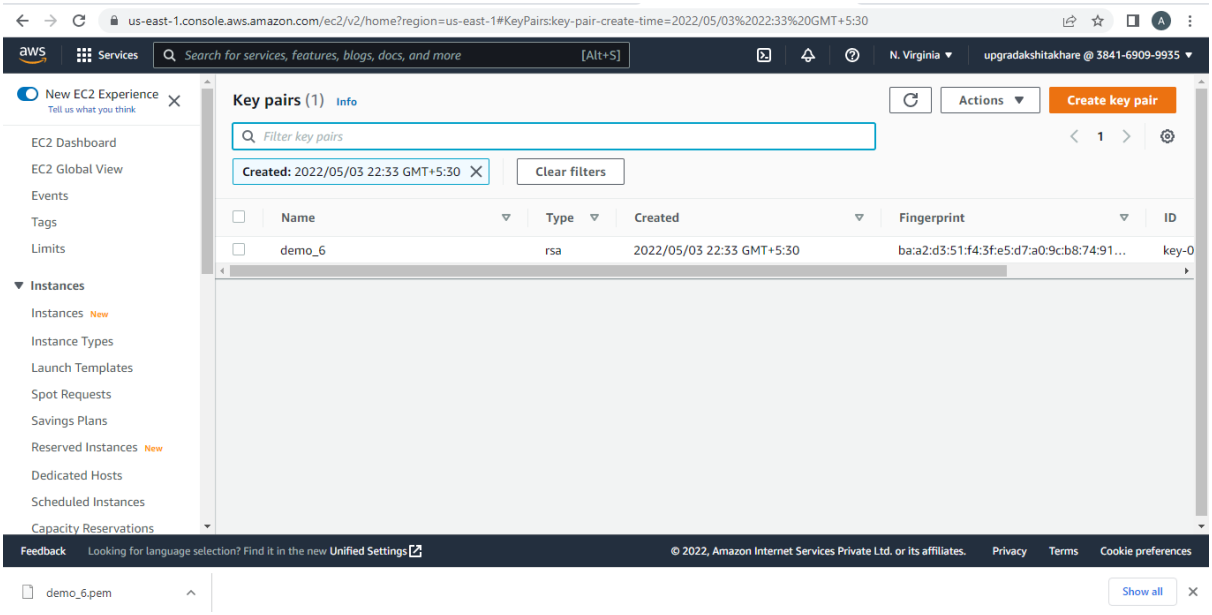
Hive E-commerce Case study

Problem Statement

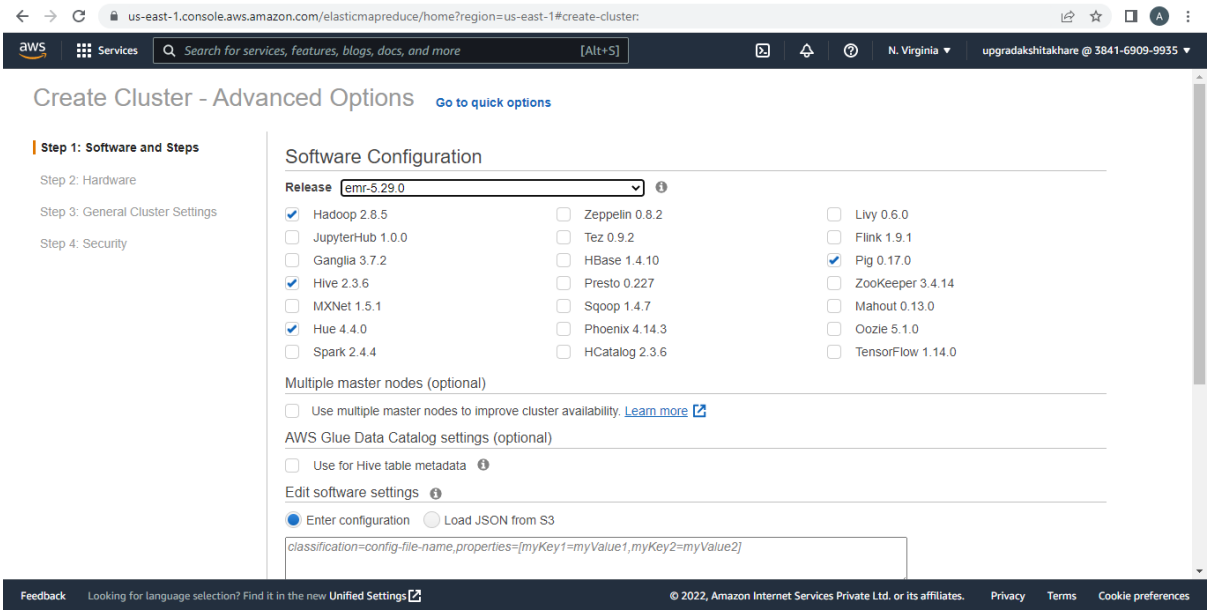
With online sales gaining popularity, tech companies are exploring ways to improve their sales by analysing customer behaviour and gaining insights about product trends. Furthermore, the websites make it easier for customers to find the products they require without much scavenging. Needless to say, the role of big data analysts is among the most sought-after job profiles of this decade. Therefore, as part of this assignment, we will be challenging you, as a big data analyst, to extract data and gather insights from a real-life data set of an e-commerce company.



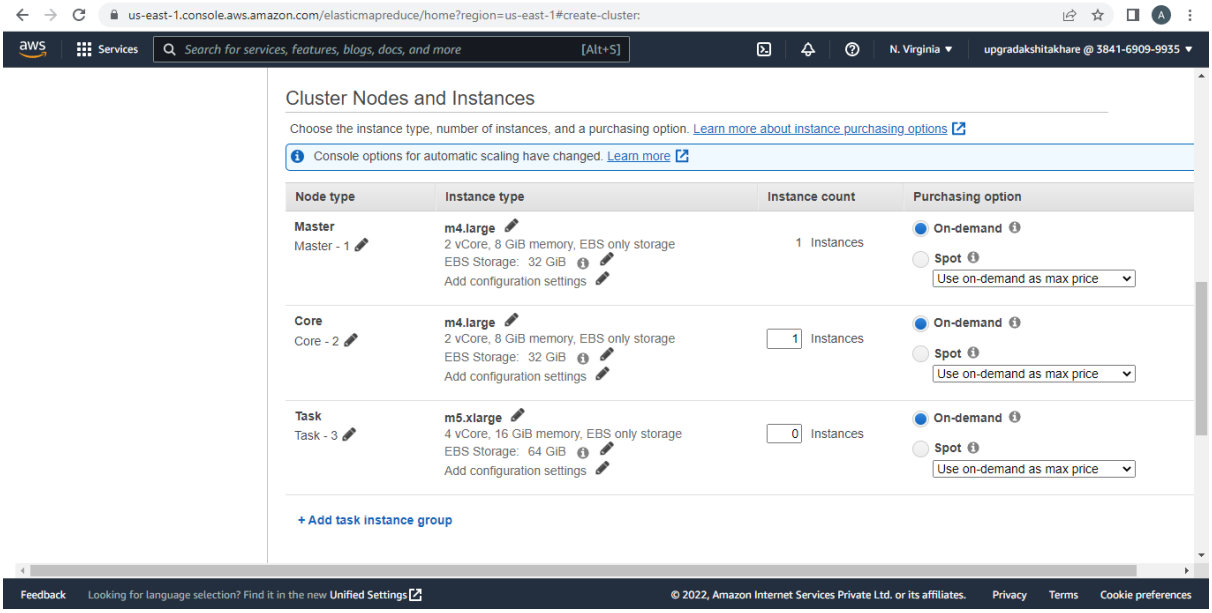
Creation of key pair

Creating the EMR version 5.29.0 as suggested.

Hive E-commerce Case study

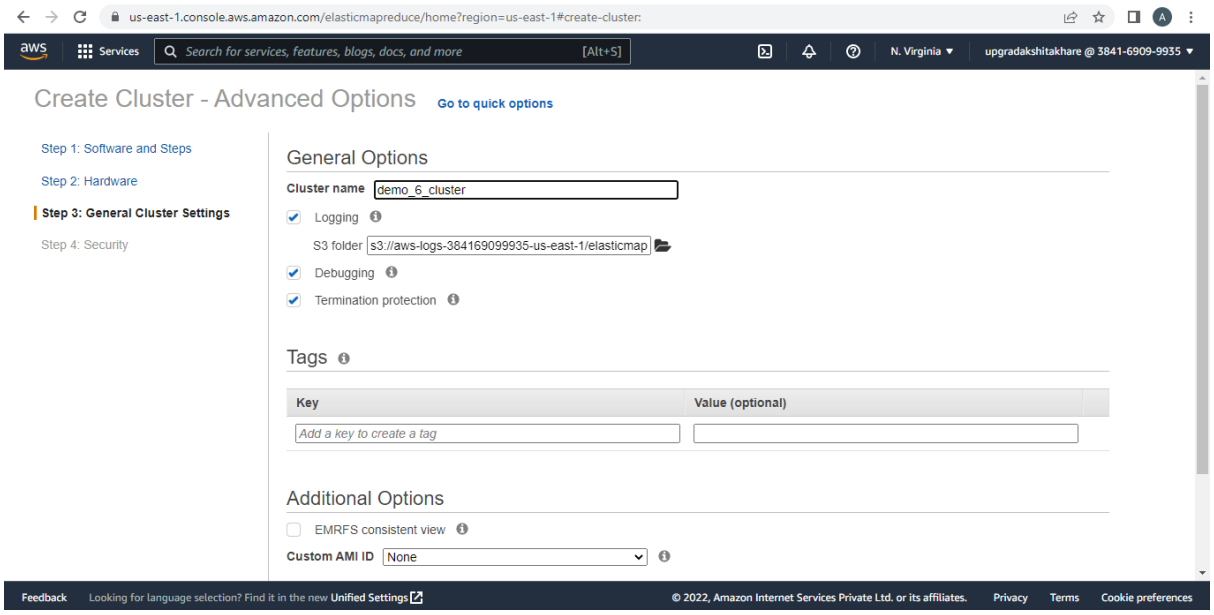


Creating the cluster with 1master node and 1 slave node.

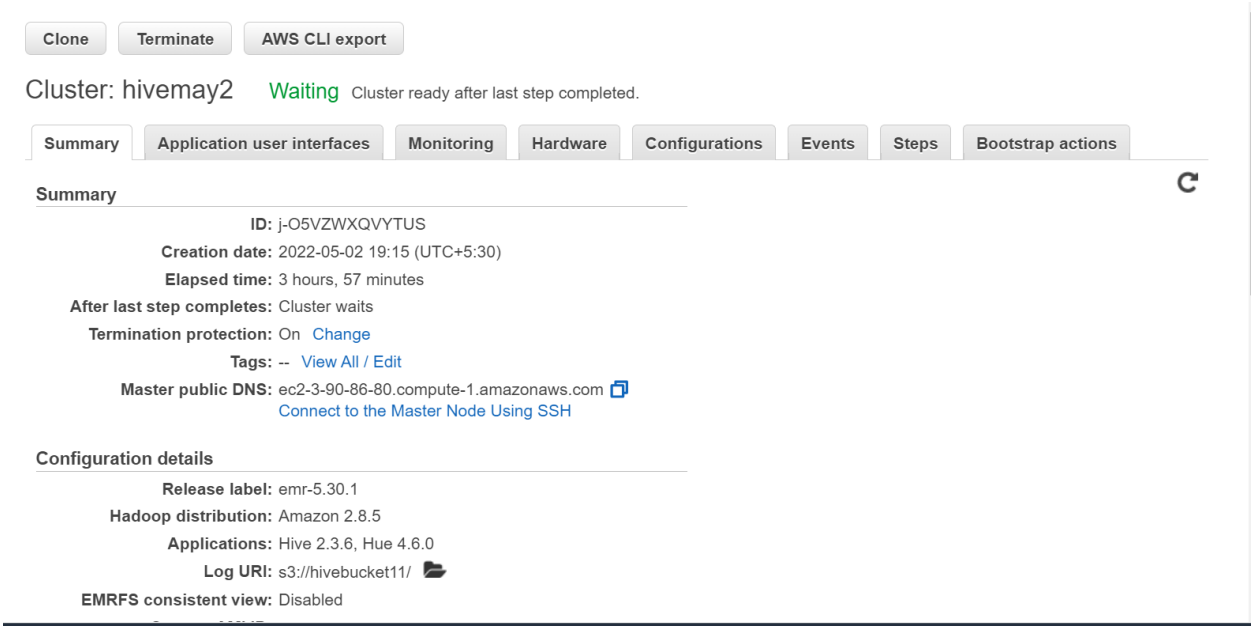


Naveli & Akshita

Hive E-commerce Case study

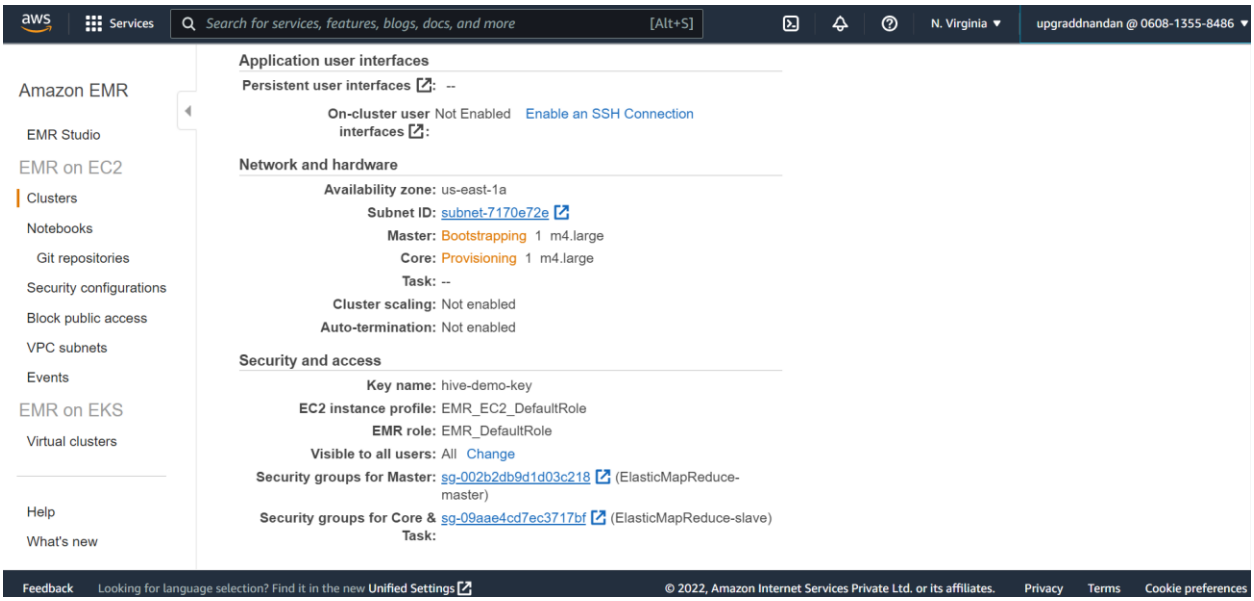


Loading the 2019-Oct.csv and 2019-Nov.csv to s3 hivebucket 11



Naveli & Akshita

Hive E-commerce Case study



We loaded the files in S3 and then loaded the hadoop EMR Cluster

```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-44-20 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-44-20 ~]$ aws s3 cp s3://hivebucket11/2019-Oct.csv .
download: s3://hivebucket11/2019-Oct.csv to ./2019-Oct.csv
[hadoop@ip-172-31-44-20 ~]$ ls
2019-Oct.csv
[hadoop@ip-172-31-44-20 ~]$ aws s3 cp s3://hivebucket11/2019-Nov.csv .
download: s3://hivebucket11/2019-Nov.csv to ./2019-Nov.csv
[hadoop@ip-172-31-44-20 ~]$ ls
2019-Nov.csv  2019-Oct.csv
[hadoop@ip-172-31-44-20 ~]$ create database if not exixts
```

Naveli & Akshita

Hive E-commerce Case study

Creating the database ,table and loading the Oct and Nov data to the same table.
Altering the table with time stamp format.

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>  create database if not exists edata;
OK
Time taken: 0.895 seconds
hive> use edata;
OK
Time taken: 0.12 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecom_events (
    > 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
    >  )
    >  COMMENT 'ecom_events Table'
    >  ROW FORMAT DELIMITED
    >  FIELDS TERMINATED BY ','
    >  LOCATION '/tmp/ecomdata';  ;
OK
Time taken: 1.449 seconds
hive> show tables;
OK
ecom_events
Time taken: 0.137 seconds, Fetched: 1 row(s)
hive> Load data local inpath '/home/hadoop/2019-Oct.csv' into table ecom_events;
Loading data to table edata.ecom_events
OK
Time taken: 9.987 seconds
hive> ALTER TABLE ecom_events SET TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.14 seconds
hive>
    >  ALTER TABLE ecom_events SET TBLPROPERTIES ("timestamp.formats"="yyyy-MM-d
d HH:mm:ss 'UTC'");
OK
Time taken: 0.114 seconds
hive>
    >  set hive.cli.print.header=true;
hive>
```

Describing the table:

```
hive> describe ecom_events;
OK
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
Time taken: 0.103 seconds, Fetched: 9 row(s)
hive> describe ecom_events;
```

Checking for the loaded table by using cmd
select * from ecom_events limit 10;

Hive E-commerce Case study

```
hive> use edata;
OK
Time taken: 0.118 seconds
hive> show tables;
OK
ecom_events
Time taken: 0.343 seconds, Fetched: 1 row(s)
hive> load data local inpath'/home/hadoop/2019-Oct.csv' into table ecom_events ;
Loading data to table edata.ecom_events
OK
Time taken: 9.365 seconds
hive> select * from ecom_events limit 10;
OK
NULL    event_type      product_id      category_id     category_code   brand   NULL    NULL    user_session
NULL    cart    5773203 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
NULL    cart    5773353 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
NULL    cart    5881589 2151191071051219817             lovely  13.48   429681830       49e8d843-adf3-428b-a2c3-fe8bc6a307c9
NULL    cart    5723490 1487580005134238553             runail  2.62    463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885
NULL    cart    5881449 1487580013522845895             lovely  0.56    429681830       49e8d843-adf3-428b-a2c3-fe8bc6a307c9
NULL    cart    5857269 1487580005134238553             runail  2.62    430174032       73dea1e7-664e-43f4-8b30-d32b9d5af04f
NULL    cart    5739055 1487580008246412266             kapous  4.75    377667011       81326ac6-daa4-4f0a-b488-fd0956a78733
NULL    cart    5825598 1487580009445982239                     0.56    467916806       2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
NULL    cart    5698989 1487580006317032337                     1.27    385985999       d30965e8-1101-44ab-b45d-cc1bb9fae694
Time taken: 2.362 seconds, Fetched: 10 row(s)
hive>
```

Hive Assignment Questions :

Running the queries before partition and screen shots for same:

QUERY1: Find the total revenue generated due to purchases made in October.

select sum(price) as oct_revenue from ecom_events where event_type='purchase' and month(event_time)=10;

```
hive> select sum(price) as oct_revenue from ecom_events where event_type='purchase' and month(event_time)=10;
Query ID = hadoop_20220503163341_f1c35306-5239-4204-97de-a656ea6dfc50
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_1651587703400_0004)

--------------------------------------------------------------------------------
        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: 13.72 s
--------------------------------------------------------------------------------
OK
oct_revenue
1211538.4295325726
Time taken: 27.1 seconds, Fetched: 1 row(s)
hive>
```

QUERY2: Write a query to yield the total sum of purchases per month in a single output.

select month(event_time), sum(price) as purchase_sum from ecom_events where event_type='purchase' group by month(event_time);

Naveli & Akshita

Hive E-commerce Case study

```
hive> select month(event_time), sum(price) as purchase_sum from ecom_events where event_type='purchase' group by month(event_time);
Query ID = hadoop_20220503170632_55c31c97-beb8-483f-bf31-5b878e6b0f71
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_1651587703400_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.02 s
--------------------------------------------------------------------------------
OK
_c0     purchase_sum
11      1531016.8991247676
10      1211538.4295325726
Time taken: 40.118 seconds, Fetched: 2 row(s)
hive>
```

QUERY3: Write a query to find the change in revenue generated due to purchases 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 ecom_events
   WHERE event_type= 'purchase'
    AND MONTH(event_time) in ('10', '11')
 )
SELECT (Oct_Revenue - Nov_Revenue) FROM month_revenue ;
```

```
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 ecom_events
    >      WHERE event_type= 'purchase'
    >        AND MONTH(event_time) in ('10', '11')
    >   )
    >
    >    SELECT (Oct_Revenue - Nov_Revenue) FROM month_revenue ;
Query ID = hadoop_20220503174218_afe197b0-3f5e-44d6-951f-74080def4716
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_1651587703400_0007)

--------------------------------------------------------------------------------
        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: 34.82 s
--------------------------------------------------------------------------------
OK
_c0
-319478.469592195
Time taken: 46.791 seconds, Fetched: 1 row(s)
hive>
```

Naveli & Akshita

Hive E-commerce Case study

QUERY4: Find distinct categories of products. Categories with null category code can be ignored.

SELECT DISTINCT split(category_code,'\\.')[0] AS category FROM ecom_events WHERE split(category_code,'\\.')[0]<>'' ;

```
hive> SELECT DISTINCT split(category_code,'\\.')[0] AS category FROM ecom_events WHERE split(category_code,'\\.')[0]<>'' ;
Query ID = hadoop_20220504110656_3db471e1-d597-412d-ac35-7717f8a249e4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651644824663_0005)

--------------------------------------------------------------------------------
        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: 25.71 s
--------------------------------------------------------------------------------
OK
accessories
apparel
appliances
furniture
sport
stationery
Time taken: 26.965 seconds, Fetched: 6 row(s)
hive>
```

QUERY5: Find the total number of products available under each category.

SELECT split(category_code,'\\.')[0] AS category, COUNT(product_id) AS prd FROM ecom_events GROUP BY split(category_code,'\\.')[0] ORDER BY prd DESC ;

```
hive> SELECT split(category_code,'\\.')[0] AS category, COUNT(product_id) AS prd FROM ecom_events GROUP BY split(category_code,'\\.')[0] ORDER BY prd DESC ;
Query ID = hadoop_20220504113532_af1e00e8-60a9-4a2a-add5-eba3bd16ccdd
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_1651644824663_0006)

--------------------------------------------------------------------------------
        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: 29.19 s
--------------------------------------------------------------------------------
OK
        8594895
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 40.02 seconds, Fetched: 7 row(s)
hive>
```

QUERY6: Which brand had the maximum sales in October and November combined?

SELECT brand, SUM(price) AS Sales FROM ecom_events WHERE brand <>'' AND event_type='purchase' GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;

Naveli & Akshita

Hive E-commerce Case study

```
hive>
    > SELECT brand, SUM(price) AS Sales FROM ecom_events WHERE brand <>'' AND event_type='purchase' GROUP BY brand ORDER BY Sales DESC LIMIT 1 ;
Query ID = hadoop_20220504120522_0a9d8757-dfc1-4858-8a72-38f1fb2d9a55
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_1651644824663_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 28.83 s
--------------------------------------------------------------------------------
OK
runail  148297.93996394053
Time taken: 40.739 seconds, Fetched: 1 row(s)
hive>
```

QUERY7: Which brands increased their sales from October to November?

WITH monthly_diff AS (SELECT brand, SUM(CASE WHEN date_format(event_time,'MM')=10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format(event_time,'MM')=11 THEN price ELSE 0 END) AS November FROM ecom_events WHERE event_type='purchase' GROUP BY brand) SELECT brand, October, November, (November-October) as Sales_diff FROM monthly_diff WHERE (November-October) >0 ORDER BY Sales_diff ;

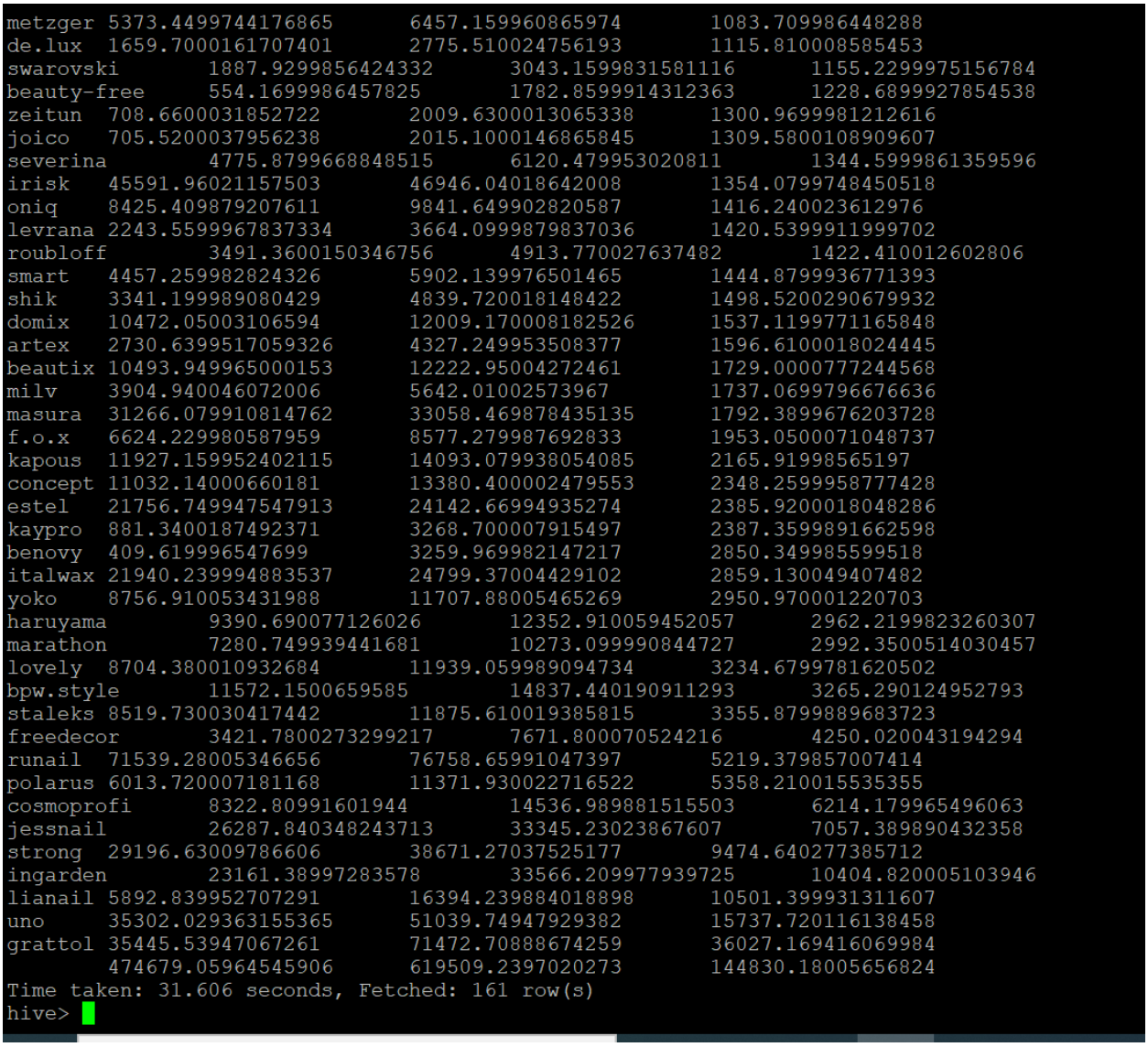
```
hive> WITH monthly_diff AS ( SELECT brand, SUM(CASE WHEN date_format(event_time,'MM')=10 THEN price ELSE 0 END) AS October, SUM(CASE WHEN date_format(event_t
ime,'MM')=11 THEN price ELSE 0 END) AS November FROM ecom_events WHERE event_type='purchase' GROUP BY brand) SELECT brand, October, November, (November-Octob
er) as Sales_diff FROM monthly_diff WHERE (November-October) >0 ORDER BY Sales_diff ;
Query ID = hadoop_20220504120847_97c906e2-1d7d-4016-a2fe-fa5865cfecf7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1651644824663_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.59 s
--------------------------------------------------------------------------------
OK
ovale   2.5399999618530273      3.0999999046325684      0.559999942779541
cosima  20.230000972747803      20.930000603199005      0.6999996304512024
grace   100.9200005531311       102.61000108718872      1.6900005340576172
helloganic      0.0     3.0999999046325684      3.0999999046325684
skinity 8.880000114440918       12.440000057220459      3.559999942779541
bodyton 1376.3399817943573      1380.639987230301       4.3000054359436035
moyou   5.710000038146973       10.28000020980835       4.570000171661377
neoleor 43.40999984741211       51.70000076293945       8.290000915527344
soleo   204.1999952197075       212.52999597787857      8.330000758171082
jaguar  1102.110021829605       1110.6500117778778      8.539989948272705
tertio  236.15999841690063      245.80000019073486      9.640001773834229
fly     17.139999389648438      27.170000553131104      10.030001163482666
rasyan  18.799999952316284      28.940000295639038      10.140000343322754
deoproce        316.8399999141693       329.1699993610382       12.329999446868896
barbie  0.0     12.390000343322754      12.390000343322754
supertan        50.37000048160553       66.51000016927719       16.13999968767166
treaclemoon     163.36999654769897      181.48999691009521      18.12000036239624
kamill  63.010000228881836      81.48999953269958       18.47999930381775
juno    0.0     21.079999923706055      21.079999923706055
veraclara       50.11000084877014       71.21000015735626       21.09999930858612
glysolid        69.73000013828278       91.59000062942505       21.860000491142273
godefroy        401.22000312805176      425.1200022697449       23.899999141693115
binacil 0.0     24.260000228881836      24.260000228881836
blixz   38.94999921321869       63.400001764297485      24.450002551078796
profepil        93.36000156402588       118.01999974250793      24.659998178482056
estelare        444.80999556183815      471.86999905109406      27.060003489255905
```

Hive E-commerce Case study

```
orly    902.3799939155579       931.0899903774261       28.709996461868286
biore   60.650001525878906      90.30999946594238       29.659997940063477
beautyblender   78.73999977111816       109.41000175476074      30.670001983642578
vilenta 197.59999787807465      231.20999908447266      33.61000120639801
mavala  409.0400023460388       446.32000255584717      37.28000020980835
likato  296.0599980354309       340.9699954986572       44.90999746322632
ladykin 125.64999961853027      170.56999969482422      44.920000076293945
foamie  35.03999996185303       80.48999977111816       45.44999980926514
elskin  251.0900001525879       307.6499996781349       56.55999952554703
balbcare        155.33000373840332      212.3800015449524       57.04999780654907
koelcia 55.5    112.75  57.25
profhenna       679.2300038337708       736.8500001430511       57.619996309280396
kares   0.0     59.45000076293945       59.45000076293945
marutaka-foot   49.21999979019165       109.33000040054321      60.11000061035156
dewal   0.0     61.28999876976013       61.28999876976013
inm     288.01999855041504      351.2099983692169       63.18999981880188
laboratorium    246.49999952316284      312.5199975967407       66.01999807357788
cutrin  299.3700017929077       367.6199998855591       68.24999809265137
egomania        77.46999835968018       146.04000091552734      68.57000255584717
konad   739.8300001621246       810.6699978709221       70.83999770879745
nirvel  163.04000329971313      234.33000826835632      71.29000496864319
koelf   422.7300081253052       507.29000186920166      84.55999374389648
plazan  101.37000036239624      194.010000705719        92.64000034332275
aura    83.95000076293945       177.5100040435791       93.56000328063965
kerasys 430.9100044965744       525.2000050544739       94.29000055789948
enjoy   41.34999966621399       136.57000184059143      95.22000217437744
depilflax       2707.0699973106384      2803.7799961566925      96.70999884605408
eos     54.34000015258789       152.60999727249146      98.26999711990356
carmex  145.07999897003174      243.3599967956543       98.27999782562256
batiste 772.400013923645        874.1700088977814       101.76999497413635
osmo    645.5800037384033       762.3100028038025       116.72999906539917
dizao   819.1300112009048       945.5100176334381       126.38000643253326
igrobeauty      513.6600003838539       645.0699995160103       131.40999913215637
finish  98.37999773025513       230.37999820709229      132.00000047683716
nefertiti       233.51999759674072      366.64000034332275      133.12000274658203
elizavecca      70.52999973297119       204.29999923706055      133.76999950408936
miskin  158.04000186920166      293.0700011253357       135.02999925613403
latinoil        249.5199966430664       384.5899987220764       135.07000207901
farmona 1692.46000289917        1843.4299907684326      150.9699878692627
cristalinas     427.63000297546387      584.950008392334        157.32000541687012
chi     358.93999576568604      538.6099972724915       179.67000150680542
matreshka       0.0     182.66999757289886      182.66999757289886
freshbubble     318.69999980926514      502.3399975299835       183.63999772071838
mane    66.79000186920166       260.26000118255615      193.4699993133545
```

Naveli & Akshita

Hive E-commerce Case study



QUERY8: 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 expense FROM ecom_events WHERE event_type='purchase' GROUP BY user_id ORDER BY expense DESC LIMIT 10 ;

Hive E-commerce Case study

```
hive> SELECT user_id, SUM(price) AS expense FROM ecom_events WHERE event_type='purchase' GROUP BY user_id ORDER BY expense DESC LIMIT 10 ;
Query ID = hadoop_20220504121614_4132da98-6f40-4796-bff3-837deb75417b
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_1651644824663_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.53 s
--------------------------------------------------------------------------------
OK
557790271       2715.8699957430363
150318419       1645.970008611679
562167663       1352.8499938696623
531900924       1329.4499949514866
557850743       1295.4800310581923
522130011       1185.3899966478348
561592095       1109.700007289648
431950134       1097.5900000333786
566576008       1056.3600097894669
521347209       1040.9099964797497
Time taken: 41.267 seconds, Fetched: 10 row(s)
hive>
```

Performance analysis running query before partition and after partition and screen shots for same:

```
hive> select * from clickStream2019 limit 5;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0
.32     562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2
.38     553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     2
2.22    556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail
3.16    564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900     3
.33     553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 2.292 seconds, Fetched: 5 row(s)
```

Select * on non-partitioned table took 2.292 sec

```
hive> select * from part_cs_2019 limit 5;
OK
NULL    remove_from_cart        5825586 1487580005754995573                     4.440   516935834       8bbfecd8-d688-4134-bb54-246b8b7dad85    10
NULL    cart    5773203 1487580005134238553             runail  2.620   463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885    10
NULL    cart    5663062 1487580009622143014                     1.430   251478914       a99a5589-0f7a-40a5-9748-b19961fc4d30    10
NULL    cart    5773353 1487580005134238553             runail  2.620   463240011       26dd6e6e-4dac-4778-8d2c-92e149dab885    10
NULL    view    5826000 1487580005092955511             lianail 6.330   557157386       04a21b33-6629-41f5-846e-602d4d894209    10
Time taken: 0.227 seconds, Fetched: 5 row(s)
```

Select * on partitioned table took 0.229 sec

QUERY2: Write a query to yield the total sum of purchases per month in a single output.

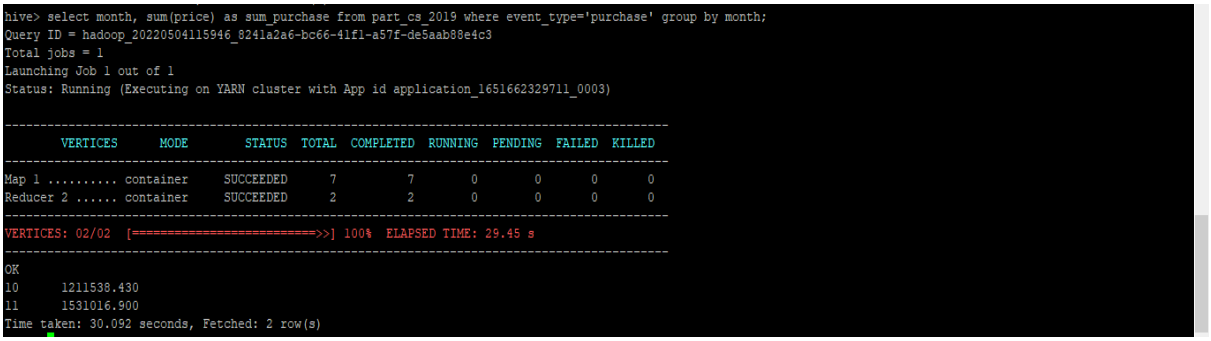
select month(event_time), sum(price) as purchase_sum from ecom_events where event_type='purchase' group by month(event_time);

```
hive>  select month(event_time) as month, sum(price) as sum_purchase from clickstream2019 where event_type='purchase' group by month(event_time);
Query ID = hadoop_20220504135554_2e4de3a4-d376-4447-9946-92a5845d2a6d
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_1651670928154_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 62.71 s
--------------------------------------------------------------------------------
OK
10      1211538.4299997438
11      1531016.900000122
Time taken: 74.247 seconds, Fetched: 2 row(s)
```

2nd query on non-partitioned table took 74.247 sec

Naveli & Akshita

Hive E-commerce Case study



2nd query on partitioned table took 30.092 sec

Partitioning on the basis of month:

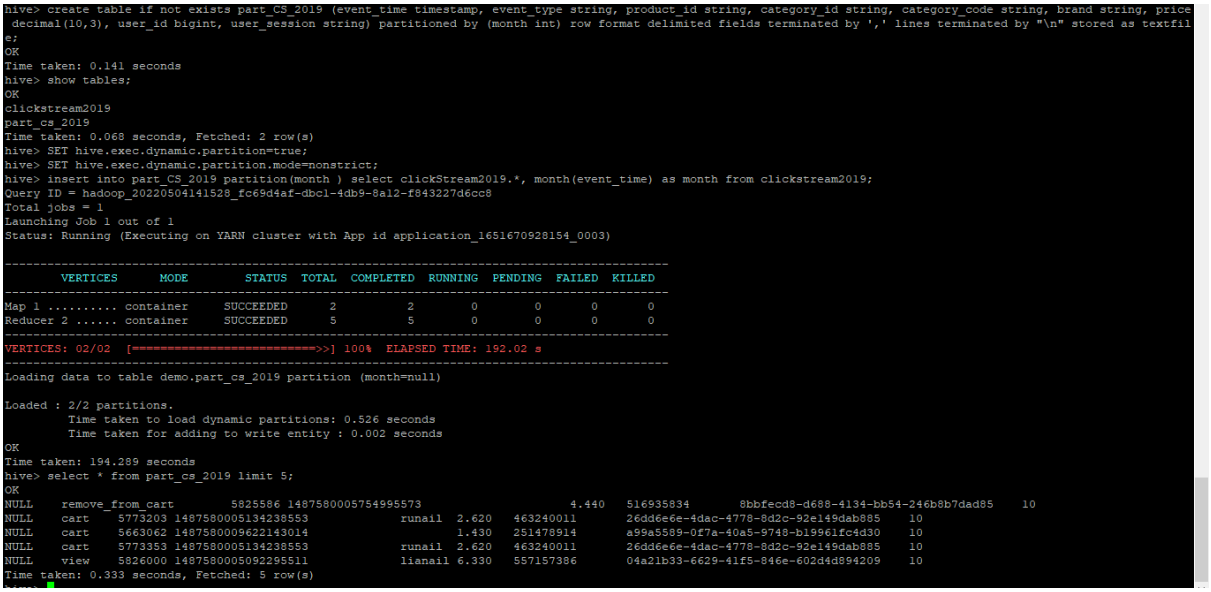
Partitioning segregates all the entries for the various columns of the dataset on the basis of the parameter chosen for partitioning and stores the data in their respective partitions. Hence, While we write the query to fetch the values from the table, only the required partitions of the table are queried. Thus, it reduces the time taken by the query to yield the result. In dynamic partitioning, the values of partitioned columns exist within the table. So, it is not required to pass the values of partitioned columns manually. It works well for columns having low cardinality.

There are two modes of **dynamic partitioning:**
Strict: This needs at least one column to be static while loading the data.
Non-strict: This allows us to have dynamic values of all the partition columns.

In the next snapshot, we have created a dynamic partitioned table named “part_cs_2019” and how data has been loaded into it using the “insert” clause:

By default, dynamic partitioning is not allowed in Hive. So it has to be enabled .

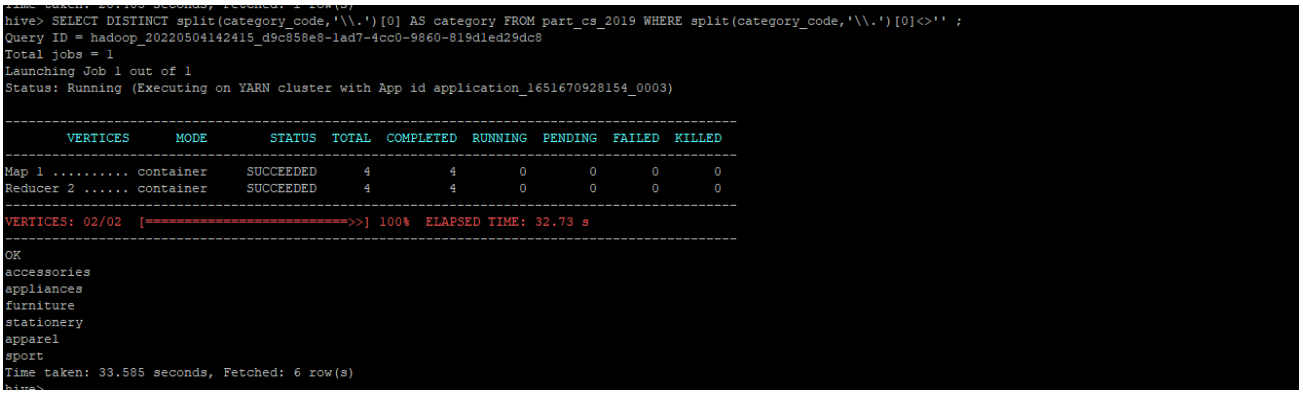
Hive E-commerce Case study



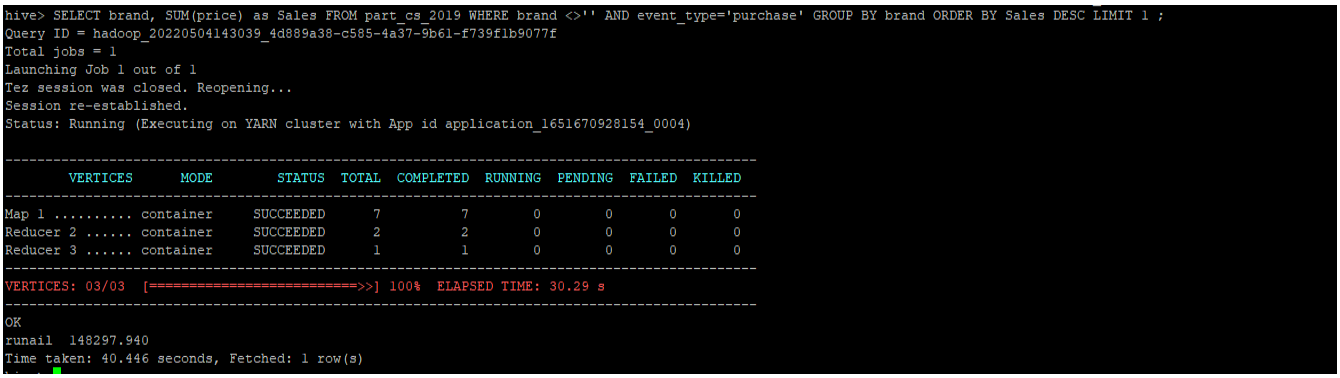
After partitioning the time taken for the execution of query is less.

Accessing the content of the table is much faster after partition and buckecting.

Query 4) after partitioning



Query 6) after partitioning

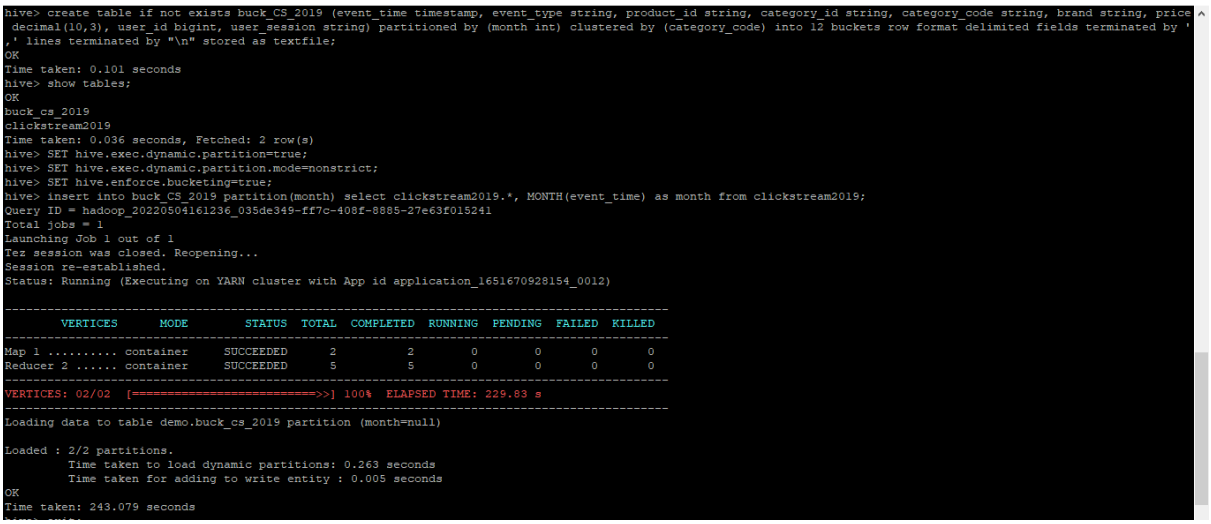


Naveli & Akshita

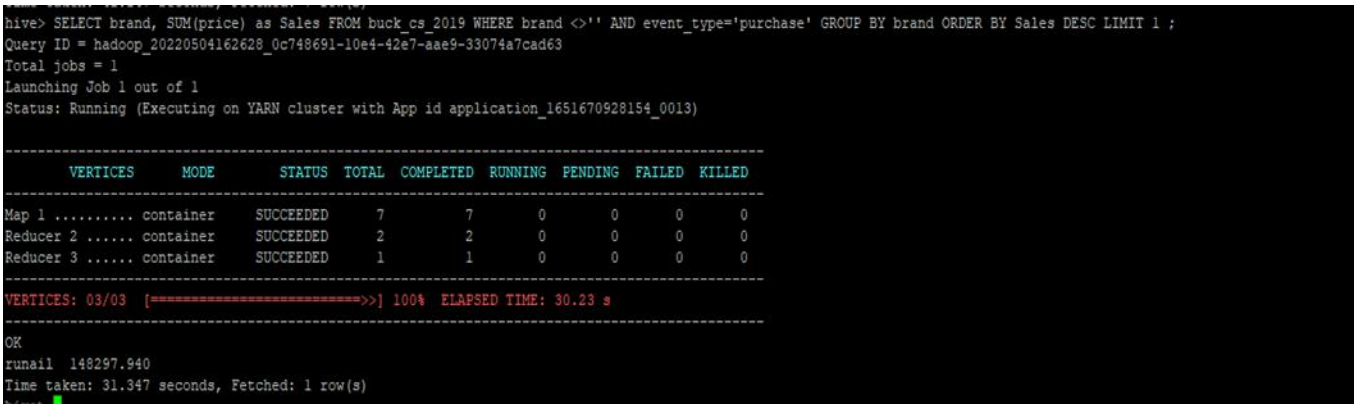
Hive E-commerce Case study

Bucketing :

Bucketing provides flexibility to further segregate the data into more manageable sections called buckets or clusters. **CLUSTERED BY** clause is used to divide the table into buckets. It works well for the columns having high cardinality.



Running query 6 after bucketing



Naveli & Akshita