

### Nhóm 01:

Họ và tên	Mã sinh viên
Đỗ Phương Mai Anh	2121051257
Nguyễn Hồng Khanh	2121050122

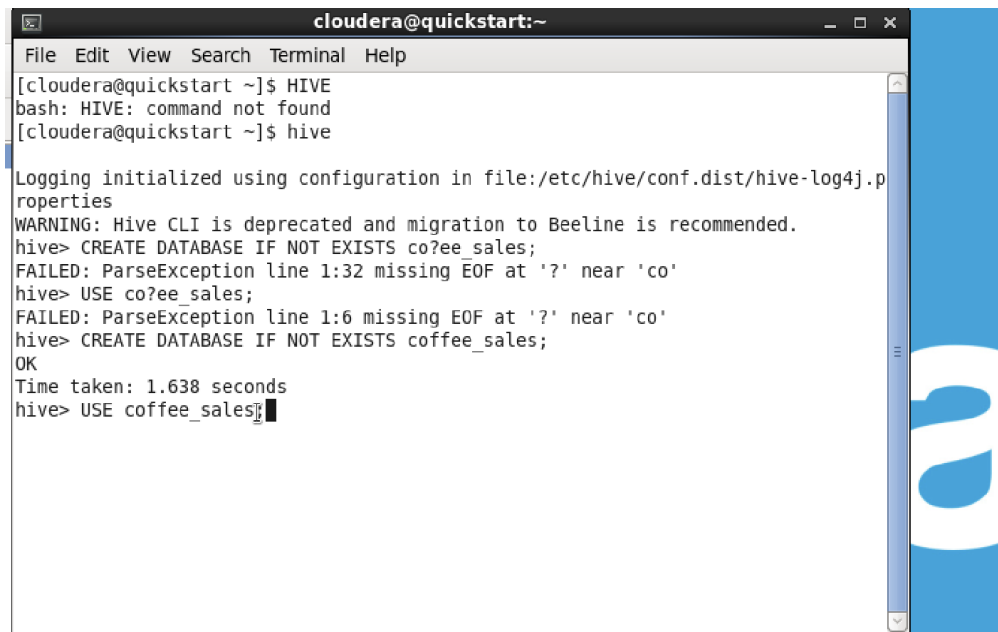
#### Bước 1: Chuẩn bị Tập Dữ Liệu

- Sử dụng tập dữ liệu Coffee Sales “index.csv”

#### Bước 2: Tạo Database và Bảng trong Hive

- Tạo Database:** Tạo database với tên coffee\_sales để quản lý dữ liệu.

CREATE DATABASE coffee\_sales;

A screenshot of a terminal window titled "cloudera@quickstart:~". The terminal shows the following commands and output:

```
[cloudera@quickstart ~]$ HIVE
bash: HIVE: command not found
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE DATABASE IF NOT EXISTS co?ee_sales;
FAILED: ParseException line 1:32 missing EOF at '?' near 'co'
hive> USE co?ee_sales;
FAILED: ParseException line 1:6 missing EOF at '?' near 'co'
hive> CREATE DATABASE IF NOT EXISTS coffee_sales;
OK
Time taken: 1.638 seconds
hive> USE coffee_sales;
```

- Tạo bảng** cho tập dữ liệu:

CREATE TABLE coffee\_sales (

order\_id STRING,

order\_date DATE,

product\_id STRING,

quantity INT,

customer\_id STRING,

```
sales DECIMAL(10,2)
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

```
FAILED: ParseException line 1:6 missing EOF at '?' near 'co'
hive> CREATE DATABASE IF NOT EXISTS coffee_sales;
OK
Time taken: 1.638 seconds
hive> CREATE TABLE coffee_sales (
  > order_id STRING,
  > order_date DATE,
  > product_id STRING,
  > quantity INT,
  > customer_id STRING,
  > sales DECIMAL(10,2)
  > )
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ','
  > STORED AS TEXTFILE;
OK
Time taken: 3.787 seconds
hive>
```

### Bước 3: Nhập Dữ Liệu vào Hive

Tải dữ liệu index.csv vào các bảng coffee\_sales trong Hive:

```
LOAD DATA LOCAL INPATH '/home/cloudera/index.csv' INTO TABLE coffee_sales;
```

```
V': No files matching path file:/path/to/your/coffee_sales.csv
hive> LOAD DATA LOCAL INPATH '/home/cloudera/index.csv' INTO TABLE coffee_sales;
Loading data to table default.coffee_sales
Table default.coffee_sales stats: [numFiles=1, totalSize=166530]
OK
Time taken: 7.415 seconds
hive>
```

### Bước 4: Thực hiện các Truy vấn

#### 1. Phân tích doanh số theo thời gian:

```
SELECT
  YEAR(order_date) as year,
  MONTH(order_date) as month,
  SUM(sales) as total_sales,
  COUNT(DISTINCT customer_id) as unique_customers,
```

```

SUM(sales)/COUNT(DISTINCT customer_id) as avg_sales_per_customer
FROM coffee_sales
GROUP BY YEAR(order_date), MONTH(order_date)
ORDER BY year, month;

```

```

hive> SELECT
  > YEAR(order_date) as year,
  > MONTH(order_date) as month,
  > SUM(sales) as total_sales,
  > COUNT(DISTINCT customer_id) as unique_customers,
  > SUM(sales)/COUNT(DISTINCT customer_id) as avg_sales_per_customer
  > FROM coffee_sales
  > GROUP BY YEAR(order_date), MONTH(order_date)
  > ORDER BY year, month;

```

```

3 HDFS Write: 114 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 58.37 sec HDFS Read: 6286 H
DFS Write: 15 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 38 seconds 440 msec
OK
NULL NULL NULL 21 NULL
Time taken: 799.693 seconds, Fetched: 1 row(s)

```

- Ý nghĩa:
  - Theo dõi doanh số theo từng tháng, năm
  - Biết được số lượng khách hàng active trong mỗi tháng
  - Tính được mức chi tiêu trung bình của mỗi khách hàng

## 2. Phân tích thời điểm bán hàng trong ngày:

```

SELECT
CASE
  WHEN HOUR(order_date) BETWEEN 6 AND 10 THEN 'Buổi sáng'
  WHEN HOUR(order_date) BETWEEN 11 AND 14 THEN 'Buổi trưa'
  WHEN HOUR(order_date) BETWEEN 15 AND 18 THEN 'Buổi chiều'
  ELSE 'Buổi tối'
END as time_of_day,
COUNT(*) as number_of_orders,
ROUND(AVG(quantity), 2) as avg_items_per_order,
ROUND(SUM(sales), 2) as total_sales

```

FROM coffee\_sales

GROUP BY CASE

WHEN HOUR(order\_date) BETWEEN 6 AND 10 THEN 'Buổi sáng'

WHEN HOUR(order\_date) BETWEEN 11 AND 14 THEN 'Buổi trưa'

WHEN HOUR(order\_date) BETWEEN 15 AND 18 THEN 'Buổi chiều'

ELSE 'Buổi tối'

END

ORDER BY number\_of\_orders DESC;

```
hive> SELECT
  > CASE
  > WHEN HOUR(order_date) BETWEEN 6 AND 10 THEN 'Buổi sáng'
  > WHEN HOUR(order_date) BETWEEN 11 AND 14 THEN 'Buổi trưa'
  > WHEN HOUR(order_date) BETWEEN 15 AND 18 THEN 'Buổi chiều'
  > ELSE 'Buổi tối'
  > END as time_of_day,
  > COUNT(*) as number_of_orders,
  > ROUND(AVG(quantity), 2) as avg_items_per_order,
  > ROUND(SUM(sales), 2) as total_sales
  > FROM coffee_sales
  > GROUP BY CASE
  > WHEN HOUR(order_date) BETWEEN 6 AND 10 THEN 'Buổi sáng'
  > WHEN HOUR(order_date) BETWEEN 11 AND 14 THEN 'Buổi trưa'
  > WHEN HOUR(order_date) BETWEEN 15 AND 18 THEN 'Buổi chiều'
  > ELSE 'Buổi tối'
  > END
  > ORDER BY number_of_orders DESC;
Query ID = cloudera_20241111194141_7d7d1f47-66e8-4def-a027-4eac568ca98b

Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 65.3 sec HDFS Read: 6104 HD
FS Write: 30 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 40 seconds 740 msec
OK
Buổi tối          2176      NULL      NULL
Time taken: 642.367 seconds, Fetched: 1 row(s)
hive>
```

- Ý nghĩa:

- Xác định khung giờ bán hàng tốt nhất
- Biết được thói quen mua sắm của khách hàng theo giờ
- Hỗ trợ việc sắp xếp nhân viên theo ca làm việc

### 3. Phân tích xu hướng sản phẩm:

SELECT

product\_id,

```

COUNT(*) as times_ordered,
SUM(quantity) as total_quantity_sold,
SUM(sales) as total_revenue,
AVG(sales/quantity) as avg_price_per_unit
FROM coffee_sales
GROUP BY product_id
HAVING COUNT(*) > 10
ORDER BY total_revenue DESC;

```

```

hive> SELECT
>   product_id,
>   COUNT(*) as times_ordered,
>   SUM(quantity) as total_quantity_sold,
>   SUM(sales) as total_revenue,
>   AVG(sales/quantity) as avg_price_per_unit
> FROM coffee_sales
> GROUP BY product_id
> HAVING COUNT(*) > 10
> ORDER BY total_revenue DESC;
Query ID = cloudera 20241111191919 c83b4e80-4b33-45d8-b2b5-036365859138

Total MapReduce CPU Time Spent: 2 minutes 29 seconds 200 msec
OK
cash      89      NULL      NULL      NULL
card     2086     NULL      NULL      NULL
Time taken: 613.424 seconds, Fetched: 2 row(s)

```

- Ý nghĩa:
  - Xác định sản phẩm bán chạy nhất
  - Phân tích giá bán trung bình của từng sản phẩm
  - Đánh giá hiệu suất bán hàng của từng sản phẩm

#### 4. Phân tích mùa vụ:

```

SELECT
CASE
    WHEN MONTH(order_date) IN (12,1,2) THEN 'Mùa Đông'
    WHEN MONTH(order_date) IN (3,4,5) THEN 'Mùa Xuân'
    WHEN MONTH(order_date) IN (6,7,8) THEN 'Mùa Hè'
    ELSE 'Mùa Thu'

```

```

END as season,
COUNT(*) as total_orders,
SUM(sales) as total_sales,
AVG(quantity) as avg_quantity_per_order
FROM coffee_sales
GROUP BY CASE
    WHEN MONTH(order_date) IN (12,1,2) THEN 'Mùa Đông'
    WHEN MONTH(order_date) IN (3,4,5) THEN 'Mùa Xuân'
    WHEN MONTH(order_date) IN (6,7,8) THEN 'Mùa Hè'
    ELSE 'Mùa Thu'
END;

```

```

hive> SELECT
> CASE
>   WHEN MONTH(order_date) IN (12,1,2) THEN 'Mùa Đông'
>   WHEN MONTH(order_date) IN (3,4,5) THEN 'Mùa Xuân'
>   WHEN MONTH(order_date) IN (6,7,8) THEN 'Mùa Hè'
>   ELSE 'Mùa Thu'
> END as season,
> COUNT(*) as total_orders,
> SUM(sales) as total_sales,
> AVG(quantity) as avg_quantity_per_order
> FROM coffee_sales
> GROUP BY CASE
>   WHEN MONTH(order_date) IN (12,1,2) THEN 'Mùa Đông'
>   WHEN MONTH(order_date) IN (3,4,5) THEN 'Mùa Xuân'
>   WHEN MONTH(order_date) IN (6,7,8) THEN 'Mùa Hè'
>   ELSE 'Mùa Thu'
> END;

HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 25 seconds 600 msec
OK
Mùa Thu 2176      NULL      NULL
Time taken: 337.012 seconds, Fetched: 1 row(s)
hive>

```

- Ý nghĩa:
  - Phân tích tính thời vụ trong việc bán hàng
  - Xác định mùa bán hàng tốt nhất
  - Hiểu được hành vi mua hàng theo mùa của khách hàng

