Big Data - Case Study

Subject - Big Data Analytics and Architecture

PROJECT

Foodpanda Data Set Overview

Foodpanda Data Set Overview

Project Overview

This project focuses on analyzing and extracting insights from a Foodpanda customer and order dataset using Apache Hive. The main goal is to utilize Hive's SQL-like querying capabilities to explore customer behavior, restaurant performance, ordering patterns, and payment trends. It demonstrates how structured transactional data can be effectively managed in a Big Data ecosystem (Hadoop/Cloudera) and how HiveQL can be used to perform large-scale analytical queries for actionable business insights.

Dataset Description

The dataset, titled Foodpanda Analysis Dataset.csv, provides detailed information on customer demographics, order details, and restaurant data. It includes the following fields:

- Customer ID
- Gender
- Age
- City
- Signup Date
- Order Date
- Restaurant Name
- Dish Name
- Category
- Quantity
- Price
- Payment Type
- Order Frequency
- Last Order
- Loyalty Policy

- Churned
- Rating
- Rating Date
- Delivery Status

Objectives

The project's main objectives are as follows:

- To import and store the Foodpanda dataset from CSV format into Hive tables.
- To perform analytical queries that uncover trends in customer behavior, restaurant performance, and payment modes.
- To extract valuable insights such as:
    - Top restaurants by order volume and revenue.
    - Average order value by city and category.
    - Customer loyalty and churn rate analysis.
    - Most popular dishes and cuisines.
    - Payment method preferences and delivery success rate.

Technologies Used

- Apache Hive
- Hadoop (Cloudera Environment)
- HiveQL (SQL-like Query Language)
- CSV File Data Loading
- HDFS (Hadoop Distributed File System)

Steps Involved

1. Created a new Hive database and defined a table schema for the Foodpanda dataset.
2. Loaded the CSV file from local storage or HDFS into the Hive table.

3. Executed various analytical queries to derive insights, including:
    - SELECT COUNT(*) to find total orders.
    - GROUP BY to analyze city-wise or restaurant-wise performance.
    - AVG() and SUM() to calculate average revenue and order values.
    - ORDER BY and LIMIT to identify top-performing restaurants and popular dishes.
    - WHERE and CASE statements to analyze churned vs. loyal customers.
4. Generated analytical reports summarizing key findings and visual insights.

Key Insights

- Identified the top-performing restaurants with the highest number of orders.
- Determined the most popular dishes and food categories among customers.
- Analyzed customer loyalty trends and churn patterns.
- Revealed revenue patterns based on city and age demographics.
- Highlighted preferred payment methods and delivery success rates.

Conclusion

This project demonstrates how Apache Hive can be used effectively to manage and analyze large-scale Foodpanda transactional data. By leveraging HiveQL within the Hadoop ecosystem, the project extracts meaningful insights into customer preferences, restaurant performance, and order behavior. These findings can support data-driven decision-making for marketing, operations, and customer engagement strategies.

Use Database:

```
hive> CREATE TABLE foodpanda_data (
    >     Customer_ID STRING,
    >     Gender STRING,
    >     Age INT,
    >     City STRING,
    >     Signup_Date STRING,
    >     Order_Date STRING,
    >     Restaurant_Name STRING,
    >     Dish_Name STRING,
    >     Category STRING,
    >     Quantity INT,
    >     Price FLOAT,
    >     Payment_Type STRING,
    >     Order_Frequency INT,
    >     Last_Order STRING,
    >     Loyalty_Policy STRING,
    >     Churned STRING,
    >     Rating FLOAT,
    >     Rating_Date STRING,
    >     Delivery_Status STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.354 seconds
```

Load Data:

```
hive> load data local inpath '/home/cloudera/Desktop/Foodpanda Analysis Dataset.csv' into table foodpanda_data;
Loading data to table upendra.foodpanda_data
Table upendra.foodpanda_data stats: [numFiles=1, totalSize=545471]
OK
Time taken: 0.664 seconds
```

Q.1 Which gender spends more on average per order?

SELECT Gender, ROUND(AVG(Price),2) AS Avg_Spending

FROM foodpanda_data

GROUP BY Gender;

```
hive> SELECT Gender, ROUND(AVG(Price),2) AS Avg_Spending
    > FROM foodpanda_data
    > GROUP BY Gender;
Query ID = cloudera_20251030220808_89553b80-3d56-45f7-94d0-4ab7241b434c
```

Out Put:

```
Total MapReduce CPU Time Spent: 2 seconds 440 msec
OK
         NULL
Female   3.01
Male     2.99
Other    2.96
gender   NULL
Time taken: 25.701 seconds, Fetched: 5 row(s)
hive>
```

Q.2 Which dish is ordered the most times overall?

SELECT Dish_Name, COUNT(*) AS Total_Orders

FROM foodpanda_data

GROUP BY Dish_Name

ORDER BY Total_Orders DESC

LIMIT 1;

```
hive> SELECT Dish_Name, COUNT(*) AS Total_Orders
    > FROM foodpanda_data
    > GROUP BY Dish_Name
    > ORDER BY Total_Orders DESC
    > LIMIT 1;
Query ID = cloudera_20251030221212_d4188bdf-8334-43c0-8a34-431012c84204
```

Out put:

```
Total MapReduce CPU Time Spent: 3 seconds 660 msec
OK
        2536
Time taken: 38.984 seconds, Fetched: 1 row(s)
hive>
```

Q.3 Which city generates the highest total revenue?

SELECT City, SUM(Price * Quantity) AS Total_Revenue

FROM foodpanda_data

GROUP BY City

ORDER BY Total_Revenue DESC;

```
hive> SELECT City, SUM(Price * Quantity) AS Total_Revenue
    > FROM foodpanda_data
    > GROUP BY City
    > ORDER BY Total_Revenue DESC;
Query ID = cloudera_20251030221616_0514db25-c3e8-4daa-b2fb-0373b860f8a3
```

Out put:

```
Total MapReduce CPU Time Spent: 3 seconds 960 msec
OK
city    NULL
Peshawar        NULL
Multan  NULL
Lahore  NULL
Karachi NULL
Islamabad       NULL
        NULL
Time taken: 38.639 seconds, Fetched: 7 row(s)
hive>
```

Q.4 Which payment type is used most frequently by customers?

SELECT Payment_Type, COUNT(*) AS Payment_Count

FROM foodpanda_data

GROUP BY Payment_Type

ORDER BY Payment_Count DESC LIMIT 10;

```
hive> SELECT Payment_Type, COUNT(*) AS Payment_Count
    > FROM foodpanda_data
    > GROUP BY Payment_Type
    > ORDER BY Payment_Count DESC LIMIT 10;
Query ID = cloudera_20251030222222_fa291fe8-047c-4410-8106-0252fbcc8bd6
```

Out put:

```
Total MapReduce CPU Time Spent: 4 seconds 660 msec
OK
        2536
235.82  3
661.62  2
1337.62 2
1114.25 2
1044.89 2
788.9   2
1271.26 2
288.68  2
1135.77 2
Time taken: 40.114 seconds, Fetched: 10 row(s)
hive>
```

Q.5 How many churned customers had a loyalty policy active?

SELECT Loyalty_Policy, Churned, COUNT(*) AS Total_Customers

FROM foodpanda_data

GROUP BY Loyalty_Policy, Churned LIMIT 15;

```
hive> SELECT Loyalty_Policy, Churned, COUNT(*) AS Total_Customers
    > FROM foodpanda_data
    > GROUP BY Loyalty_Policy, Churned LIMIT 15;
Query ID = cloudera_20251030222626_842e9d78-6bb4-4979-9fa1-c39b4ae349b5
```

Out put:

```
Total MapReduce CPU Time Spent: 2 seconds 250 msec
OK
                2536
01-01-2025      180     1
01-01-2025      181     1
01-01-2025      31      1
01-01-2025      355     1
01-01-2025      434     1
01-01-2025      436     1
01-01-2025      440     1
01-01-2025      451     1
01-01-2025      464     1
01-01-2025      61      1
01-02-2025      148     1
01-02-2025      251     1
01-02-2025      266     1
01-02-2025      300     1
Time taken: 18.545 seconds, Fetched: 15 row(s)
hive>
```

Q.6 Which restaurants have the highest average ratings?

SELECT Restaurant_Name, ROUND(AVG(Rating),2) AS Avg_Rating

FROM foodpanda_data

GROUP BY Restaurant_Name

ORDER BY Avg_Rating DESC

LIMIT 10;

```
hive> SELECT Restaurant_Name, ROUND(AVG(Rating),2) AS Avg_Rating
    > FROM foodpanda_data
    > GROUP BY Restaurant_Name
    > ORDER BY Avg_Rating DESC
    > LIMIT 10;
Query ID = cloudera_20251030222727_75dce4ca-ed76-4abc-a624-eda4a44bbc0f
```

Output:

```
Total MapReduce CPU Time Spent: 4 seconds 190 msec
OK
order_date      NULL
9/30/2024       NULL
9/30/2023       NULL
9/29/2024       NULL
9/29/2023       NULL
9/28/2024       NULL
9/28/2023       NULL
9/27/2024       NULL
9/27/2023       NULL
9/26/2024       NULL
Time taken: 38.802 seconds, Fetched: 10 row(s)
hive>
```

Q.7 Which food category is most popular among male and female customers?

SELECT Gender, Category, COUNT(*) AS Orders

FROM foodpanda_data

GROUP BY Gender, Category

ORDER BY Gender, Orders DESC;

```
hive> SELECT Gender, Category, COUNT(*) AS Orders
    > FROM foodpanda_data
    > GROUP BY Gender, Category
    > ORDER BY Gender, Orders DESC;
Query ID = cloudera_20251030222929_a032f21e-0546-45d5-ac89-62292aacefa1
```

Out Put:

```
Total MapReduce CPU Time Spent: 3 seconds 490 msec
OK
              2536
Female  Sandwich        245
Female  Pasta   242
Female  Burger  233
Female  Pizza   211
Female  Fries   206
Male    Pasta   264
Male    Sandwich        253
Male    Burger  226
Male    Fries   216
Male    Pizza   213
Other   Fries   242
Other   Pasta   240
Other   Pizza   232
Other   Burger  224
Other   Sandwich        214
gender  dish_name       1
Time taken: 36.428 seconds, Fetched: 17 row(s)
hive>
```

Q.8 What is the average order frequency across different age groups?

SELECT CASE

WHEN Age < 20 THEN 'Teen'

WHEN Age BETWEEN 20 AND 30 THEN 'Young Adult'

WHEN Age BETWEEN 31 AND 45 THEN 'Adult'

ELSE 'Senior'

END AS Age_Group,

ROUND(AVG(Order_Frequency),2) AS Avg_Frequency

FROM foodpanda_data

GROUP BY

```
CASE
  WHEN Age < 20 THEN 'Teen'
  WHEN Age BETWEEN 20 AND 30 THEN 'Young Adult'
  WHEN Age BETWEEN 31 AND 45 THEN 'Adult'
  ELSE 'Senior'
END;
```

```
hive> SELECT
    >   CASE
    >     WHEN Age < 20 THEN 'Teen'
    >     WHEN Age BETWEEN 20 AND 30 THEN 'Young Adult'
    >     WHEN Age BETWEEN 31 AND 45 THEN 'Adult'
    >     ELSE 'Senior'
    >   END AS Age_Group,
    >   ROUND(AVG(Order_Frequency),2) AS Avg_Frequency
    > FROM foodpanda_data
    > GROUP BY
    >   CASE
    >     WHEN Age < 20 THEN 'Teen'
    >     WHEN Age BETWEEN 20 AND 30 THEN 'Young Adult'
    >     WHEN Age BETWEEN 31 AND 45 THEN 'Adult'
    >     ELSE 'Senior'
    >   END;
Query ID = cloudera_20251030223232_ca26d055-b6ec-4ea3-8645-639e54a7f7da
```

Out Put:

```
Total MapReduce CPU Time Spent: 2 seconds 710 msec
OK
Senior  NULL
Time taken: 20.684 seconds, Fetched: 1 row(s)
hive>
```

Q.9 How many orders were delivered successfully vs delayed?

```
SELECT Delivery_Status, COUNT(*) AS Total_Orders
FROM foodpanda_data
GROUP BY Delivery_Status;
```

```
hive> SELECT Delivery_Status, COUNT(*) AS Total_Orders
    > FROM foodpanda_data
    > GROUP BY Delivery_Status;
Query ID = cloudera_20251030223434_49d2c2b4-5c82-4cc3-8720-64f7ec804abe
```

Out Put:

```
Total MapReduce CPU Time Spent: 1 seconds 950 msec
OK
        2536
01-01-2025      12
01-02-2025      14
01-03-2025      7
01-04-2025      6
01-05-2025      8
01-06-2025      10
01-07-2025      5
01-08-2025      6
01-09-2025      7
01-10-2025      8
01-11-2025      7
01-12-2025      8
02-01-2025      8
02-02-2025      9
02-03-2025      5
02-04-2025      10
02-05-2025      3
02-06-2025      14
02-07-2025      3
02-08-2025      8
02-09-2025      9
02-10-2025      9
02-11-2025      6
02-12-2025      8
03-01-2025      17
03-02-2025      7
03-03-2025      6
03-04-2025      15
```

Q.10 Which city has the highest total revenue based on the sum of all orders?

```
SELECT City,
    SUM(Price * Quantity) AS Total_Revenue
FROM foodpanda_data
GROUP BY City
ORDER BY Total_Revenue DESC
LIMIT 1;
```

```
hive> SELECT City,
    >        SUM(Price * Quantity) AS Total_Revenue
    > FROM foodpanda_data
    > GROUP BY City
    > ORDER BY Total_Revenue DESC
    > LIMIT 1;
Query ID = cloudera_20251030223737_4585632f-9ff9-464d-84f6-2a49880e299b
```

Out Put:

```
Total MapReduce CPU Time Spent: 4 seconds 20 msec
OK
city    NULL
Time taken: 39.082 seconds, Fetched: 1 row(s)
hive>
```