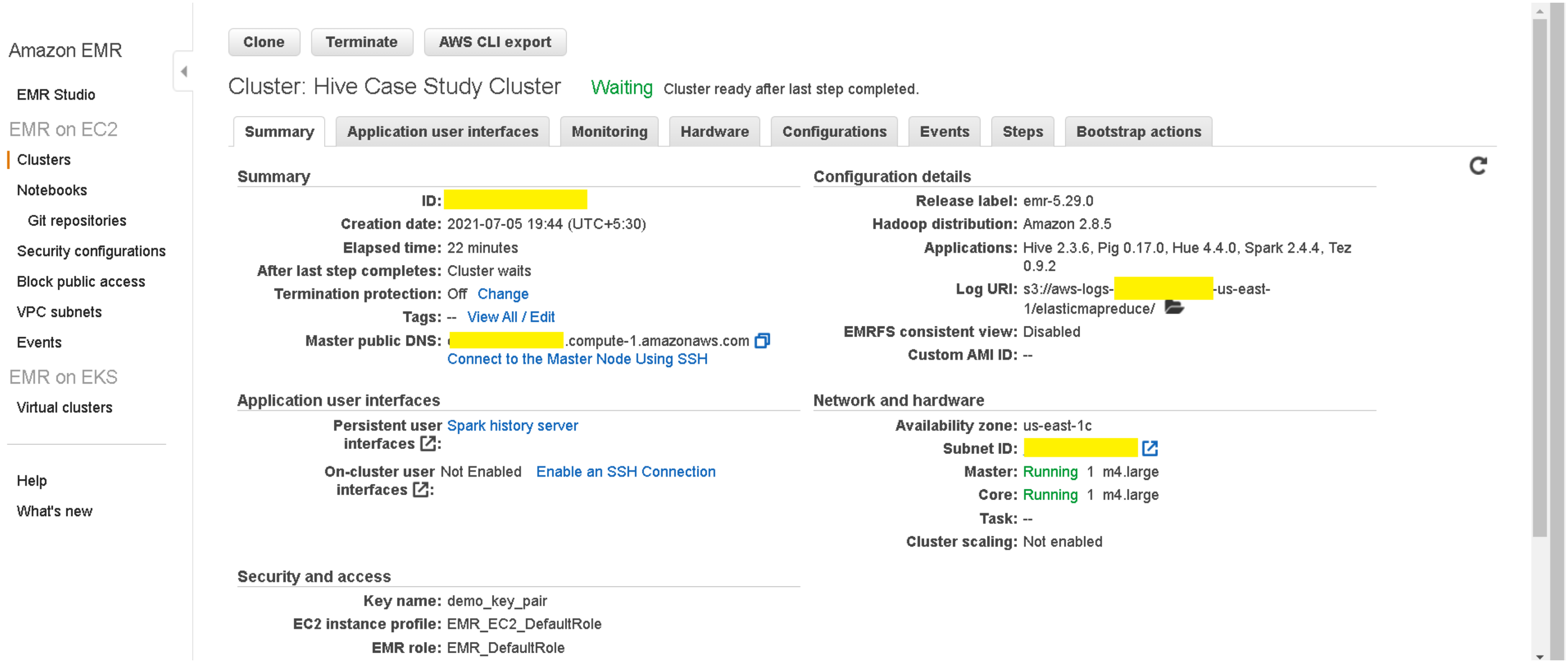
Hive Case Study

➢ Loading the EMR cluster

- Launch an EMR cluster with one master node and one core node with configuration of m4.large each, as shown below.



Cluster: Hive Case Study Cluster Waiting Cluster ready after last step completed.

Summary

ID: [redacted]
Creation date: 2021-07-05 19:44 (UTC+5:30)
Elapsed time: 22 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: [redacted].compute-1.amazonaws.com
Connect to the Master Node Using SSH

Application user interfaces

Persistent user interfaces: Spark history server
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Security and access

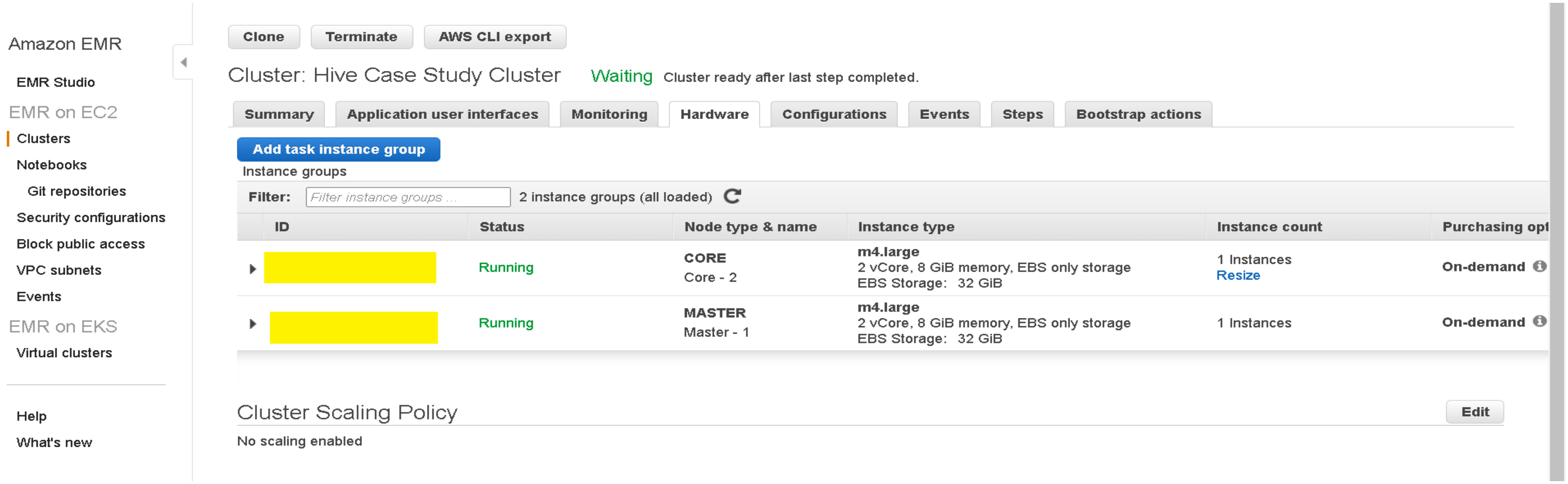
Key name: demo_key_pair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0, Spark 2.4.4, Tez 0.9.2
Log URI: s3://aws-logs-[redacted]-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-east-1c
Subnet ID: [redacted]
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled



Cluster: Hive Case Study Cluster Waiting Cluster ready after last step completed.

Instance groups

Filter: Filter instance groups ... 2 instance groups (all loaded)

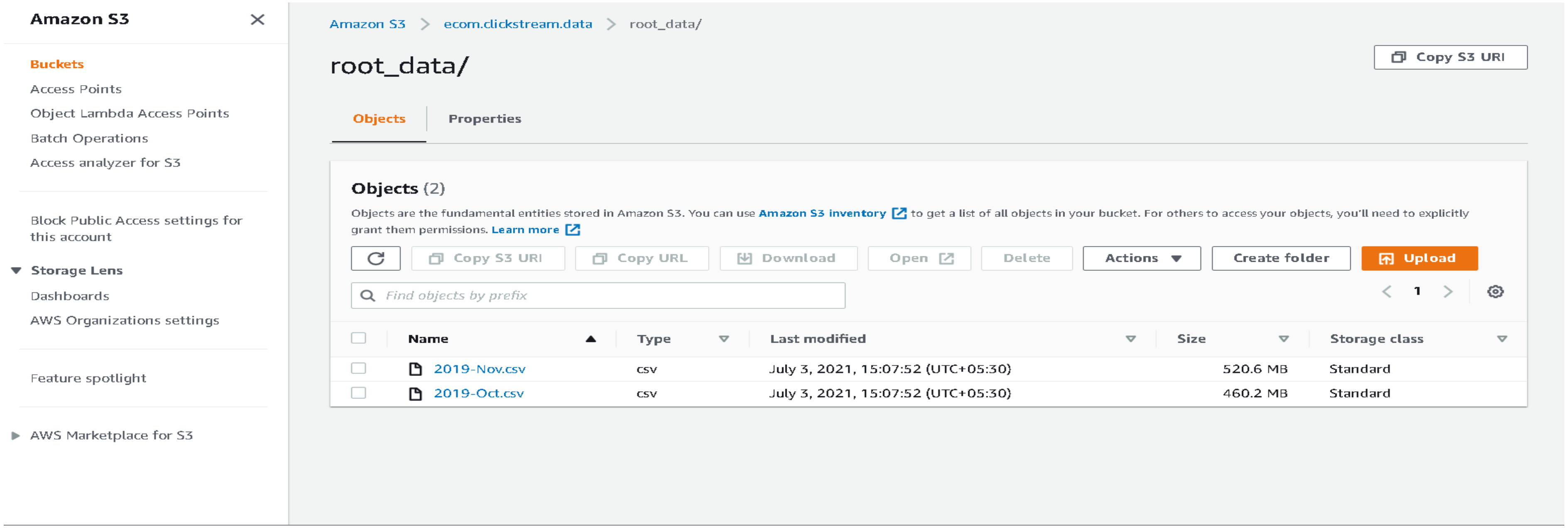
ID	Status	Node type & name	Instance type	Instance count	Purchasing opt
[redacted]	Running	CORE Core - 2	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB	1 Instances Resize	On-demand
[redacted]	Running	MASTER Master - 1	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB	1 Instances	On-demand

Cluster Scaling Policy

No scaling enabled

➢ Loading the data from S3 bucket to HDFS

- Create S3 bucket containing the two CSV files.



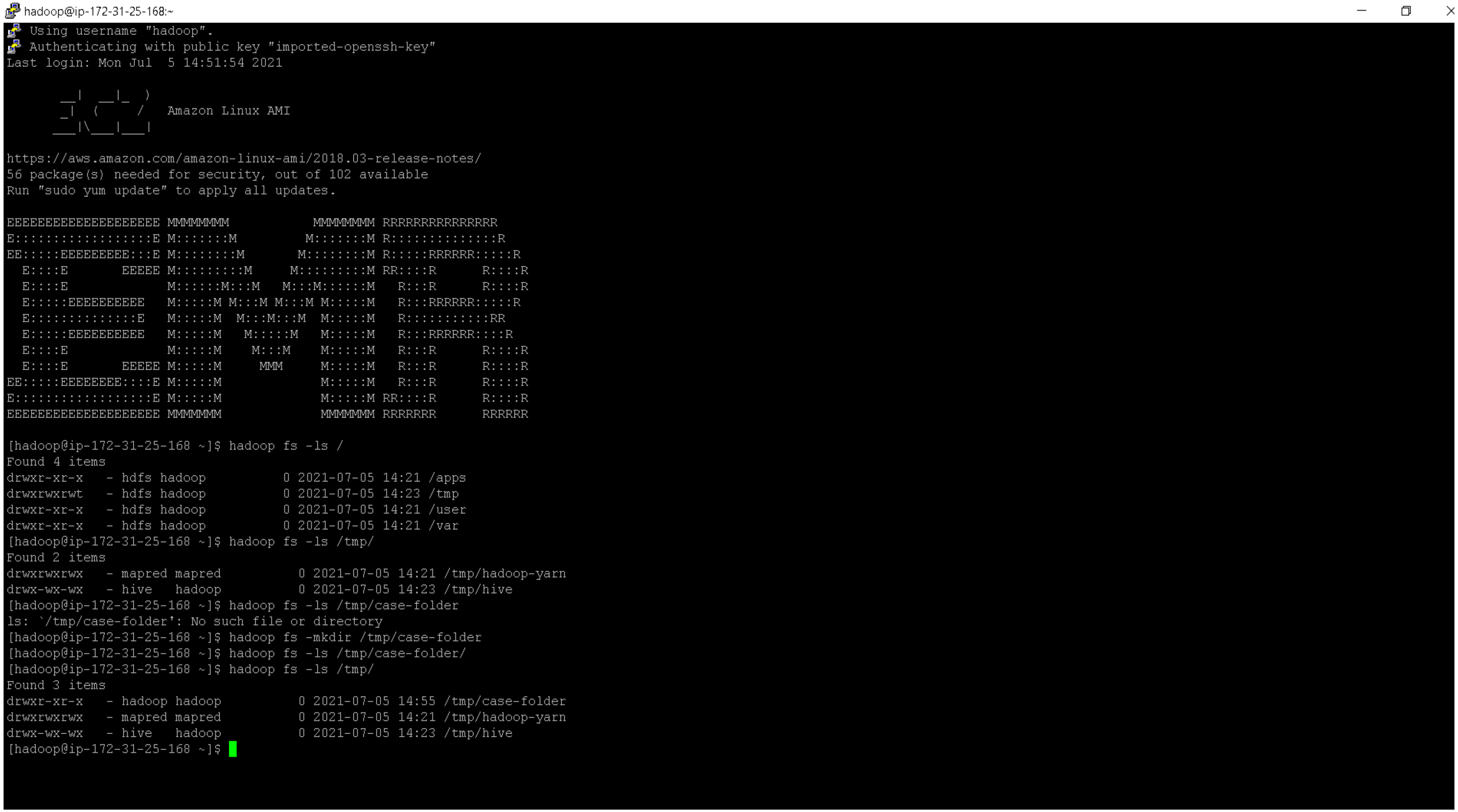
root_data/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	2019-Nov.csv	csv	July 3, 2021, 15:07:52 (UTC+05:30)	520.6 MB	Standard
□	2019-Oct.csv	csv	July 3, 2021, 15:07:52 (UTC+05:30)	460.2 MB	Standard

- Create a folder/ directory in the ‘tmp’ folder of HDFS named as ‘case-folder’.

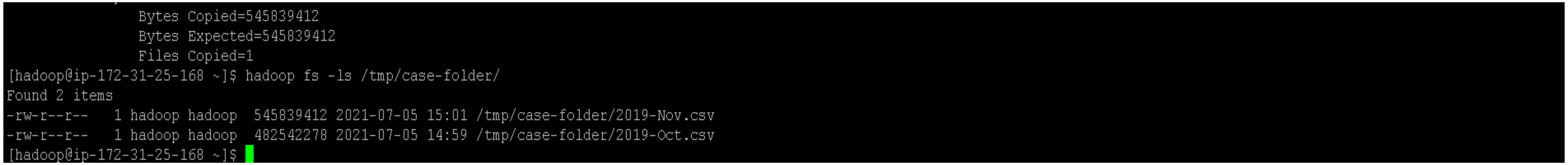


```
 Using username "hadoop".
 Authenticating with public key "imported-openssh-key"
Last login: Mon Jul  5 14:51:54 2021

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
56 package(s) needed for security, out of 102 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-25-168 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2021-07-05 14:21 /apps
drwxrwxrwt   - hdfs hadoop          0 2021-07-05 14:23 /tmp
drwxr-xr-x   - hdfs hadoop          0 2021-07-05 14:21 /user
drwxr-xr-x   - hdfs hadoop          0 2021-07-05 14:21 /var
[hadoop@ip-172-31-25-168 ~]$ hadoop fs -ls /tmp/
Found 2 items
drwxrwxrwx   - mapred mapred          0 2021-07-05 14:21 /tmp/hadoop-yarn
drwx-wx-wx   - hive   hadoop          0 2021-07-05 14:23 /tmp/hive
[hadoop@ip-172-31-25-168 ~]$ hadoop fs -ls /tmp/case-folder
ls: `/tmp/case-folder': No such file or directory
[hadoop@ip-172-31-25-168 ~]$ hadoop fs -mkdir /tmp/case-folder
[hadoop@ip-172-31-25-168 ~]$ hadoop fs -ls /tmp/case-folder/
[hadoop@ip-172-31-25-168 ~]$ hadoop fs -ls /tmp/
Found 3 items
drwxr-xr-x   - hadoop hadoop          0 2021-07-05 14:55 /tmp/case-folder
drwxrwxrwx   - mapred mapred          0 2021-07-05 14:21 /tmp/hadoop-yarn
drwx-wx-wx   - hive   hadoop          0 2021-07-05 14:23 /tmp/hive
[hadoop@ip-172-31-25-168 ~]$
```

- Load the data from S3 bucket into the HDFS

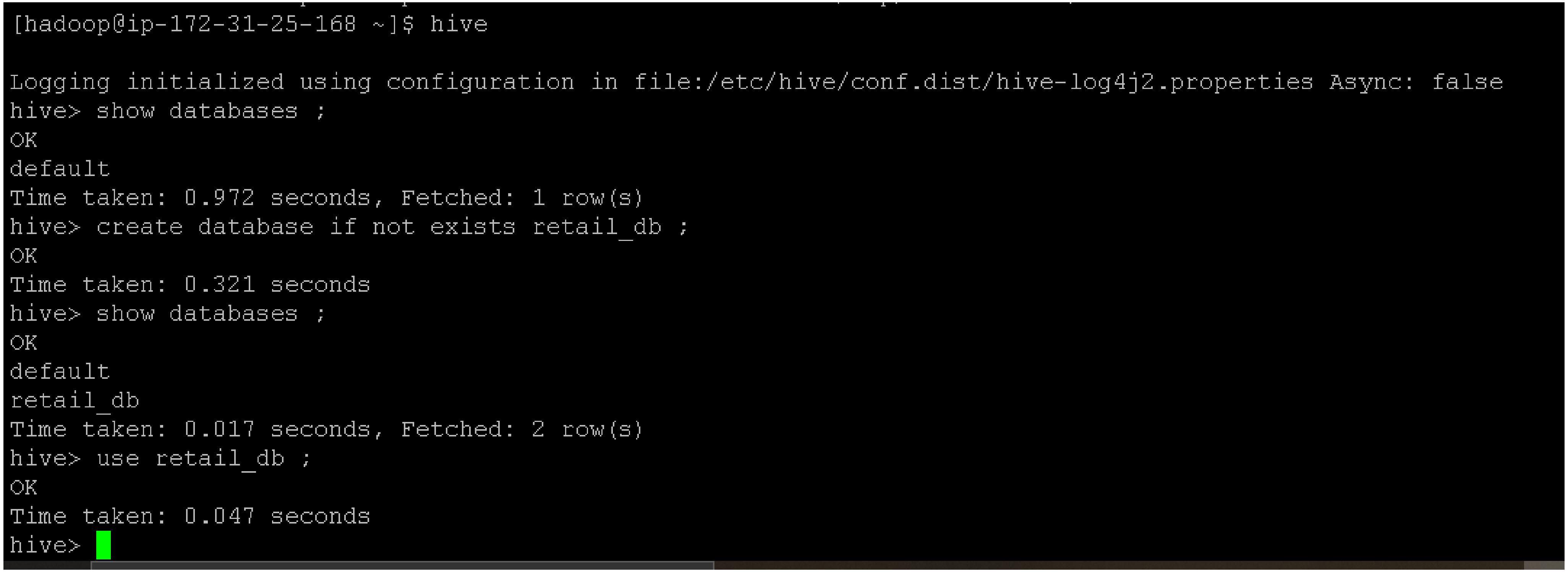
hadoop distcp s3n://ecom.clickstream.data/root_data/2019-Oct.csv /tmp/case-folder/2019-Oct.csv

hadoop distcp s3n://ecom.clickstream.data/root_data/2019-Nov.csv /tmp/case-folder/2019-Nov.csv



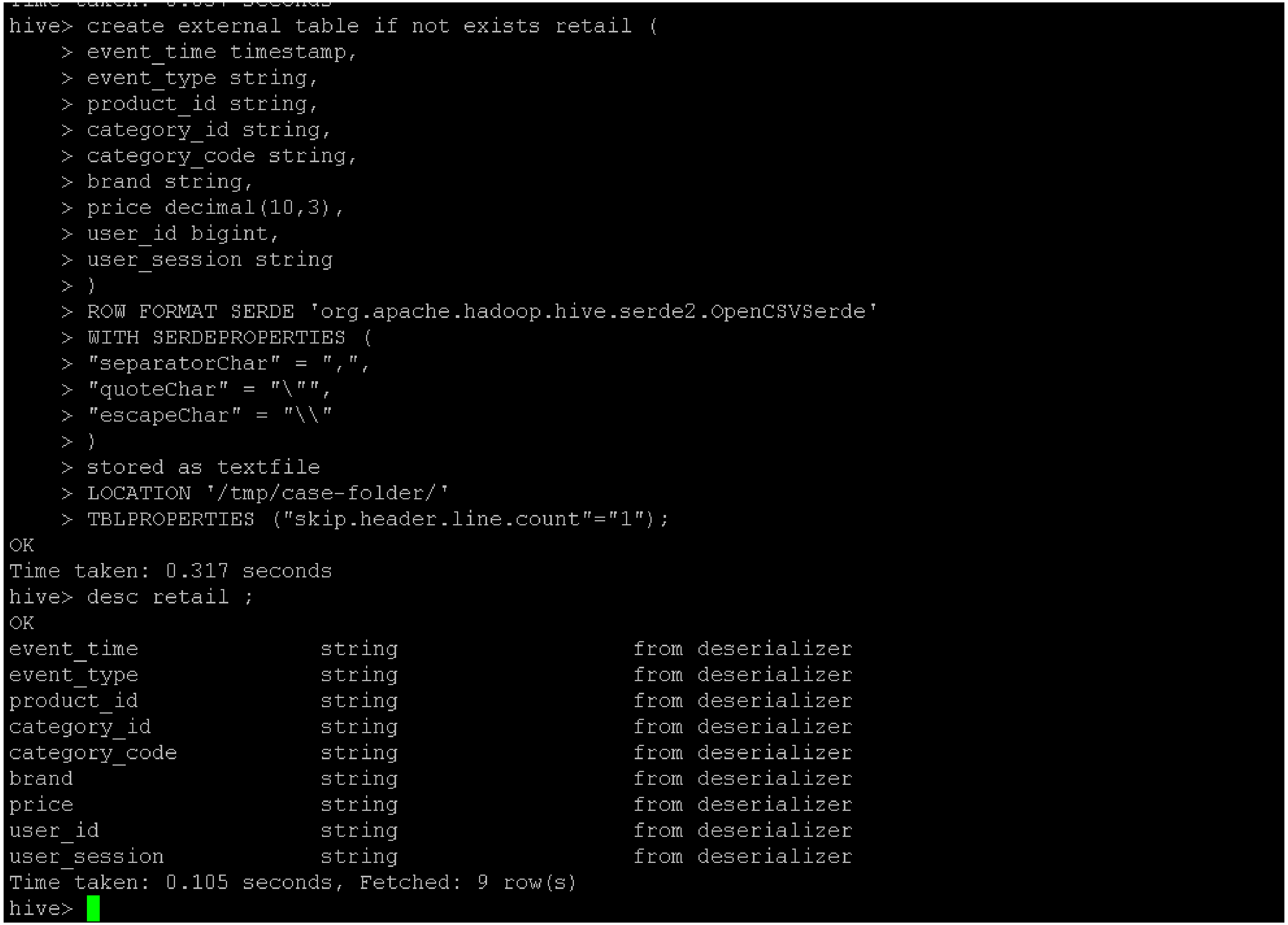
```
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
[hadoop@ip-172-31-25-168 ~]$ hadoop fs -ls /tmp/case-folder/
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-07-05 15:01 /tmp/case-folder/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-07-05 14:59 /tmp/case-folder/2019-Oct.csv
[hadoop@ip-172-31-25-168 ~]$
```

➢ **Creating the databases and required tables in Hive**

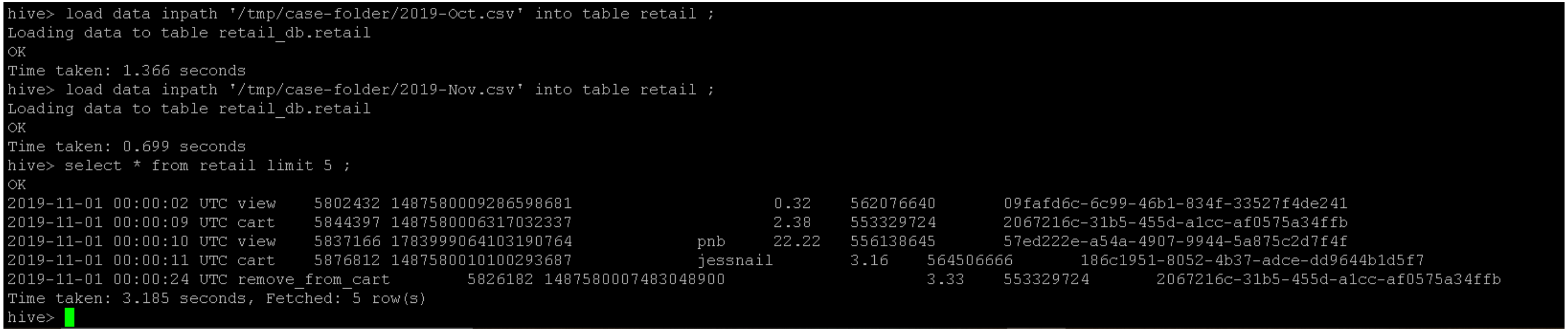


```
[hadoop@ip-172-31-25-168 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases ;
OK
default
Time taken: 0.972 seconds, Fetched: 1 row(s)
hive> create database if not exists retail_db ;
OK
Time taken: 0.321 seconds
hive> show databases ;
OK
default
retail_db
Time taken: 0.017 seconds, Fetched: 2 row(s)
hive> use retail_db ;
OK
Time taken: 0.047 seconds
hive>
```

- Creating retail table with the appropriate columns, checking schema and loading data into it

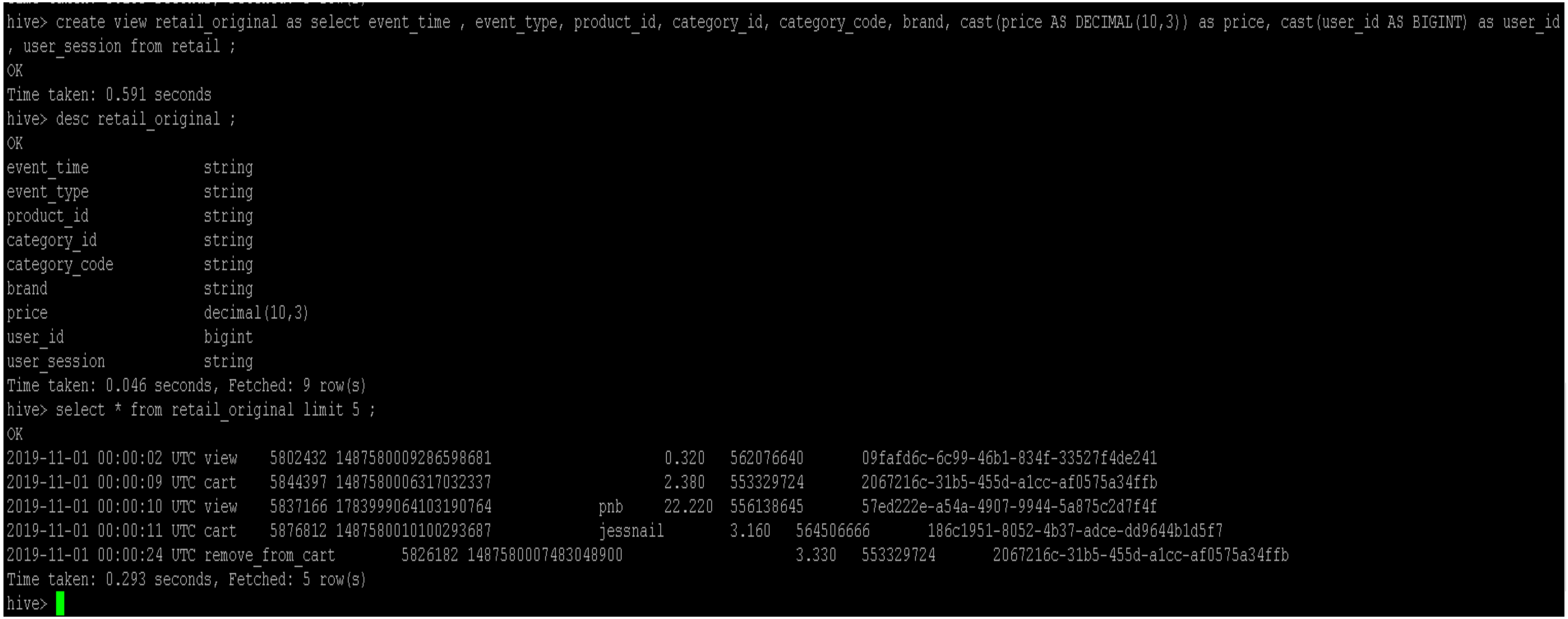


```
hive> create external table if not exists retail (
    > event_time timestamp,
    > event_type string,
    > product_id string,
    > category_id string,
    > category_code string,
    > brand string,
    > price decimal(10,3),
    > user_id bigint,
    > user_session string
    > )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    > "separatorChar" = ",",
    > "quoteChar" = "\"",
    > "escapeChar" = "\\"
    > )
    > stored as textfile
    > LOCATION '/tmp/case-folder/'
    > TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.317 seconds
hive> desc retail ;
OK
event_time              string                  from deserializer
event_type              string                  from deserializer
product_id              string                  from deserializer
category_id             string                  from deserializer
category_code           string                  from deserializer
brand                   string                  from deserializer
price                   string                  from deserializer
user_id                 string                  from deserializer
user_session            string                  from deserializer
Time taken: 0.105 seconds, Fetched: 9 row(s)
hive>
```



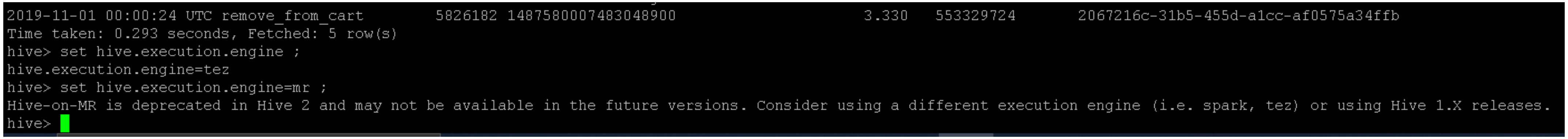
```
hive> load data inpath '/tmp/case-folder/2019-Oct.csv' into table retail ;
Loading data to table retail_db.retail
OK
Time taken: 1.366 seconds
hive> load data inpath '/tmp/case-folder/2019-Nov.csv' into table retail ;
Loading data to table retail_db.retail
OK
Time taken: 0.699 seconds
hive> select * from retail 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     22.22   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: 3.185 seconds, Fetched: 5 row(s)
hive>
```

- Creating view ‘retail_original’ with the columns as per the original datatype and checking schema



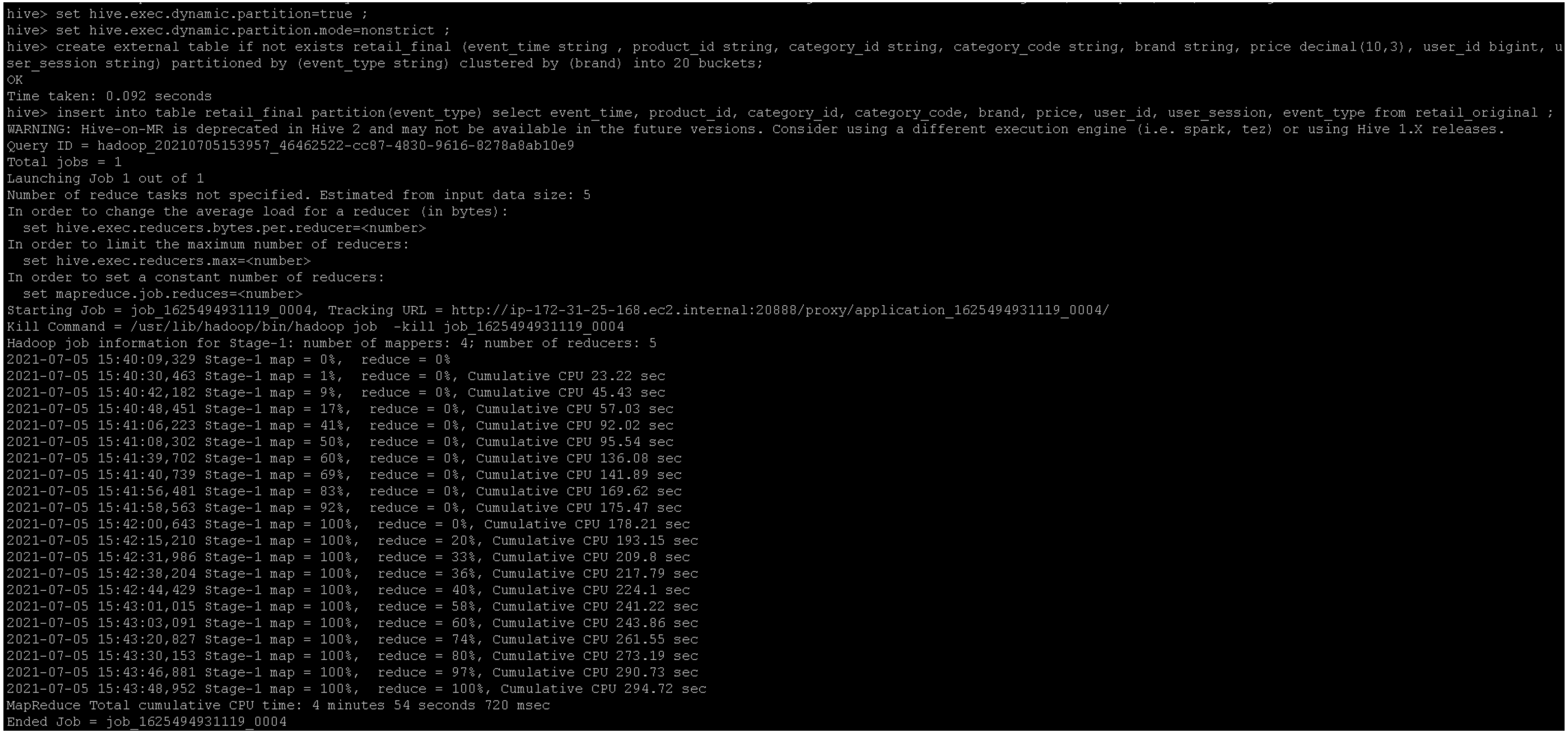
```
hive> create view retail_original as select event_time , event_type, product_id, category_id, category_code, brand, cast(price AS DECIMAL(10,3)) as price, cast(user_id AS BIGINT) as user_id
, user_session from retail ;
OK
Time taken: 0.591 seconds
hive> desc retail_original ;
OK
event_time              string
event_type              string
product_id              string
category_id             string
category_code           string
brand                   string
price                   decimal(10,3)
user_id                 bigint
user_session            string
Time taken: 0.046 seconds, Fetched: 9 row(s)
hive> select * from retail_original limit 5 ;
OK
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681                     0.320   562076640       09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337                     2.380   553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764             pnb     22.220  556138645       57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 1487580010100293687             jessnail        3.160   564506666       186c1951-8052-4b37-adce-dd9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.330   553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 0.293 seconds, Fetched: 5 row(s)
hive>
```

- Setting Hive execution engine to MapReduce

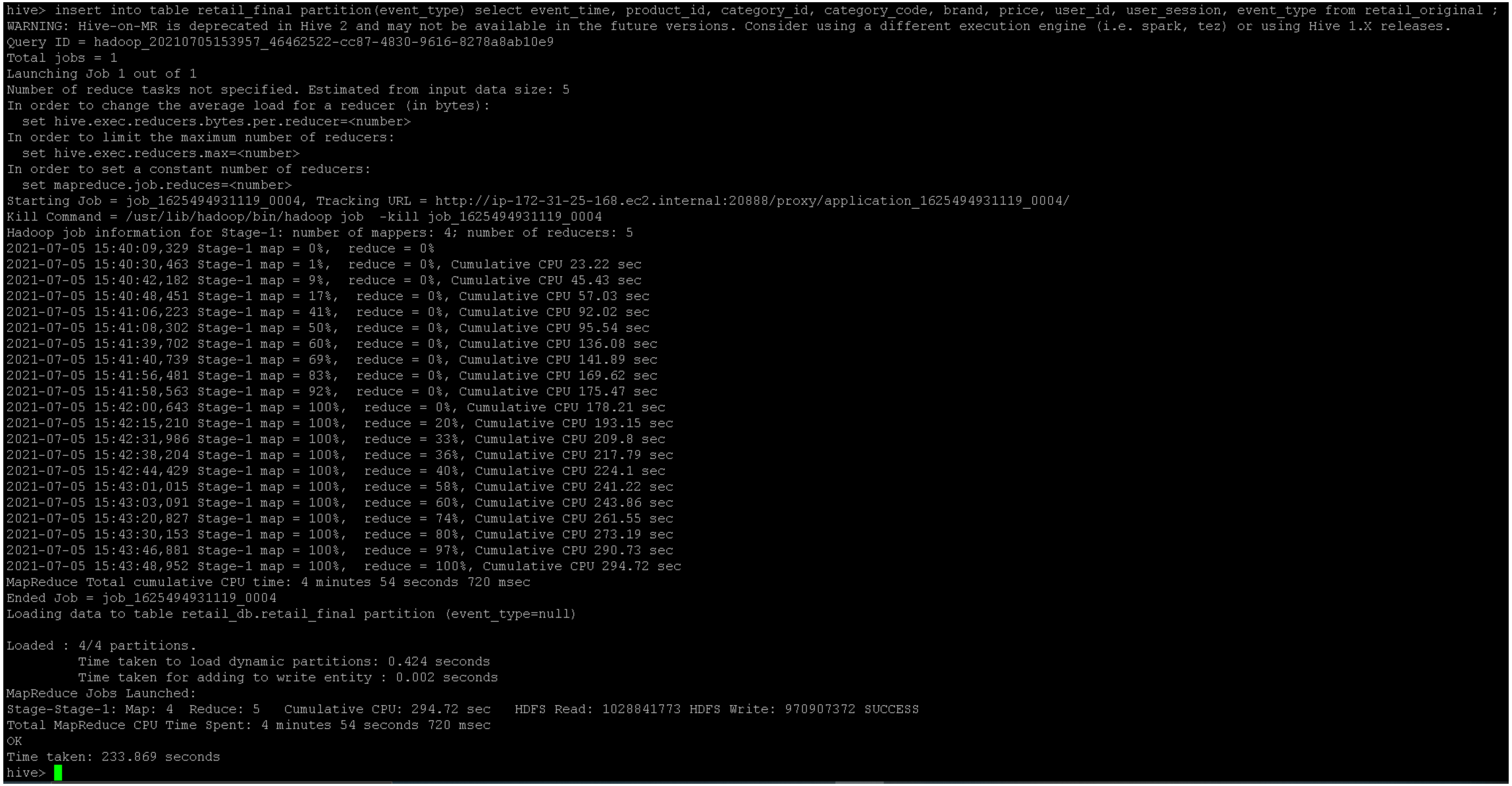


```
2019-11-01 00:00:24 UTC remove_from_cart        5826182 1487580007483048900                     3.330   553329724       2067216c-31b5-455d-a1cc-af0575a34ffb
Time taken: 0.293 seconds, Fetched: 5 row(s)
hive> set hive.execution.engine ;
hive.execution.engine=tez
hive> set hive.execution.engine=mr ;
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> 
```

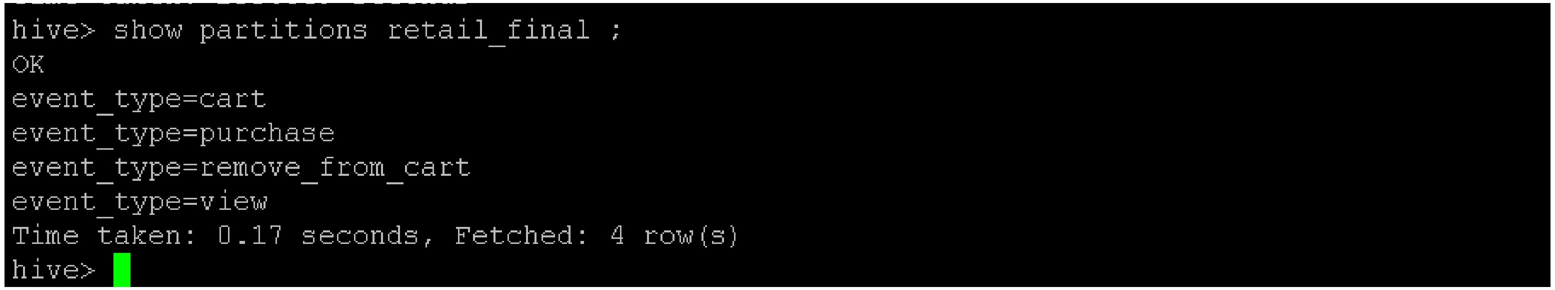
- Creating ‘retail_final’ table with partition on event_type and buckets on brand for optimization and loading data into it



```
hive> set hive.exec.dynamic.partition=true ;
hive> set hive.exec.dynamic.partition.mode=nonstrict ;
hive> create external table if not exists retail_final (event_time string , product_id string, category_id string, category_code string, brand string, price decimal(10,3), user_id bigint, u
ser_session string) partitioned by (event_type string) clustered by (brand) into 20 buckets;
OK
Time taken: 0.092 seconds
hive> insert into table retail_final partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from retail_original ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705153957_46462522-cc87-4830-9616-8278a8ab10e9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 5
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0004, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0004
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 5
2021-07-05 15:40:09,329 Stage-1 map = 0%,  reduce = 0%
2021-07-05 15:40:30,463 Stage-1 map = 1%,  reduce = 0%, Cumulative CPU 23.22 sec
2021-07-05 15:40:42,182 Stage-1 map = 9%,  reduce = 0%, Cumulative CPU 45.43 sec
2021-07-05 15:40:48,451 Stage-1 map = 17%,  reduce = 0%, Cumulative CPU 57.03 sec
2021-07-05 15:41:06,223 Stage-1 map = 41%,  reduce = 0%, Cumulative CPU 92.02 sec
2021-07-05 15:41:08,302 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 95.54 sec
2021-07-05 15:41:39,702 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 136.08 sec
2021-07-05 15:41:40,739 Stage-1 map = 69%,  reduce = 0%, Cumulative CPU 141.89 sec
2021-07-05 15:41:56,481 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 169.62 sec
2021-07-05 15:41:58,563 Stage-1 map = 92%,  reduce = 0%, Cumulative CPU 175.47 sec
2021-07-05 15:42:00,643 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 178.21 sec
2021-07-05 15:42:15,210 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 193.15 sec
2021-07-05 15:42:31,986 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 209.8 sec
2021-07-05 15:42:38,204 Stage-1 map = 100%,  reduce = 36%, Cumulative CPU 217.79 sec
2021-07-05 15:42:44,429 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 224.1 sec
2021-07-05 15:43:01,015 Stage-1 map = 100%,  reduce = 58%, Cumulative CPU 241.22 sec
2021-07-05 15:43:03,091 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 243.86 sec
2021-07-05 15:43:20,827 Stage-1 map = 100%,  reduce = 74%, Cumulative CPU 261.55 sec
2021-07-05 15:43:30,153 Stage-1 map = 100%,  reduce = 80%, Cumulative CPU 273.19 sec
2021-07-05 15:43:46,881 Stage-1 map = 100%,  reduce = 97%, Cumulative CPU 290.73 sec
2021-07-05 15:43:48,952 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 294.72 sec
MapReduce Total cumulative CPU time: 4 minutes 54 seconds 720 msec
Ended Job = job_1625494931119_0004
```



```
hive> insert into table retail_final partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from retail_original ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705153957_46462522-cc87-4830-9616-8278a8ab10e9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 5
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0004, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0004
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 5
2021-07-05 15:40:09,329 Stage-1 map = 0%,  reduce = 0%
2021-07-05 15:40:30,463 Stage-1 map = 1%,  reduce = 0%, Cumulative CPU 23.22 sec
2021-07-05 15:40:42,182 Stage-1 map = 9%,  reduce = 0%, Cumulative CPU 45.43 sec
2021-07-05 15:40:48,451 Stage-1 map = 17%,  reduce = 0%, Cumulative CPU 57.03 sec
2021-07-05 15:41:06,223 Stage-1 map = 41%,  reduce = 0%, Cumulative CPU 92.02 sec
2021-07-05 15:41:08,302 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 95.54 sec
2021-07-05 15:41:39,702 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 136.08 sec
2021-07-05 15:41:40,739 Stage-1 map = 69%,  reduce = 0%, Cumulative CPU 141.89 sec
2021-07-05 15:41:56,481 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 169.62 sec
2021-07-05 15:41:58,563 Stage-1 map = 92%,  reduce = 0%, Cumulative CPU 175.47 sec
2021-07-05 15:42:00,643 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 178.21 sec
2021-07-05 15:42:15,210 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 193.15 sec
2021-07-05 15:42:31,986 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 209.8 sec
2021-07-05 15:42:38,204 Stage-1 map = 100%,  reduce = 36%, Cumulative CPU 217.79 sec
2021-07-05 15:42:44,429 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 224.1 sec
2021-07-05 15:43:01,015 Stage-1 map = 100%,  reduce = 58%, Cumulative CPU 241.22 sec
2021-07-05 15:43:03,091 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 243.86 sec
2021-07-05 15:43:20,827 Stage-1 map = 100%,  reduce = 74%, Cumulative CPU 261.55 sec
2021-07-05 15:43:30,153 Stage-1 map = 100%,  reduce = 80%, Cumulative CPU 273.19 sec
2021-07-05 15:43:46,881 Stage-1 map = 100%,  reduce = 97%, Cumulative CPU 290.73 sec
2021-07-05 15:43:48,952 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 294.72 sec
MapReduce Total cumulative CPU time: 4 minutes 54 seconds 720 msec
Ended Job = job_1625494931119_0004
Loading data to table retail_db.retail_final partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.424 seconds
         Time taken for adding to write entity : 0.002 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4  Reduce: 5   Cumulative CPU: 294.72 sec   HDFS Read: 1028841773 HDFS Write: 970907372 SUCCESS
Total MapReduce CPU Time Spent: 4 minutes 54 seconds 720 msec
OK
Time taken: 233.869 seconds
hive>
```



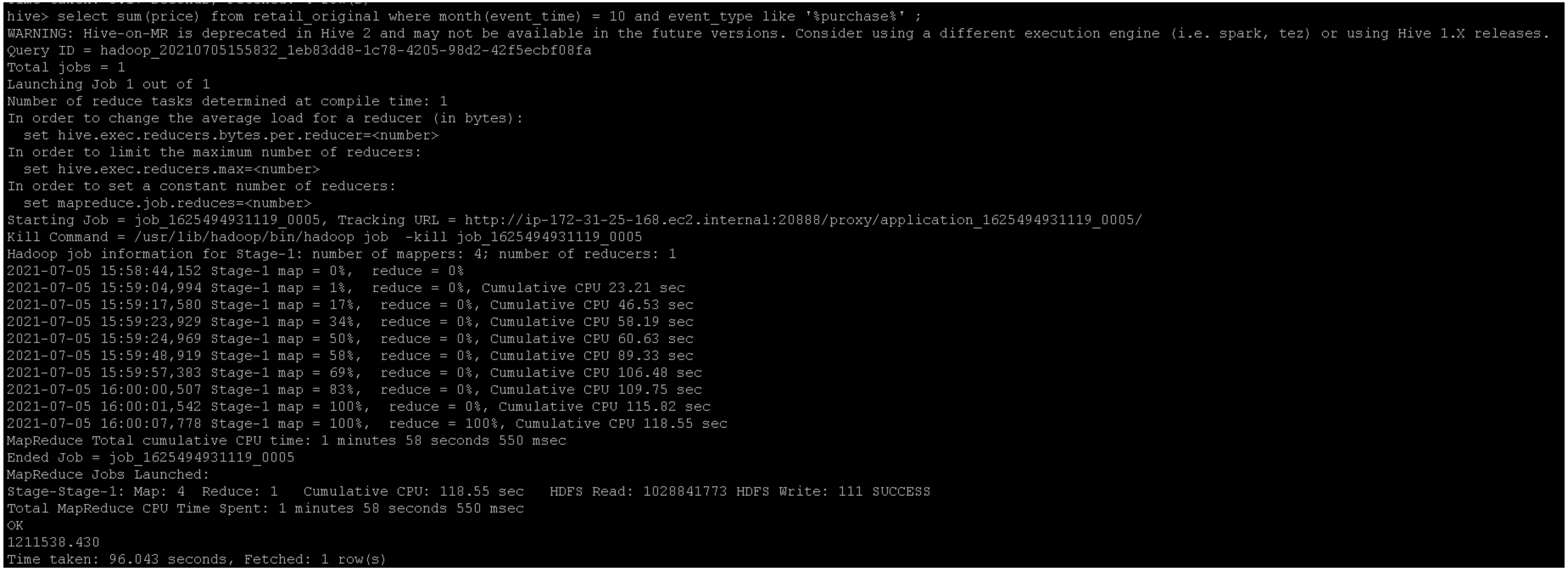
```
hive> show partitions retail_final ;
OK
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.17 seconds, Fetched: 4 row(s)
hive>
```

➢ Hive Queries on the given data

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

- Query on non-optimized table ‘retail_original’

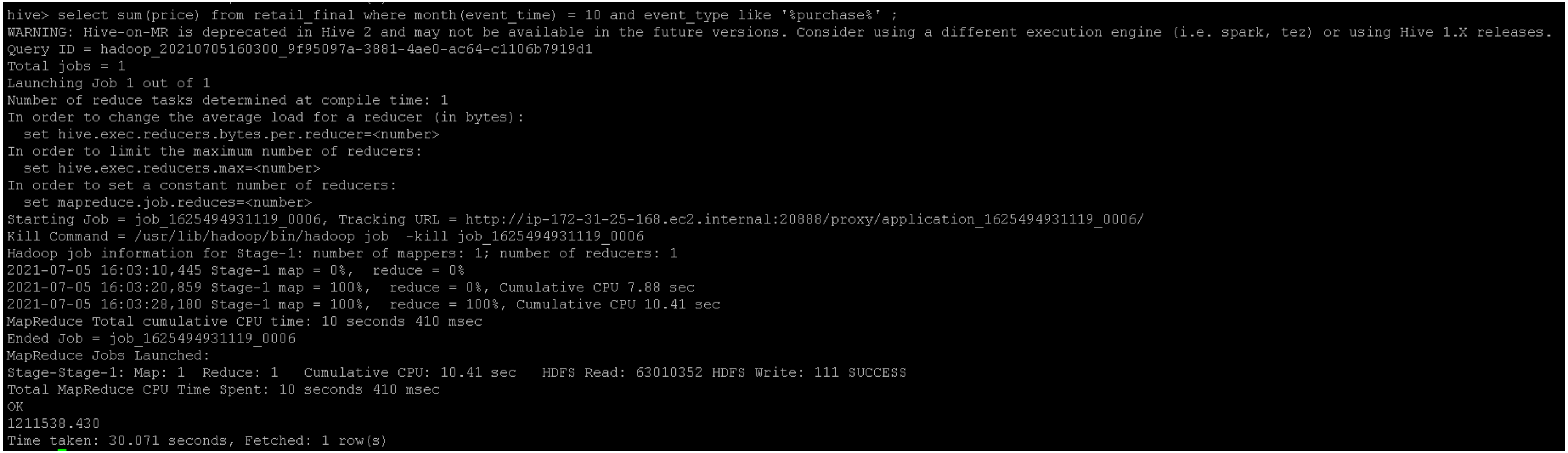
select sum(price) from retail_original where month(event_time) = 10 and event_type like '%purchase%' ;



```
hive> select sum(price) from retail_original where month(event_time) = 10 and event_type like '%purchase%' ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705155832_1eb83dd8-1c78-4205-98d2-42f5ecbf08fa
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0005, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0005
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 1
2021-07-05 15:58:44,152 Stage-1 map = 0%,  reduce = 0%
2021-07-05 15:59:04,994 Stage-1 map = 1%,  reduce = 0%, Cumulative CPU 23.21 sec
2021-07-05 15:59:17,580 Stage-1 map = 17%,  reduce = 0%, Cumulative CPU 46.53 sec
2021-07-05 15:59:23,929 Stage-1 map = 34%,  reduce = 0%, Cumulative CPU 58.19 sec
2021-07-05 15:59:24,969 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 60.63 sec
2021-07-05 15:59:48,919 Stage-1 map = 58%,  reduce = 0%, Cumulative CPU 89.33 sec
2021-07-05 15:59:57,383 Stage-1 map = 69%,  reduce = 0%, Cumulative CPU 106.48 sec
2021-07-05 16:00:00,507 Stage-1 map = 83%,  reduce = 0%, Cumulative CPU 109.75 sec
2021-07-05 16:00:01,542 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 115.82 sec
2021-07-05 16:00:07,778 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 118.55 sec
MapReduce Total cumulative CPU time: 1 minutes 58 seconds 550 msec
Ended Job = job_1625494931119_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4  Reduce: 1   Cumulative CPU: 118.55 sec   HDFS Read: 1028841773 HDFS Write: 111 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 58 seconds 550 msec
OK
1211538.430
Time taken: 96.043 seconds, Fetched: 1 row(s)
```

- Query on optimized table ‘retail_final’

select sum(price) from retail_final where month(event_time) = 10 and event_type like '%purchase%' ;

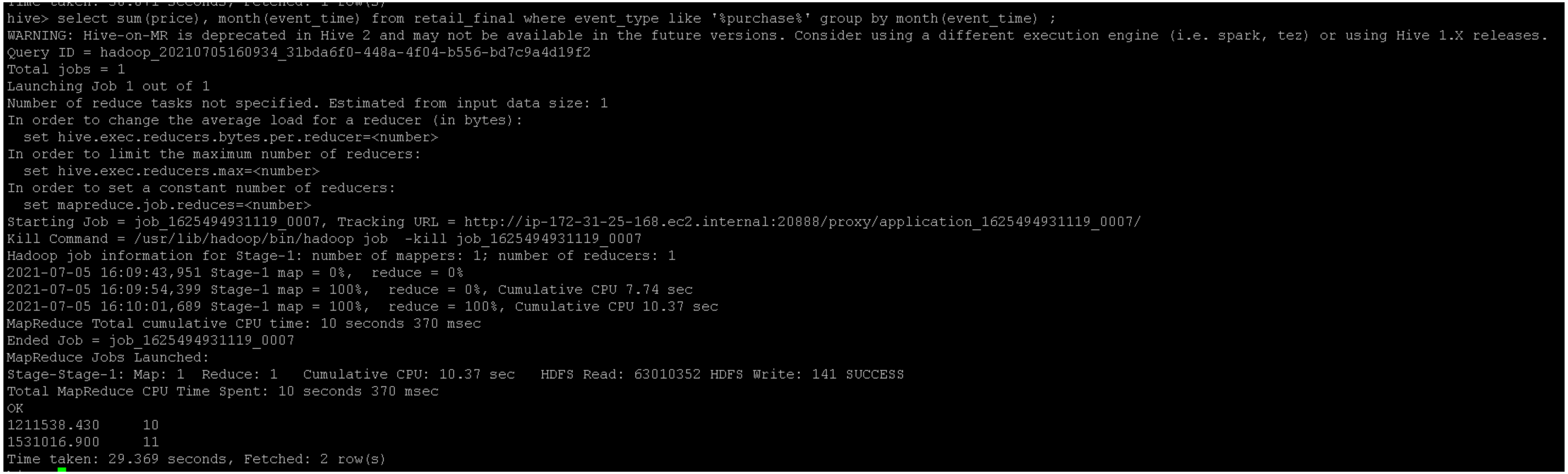


```
hive> select sum(price) from retail_final where month(event_time) = 10 and event_type like '%purchase%' ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705160300_9f95097a-3881-4ae0-ac64-c1106b7919d1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0006, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-05 16:03:10,445 Stage-1 map = 0%,  reduce = 0%
2021-07-05 16:03:20,859 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.88 sec
2021-07-05 16:03:28,180 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.41 sec
MapReduce Total cumulative CPU time: 10 seconds 410 msec
Ended Job = job_1625494931119_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.41 sec   HDFS Read: 63010352 HDFS Write: 111 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 410 msec
OK
1211538.430
Time taken: 30.071 seconds, Fetched: 1 row(s)
```

Therefore, querying on the Non-partitioned ‘retail _original’ table gave result in 96.043 seconds and querying on Partitioned and bucketed table ‘retail_final’ table gave result in 30.071 seconds.

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

select sum(price), month(event_time) from retail_final where event_type like '%purchase%' group by month(event_time) ;



```
hive> select sum(price), month(event_time) from retail_final where event_type like '%purchase%' group by month(event_time) ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705160934_31bda6f0-448a-4f04-b556-bd7c9a4d19f2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0007, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-05 16:09:43,951 Stage-1 map = 0%,  reduce = 0%
2021-07-05 16:09:54,399 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.74 sec
2021-07-05 16:10:01,689 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.37 sec
MapReduce Total cumulative CPU time: 10 seconds 370 msec
Ended Job = job_1625494931119_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.37 sec   HDFS Read: 63010352 HDFS Write: 141 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 370 msec
OK
1211538.430     10
1531016.900     11
Time taken: 29.369 seconds, Fetched: 2 row(s)
```

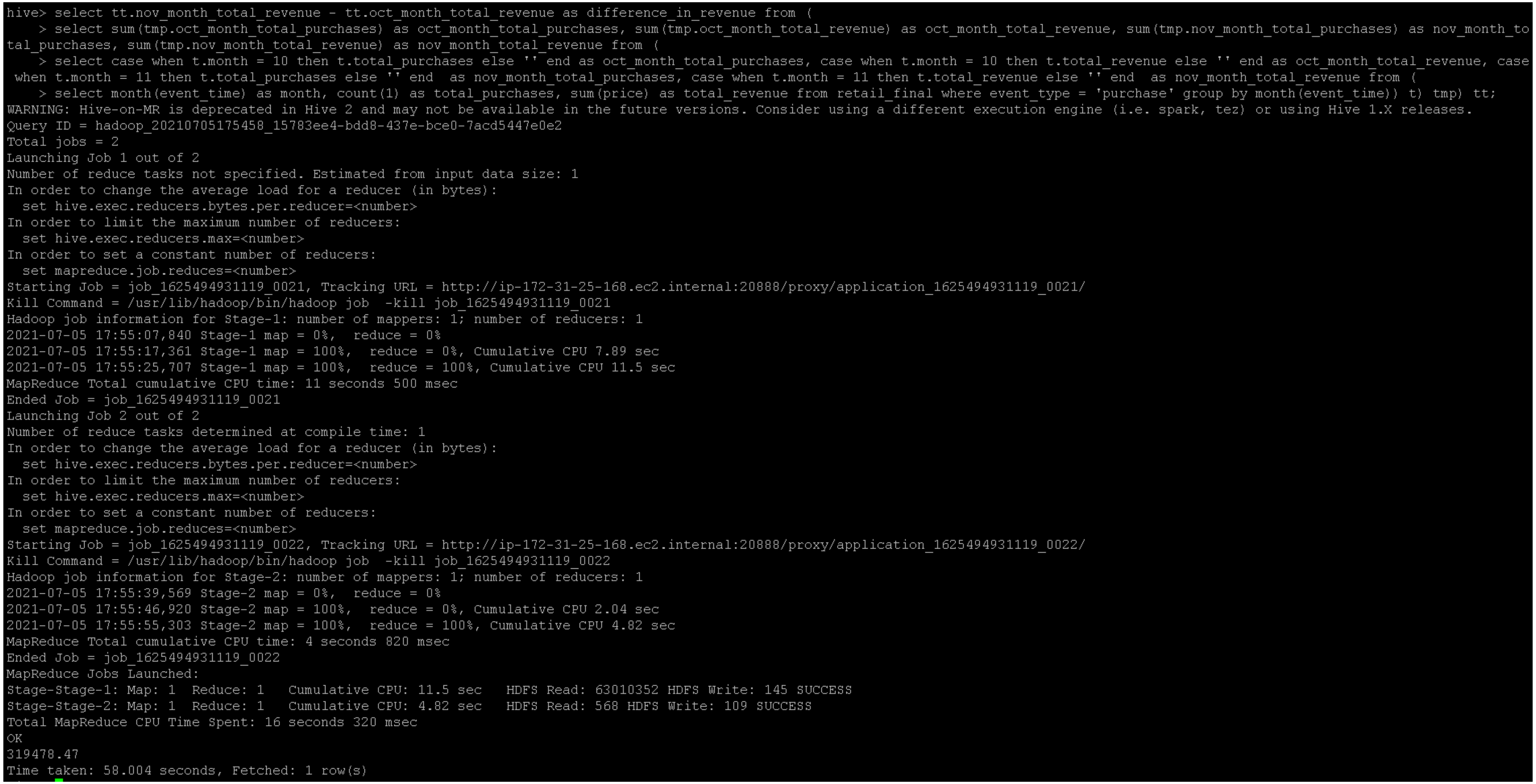
3. ***Write a query to find the change in revenue generated due to purchases from October to November.***

select tt.nov_month_total_revenue - tt.oct_month_total_revenue as difference_in_revenue from (

select sum(tmp.oct_month_total_purchases) as oct_month_total_purchases, sum(tmp.oct_month_total_revenue) as oct_month_total_revenue, sum(tmp.nov_month_total_purchases) as nov_month_total_purchases, sum(tmp.nov_month_total_revenue) as nov_month_total_revenue from (

select case when t.month = 10 then t.total_purchases else '' end as oct_month_total_purchases, case when t.month = 10 then t.total_revenue else '' end as oct_month_total_revenue, case when t.month = 11 then t.total_purchases else '' end as nov_month_total_purchases, case when t.month = 11 then t.total_revenue else '' end as nov_month_total_revenue from (

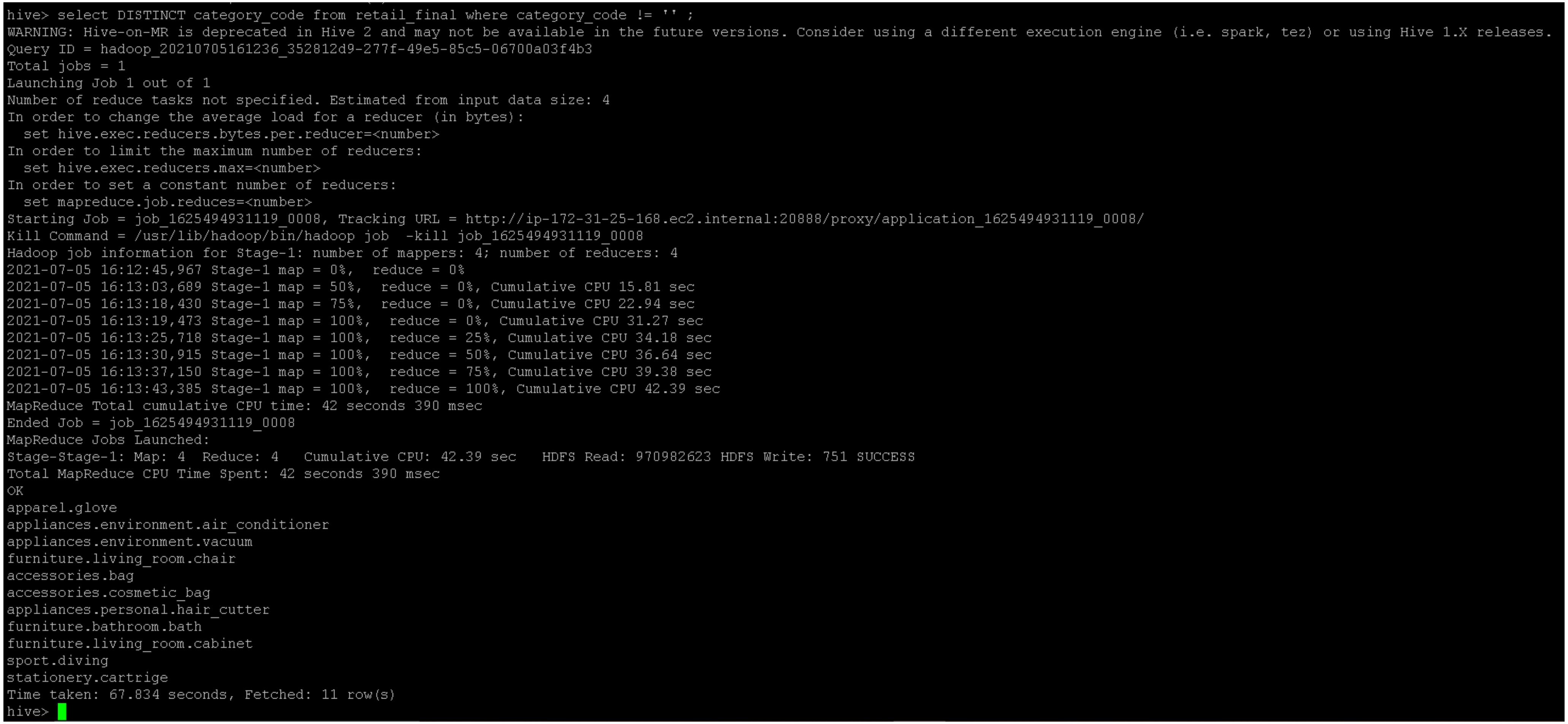
select month(event_time) as month, count(1) as total_purchases, sum(price) as total_revenue from retail_final where event_type = 'purchase' group by month(event_time)) t) tmp) tt;



```
hive> select tt.nov_month_total_revenue - tt.oct_month_total_revenue as difference_in_revenue from (
    > select sum(tmp.oct_month_total_purchases) as oct_month_total_purchases, sum(tmp.oct_month_total_revenue) as oct_month_total_revenue, sum(tmp.nov_month_total_purchases) as nov_month_to
tal_purchases, sum(tmp.nov_month_total_revenue) as nov_month_total_revenue from (
    > select case when t.month = 10 then t.total_purchases else '' end as oct_month_total_purchases, case when t.month = 10 then t.total_revenue else '' end as oct_month_total_revenue, case
 when t.month = 11 then t.total_purchases else '' end  as nov_month_total_purchases, case when t.month = 11 then t.total_revenue else '' end  as nov_month_total_revenue from (
    > select month(event_time) as month, count(1) as total_purchases, sum(price) as total_revenue from retail_final where event_type = 'purchase' group by month(event_time)) t) tmp) tt;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705175458_15783ee4-bdd8-437e-bce0-7acd5447e0e2
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0021, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-05 17:55:07,840 Stage-1 map = 0%,  reduce = 0%
2021-07-05 17:55:17,361 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.89 sec
2021-07-05 17:55:25,707 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.5 sec
MapReduce Total cumulative CPU time: 11 seconds 500 msec
Ended Job = job_1625494931119_0021
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0022, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0022/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0022
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-07-05 17:55:39,569 Stage-2 map = 0%,  reduce = 0%
2021-07-05 17:55:46,920 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.04 sec
2021-07-05 17:55:55,303 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.82 sec
MapReduce Total cumulative CPU time: 4 seconds 820 msec
Ended Job = job_1625494931119_0022
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.5 sec   HDFS Read: 63010352 HDFS Write: 145 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.82 sec   HDFS Read: 568 HDFS Write: 109 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 320 msec
OK
319478.47
Time taken: 58.004 seconds, Fetched: 1 row(s)
```

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

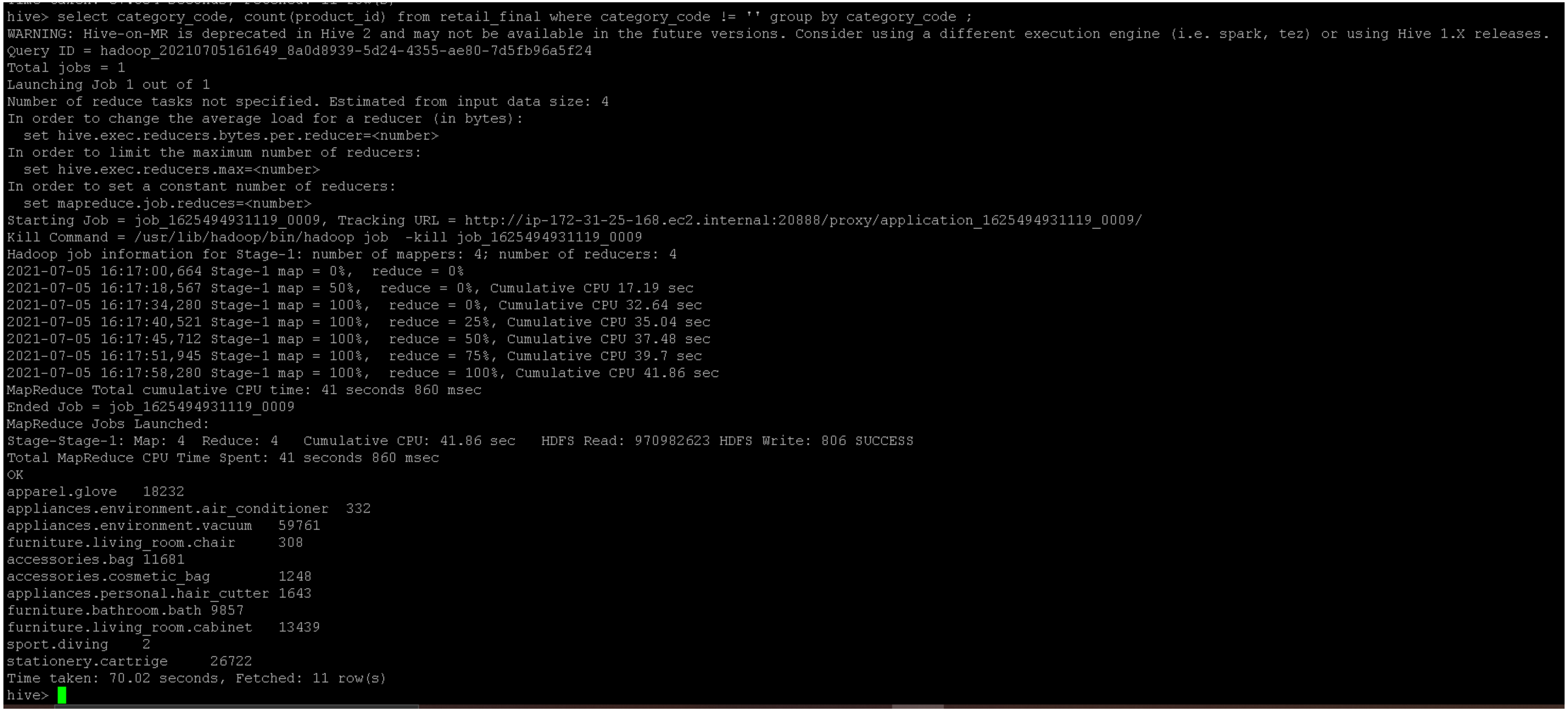
select DISTINCT category_code from retail_final where category_code != '' ;



```
hive> select DISTINCT category_code from retail_final where category_code != '' ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705161236_352812d9-277f-49e5-85c5-06700a03f4b3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0008, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0008
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 4
2021-07-05 16:12:45,967 Stage-1 map = 0%,  reduce = 0%
2021-07-05 16:13:03,689 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 15.81 sec
2021-07-05 16:13:18,430 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 22.94 sec
2021-07-05 16:13:19,473 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 31.27 sec
2021-07-05 16:13:25,718 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 34.18 sec
2021-07-05 16:13:30,915 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 36.64 sec
2021-07-05 16:13:37,150 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 39.38 sec
2021-07-05 16:13:43,385 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 42.39 sec
MapReduce Total cumulative CPU time: 42 seconds 390 msec
Ended Job = job_1625494931119_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4  Reduce: 4   Cumulative CPU: 42.39 sec   HDFS Read: 970982623 HDFS Write: 751 SUCCESS
Total MapReduce CPU Time Spent: 42 seconds 390 msec
OK
apparel.glove
appliances.environment.air_conditioner
appliances.environment.vacuum
furniture.living_room.chair
accessories.bag
accessories.cosmetic_bag
appliances.personal.hair_cutter
furniture.bathroom.bath
furniture.living_room.cabinet
sport.diving
stationery.cartrige
Time taken: 67.834 seconds, Fetched: 11 row(s)
hive>
```

5. ***Find the total number of products available under each category.***

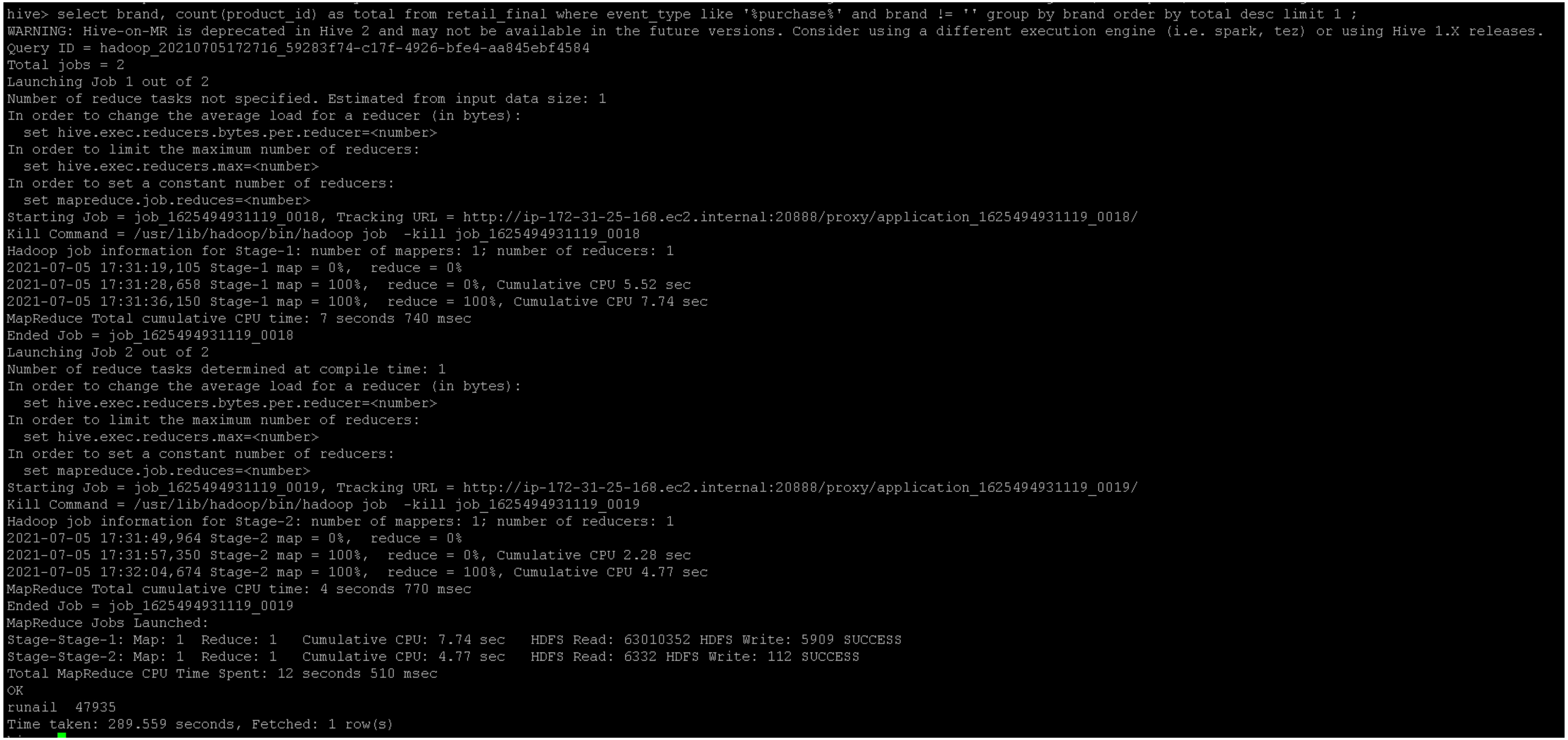
select category_code, count(product_id) from retail_final where category_code != '' group by category_code ;



```
hive> select category_code, count(product_id) from retail_final where category_code != '' group by category_code ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705161649_8a0d8939-5d24-4355-ae80-7d5fb96a5f24
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0009, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0009
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 4
2021-07-05 16:17:00,664 Stage-1 map = 0%,  reduce = 0%
2021-07-05 16:17:18,567 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 17.19 sec
2021-07-05 16:17:34,280 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 32.64 sec
2021-07-05 16:17:40,521 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 35.04 sec
2021-07-05 16:17:45,712 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 37.48 sec
2021-07-05 16:17:51,945 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 39.7 sec
2021-07-05 16:17:58,280 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 41.86 sec
MapReduce Total cumulative CPU time: 41 seconds 860 msec
Ended Job = job_1625494931119_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4  Reduce: 4   Cumulative CPU: 41.86 sec   HDFS Read: 970982623 HDFS Write: 806 SUCCESS
Total MapReduce CPU Time Spent: 41 seconds 860 msec
OK
apparel.glove   18232
appliances.environment.air_conditioner  332
appliances.environment.vacuum   59761
furniture.living_room.chair     308
accessories.bag 11681
accessories.cosmetic_bag        1248
appliances.personal.hair_cutter 1643
furniture.bathroom.bath 9857
furniture.living_room.cabinet   13439
sport.diving    2
stationery.cartrige     26722
Time taken: 70.02 seconds, Fetched: 11 row(s)
hive>
```

6. *Which brand had the maximum sales in October and November combined?*

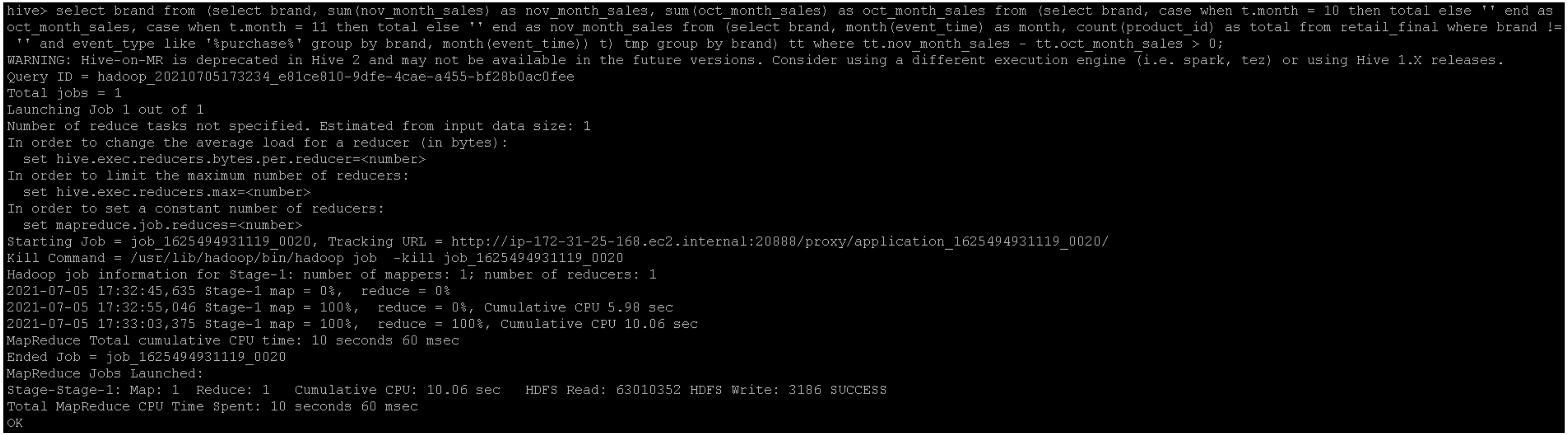
select brand, count(product_id) as total from retail_final where event_type like '%purchase%' and brand != '' group by brand order by total desc limit 1 ;



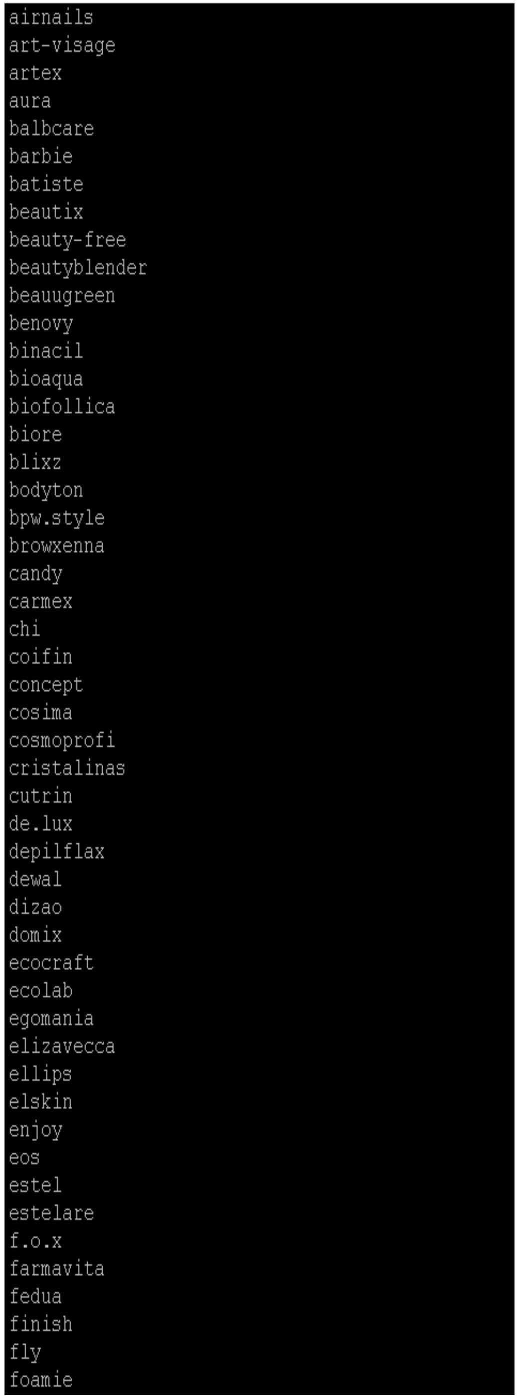
```
hive> select brand, count(product_id) as total from retail_final where event_type like '%purchase%' and brand != '' group by brand order by total desc limit 1 ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705172716_59283f74-c17f-4926-bfe4-aa845ebf4584
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0018, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0018
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-05 17:31:19,105 Stage-1 map = 0%,  reduce = 0%
2021-07-05 17:31:28,658 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.52 sec
2021-07-05 17:31:36,150 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.74 sec
MapReduce Total cumulative CPU time: 7 seconds 740 msec
Ended Job = job_1625494931119_0018
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0019, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0019
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-07-05 17:31:49,964 Stage-2 map = 0%,  reduce = 0%
2021-07-05 17:31:57,350 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.28 sec
2021-07-05 17:32:04,674 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.77 sec
MapReduce Total cumulative CPU time: 4 seconds 770 msec
Ended Job = job_1625494931119_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.74 sec   HDFS Read: 63010352 HDFS Write: 5909 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.77 sec   HDFS Read: 6332 HDFS Write: 112 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 510 msec
OK
runail  47935
Time taken: 289.559 seconds, Fetched: 1 row(s)
```

7. *Which brands increased their sales from October to November?*

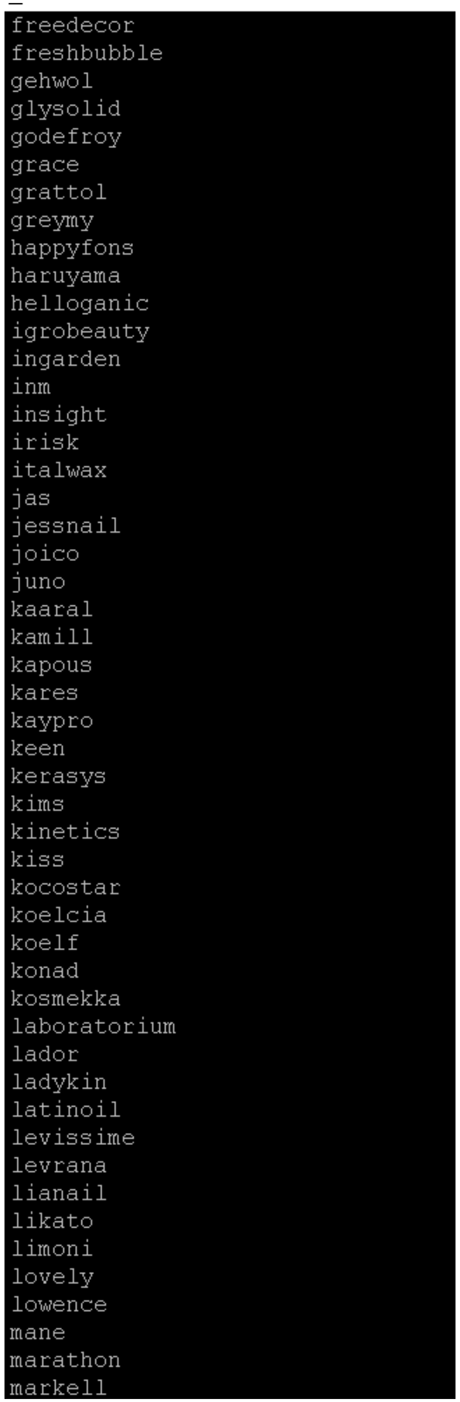
select brand from (select brand, sum(nov_month_sales) as nov_month_sales, sum(oct_month_sales) as oct_month_sales from (select brand, case when t.month = 10 then total else '' end as oct_month_sales, case when t.month = 11 then total else '' end as nov_month_sales from (select brand, month(event_time) as month, count(product_id) as total from retail_final where brand != '' and event_type like '%purchase%' group by brand, month(event_time)) t) tmp group by brand) tt where tt.nov_month_sales - tt.oct_month_sales > 0;



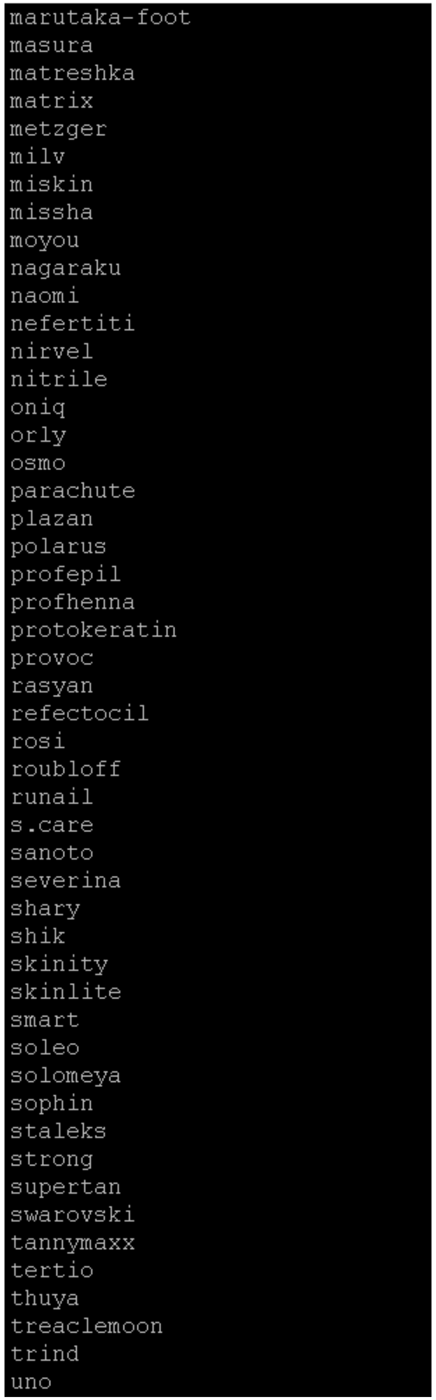
```
hive> select brand from (select brand, sum(nov_month_sales) as nov_month_sales, sum(oct_month_sales) as oct_month_sales from (select brand, case when t.month = 10 then total else '' end as
oct_month_sales, case when t.month = 11 then total else '' end as nov_month_sales from (select brand, month(event_time) as month, count(product_id) as total from retail_final where brand !=
 '' and event_type like '%purchase%' group by brand, month(event_time)) t) tmp group by brand) tt where tt.nov_month_sales - tt.oct_month_sales > 0;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705173234_e81ce810-9dfe-4cae-a455-bf28b0ac0fee
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0020, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-05 17:32:45,635 Stage-1 map = 0%,  reduce = 0%
2021-07-05 17:32:55,046 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.98 sec
2021-07-05 17:33:03,375 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.06 sec
MapReduce Total cumulative CPU time: 10 seconds 60 msec
Ended Job = job_1625494931119_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.06 sec   HDFS Read: 63010352 HDFS Write: 3186 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 60 msec
OK
```



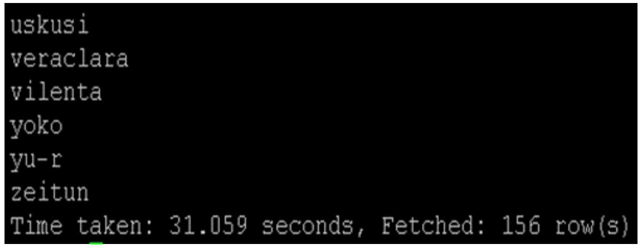
```
airnails
art-visage
artex
aura
balbcare
barbie
batiste
beautix
beauty-free
beautyblender
beauugreen
benovy
binacil
bioaqua
biofollica
biore
blixz
bodyton
bpw.style
browxenna
candy
carmex
chi
coifin
concept
cosima
cosmoprofi
cristalinas
cutrin
de.lux
depilflax
dewal
dizao
domix
ecocraft
ecolab
egomania
elizavecca
ellips
elskin
enjoy
eos
estel
estelare
f.o.x
farmavita
fedua
finish
fly
foamie
```



```
freedecor
freshbubble
gehwol
glysolid
godefroy
grace
grattol
greymy
happyfons
haruyama
helloganic
igrobeauty
ingarden
inm
insight
irisk
italwax
jas
jessnail
joico
juno
kaaral
kamill
kapous
kares
kaypro
keen
kerasys
kims
kinetics
kiss
kocostar
koelcia
koelf
konad
kosmekka
laboratorium
lador
ladykin
latinoil
levissime
levrana
lianail
likato
limoni
lovely
lowence
mane
marathon
markell
```



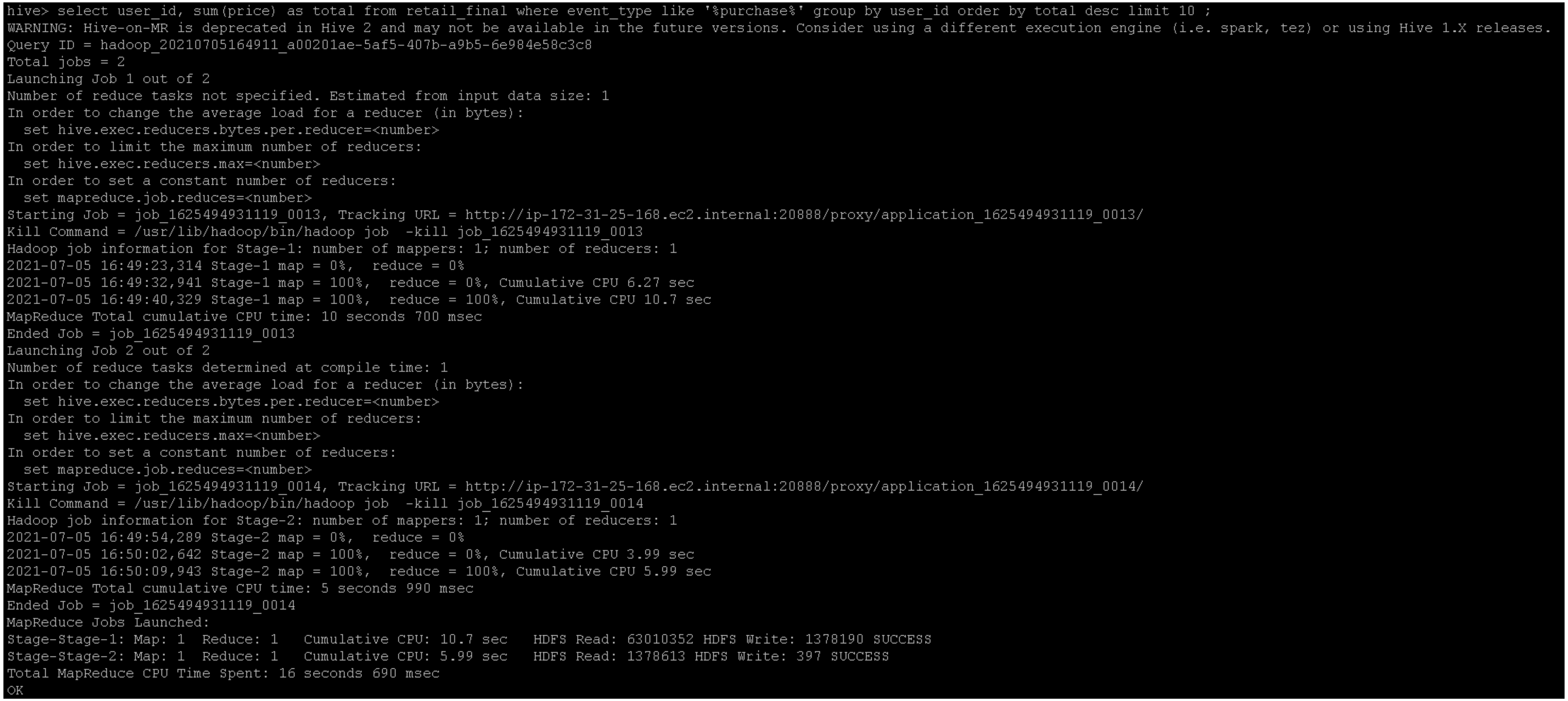
```
marutaka-foot
masura
matreshka
matrix
metzger
milv
miskin
missha
moyou
nagaraku
naomi
nefertiti
nirvel
nitrile
oniq
orly
osmo
parachute
plazan
polarus
profepil
profhenna
protokeratin
provoc
rasyan
refectocil
rosi
roubloff
runail
s.care
sanoto
severina
shary
shik
skinity
skinlite
smart
soleo
solomeya
sophin
staleks
strong
supertan
swarovski
tannymaxx
tertio
thuya
treaclemoon
trind
uno
```



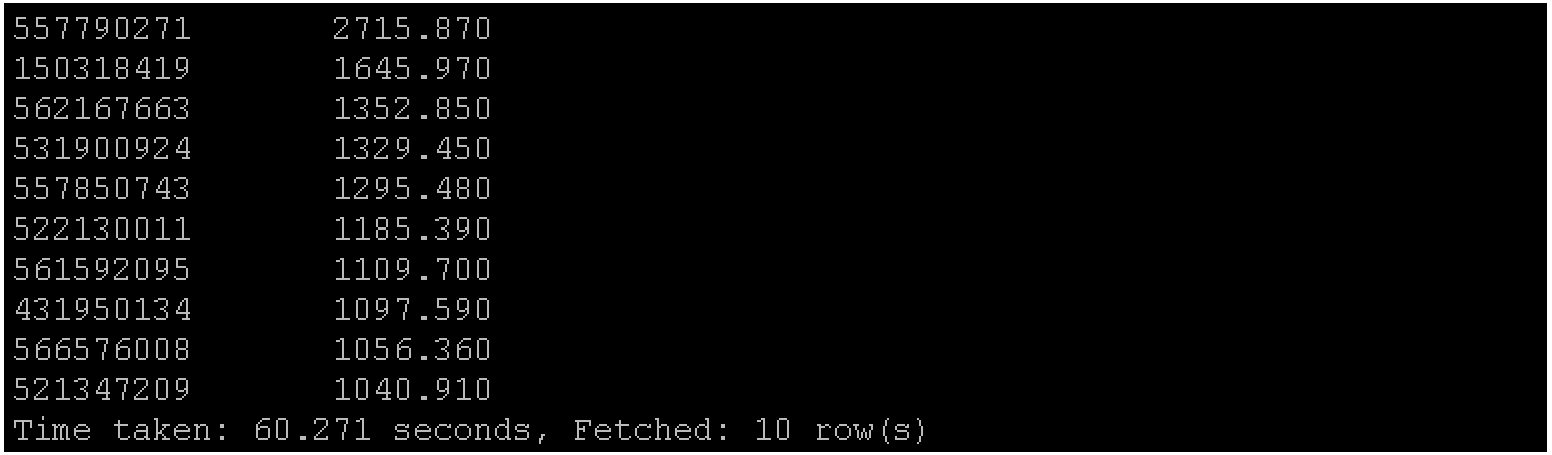
```
uskusi
veraclara
vilenta
yoko
yu-r
zeitun
Time taken: 31.059 seconds, Fetched: 156 row(s)
```

8. ***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 from retail_final where event_type like '%purchase%' group by user_id order by total desc limit 10 ;

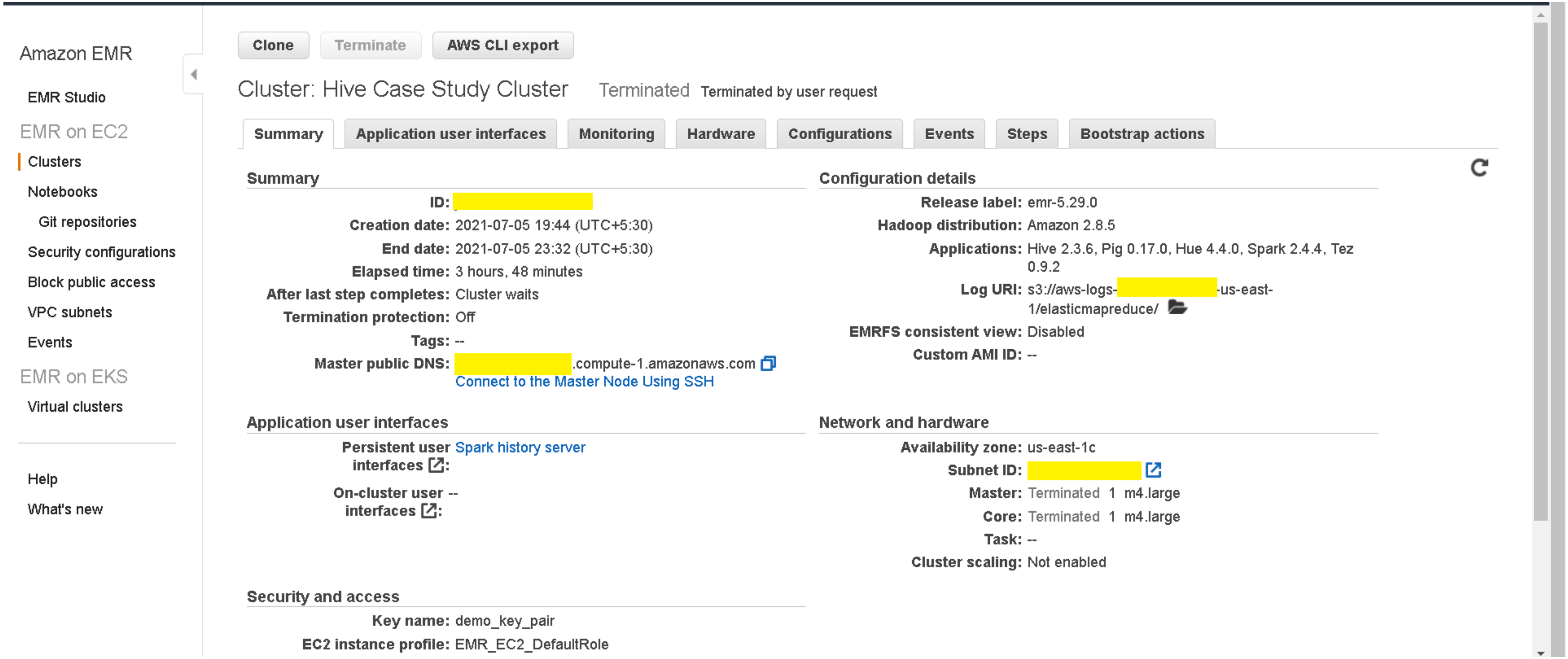


```
hive> select user_id, sum(price) as total from retail_final where event_type like '%purchase%' group by user_id order by total desc limit 10 ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hadoop_20210705164911_a00201ae-5af5-407b-a9b5-6e984e58c3c8
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0013, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-05 16:49:23,314 Stage-1 map = 0%,  reduce = 0%
2021-07-05 16:49:32,941 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.27 sec
2021-07-05 16:49:40,329 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.7 sec
MapReduce Total cumulative CPU time: 10 seconds 700 msec
Ended Job = job_1625494931119_0013
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1625494931119_0014, Tracking URL = http://ip-172-31-25-168.ec2.internal:20888/proxy/application_1625494931119_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1625494931119_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-07-05 16:49:54,289 Stage-2 map = 0%,  reduce = 0%
2021-07-05 16:50:02,642 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.99 sec
2021-07-05 16:50:09,943 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.99 sec
MapReduce Total cumulative CPU time: 5 seconds 990 msec
Ended Job = job_1625494931119_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.7 sec   HDFS Read: 63010352 HDFS Write: 1378190 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.99 sec   HDFS Read: 1378613 HDFS Write: 397 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 690 msec
OK
```



```
557790271       2715.870
150318419       1645.970
562167663       1352.850
531900924       1329.450
557850743       1295.480
522130011       1185.390
561592095       1109.700
431950134       1097.590
566576008       1056.360
521347209       1040.910
Time taken: 60.271 seconds, Fetched: 10 row(s)
```

➢ Terminating the EMR cluster



Cluster: Hive Case Study Cluster Terminated Terminated by user request

Summary

ID: [redacted]
Creation date: 2021-07-05 19:44 (UTC+5:30)
End date: 2021-07-05 23:32 (UTC+5:30)
Elapsed time: 3 hours, 48 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: [redacted].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: Hive 2.3.6, Pig 0.17.0, Hue 4.4.0, Spark 2.4.4, Tez 0.9.2
Log URI: s3://aws-logs-[redacted]-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: Spark history server
On-cluster user interfaces: --

Network and hardware

Availability zone: us-east-1c
Subnet ID: [redacted]
Master: Terminated 1 m4.large
Core: Terminated 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: demo_key_pair
EC2 instance profile: EMR_EC2_DefaultRole