Ecommerce Assignment (Hive)

Steps

- Uploaded data to S3

- Created EMR cluster and ssh using CLI command

```
ssh -i ec2_alpa2.pem hadoop@ec2-100-25-200-54.compute-1.amazonaws.com
```

- Moved data from S3 to HDFS

```
hadoop fs -mkdir /user/hadoop/assignment/
hadoop fs -mkdir /user/hadoop/assignment/ecommerce
hadoop fs -ls /user/hadoop/assignment/ecommerce/
hadoop distcp s3a://alpaupgrad1/2019-Nov.csv /user/hadoop/assignment/ecommerce/2019-Nov.csv
hadoop distcp s3a://alpaupgrad1/2019-Oct.csv /user/hadoop/assignment/ecommerce/2019-Oct.csv
```

Steps (Continued)

Move data to hive

- Create database and use it using commands

```
create database if not exists assignment;
use assignment;
```

- Created external table with all string values for now

```
create external table if not exists product_data_external_table (event_time string, event_type string, product_id string, category_id string,category_code string,brand string,price string,user_id string,user_session string) row format delimited fields terminated by ',' lines terminated by '\n' tblproperties("skip.header.line.count"="1");
```

- Alter table to skip CSV headers

```
ALTER TABLE tablename SET ecommerce_events ("skip.header.line.count"="1");
```

- Load data from HDFS CSV to hive tables

```
load data inpath '/user/hadoop/assignment/ecommerce/2019-Nov.csv' into table product_data_external_table;
load data inpath '/user/hadoop/assignment/ecommerce/2019-Nov.csv' into table product_data_external_table
```

Steps (continued)

- Create product table

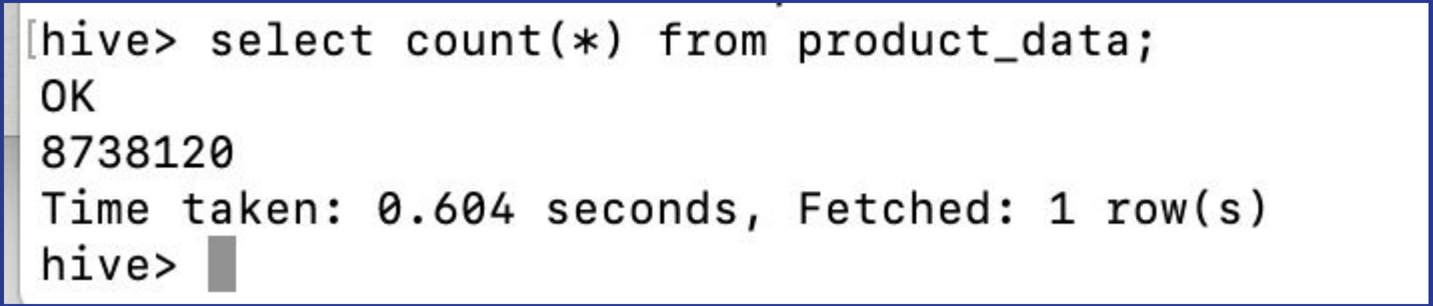
```
create table if not exists product_data(event_time timestamp, event_type string, product_id string,  category_id string, category_code string, brand string, price float, user_id bigint, user_session string) stored as parquet;
```

- Insert data from external table to product_data table

```
insert into product_data select cast(from_unixtime(unix_timestamp(event_time,'yyyy-MM-dd HH:mm:ss Z'),'yyyy-MM-dd HH:mm:ss') as timestamp) as event_time, event_type, product_id, category_id, category_code, brand, cast(price as float) as price, cast(user_id as bigint) as user_id,user_session from product_data_external_table;
```

- I created another table with partition on event_type and buckets on category id to compare performance using normal table and partitioned table.

Screenshot to check all data was uploaded correctly to table



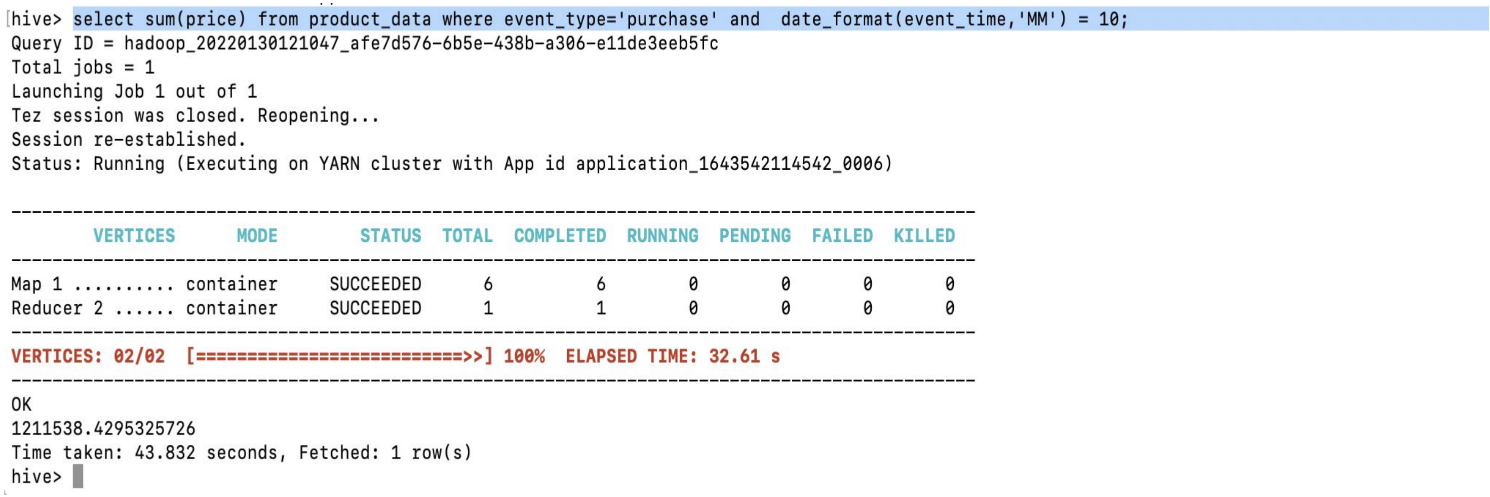
```
hive> select count(*) from product_data;
OK
8738120
Time taken: 0.604 seconds, Fetched: 1 row(s)
hive>
```

Find the total revenue generated due to the purchases made in October.

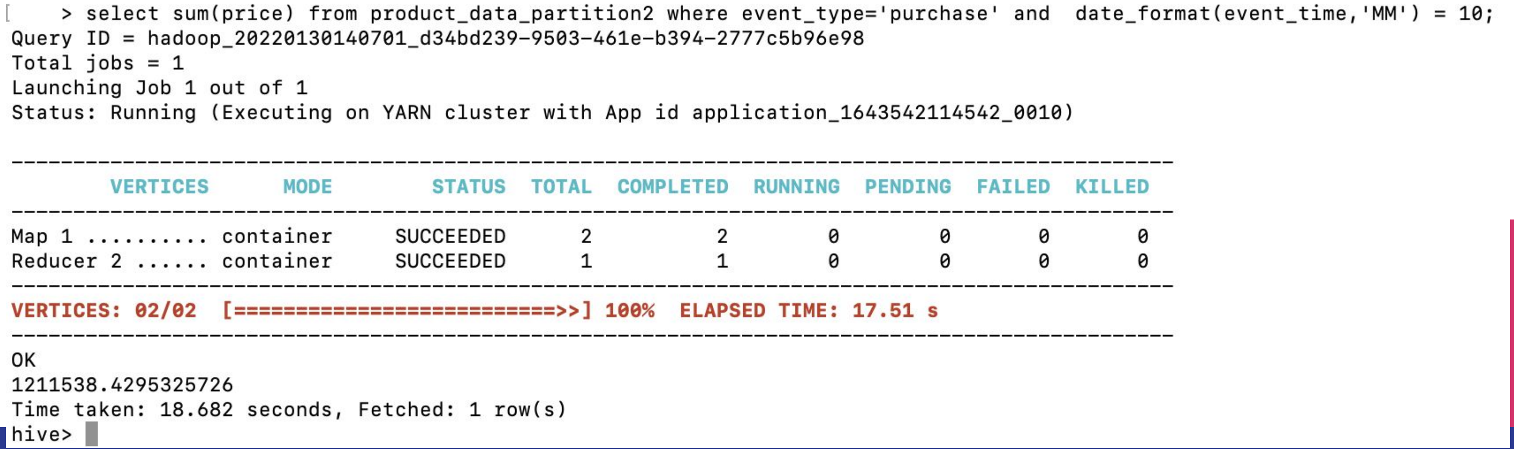
Query -select sum(price) from product_data where event_type='purchase' and date_format(event_time,’MM’) = 10

	Query Result	Time Taken
Regular table	1211538.4295325726	43.832 seconds
Partitioned table	1211538.4295325726	18.682 seconds

Screenshots



```
hive> select sum(price) from product_data where event_type='purchase' and  date_format(event_time,'MM') = 10;
Query ID = hadoop_20220130121047_afe7d576-6b5e-438b-a306-e11de3eeb5fc
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1643542114542_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 32.61 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 43.832 seconds, Fetched: 1 row(s)
hive>
```



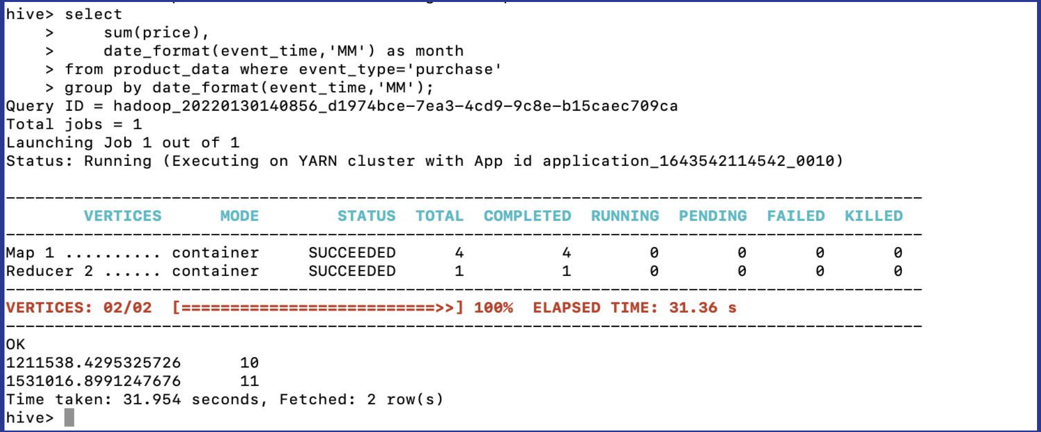
```
    > select sum(price) from product_data_partition2 where event_type='purchase' and  date_format(event_time,'MM') = 10;
Query ID = hadoop_20220130140701_d34bd239-9503-461e-b394-2777c5b96e98
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 17.51 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726
Time taken: 18.682 seconds, Fetched: 1 row(s)
hive>
```

Write a query to yield the total sum of purchases per month in a single output.

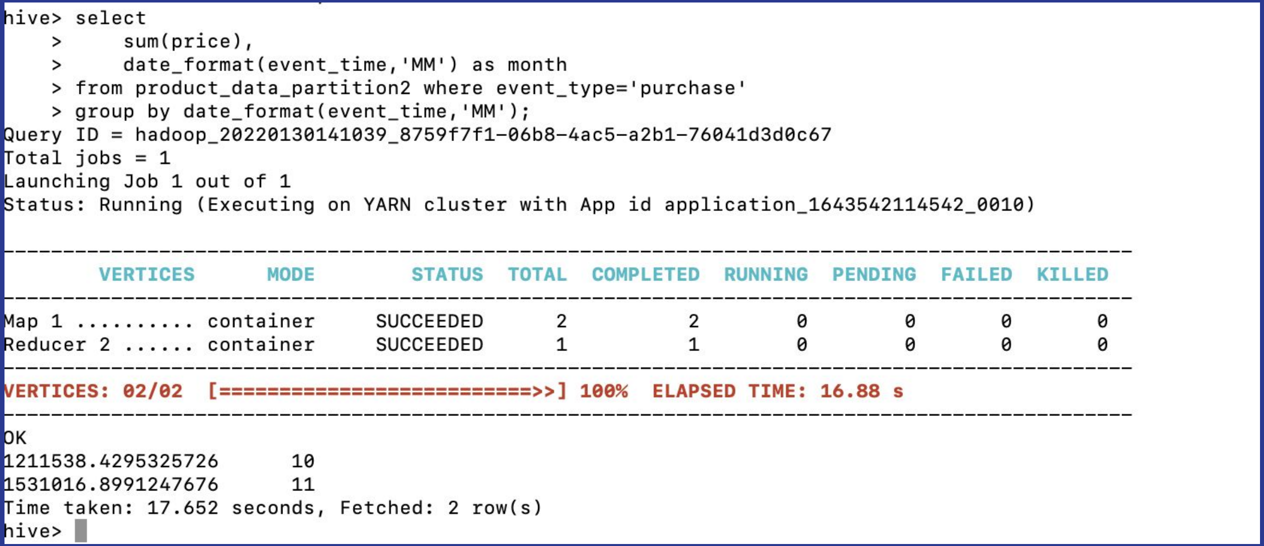
Query - select sum(price), date_format(event_time,'MM') as month from product_data where event_type='purchase' group by date_format(event_time,'MM');

	Query Result	Time Taken
Regular table	1211538.4295325726 10 1531016.8991247676 11	31.954 seconds

Partitioned table	1531016.8991247676 11
	1211538.4295325726 10
	1531016.8991247676 11
		17.652 seconds
		30.256 seconds



```
hive> select
    >     sum(price),
    >     date_format(event_time,'MM') as month
    > from product_data where event_type='purchase'
    > group by date_format(event_time,'MM');
Query ID = hadoop_20220130140856_d1974bce-7ea3-4cd9-9c8e-b15caec709ca
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 31.36 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726      10
1531016.8991247676      11
Time taken: 31.954 seconds, Fetched: 2 row(s)
hive>
```

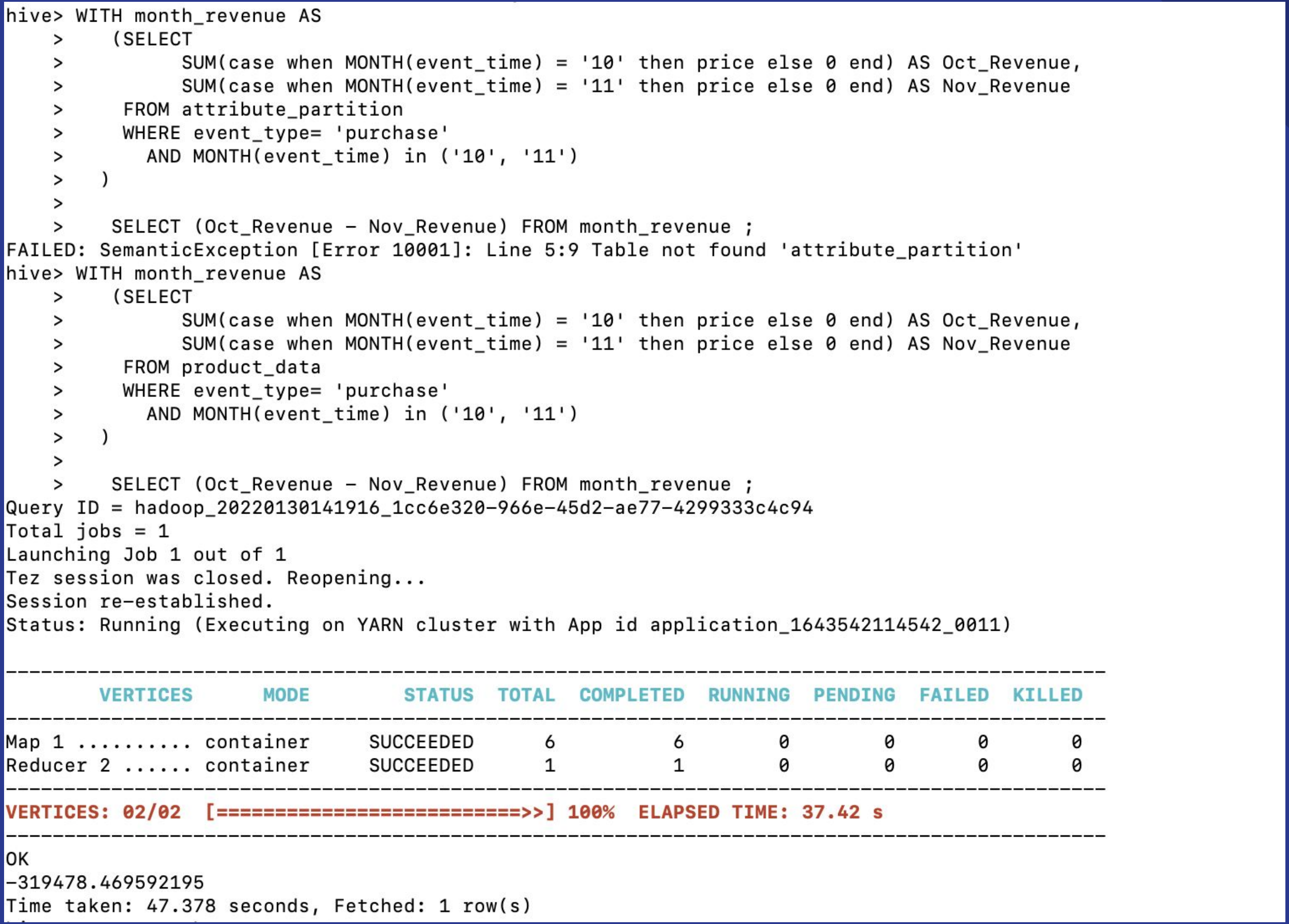


```
hive> select
    >     sum(price),
    >     date_format(event_time,'MM') as month
    > from product_data_partition2 where event_type='purchase'
    > group by date_format(event_time,'MM');
Query ID = hadoop_20220130141039_8759f7f1-06b8-4ac5-a2b1-76041d3d0c67
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 16.88 s
----------------------------------------------------------------------------------------------
OK
1211538.4295325726      10
1531016.8991247676      11
Time taken: 17.652 seconds, Fetched: 2 row(s)
hive>
```

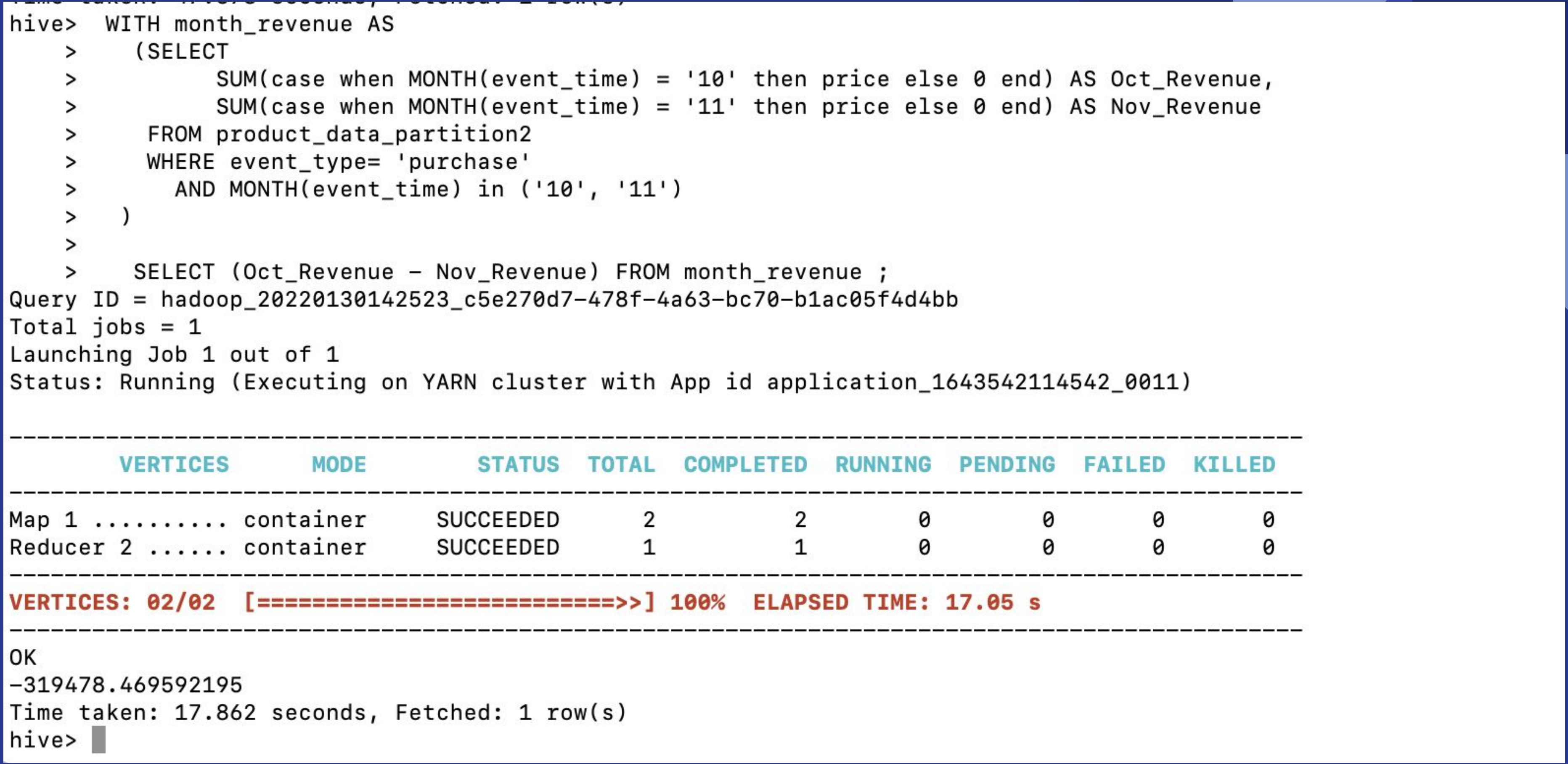
Write a query to find the change in the revenue generated due to purchases made from October to November.

```
WITH month_revenue AS (SELECT SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_Revenue,

SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_Revenue FROM product_data WHERE event_type= 'purchase' AND

MONTH(event_time) in ('10', '11')

SELECT (Oct_Revenue - Nov_Revenue) FROM month_revenue ;
```

	Query Result	**Time Taken**
Regular table	-319478.469592195	47.378 seconds
Partitioned table	-319478.469592195	17.862 seconds



```
hive> WITH month_revenue AS
    >    (SELECT
    >          SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_Revenue,
    >          SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_Revenue
    >     FROM attribute_partition
    >     WHERE event_type= 'purchase'
    >       AND MONTH(event_time) in ('10', '11')
    >   )
    >
    >    SELECT (Oct_Revenue - Nov_Revenue) FROM month_revenue ;
FAILED: SemanticException [Error 10001]: Line 5:9 Table not found 'attribute_partition'
hive> WITH month_revenue AS
    >    (SELECT
    >          SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_Revenue,
    >          SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_Revenue
    >     FROM product_data
    >     WHERE event_type= 'purchase'
    >       AND MONTH(event_time) in ('10', '11')
    >   )
    >
    >    SELECT (Oct_Revenue - Nov_Revenue) FROM month_revenue ;
Query ID = hadoop_20220130141916_1cc6e320-966e-45d2-ae77-4299333c4c94
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1643542114542_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 37.42 s
----------------------------------------------------------------------------------------------
OK
-319478.469592195
Time taken: 47.378 seconds, Fetched: 1 row(s)
```



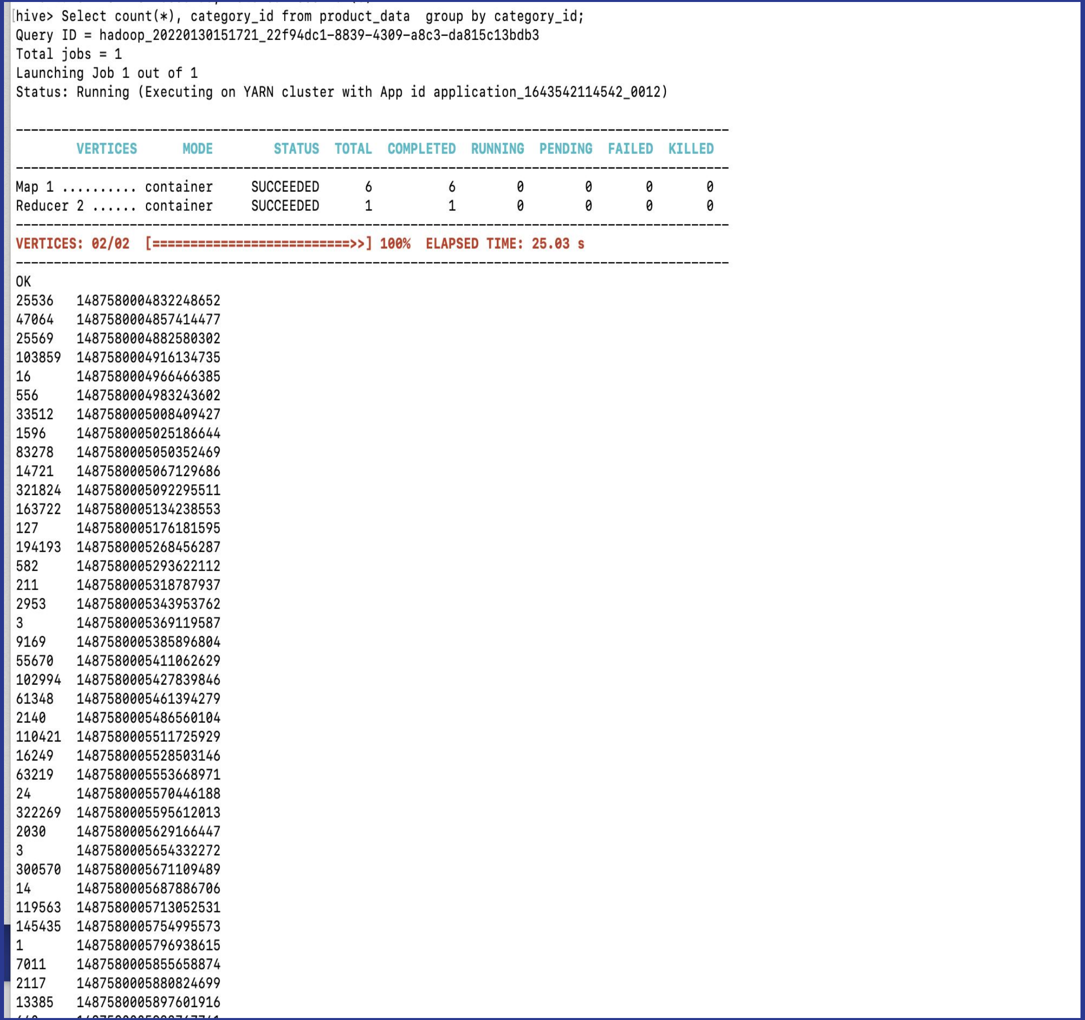
```
hive>  WITH month_revenue AS
    >    (SELECT
    >          SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_Revenue,
    >          SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_Revenue
    >     FROM product_data_partition2
    >     WHERE event_type= 'purchase'
    >       AND MONTH(event_time) in ('10', '11')
    >   )
    >
    >    SELECT (Oct_Revenue - Nov_Revenue) FROM month_revenue ;
Query ID = hadoop_20220130142523_c5e270d7-478f-4a63-bc70-b1ac05f4d4bb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 17.05 s
----------------------------------------------------------------------------------------------
OK
-319478.469592195
Time taken: 17.862 seconds, Fetched: 1 row(s)
hive>
```

- Find distinct categories of products.
- Query - select distinct category_id from product_data;

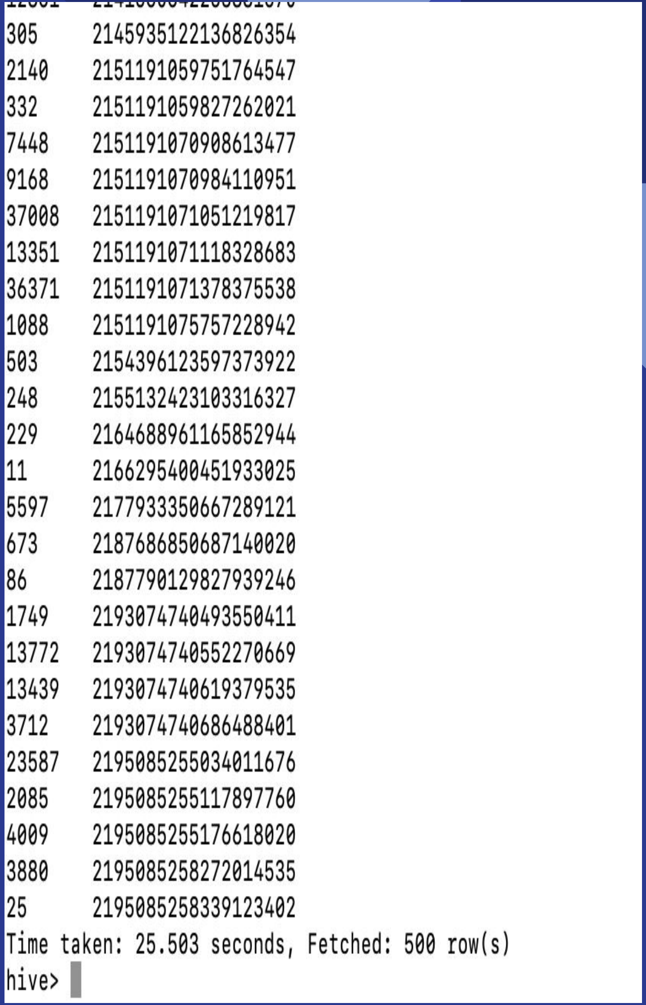
	Query Result	**Time Taken**
Regular table	500 category ids	26.124 seconds
Partitioned table	500 category ids	34.971 seconds

- Find the total number of products available under each category.
- Query - Select count(*), category_id from product_data group by category_id

	Query Result	**Time Taken**
Regular table	500 category ids	25.503 second
Partitioned table	500 category ids	35.614 seconds



```
hive> Select count(*), category_id from product_data  group by category_id;
Query ID = hadoop_20220130151721_22f94dc1-8839-4309-a8c3-da815c13bdb3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.03 s
----------------------------------------------------------------------------------------------
OK
25536   1487580004832248652
47064   1487580004857414477
25569   1487580004882580302
103859  1487580004916134735
16      1487580004966466385
556     1487580004983243602
33512   1487580005008409427
1596    1487580005025186644
83278   1487580005050352469
14721   1487580005067129686
321824  1487580005092295511
163722  1487580005134238553
127     1487580005176181595
194193  1487580005268456287
582     1487580005293622112
211     1487580005318787937
2953    1487580005343953762
3       1487580005369119587
9169    1487580005385896804
55670   1487580005411062629
102994  1487580005427839846
61348   1487580005461394279
2140    1487580005486560104
110421  1487580005511725929
16249   1487580005528503146
63219   1487580005553668971
24      1487580005570446188
322269  1487580005595612013
2030    1487580005629166447
3       1487580005654332272
300570  1487580005671109489
14      1487580005687886706
119563  1487580005713052531
145435  1487580005754995573
1       1487580005796938615
7011    1487580005855658874
2117    1487580005880824699
13385   1487580005897601916
```

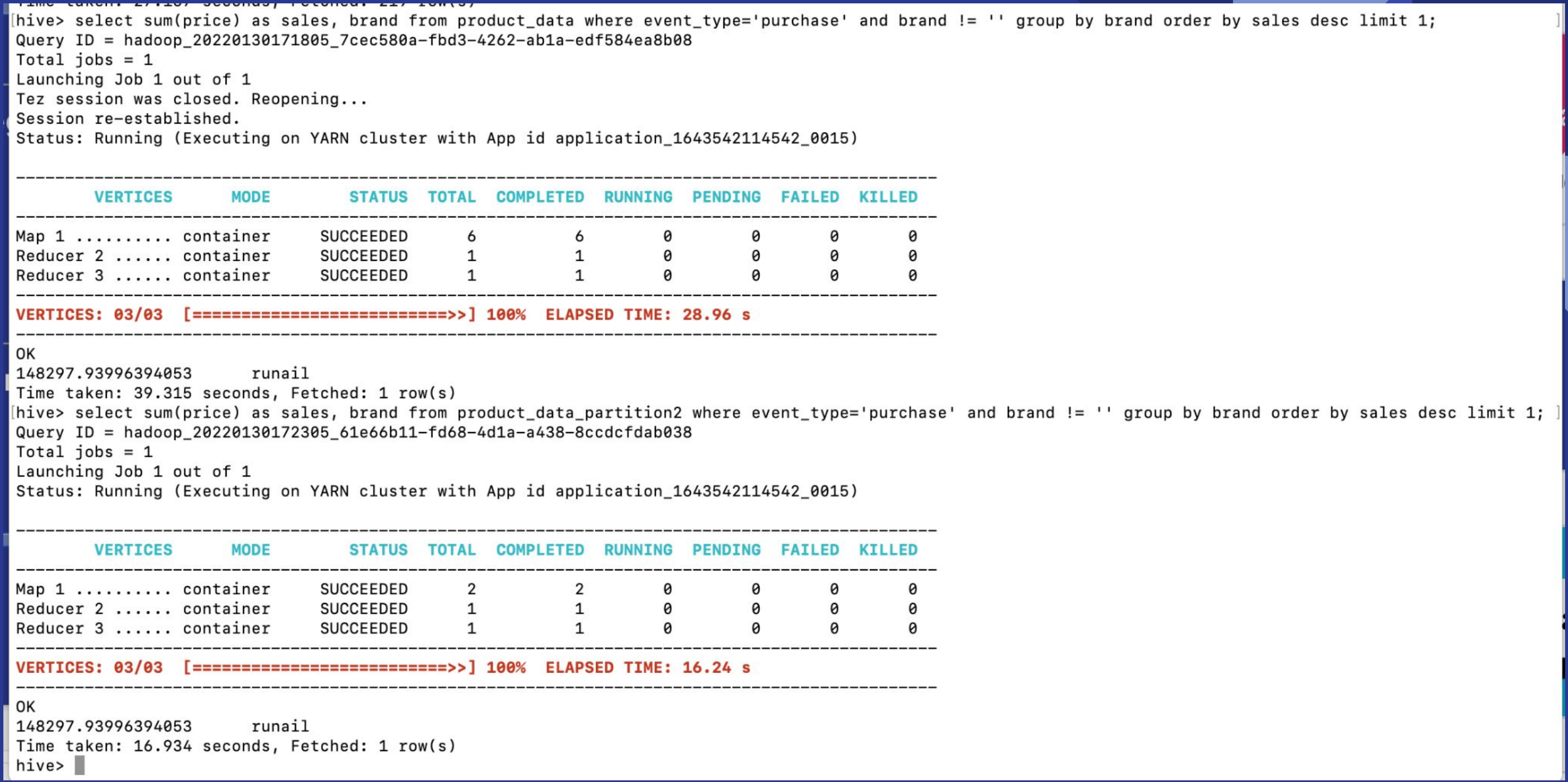


```
305     2145935122136826354
2140    2151191059751764547
332     2151191059827262021
7448    2151191070908613477
9168    2151191070984110951
37008   2151191071051219817
13351   2151191071118328683
36371   2151191071378375538
1088    2151191075757228942
503     2154396123597373922
248     2155132423103316327
229     2164688961165852944
11      2166295400451933025
5597    2177933350667289121
673     2187686850687140020
86      2187790129827939246
1749    2193074740493550411
13772   2193074740552270669
13439   2193074740619379535
3712    2193074740686488401
23587   2195085255034011676
2085    2195085255117897760
4009    2195085255176618020
3880    2195085258272014535
25      2195085258339123402
Time taken: 25.503 seconds, Fetched: 500 row(s)
hive> 
```

- Which brand had the maximum sales in October and November combined?

- Query - select sum(price) as sales, brand from product_data where event_type='purchase' and brand != '' group by brand order by sales desc limit 1

	Query Result	Time Taken
Regular table	148297.93996394053 runail	39.315 seconds
Partitioned table	148297.93996394053 runail	16.934 seconds

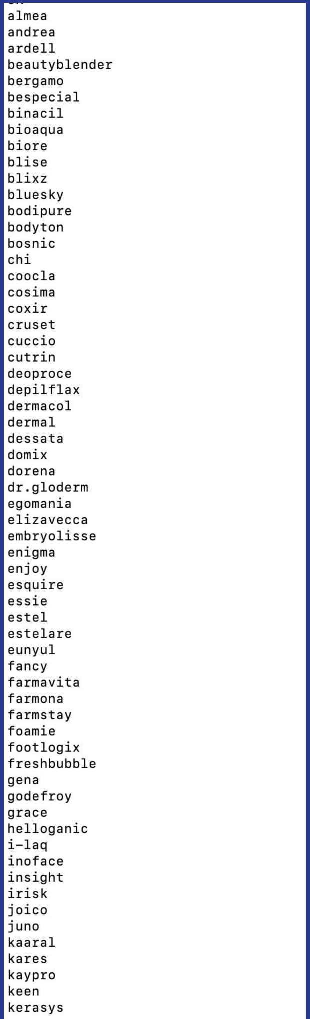


```
Time taken: 27.187 seconds, Fetched: 217 row(s)
hive> select sum(price) as sales, brand from product_data where event_type='purchase' and brand != '' group by brand order by sales desc limit 1;
Query ID = hadoop_20220130171805_7cec580a-fbd3-4262-ab1a-edf584ea8b08
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1643542114542_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        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: 28.96 s
----------------------------------------------------------------------------------------------
OK
148297.93996394053      runail
Time taken: 39.315 seconds, Fetched: 1 row(s)
hive> select sum(price) as sales, brand from product_data_partition2 where event_type='purchase' and brand != '' group by brand order by sales desc limit 1;
Query ID = hadoop_20220130172305_61e66b11-fd68-4d1a-a438-8ccdcfdab038
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0015)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        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: 16.24 s
----------------------------------------------------------------------------------------------
OK
148297.93996394053      runail
Time taken: 16.934 seconds, Fetched: 1 row(s)
hive> 
```

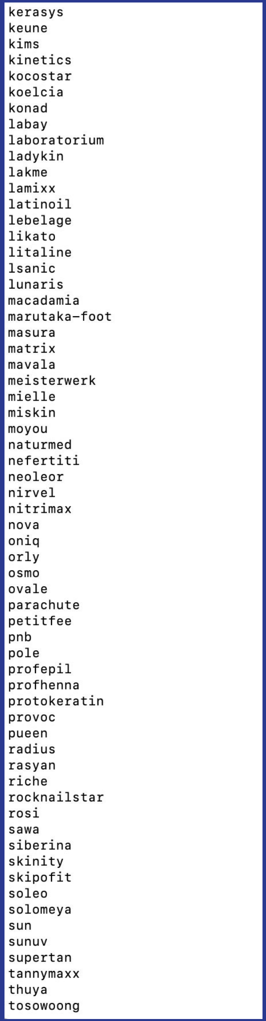
- Which brand had the maximum sales in October and November combined?

- select sum(price) as sales, brand from product_data where event_type='purchase' and brand != ''
  group by brand order by sales desc limit 1

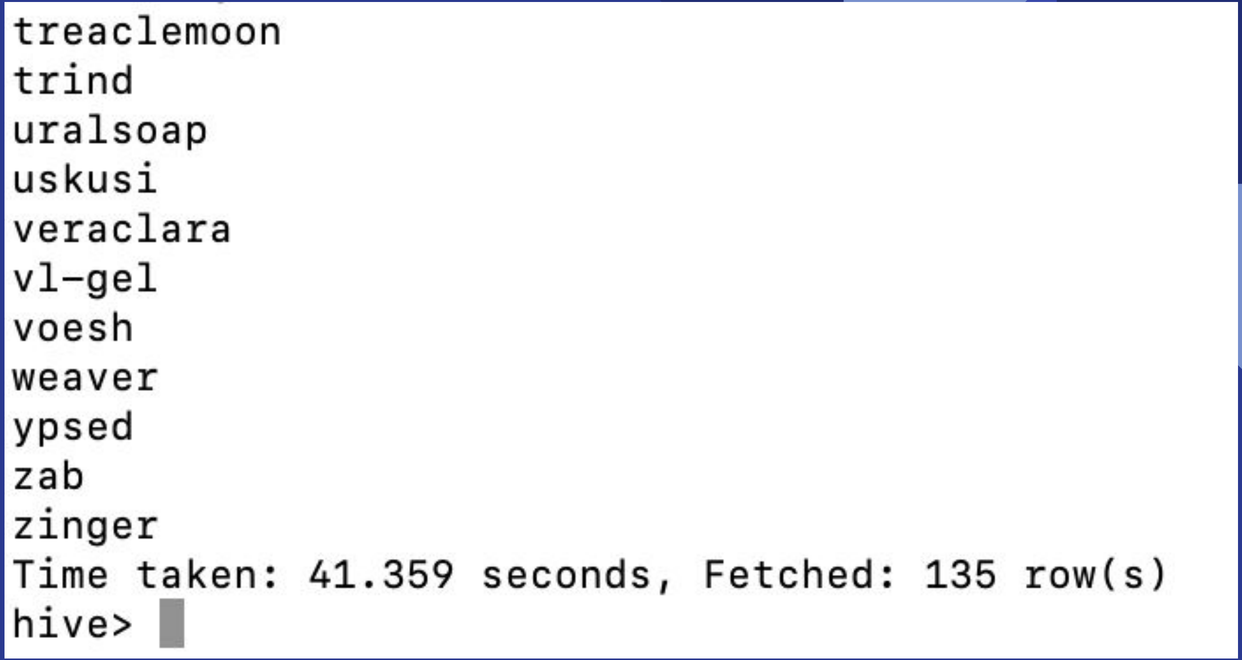
	Query Result	**Time Taken**
Regular table	Results in next slide	41.359 seconds
Partitioned table	Results in next slide	52.989 seconds



```
almea
andrea
ardell
beautyblender
bergamo
bespecial
binacil
bioaqua
biore
blise
blixz
bluesky
bodipure
bodyton
bosnic
chi
coocla
cosima
coxir
cruset
cuccio
cutrin
deoproce
depilflax
dermacol
dermal
dessata
domix
dorena
dr.gloderm
egomania
elizavecca
embryolisse
enigma
enjoy
esquire
essie
estel
estelare
eunyul
fancy
farmavita
farmona
farmstay
foamie
footlogix
freshbubble
gena
godefroy
grace
helloganic
i-laq
inoface
insight
irisk
joico
juno
kaaral
kares
kaypro
keen
kerasys
```



```
kerasys
keune
kims
kinetics
kocostar
koelcia
konad
labay
laboratorium
ladykin
lakme
lamixx
latinoil
lebelage
likato
litaline
lsanic
lunaris
macadamia
marutaka-foot
masura
matrix
mavala
meisterwerk
mielle
miskin
moyou
naturmed
nefertiti
neoleor
nirvel
nitrimax
nova
oniq
orly
osmo
ovale
parachute
petitfee
pnb
pole
profepil
profhenna
protokeratin
provoc
pueen
radius
rasyan
riche
rocknailstar
rosi
sawa
siberina
skinity
skipofit
soleo
solomeya
sun
sunuv
supertan
tannymaxx
thuya
tosowoong
```

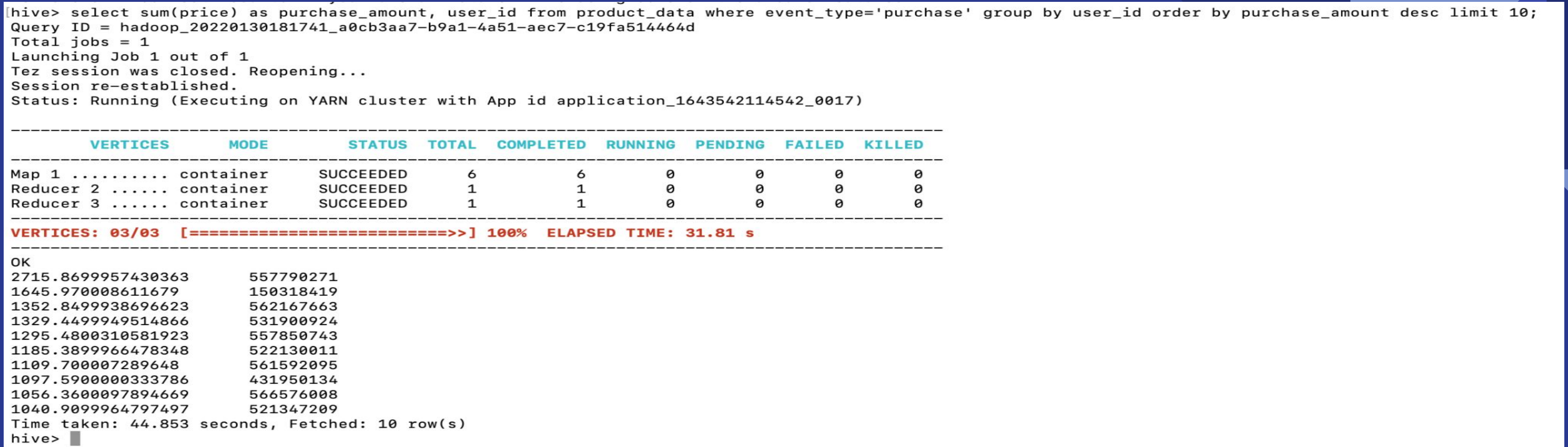


```
treaclemoon
trind
uralsoap
uskusi
veraclara
vl-gel
voesh
weaver
ypsed
zab
zinger
Time taken: 41.359 seconds, Fetched: 135 row(s)
hive>
```

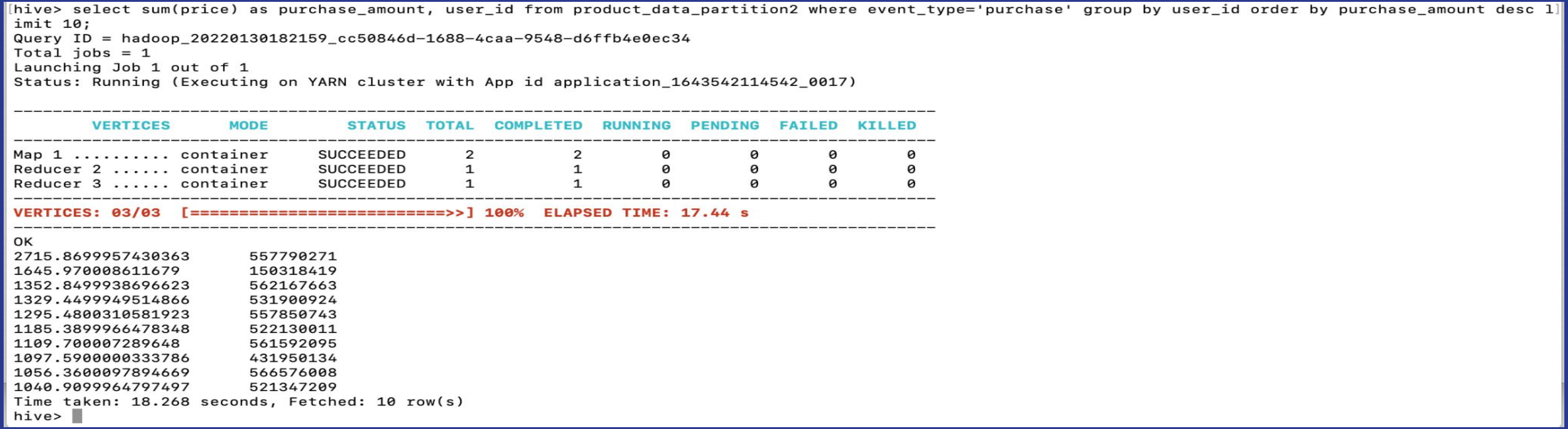
- Your company wants to reward the top 10 users of its website with a Golden Customer plan. Write a query to generate a list of top 10 users who spend the most on purchases.

```
select sum(price) as sales, brand from product_data where event_type='purchase' and brand != ''
group by brand order by sales desc limit 1
```

	Query Result	Time Taken
Regular table	Results in next slide	44.853 seconds
Partitioned table	Results in next slide	18.268 seconds



```
hive> select sum(price) as purchase_amount, user_id from product_data where event_type='purchase' group by user_id order by purchase_amount desc limit 10;
Query ID = hadoop_20220130181741_a0cb3aa7-b9a1-4a51-aec7-c19fa514464d
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1643542114542_0017)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        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: 31.81 s
----------------------------------------------------------------------------------------------
OK
2715.8699957430363      557790271
1645.970008611679       150318419
1352.8499938696623      562167663
1329.4499949514866      531900924
1295.4800310581923      557850743
1185.3899966478348      522130011
1109.700007289648       561592095
1097.5900000333786      431950134
1056.3600097894669      566576008
1040.9099964797497      521347209
Time taken: 44.853 seconds, Fetched: 10 row(s)
hive> 
```



```
hive> select sum(price) as purchase_amount, user_id from product_data_partition2 where event_type='purchase' group by user_id order by purchase_amount desc l
imit 10;
Query ID = hadoop_20220130182159_cc50846d-1688-4caa-9548-d6ffb4e0ec34
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1643542114542_0017)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        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: 17.44 s
----------------------------------------------------------------------------------------------
OK
2715.8699957430363      557790271
1645.970008611679       150318419
1352.8499938696623      562167663
1329.4499949514866      531900924
1295.4800310581923      557850743
1185.3899966478348      522130011
1109.700007289648       561592095
1097.5900000333786      431950134
1056.3600097894669      566576008
1040.9099964797497      521347209
Time taken: 18.268 seconds, Fetched: 10 row(s)
hive> 
```