Hive Case Study Of Online Cosmetic Store Using HiveQL

Subhasis Jana

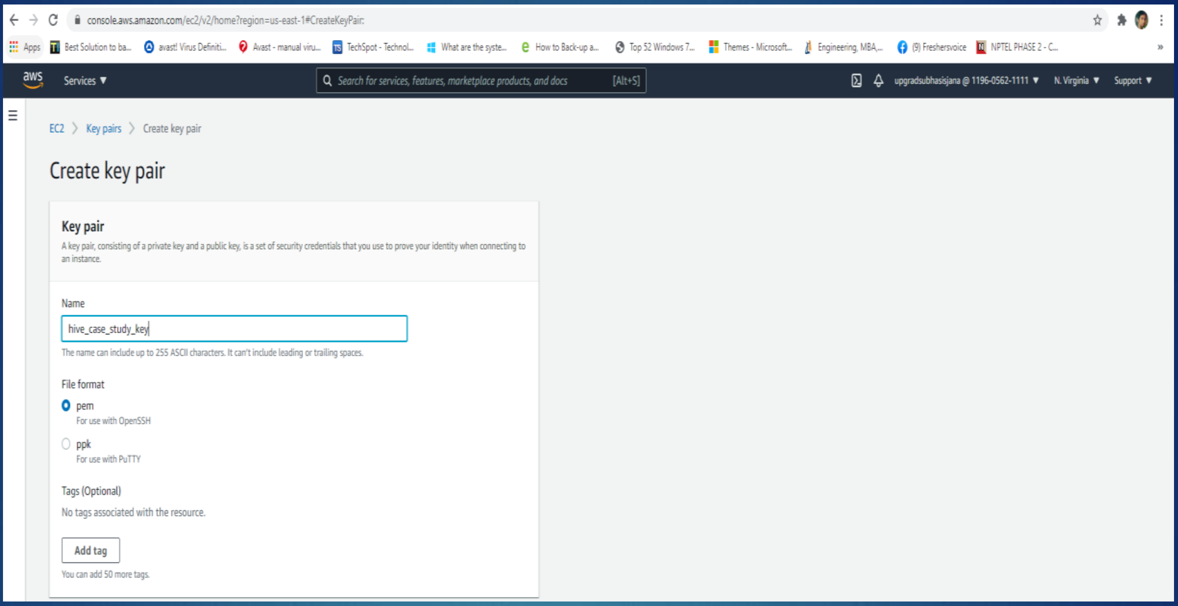
Bramhani Kottada

2/20/21

SQL & NoSQL Databases: Case Study

Case Study Steps

Key-pair .pem creation and .ppk generation



Create key pair

Key pair

A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name

hive_case_study_key

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

File format

● pem
For use with OpenSSH

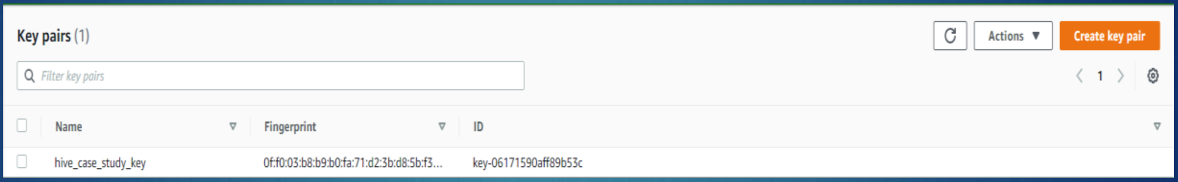
○ ppk
For use with PuTTY

Tags (Optional)

No tags associated with the resource.

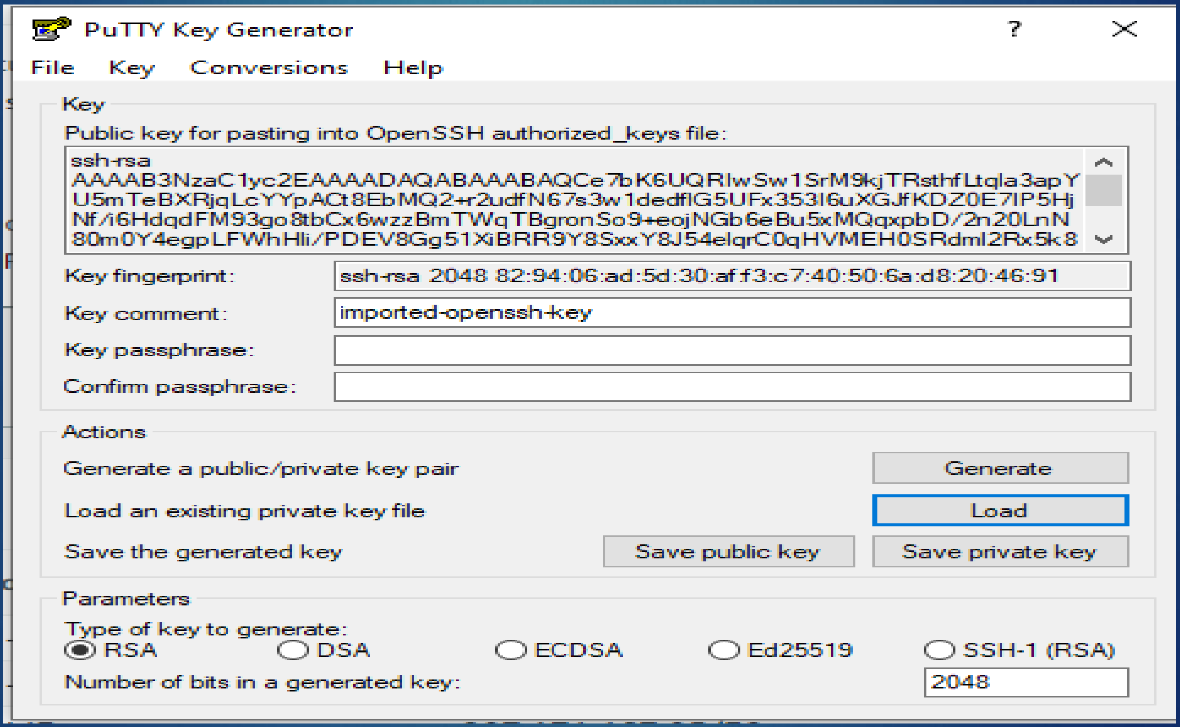
Add tag

You can add 50 more tags.



Key pairs (1)

	Name	Fingerprint	ID
□	hive_case_study_key	0f:f0:03:b8:b9:b0:fa:71:d2:3b:d8:5b:f3...	key-06171590aff89b53c



Key

Public key for pasting into OpenSSH authorized_keys file:

```
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCe7bK6UQRIwSw1SrM9kjTRsthfLtqla3apY
U5mTeBXRjqLcYYpACt8EbMQ2+r2udfN67s3w1dedfIG5UFx353I6uXGJfKDZ0E7IP5Hj
Nf/i6HdqdFM93go8tbCx6wzzBmTWqTBgronSo9+eojNGb6eBu5xMQqxpbD/2n20LnN
80m0Y4egpLFWhHli/PDEV8Gg51XiBRR9Y8SxxY8J54elqrC0qHVMEH0SRdml2Rx5k8
```

Key fingerprint: ssh-rsa 2048 82:94:06:ad:5d:30:af:f3:c7:40:50:6a:d8:20:46:91

Key comment: imported-openssh-key

Key passphrase:

Confirm passphrase:

Actions

Generate a public/private key pair — Generate

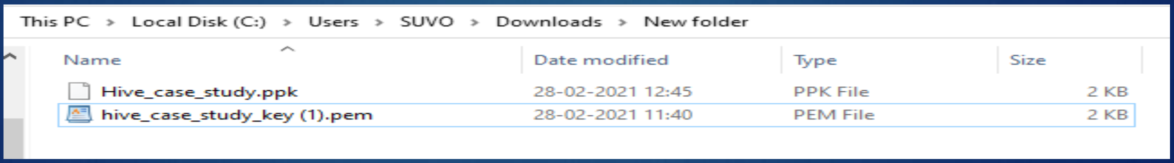
Load an existing private key file — Load

Save the generated key — Save public key — Save private key

Parameters

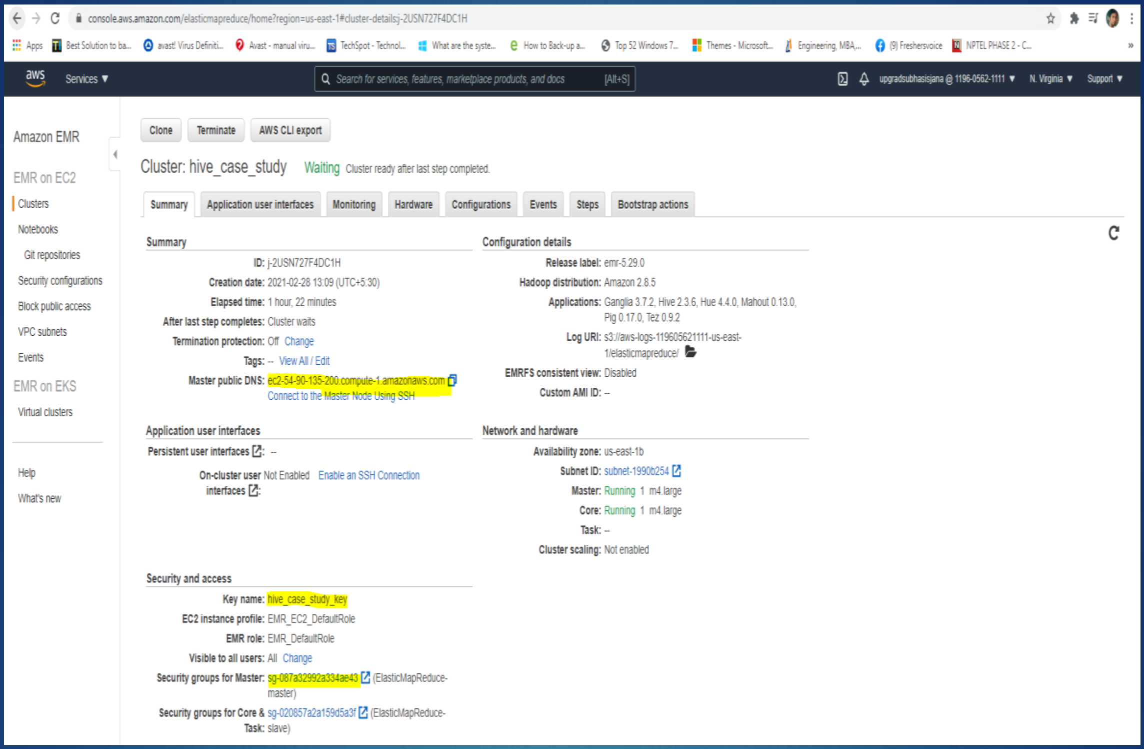
Type of key to generate:
● RSA ○ DSA ○ ECDSA ○ Ed25519 ○ SSH-1 (RSA)

Number of bits in a generated key: 2048


This PC > Local Disk (C:) > Users > SUVO > Downloads > New folder

Name	Date modified	Type	Size
Hive_case_study.ppk	28-02-2021 12:45	PPK File	2 KB
hive_case_study_key (1).pem	28-02-2021 11:40	PEM File	2 KB

EMR cluster creation and configuration ‘Hive_Case_Study’



Cluster: hive_case_study Waiting Cluster ready after last step completed.

Summary

ID: j-2USN727F4DC1H
Creation date: 2021-02-28 13:09 (UTC+5:30)
Elapsed time: 1 hour, 22 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-54-90-135-200.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-119605621111-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

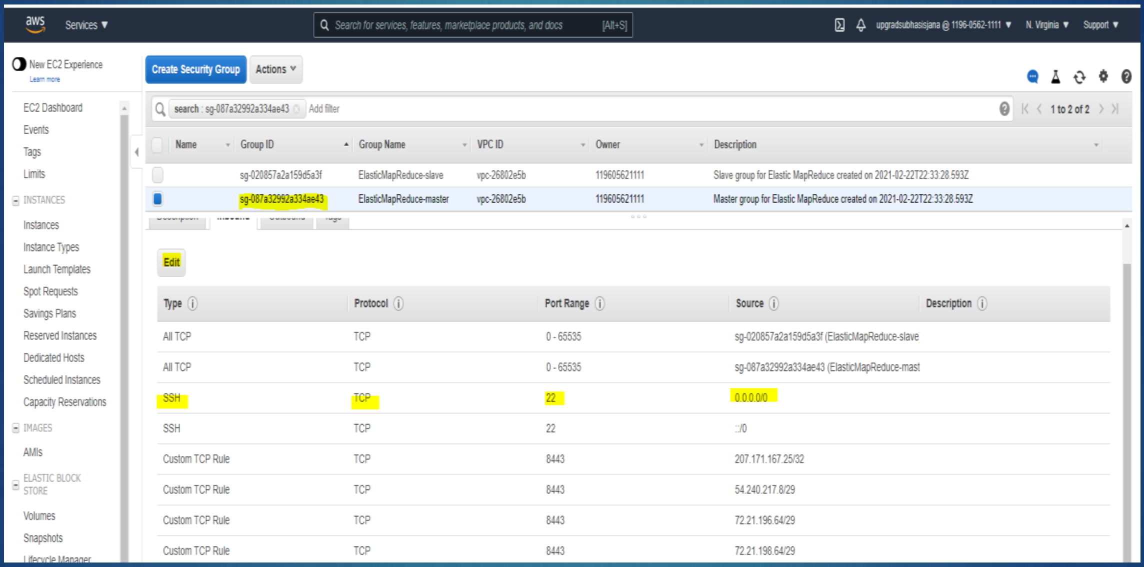
Persistent user interfaces: --
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Network and hardware

Availability zone: us-east-1b
Subnet ID: subnet-1990b254
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: hive_case_study_key
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-087a32992a334ae43 (ElasticMapReduce-master)
Security groups for Core & Task: sg-020857a2a159d5a3f (ElasticMapReduce-slave)

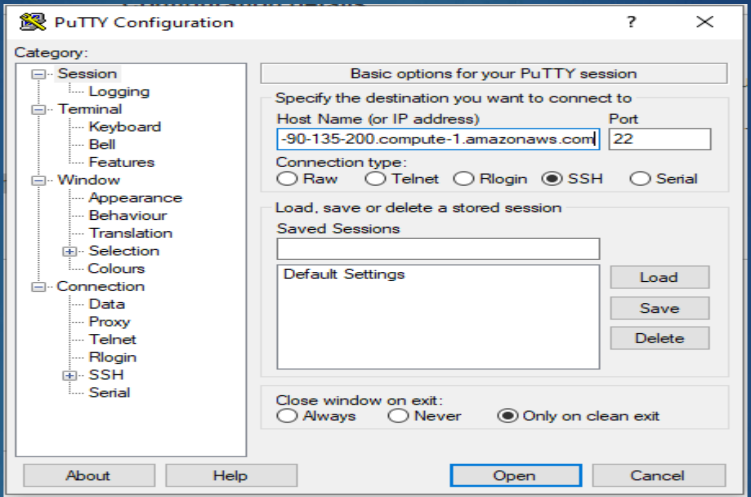


search : sg-087a32992a334ae43

Name	Group ID	Group Name	VPC ID	Owner	Description
	sg-020857a2a159d5a3f	ElasticMapReduce-slave	vpc-26802e5b	119605621111	Slave group for Elastic MapReduce created on 2021-02-22T22:33:28.593Z
	sg-087a32992a334ae43	ElasticMapReduce-master	vpc-26802e5b	119605621111	Master group for Elastic MapReduce created on 2021-02-22T22:33:28.593Z

Type	Protocol	Port Range	Source	Description
All TCP	TCP	0 - 65535	sg-020857a2a159d5a3f (ElasticMapReduce-slave	
All TCP	TCP	0 - 65535	sg-087a32992a334ae43 (ElasticMapReduce-mast	
SSH	TCP	22	0.0.0.0/0	
SSH	TCP	22	::/0	
Custom TCP Rule	TCP	8443	207.171.167.25/32	
Custom TCP Rule	TCP	8443	54.240.217.8/29	
Custom TCP Rule	TCP	8443	72.21.196.64/29	
Custom TCP Rule	TCP	8443	72.21.198.64/29	

Starting terminal using Putty



Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): -90-135-200.compute-1.amazonaws.com

Port: 22

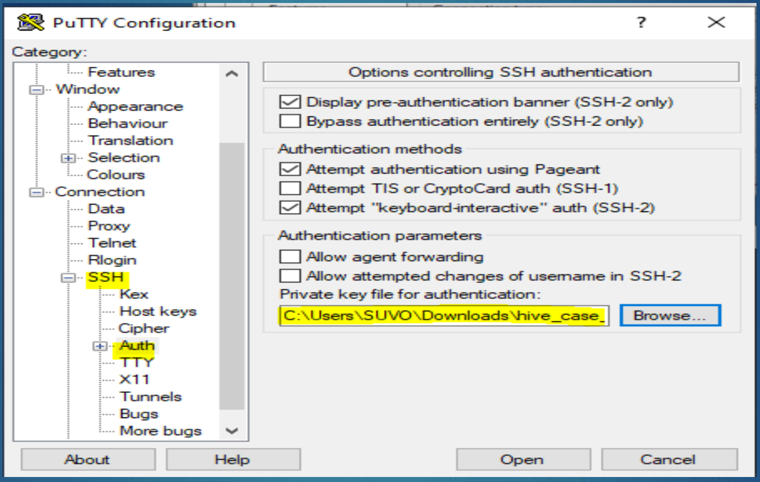
Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions:

Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit



Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)

□ Bypass authentication entirely (SSH-2 only)

Authentication methods

☑ Attempt authentication using Pageant

□ Attempt TIS or CryptoCard auth (SSH-1)

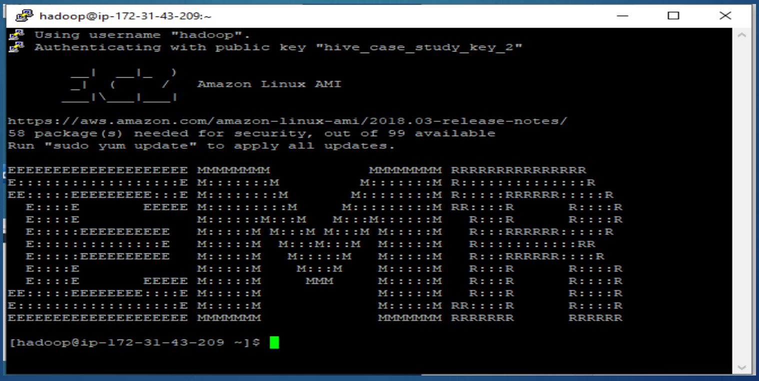
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

□ Allow agent forwarding

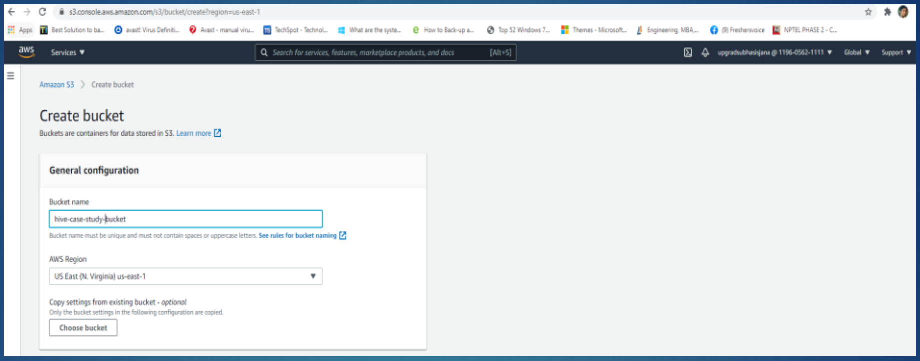
□ Allow attempted changes of username in SSH-2

Private key file for authentication: C:\Users\SUVO\Downloads\hive_case_



```
Using username "hadoop".
Authenticating with public key "hive_case_study_key_2"

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 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-172-31-43-209 ~]$
```

S3 Bucket Creation and to store data files



Create bucket

Buckets are containers for data stored in S3. Learn more

General configuration

Bucket name

hive-case-study-bucket

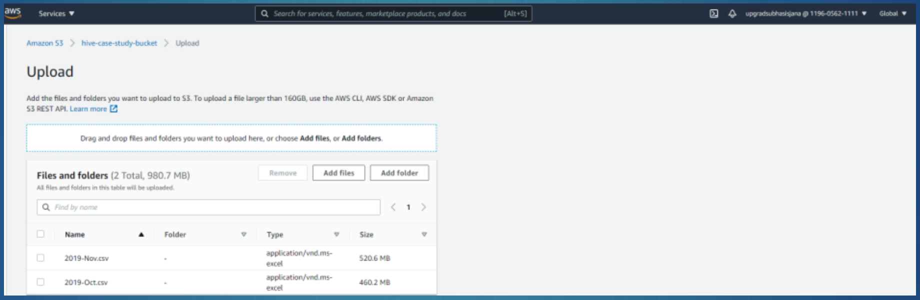
Bucket name must be unique and must not contain spaces or uppercase letters. **See rules for bucket naming**

AWS Region

US East (N. Virginia) us-east-1 ▼

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.


aws
Services ▼
Search for services, features, marketplace products, and docs [Alt+S]
upgradsubhasisjana @ 1196-0562-1111 ▼
Global ▼
Amazon S3 > hive-case-study-bucket > Upload

Upload

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (2 Total, 980.7 MB)

All files and folders in this table will be uploaded.

Remove
Add files
Add folder
Find by name
1

□	Name ▲	Folder	Type	Size
□	2019-Nov.csv	-	application/vnd.ms-excel	520.6 MB
□	2019-Oct.csv	-	application/vnd.ms-excel	460.2 MB

1. Command to check for already present directories in HDFS

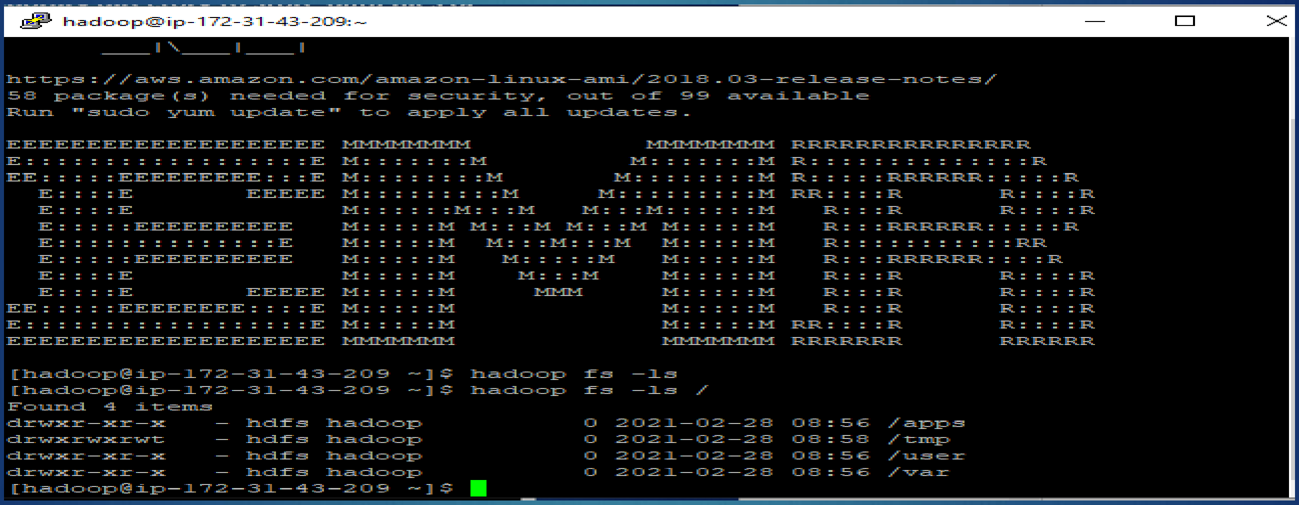
- hadoop fs -ls /

Output:

```
Found 4 items

drwxr-xr-x - hdfs hadoop          0 2021-02-28 17:34 /apps

drwxrwxrwt - hdfs hadoop          0 2021-02-28 17:36 /tmp

drwxr-xr-x - hdfs hadoop          0 2021-02-28 17:34 /user

drwxr-xr-x - hdfs hadoop          0 2021-02-28 17:34 /var
```

Insights:

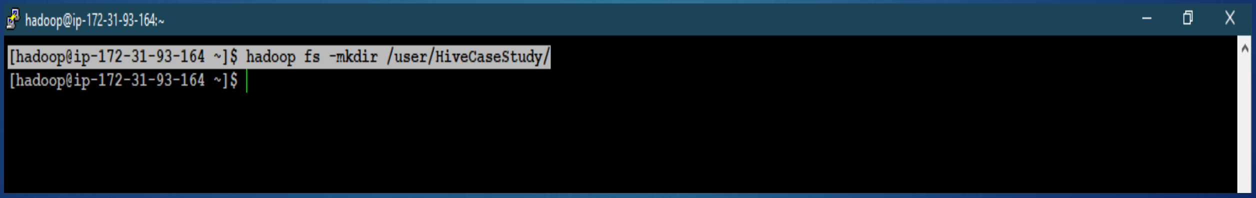
- All the **above directories are in-built in HDFS.**
- **Either these directories can be used to create our temporary directory to store data files or create a separate temporary directory.**



```
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
58 package(s) needed for security, out of 99 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-172-31-43-209 ~]$ hadoop fs -ls
[hadoop@ip-172-31-43-209 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2021-02-28 08:56 /apps
drwxrwxrwt   - hdfs hadoop          0 2021-02-28 08:58 /tmp
drwxr-xr-x   - hdfs hadoop          0 2021-02-28 08:56 /user
drwxr-xr-x   - hdfs hadoop          0 2021-02-28 08:56 /var
[hadoop@ip-172-31-43-209 ~]$
```

2. Creating new temporary directory i.e., 'HiveCaseStudy' to store data file in the already present directory (Permanent) i.e., 'user'

- hadoop fs -mkdir /user/HiveCaseStudy/



```
[hadoop@ip-172-31-93-164 ~]$ hadoop fs -mkdir /user/HiveCaseStudy/
[hadoop@ip-172-31-93-164 ~]$
```

3. Command to check creation of new temporary Directory in ‘user’ directory

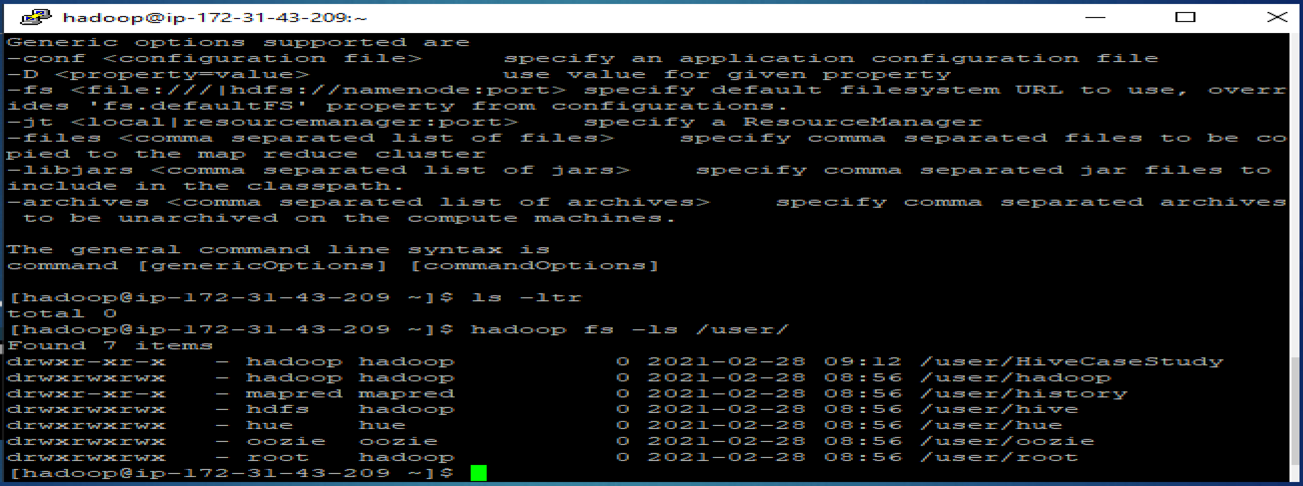
- hadoop fs -ls /user/

Output:

Found 7 items

drwxr-xr-x	-	hadoop	hadoop	0	28-02-2021	09:12	/user/HiveCaseStudy
drwxrwxrwx	-	hadoop	hadoop	0	28-02-2021	08:56	/user/hadoop
drwxr-xr-x	-	mapred	mapred	0	28-02-2021	08:56	/user/history
drwxrwxrwx	-	hdfs	hadoop	0	28-02-2021	08:56	/user/hive
drwxrwxrwx	-	hue	hue	0	28-02-2021	08:56	/user/hue
drwxrwxrwx	-	oozie	oozie	0	28-02-2021	08:56	/user/oozie
drwxrwxrwx	-	root	hadoop	0	28-02-2021	08:56	/user/root

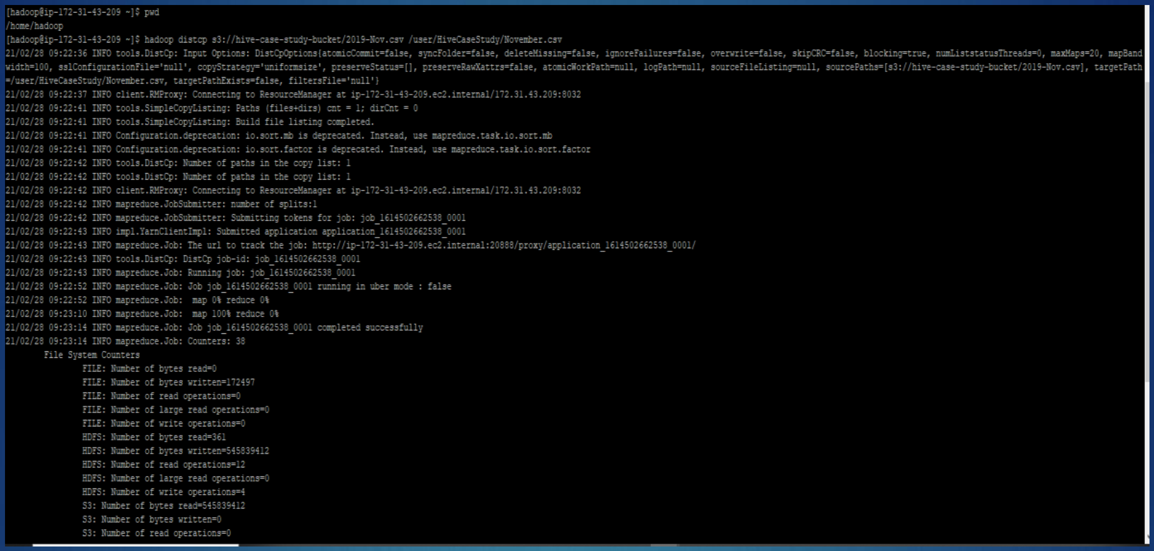
- There will always be some files within the permanent directories of the HDFS.



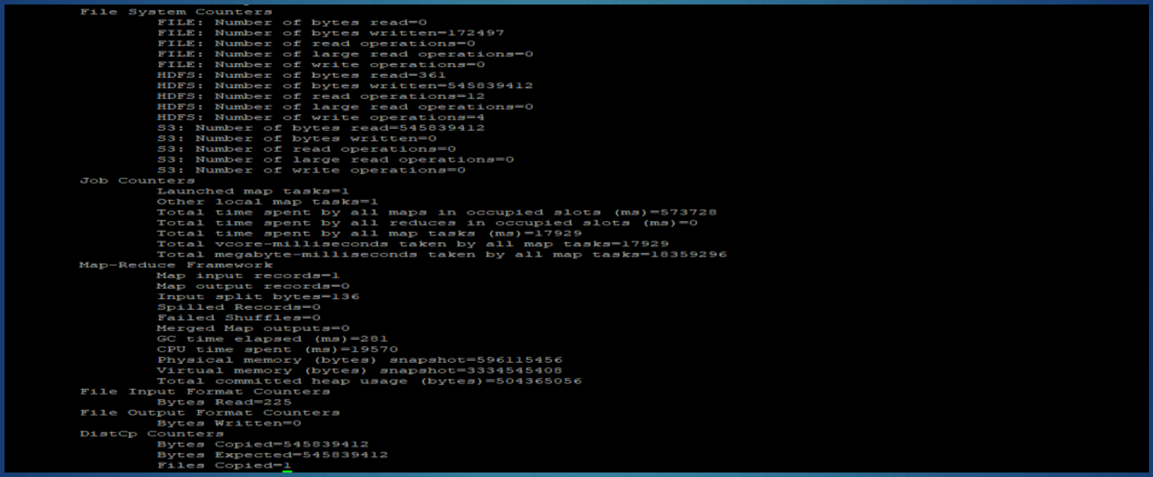
```
Generic options supported are
-conf <configuration file>     specify an application configuration file
-D <property=value>            use value for given property
-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overr
ides 'fs.defaultFS' property from configurations.
-jt <local|resourcemanager:port>    specify a ResourceManager
-files <comma separated list of files>    specify comma separated files to be co
pied to the map reduce cluster
-libjars <comma separated list of jars>    specify comma separated jar files to
include in the classpath.
-archives <comma separated list of archives>    specify comma separated archives
 to be unarchived on the compute machines.

The general command line syntax is
command [genericOptions] [commandOptions]

[hadoop@ip-172-31-43-209 ~]$ ls -ltr
total 0
[hadoop@ip-172-31-43-209 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x   - hadoop hadoop          0 2021-02-28 09:12 /user/HiveCaseStudy
drwxrwxrwx   - hadoop hadoop          0 2021-02-28 08:56 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2021-02-28 08:56 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2021-02-28 08:56 /user/hive
drwxrwxrwx   - hue    hue             0 2021-02-28 08:56 /user/hue
drwxrwxrwx   - oozie  oozie           0 2021-02-28 08:56 /user/oozie
drwxrwxrwx   - root   hadoop          0 2021-02-28 08:56 /user/root
[hadoop@ip-172-31-43-209 ~]$
```

4. Command to load 1st data file '2019-Nov.csv' from S3 storage into HDFS storage as ‘Novemver.csv'

hadoop distcp s3://hive-case-study-bucket/2019-Nov.csv /user/HiveCaseStudy/November.csv



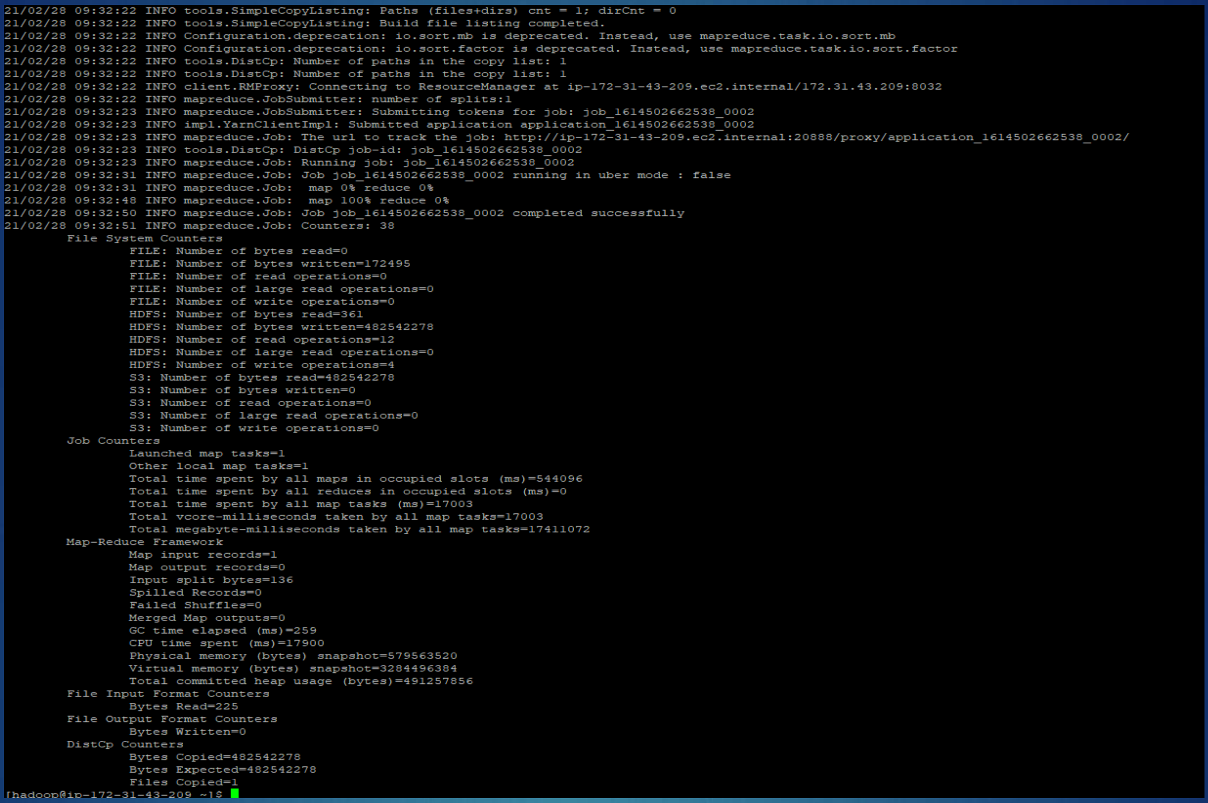
```
[hadoop@ip-172-31-43-209 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-43-209 ~]$ hadoop distcp s3://hive-case-study-bucket/2019-Nov.csv /user/HiveCaseStudy/November.csv
21/02/28 09:22:36 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBand
width=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://hive-case-study-bucket/2019-Nov.csv], targetPath
=/user/HiveCaseStudy/November.csv, targetPathExists=false, filtersFile='null'}
21/02/28 09:22:37 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-43-209.ec2.internal/172.31.43.209:8032
21/02/28 09:22:41 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/02/28 09:22:41 INFO tools.SimpleCopyListing: Build file listing completed.
21/02/28 09:22:41 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/02/28 09:22:41 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/02/28 09:22:42 INFO tools.DistCp: Number of paths in the copy list: 1
21/02/28 09:22:42 INFO tools.DistCp: Number of paths in the copy list: 1
21/02/28 09:22:42 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-43-209.ec2.internal/172.31.43.209:8032
21/02/28 09:22:42 INFO mapreduce.JobSubmitter: number of splits:1
21/02/28 09:22:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1614502662538_0001
21/02/28 09:22:43 INFO impl.YarnClientImpl: Submitted application application_1614502662538_0001
21/02/28 09:22:43 INFO mapreduce.Job: The url to track the job: http://ip-172-31-43-209.ec2.internal:20888/proxy/application_1614502662538_0001/
21/02/28 09:22:43 INFO tools.DistCp: DistCp job-id: job_1614502662538_0001
21/02/28 09:22:43 INFO mapreduce.Job: Running job: job_1614502662538_0001
21/02/28 09:22:52 INFO mapreduce.Job: Job job_1614502662538_0001 running in uber mode : false
21/02/28 09:22:52 INFO mapreduce.Job:  map 0% reduce 0%
21/02/28 09:23:10 INFO mapreduce.Job:  map 100% reduce 0%
21/02/28 09:23:14 INFO mapreduce.Job: Job job_1614502662538_0001 completed successfully
21/02/28 09:23:14 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172497
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=361
                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
```



```
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172497
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=361
                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)=573728
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=17929
                Total vcore-milliseconds taken by all map tasks=17929
                Total megabyte-milliseconds taken by all map tasks=18359296
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=281
                CPU time spent (ms)=19570
                Physical memory (bytes) snapshot=596115456
                Virtual memory (bytes) snapshot=3334545408
                Total committed heap usage (bytes)=504365056
        File Input Format Counters
                Bytes Read=225
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
```

5. Command to load 2nd data file '2019-Oct.csv' from S3 storage into HDFS storage as ‘October.csv'

-hadoop distcp s3://hive-case-study-bucket/2019-Oct.csv /user/HiveCaseStudy/October.csv



```
21/02/28 09:32:22 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/02/28 09:32:22 INFO tools.SimpleCopyListing: Build file listing completed.
21/02/28 09:32:22 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/02/28 09:32:22 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/02/28 09:32:22 INFO tools.DistCp: Number of paths in the copy list: 1
21/02/28 09:32:22 INFO tools.DistCp: Number of paths in the copy list: 1
21/02/28 09:32:22 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-43-209.ec2.internal/172.31.43.209:8032
21/02/28 09:32:22 INFO mapreduce.JobSubmitter: number of splits:1
21/02/28 09:32:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1614502662538_0002
21/02/28 09:32:23 INFO impl.YarnClientImpl: Submitted application application_1614502662538_0002
21/02/28 09:32:23 INFO mapreduce.Job: The url to track the job: http://ip-172-31-43-209.ec2.internal:20888/proxy/application_1614502662538_0002/
21/02/28 09:32:23 INFO tools.DistCp: DistCp job-id: job_1614502662538_0002
21/02/28 09:32:23 INFO mapreduce.Job: Running job: job_1614502662538_0002
21/02/28 09:32:31 INFO mapreduce.Job: Job job_1614502662538_0002 running in uber mode : false
21/02/28 09:32:31 INFO mapreduce.Job:  map 0% reduce 0%
21/02/28 09:32:48 INFO mapreduce.Job:  map 100% reduce 0%
21/02/28 09:32:50 INFO mapreduce.Job: Job job_1614502662538_0002 completed successfully
21/02/28 09:32:51 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172495
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=361
                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)=544096
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=17003
                Total vcore-milliseconds taken by all map tasks=17003
                Total megabyte-milliseconds taken by all map tasks=17411072
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=259
                CPU time spent (ms)=17900
                Physical memory (bytes) snapshot=579563520
                Virtual memory (bytes) snapshot=3284496384
                Total committed heap usage (bytes)=491257856
        File Input Format Counters
                Bytes Read=225
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=482542278
                Bytes Expected=482542278
                Files Copied=1
[hadoop@ip-172-31-43-209 ~]$
```

6. Command to check successful loading of data files into the already created new temporary directory of HDFS i.e., 'HiveCaseStudy'

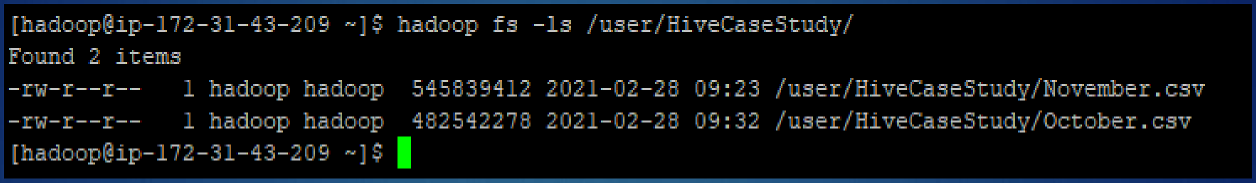
- hadoop fs -ls /user/HiveCaseStudy/

Output:

Found 2 items

-rw-r--r-- 1 hadoop hadoop 545839412 2021-02-28 14:54 /user/HiveCaseStudy/November.csv

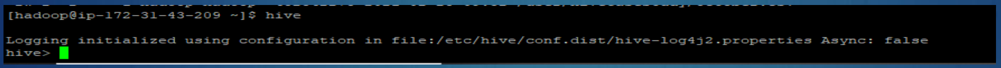
-rw-r--r-- 1 hadoop hadoop 482542278 2021-02-28 14:51 /user/HiveCaseStudy/October.csv



```
[hadoop@ip-172-31-43-209 ~]$ hadoop fs -ls /user/HiveCaseStudy/
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2021-02-28 09:23 /user/HiveCaseStudy/November.csv
-rw-r--r--   1 hadoop hadoop  482542278 2021-02-28 09:32 /user/HiveCaseStudy/October.csv
[hadoop@ip-172-31-43-209 ~]$
```

7. Command to start Hive system

- hive



```
[hadoop@ip-172-31-43-209 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

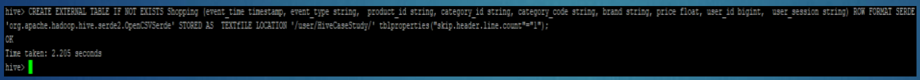
8. Creating an External table i.e., 'Shopping' which will hold the data for both the data files stored in temporary directory of HDFS.

- CREATE EXTERNAL TABLE IF NOT EXISTS Shopping (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/HiveCaseStudy/' tblproperties("skip.header.line.count"="1");

Output:

OK

Time taken: 2.205 seconds



```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Shopping (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/HiveCaseStudy/' tblproperties("skip.header.line.count"="1");
OK
Time taken: 2.205 seconds
hive>
```

9. Command to enable heading in the output

- set hive.cli.print.header=True;

10. Simple HiveQL command to check successful creation of table and storage of data from both data files into table

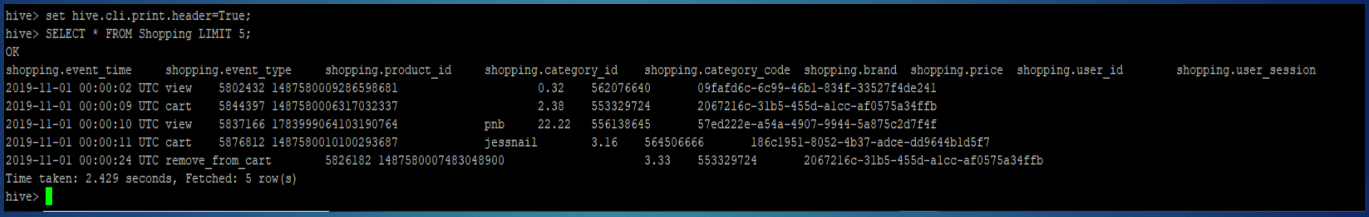
Query:

SELECT * FROM Shopping LIMIT 5;

Output:

OK

```
shopping.event_time     shopping.event_type     shopping.product_id     shopping.category_id     shopping.category_code  shopping.brand
shopping.price  shopping.user_id       shopping.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-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: 2.429 seconds, Fetched: 5 row(s)
```



```
hive> set hive.cli.print.header=True;
hive> SELECT * FROM Shopping LIMIT 5;
OK
shopping.event_time     shopping.event_type     shopping.product_id     shopping.category_id    shopping.category_code  shopping.brand  shopping.price  shopping.user_id        shopping.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-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: 2.429 seconds, Fetched: 5 row(s)
hive>
```

Questions

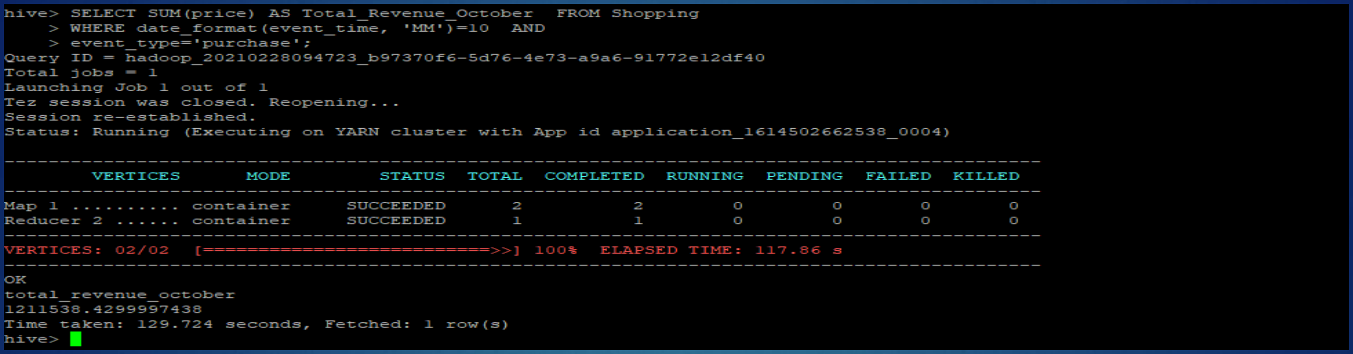
Question 1: Find the total revenue generated due to purchases made in October.

Query:

```
SELECT SUM(price) AS Total_Revenue_October

FROM Shopping

WHERE date_format(event_time, 'MM')=10

AND

event_type='purchase';
```

Output:

```
Query ID = hadoop_20210228094723_b97370f6-5d76-4e73-a9a6-91772e12df40
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_1614502662538_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: 117.86 s
----------------------------------------------------------------------------------------------
OK
total_revenue_october
1211538.4299997438
Time taken: 129.724 seconds, Fetched: 1 row(s)
```



```
hive> SELECT SUM(price) AS Total_Revenue_October  FROM Shopping
    > WHERE date_format(event_time, 'MM')=10  AND
    > event_type='purchase';
Query ID = hadoop_20210228094723_b97370f6-5d76-4e73-a9a6-91772e12df40
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_1614502662538_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: 117.86 s
----------------------------------------------------------------------------------------------
OK
total_revenue_october
1211538.4299997438
Time taken: 129.724 seconds, Fetched: 1 row(s)
hive>
```

Insights:

- The **total revenue** generated **based on Purchase in** the month of **October of 2019 was 1,211,538.43** /-.

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

Query:

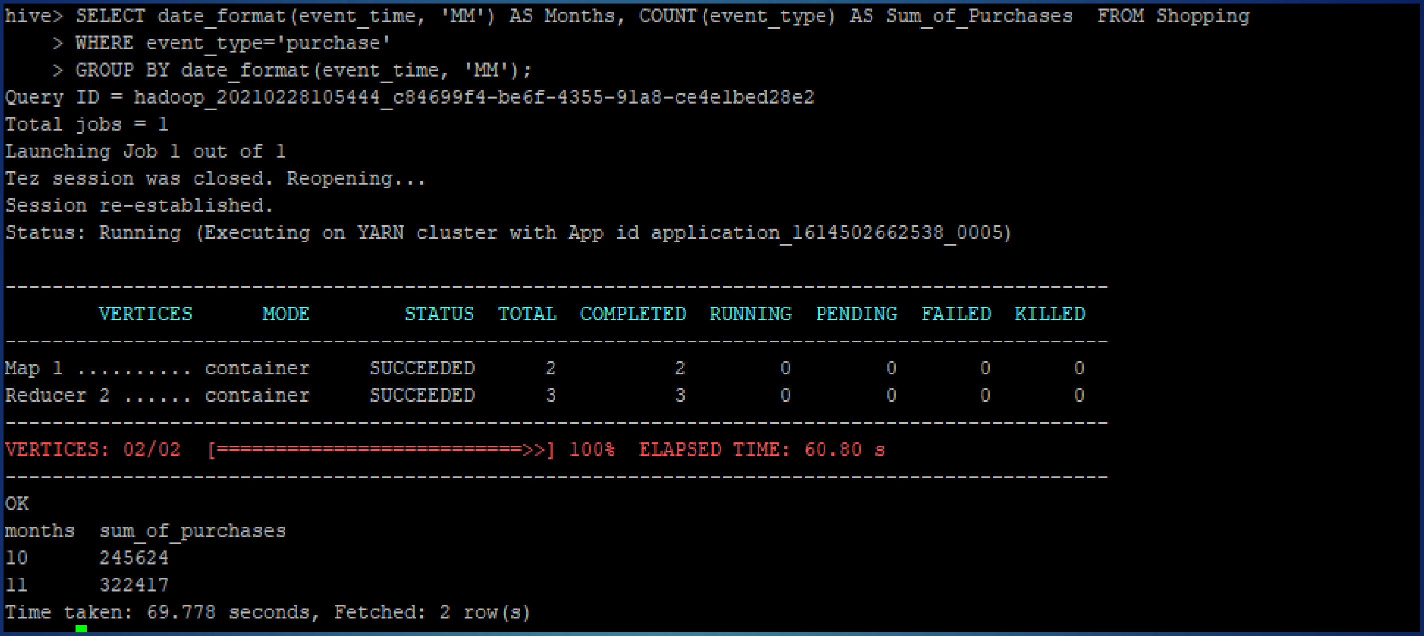
SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases FROM

Shopping

WHERE event_type='purchase'

GROUP BY date_format(event_time, 'MM');

```
Query ID = hadoop_20210228105444_c84699f4-be6f-4355-91a8-ce4e1bed28e2
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_1614502662538_0005)

----------------------------------------------------------------------------------------------
    VERTICES     MODE     STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container    SUCCEEDED    2       2      0      0      0      0
Reducer 2 ...... container    SUCCEEDED    3       3      0      0      0      0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 60.80 s
----------------------------------------------------------------------------------------------
OK
months  sum_of_purchases
10      245624
11      322417
Time taken: 69.778 seconds, Fetched: 2 row(s)
```



```
hive> SELECT date_format(event_time, 'MM') AS Months, COUNT(event_type) AS Sum_of_Purchases  FROM Shopping
    > WHERE event_type='purchase'
    > GROUP BY date_format(event_time, 'MM');
Query ID = hadoop_20210228105444_c84699f4-be6f-4355-91a8-ce4e1bed28e2
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_1614502662538_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 60.80 s
----------------------------------------------------------------------------------------------
OK
months  sum_of_purchases
10      245624
11      322417
Time taken: 69.778 seconds, Fetched: 2 row(s)
```

Insights:

- It seems to be that there was **more purchase made in** the month of **November (11) i.e., 322,417 than** in the month of **October (10) i.e., 245,624**.
- Looking at these figures **we could assume that the month of November must be more profitable than the month of October**. But we can verify our assumption by conducting further investigations.

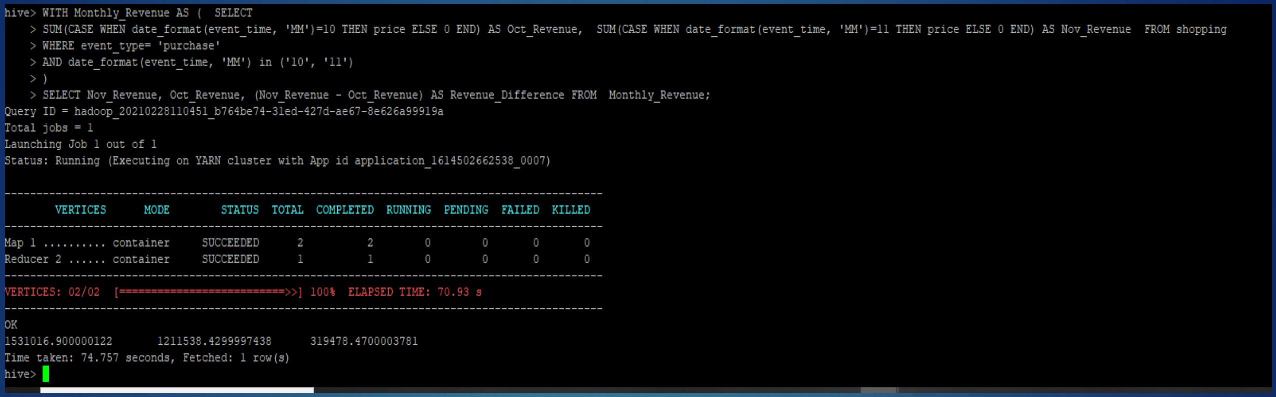
Question 3: Write a query to find the change in revenue generated due to purchases from October to November.

Query:

```
WITH Monthly_Revenue AS (

SELECT

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 shopping

WHERE event_type= 'purchase'

AND date_format(event_time, 'MM') in ('10', '11')

)

SELECT Nov_Revenue, Oct_Revenue, (Nov_Revenue - Oct_Revenue) AS Revenue_Difference FROM
Monthly_Revenue;
```

Output:

```
Query ID = hadoop_20210228110451_b764be74-31ed-427d-ae67-8e626a99919a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614502662538_0007)

----------------------------------------------------------------------------------------------
     VERTICES     MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2        2       0       0       0       0
Reducer 2 ...... container     SUCCEEDED      1        1       0       0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.93 s
----------------------------------------------------------------------------------------------
OK
1531016.900000122       1211538.4299997438      319478.4700003781
Time taken: 74.757 seconds, Fetched: 1 row(s)
```



```
hive> WITH Monthly_Revenue AS (  SELECT
    > 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 shopping
    > WHERE event_type= 'purchase'
    > AND date_format(event_time, 'MM') in ('10', '11')
    > )
    > SELECT Nov_Revenue, Oct_Revenue, (Nov_Revenue - Oct_Revenue) AS Revenue_Difference FROM  Monthly_Revenue;
Query ID = hadoop_20210228110451_b764be74-31ed-427d-ae67-8e626a99919a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614502662538_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 70.93 s
----------------------------------------------------------------------------------------------
OK
1531016.900000122       1211538.4299997438      319478.4700003781
Time taken: 74.757 seconds, Fetched: 1 row(s)
hive>
```

Insights:

- On the basis of the results **considering purchase as event**, we could conclude that the **revenue** generated **in November of 2019 was more than** the revenue generated **in** the month of **October**. In other words, **November was more profitable for the company than October**.
- Company had a better sale in November, 2019.

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

Query:

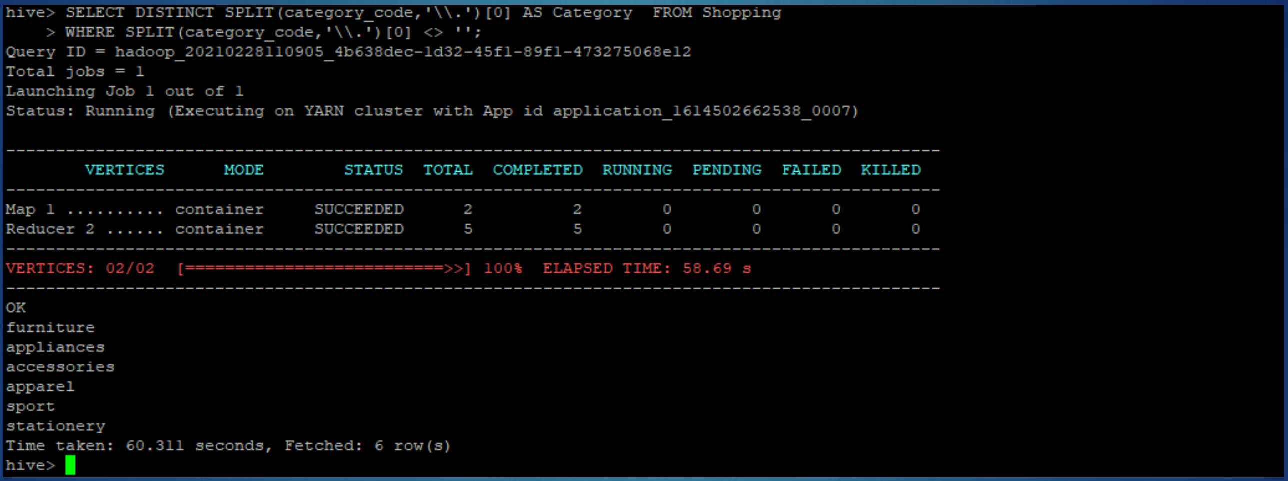
SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category

FROM Shopping

WHERE SPLIT(category_code,'\\.')[0] <> '';

Output

```
Query ID = hadoop_20210228110905_4b638dec-1d32-45f1-89f1-473275068e12
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614502662538_0007)

----------------------------------------------------------------------------------------------
      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: 02/02  [==========================>>] 100%  ELAPSED TIME: 58.69 s
----------------------------------------------------------------------------------------------
OK
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 60.311 seconds, Fetched: 6 row(s)
```



```
hive> SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category  FROM Shopping
    > WHERE SPLIT(category_code,'\\.')[0] <> '';
Query ID = hadoop_20210228110905_4b638dec-1d32-45f1-89f1-473275068e12
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614502662538_0007)

----------------------------------------------------------------------------------------------
        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: 02/02  [==========================>>] 100%  ELAPSED TIME: 58.69 s
----------------------------------------------------------------------------------------------
OK
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 60.311 seconds, Fetched: 6 row(s)
hive>
```

Insights:

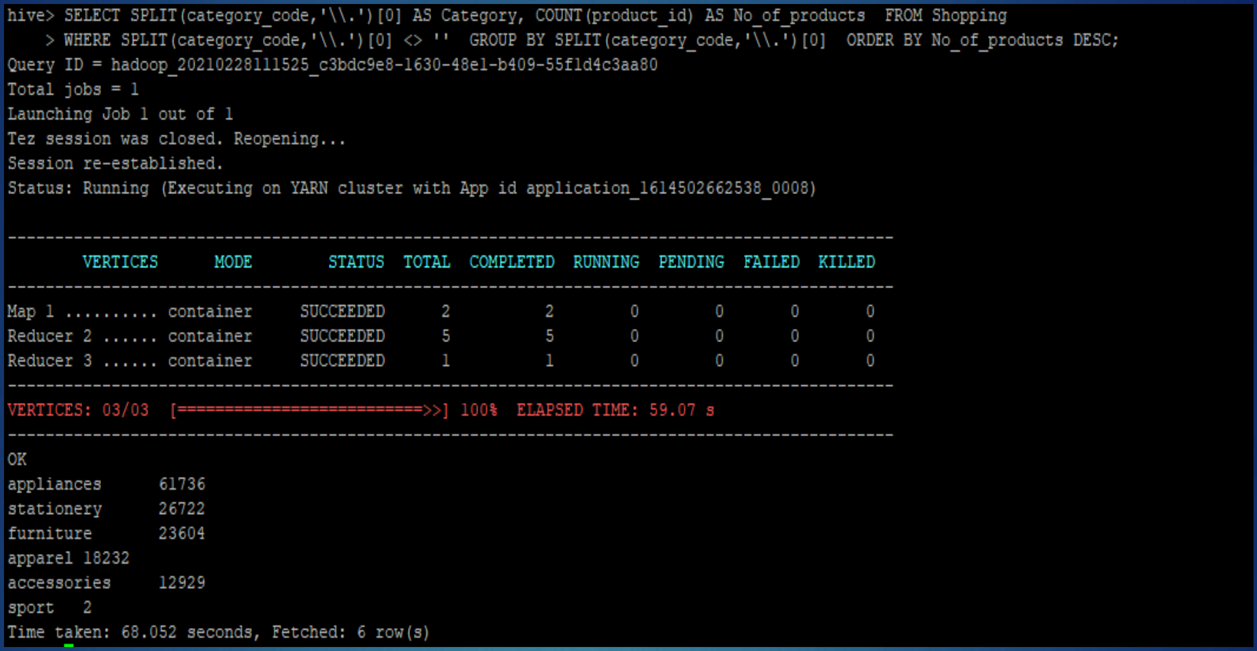
- There is total **6 different categories** under which company sells their different products.

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

Query:

```
SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products

FROM Shopping

WHERE SPLIT(category_code,'\\.')[0] <> ''

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

ORDER BY No_of_products DESC;
```

```
Query ID = hadoop_20210228111525_c3bdc9e8-1630-48e1-b409-55f1d4c3aa80
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_1614502662538_0008)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 59.07 s
----------------------------------------------------------------------------------------------
OK
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 68.052 seconds, Fetched: 6 row(s)
```



```
hive> SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_products  FROM Shopping
    > WHERE SPLIT(category_code,'\\.')[0] <> ''  GROUP BY SPLIT(category_code,'\\.')[0]  ORDER BY No_of_products DESC;
Query ID = hadoop_20210228111525_c3bdc9e8-1630-48e1-b409-55f1d4c3aa80
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_1614502662538_0008)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 59.07 s
----------------------------------------------------------------------------------------------
OK
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 68.052 seconds, Fetched: 6 row(s)
```

Insights:

- Company has **more products registered under Appliances category i.e., 61,736 products** than any other categories.
- Then it is followed by **stationery as second with 26,722 products, furniture as third with 23,604 products, apparel as fourth with 18232 products** registered, **accessories as fifth with 12929 products**.
- **Sports category has only 2 products** registered. This **must be due to low cosmetic products in the sports market**.

Question 6: Which brand had the maximum sales in October and November combined?

Query:

WITH Max_Sales_Brand AS (

SELECT brand,

```
SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Sales,

SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Sales

FROM Shopping

WHERE (

event_type='purchase'

AND

date_format(event_time, 'MM') in ('10','11')

AND

brand <> '')

GROUP BY brand

)

SELECT brand, Nov_Sales + Oct_Sales AS Total_Sales

FROM Max_Sales_Brand

ORDER BY Total_Sales DESC

LIMIT 1;
```

Output:

```
Query ID = hadoop_20210220155441_e5643e59-8162-4068-a271-a8e536398dbc

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1610894517504_0006)


----------------------------------------------------------------------------------------------

        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED

----------------------------------------------------------------------------------------------

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

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

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

----------------------------------------------------------------------------------------------
```

VERTICES: 03/03 [==========================>>] 100% ELAPSED TIME: 63.74 s

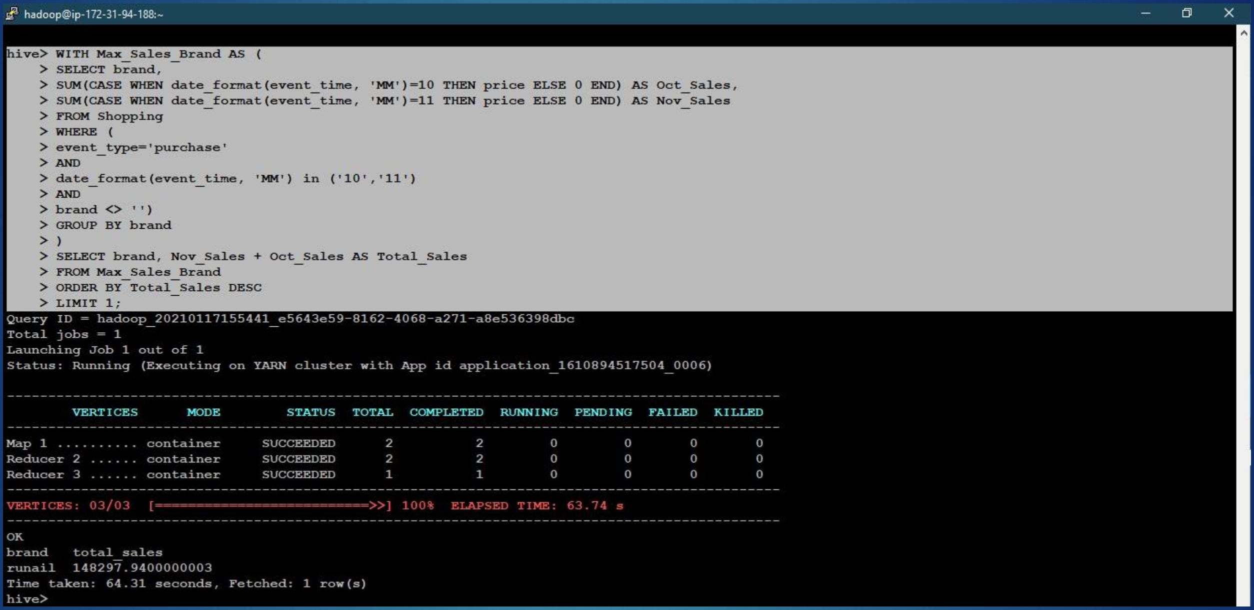
- -

OK

brand total_sales

runail 148297.9400000003

Time taken: 64.31 seconds, Fetched: 1 row(s)



```
hive> WITH Max_Sales_Brand AS (
    > SELECT brand,
    > SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Sales,
    > SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Sales
    > FROM Shopping
    > WHERE (
    > event_type='purchase'
    > AND
    > date_format(event_time, 'MM') in ('10','11')
    > AND
    > brand <> '')
    > GROUP BY brand
    > )
    > SELECT brand, Nov_Sales + Oct_Sales AS Total_Sales
    > FROM Max_Sales_Brand
    > ORDER BY Total_Sales DESC
    > LIMIT 1;
Query ID = hadoop_20210117155441_e5643e59-8162-4068-a271-a8e536398dbc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610894517504_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 63.74 s
----------------------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.9400000003
Time taken: 64.31 seconds, Fetched: 1 row(s)
hive>
```

Insights:

- **Runail** is the **brand** that **has highest / maximum sales** in the month of **October and November of 2019 combined.**
- It seems that **Runail brand has high popularity among cosmetic lovers** and bringing in **more products related to Runail brand could help in increasing their profit**.

Question 7: Which brands increased their sales from October to November?

Query:

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 Shopping
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;
```

Output:

```
Query ID = hadoop_20210220155852_282b0369-324c-4c04-91c0-102abc59add0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610894517504_0006)

-----------------------------------------------------------------------------------
     VERTICES    MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----------------------------------------------------------------------------------
Map 1 .......... container    SUCCEEDED     2        2      0       0      0      0
Reducer 2 ...... container    SUCCEEDED     2        2      0       0      0      0
Reducer 3 ...... container    SUCCEEDED     1        1      0       0      0      0
-----------------------------------------------------------------------------------
VERTICES: 03/03 [==========================>>] 100% ELAPSED TIME: 69.69 s
-----------------------------------------------------------------------------------
OK
brand  oct_revenue        nov_revenue      sales_difference
ovale    2.54             3.1              0.56
```

cosima	20.23	20.929999999999993	0.6999999999999922
grace	100.92000000000002	102.61000000000001	1.6899999999999977
helloganic	0.0	3.1	3.1
skinity	8.88	12.440000000000001	3.5600000000000005
bodyton	1376.3399999999974	1380.6399999999992	4.3000000000017735
moyou	5.71	10.280000000000001	4.570000000000001
neoleor	43.41	51.7	8.290000000000006
soleo	204.2000000000003	212.5299999999998	8.329999999999501
jaguar	1102.11	1110.6500000000003	8.540000000000418
tertio	236.16000000000008	245.79999999999978	9.639999999999702
fly	17.14	27.17	10.030000000000001
rasyan	18.799999999999997	28.939999999999994	10.139999999999997
deoproce	316.84	329.1700000000001	12.330000000000098
barbie	0.0	12.39	12.39
supertan	50.37000000000001	66.51000000000002	16.140000000000008
treaclemoon	163.36999999999995	181.48999999999995	18.120000000000005
kamill	63.00999999999999	81.49000000000002	18.480000000000032
juno	0.0	21.08	21.08
veraclara	50.109999999999985	71.21000000000001	21.100000000000023
glysolid	69.72999999999998	91.58999999999997	21.86
godefroy	401.2200000000002	425.12000000000006	23.899999999999864
binacil	0.0	24.259999999999998	24.259999999999998
blixz	38.949999999999996	63.39999999999998	24.44999999999998
profepil	93.36000000000003	118.02000000000005	24.660000000000025
estelare	444.80999999999943	471.8700000000009	27.06000000000148
orly	902.3800000000005	931.0900000000003	28.70999999999981
biore	60.650000000000006	90.31	29.659999999999997
beautyblender	78.74000000000001	109.41	30.669999999999987
vilenta	197.6000000000002	231.2100000000002	33.610000000000014

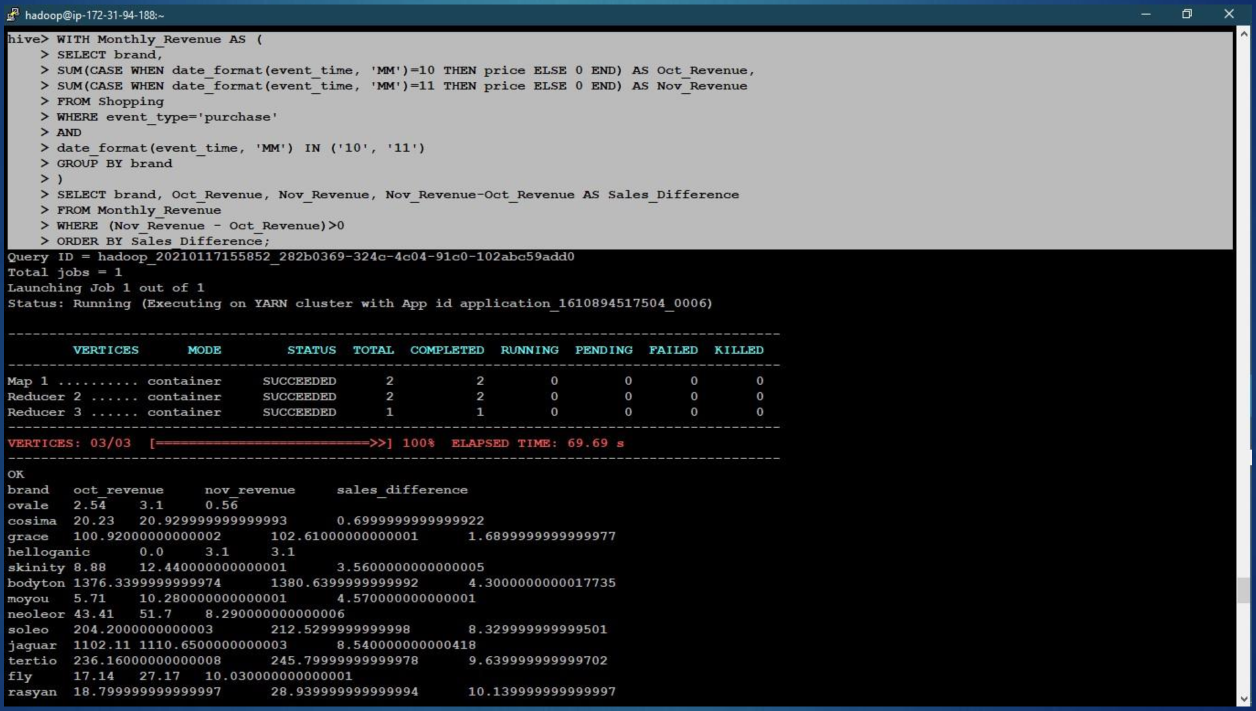
mavala	409.03999999999985	446.32	37.28000000000014
likato	296.0599999999999	340.9699999999999	44.910000000000025
ladykin	125.64999999999999	170.57	44.92
foamie	35.04	80.49	45.449999999999996
elskin	251.09000000000057	307.65000000000055	56.559999999999974
balbcare	155.32999999999996	212.38000000000025	57.050000000000296
koelcia	55.5	112.75000000000003	57.25000000000003
profhenna	679.2299999999999	736.8500000000005	57.62000000000057
kares	0.0	59.45	59.45
marutaka-foot	49.21999999999999	109.33	60.11000000000001
dewal	0.0	61.29	61.29
inm	288.02	351.2100000000001	63.19000000000011
laboratorium	246.49999999999991	312.52	66.02000000000007
cutrin	299.36999999999995	367.62	68.25000000000006
egomania	77.47	146.04000000000002	68.57000000000002
konad	739.8299999999991	810.6700000000003	70.84000000000117
nirvel	163.03999999999996	234.32999999999984	71.28999999999988
koelf	422.72999999999985	507.2900000000002	84.56000000000034
plazan	101.37	194.01000000000002	92.64000000000001
aura	83.95	177.51	93.55999999999999
kerasys	430.90999999999985	525.2000000000002	94.2900000000003
enjoy	41.349999999999994	136.57000000000002	95.22000000000003
depilflax	2707.069999999994	2803.7799999999975	96.71000000000367
eos	54.339999999999996	152.61	98.27000000000001
carmex	145.08	243.36	98.28
batiste	772.3999999999999	874.1699999999994	101.76999999999953
osmo	645.58	762.3100000000002	116.73000000000013
dizao	819.1300000000012	945.5099999999998	126.37999999999852
igrobeauty	513.6600000000009	645.0699999999999	131.40999999999906

finish	98.38	230.38000000000008	132.00000000000009
nefertiti	233.52000000000007	366.64	133.11999999999992
elizavecca	70.53	204.3	133.77
miskin	158.04	293.07000000000005	135.03000000000006
latinoil	249.52	384.59	135.06999999999996
farmona	1692.4599999999996	1843.4300000000007	150.97000000000116
cristalinas	427.6299999999999	584.949999999999	157.31999999999914
chi	358.9400000000002	538.6100000000002	179.67000000000002
matreshka	0.0	182.67000000000002	182.67000000000002
freshbubble	318.7000000000001	502.34000000000015	183.64000000000004
mane	66.78999999999999	260.26	193.47
keen	236.35000000000005	435.62	199.26999999999995
ecocraft	41.160000000000004	241.95	200.79
fedua	52.38	263.81000000000006	211.43000000000006
provoc	827.9900000000009	1063.8200000000006	235.8299999999997
skinlite	651.9400000000002	890.4499999999979	238.50999999999772
entity	479.7100000000015	719.2599999999993	239.5499999999978
trind	298.07000000000005	542.9600000000002	244.8900000000001
protokeratin	201.25000000000003	456.79000000000013	255.5400000000001
beauugreen	511.5099999999999	768.35	256.84000000000015
bluesky	10307.239999999858	10565.529999999713	258.28999999985535
candy	534.9599999999999	799.3799999999993	264.4199999999994
insight	1443.7000000000012	1721.9600000000003	278.2599999999991
kocostar	310.8500000000001	594.9300000000003	284.0800000000002
happyfons	801.9200000000006	1091.5900000000001	289.6699999999995
kims	330.03999999999996	632.0400000000001	302.0000000000001
shary	871.9599999999994	1176.4899999999989	304.5299999999995
nitrile	847.279999999999	1162.679999999999	315.4
lowence	242.84	567.7499999999997	324.9099999999996

jas	3318.959999999995	3657.4300000000026	338.47000000000753
ellips	245.8499999999999	606.0399999999996	360.1899999999997
lador	2083.610000000004	2471.530000000007	387.9200000000028
naomi	0.0	389.0	389.0
kiss	421.54999999999944	817.3299999999994	395.7799999999999
yu-r	271.41	673.7099999999998	402.2999999999998
sophin	1067.8600000000001	1515.5200000000011	447.660000000001
farmavita	837.3699999999984	1291.9700000000003	454.60000000000184
bioaqua	942.8899999999996	1398.1199999999997	455.23
greymy	29.21	489.49	460.28000000000003
gehwol	1089.07	1557.6799999999982	468.6099999999983
matrix	3243.249999999999	3726.7400000000007	483.4900000000016
limoni	1308.9000000000003	1796.5999999999997	487.69999999999936
s.care	412.68	913.0699999999999	500.38999999999993
coifin	903.0000000000001	1428.4899999999998	525.4899999999997
uskusi	5142.270000000017	5690.310000000005	548.0399999999881
airnails	5118.899999999939	5691.519999999996	572.6200000000572
browxenna	14331.36999999995	14916.729999999976	585.360000000026
kinetics	6334.2499999999945	6945.260000000017	611.010000000022
kosmekka	1181.4400000000003	1813.37	631.9299999999996
kaaral	4412.4299999999985	5086.069999999992	673.639999999994
refectocil	2716.180000000005	3475.580000000007	759.4000000000024
rosi	3077.0399999999927	3841.560000000013	764.5200000000204
solomeya	1899.699999999992	2685.799999999991	786.099999999999
missha	1293.8299999999995	2150.2799999999984	856.4499999999989
levissime	2227.5000000000064	3085.3099999999977	857.8099999999913
art-visage	2092.71000000001	2997.800000000011	905.090000000001
ecolab	262.8500000000001	1214.2999999999988	951.4499999999987
nagaraku	4369.740000000054	5327.680000000063	957.9400000000087

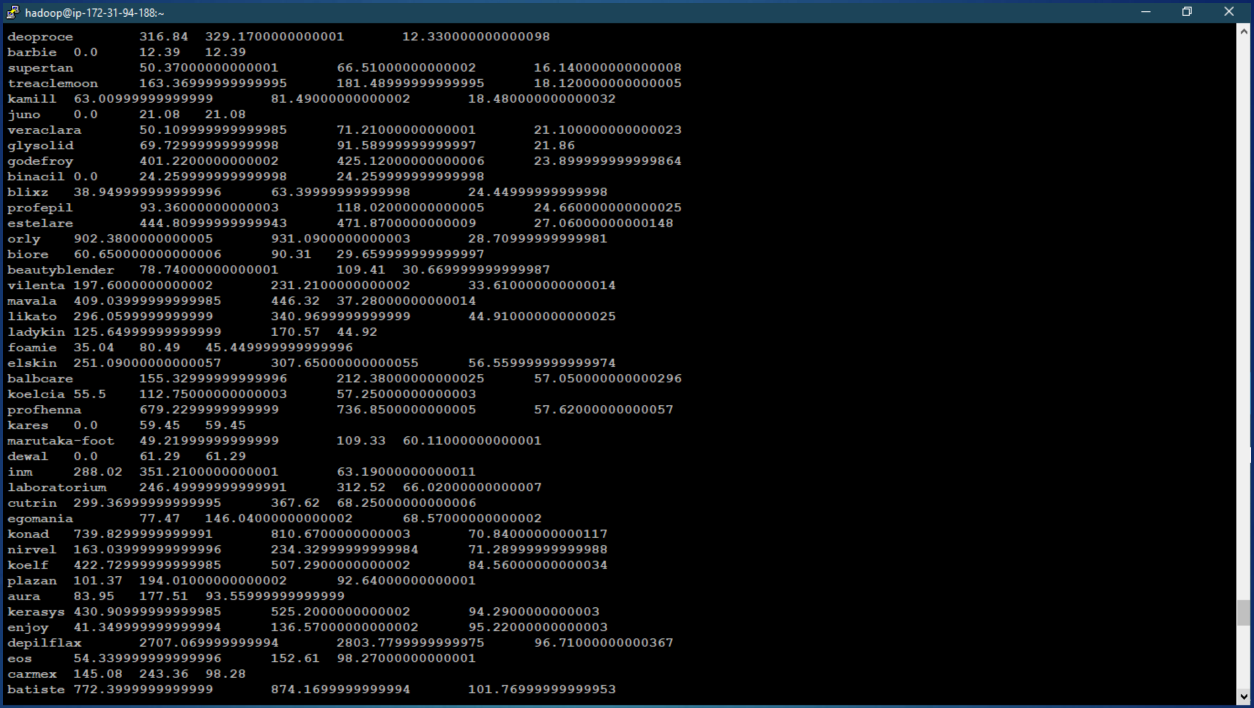
sanoto	157.14	1209.6799999999998	1052.54
markell	1768.7499999999989	2834.4300000000007	1065.6800000000019
metzger	5373.450000000006	6457.159999999988	1083.7099999999818
de.lux	1659.699999999967	2775.509999999968	1115.8100000000009
swarovski	1887.9299999999873	3043.160000000003	1155.2300000000157
beauty-free	554.1700000000006	1782.8600000000163	1228.6900000000155
zeitun	708.6600000000004	2009.63	1300.9699999999998
joico	705.52	2015.1000000000015	1309.5800000000015
severina	4775.88	6120.480000000023	1344.600000000023
irisk	45591.96000000588	46946.040000002184	1354.0799999963056
oniq	8425.41000000003	9841.650000000018	1416.239999999987
levrana	2243.560000000002	3664.099999999998	1420.5399999999959
roubloff	3491.360000000003	4913.769999999991	1422.4099999999885
smart	4457.260000000004	5902.140000000017	1444.8800000000128
shik	3341.2	4839.720000000007	1498.5200000000068
domix	10472.04999999994	12009.170000000022	1537.1200000000827
artex	2730.639999999998	4327.250000000017	1596.6100000000192
beautix	10493.949999999966	12222.949999999913	1728.9999999999472
milv	3904.9399999999964	5642.01000000008	1737.0700000000838
masura	31266.07999999821	33058.46999999708	1792.3899999988753
f.o.x	6624.229999999982	8577.280000000004	1953.050000000022
kapous	11927.159999999898	14093.080000000158	2165.92000000026
concept	11032.139999999925	13380.39999999993	2348.2600000000057
estel	21756.750000000342	24142.67000000022	2385.919999999878
kaypro	881.3399999999998	3268.699999999995	2387.359999999995
benovy	409.6200000000002	3259.970000000001	2850.350000000001
italwax	21940.239999999732	24799.369999999893	2859.130000000161
yoko	8756.909999999949	11707.879999999996	2950.9700000000466
haruyama	9390.689999999991	12352.91000000013	2962.2200000001394

```
marathon        7280.749999999997       10273.1         2992.350000000003
lovely  8704.379999999952       11939.060000000045      3234.680000000093
bpw.style       11572.150000001699      14837.440000000812      3265.289999999113
staleks 8519.730000000003       11875.61000000008       3355.8800000000774
freedecor       3421.779999999971       7671.800000000175       4250.020000000204
runail  71539.27999999933       76758.66000000098       5219.380000001649
polarus 6013.720000000003       11371.930000000018      5358.2100000000155
cosmoprofi      8322.81000000007        14536.99000000016       6214.180000000089
jessnail        26287.839999999916      33345.22999999992       7057.390000000007
strong  29196.62999999994       38671.269999999924      9474.639999999985
ingarden        23161.390000000138      33566.21000000009       10404.819999999949
lianail 5892.839999999975       16394.240000000245      10501.40000000027
uno     35302.02999999977       51039.749999998035      15737.719999998262
grattol 35445.5400000011        71472.71000000068       36027.169999999576
        474679.0599999623       619509.2399999934       144830.18000003108
Time taken: 70.259 seconds, Fetched: 161 row(s)
```

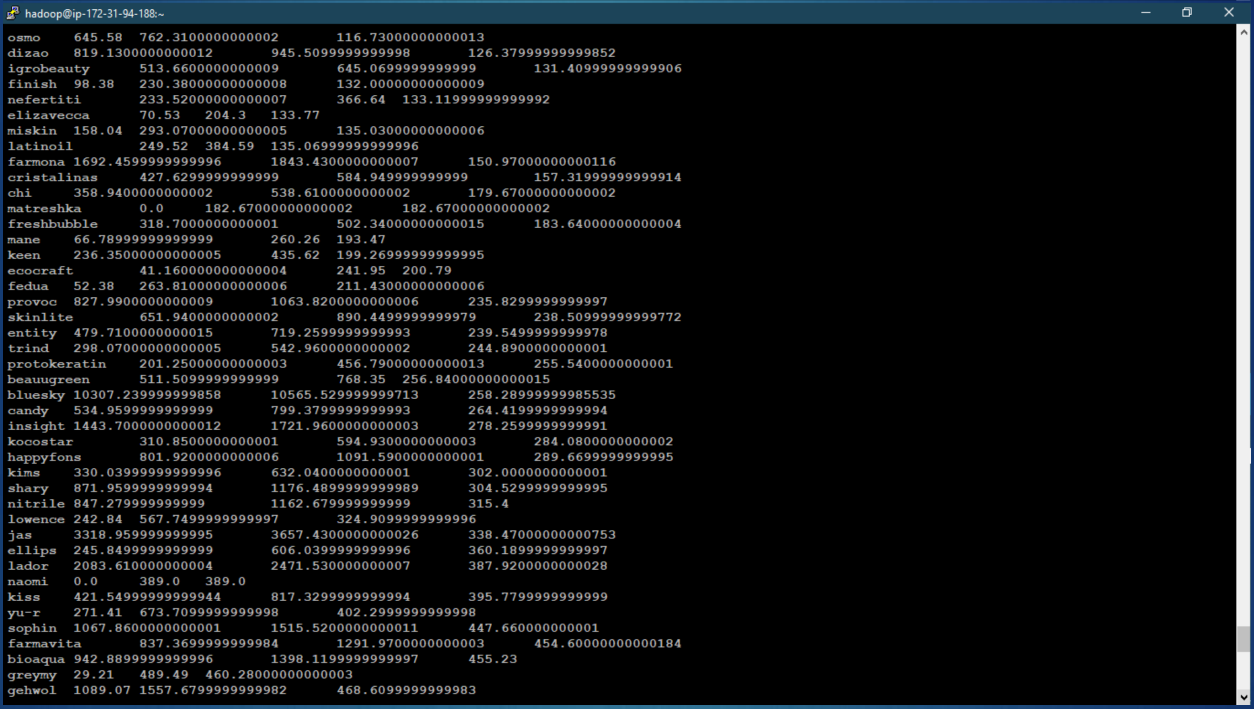


hadoop@ip-172-31-94-188:~

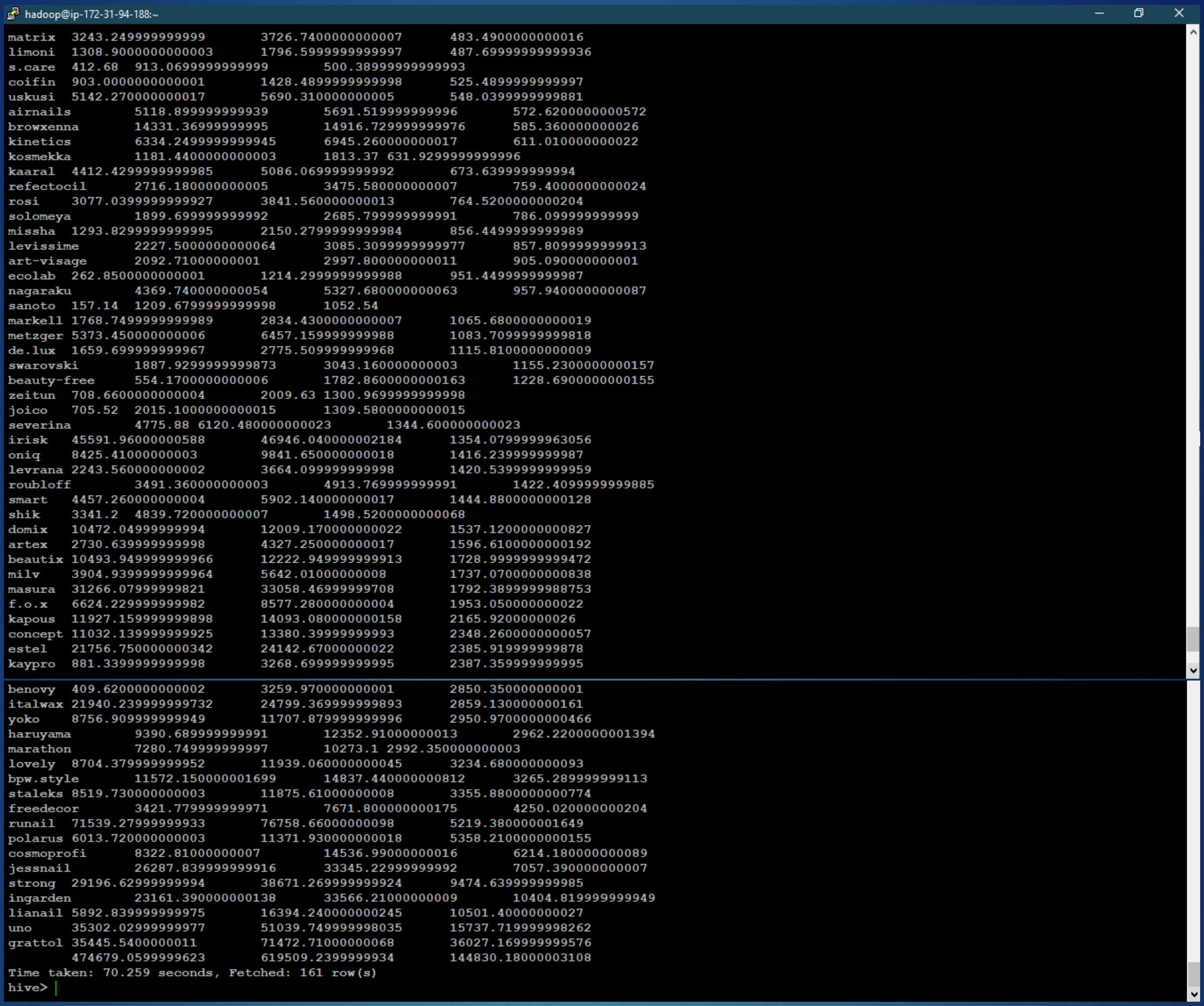
```
hive> WITH Monthly_Revenue AS (
    > SELECT brand,
    > SUM(CASE WHEN date_format(event_time, 'MM')=10 THEN price ELSE 0 END) AS Oct_Revenue,
    > SUM(CASE WHEN date_format(event_time, 'MM')=11 THEN price ELSE 0 END) AS Nov_Revenue
    > FROM Shopping
    > WHERE event_type='purchase'
    > AND
    > 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;
Query ID = hadoop_20210117155852_282b0369-324c-4c04-91c0-102abc59add0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610894517504_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 69.69 s
----------------------------------------------------------------------------------------------
OK
brand   oct_revenue     nov_revenue     sales_difference
ovale   2.54    3.1     0.56
cosima  20.23   20.929999999999993      0.6999999999999922
grace   100.92000000000002      102.61000000000001      1.6899999999999977
helloganic      0.0     3.1     3.1
skinity 8.88    12.440000000000001      3.5600000000000005
bodyton 1376.3399999999974      1380.6399999999992      4.3000000000017735
moyou   5.71    10.280000000000001      4.570000000000001
neoleor 43.41   51.7    8.290000000000006
soleo   204.2000000000003       212.5299999999998       8.329999999999501
jaguar  1102.11 1110.6500000000003      8.540000000000418
tertio  236.1600000000008       245.79999999999978      9.639999999999702
fly     17.14   27.17   10.030000000000001
rasyan  18.799999999999997      28.939999999999994      10.139999999999997
```



```
deoproce        316.84  329.1700000000001       12.330000000000098
barbie  0.0     12.39   12.39
supertan        50.37000000000001       66.51000000000002       16.140000000000008
treaclemoon     163.36999999999995      181.48999999999995      18.120000000000005
kamill  63.00999999999999       81.49000000000002       18.480000000000032
juno    0.0     21.08   21.08
veraclara       50.109999999999985      71.21000000000001       21.100000000000023
glysolid        69.72999999999998       91.58999999999997       21.86
godefroy        401.2200000000002       425.12000000000006      23.899999999999864
binacil 0.0     24.259999999999998      24.259999999999998
blixz   38.949999999999996      63.39999999999998       24.44999999999998
profepil        93.36000000000003       118.02000000000005      24.660000000000025
estelare        444.80999999999943      471.8700000000009       27.06000000000148
orly    902.3800000000005       931.0900000000003       28.70999999999981
biore   60.650000000000006      90.31   29.659999999999997
beautyblender   78.74000000000001       109.41  30.669999999999987
vilenta 197.6000000000002       231.2100000000002       33.61000000000014
mavala  409.03999999999985      446.32  37.28000000000014
likato  296.0599999999999       340.9699999999999       44.91000000000025
ladykin 125.64999999999999      170.57  44.92
foamie  35.04   80.49   45.449999999999996
elskin  251.09000000000057      307.65000000000055      56.559999999999974
balbcare        155.32999999999996      212.38000000000025      57.050000000000296
koelcia 55.5    112.75000000000003      57.25000000000003
profhenna       679.2299999999999       736.8500000000005       57.62000000000057
kares   0.0     59.45   59.45
marutaka-foot   49.21999999999999       109.33  60.11000000000001
dewal   0.0     61.29   61.29
inm     288.02  351.2100000000001       63.19000000000011
laboratorium    246.49999999999991      312.52  66.02000000000007
cutrin  299.36999999999995      367.62  68.25000000000006
egomania        77.47   146.04000000000002      68.57000000000002
konad   739.8299999999991       810.6700000000003       70.84000000000117
nirvel  163.03999999999996      234.32999999999984      71.28999999999988
koelf   422.72999999999985      507.2900000000002       84.56000000000034
plazan  101.37  194.01000000000002      92.64000000000001
aura    83.95   177.51  93.55999999999999
kerasys 430.90999999999985      525.2000000000002       94.2900000000003
enjoy   41.349999999999994      136.57000000000002      95.22000000000003
depilflax       2707.069999999994       2803.7799999999975      96.71000000000367
eos     54.339999999999996      152.61  98.27000000000001
carmex  145.08  243.36  98.28
batiste 772.3999999999999       874.1699999999994       101.76999999999953
```



```
osmo    645.58  762.3100000000002       116.73000000000013
dizao   819.1300000000012       945.5099999999998       126.37999999999852
igrobeauty      513.6600000000009       645.0699999999999       131.40999999999906
finish  98.38   230.38000000000008      132.00000000000009
nefertiti       233.52000000000007      366.64  133.11999999999992
elizavecca      70.53   204.3   133.77
miskin  158.04  293.07000000000005      135.03000000000006
latinoil        249.52  384.59  135.06999999999996
farmona 1692.4599999999996      1843.4300000000007      150.97000000000116
cristalinas     427.6299999999999       584.949999999999        157.31999999999914
chi     358.9400000000002       538.6100000000002       179.67000000000002
matreshka       0.0     182.67000000000002      182.67000000000002
freshbubble     318.7000000000001       502.34000000000015      183.64000000000004
mane    66.78999999999999       260.26  193.47
keen    236.35000000000005      435.62  199.26999999999995
ecocraft        41.160000000000004      241.95  200.79
fedua   52.38   263.81000000000006      211.43000000000006
provoc  827.9900000000009       1063.8200000000006      235.8299999999997
skinlite        651.9400000000002       890.4499999999979       238.50999999999772
entity  479.7100000000015       719.2599999999993       239.5499999999978
trind   298.07000000000005      542.9600000000002       244.8900000000001
protokeratin    201.25000000000003      456.79000000000013      255.540000000001
beauugreen      511.5099999999999       768.35  256.84000000000015
bluesky 10307.239999999858      10565.529999999713      258.28999999985535
candy   534.9599999999999       799.3799999999993       264.4199999999994
insight 1443.7000000000012      1721.9600000000003      278.2599999999991
kocostar        310.8500000000001       594.9300000000003       284.0800000000002
happyfons       801.9200000000006       1091.5900000000001      289.6699999999995
kims    330.03999999999996      632.0400000000001       302.0000000000001
shary   871.9599999999994       1176.4899999999989      304.5299999999995
nitrile 847.279999999999        1162.679999999999       315.4
lowence 242.84  567.7499999999997       324.9099999999996
jas     3318.959999999995       3657.4300000000026      338.47000000000753
ellips  245.8499999999999       606.0399999999996       360.1899999999997
lador   2083.610000000004       2471.530000000007       387.920000000028
naomi   0.0     389.0   389.0
kiss    421.54999999999944      817.3299999999994       395.7799999999999
yu-r    271.41  673.7099999999998       402.2999999999998
sophin  1067.8600000000001      1515.5200000000011      447.660000000001
farmavita       837.3699999999984       1291.9700000000003      454.6000000000184
bioaqua 942.8899999999996       1398.1199999999997      455.23
greymy  29.21   489.49  460.28000000000003
gehwol  1089.07 1557.6799999999982      468.6099999999983
```



```
matrix  3243.249999999999       3726.7400000000007      483.4900000000016
limoni  1308.9000000000003      1796.5999999999997      487.69999999999936
s.care  412.68  913.0699999999999       500.38999999999993
coifin  903.0000000000001       1428.4899999999998      525.4899999999997
uskusi  5142.270000000017       5690.310000000005       548.0399999999881
airnails        5118.899999999939       5691.519999999996       572.6200000000572
browxenna       14331.36999999995       14916.729999999976      585.360000000026
kinetics        6334.2499999999945      6945.260000000017       611.010000000022
kosmekka        1181.4400000000003      1813.37 631.9299999999996
kaaral  4412.4299999999985      5086.069999999992       673.639999999994
refectocil      2716.180000000005       3475.580000000007       759.4000000000024
rosi    3077.0399999999927      3841.560000000013       764.5200000000204
solomeya        1899.699999999992       2685.799999999991       786.099999999999
missha  1293.8299999999995      2150.2799999999984      856.4499999999989
levissime       2227.5000000000064      3085.3099999999977      857.8099999999913
art-visage      2092.71000000001        2997.800000000011       905.090000000001
ecolab  262.8500000000001       1214.2999999999988      951.4499999999987
nagaraku        4369.740000000054       5327.680000000063       957.9400000000087
sanoto  157.14  1209.6799999999998      1052.54
markell 1768.7499999999989      2834.4300000000007      1065.6800000000019
metzger 5373.450000000006       6457.159999999988       1083.7099999999818
de.lux  1659.699999999967       2775.509999999968       1115.8100000000009
swarovski       1887.9299999999873      3043.160000000003       1155.2300000000157
beauty-free     554.1700000000006       1782.8600000000163      1228.6900000000155
zeitun  708.6600000000004       2009.63 1300.9699999999998
joico   705.52  2015.1000000000015      1309.5800000000015
severina        4775.88 6120.480000000023       1344.600000000023
irisk   45591.96000000588       46946.040000002184      1354.0799999963056
oniq    8425.41000000003        9841.650000000018       1416.239999999987
levrana 2243.560000000002       3664.099999999998       1420.5399999999959
roubloff        3491.360000000003       4913.769999999991       1422.4099999999885
smart   4457.260000000004       5902.140000000017       1444.880000000128
shik    3341.2  4839.720000000007       1498.5200000000068
domix   10472.04999999994       12009.170000000022      1537.1200000000827
artex   2730.639999999998       4327.250000000017       1596.6100000000192
beautix 10493.949999999966      12222.949999999913      1728.9999999999472
milv    3904.9399999999964      5642.01000000008        1737.0700000000838
masura  31266.07999999821       33058.46999999708       1792.389999988753
f.o.x   6624.229999999982       8577.280000000004       1953.050000000022
kapous  11927.159999999898      14093.080000000158      2165.92000000026
concept 11032.139999999925      13380.39999999993       2348.2600000000057
estel   21756.750000000342      24142.67000000022       2385.919999999878
kaypro  881.3399999999998       3268.699999999995       2387.359999999995
```

```
benovy  409.6200000000002       3259.970000000001       2850.350000000001
italwax 21940.239999999732      24799.369999999893      2859.130000000161
yoko    8756.909999999949       11707.879999999996      2950.9700000000466
haruyama        9390.689999999991       12352.91000000013       2962.2200000001394
marathon        7280.749999999997       10273.1 2992.350000000003
lovely  8704.379999999952       11939.060000000045      3234.680000000093
bpw.style       11572.150000001699      14837.440000000812      3265.289999999113
staleks 8519.730000000003       11875.61000000008       3355.8800000000774
freedecor       3421.779999999971       7671.800000000175       4250.020000000204
runail  71539.27999999933       76758.66000000098       5219.380000001649
polarus 6013.720000000003       11371.930000000018      5358.2100000000155
cosmoprofi      8322.81000000007        14536.99000000016       6214.180000000089
jessnail        26287.839999999916      33345.22999999992       7057.390000000007
strong  29196.62999999994       38671.269999999924      9474.639999999985
ingarden        23161.390000000138      33566.21000000009       10404.819999999949
lianail 5892.839999999975       16394.240000000245      10501.40000000027
uno     35302.02999999977       51039.749999998035      15737.719999998262
grattol 35445.5400000011        71472.71000000068       36027.169999999576
        474679.0599999623       619509.2399999934       144830.18000003108
Time taken: 70.259 seconds, Fetched: 161 row(s)
hive>
```

Insights:

- Here are some **161 brands with increment** in the selling from October to November.
- **'Grattol' brand has the highest total increment i.e., 36,027 /-** and **'Ovale' seems to have least increment of 0.56 /-** from October to November.
- Among all these brands list, **'Runail'** which was the best brand in terms of selling in October and November combined **is also in the top 10 brands with high increment** for **October (71539.28 /-)** to **November (76758.61 /-)** i.e., **increment of total 5219.38 /-**.
- This implies that **'Runail' is the best and popular brand among all other brands within people.**

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.

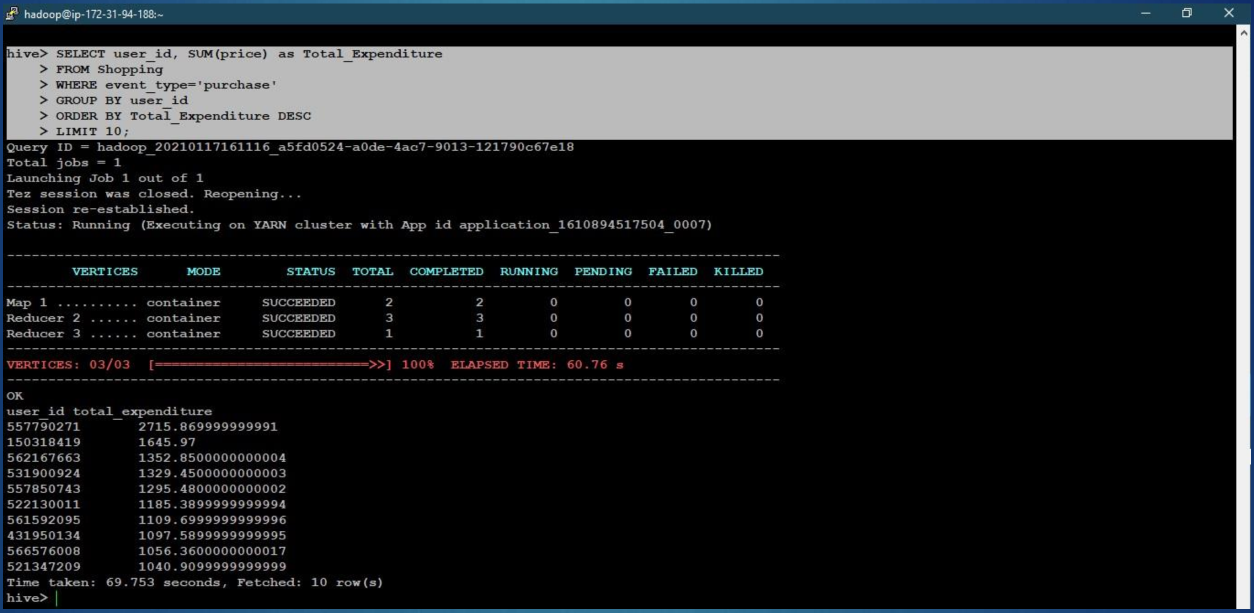
Query:

```
SELECT user_id, SUM(price) as Total_Expenditure

FROM Shopping

WHERE event_type='purchase'

GROUP BY user_id

ORDER BY Total_Expenditure DESC

LIMIT 10;
```

Output:

```
Query ID = hadoop_20210220161116_a5fd0524-a0de-4ac7-9013-121790c67e18

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_1610894517504_0007)

----------------------------------------------------------------------------------------------

        VERTICES      MODE        STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED

----------------------------------------------------------------------------------------------

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

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

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

----------------------------------------------------------------------------------------------

VERTICES: 03/03 [==========================>>] 100% ELAPSED TIME: 60.76 s

----------------------------------------------------------------------------------------------
```

```
OK

user_id total_expenditure

557790271       2715.869999999991

150318419       1645.97

562167663       1352.8500000000004

531900924       1329.4500000000003

557850743       1295.4800000000002

522130011       1185.3899999999994

561592095       1109.6999999999996

431950134       1097.5899999999995

566576008       1056.3600000000017

521347209       1040.9099999999999

Time taken: 69.753 seconds, Fetched: 10 row(s)
```



```
hive> SELECT user_id, SUM(price) as Total_Expenditure
    > FROM Shopping
    > WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expenditure DESC
    > LIMIT 10;
Query ID = hadoop_20210117161116_a5fd0524-a0de-4ac7-9013-121790c67e18
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_1610894517504_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      3          3        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 60.76 s
----------------------------------------------------------------------------------------------
OK
user_id total_expenditure
557790271       2715.869999999991
150318419       1645.97
562167663       1352.8500000000004
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3899999999994
561592095       1109.6999999999996
431950134       1097.5899999999995
566576008       1056.3600000000017
521347209       1040.9099999999999
Time taken: 69.753 seconds, Fetched: 10 row(s)
hive>
```

Insights:

- Here is the list of the top 10 users or buyers who have spend the most and could be rewarded with a Golden Customer plan to attract more people in the coming future.
- We are **selecting this query to be executed using Optimized table** to check that does optimized table reduces execution time with proper partitioning and bucketing.
- **Time taken to execute this query on Base table (non-optimized table) is 69.753 seconds.**

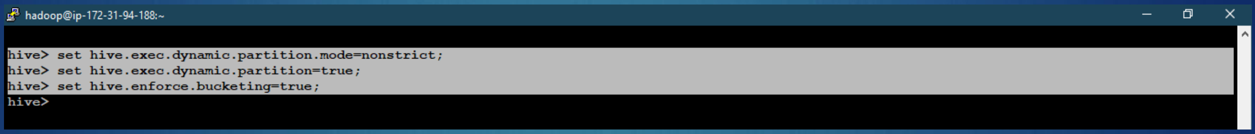
Optimized Table

To create table with Partitioning and Bucketing below commands need to be executed one by one separately.

- set hive.exec.dynamic.partition.mode=nonstrict;

- set hive.exec.dynamic.partition=true;

- set hive.enforce.bucketing=true;



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

Table optimization steps:-

1. Command to create table 'Dyn_Part_Buck_Shopping' with partition on 'event_type' attribute and bucket(cluster) on 'price' attribute.

Query:

CREATE TABLE IF NOT EXISTS Dyn_Part_Buck_Shopping(

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 (price) INTO 7 BUCKETS

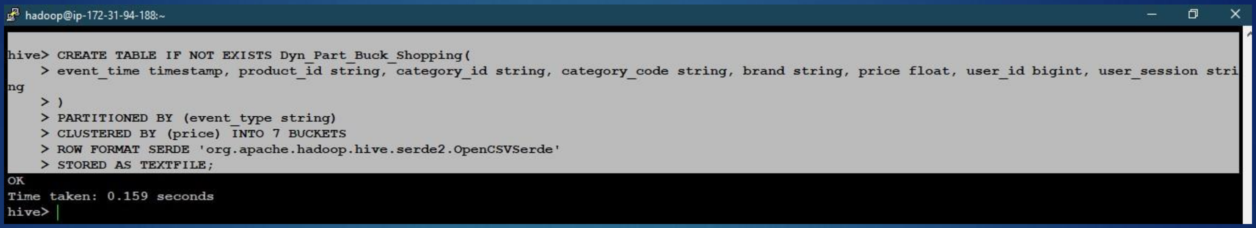
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

STORED AS TEXTFILE;

Output:

OK

Time taken: 0.159 seconds



```
hive> CREATE TABLE IF NOT EXISTS Dyn_Part_Buck_Shopping(
    > event_time timestamp, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session stri
ng
    > )
    > PARTITIONED BY (event_type string)
    > CLUSTERED BY (price) INTO 7 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE;
OK
Time taken: 0.159 seconds
hive>
```

2. To add data into partitioned and bucketed table we need to get it from already created table i.e., 'Shopping'

Query:

```
INSERT INTO TABLE Dyn_Part_Buck_Shopping

PARTITION (event_type)

SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session,
event_type

FROM Shopping;
```

Output:

```
Query ID = hadoop_20210220162425_57023bb0-e16e-4665-8c81-ab7f87859fd7

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1610894517504_0011)


----------------------------------------------------------------------------------------------

        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: 02/02 [==========================>>] 100% ELAPSED TIME: 163.41 s

----------------------------------------------------------------------------------------------
```

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

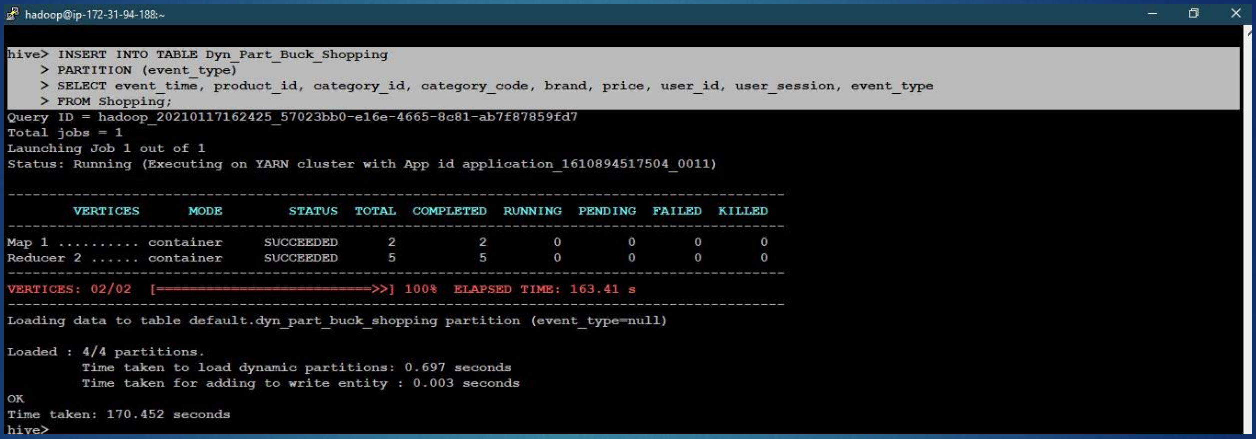
Loaded : 4/4 partitions.

Time taken to load dynamic partitions: 0.697 seconds

Time taken for adding to write entity : 0.003 seconds

OK

Time taken: 170.452 seconds

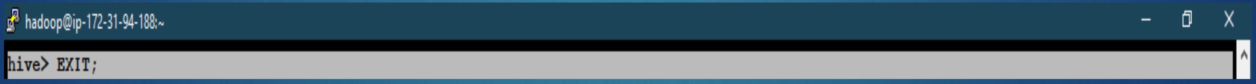


```
hive> INSERT INTO TABLE Dyn_Part_Buck_Shopping
    > PARTITION (event_type)
    > SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
    > FROM Shopping;
Query ID = hadoop_20210117162425_57023bb0-e16e-4665-8c81-ab7f87859fd7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610894517504_0011)

----------------------------------------------------------------------------------------------
        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: 02/02  [==========================>>] 100%  ELAPSED TIME: 163.41 s
----------------------------------------------------------------------------------------------
Loading data to table default.dyn_part_buck_shopping partition (event_type=null)

Loaded : 4/4 partitions.
         Time taken to load dynamic partitions: 0.697 seconds
         Time taken for adding to write entity : 0.003 seconds
OK
Time taken: 170.452 seconds
hive>
```

3.Command to check the successful creation of partitioned and bucketed table first we need to exit from Hive environment by executing 'EXIT;' command and then run below mentioned commands

1. Command to exit Hive environment

- EXIT;



```
hive> EXIT;
```

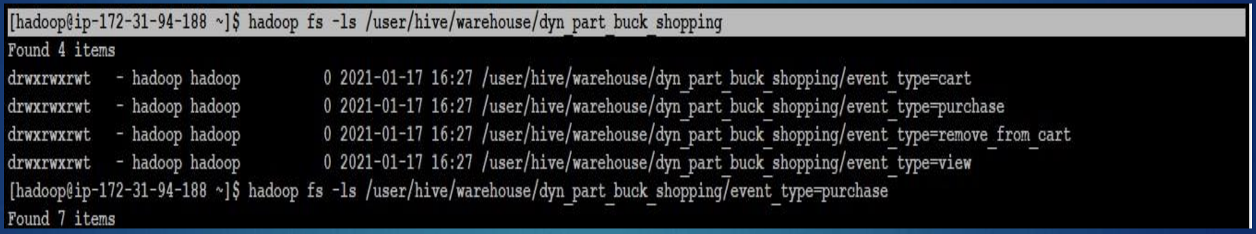
3.2. Command to check successful existence of Partitioned and Bucketed table ‘Dyn_Part_Buck_Shopping’ in hive warehouse.

- hadoop fs -ls /user/hive/warehouse/Dyn_Part_Buck_Shopping

Output:

Fount 4 items

```
drwxrwxrwt   - hadoop hadoop          0 2021-02-28 16:27
/user/hive/warehouse/dyn_part_buck_shopping/event_type=cart

drwxrwxrwt   - hadoop hadoop          0 2021-02-28 16:27
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase

drwxrwxrwt   - hadoop hadoop          0 2021-02-28 16:27
/user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart

drwxrwxrwt   - hadoop hadoop          0 2021-02-28 16:27
/user/hive/warehouse/dyn_part_buck_shopping/event_type=view
```



```
[hadoop@ip-172-31-94-188 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping
Found 4 items
drwxrwxrwt   - hadoop hadoop          0 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart
drwxrwxrwt   - hadoop hadoop          0 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase
drwxrwxrwt   - hadoop hadoop          0 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart
drwxrwxrwt   - hadoop hadoop          0 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view
[hadoop@ip-172-31-94-188 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase
Found 7 items
```

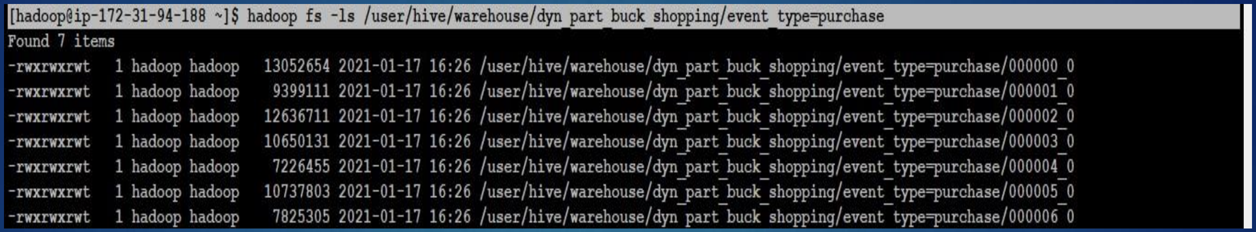
3.3. Command to check existence of partitions (event_type = purchase) in the table

```
hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase
```

Output:

```
Found 7 items

-rwxrwxrwt 1 hadoop hadoop 13052654 2021-02-28 16:26
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000000_0

-rwxrwxrwt 1 hadoop hadoop 9399111 2021-02-28 16:26
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000001_0

-rwxrwxrwt 1 hadoop hadoop 12636711 2021-02-28 16:26
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000002_0

-rwxrwxrwt 1 hadoop hadoop 10650131 2021-02-28 16:26
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000003_0

-rwxrwxrwt 1 hadoop hadoop 7226455 2021-02-28 16:26
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000004_0

-rwxrwxrwt 1 hadoop hadoop 10737803 2021-02-28 16:26
/user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000005_0
```

-rwxrwxrwt 1 hadoop hadoop 7825305 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000006_0



```
[hadoop@ip-172-31-94-188 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   13052654 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000000_0
-rwxrwxrwt   1 hadoop hadoop    9399111 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000001_0
-rwxrwxrwt   1 hadoop hadoop   12636711 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000002_0
-rwxrwxrwt   1 hadoop hadoop   10650131 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000003_0
-rwxrwxrwt   1 hadoop hadoop    7226455 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000004_0
-rwxrwxrwt   1 hadoop hadoop   10737803 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000005_0
-rwxrwxrwt   1 hadoop hadoop    7825305 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=purchase/000006_0
```

3.4. Command to check existence of partitions (event_type = cart) in the table

hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart

Output:

Found 7 items

-rwxrwxrwt 1 hadoop hadoop 57724286 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000000_0

-rwxrwxrwt 1 hadoop hadoop 43094161 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000001_0

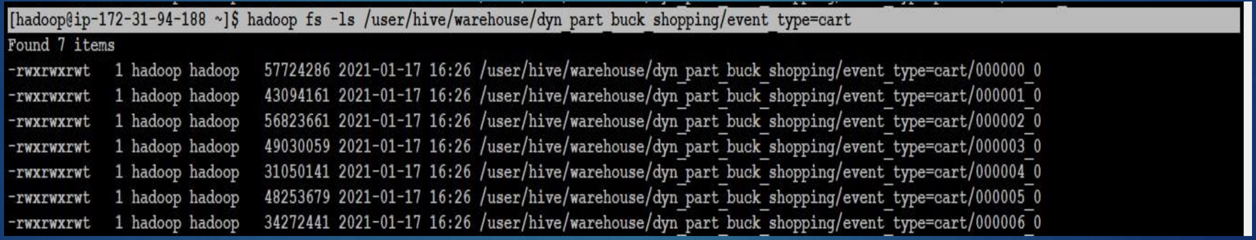
-rwxrwxrwt 1 hadoop hadoop 56823661 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000002_0

-rwxrwxrwt 1 hadoop hadoop 49030059 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000003_0

-rwxrwxrwt 1 hadoop hadoop 31050141 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000004_0

-rwxrwxrwt 1 hadoop hadoop 48253679 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000005_0

-rwxrwxrwt 1 hadoop hadoop 34272441 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000006_0



```
[hadoop@ip-172-31-94-188 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   57724286 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   43094161 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000001_0
-rwxrwxrwt   1 hadoop hadoop   56823661 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000002_0
-rwxrwxrwt   1 hadoop hadoop   49030059 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000003_0
-rwxrwxrwt   1 hadoop hadoop   31050141 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000004_0
-rwxrwxrwt   1 hadoop hadoop   48253679 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000005_0
-rwxrwxrwt   1 hadoop hadoop   34272441 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=cart/000006_0
```

3.5. Command to check existence of partitions (event_type = remove_from_cart) in the table

hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart

Output:

Found 7 items

-rwxrwxrwt 1 hadoop hadoop 39017824 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000000_0

-rwxrwxrwt 1 hadoop hadoop 29421828 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000001_0

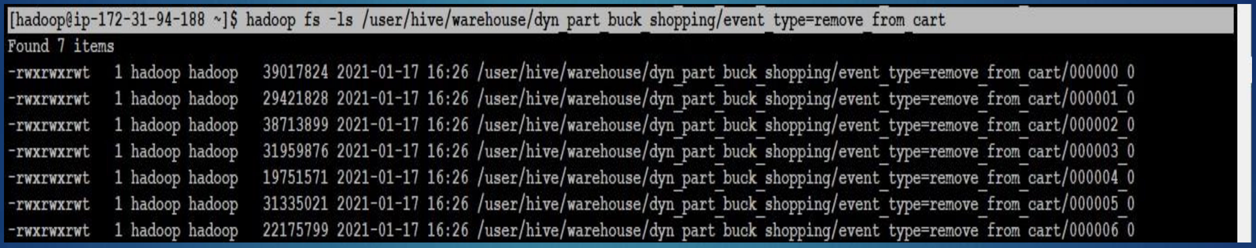
-rwxrwxrwt 1 hadoop hadoop 38713899 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000002_0

-rwxrwxrwt 1 hadoop hadoop 31959876 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000003_0

-rwxrwxrwt 1 hadoop hadoop 19751571 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000004_0

-rwxrwxrwt 1 hadoop hadoop 31335021 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000005_0

-rwxrwxrwt 1 hadoop hadoop 22175799 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000006_0



```
[hadoop@ip-172-31-94-188 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   39017824 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000000_0
-rwxrwxrwt   1 hadoop hadoop   29421828 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000001_0
-rwxrwxrwt   1 hadoop hadoop   38713899 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000002_0
-rwxrwxrwt   1 hadoop hadoop   31959876 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000003_0
-rwxrwxrwt   1 hadoop hadoop   19751571 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000004_0
-rwxrwxrwt   1 hadoop hadoop   31335021 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000005_0
-rwxrwxrwt   1 hadoop hadoop   22175799 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=remove_from_cart/000006_0
```

3.6. Command to check existence of partitions (event_type = view) in the table

hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=view

Output:

Found 7 items

-rwxrwxrwt 1 hadoop hadoop 88831872 2021-02-28 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000000_0

-rwxrwxrwt 1 hadoop hadoop 73953212 2021-02-28 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000001_0

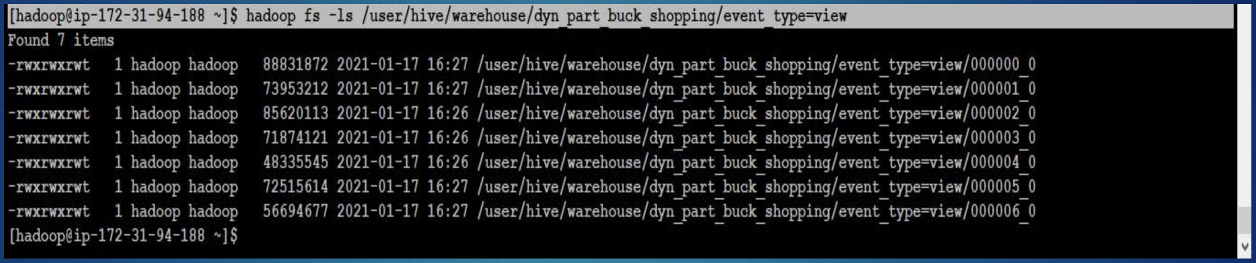
-rwxrwxrwt 1 hadoop hadoop 85620113 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000002_0

-rwxrwxrwt 1 hadoop hadoop 71874121 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000003_0

-rwxrwxrwt 1 hadoop hadoop 48335545 2021-02-28 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000004_0

-rwxrwxrwt 1 hadoop hadoop 72515614 2021-02-28 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000005_0

-rwxrwxrwt 1 hadoop hadoop 56694677 2021-02-28 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000006_0



```
[hadoop@ip-172-31-94-188 ~]$ hadoop fs -ls /user/hive/warehouse/dyn_part_buck_shopping/event_type=view
Found 7 items
-rwxrwxrwt   1 hadoop hadoop   88831872 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000000_0
-rwxrwxrwt   1 hadoop hadoop   73953212 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000001_0
-rwxrwxrwt   1 hadoop hadoop   85620113 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000002_0
-rwxrwxrwt   1 hadoop hadoop   71874121 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000003_0
-rwxrwxrwt   1 hadoop hadoop   48335545 2021-01-17 16:26 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000004_0
-rwxrwxrwt   1 hadoop hadoop   72515614 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000005_0
-rwxrwxrwt   1 hadoop hadoop   56694677 2021-01-17 16:27 /user/hive/warehouse/dyn_part_buck_shopping/event_type=view/000006_0
[hadoop@ip-172-31-94-188 ~]$
```

4. Now we need to re-enter the Hive environment to execute Query No 8 which we have selected to run on Optimized table.

- hive

5. Running the same query for Question 8 on Optimized as executed on Base table to understand the execution time of same query on Base table and Optimized table.

(Optimized) 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.

Query:

SELECT user_id, SUM(price) AS Total_Expenditure

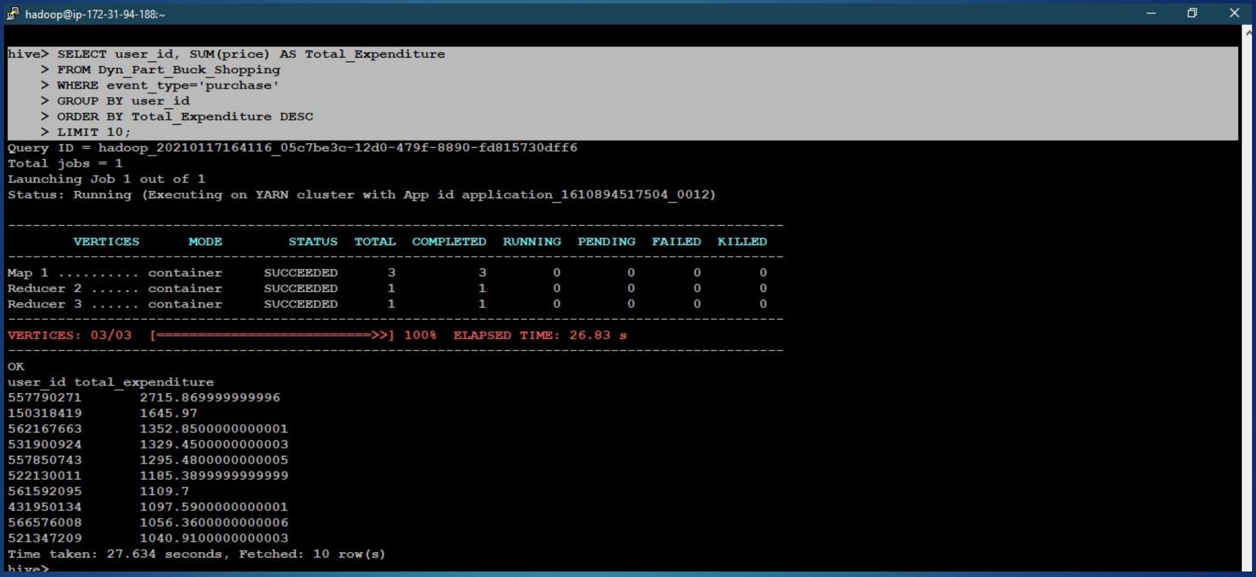
FROM Dyn_Part_Buck_Shopping

```
WHERE event_type='purchase'
GROUP BY user_id
ORDER BY Total_Expenditure DESC
LIMIT 10;
```

Output:

```
Query ID = hadoop_20210220164116_05c7be3c-12d0-479f-8890-fd815730dff6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610894517504_0012)


----------------------------------------------------------------------
     VERTICES     MODE       STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----------------------------------------------------------------------
Map 1 .......... container    SUCCEEDED     3        3      0      0     0     0
Reducer 2 ...... container    SUCCEEDED     1        1      0      0     0     0
Reducer 3 ...... container    SUCCEEDED     1        1      0      0     0     0
----------------------------------------------------------------------
VERTICES: 03/03 [==========================>>] 100% ELAPSED TIME: 26.83 s
----------------------------------------------------------------------
OK
user_id         total_expenditure
557790271       2715.869999999996
150318419       1645.97
562167663       1352.8500000000001
531900924       1329.4500000000003
557850743       1295.4800000000005
522130011       1185.3899999999999
561592095       1109.7
```

```
431950134       1097.5900000000001

566576008       1056.3600000000006

521347209       1040.9100000000003

Time taken: 27.634 seconds, Fetched: 10 row(s)
```



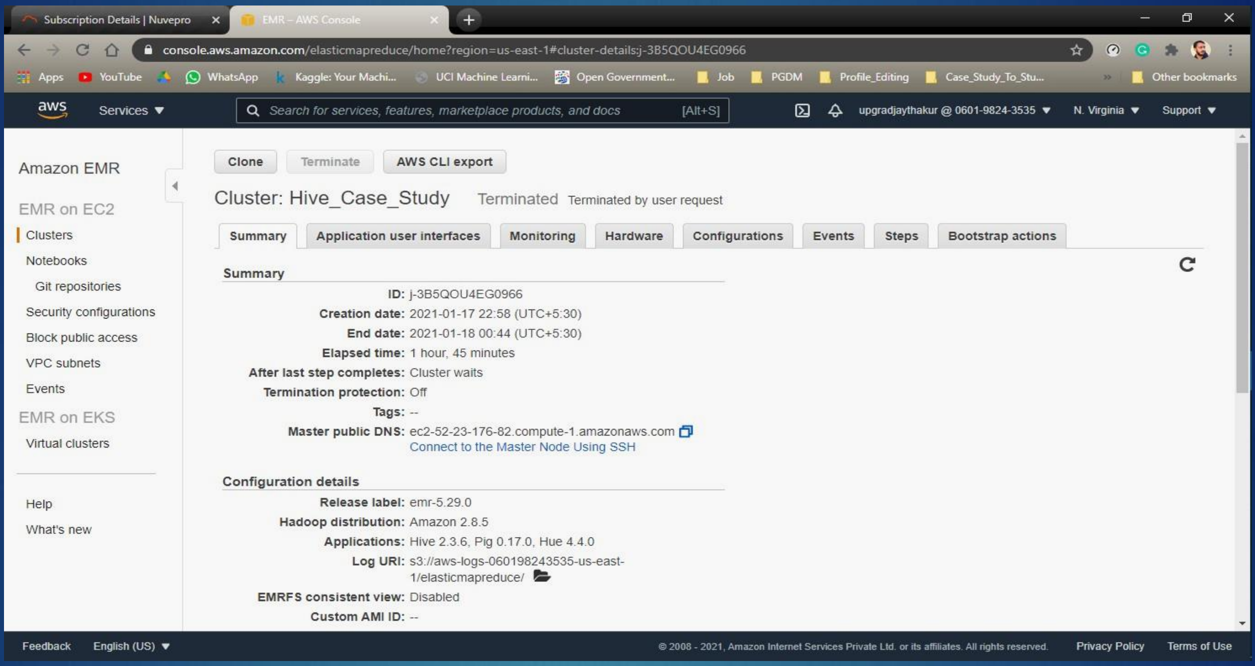
hadoop@ip-172-31-94-188:~

```
hive> SELECT user_id, SUM(price) AS Total_Expenditure
    > FROM Dyn_Part_Buck_Shopping
    > WHERE event_type='purchase'
    > GROUP BY user_id
    > ORDER BY Total_Expenditure DESC
    > LIMIT 10;
Query ID = hadoop_20210117164116_05c7be3c-12d0-479f-8890-fd815730dff6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1610894517504_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 26.83 s
----------------------------------------------------------------------------------------------
OK
user_id total_expenditure
557790271       2715.869999999996
150318419       1645.97
562167663       1352.8500000000001
531900924       1329.4500000000003
557850743       1295.4800000000005
522130011       1185.3899999999999
561592095       1109.7
431950134       1097.5900000000001
566576008       1056.3600000000006
521347209       1040.9100000000003
Time taken: 27.634 seconds, Fetched: 10 row(s)
hive>
```

Insights:

- After creating an optimized table by **Partitioning on 'event_type'** attribute and **Bucketing (Clustering) on 'price'** we have executed same query of Question No. 8 on this table.
- We can the result is same as we have got when executed on Base table (Non-Optimized table).
- Secondly, most importantly we can see there is significant drop in the execution time of the same query i.e., **previously the execution was measured as 69.753 seconds** and **now it is 27.634 seconds** with the **difference of 42.119 seconds**.
- **Hence, with proper partitioning and bucketing on table we can reduce execution time of the query.**

Terminating EMR Cluster (Hive_Case_Study)



Cluster: Hive_Case_Study Terminated Terminated by user request

Summary

ID: j-3B5QOU4EG0966
Creation date: 2021-01-17 22:58 (UTC+5:30)
End date: 2021-01-18 00:44 (UTC+5:30)
Elapsed time: 1 hour, 45 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-52-23-176-82.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
Log URI: s3://aws-logs-060198243535-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

THANK YOU