Table of Contents

Appendix

Introduction

The objective of my project is to analyze and visualize hotel booking data using bigdata concepts like Hadoop Distributed File System (HDFS) and Apache Hive with an aim to identify patterns and trends including booking preferences, cancellations, family setup of the bookings (adults with kids), performance of an agent (if used), weekend vs weekday stays insights, performance of hotel type (resort vs city), performance by month, preference for meals, market segment (online, corporate, direct), room type preference, whether the guest got his or her reserved room type, number of special requests, which can be utilized by a hotel management organization to effectively plan its operations.

Methodology

The dataset that I am using is from Kaggle, ((Hotel booking demand, n.d.). I stored the file using HDFS (Figure 1). A distributed system like HDFS is best suited for analytical systems. I used Hive for analysis and querying. Hive is an open source Datawarehouse. Hive is meant to solve analytical problems. Hive will be based on a cluster of machines and not monolithic systems. Since we have structured data, zi to visualized by writing SQL like queries or Hive query language to obtain data in the form of tables. Actual data is stored in HDFS and therefore, it would not support updates and cannot be retrieved very quickly. Metadata is stored in a database like MySQL or Hive Metastore. Therefore, I stored in Hive Metastore (Figure 2).

I then removed country column having null values as other columns such as agent company having null value meant that there was no agent used and was therefore relevant to the analysis (Figure 3).

Results and Discussion

1. Number of nationalities represented in dataset (Figure 4).

   I noticed that the nationalities represented in the data were comprehensive and representative of 178 countries.

2. Preference for room types (Figure 5)

   There is a clear preference for room type A (count of 85k+ instances) followed by a wide gap by D (19k+ instances). Other room types were aggregated less than 3,000 instances.

3. If requested room type was allocated (Figure 6)

The comparison data is observed to indicate that the requested room type was mostly allocated to the visitors.

4. Sum of previous cancellations by hotel type (Figure 7)

Reservations to city hotels were cancelled at a rate 150% higher than those for Resort hotels.

5. Preference for meal types (Figure 8)

Analysis of data indicates a clear preference for BB food type followed by a wide gap of almost 80k instances by HB type. Other food types were even lower.

6. Sum of weekend stays vs weekday stays (Figure 9)

The sum of weekend nights stayed were observed to be about 110k and weeknights were observed to be 298k. However, I rationalized these results by dividing them by number of days (5 for weeknights and 2 for weekend nights. This resulted in about 60k weeknights and 55k for weekend nights indicating better occupancy for weeknights.

7. Count of families with kids (Figure 10)

The count of families staying in these hotels was less than 10% of the dataset.

8. Performance of agent (Figure 11)

Identification of agents by number of bookings clearly indicated that agent 9 was making most of the bookings (approximately 30% of all reservations).

9. Performance of market segment (Figure 12)

Online travel agents were responsible for about 50% of all reservations followed by offline travel agents or tour operators representing about 20% of reservations. Group reservations indicated a significant >15% reservations as well.

10. Arrival month pattern (Figure 13)

Arrival month was evenly distributed with January, November, and December appearing to be low occupancy months.

11. Average day rate by market segment (Figure 14)

Online bookings had the highest average day rates followed by direct booking, aviation, offline, groups, and corporate. Corporates were observed to get the best rates if undefined and complementary bookings were ignored.

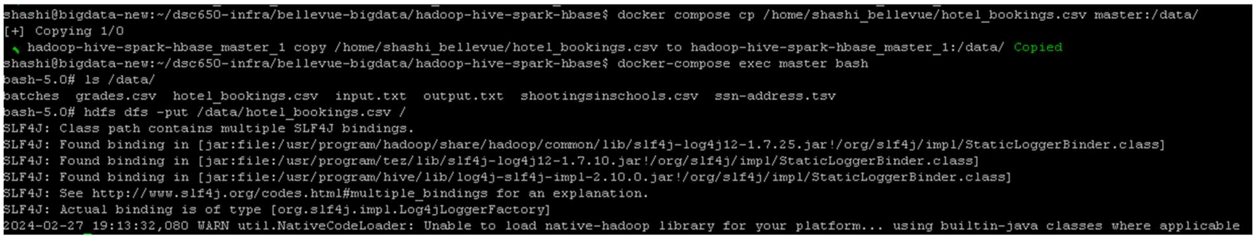
Conclusion

Following operational conclusions of significance were drawn:

1. Given the demand for type A rooms, hotels can look to convert other room types to type A.
2. City hotels should investigate ways to reduce the number of cancellations.
3. Hotels must include food type BB in their menu considering overwhelming demand for this food type.
4. Since occupancy rates for weekend nights were found to be lower, the hotels could focus on marketing in ways to attract visitors during weekend nights like involving agent 9 who was a high performing booking agent.
5. If a hotel did not have an online portal it must invest in this booking platform as about 50% of reservations were made using this platform. In addition, the day rates were observed to be better though this platform.
6. Hotels could use November, December, and January months for renovations were number of bookings were significantly lower for these months than others.

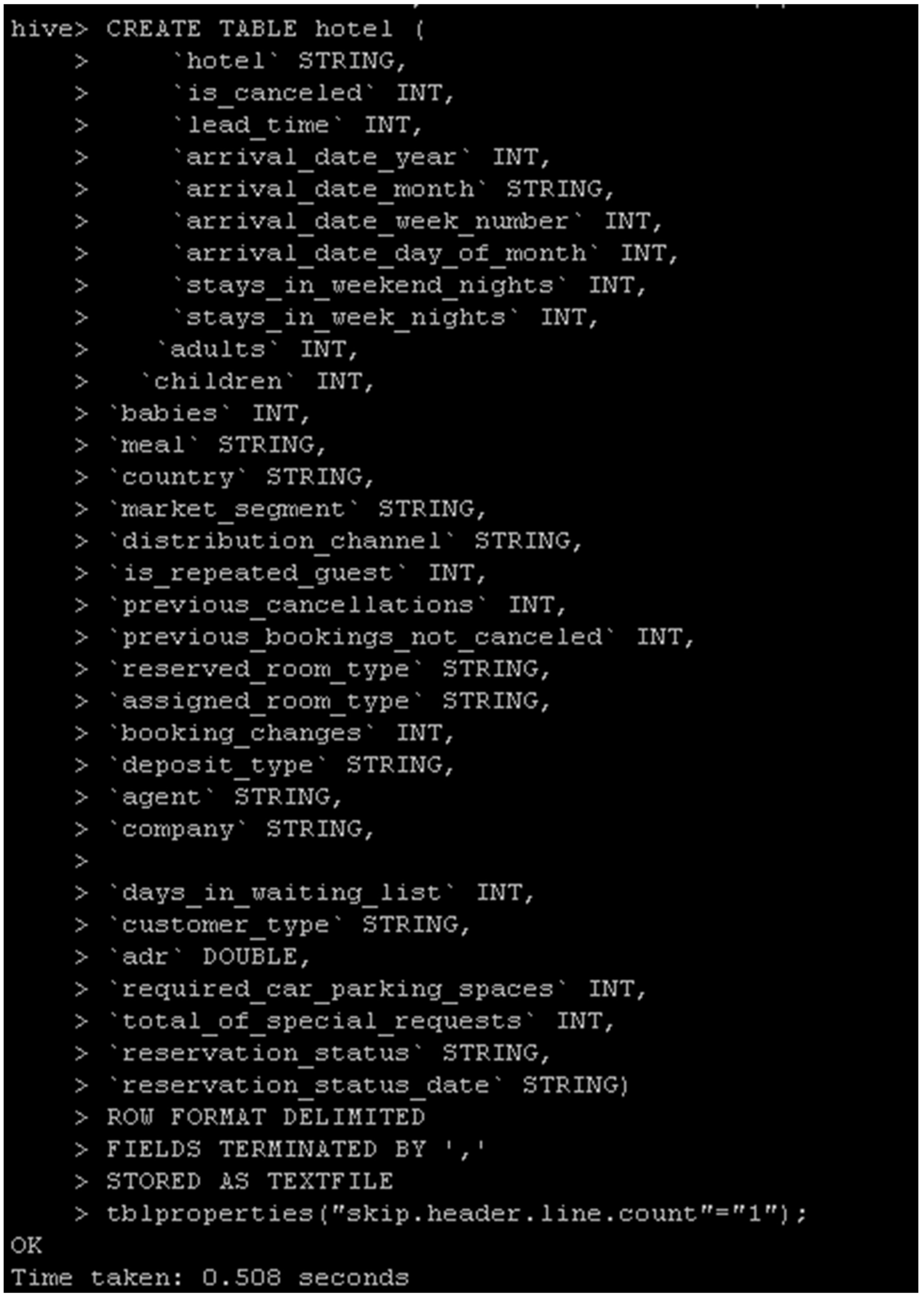
Appendix

Figure 1

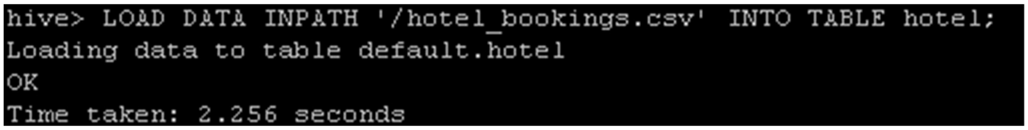


```
shashi@bigdata-new:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker compose cp /home/shashi_bellevue/hotel_bookings.csv master:/data/
[+] Copying 1/0
 ✔ hadoop-hive-spark-hbase_master_1 copy /home/shashi_bellevue/hotel_bookings.csv to hadoop-hive-spark-hbase_master_1:/data/ Copied
shashi@bigdata-new:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker-compose exec master bash
bash-5.0# ls /data/
batches  grades.csv  hotel_bookings.csv  input.txt  output.txt  shootingsinschools.csv  ssn-address.tsv
bash-5.0# hdfs dfs -put /data/hotel_bookings.csv /
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-02-27 19:13:32,080 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Figure 2

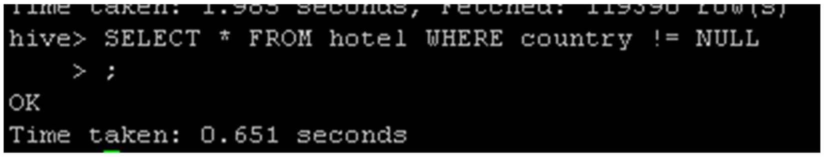


```
hive> CREATE TABLE hotel (
    >     `hotel` STRING,
    >     `is_canceled` INT,
    >     `lead_time` INT,
    >     `arrival_date_year` INT,
    >     `arrival_date_month` STRING,
    >     `arrival_date_week_number` INT,
    >     `arrival_date_day_of_month` INT,
    >     `stays_in_weekend_nights` INT,
    >     `stays_in_week_nights` INT,
    >    `adults` INT,
    >   `children` INT,
    > `babies` INT,
    > `meal` STRING,
    > `country` STRING,
    > `market_segment` STRING,
    > `distribution_channel` STRING,
    > `is_repeated_guest` INT,
    > `previous_cancellations` INT,
    > `previous_bookings_not_canceled` INT,
    > `reserved_room_type` STRING,
    > `assigned_room_type` STRING,
    > `booking_changes` INT,
    > `deposit_type` STRING,
    > `agent` STRING,
    > `company` STRING,
    >
    > `days_in_waiting_list` INT,
    > `customer_type` STRING,
    > `adr` DOUBLE,
    > `required_car_parking_spaces` INT,
    > `total_of_special_requests` INT,
    > `reservation_status` STRING,
    > `reservation_status_date` STRING)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.508 seconds
```



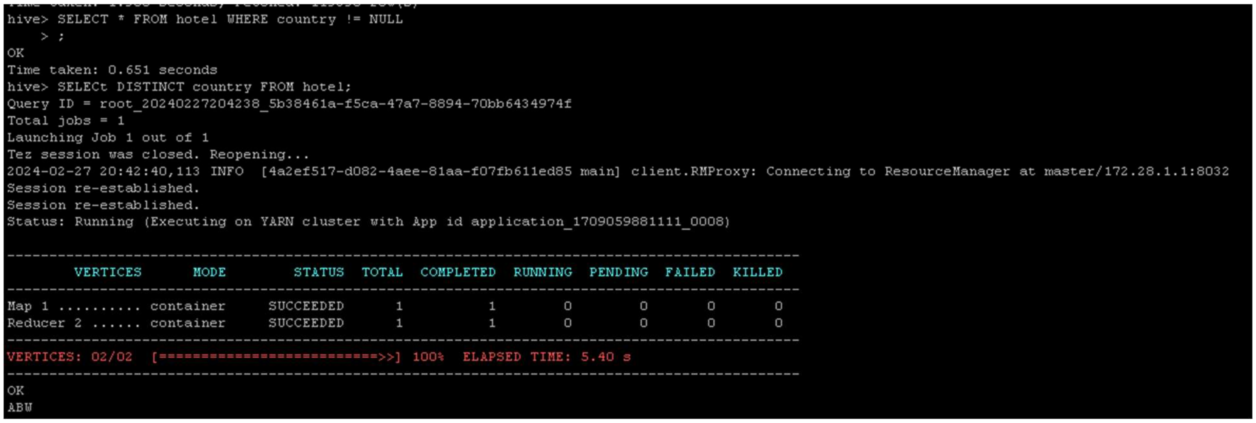
```
hive> LOAD DATA INPATH '/hotel_bookings.csv' INTO TABLE hotel;
Loading data to table default.hotel
OK
Time taken: 2.256 seconds
```

Figure 3

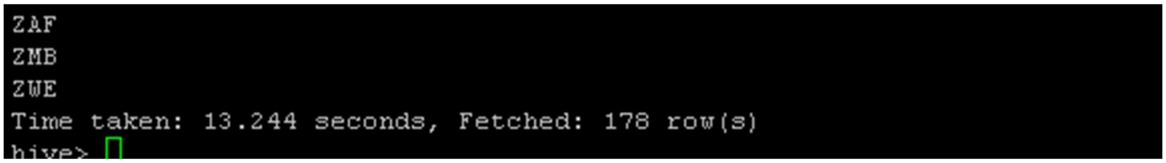


```
Time taken: 1.985 seconds, Fetched: 119390 row(s)
hive> SELECT * FROM hotel WHERE country != NULL
    > ;
OK
Time taken: 0.651 seconds
```

Figure 4

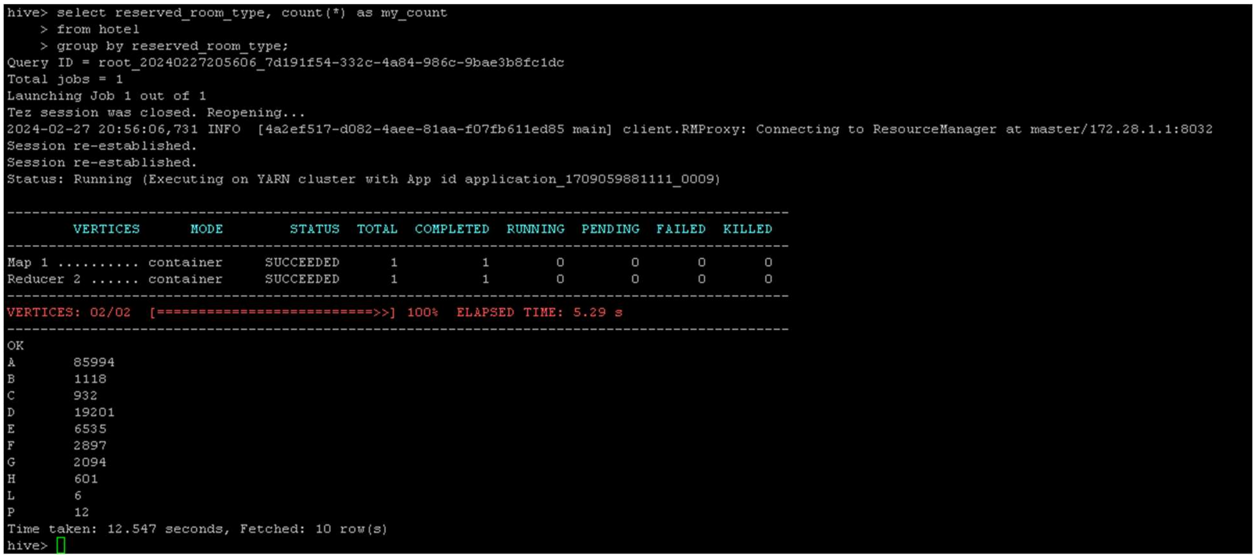


```
hive> SELECT * FROM hotel WHERE country != NULL
    > ;
OK
Time taken: 0.651 seconds
hive> SELECt DISTINCT country FROM hotel;
Query ID = root_20240227204238_5b38461a-f5ca-47a7-8894-70bb6434974f
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-02-27 20:42:40,113 INFO  [4a2ef517-d082-4aee-81aa-f07fb611ed85 main] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1709059881111_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.40 s
----------------------------------------------------------------------------------------------
OK
ABW
```



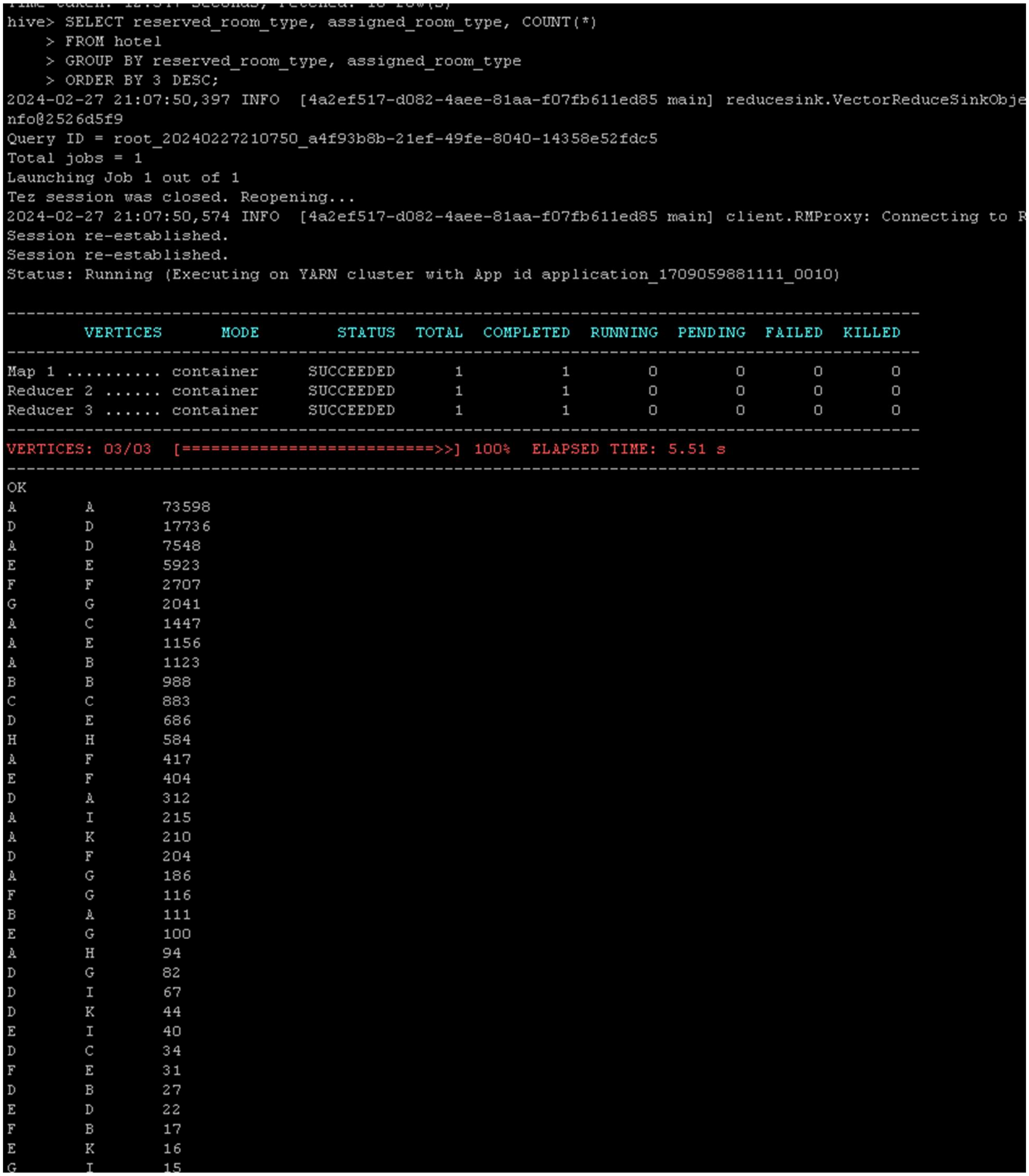
```
ZAF
ZMB
ZWE
Time taken: 13.244 seconds, Fetched: 178 row(s)
hive>
```

Figure 5

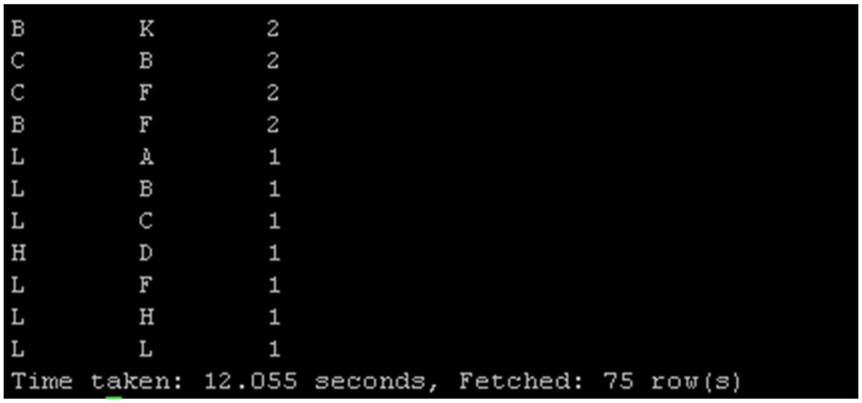


```
hive> select reserved_room_type, count(*) as my_count
    > from hotel
    > group by reserved_room_type;
Query ID = root_20240227205606_7d191f54-332c-4a84-986c-9bae3b8fc1dc
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-02-27 20:56:06,731 INFO  [4a2ef517-d082-4aee-81aa-f07fb611ed85 main] client.RMProxy: Connecting to ResourceManager at master/172.28.1.1:8032
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1709059881111_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.29 s
----------------------------------------------------------------------------------------------
OK
A       85994
B       1118
C       932
D       19201
E       6535
F       2897
G       2094
H       601
L       6
P       12
Time taken: 12.547 seconds, Fetched: 10 row(s)
hive>
```

Figure 6

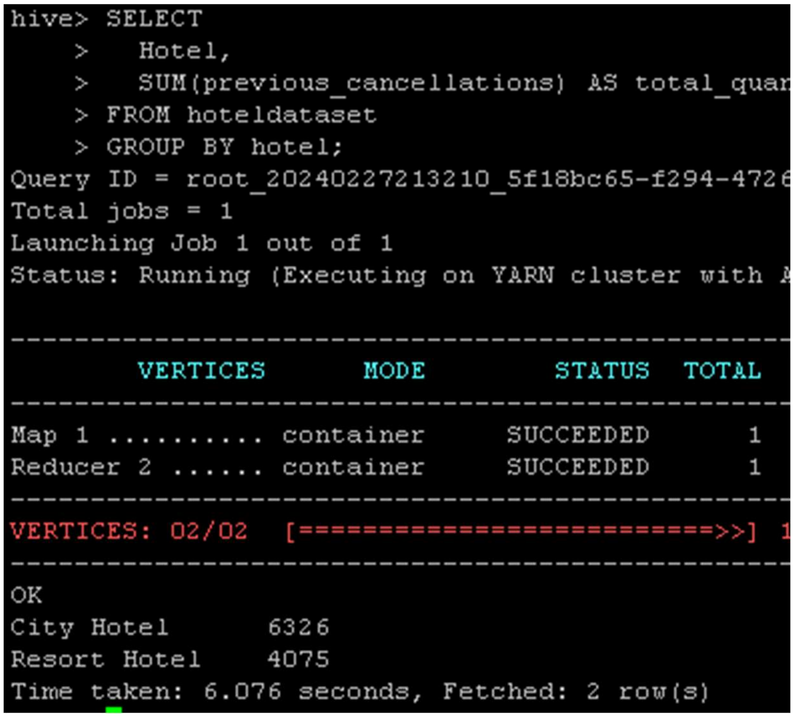


```
Time taken: 12.517 seconds, Fetched: 18 row(s)
hive> SELECT reserved_room_type, assigned_room_type, COUNT(*)
    > FROM hotel
    > GROUP BY reserved_room_type, assigned_room_type
    > ORDER BY 3 DESC;
2024-02-27 21:07:50,397 INFO  [4a2ef517-d082-4aee-81aa-f07fb611ed85 main] reducesink.VectorReduceSinkObje
nfo@2526d5f9
Query ID = root_20240227210750_a4f93b8b-21ef-49fe-8040-14358e52fdc5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-02-27 21:07:50,574 INFO  [4a2ef517-d082-4aee-81aa-f07fb611ed85 main] client.RMProxy: Connecting to R
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1709059881111_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.51 s
----------------------------------------------------------------------------------------------
OK
A       A       73598
D       D       17736
A       D       7548
E       E       5923
F       F       2707
G       G       2041
A       C       1447
A       E       1156
A       B       1123
B       B       988
C       C       883
D       E       686
H       H       584
A       F       417
E       F       404
D       A       312
A       I       215
A       K       210
D       F       204
A       G       186
F       G       116
B       A       111
E       G       100
A       H       94
D       G       82
D       I       67
D       K       44
E       I       40
D       C       34
F       E       31
D       B       27
E       D       22
F       B       17
E       K       16
G       I       15
```



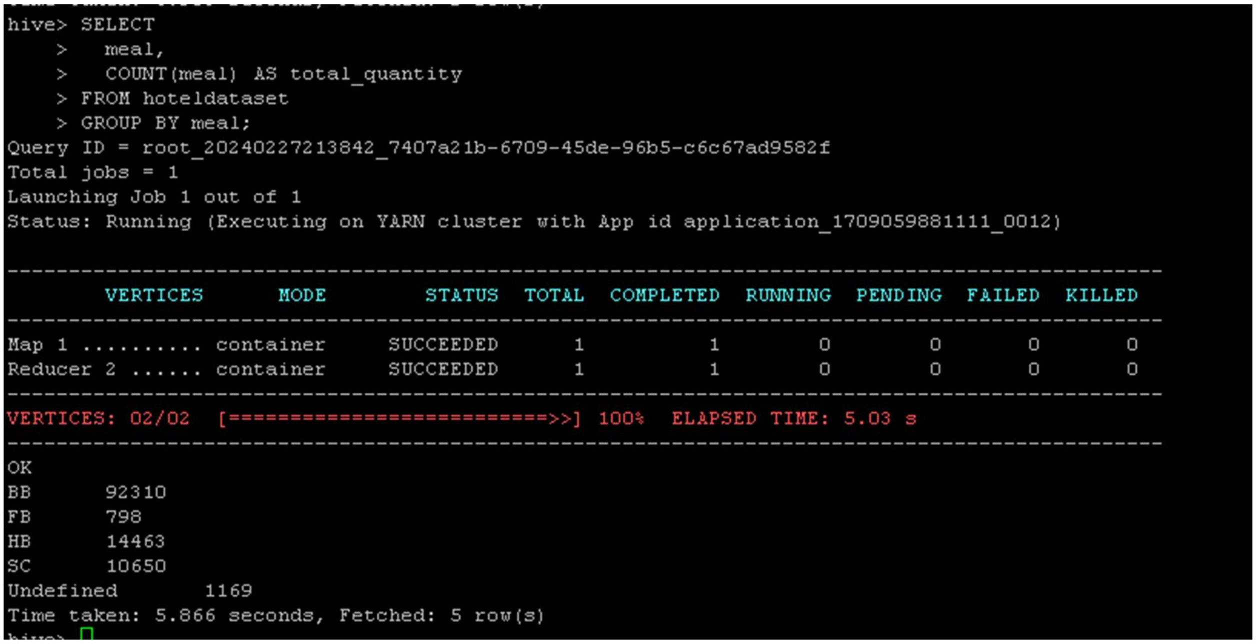
```
B       K       2
C       B       2
C       F       2
B       F       2
L       A       1
L       B       1
L       C       1
H       D       1
L       F       1
L       H       1
L       L       1
Time taken: 12.055 seconds, Fetched: 75 row(s)
```

Figure 7



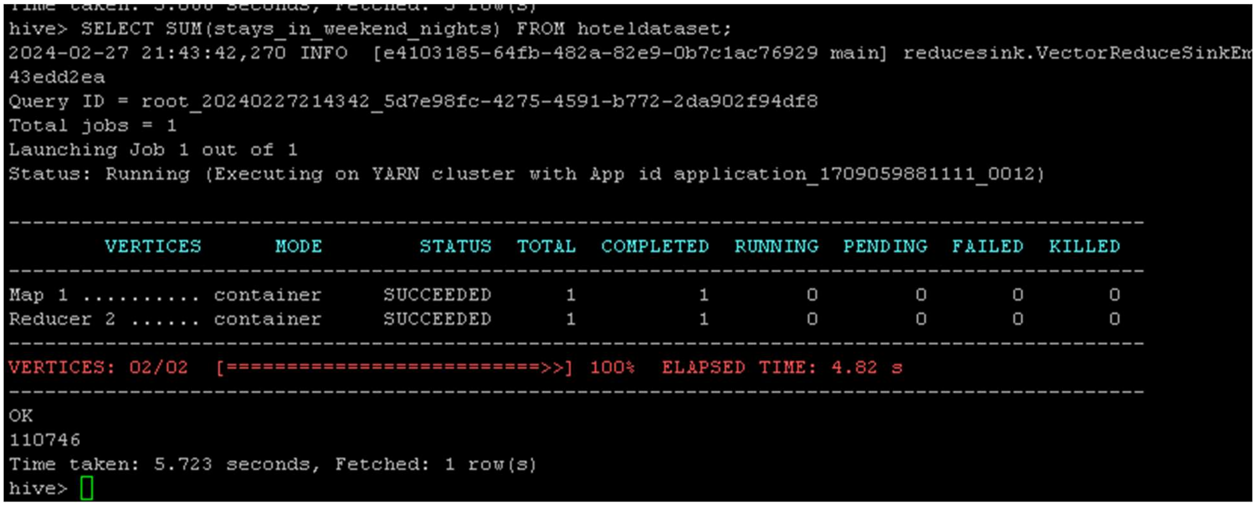
```
hive> SELECT
    >   Hotel,
    >   SUM(previous_cancellations) AS total_quan
    > FROM hoteldataset
    > GROUP BY hotel;
Query ID = root_20240227213210_5f18bc65-f294-4726
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with A

-------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL
-------------------------------------------------
Map 1 .......... container     SUCCEEDED      1
Reducer 2 ...... container     SUCCEEDED      1
-------------------------------------------------
VERTICES: 02/02  [==========================>>]  1
-------------------------------------------------
OK
City Hotel      6326
Resort Hotel    4075
Time taken: 6.076 seconds, Fetched: 2 row(s)
```

Figure 8

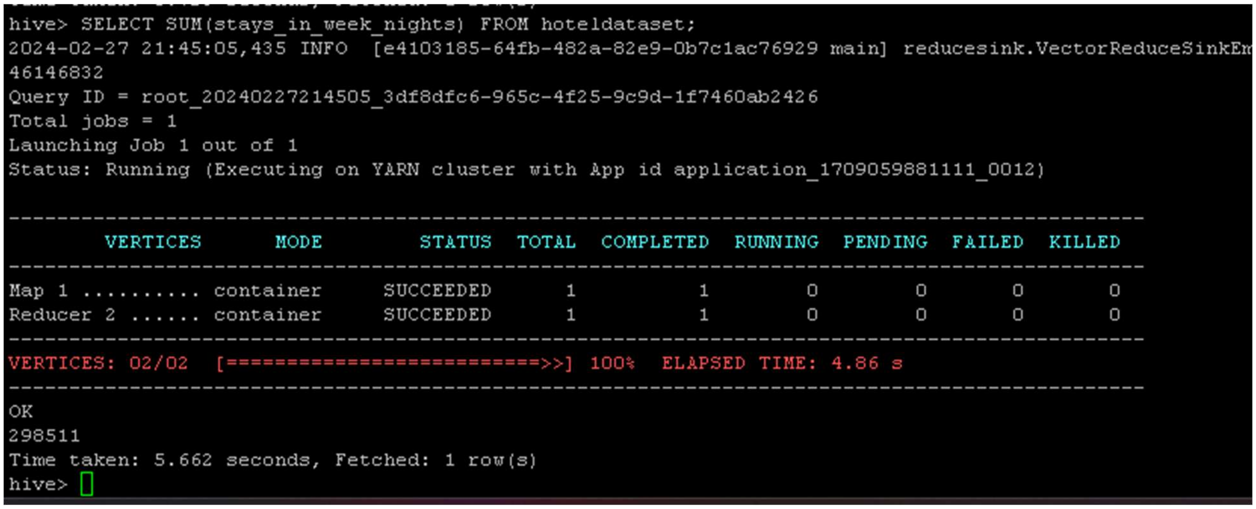


```
hive> SELECT
    >   meal,
    >   COUNT(meal) AS total_quantity
    > FROM hoteldataset
    > GROUP BY meal;
Query ID = root_20240227213842_7407a21b-6709-45de-96b5-c6c67ad9582f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709059881111_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.03 s
----------------------------------------------------------------------------------------------
OK
BB      92310
FB      798
HB      14463
SC      10650
Undefined       1169
Time taken: 5.866 seconds, Fetched: 5 row(s)
hive>
```

Figure 9

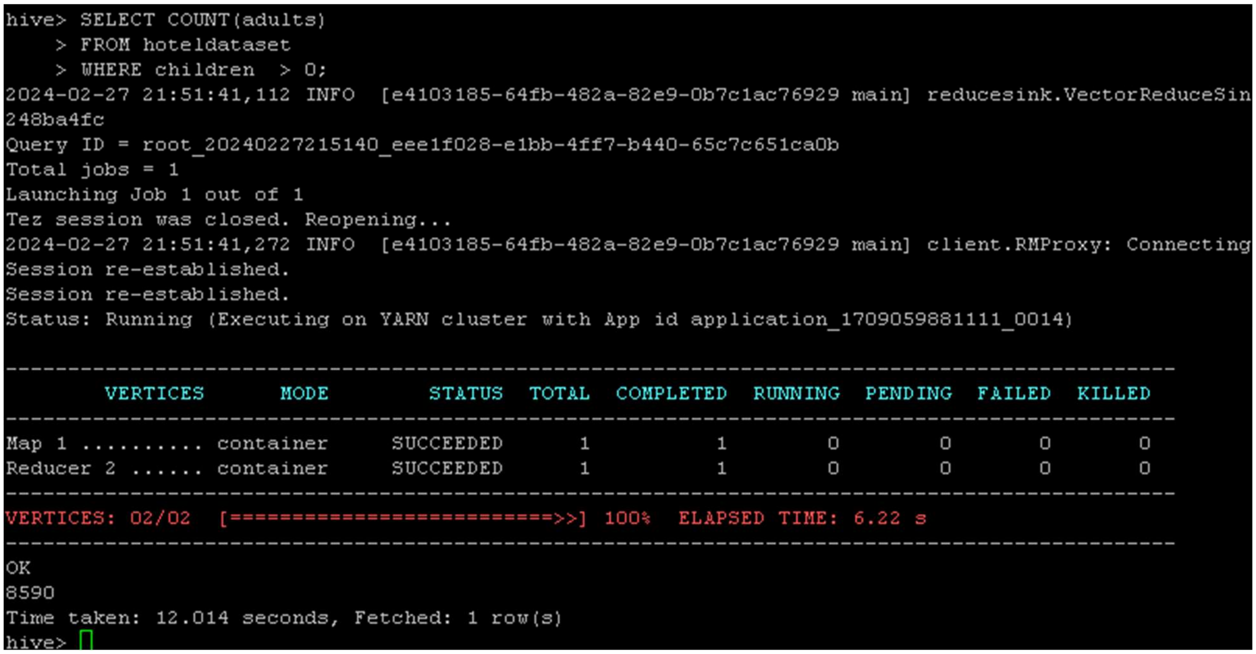


```
hive> SELECT SUM(stays_in_weekend_nights) FROM hoteldataset;
2024-02-27 21:43:42,270 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduceSinkEm
43edd2ea
Query ID = root_20240227214342_5d7e98fc-4275-4591-b772-2da902f94df8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709059881111_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.82 s
----------------------------------------------------------------------------------------------
OK
110746
Time taken: 5.723 seconds, Fetched: 1 row(s)
hive>
```



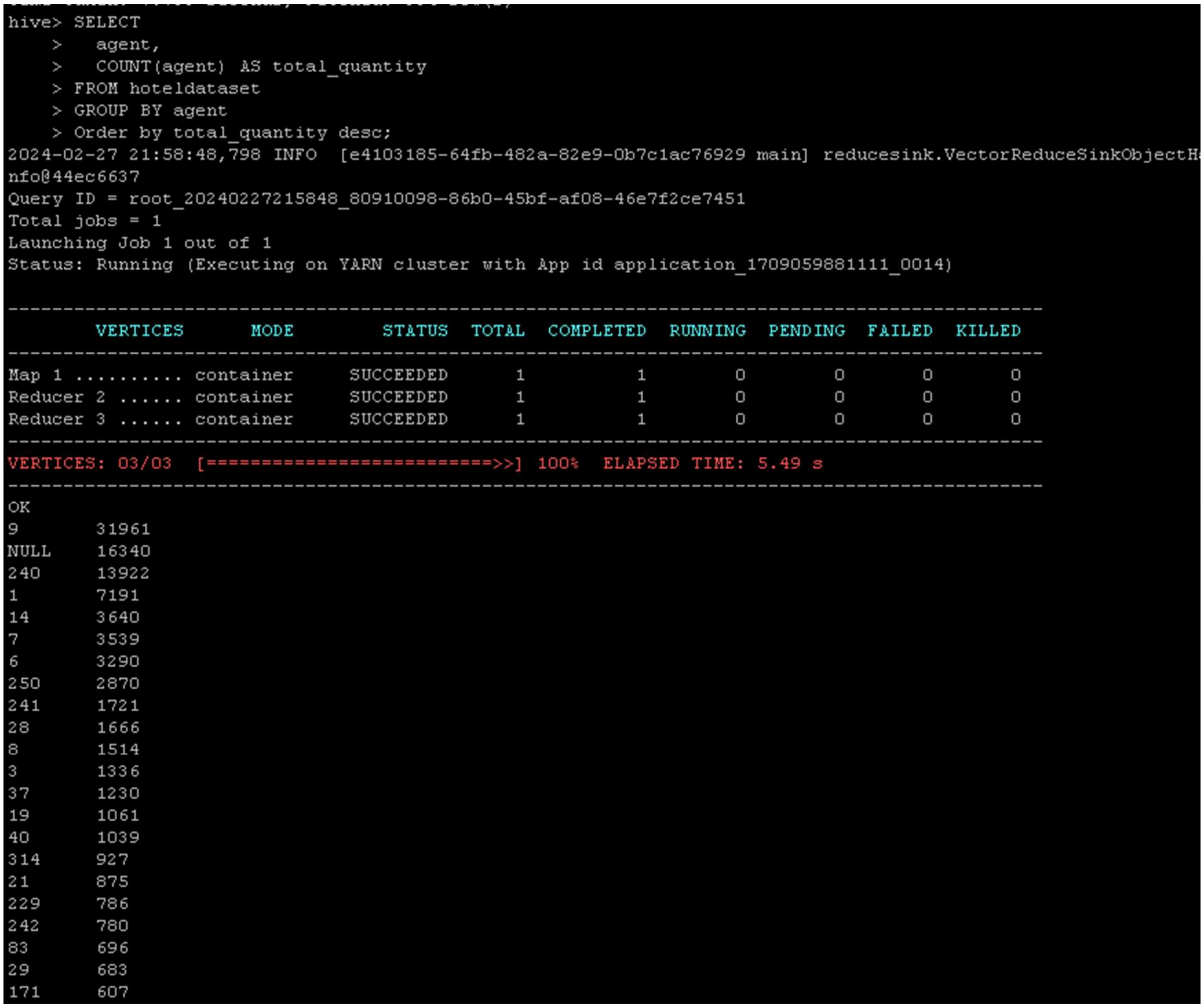
```
hive> SELECT SUM(stays_in_week_nights) FROM hoteldataset;
2024-02-27 21:45:05,435 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduceSinkEm
46146832
Query ID = root_20240227214505_3df8dfc6-965c-4f25-9c9d-1f7460ab2426
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709059881111_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.86 s
----------------------------------------------------------------------------------------------
OK
298511
Time taken: 5.662 seconds, Fetched: 1 row(s)
hive> 
```

Figure 10



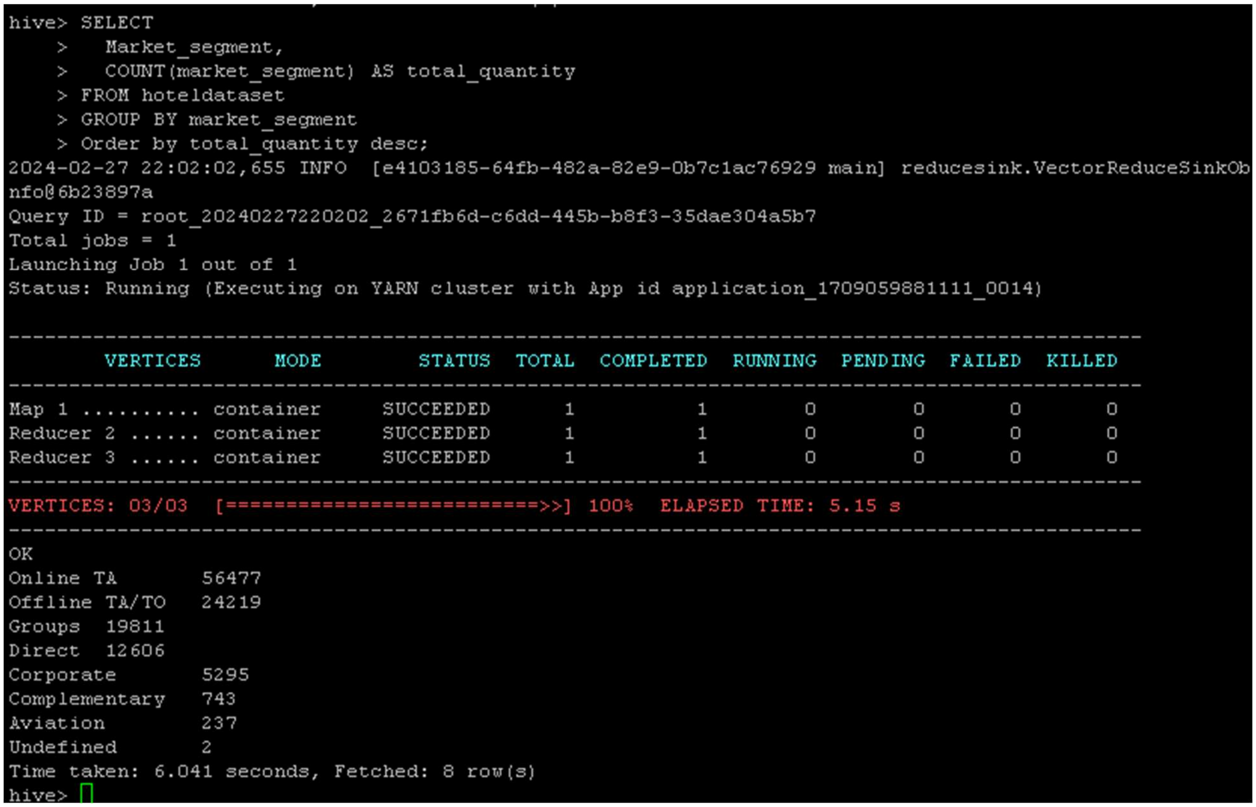
```
hive> SELECT COUNT(adults)
    > FROM hoteldataset
    > WHERE children  > 0;
2024-02-27 21:51:41,112 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduceSin
248ba4fc
Query ID = root_20240227215140_eee1f028-e1bb-4ff7-b440-65c7c651ca0b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-02-27 21:51:41,272 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] client.RMProxy: Connecting
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1709059881111_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.22 s
----------------------------------------------------------------------------------------------
OK
8590
Time taken: 12.014 seconds, Fetched: 1 row(s)
hive> 
```

Figure 11



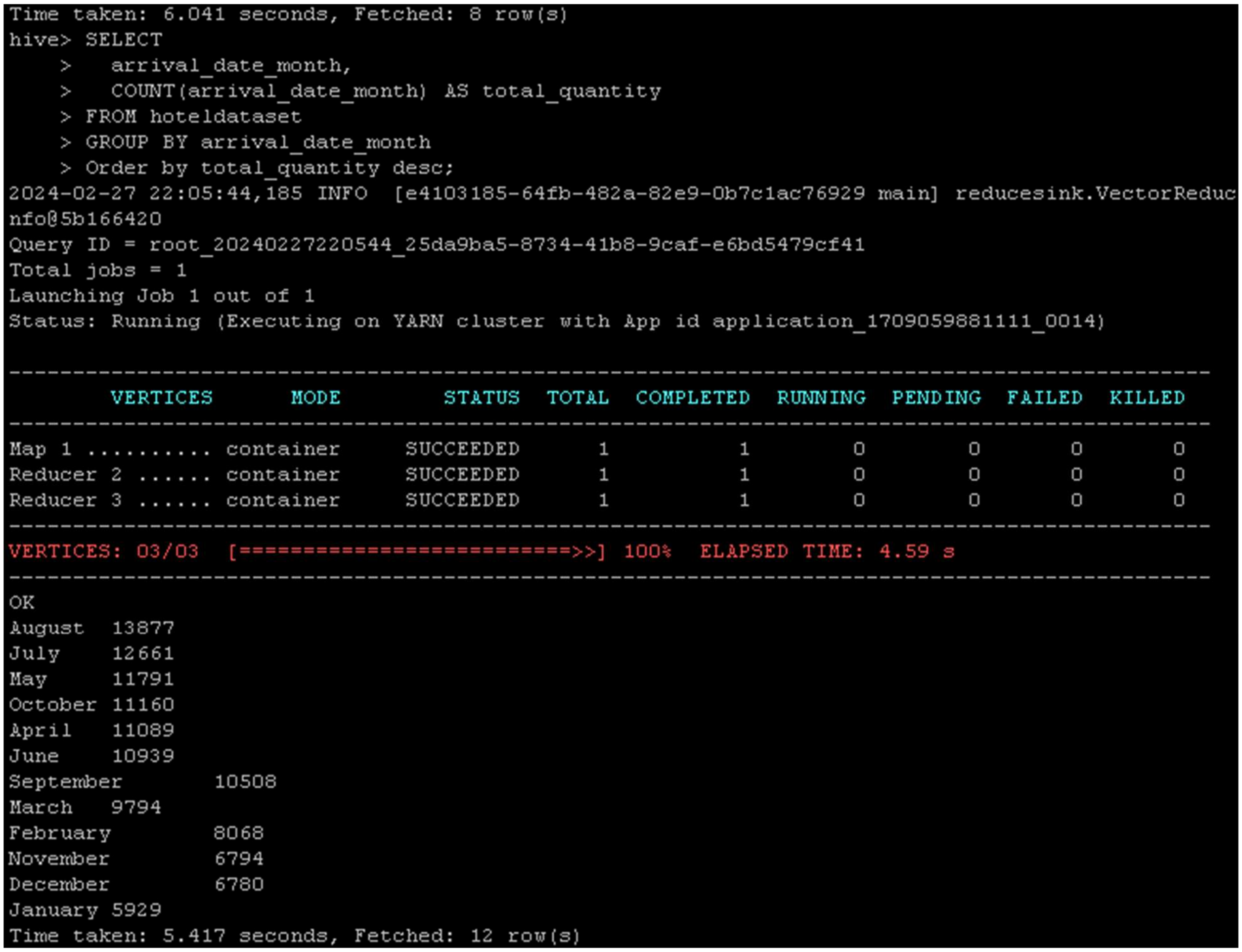
```
hive> SELECT
    >   agent,
    >   COUNT(agent) AS total_quantity
    > FROM hoteldataset
    > GROUP BY agent
    > Order by total_quantity desc;
2024-02-27 21:58:48,798 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduceSinkObjectH
nfo@44ec6637
Query ID = root_20240227215848_80910098-86b0-45bf-af08-46e7f2ce7451
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709059881111_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.49 s
----------------------------------------------------------------------------------------------
OK
9       31961
NULL    16340
240     13922
1       7191
14      3640
7       3539
6       3290
250     2870
241     1721
28      1666
8       1514
3       1336
37      1230
19      1061
40      1039
314     927
21      875
229     786
242     780
83      696
29      683
171     607
```

Figure 12



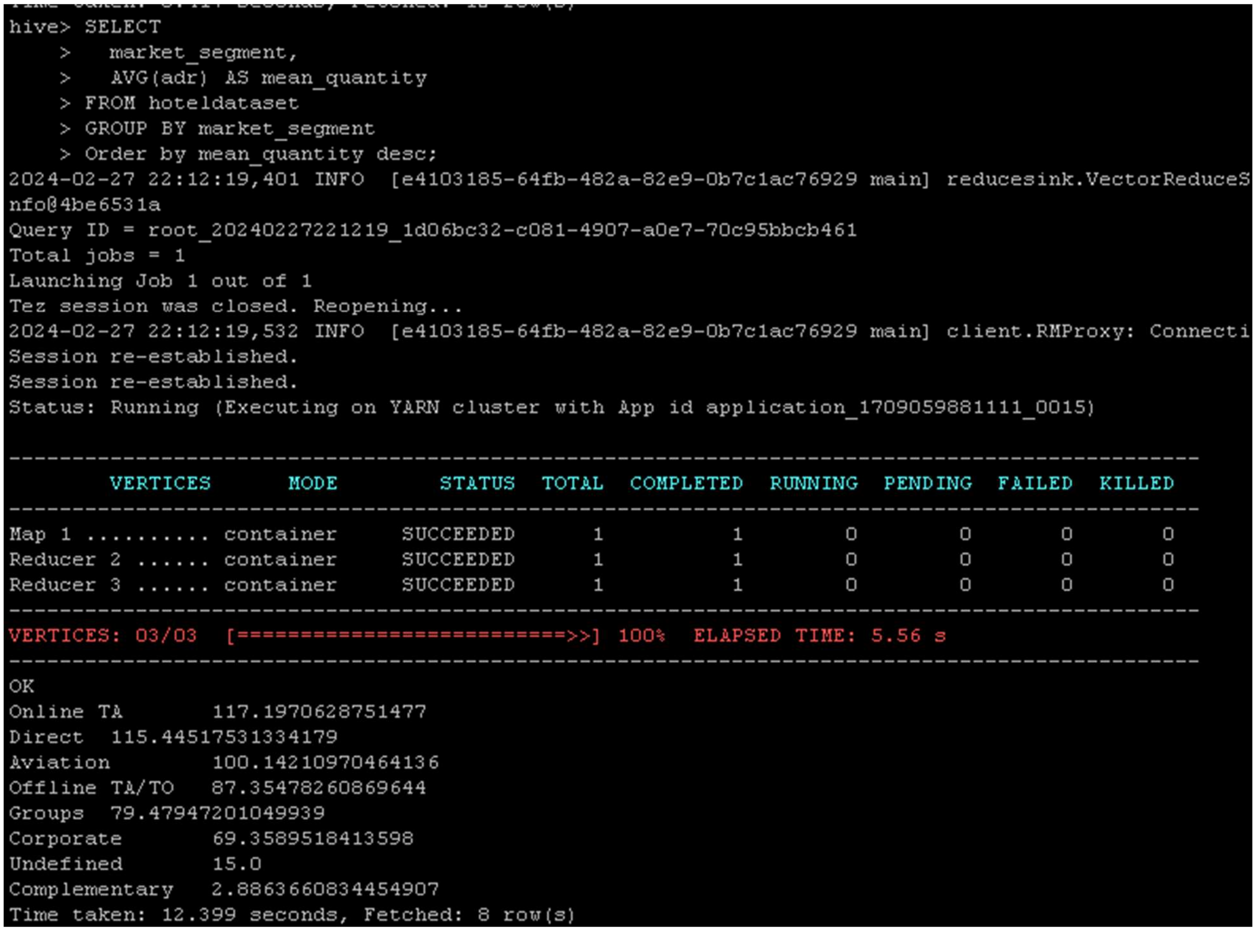
```
hive> SELECT
    >   Market_segment,
    >   COUNT(market_segment) AS total_quantity
    > FROM hoteldataset
    > GROUP BY market_segment
    > Order by total_quantity desc;
2024-02-27 22:02:02,655 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduceSinkOb
nfo@6b23897a
Query ID = root_20240227220202_2671fb6d-c6dd-445b-b8f3-35dae304a5b7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709059881111_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.15 s
----------------------------------------------------------------------------------------------
OK
Online TA       56477
Offline TA/TO   24219
Groups  19811
Direct  12606
Corporate       5295
Complementary   743
Aviation        237
Undefined       2
Time taken: 6.041 seconds, Fetched: 8 row(s)
hive> 
```

Figure 13



```
Time taken: 6.041 seconds, Fetched: 8 row(s)
hive> SELECT
    >   arrival_date_month,
    >   COUNT(arrival_date_month) AS total_quantity
    > FROM hoteldataset
    > GROUP BY arrival_date_month
    > Order by total_quantity desc;
2024-02-27 22:05:44,185 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduc
nfo@5b166420
Query ID = root_20240227220544_25da9ba5-8734-41b8-9caf-e6bd5479cf41
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1709059881111_0014)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.59 s
----------------------------------------------------------------------------------------------
OK
August  13877
July    12661
May     11791
October 11160
April   11089
June    10939
September       10508
March   9794
February        8068
November        6794
December        6780
January 5929
Time taken: 5.417 seconds, Fetched: 12 row(s)
```

Figure 14



```
hive> SELECT
    >   market_segment,
    >   AVG(adr) AS mean_quantity
    > FROM hoteldataset
    > GROUP BY market_segment
    > Order by mean_quantity desc;
2024-02-27 22:12:19,401 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] reducesink.VectorReduceS
nfo@4be6531a
Query ID = root_20240227221219_1d06bc32-c081-4907-a0e7-70c95bbcb461
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-02-27 22:12:19,532 INFO  [e4103185-64fb-482a-82e9-0b7c1ac76929 main] client.RMProxy: Connecti
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1709059881111_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.56 s
----------------------------------------------------------------------------------------------
OK
Online TA       117.1970628751477
Direct  115.44517531334179
Aviation        100.14210970464136
Offline TA/TO   87.35478260869644
Groups  79.47947201049939
Corporate       69.3589518413598
Undefined       15.0
Complementary   2.8863660834454907
Time taken: 12.399 seconds, Fetched: 8 row(s)
```