HIVE Case Study – BI Track

Submitted by:
Prashant Kapoor
prashantkap1997@gmail.com

Problem Statement

We were asked to work with a public clickstream dataset of a cosmetics store and were given the task of extracting valuable insights from it using HIVE.

Here are the links for our datasets:

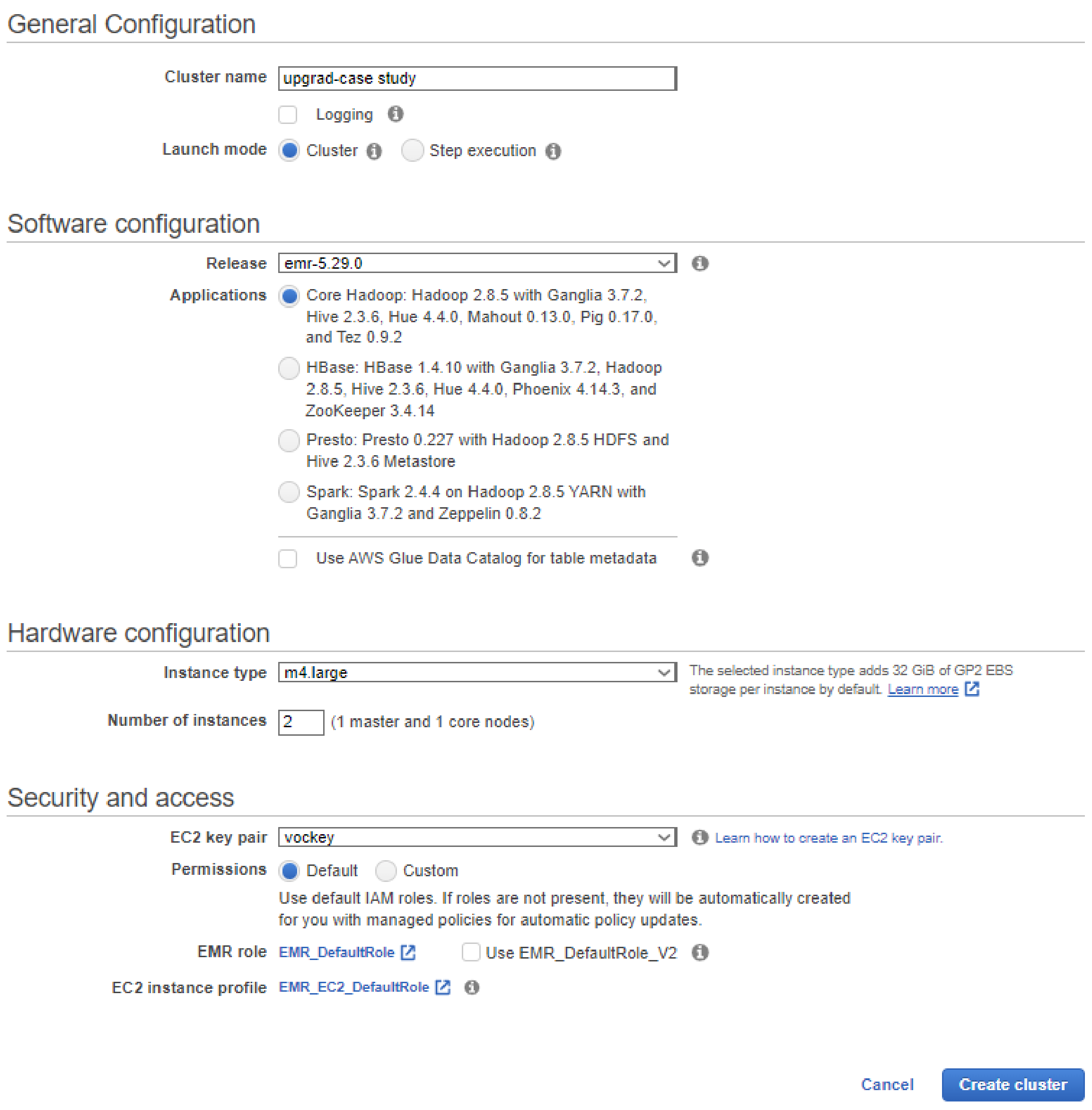
https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv
https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv

Note: In this report, all the steps taken have been outlined along with the screenshots. All the commands in text format are present at the end in the appendix section.

Solution/Implementation

Launching EMR and loading Data on Hive

emr 5.29.0 cluster was launched with 1 master and 1 core nodes of instance m4.large



General Configuration

Cluster name: upgrad-case study
Logging
Launch mode: [✓ Cluster] Step execution

Software configuration

Release: emr-5.29.0
Applications:
[✓ Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2]
HBase: HBase 1.4.10 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.6, Hue 4.4.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
Presto: Presto 0.227 with Hadoop 2.8.5 HDFS and Hive 2.3.6 Metastore
Spark: Spark 2.4.4 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.2
Use AWS Glue Data Catalog for table metadata

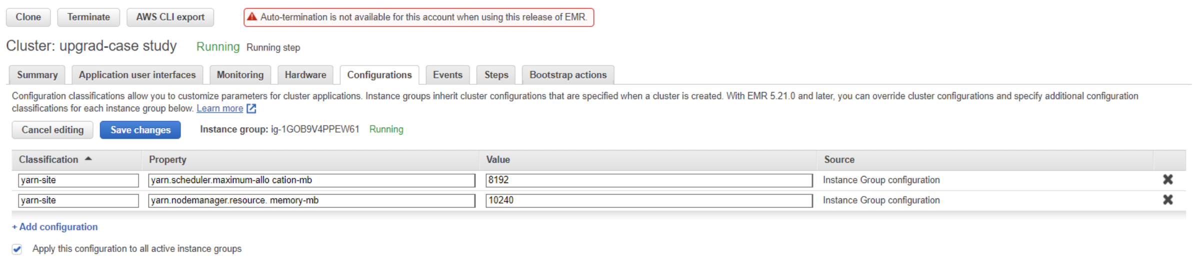
Hardware configuration

Instance type: m4.large — The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. Learn more
Number of instances: 2 (1 master and 1 core nodes)

Security and access

EC2 key pair: vockey — Learn how to create an EC2 key pair.
Permissions: [✓ Default] Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
EMR role: EMR_DefaultRole — Use EMR_DefaultRole_V2
EC2 instance profile: EMR_EC2_DefaultRole

Following that, YARN Configuration was done on the cluster



Auto-termination is not available for this account when using this release of EMR.

Cluster: upgrad-case study Running Running step

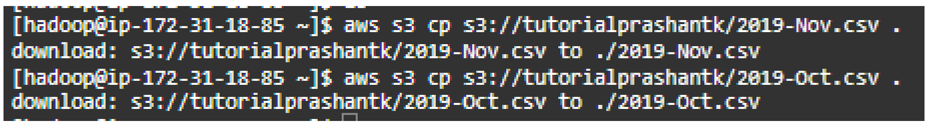
Configuration classifications allow you to customize parameters for cluster applications. Instance groups inherit cluster configurations that are specified when a cluster is created. With EMR 5.21.0 and later, you can override cluster configurations and specify additional configuration classifications for each instance group below. Learn more

Instance group: ig-1GOB9V4PPEW61 Running

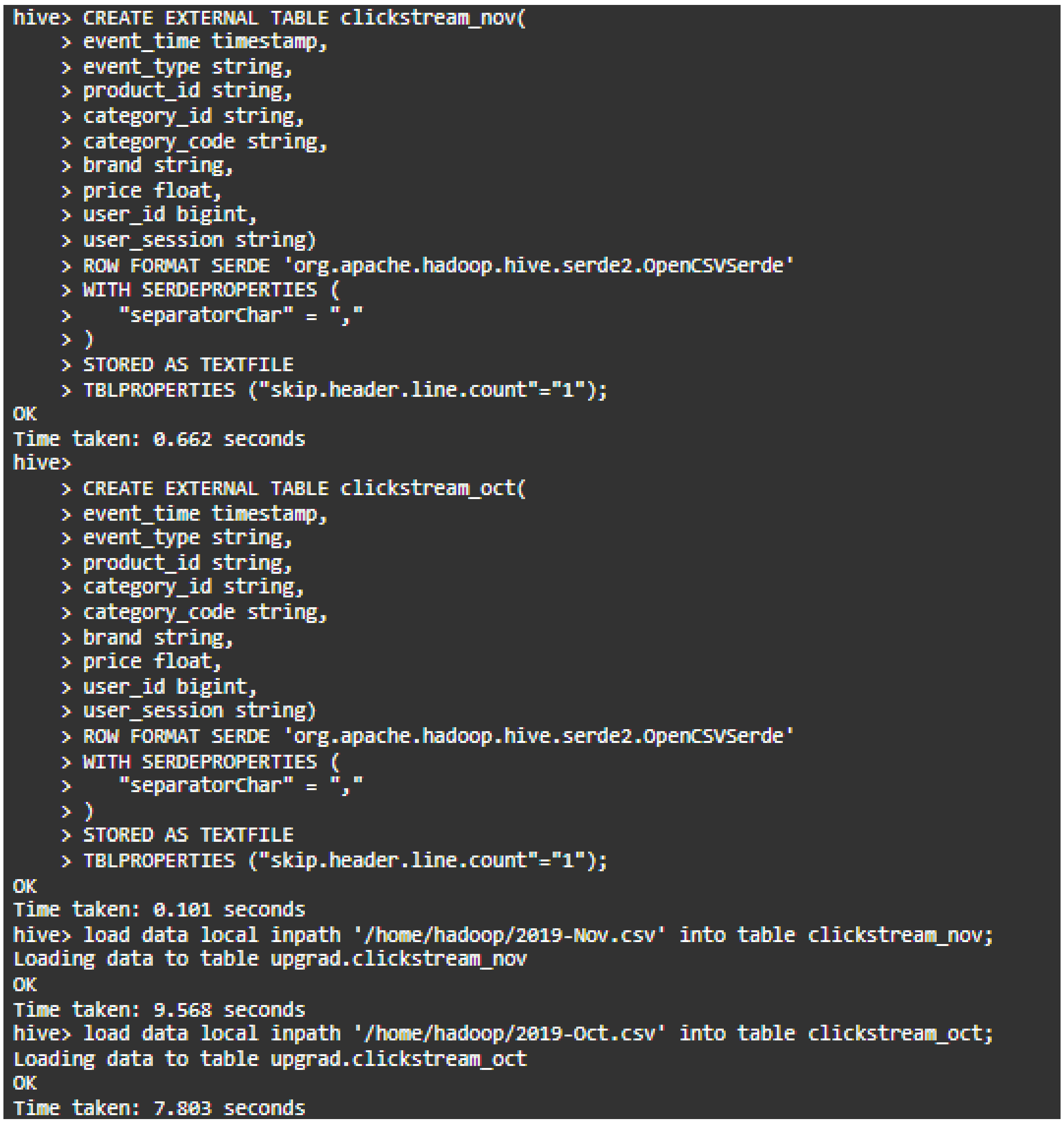
Classification	Property	Value	Source
yarn-site	yarn.scheduler.maximum-allo cation-mb	8192	Instance Group configuration
yarn-site	yarn.nodemanager.resource. memory-mb	10240	Instance Group configuration

✓ Apply this configuration to all active instance groups

Moving data from S3 bucket into the HDFS

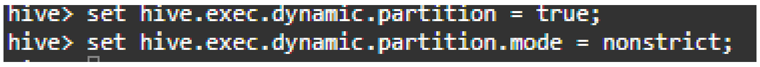

```
[hadoop@ip-172-31-18-85 ~]$ aws s3 cp s3://tutorialprashantk/2019-Nov.csv .
download: s3://tutorialprashantk/2019-Nov.csv to ./2019-Nov.csv
[hadoop@ip-172-31-18-85 ~]$ aws s3 cp s3://tutorialprashantk/2019-Oct.csv .
download: s3://tutorialprashantk/2019-Oct.csv to ./2019-Oct.csv
```

Tables clickstream_nov and clickstream_oct are created using CSVSerde and the data is loaded from CSV files in HDFS


```
hive> CREATE EXTERNAL TABLE clickstream_nov(
    > 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)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    >    "separatorChar" = ","
    > )
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.662 seconds
hive>
    > CREATE EXTERNAL TABLE clickstream_oct(
    > 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)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    >    "separatorChar" = ","
    > )
    > STORED AS TEXTFILE
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.101 seconds
hive> load data local inpath '/home/hadoop/2019-Nov.csv' into table clickstream_nov;
Loading data to table upgrad.clickstream_nov
OK
Time taken: 9.568 seconds
hive> load data local inpath '/home/hadoop/2019-Oct.csv' into table clickstream_oct;
Loading data to table upgrad.clickstream_oct
OK
Time taken: 7.803 seconds
```

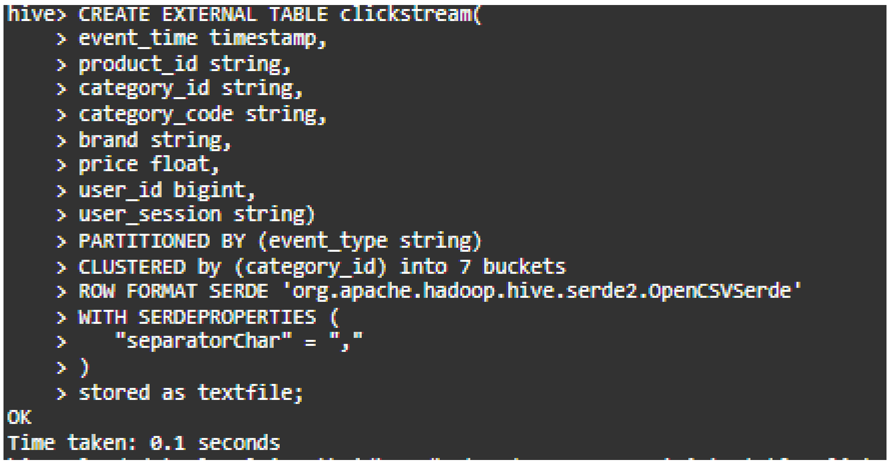
Using optimized techniques to run queries efficiently

Enabling Dynamic Partitioning



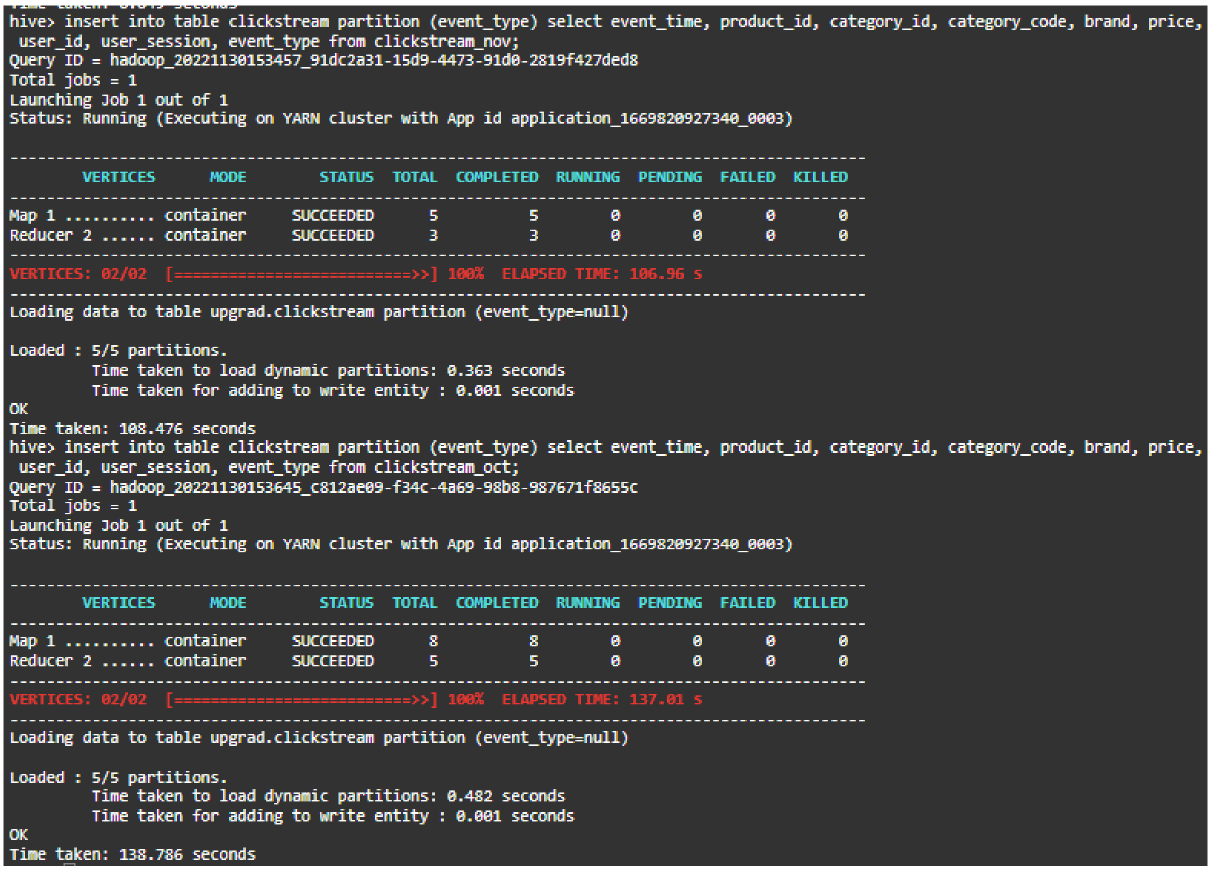
```
hive> set hive.exec.dynamic.partition = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
```

Creating a partitioned and a bucketed table ‘clickstream’ to put all the data together



```
hive> CREATE EXTERNAL TABLE clickstream(
    > event_time timestamp,
    > product_id string,
    > category_id string,
    > category_code string,
    > brand string,
    > price float,
    > user_id bigint,
    > user_session string)
    > PARTITIONED BY (event_type string)
    > CLUSTERED by (category_id) into 7 buckets
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    >    "separatorChar" = ","
    > )
    > stored as textfile;
OK
Time taken: 0.1 seconds
```

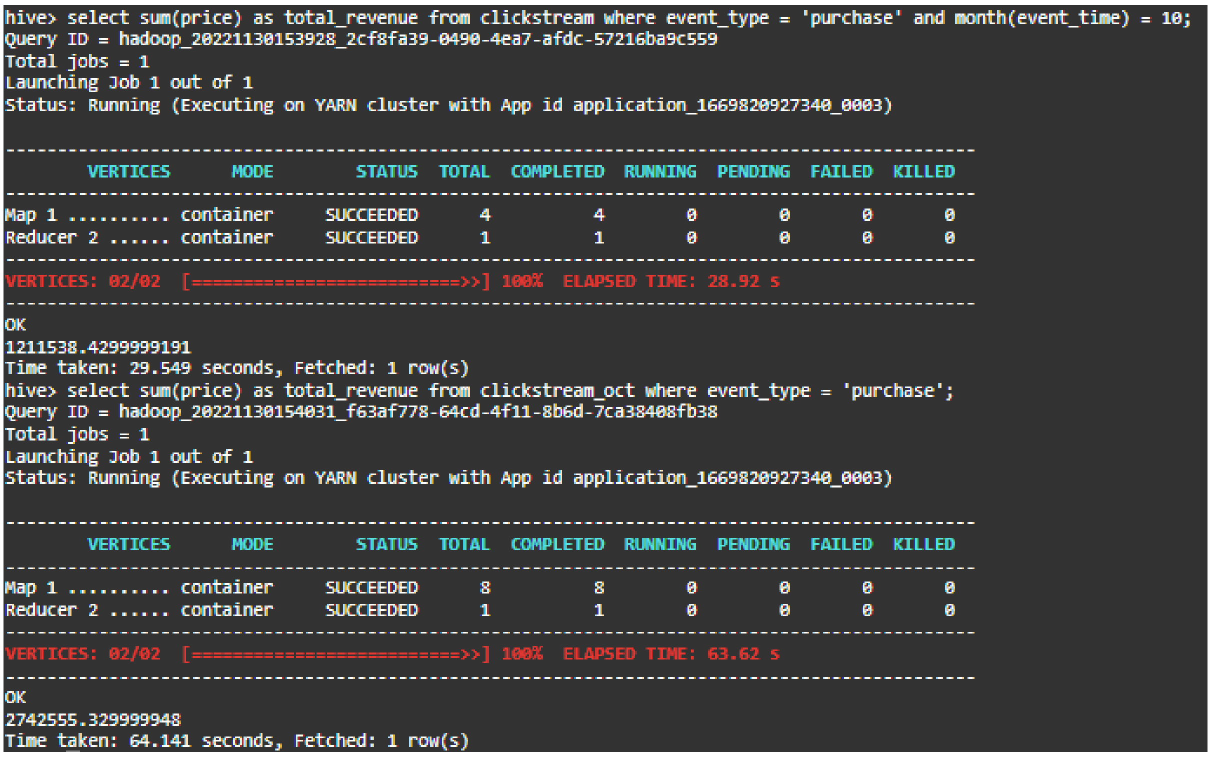
Loading data into ‘clickstream’ table



```
hive> insert into table clickstream partition (event_type) select event_time, product_id, category_id, category_code, brand, price,
 user_id, user_session, event_type from clickstream_nov;
Query ID = hadoop_20221130153457_91dc2a31-15d9-4473-91d0-2819f427ded8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 106.96 s
----------------------------------------------------------------------------------------------
Loading data to table upgrad.clickstream partition (event_type=null)

Loaded : 5/5 partitions.
         Time taken to load dynamic partitions: 0.363 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 108.476 seconds
hive> insert into table clickstream partition (event_type) select event_time, product_id, category_id, category_code, brand, price,
 user_id, user_session, event_type from clickstream_oct;
Query ID = hadoop_20221130153645_c812ae09-f34c-4a69-98b8-987671f8655c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 137.01 s
----------------------------------------------------------------------------------------------
Loading data to table upgrad.clickstream partition (event_type=null)

Loaded : 5/5 partitions.
         Time taken to load dynamic partitions: 0.482 seconds
         Time taken for adding to write entity : 0.001 seconds
OK
Time taken: 138.786 seconds
```

Improvement of performance after using optimization

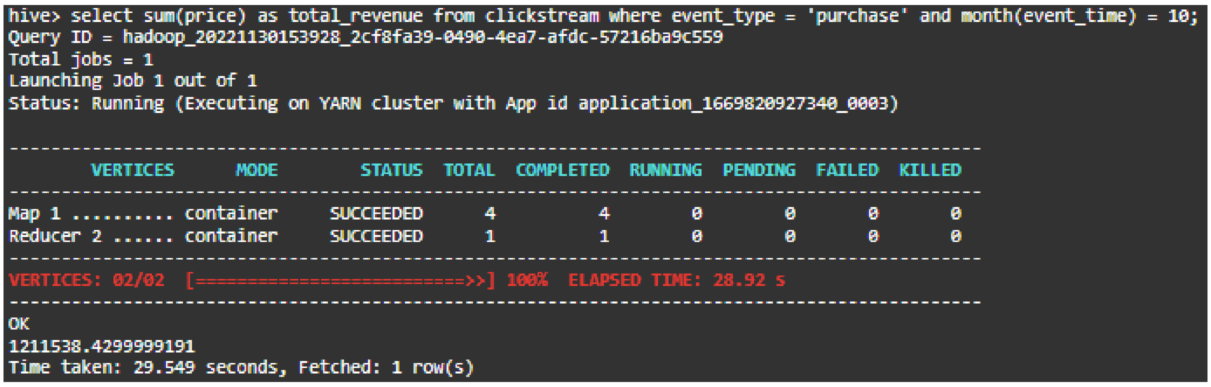
- Query for getting total revenue in the month of October was launched on clickstream, our partitioned and bucketed table, and clickstream_oct, the table we used to load data from csv files.
- Query on the clickstream table took 29.54 seconds whereas query on clickstream_oct table took 64.1 seconds.



```
hive> select sum(price) as total_revenue from clickstream where event_type = 'purchase' and month(event_time) = 10;
Query ID = hadoop_20221130153928_2cf8fa39-0490-4ea7-afdc-57216ba9c559
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0003)

----------------------------------------------------------------------------------------------
        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: 28.92 s
----------------------------------------------------------------------------------------------
OK
1211538.4299999191
Time taken: 29.549 seconds, Fetched: 1 row(s)
hive> select sum(price) as total_revenue from clickstream_oct where event_type = 'purchase';
Query ID = hadoop_20221130154031_f63af778-64cd-4f11-8b6d-7ca38408fb38
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 63.62 s
----------------------------------------------------------------------------------------------
OK
2742555.329999948
Time taken: 64.141 seconds, Fetched: 1 row(s)
```

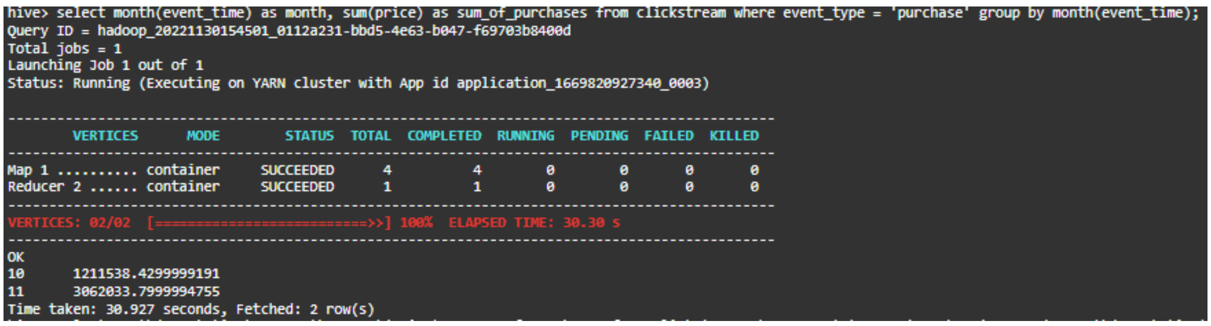
Analysis – Answering Questions using HIVE Queries

Find the total revenue generated due to purchases made in October.
Answer: 12,11,538.43



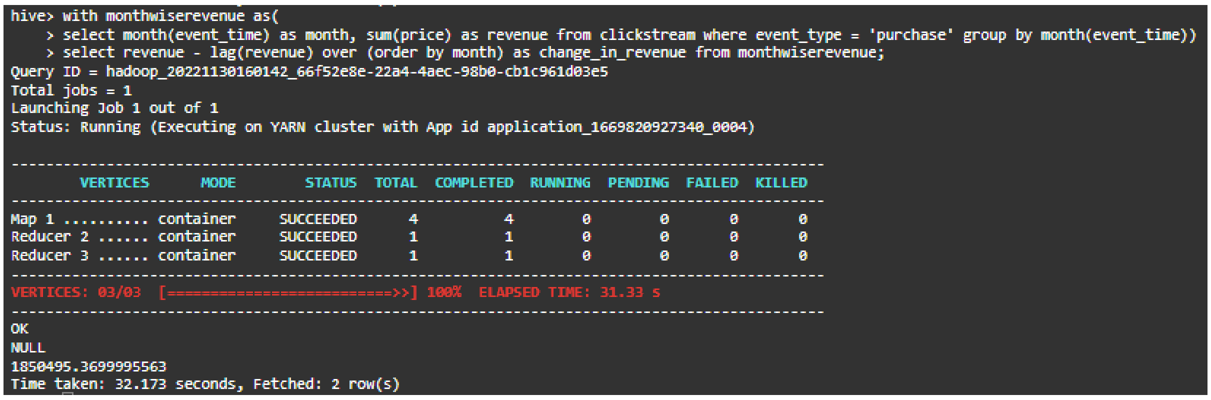
```
hive> select sum(price) as total_revenue from clickstream where event_type = 'purchase' and month(event_time) = 10;
Query ID = hadoop_20221130153928_2cf8fa39-0490-4ea7-afdc-57216ba9c559
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0003)

----------------------------------------------------------------------------------------------
        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: 28.92 s
----------------------------------------------------------------------------------------------
OK
1211538.4299999191
Time taken: 29.549 seconds, Fetched: 1 row(s)
```

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



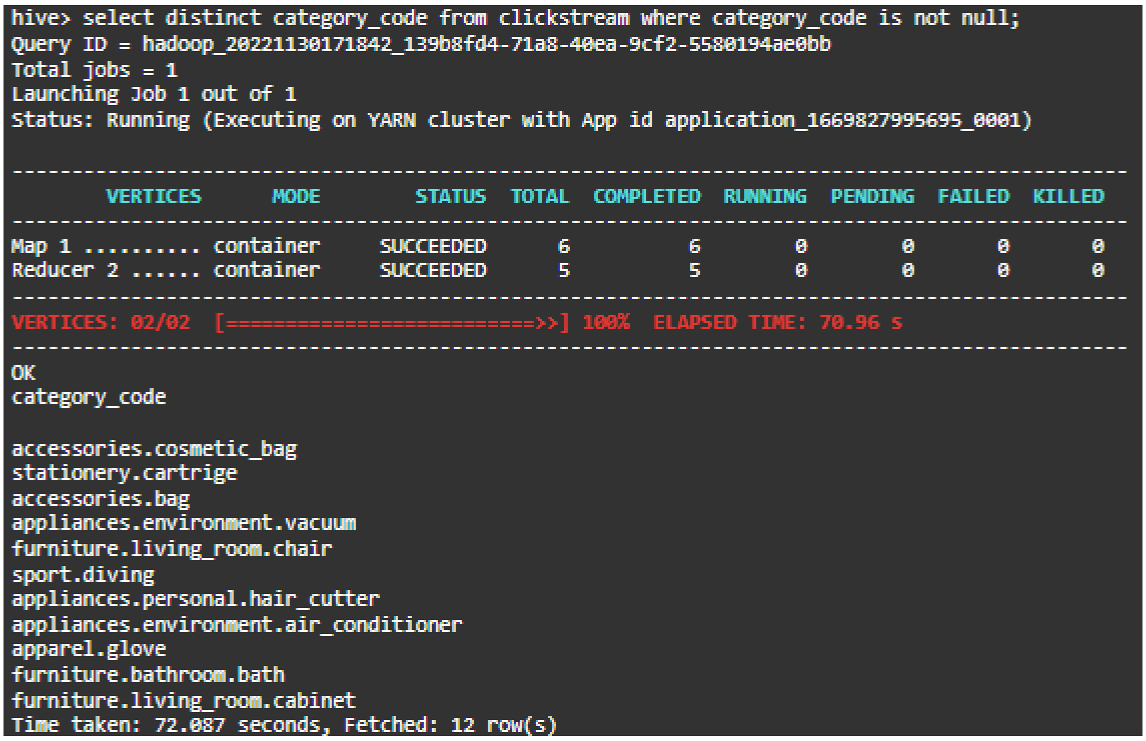
```
hive> select month(event_time) as month, sum(price) as sum_of_purchases from clickstream where event_type = 'purchase' group by month(event_time);
Query ID = hadoop_20221130154501_0112a231-bbd5-4e63-b047-f69703b8400d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0003)

----------------------------------------------------------------------------------------------
        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: 30.30 s
----------------------------------------------------------------------------------------------
OK
10      1211538.4299999191
11      3062033.7999994755
Time taken: 30.927 seconds, Fetched: 2 row(s)
```

Write a query to find the change in revenue generated due to purchases from October to November.
Answer: 1850495.37



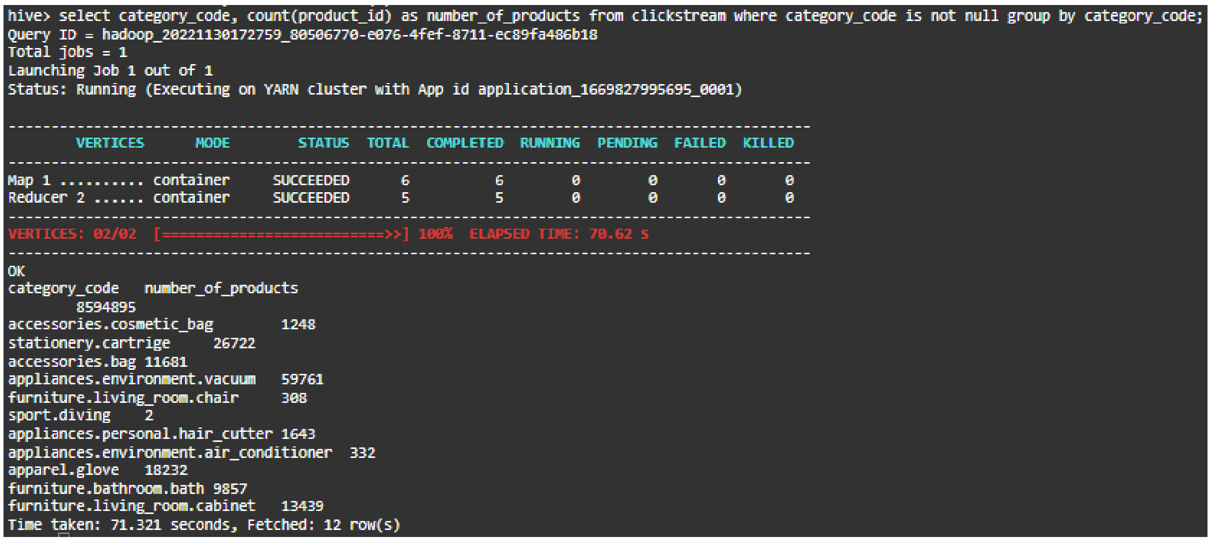
```
hive> with monthwiserevenue as(
    > select month(event_time) as month, sum(price) as revenue from clickstream where event_type = 'purchase' group by month(event_time))
    > select revenue - lag(revenue) over (order by month) as change_in_revenue from monthwiserevenue;
Query ID = hadoop_20221130160142_66f52e8e-22a4-4aec-98b0-cb1c961d03e5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669820927340_0004)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 31.33 s
----------------------------------------------------------------------------------------------
OK
NULL
1850495.3699995563
Time taken: 32.173 seconds, Fetched: 2 row(s)
```

Find distinct categories of products. Categories with null category code can be ignored.



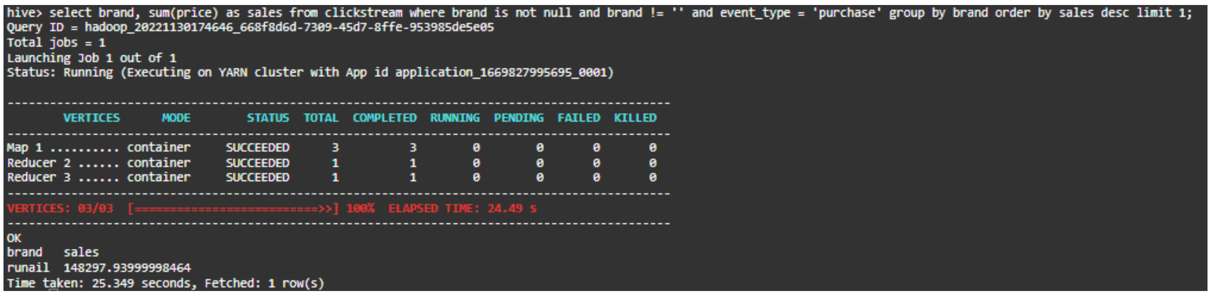
```
hive> select distinct category_code from clickstream where category_code is not null;
Query ID = hadoop_20221130171842_139b8fd4-71a8-40ea-9cf2-5580194ae0bb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669827995695_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.96 s
----------------------------------------------------------------------------------------------
OK
category_code

accessories.cosmetic_bag
stationery.cartrige
accessories.bag
appliances.environment.vacuum
furniture.living_room.chair
sport.diving
appliances.personal.hair_cutter
appliances.environment.air_conditioner
apparel.glove
furniture.bathroom.bath
furniture.living_room.cabinet
Time taken: 72.087 seconds, Fetched: 12 row(s)
```

Find the total number of products available under each category.



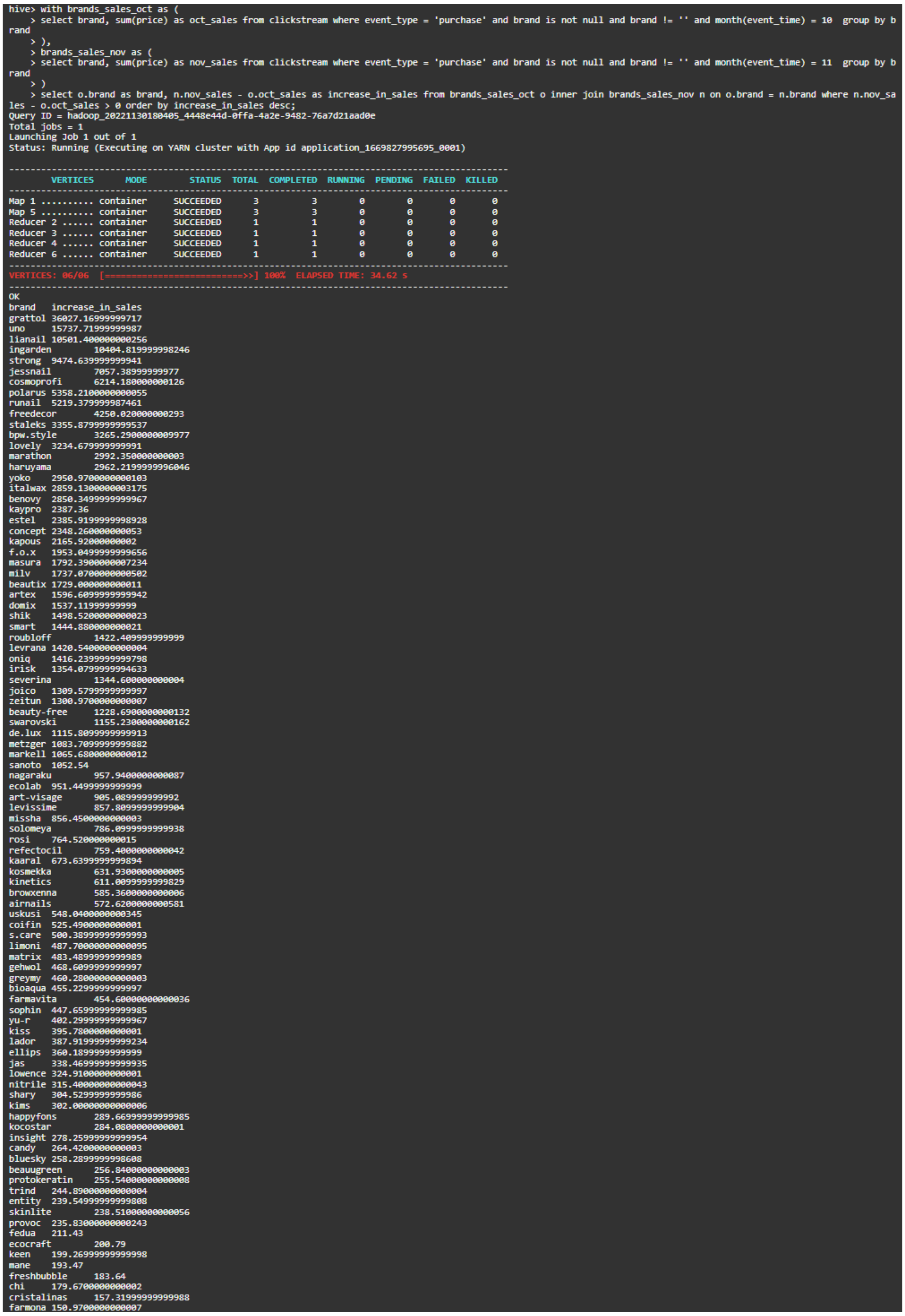
```
hive> select category_code, count(product_id) as number_of_products from clickstream where category_code is not null group by category_code;
Query ID = hadoop_20221130172759_80506770-e076-4fef-8711-ec89fa486b18
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669827995695_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.62 s
----------------------------------------------------------------------------------------------
OK
category_code   number_of_products
        8594895
accessories.cosmetic_bag        1248
stationery.cartrige     26722
accessories.bag 11681
appliances.environment.vacuum   59761
furniture.living_room.chair     308
sport.diving    2
appliances.personal.hair_cutter 1643
appliances.environment.air_conditioner  332
apparel.glove   18232
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
Time taken: 71.321 seconds, Fetched: 12 row(s)
```

Which brand had the maximum sales in October and November combined?
Answer: runail

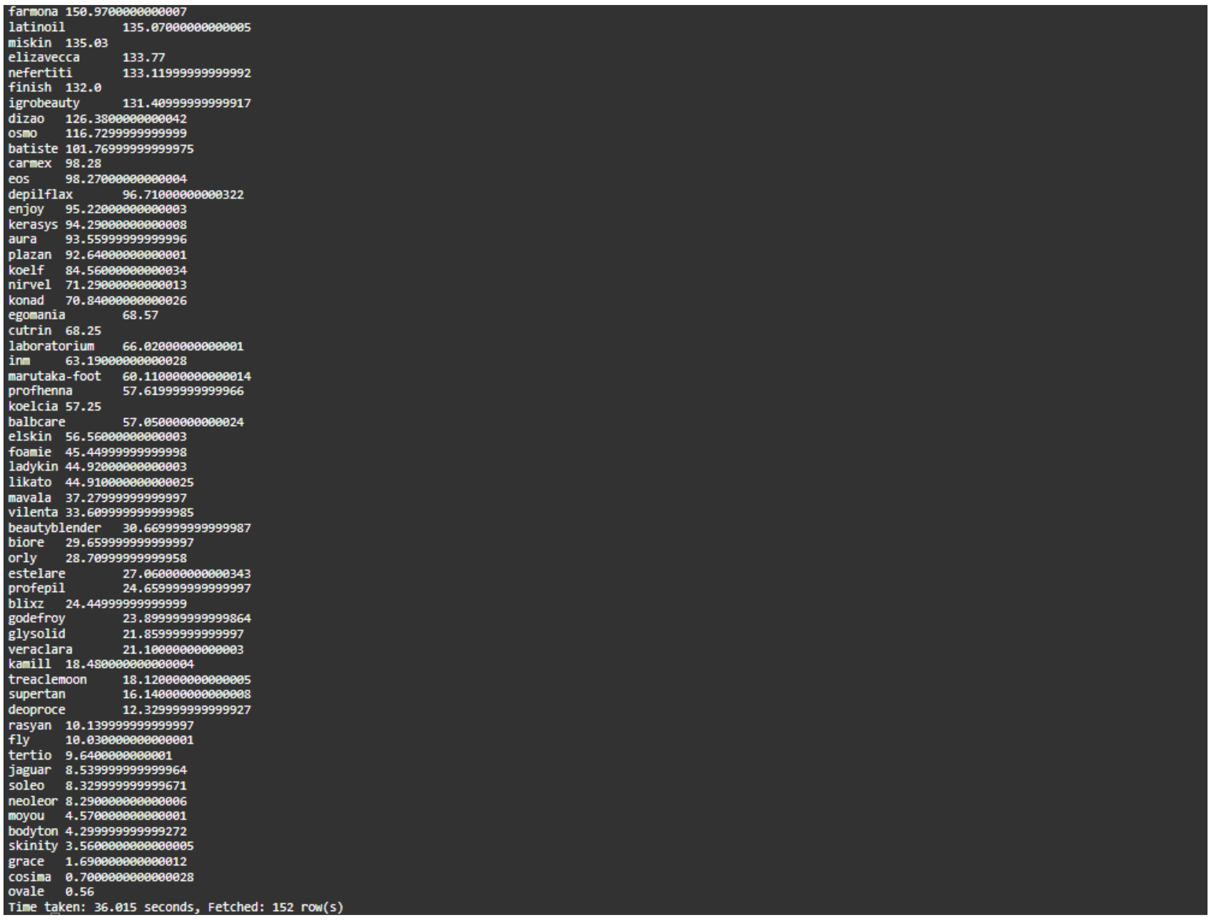


```
hive> select brand, sum(price) as sales from clickstream where brand is not null and brand != '' and event_type = 'purchase' group by brand order by sales desc limit 1;
Query ID = hadoop_20221130174646_668f8d6d-7309-45d7-8ffe-953985de5e05
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669827995695_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        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: 24.49 s
----------------------------------------------------------------------------------------------
OK
brand   sales
runail  148297.93999998464
Time taken: 25.349 seconds, Fetched: 1 row(s)
```

Which brands increased their sales from October to November?
There are a total of 152 brands that increased sales from October to November.

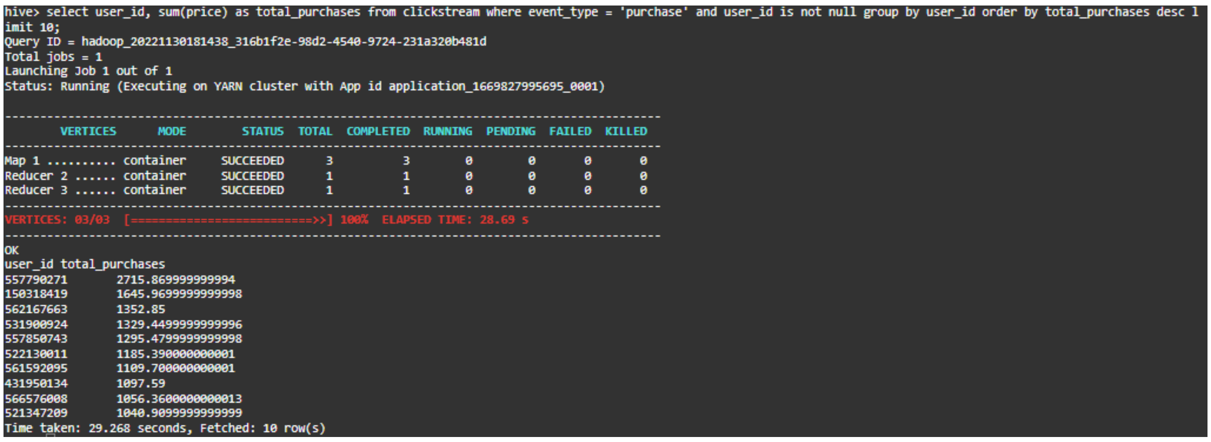


```
hive> with brands_sales_oct as (
    > select brand, sum(price) as oct_sales from clickstream where event_type = 'purchase' and brand is not null and brand != '' and month(event_time) = 10  group by b
rand
    > ),
    > brands_sales_nov as (
    > select brand, sum(price) as nov_sales from clickstream where event_type = 'purchase' and brand is not null and brand != '' and month(event_time) = 11  group by b
rand
    > )
    > select o.brand as brand, n.nov_sales - o.oct_sales as increase_in_sales from brands_sales_oct o inner join brands_sales_nov n on o.brand = n.brand where n.nov_sa
les - o.oct_sales > 0 order by increase_in_sales desc;
Query ID = hadoop_20221130180405_4448e44d-0ffa-4a2e-9482-76a7d21aad0e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669827995695_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Map 5 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 6 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 06/06  [==========================>>] 100%  ELAPSED TIME: 34.62 s
----------------------------------------------------------------------------------------------
OK
brand   increase_in_sales
grattol 36027.16999999717
uno     15737.71999999987
lianail 10501.400000000256
ingarden        10404.819999998246
strong  9474.639999999941
jessnail        7057.38999999977
cosmoprofi      6214.180000000126
polarus 5358.2100000000055
runail  5219.379999987461
freedecor       4250.020000000293
staleks 3355.8799999999537
bpw.style       3265.2900000009977
lovely  3234.679999999991
marathon        2992.350000000003
haruyama        2962.2199999996046
yoko    2950.9700000000103
italwax 2859.1300000003175
benovy  2850.3499999999967
kaypro  2387.36
estel   2385.9199999998928
concept 2348.260000000053
kapous  2165.92000000002
f.o.x   1953.0499999999656
masura  1792.3900000007234
milv    1737.0700000000502
beautix 1729.000000000011
artex   1596.6099999999942
domix   1537.11999999999
shik    1498.5200000000023
smart   1444.880000000021
roubloff        1422.409999999999
levrana 1420.5400000000004
oniq    1416.2399999999798
irisk   1354.0799999994633
severina        1344.600000000004
joico   1309.5799999999997
zeitun  1300.9700000000007
beauty-free     1228.6900000000132
swarovski       1155.2300000000162
de.lux  1115.8099999999913
metzger 1083.7099999999882
markell 1065.6800000000012
sanoto  1052.54
nagaraku        957.9400000000087
ecolab  951.4499999999999
art-visage      905.089999999992
levissime       857.8099999999904
missha  856.4500000000003
solomeya        786.0999999999938
rosi    764.520000000015
refectocil      759.4000000000042
kaaral  673.6399999999894
kosmekka        631.9300000000005
kinetics        611.0099999999829
browxenna       585.3600000000006
airnails        572.6200000000581
uskusi  548.0400000000345
coifin  525.4900000000001
s.care  500.38999999999993
limoni  487.70000000000095
matrix  483.4899999999989
gehwol  468.6099999999997
greymy  460.28000000000003
bioaqua 455.2299999999997
farmavita       454.60000000000036
sophin  447.65999999999985
yu-r    402.29999999999967
kiss    395.7800000000001
lador   387.91999999999234
ellips  360.1899999999999
jas     338.46999999999935
lowence 324.9100000000001
nitrile 315.40000000000043
shary   304.5299999999986
kims    302.00000000000006
happyfons       289.66999999999985
kocostar        284.0800000000001
insight 278.25999999999954
candy   264.4200000000003
bluesky 258.2899999998608
beauugreen      256.84000000000003
protokeratin    255.54000000000008
trind   244.89000000000004
entity  239.54999999999808
skinlite        238.51000000000056
provoc  235.83000000000243
fedua   211.43
ecocraft        200.79
keen    199.26999999999998
mane    193.47
freshbubble     183.64
chi     179.6700000000002
cristalinas     157.31999999999988
farmona 150.9700000000007
```



```
farmona 150.9700000000007
latinoil        135.07000000000005
miskin  135.03
elizavecca      133.77
nefertiti       133.11999999999992
finish  132.0
igrobeauty      131.40999999999917
dizao   126.3800000000042
osmo    116.7299999999999
batiste 101.76999999999975
carmex  98.28
eos     98.27000000000004
depilflax       96.71000000000322
enjoy   95.22000000000003
kerasys 94.29000000000008
aura    93.55999999999996
plazan  92.64000000000001
koelf   84.56000000000034
nirvel  71.29000000000013
konad   70.84000000000026
egomania        68.57
cutrin  68.25
laboratorium    66.02000000000001
inm     63.19000000000028
marutaka-foot   60.110000000000014
profhenna       57.61999999999966
koelcia 57.25
balbcare        57.05000000000024
elskin  56.56000000000003
foamie  45.44999999999998
ladykin 44.92000000000003
likato  44.910000000000025
mavala  37.27999999999997
vilenta 33.609999999999985
beautyblender   30.669999999999987
biore   29.659999999999997
orly    28.70999999999958
estelare        27.060000000000343
profepil        24.659999999999997
blixz   24.44999999999999
godefroy        23.899999999999864
glysolid        21.85999999999997
veraclara       21.10000000000003
kamill  18.480000000000004
treaclemoon     18.120000000000005
supertan        16.140000000000008
deoproce        12.329999999999927
rasyan  10.139999999999997
fly     10.030000000000001
tertio  9.6400000000001
jaguar  8.539999999999964
soleo   8.329999999999671
neoleor 8.290000000000006
moyou   4.570000000000001
bodyton 4.299999999999272
skinity 3.5600000000000005
grace   1.690000000000012
cosima  0.7000000000000028
ovale   0.56
Time taken: 36.015 seconds, Fetched: 152 row(s)
```

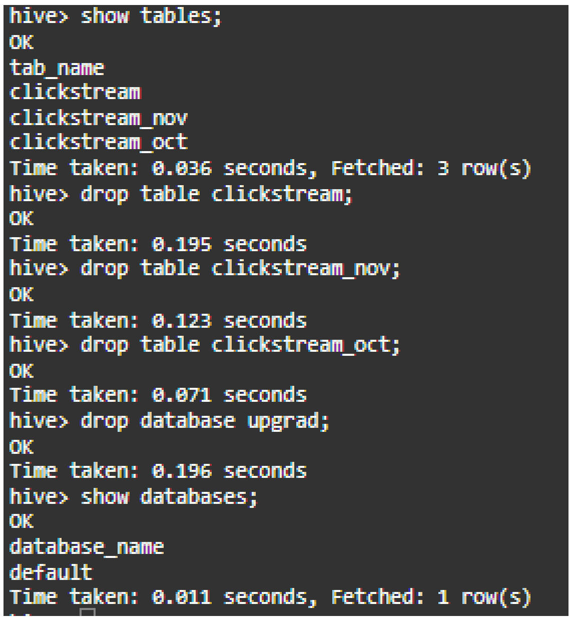
Write a query to generate a list of top 10 users who spend the most.



```
hive> select user_id, sum(price) as total_purchases from clickstream where event_type = 'purchase' and user_id is not null group by user_id order by total_purchases desc l
imit 10;
Query ID = hadoop_20221130181438_316b1f2e-98d2-4540-9724-231a320b481d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669827995695_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        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: 28.69 s
----------------------------------------------------------------------------------------------
OK
user_id total_purchases
557790271       2715.869999999994
150318419       1645.9699999999998
562167663       1352.85
531900924       1329.4499999999996
557850743       1295.4799999999998
522130011       1185.390000000001
561592095       1109.700000000001
431950134       1097.59
566576008       1056.3600000000013
521347209       1040.9099999999999
Time taken: 29.268 seconds, Fetched: 10 row(s)
```

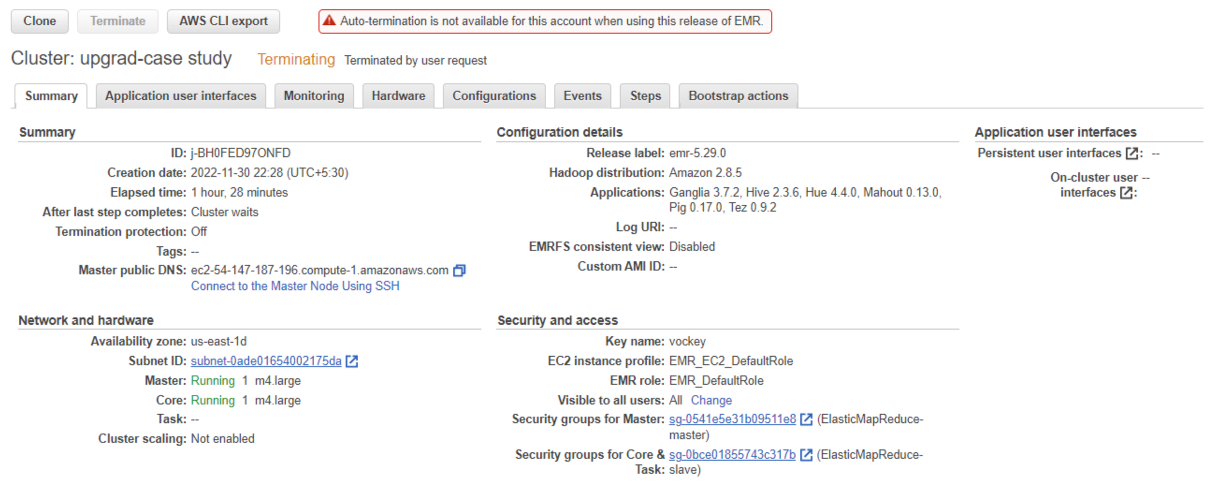
Cleaning up

Dropping the Database



```
hive> show tables;
OK
tab_name
clickstream
clickstream_nov
clickstream_oct
Time taken: 0.036 seconds, Fetched: 3 row(s)
hive> drop table clickstream;
OK
Time taken: 0.195 seconds
hive> drop table clickstream_nov;
OK
Time taken: 0.123 seconds
hive> drop table clickstream_oct;
OK
Time taken: 0.071 seconds
hive> drop database upgrad;
OK
Time taken: 0.196 seconds
hive> show databases;
OK
database_name
default
Time taken: 0.011 seconds, Fetched: 1 row(s)
```

Terminating the Cluster



Auto-termination is not available for this account when using this release of EMR.

Cluster: upgrad-case study Terminating Terminated by user request

Summary

ID: j-BH0FED97ONFD
Creation date: 2022-11-30 22:28 (UTC+5:30)
Elapsed time: 1 hour, 28 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-54-147-187-196.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: --
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: --

Network and hardware

Availability zone: us-east-1d
Subnet ID: subnet-0ade01654002175da
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: vockey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0541e5e31b09511e8 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0bce01855743c317b (ElasticMapReduce-slave)

Appendix

Script of all the commands used:

```
#Move Data from S3 to HDFS
aws s3 cp s3://tutorialprashantk/2019-Nov.csv .
aws s3 cp s3://tutorialprashantk/2019-Oct.csv .

#Creating a database on Hive
create database upgrad;
use upgrad;

#Printing Headers
set hive.cli.print.header=true ;

#Creating a table
CREATE EXTERNAL TABLE clickstream_nov(
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)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = ","
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

```
CREATE EXTERNAL TABLE clickstream_oct(
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)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = ","
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

#Loading Data
load data local inpath '/home/hadoop/2019-Nov.csv' into table clickstream_nov;
load data local inpath '/home/hadoop/2019-Oct.csv' into table clickstream_oct;

#Enabling Dynamic Partitioning
set hive.exec.dynamic.partition = true;
set hive.exec.dynamic.partition.mode = nonstrict;
```

```
#Creating our final table with partitions and buckets
CREATE EXTERNAL TABLE clickstream(
event_time timestamp,
product_id string,
category_id string,
category_code string,
brand string,
price float,
user_id bigint,
user_session string)
PARTITIONED BY (event_type string)
CLUSTERED by (category_id) into 7 buckets
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = ","
)
stored as textfile
TBLPROPERTIES ('serialization.null.format'='');


#Inserting data
insert into table clickstream partition (event_type) select event_time, product_id, category_id,
category_code, brand, price, user_id, user_session, event_type from clickstream_nov;


insert into table clickstream partition (event_type) select event_time, product_id, category_id,
category_code, brand, price, user_id, user_session, event_type from clickstream_oct;


--------- Getting Answers for questions -----------
#Find the total revenue generated due to purchases made in October.
select sum(price) as total_revenue from clickstream where event_type = 'purchase' and
month(event_time) = 10;


#Write a query to yield the total sum of purchases per month in a single output.
select month(event_time) as month, sum(price) as sum_of_purchases from clickstream where
event_type = 'purchase' group by month(event_time);


#Write a query to find the change in revenue generated due to purchases from October to November.
with monthwiserevenue as(
select month(event_time) as month, sum(price) as revenue from clickstream where event_type =
'purchase' group by month(event_time))
select revenue - lag(revenue) over (order by month) as change_in_revenue from monthwiserevenue;
```

```
#Find distinct categories of products. Categories with null category code can be ignored

select distinct category_code from clickstream where category_code is not null;


#Find the total number of products available under each category.

select category_code, count(product_id) as number_of_products from clickstream where
category_code is not null group by category_code;


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

select brand, sum(price) as sales from clickstream where brand is not null and brand != '' and
event_type = 'purchase' group by brand order by sales desc limit 1;


#Which brands increased their sales from October to November?

with brands_sales_oct as (

select brand, sum(price) as oct_sales from clickstream where event_type = 'purchase' and brand is not
null and brand != '' and month(event_time) = 10  group by brand

),

brands_sales_nov as (

select brand, sum(price) as nov_sales from clickstream where event_type = 'purchase' and brand is not
null and brand != '' and month(event_time) = 11  group by brand

)

select o.brand as brand, n.nov_sales - o.oct_sales as increase_in_sales from brands_sales_oct o inner
join brands_sales_nov n on o.brand = n.brand where n.nov_sales - o.oct_sales > 0 order by
increase_in_sales desc;


#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.

select user_id, sum(price) as total_purchases from clickstream where event_type = 'purchase' and
user_id is not null group by user_id order by total_purchases desc limit 10;
```