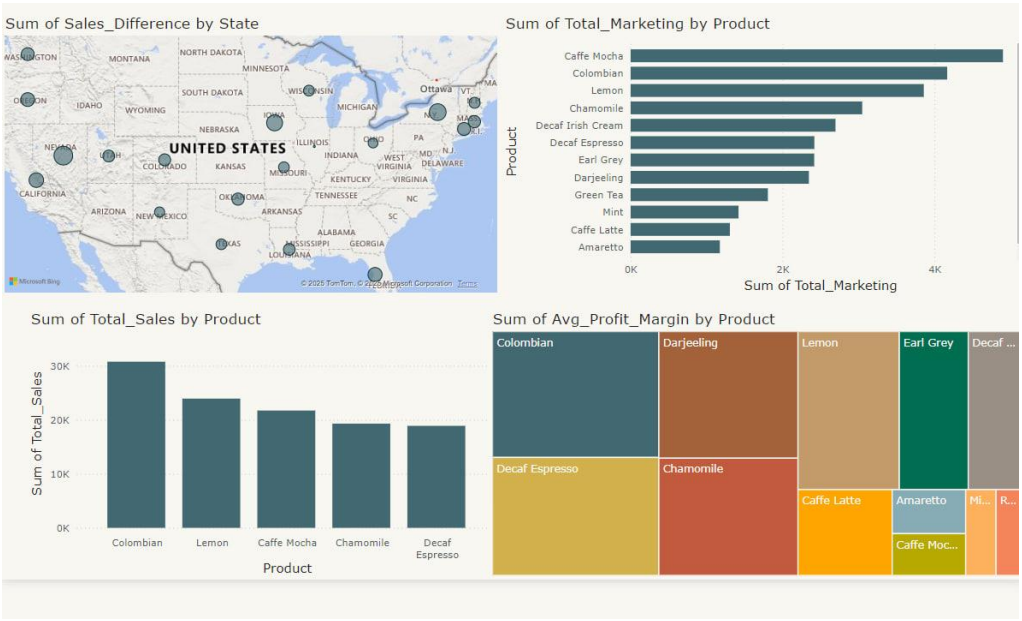


Figure 5.1 Power BI Visualization Graphs

The Power BI visualization effectively provides insights into coffee sales and marketing data across various dimensions. The map highlights the Sales Difference by State, showcasing geographical disparities in sales performance. The bar chart titled Sum of Total_Marketing by Product reveals that products like Caffe Mocha and Colombian coffee received the highest marketing spend. The Sum of Total_Sales by Product bar chart indicates that Colombian coffee leads in sales, followed by Lemon and Caffe Mocha. Lastly, the treemap visual, Sum of Avg_Profit_Margin by Product, identifies Colombian coffee and Decaf Espresso as the most profitable products, while products like Caffe Latte

and Chamomile exhibit moderate profit margins. This visualization provides actionable insights to optimize marketing efforts and product profitability.

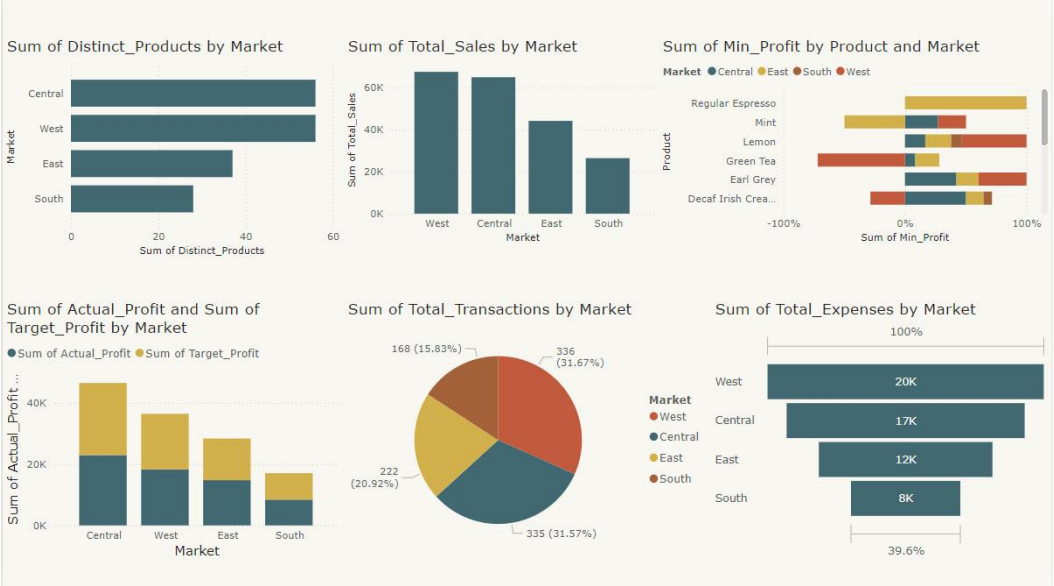


Figure 5.2 Power BI Visualization Graphs

This Power BI visualization provides a detailed analysis of various metrics across different markets. The Sum of Distinct_Products by Market bar chart shows that the Central market offers the highest variety of products, followed by the West market. The Sum of Total_Sales by Market highlights the West market as the leader in sales, with the Central market closely trailing, while the East and South markets lag behind.The Sum of Min_Profit by Product and Market chart reveals the lowest profitability by product within each market, indicating areas where performance may need improvement. The Sum of Actual_Profit and Sum of Target_Profit by Market comparison demonstrates that while the Central and West markets meet or exceed their profit targets, the East and South markets fall short.The Sum of Total_Transactions by Market pie chart emphasizes that the West and Central markets account for the majority of transactions, with the East and South markets contributing less. Lastly, the Sum of Total_Expenses by Market bar chart indicates that the West market incurs the highest expenses, followed by the Central market, with the South market having the lowest expenditure.These insights can help in identifying strong-performing markets, addressing underperforming ones, and optimizing operations across all regions.

6.0 Conclusion

In conclusion, this project successfully demonstrates the use of data analytics to enhance decision-making in the coffee sales industry. By integrating big data tools like Apache Hive and Hadoop, along with a well-designed database structure, we were able to analyze key metrics such as sales trends, market performance, and product profitability. The Entity-Relationship Diagram (ERD) effectively represents the relationships between markets, products, and sales, providing a solid foundation for querying and analyzing data. The insights derived from this project, such as identifying top-performing markets and products, optimizing profit margins, and evaluating marketing impact,

highlight the value of data-driven strategies in improving operational efficiency and maximizing profitability. This project underscores the importance of leveraging modern data technologies to address real-world business challenges effectively.

7.0 Acknowledgement

We would like to express our sincere gratitude to our lecturer, Dr Khairul Anuar B. Sedek for their guidance, support, and valuable insights throughout this course. His dedication and encouragement have been instrumental in helping us overcome challenges and complete this project successfully. The knowledge and skills we have gained under your mentorship are truly appreciated. We also wish to thank our group members for their commitment, collaboration, and hard work. Each member's contributions, from brainstorming ideas to implementing solutions, played a vital role in ensuring the project's success. The teamwork and shared effort made this journey both productive and enjoyable.

8.0 References

Azeroual, O. (2020). Data Wrangling in Database Systems: Purging of Dirty Data. *Data*, *5*, 50.

B. Małysiak-Mrozek, J. Wieszok, W. Pedrycz, W. Ding and D. Mrozek, "High-Efficient Fuzzy Querying With HiveQL for Big Data Warehousing," in IEEE Transactions on Fuzzy Systems, vol. 30, no. 6, pp. 1823-1837, June 2022, doi: 10.1109/TFUZZ.2021.3069332.

Costa, E., Costa, C.A., & Santos, M.Y. (2019). Evaluating partitioning and bucketing strategies for Hive-based Big Data Warehousing systems. *Journal of Big Data, 6*.

Chugh, A. (2021). Peer Review #2 of "Efficient processing of complex XSD using Hive and Spark (v0.2)". *PeerJ*.

Mahalle, P.N., Shinde, G.R., Pise, P.D., & Deshmukh, J.Y. (2021). Data Collection and Preparation. *Studies in Big Data*.

Taylor, K.S., Mahtani, K.R., & Aronson, J.K. (2021). Summarising good practice guidelines for data extraction for systematic reviews and meta-analysis. *BMJ Evidence-Based Medicine, 26*, 88 - 90.

Fernandes, A.A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N.W., & Sakellariou, R. (2023). Data Preparation: A Technological Perspective and Review. *SN Computer Science*, *4*.

Małysiak-Mrozek, B., Wieszok, J., Pedrycz, W., Ding, W., & Mrozek, D. (2022). High-Efficient Fuzzy Querying With HiveQL for Big Data Warehousing. *IEEE Transactions on Fuzzy Systems, 30*, 1823-1837.