

We have created keypair with a name "cloudkeypair".

The screenshot shows the AWS EC2 Management Console. The left sidebar has 'Services' selected, and the main content area is titled 'Key pairs (2)'. A table lists two key pairs:

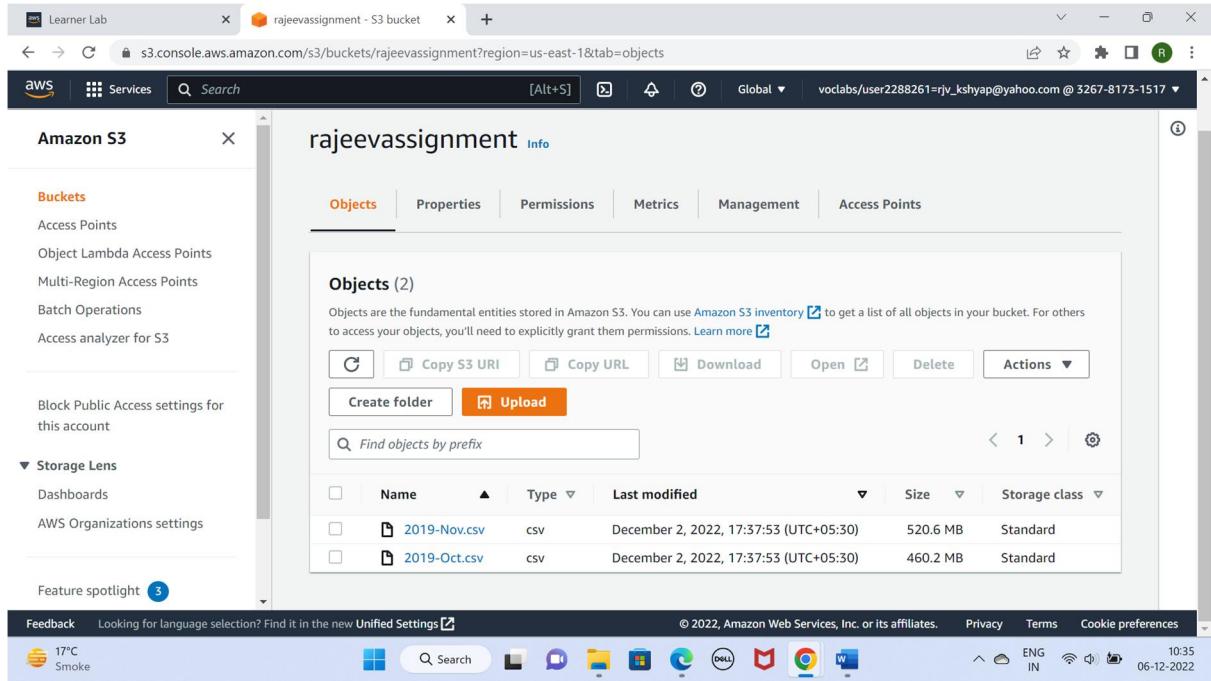
Name	Type	Created	Fingerprint
cloudkeypair	rsa	2022/11/27 09:34 GMT+5:30	c7:1e:b0:cb:5e:11:dd:5d:cb:d8:00:5b
vockey	rsa	2022/11/24 10:41 GMT+5:30	2a:b9:9d:8c:18:84:28:9c:97:1e:93:c3

We have created bucket with name "rajeevassignment".

The screenshot shows the AWS S3 Management Console. The left sidebar has 'Services' selected, and the main content area is titled 'Buckets (2)'. A table lists two buckets:

Name	AWS Region	Access	Creation date
aws-logs-326781731517-us-east-1	US East (N. Virginia) us-east-1	Objects can be public	November 27, 2022, 09:55:08 (
rajeevassignment	US East (N. Virginia) us-east-1	Bucket and objects not public	November 26, 2022, 11:30:36 (

We have uploaded October and November dataset as given in case study.



The screenshot shows the AWS S3 console interface. The left sidebar has sections for Buckets, Storage Lens, and Feature spotlight. The main area is titled 'Objects (2)' and displays two CSV files: '2019-Nov.csv' and '2019-Oct.csv'. The table includes columns for Name, Type, Last modified, Size, and Storage class. The 'Actions' dropdown menu is visible above the table. The top navigation bar shows the URL s3.console.aws.amazon.com and the user's email address.

Name	Type	Last modified	Size	Storage class
2019-Nov.csv	csv	December 2, 2022, 17:37:53 (UTC+05:30)	520.6 MB	Standard
2019-Oct.csv	csv	December 2, 2022, 17:37:53 (UTC+05:30)	460.2 MB	Standard

We have created cluster named “casestudy”.It is running now as shown below:

The screenshot shows the Amazon EMR AWS Console interface. On the left, there's a sidebar with navigation links for Learner Lab, Services (selected), Search, and various EMR-related sections like Studio, Serverless, and EC2. The main content area displays a success message about EMR Serverless being GA, followed by a summary of a cluster named 'casetudy'. The cluster is currently 'Running' and has completed a 'Running step'. A summary tab is selected, showing details such as ID (j-1NWNOOMW9C44O), Creation date (2022-12-06 10:38 UTC+5:30), Elapsed time (9 minutes), and termination protection status (Off). It also lists the Master public DNS (ec2-44-202-207-159.compute-1.amazonaws.com) and provides a link to connect via SSH. Below this, configuration details are shown, including release label (mrn-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), and log URI (s3://aws-logs-326781731517-us-east-1). The bottom of the screen includes a feedback bar, language selection, and standard browser navigation controls.

We have connected putty and it is activated now.

Data is copied into HDFS.

Accessing the S3 bucket.

Code: aws s3 ls rajeevassignment

```
[hadoop@ip-172-31-13-215:~]
Using username "hadoop".
Authenticating with public key "cloudkeypair"
Last login: Tue Dec  6 05:17:34 2022
[ ](hadoop@ip-172-31-13-215:~) ~ Amazon Linux AMI
[ ](hadoop@ip-172-31-13-215:~) ~

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
60 package(s) found for security, out of 88 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEE MMMMMMM MBBBBBBBBB RRRRRRRRRRRRRRRR
E: :::::::::::::::E M:::::::M M:::::::M R:::::::::::R
EE:::::;E:::::;E E: M:::::::M M:::::::M R:::::RKKKKK:::::R
E:;:::;E EEEEEE M:::::::M M:::::::M RR:::::R R:::::R
E:;:::;E M:::::::M M:::::::M M:::::::M R:::R R:::R
E:;:::;E EEEEEEE M:::::::M M:::::::M M:::::::M R:::::RKKKKR:::::R
E:;:::;E EEEEEE M:::::::M M:::::::M M:::::::M R:::::R R:::::R
EE:;:::;E EEEEEE E:::::::E M:::::::M M:::::::M R:::::R R:::::R
E:;:::;E EEEEEE M:::::::M M:::::::M M:::::::M R:::::R R:::::R
E:;:::;E EEEEEE M:::::::M M:::::::M R:::::R R:::::R
E:;:::;E EEEEEE M:::::::M M:::::::M R:::::R R:::::R
EE:;:::;E EEEEEE MMMMMMM MBBBBBBBBB RRRRRRRRRRRRRRRR
EEEEEEEEEEEEEEEEEEEEE MMMMMMM MBBBBBBBBB RRRRRRRRRRRRRRRR

[hadoop@ip-172-31-13-215:~]$ aws s3 ls rajeavassignment
2022-12-02 12:07:53 545839412 2019-Nov.csv
2022-12-02 12:07:53 482542278 2019-Oct.csv
[hadoop@ip-172-31-13-215:~]$ [ ](hadoop@ip-172-31-13-215:~) ~
```

Checking the availability of directory

Code: Hadoop fs –ls /user/hive

Loading the s3 public data set to created directory “casestudy” in hadoop .

Code: hadoop distcp 's3://rajeevassignment/2019-Oct.csv' /user/hive/casestudy/oct.csv

```
hadoop distcp 's3://rajeevassignment/2019-Nov.csv' /user/hive/casestudy/nov.csv
```

```
[hadoop@ip-172-31-13-215 ~]$ hadoop distcp 's3://rajeevassignment/2019-Oct.csv' /user/hive/casestudy/oct.csv
22/12/06 05:29:40 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListStatusThreads=20, maxMaps=20, mapbandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preservesStatus=[]}, preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourcefilelisting=null, sourcePaths=[s3://rajeevassignment/2019-Oct.csv], targetPath=/user/hive/casestudy/oct.csv, targetPartExists=false, filtersFile='null'
22/12/06 05:29:40 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-13-215.ec2.internal/172.31.13.215:8032
22/12/06 05:29:43 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
22/12/06 05:29:43 INFO tools.SimpleCopyListing: Build file listing completed.
22/12/06 05:29:43 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/12/06 05:29:43 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/12/06 05:29:43 INFO tools.DistCp: Number of paths in the copy list: 1
22/12/06 05:29:43 INFO tools.DistCp: Number of paths in the copy list: 1
22/12/06 05:29:43 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-13-215.ec2.internal/172.31.13.215:8032
22/12/06 05:29:44 INFO mapreduce.JobSubmitter: number of splits:1
22/12/06 05:29:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1670303708630_0001
22/12/06 05:29:44 INFO impl.YarnClientImpl: Submitted application application_1670303708630_0001
22/12/06 05:29:44 INFO mapreduce.Job: The url to track the job: http://ip-172-31-13-215.ec2.internal:20888/proxy/application_1670303708630_0001/
22/12/06 05:29:44 INFO tools.DistCp: DistCp job-id: job_1670303708630_0001
22/12/06 05:29:44 INFO mapreduce.Job: Running job: job_1670303708630_0001
22/12/06 05:29:52 INFO mapreduce.Job: Job job_1670303708630_0001 running in uber mode : false
22/12/06 05:29:52 INFO mapreduce.Job: map 0% reduce 0%
22/12/06 05:30:10 INFO mapreduce.Job: map 100% reduce 0%
22/12/06 05:30:11 INFO mapreduce.Job: Job job_1670303708630_0001 completed successfully
22/12/06 05:30:11 INFO mapreduce.Job: Counters
    File System Counters
        FILE: Number of bytes read=0
        FILE: Number of bytes written=172972
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=355
        HDFS: Number of bytes written=482542278
        HDFS: Number of read operations=12
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=4
        S3: Number of bytes read=482542278
        S3: Number of bytes written=0
        S3: Number of read operations=0
        S3: Number of large read operations=0
        S3: Number of write operations=0
    Job Counters
        Launched map tasks=1
        Other local map tasks=1
        Total time spent by all maps in occupied slots (ms)=484256
```

```
[hadoop@ip-172-31-13-215:~]
FILE: Number of bytes read=0
FILE: Number of bytes written=172976
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=356
HDFS: Number of bytes written=545839412
HDFS: Number of read operations=12
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
S3: Number of bytes read=545839412
S3: Number of bytes written=0
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0
Job Counters
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=517344
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=16167
Total vcore-milliseconds taken by all map tasks=16167
Total megabyte-milliseconds taken by all map tasks=16555008
Map-Reduce Framework
Map input records=1
Map output records=0
Input split bytes=137
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=211
CPU time spent (ms)=18310
Physical memory (bytes) snapshot=555503616
Virtual memory (bytes) snapshot=3289276416
Total committed heap usage (bytes)=560988160
File Input Format Counters
Bytes Read=219
File Output Format Counters
Bytes Written=0
DistCp Counters
Bytes Copied=545839412
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-13-215 ~]$
```

```
[hadoop@ip-172-31-13-215:~]
FILE: Number of write operations=0
HDFS: Number of bytes read=356
HDFS: Number of bytes written=545839412
HDFS: Number of read operations=12
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
S3: Number of bytes read=545839412
S3: Number of bytes written=0
S3: Number of read operations=0
S3: Number of large read operations=0
S3: Number of write operations=0
Job Counters
Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=517344
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=16167
Total vcore-milliseconds taken by all map tasks=16167
Total megabyte-milliseconds taken by all map tasks=16555008
Map-Reduce Framework
Map input records=1
Map output records=0
Input split bytes=137
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=211
CPU time spent (ms)=18310
Physical memory (bytes) snapshot=555503616
Virtual memory (bytes) snapshot=3289276416
Total committed heap usage (bytes)=560988160
File Input Format Counters
Bytes Read=219
File Output Format Counters
Bytes Written=0
DistCp Counters
Bytes Copied=545839412
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$
```

After loading the data set we check the data set file and data set in the hadoop directory.

Code: hadoop fs -ls /user/hive/casestudy/

```
hadoop fs -cat /user/hive/casestudy/oct.csv | head
```

```
hadoop fs -cat /user/hive/casestudy/nov.csv | head
```

```
hadoop@ip-172-31-13-215:~$ Launched map tasks=1
Other local map tasks=1
Total time spent by all maps in occupied slots (ms)=517344
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=16167
Total vcore-milliseconds taken by all map tasks=16167
Total megabyte-milliseconds taken by all map tasks=16555008
Map-Reduce Framework
  Map Input Records=1
  Map Output Records=0
  Input Split bytes=137
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=211
  CPU time spent (ms)=18310
  Physical memory (bytes) snapshot=555503616
  Virtual memory (bytes) snapshot=3289276416
  Total committed heap usage (bytes)=560988160
File Input Format Counters
  Bytes Read=219
File Output Format Counters
  Bytes Written=0
DistCp Counters
  Bytes Copied=545839412
  Bytes Expected=545839412
  Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:15 UTC,cart,5881449,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73de1e7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,14875800013522845895,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cclbb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$
```

19°C
Polluted air

Search

11:11
ENG IN

06-12-2022

```
hadoop@ip-172-31-13-215:~$ Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=211
CPU time spent (ms)=18310
Physical memory (bytes) snapshot=555503616
Virtual memory (bytes) snapshot=3289276416
Total committed heap usage (bytes)=560988160
File Input Format Counters
  Bytes Read=219
File Output Format Counters
  Bytes Written=0
DistCp Counters
  Bytes Copied=545839412
  Bytes Expected=545839412
  Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:15 UTC,cart,5881449,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73de1e7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,14875800013522845895,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cclbb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/nov.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-11-01 00:00:02 UTC,view,5802432,1487580009286598681,,0.32,562076640,09fafdf6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:05 UTC,cart,5844397,1487580006317032337,,,2,38,55322972,2067216c-31b5-455d-alcc-af0575a34fffb
2019-11-01 00:00:10 UTC,view,5837166,1783999064103190764,,pnb,22,22,556138645,57ed222a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC,cart,5876812,1487580010100293687,,jansail,3,16,564506666,12c1951-8052-4b37-addc-d4964b1d5f7
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3,33,553329724,2067216c-31b5-455d-alcc-af0575a34fffb
2019-11-01 00:00:25 UTC,remove_from_cart,5826182,1487580007483048900,,,3,33,553329724,2067216c-31b5-455d-alcc-af0575a34fffb
2019-11-01 00:00:32 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafdf6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675998693,,milv,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$
```

19°C
Smoke

Search

11:12
ENG IN

06-12-2022

Creating the database

The screenshot shows a terminal window titled 'hadoop@ip-172-31-13-215:~'. The command 'hadoop fs -ls /user/hive/casestudy/' is run, listing two files: 'nov.csv' and 'oct.csv'. Then, 'hadoop fs -cat /user/hive/casestudy/oct.csv | head' is run, displaying the first few lines of the CSV file. The output includes columns like event_time, event_type, product_id, category_id, category_code, brand, price, user_id, and user_session. The terminal also shows the configuration of the hive-log4j2.properties file, setting 'Async: false' and creating a database named 'ecomdb'. The session ends with 'hive> OK' and a timestamp '06-12-2022'.

```
Total committed heap usage (bytes)=560988160
File Input Format Counters
  Bytes Read=219
File Output Format Counters
  Bytes Written=0
DistCp Counters
  Bytes Copied=545839412
  Bytes Expected=545839412
  Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/oct.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26d6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773353,1487580005134238553,,runail,2.62,463240011,26d6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26d6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,1487580013522845895,,lovey,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-8b30-d329d5af04f
2019-10-01 00:00:19 UTC,cart,5739051,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4fa0-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,0.56,467916806,2f5b5546-58cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cclbb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/nov.csv | head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
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-alcc-af0575a34ff
2019-11-01 00:00:10 UTC,view,5837166,1783999064103190764,,pnk,22.22,556138649,57ed22ae-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC,cart,5876812,1487580010100293687,,jesmail,3.16,56450666,186c1951-8052-4b37-adce-d964bd1d5f7
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:25 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:32 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:34 UTC,remove_from_cart,58700838,1487580007675986893,,mivil,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hive>
```

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database ecomdb;
OK
Time taken: 0.887 seconds
hive>

19°C Smoke ENG IN 11:14 06-12-2022

After moving the data to the directory, we create the base table(ecom) and check for the data in the table.

Code: CREATE TABLE IF NOT EXISTS ecom(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'

STORED AS TEXTFILE

LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'='1');

```
[hadoop@ip-172-31-13-215:~]
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -ls /user/hive/casestudy/
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2022-12-06 05:35 /user/hive/casestudy/nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2022-12-06 05:30 /user/hive/casestudy/oct.csv
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/oct.csv | head
event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session
2019-10-01 00:00:00 UTC, cart, 5773203, 1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC, cart, 5773353, 1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC, cart, 5881589, 2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC, cart, 5723490, 1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC, cart, 5881449, 148758000513522845896,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC, cart, 5887269, 1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-b830-d329d5af04f
2019-10-01 00:00:19 UTC, cart, 5739055, 1487580005246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:25 UTC, cart, 5825598, 1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC, cart, 5698989, 1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cclbb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database ecomdb;
OK
Time taken: 0.887 seconds
hive> CREATE TABLE IF NOT EXISTS ecom(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'
> STORED AS TEXTFILE
> LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'='1');
OK
Time taken: 0.447 seconds
hive>
```

```
[hadoop@ip-172-31-13-215:~]
2019-10-01 00:00:00 UTC, cart, 5773203, 1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC, cart, 5773353, 1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC, cart, 5881589, 2151191071051219817,,lovely,13.48,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC, cart, 5723490, 1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC, cart, 5881449, 148758000513522845895,,lovely,0.56,429681830,49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC, cart, 5887269, 1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-b830-d329d5af04f
2019-10-01 00:00:24 UTC, cart, 5825598, 1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC, cart, 5698989, 1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cclbb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hadoop fs -cat /user/hive/casestudy/nov.csv | head
event_time, event_type, product_id, category_id, category_code, brand, price, user_id, user_session
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-alcc-af0575a34ffb
2019-11-01 00:00:10 UTC, view, 5837166, 1783999064103190764,,pnb,22.22,556138645,57ed22e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC, cart, 5876812, 148758000100293687,,jessnail,3.16,564506666,186c1951-8052-4b37-adce-d9d644b1d5f7
2019-11-01 00:00:24 UTC, remove_from_cart, 5826182, 1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:25 UTC, view, 5856189, 1487580009026551821,,runail,15.71,562076640,09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:34 UTC, remove_from_cart, 5870838, 1487580007675986893,,milv,0.79,429913900,2f0bfff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database ecomdb;
OK
Time taken: 0.887 seconds
hive> CREATE TABLE IF NOT EXISTS ecom(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'
> STORED AS TEXTFILE
> LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'='1');
OK
Time taken: 0.447 seconds
hive> select * from ecom 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-alcc-af0575a34ffb
2019-11-01 00:00:10 UTC view 5837166 1783999064103190764 pnb 22.22 556138645 57ed22e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart 5876812 148758000100293687 jessnail 3.16 564506666 186c1951-8052-4b37-adce-d9d644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart 5826182 1487580007483048900 3.33 553329724 2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:25 UTC view 5856189 1487580009026551821 runail 15.71 562076640 09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:32 UTC view, 5837835, 1933472286753424063, 3.49, 514649199, 432a4e95-375c-4b40-bd36-0fc039e77580
2019-11-01 00:00:34 UTC, remove_from_cart, 5870838, 1487580007675986893,,milv,0.79,429913900,2f0bfff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-13-215 ~]$ hive
```

```

hadoop@ip-172-31-13-215:~ 
2019-10-01 00:00:07 UTC, cart, 5723490, 1487580005134238553,, runail, 2.62, 463240011, 26dd6e6e-4dac-4778-8d2c-92e1494ab85 
2019-10-01 00:00:18 UTC, cart, 5881149, 14875800013522845898,, lovely, 0.56, 429681830, 49e8d843-adf3-420b-a2c3-fe8bc6a307c9 
2019-10-01 00:00:19 UTC, cart, 5857269, 1487580005134238553,, runail, 2.62, 430174032, 73deale7-664e-43f4-8b30-d329d5af04f 
2019-10-01 00:00:24 UTC, cart, 5739055, 1487580008246412266,, kapous, 4.75, 377667011, 81326acd-daa4-4f0a-b488-fd0596a78733 
2019-10-01 00:00:25 UTC, cart, 5825598, 148758000445982239,, 0.56, 467916806, 2f5b5546-58cb-9ee7-7ecd-84276f8ef486 
2019-10-01 00:00:25 UTC, cart, 5698989, 1487580006317032337,,, 1.27, 385985599, d50965e8-1101-44ab-b45d-cc1bb5fae694 
cat: Unable to write to output stream. 
[hadoop@ip-172-31-13-215 ~]$ hive 
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false 
hive> create database ecomdb; 
OK 
Time taken: 0.887 seconds 
hive> CREATE TABLE IF NOT EXISTS ecom(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' 
> STORED AS TEXTFILE 
> LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'='1'); 
OK 
Time taken: 0.447 seconds 
hive> select * from ecom 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-alcc-af0575a34ffb 
2019-11-01 00:00:10 UTC view 5844393 1487580006317032337,, 2.38, 553329724, 2067216c-31b5-455d-alcc-af0575a34ffb 
2019-11-01 00:00:11 UTC cart, 5837166, 1783999064103190764,, pnb, 22.22, 556138645, 57ed22ae-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, 1487580007483049900,, 3.33, 553329724, 2067216c-31b5-455d-alcc-af0575a34ffb 
2019-11-01 00:00:25 UTC, view, 5856189, 148758000926551821,, runail, 15.71, 562076640, 09fafd6c-6c99-46b1-834f-33527f4de241 
2019-11-01 00:00:32 UTC, view, 5837835, 1933472286753424063,, 3.49, 514649199, 432a4e95-375c-4b40-bd36-0fc039e77f580 
2019-11-01 00:00:34 UTC, remove_from_cart, 5870838, 14875800076765986893,, milv, 0.79, 429913900, 2f0bff3c-252f-4fee-afcd-5d8a6a92839a 
cat: Unable to write to output stream. 
[hadoop@ip-172-31-13-215 ~]$ 

```

Once the base table is created, we need to optimize the table for quick query result through partitioning and bucketing. Our optimized table name is ecom_part.

Code:

```

CREATE TABLE IF NOT EXISTS ecom_part (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_code) into 12 buckets 
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' 
STORED AS TEXTFILE LOCATION '/user/hive/casestudy/'tblproperties('skip.header.line.count'='1'); 
AND 
Loading the data into optimize table from base table. 

```

Code:

```

Insert into table ecom_part partition (event_type) select event_time , product_id , category_id , category_code , brand , price , user_id , user_session , event_type from ecom ; 

```

```

hadoop@ip-172-31-13-215:~$ hive> select * from ecom limit 5;
OK
2019-11-01 00:00:02 UTC view 5802432 1487580009286590681 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-af0575a34ff8
2019-11-01 00:00:10 UTC view 5837166 1783999064103190764 pnb 22.22 556138645 57e4222a-a54a-4907-8944-5a875c2d71f
2019-11-01 00:00:11 UTC cart 5876812 1487580001010029367 jessnail 3.16 564506666 186c1951-9052-4b37-adce-9d9644b1d5f7
2019-11-01 00:00:24 UTC remove_from_cart 5826182 1487580007483048900 3.33 553329724 2067216c-31b5-455d-a1cc-af0575a34ff8
Time taken: 2.098 seconds. Fetched 5 row(s)
hive> SET hive.exec.dynamic.partition=true;
hive> SET hive.exec.dynamic.partition.mode=nonstrict;
hive> SET hive.exec.reducers.bytes.per.reducer=100000000;
hive> CREATE TABLE IF NOT EXISTS ecom_part (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_code) into 12 buckets
> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
> STORED AS TEXTFILE LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'=1');
OK
Time taken: 0.059 seconds
hive> Insert into table ecom_part partition (event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from ecom;
Query ID = hadoop_20221206060533_645b97e0-90a6-436e-bd9f-5d34cb8ef62e
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_1670303708630_0004)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container SUCCEEDED 2 2 0 0 0 0  

Reducer 2 ..... container SUCCEEDED 5 5 0 0 0 0  

-----  

VERTICES: 0/2 [=====>>>] 100% ELAPSED TIME: 82.88 s  

-----  

Loading data to table default.ecom_part partition (event_type=null)  

-----  

Loaded : 4/4 partitions.  

Time taken to load dynamic partitions: 0.723 seconds  

Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 92.222 seconds
hive>
```

19°C Smoke ENG IN 11:37 06-12-2022

```

hadoop@ip-172-31-13-215:~$ hive> CREATE TABLE IF NOT EXISTS ecom_part (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_code) into 12 buckets
> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
> STORED AS TEXTFILE LOCATION '/user/hive/casestudy/' tblproperties('skip.header.line.count'=1);
OK
Time taken: 0.059 seconds
hive> Insert into table ecom_part partition (event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from ecom;
Query ID = hadoop_20221206060533_645b97e0-90a6-436e-bd9f-5d34cb8ef62e
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_1670303708630_0004)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container SUCCEEDED 2 2 0 0 0 0  

Reducer 2 ..... container SUCCEEDED 5 5 0 0 0 0  

-----  

VERTICES: 0/2 [=====>>>] 100% ELAPSED TIME: 82.88 s  

-----  

Loading data to table default.ecom_part partition (event_type=null)  

-----  

Loaded : 4/4 partitions.  

Time taken to load dynamic partitions: 0.723 seconds  

Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 92.222 seconds
hive> set hive.cli.print.header=true;
hive> select * from ecom_part limit 5;
OK
ecom_part.event_time ecom_part.product_id ecom_part.category_id ecom_part.category_code ecom_part.brand ecom_part.price ecom_part.user_id ecom_
part.user_session ecom_part.event_type
2019-10-08 11:47:14 UTC 5689725 1487580007852147670 staleks 13.17 404502068 928c919b-42de-4b94-afdd-19423944f5f0 cart
2019-10-08 18:31:54 UTC 5870696 148758000824612266 4.60 100787781 188a44b5-83f1-4f19-8a93-2fa670f2ec08 cart
2019-10-07 21:38:36 UTC 5797252 1638456119066100510 pole 4.11 533267875 4d44c69a-ea11-4fa6-8f97-39a72e6931cb cart
2019-10-08 18:31:55 UTC 5887003 1487580006317032337 7.94 459127083 76f0c023-c35e-4ca9-8146-34bc6c94382e cart
2019-10-08 18:31:55 UTC 5861279 1487580006317032337 30.95 558176613 6bcac932-1da0-46bb-bea6-6cd19ac6be00 cart
Time taken: 0.171 seconds, Fetched: 5 row(s)
hive>
```

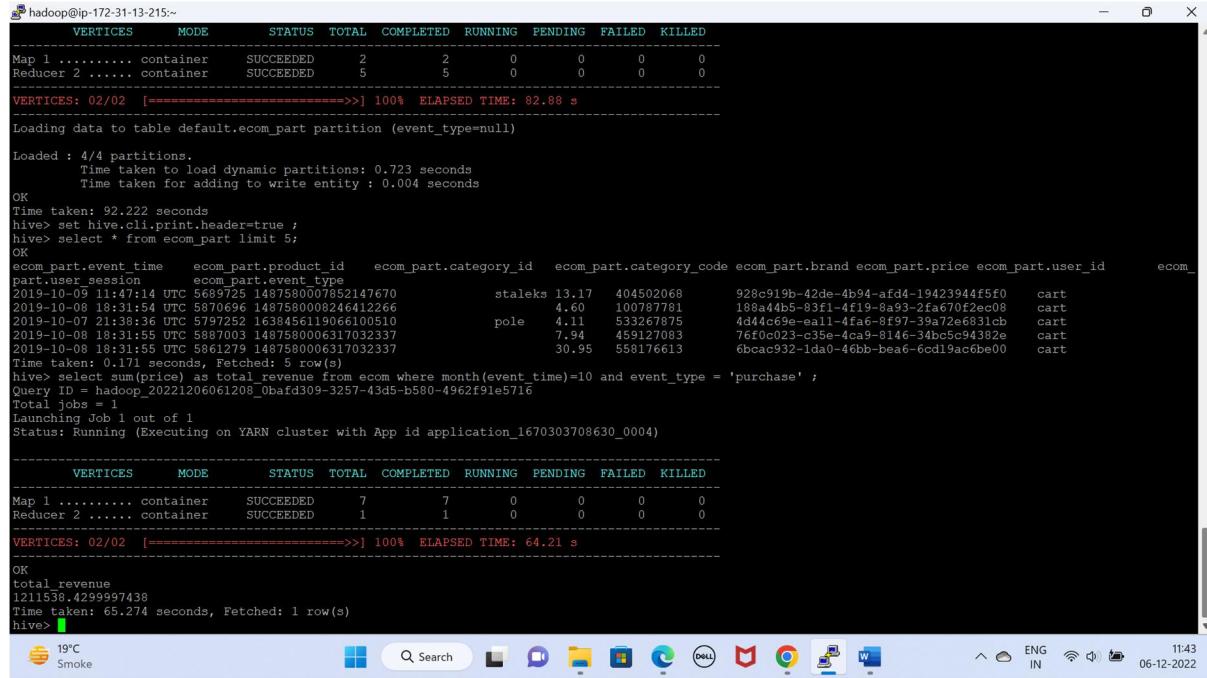
19°C Smoke ENG IN 11:40 06-12-2022

Analysis of Queries

QUESTION 1

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

Code: select sum(price) as total_revenue from ecom where month(event_time)=10 and event_type = 'purchase' ;



The screenshot shows a terminal window with the following output:

```
hadoop@ip-172-31-13-215:~$ hadoop@ip-172-31-13-215:~$ 
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 2 2 0 0 0 0 0
Reducer 2 ..... container SUCCEEDED 5 5 0 0 0 0 0
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 82.88 s
Loading data to table default.ecom_part partition (event_type=null)
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 0.723 seconds
Time taken for adding to write entity : 0.004 seconds
OK
Time taken: 92.222 seconds
hive> set hive.cli.print.header=true ;
hive> select * from ecom_part limit 5;
OK
ecom_part.event_time ecom_part.product_id ecom_part.category_id ecom_part.category_code ecom_part.brand ecom_part.price ecom_part.user_id ecom_
part.user_session ecom_part.event_type
2019-10-09 11:47:14 UTC 5689725 1487580007852147670 staleks 13.17 404502068 928c919b-42de-4b94-afcd4-19423944f5f0 cart
2019-10-09 18:31:54 UTC 5870696 1487580008246412266 4.60 100787781 188a44b5-83f1-4f19-8a93-2fa670f2ec08 cart
2019-10-07 21:38:36 UTC 5797252 1638456119066100510 pole 4.11 533267875 4d44c69e-ea11-4fa6-8f97-39a72e6831cb cart
2019-10-08 18:31:55 UTC 5887003 1487580006317032337 7.94 459127083 76f0c023-c35e-4ca9-8146-34bc5c94382e cart
2019-10-08 18:31:55 UTC 5861279 1487580006317032337 30.95 558176613 6bcac932-1da0-46bb-bea6-6cd19ac6be00 cart
Time taken: 0.171 seconds, Fetched: 5 row(s)
hive> select sum(price) as total_revenue from ecom where month(event_time)=10 and event_type = 'purchase' ;
Query ID = hadoop_20221206061208_0baids09-3257-43d5-b580-4962f91e5716
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1670303708630_0004)
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 7 7 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 64.21 s
OK
total_revenue
1211538.429997438
Time taken: 65.274 seconds, Fetched: 1 row(s)
hive>
```

The terminal window also shows system status icons at the bottom, including battery level (19°C), network, and system time (11:43, 06-12-2022).

Note: The below screenshot of the same query from both the base table and the bucketed table. When compared the bucketed table takes less time to query the result than the base table.

Code:

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

```

hadoop@ip-172-31-13-215:~ part.user_session ecom_part.event_type
2019-10-09 11:47:14 UTC 5689725 1487580007852147670 staleks 13.17 404502068 928c919b-42de-4b94-afdd-19423944f5f0 cart
2019-10-08 18:31:54 UTC 5870696 1487580008246412266 4.60 100787781 188a44b5-63f1-4f19-8a93-2fa670f2ec08 cart
2019-10-07 21:38:18 UTC 5797252 1638456119066100510 pole 4.11 533267875 4d44c69e-ea11-4fa6-8f97-39a72e6631cb cart
2019-10-07 18:31:58 UTC 5887000 1487580006317032337 7.94 459127083 76f0c023-c35e-4ca9-8146-34bc5c94382e cart
2019-10-08 18:31:55 UTC 5861279 1487580006317032337 30.95 558176613 6bcac932-1da0-46bb-bea6-6cd19ac6be00 cart
Time taken: 0.171 seconds, Fetched: 5 row(s)
hive> select sum(price) as total_revenue from ecom where month(event_time)=10 and event_type = 'purchase' ;
Query ID = hadoop_20221206061510_0baf309-3257-43d5-b580-4962f91e5716
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1670303708630_0004)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 7 7 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
----- VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 64.21 s
----- OK
total_revenue
1211538.4299997438
Time taken: 65.274 seconds, Fetched: 1 row(s)
hive> select sum(price) as total_revenue from ecom_part where month(event_time)=10 and event_type = 'purchase' ;
Query ID = hadoop_20221206061510_bc662aab-5e67-4903-8388-c1198367a372
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1670303708630_0004)

----- 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: 9.48 s
----- OK
total_revenue
1211525.179999663
Time taken: 10.36 seconds, Fetched: 1 row(s)
hive>

```

Insights:

Total revenue generated on the basis of Purchase in the month of October 2019 was 1,211,525/- We tried getting the result using both base table[ecom] and optimized table[ecom_part].We figured out that Bucketed table reduces the query time when compared to the base table approximately by 55 seconds.

QUESTION 2:

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

Code:

```

SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases
FROM ecom_data_part WHERE event_type='purchase'
GROUP BY date_format(event_time, 'MM');

```

QUESTION 3:

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

Code:

```
WITH rev_difference AS (SELECT SUM(case when MONTH(event_time) = '10' then price else 0 end)
AS Oct_purchase, SUM(case when MONTH(event_time) = '11' then price else 0 end) AS
Nov_purchase FROM ecom WHERE event_type= 'purchase') SELECT (Nov_purchase - Oct_purchase)
as difference_revenue FROM rev_difference ;
```

```
hadoop@ip-172-31-13-215:~$ hadoop hive -e "WITH rev_difference AS (SELECT SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase, SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase FROM ecom WHERE event_type= 'purchase') SELECT (Nov_purchase - Oct_purchase) as difference_revenue FROM rev_difference ;"
```

FAILED: ParseException line 2:0 extraneous input 'SUM' expecting (near '(' in statement

```
hive> WITH rev_difference AS (SELECT SUM(case when MONTH(event_time) = '10' then price else 0 end) AS Oct_purchase, SUM(case when MONTH(event_time) = '11' then price else 0 end) AS Nov_purchase FROM ecom WHERE event_type= 'purchase') SELECT (Nov_purchase - Oct_purchase) as difference_revenue FROM rev_difference ;
```

Query ID = hadoop_2022120602833_d455b781-3ccf-4a9a-b5b0-24a06f87e97e

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_1670303708630_0005)

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	7	7	0	0	0	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 57.19 s

OK

```
difference_revenue
319478.470003781
Time taken: 64.348 seconds, Fetched: 1 row(s)
```

hive>

20°C
Smoke

Search

File

Folder

Task View

Taskbar Icons

Network

Cloud

Power

System

11:59
06-12-2022

Insights:

The difference in revenue between month of October and November

is 319478.4700. So, we conclude that the revenue generated in the month of November was more than the revenue generated in the month of October.

QUESTION 4:

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

Code:

```
SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category  
FROM ecom_part  
WHERE SPLIT(category_code,'\\.')[0] <> ";
```

```
hadoop@ip-172-31-215:~  
kamill  
kapous  
kares  
keume  
kiss  
kocostar  
laiseven  
lowence  
mavalala  
profepil  
chik  
ekintiy  
emart  
sport  
stationery  
tosowuong  
uskusi  
vl-gel  
ypsed  
Time taken: 56.229 seconds, Fetched: 250 row(s)  
hive> SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category FROM ecom_part WHERE SPLIT(category_code,'\\.')[0] <> '';  
Query ID = hadoop_20221206063553_23d14707-6f8d-4e27-ae31-f0b3af486cd3  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1670303708630_0005)  
-----  
     VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 ..... container    SUCCEEDED      5      5      0      0      0      0  
Reducer 2 ..... container    SUCCEEDED      5      5      0      0      0      0  
-----  
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 34.51 s  
-----  
OK  
category  
furniture  
appliances  
accessories  
apparel  
sport  
stationery  
Time taken: 35.032 seconds, Fetched: 6 row(s)  
hive>   
20°C Smoke
```

Insights:

There are six different categories under which company sells it's different products and they are furniture, appliances, accessories, apparel, sport and stationery.

QUESTION 5:

Find the total number of products available under each category.

Code:

```
SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products  
FROM ecom_part  
WHERE SPLIT(category_code,'\\.')[0] <> "
```

```
GROUP BY SPLIT(category_code,'\\.'[0]
```

```
ORDER BY No_of_products DESC;
```

The screenshot shows a terminal window with the following content:

```
hadoop@ip-172-31-13-215:~  
embryolisse 25  
uralsoap 20  
cuccio 15  
bodipure 14  
macadamia 13  
nova 13  
invisibobble 10  
footlogix 5  
ikoo 5  
dessata 4  
shifei 4  
voesh 4  
vl-gel 3  
sport 2  
genie 2  
queen 1  
lbd 1  
Time taken: 59.652 seconds, Fetched: 250 row(s)  
hive> SELECT SPLIT(category_code,'\\.'[0] AS Category, COUNT(product_id) AS No_of_products FROM ecom_part WHERE SPLIT(category_code,'\\.'[0] <> '' GROUP BY SPLIT(category_code,'\\.'[0]) ORDER BY No_of_products DESC;  
Query ID = hadoop_20221206064656_4c437c15-e651-4307-83a8-a9b100a7b6ac  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1670303708630_0005)  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 ..... container SUCCEEDED 5 5 0 0 0 0  
Reducer 2 ..... container SUCCEEDED 5 5 0 0 0 0  
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0  
-----  
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 43.87 s  
-----  
OK  
category no_of_products  
appliances 61731  
stationery 26718  
furniture 23596  
apparel 18232  
accessories 12922  
sport 1  
Time taken: 44.354 seconds, Fetched: 6 row(s)  
hive>   
20°C Smoke Search DELL M 12:18 IN 06-12-2022
```

Insights:

The highest no. of products are available under appliances category which is 61731 and lowest products under sport category which is 1.

QUESTION 6:

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

Code:

```
SELECT brand,sum(price) as total_price from ecom  
where brand !='' and event_type ='purchase'  
group by brand order by total_price desc limit 1;
```

```

Time taken: 44.354 seconds, Fetched: 6 row(s)
hive> SELECT brand,sum(price) as total_price from ecom where brand !='' and event_type ='purchase' group by brand order by total_price desc limit 1;
Query ID = hadoop_20221206065005_e5971073-3c91-46ed-977b-edc29537a058
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1670303708630_0005)

-----  

  VERTICES   MODE     STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED    7      7      0      0      0      0  

Reducer 2 ..... container SUCCEEDED    3      3      0      0      0      0  

Reducer 3 ..... container SUCCEEDED    1      1      0      0      0      0  

-----  

VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 61.20 s  

-----  

OK  

brand    total_price  

runail  148297.9400000003  

Time taken: 61.75 seconds, Fetched: 1 row(s)
hive> 
```

Insights:

Runail is the brand which has maximum sales in the month of October and November of 2019 combined.

QUESTION 7:

Which brands increased their sales from October to November?

Code:

```

WITH Monthly_Revenue AS (
SELECT brand,
SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END)
AS Oct_Revenue, SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN
price ELSE 0 END) AS Nov_Revenue FROM ecom WHERE event_type='purchase' AND
date_format(event_time, 'MM') IN ('10', '11') GROUP BY brand)

SELECT brand, Oct_Revenue, Nov_Revenue, Nov_Revenue-Oct_Revenue AS
Sales_Difference FROM Monthly_Revenue
WHERE (Nov_Revenue - Oct_Revenue)>0
ORDER BY Sales_Difference;
```

hadoop@ip-172-31-13-215:~

```
treaclemoon      163.36999999999995   181.4899999999995   18.12000000000005
kamill       63.00999999999999     81.49000000000002   18.480000000000032
juno        0.0                   21.08          21.08
veraclara    50.109999999999985   71.21000000000001   21.10000000000023
glysophil   69.72999999999998    91.58999999999997   21.86
godefroy     401.2200000000002   425.12000000000006   23.899999999999864
binacil     0.0                   24.25999999999998   24.25999999999998
blinx       38.949999999999986   63.39999999999998   24.44999999999998
profepil     15.300000000000001   118.00000000000005   24.66000000000025
estrelade   444.80999999999943   471.87000000000009   27.06000000000148
momo       902.3800000000005   321.09000000000003   28.70995999999981
bllore     60.65000000000006   30.31          29.65999999999997
beautyblender 78.74000000000001   109.41          30.669999999999987
vilenta     197.6000000000002   231.210000000002   33.61000000000014
mavala     409.0399999999985   446.32          37.28000000000014
likato      296.0599999999998   340.969999999999   44.91000000000025
ladykin    125.6499999999999   170.57          44.92
foamie      35.04          80.49          45.44999999999996
balskin    251.0900000000057   307.6500000000055   56.55999999999974
balbcare    155.3299999999996   212.3800000000025   57.05000000000296
koelcia    112.7500000000003   57.25000000000003
profheyna  679.229999999999   736.850000000005   57.6200000000057
kares      0.0                   59.45          59.45
marutaku-foot 49.2199999999999   109.33          60.11000000000001
dewal      0.0                   61.29          61.29
imm        288.02          351.210000000001   63.19000000000011
laboratorium 246.4499999999991   312.52          66.02000000000007
cutrin      299.3699999999995   367.62          68.2500000000006
egomania    77.47          146.0400000000002   68.57000000000002
konad      739.829999999991   810.6700000000003   70.84000000000117
nirvel      163.0399999999996   234.3299999999984   71.28999999999988
koell     422.7299999999985   507.2900000000002   84.5600000000034
plazan     101.37          194.010000000002   92.6400000000001
aura       83.95          177.51          93.5599999999999
kerasys    430.909999999985   525.2000000000002   94.2900000000003
enjoy      41.3499999999994   136.5700000000002   95.22000000000003
depiliflax  2707.06999999994   2803.779999999975   96.71000000000367
eos        54.3399999999996   152.61          98.2700000000001
carmex     145.08          243.36          98.28
batiste     772.399999999999   874.1699999999953
osmo       645.58          762.310000000002   116.7300000000013
dizao     819.1300000000012   945.5099999999998   126.37999999999852
igrobeauty  513.660000000009   645.069999999999   131.40999999999906
finish     98.38          230.3800000000008   132.0000000000009
```

```

[hadoop@ip-172-31-161-215:~]
metzger 5373.450000000006 6457.159999999988 1083.7099999999818
de.lux 1659.699999999967 2775.509999999988 1115.8100000000009
swarovski 1897.9299999999873 3043.160000000003 1155.2300000000157
beauty-free 554.1700000000000 1782.8600000000163 1228.6900000000155
zeitun 708.660000000004 2009.63 1300.969999999999
jio.co 705.52 2015.180000000015 1309.580000000015
severine 1775.88 6120.480000000023 1344.60000000023
irisk 45591.96000000588 46946.040000002184 1354.0799999963056
onix 8425.41000000003 9841.650000000018 1416.239999999987
ileana 2243.56000000002 3664.099999999998 1420.5399999999959
rouboloff 3491.360000000003 4913.769999999991 1422.4099999999885
smart 4457.26000000004 5902.140000000017 1444.8800000000128
shik 3341.2 4839.72000000007 1498.520000000068
domix 10472.04999999994 12009.17000000022 1537.1200000000827
artax 2730.63999999998 4327.250000000017 1596.6100000000192
beautix 10493.94999999996 12222.949999999913 1728.999999999472
mily 3904.939999999964 5642.010000000017 1737.0700000000838
masura 31266.07999999821 33058.46999999708 1792.389999998753
f.o.x 6624.22999999982 8577.28000000004 1953.050000000022
kapous 11927.15999999989 14093.080000000158 2165.920000000026
concept 11032.139999999925 13380.3999999993 2348.2600000000057
estel 21756.750000000342 24142.6700000002 2385.919999999878
kaypro 881.339999999998 3268.6999999995 2387.359999999995
benovy 409.620000000002 3259.970000000001 2850.350000000001
italwax 21940.23999999732 24799.369999999893 2859.130000000161
yoko 8756.90999999949 11707.87999999996 2950.9700000000466
haruyama 9390.68999999991 12352.91000000013 2962.2200000001394
marathon 7280.74999999997 10273.1 2992.350000000003
lovelv 8704.37999999952 11939.06000000045 3234.680000000093
bpw.style 11572.150000001699 14837.440000000812 3265.289999999113
staleks 8519.73000000003 11875.6100000008 3355.880000000074
freedecor 3421.7799999971 7671.800000000175 4250.02000000204
runail 71539.279999993 76758.66000000098 5219.380000001649
polarus 6013.72000000003 11371.93000000018 5358.210000000155
cosmoprofi 8322.81000000007 14536.99000000016 6214.180000000089
jessnail 26287.83999999916 33345.2299999992 7057.390000000007
strong 29196.6299999994 38671.26999999924 9474.63999999985
ingarden 23161.390000000138 33566.21000000009 10404.81999999949
lilemail 5892.83999999975 16394.240000000245 10501.40000000027
uno 35302.0299999977 51039.749999998035 15737.719999998262
gratolo 35445.5400000011 71472.7100000068 36027.16999999576
474679.0599999623 619509.2399999934 144830.18000003108
Time taken: 65.917 seconds, Fetched: 161 row(s)
hive> [REDACTED]

```

Insights:

'Grattol' brand has maximum increment of 36,027.16 /- and 'Ovale' has least increment of 0.56 /- from October to November.

QUESTION 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.

Code:

```

SELECT user_id, SUM(price) as Total_Expense
FROM ecom
WHERE event_type='purchase'
GROUP BY user_id
ORDER BY Total_Expense DESC LIMIT 10;

```

```

hadoop@ip-172-31-13-215:~$ hadoop@ip-172-31-13-215:~$ 
staleks 8519.730000000003 11875.61000000008 3355.8800000000774
freedecor 3421.779999999971 7671.8000000000175 4250.020000000204
runail 71539.719999999933 76758.660000000098 5219.3800000001649
polarus 6013.720000000003 11371.930000000018 5358.2100000001055
cosmoprofi 1322.400000000007 14536.99000000016 6214.160000000089
jessnail 26287.839999999516 33345.23999999982 1057.360000000007
strong 29196.62999999994 30671.25999999994 9474.639999999949
ingarden 23161.3900000000138 33566.21000000009 10404.819999999949
lianmail 5892.839999999975 16394.240000000245 10501.40000000027
uno 35302.02999999977 51039.749999998035 15737.719999998262
grattol 35445.5400000011 71472.71000000068 36027.16999999576
474679.0599999623 619509.2399999934 144830.18000003108
Time taken: 65.917 seconds, Fetched: 161 row(s)
hive> SELECT user_id, SUM(price) as Total_Expense FROM ecom WHERE event_type='purchase' GROUP BY user_id ORDER BY Total_Expense DESC LIMIT 10;
Query ID: 20221206070415_2d49e900-9064-4528-a69c-bc73cc37efaa
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_1670303708630_0007)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 7 7 0 0 0 0  

Reducer 2 ..... container SUCCEEDED 3 3 0 0 0 0  

Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0  

-----  

VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 63.44 s  

-----  

OK  

user_id total_expense  

557790271 2715.869999999991  

150318419 1645.97  

562167663 1352.850000000004  

531900924 1329.450000000003  

557850743 1295.480000000002  

522130011 1185.389999999994  

561592095 1109.699999999996  

431950134 1097.589999999995  

566576008 1056.360000000017  

521347209 1040.909999999999  

Time taken: 69.989 seconds, Fetched: 10 row(s)
hive>
```

Insight:

Above is the list of top 10 users to which company can reward golden customer plan.

Once the analysis is done, we can terminate the cluster by changing the Termination protection from ON to OFF and then click on the terminate button.

Learner Lab EMR – AWS Console EC2 Management Console

us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1NWNOOMW9C44O

aws | Services | Search | [Alt+S] | N. Virginia | vclabs/user2288261=rjv_kshyap@yahoo.com @ 3267-8173-1517

Amazon EMR

EMR Studio
EMR Serverless New
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters

Cluster: casestudy Waiting Cluster ready after last step completed.

Clone Terminate AWS CLI export Auto-termination is not available for this account when using this release of EMR.

Summary

ID: j-1NWNOOMW9C44O
Creation date: 2022-12-06 10:38 (UTC+5:30)
Elapsed time: 1 hour, 57 minutes
After last step completes: Cluster waits
Termination protection: On Off ✓ ✗
Tags: -- View All / Edit
Master public DNS: ec2-44-202-207-159.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

Feedback Looking for language selection? Find it in the new Unified Settings [?](#)

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

ENG IN 12:41 06-12-2022

Learner Lab EMR – AWS Console EC2 Management Console

us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1NWNOOMW9C44O

aws | Services | Search | [Alt+S] | N. Virginia | vclabs/user2288261=rjv_kshyap@yahoo.com @ 3267-8173-1517

Amazon EMR

EMR Studio
EMR Serverless New
EMR on EC2
Clusters
Notebooks
Git repositories
Security configurations
Block public access
VPC subnets
Events
EMR on EKS
Virtual clusters

Cluster: casestudy Terminating Terminated by user request

Clone Terminate AWS CLI export Auto-termination is not available for this account when using this release of EMR.

Summary

ID: j-1NWNOOMW9C44O
Creation date: 2022-12-06 10:38 (UTC+5:30)
Elapsed time: 2 hours, 3 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-44-202-207-159.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

Feedback Looking for language selection? Find it in the new Unified Settings [?](#)

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

ENG IN 12:42 06-12-2022

The screenshot shows the AWS EMR console interface. On the left, a sidebar lists services: Amazon EMR, EMR Studio, EMR Serverless (with a 'New' button), EMR on EC2 (selected), Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, Events, and EMR on EKS. The main content area displays a cluster named 'casestudy' which is 'Terminated' due to a user request. A prominent message at the top says 'EMR Serverless is now GA.' It encourages users to get started with EMR Serverless. Below this, there are tabs for Summary, Application user interfaces, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. The Summary tab is selected. Key details shown include:

- ID: j-1NWNOOMW9C44O
- Creation date: 2022-12-06 10:38 (UTC+5:30)
- End date: 2022-12-06 12:44 (UTC+5:30)
- Elapsed time: 2 hours, 6 minutes
- After last step completes: Cluster waits
- Termination protection: Off
- Tags: --
- Master public DNS: ec2-44-202-207-159.compute-1.amazonaws.com [Copy](#)
- [Connect to the Master Node Using SSH](#)

At the bottom of the page, there are links for Feedback, Unified Settings, © 2022, Amazon Web Services, Inc. or its affiliates., Privacy, Terms, and Cookie preferences. The status bar at the bottom right shows ENG IN, 12:47, and 06-12-2022.

SUMMARY

For this case study we worked with public clickstream dataset of a cosmetics store.

The complete steps are given below:

- 1.Copying the data set into the HDFS from S3 Bucket.
- 2.Creating the database and launching Hive queries on EMR cluster.
[use CSVSerde with the default properties value for loading the dataset into Hive table]
- 3.Cleaning up.