Big Data - Case Study

Subject - Big Data Analytics and Architecture

PROJECT

Coffee Sales Data

Coffee Sales Data Analysis Using Apache Hive

Project Overview

This project focuses on performing data analysis and generating business insights from a Coffee Sales dataset using Apache Hive.

The objective is to utilize Hive's SQL-like querying capabilities to analyze key business metrics such as:

- Customer country-wise sales distribution
- Most selling coffee products
- Daily and total revenue
- Price & demand trends

This project demonstrates how to store, manage, and analyze structured sales data on a Big Data platform (Hadoop/Cloudera) using HiveQL for efficient analytical processing.

Dataset Description

The dataset coffe.txt.csv contains detailed daily coffee sales transactions, including:

Column Name	Description
date	Transaction date
country	Customer country
coffee_type	Type of coffee

Column Name	Description
quantity	Units sold
unit_price	Price per unit
total_price	Total sale amount

Objectives

Key goals of this project include:

- To import and store CSV-based sales data into Hive tables
- To perform analytical queries for business-driven insights
- To identify sales patterns such as:
    - Top selling coffee types
    - Country-wise sales & popularity
    - Daily revenue trends
    - Quantity sold vs revenue analysis
    - Highest revenue generating days

Technologies Used

- Apache Hive
- Hadoop HDFS (Cloudera)
- HiveQL

- CSV File Data Loading
- Big Data Storage & Query Engine

Use Database

```
hive> CREATE EXTERNAL TABLE coffee_sales (
    >     date STRING,
    >     country STRING,
    >     coffee_type STRING,
    >     quantity INT,
    >     unit_price FLOAT,
    >     total_price FLOAT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.577 seconds
hive>
```

Load Data:

```
hive> load data local inpath '/home/cloudera/Desktop/coffe.csv'into table coffee_sales;
Loading data to table sakshi.coffee_sales
Table sakshi.coffee_sales stats: [numFiles=1, totalSize=232264]
OK
Time taken: 0.776 seconds
hive>
```

Show first 10 records

Q.1 Dataset ke first 10 rows dikhaye

SELECT * FROM coffee_sales LIMIT 10;

```
Time taken: 0.071 seconds
hive> SELECT * FROM coffee_sales LIMIT 10;
OK
hour_of_day     cash_type       money   NULL    NULL    NULL
10      card    38.7    NULL    NULL    NULL
12      card    38.7    NULL    NULL    NULL
12      card    38.7    NULL    NULL    NULL
13      card    28.9    NULL    NULL    NULL
13      card    38.7    NULL    NULL    NULL
15      card    33.8    NULL    NULL    NULL
16      card    38.7    NULL    NULL    NULL
18      card    33.8    NULL    NULL    NULL
19      card    38.7    NULL    NULL    NULL
Time taken: 0.761 seconds, Fetched: 10 row(s)
```

Total Sales (Revenue) Calculate karo

Q.2 Total revenue kitna generate hua?

SELECT SUM(total_price) AS total_revenue

FROM coffee_sales;

Output: Total coffee business revenue

```
hive> SELECT SUM(total_price) AS total_revenue
    > FROM coffee_sales;
Query ID = cloudera 20251030005959 8e02832e-ca63-4e85-a003-299b6c69a510
```

Out Put:

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.99 sec   HDFS
Total MapReduce CPU Time Spent: 2 seconds 990 msec
OK
NULL
Time taken: 36.125 seconds, Fetched: 1 row(s)
hive>
```

Total Quantity Sold

Q.3 Kitne cups/units beche gaye?

SELECT SUM(quantity) AS total_quantity

FROM coffee_sales;

```
hive> SELECT SUM(quantity) AS total_quantity
    > FROM coffee_sales;
Query ID = cloudera_20251030010505_87cb7c3a-01ec-4909-8f3f-fcf16a07eac2
```

Out Put:

```
Total MapReduce CPU Time Spent: 2 seconds 670 msec
OK
NULL
Time taken: 27.592 seconds, Fetched: 1 row(s)
hive>
```

Most Selling Coffee Type

Q.4 Konsa coffee type sabse zyada bikta hai?

SELECT coffee_type, SUM(quantity) AS total_sold

FROM coffee_sales

GROUP BY coffee_type

ORDER BY total_sold DESC

LIMIT 1;

```
hive> SELECT coffee_type, SUM(quantity) AS total_qty
    > FROM coffee_sales
    > GROUP BY coffee_type
    > ORDER BY total_qty DESC
    > LIMIT 1;
Query ID = cloudera_20251030010707_228a8572-a794-428b-8475-3c7d4c0ecefe
```

Out Put:

```
Total MapReduce CPU Time Spent: 6 seconds 40 msec
OK
money   NULL
Time taken: 54.989 seconds, Fetched: 1 row(s)
hive>
```

Country-wise Revenue

Q.5 Har country ka total revenue kitna hai?

SELECT country, SUM(total_price) AS country_revenue

FROM coffee_sales

GROUP BY country

ORDER BY country_revenue DESC;

```
Time taken: 54.989 seconds, Fetched: 1 row(s)
hive> SELECT country, SUM(total_price) AS sales
    > FROM coffee_sales
    > GROUP BY country
    > ORDER BY sales DESC;
Query ID = cloudera 20251030011010 996c6eff-3849-4f19-9915-5e0d3ed06654
```

Out Put:

```
Total MapReduce CPU Time Spent: 5 seconds 420 msec
OK
cash_type       NULL
card    NULL
Time taken: 51.993 seconds, Fetched: 2 row(s)
hive>
```

Daily Sales Report

Q.6 Har date par total sales kitna hua?

SELECT date, SUM(total_price) AS daily_sales

FROM coffee_sales

GROUP BY date

ORDER BY date;

```
card    NULL
Time taken: 51.993 seconds, Fetched: 2 row(s)
hive> SELECT date, SUM(total_price) AS daily_sales
    > FROM coffee_sales
    > GROUP BY date
    > ORDER BY date;
Query ID = cloudera_20251030011313_7e5f98ef-fd2c-4e69-88f2-5ce31f98bb82
```

Out Put:

```
Stage-Stage-2: Map
Total MapReduce CP
OK
10      NULL
11      NULL
12      NULL
13      NULL
14      NULL
```

Q.7 Average Unit Price for Each Coffee

SELECT coffee_type, AVG(unit_price) AS avg_price

FROM coffee_sales

GROUP BY coffee_type;

```
hive> SELECT coffee_type, AVG(unit_price) AS avg_price
    > FROM coffee_sales
    > GROUP BY coffee_type;
Query ID = cloudera_20251030011717_e811a71c-736a-450f-a162-767be5453599
```

Out put:

```
Total MapReduce CPU Time Spent: 2 seconds 870 msec
OK
18.12   NULL
21.06   NULL
23.02   NULL
24      NULL
25.96   NULL
27.92   NULL
28.9    NULL
30.86   NULL
32.82   NULL
33.8    NULL
35.76   NULL
37.72   NULL
38.7    NULL
money   NULL
Time taken: 28.169 seconds, Fetched: 14 row(s)
hive>
```

Q.8 Count Total Orders

SELECT COUNT(*) AS total_orders

FROM coffee_sales;

```
hive> SELECT COUNT(*) AS total_orders
    > FROM coffee_sales;
Query ID = cloudera_20251030011919_7300b337-967b-491d-a30a-27382878ff44
```

Out put:

```
Total MapReduce CPU Time Spent: 2 seconds 860 msec
OK
3548
Time taken: 26.114 seconds, Fetched: 1 row(s)
hive>
```

Q.9 Top 3 Countries by Quantity Sold

SELECT country, SUM(quantity) AS total_qty

FROM coffee_sales

GROUP BY country

ORDER BY total_qty DESC

LIMIT 3;

```
hive> SELECT country, SUM(quantity) AS total_qty
    > FROM coffee_sales
    > GROUP BY country
    > ORDER BY total_qty DESC
    > LIMIT 3;
Query ID = cloudera_20251030012020_bd8303b2-dd46-49a1-b668-f24b26a24034
```

Out Put:

```
Total MapReduce CPU Time Spent: 4 seconds 820 msec
OK
cash_type       NULL
card    NULL
Time taken: 54.76 seconds, Fetched: 2 row(s)
hive>
```

Q.10 Bonus (Partition Example Query)

SELECT country, date, SUM(total_price) AS revenue

FROM coffee_sales

GROUP BY country, date

ORDER BY revenue DESC;

```
hive> SELECT country, date, SUM(total_price) AS revenue
    > FROM coffee_sales
    > GROUP BY country, date
    > ORDER BY revenue DESC;
Query ID = cloudera_20251030012222_8033737a-68f1-4800-acb3-8efd684bfbad
```

Out Put:

```
Total MapReduce CPU Time Spent: 4 seconds 800 msec
OK
cash_type       hour_of_day     NULL
card    9       NULL
card    8       NULL
card    7       NULL
card    6       NULL
card    22      NULL
card    21      NULL
card    20      NULL
card    19      NULL
card    18      NULL
card    17      NULL
card    16      NULL
card    15      NULL
card    14      NULL
card    13      NULL
card    12      NULL
card    11      NULL
card    10      NULL
Time taken: 51.14 seconds, Fetched: 18 row(s)
hive>
```