HIVE CASE STUDY

Submitted by Payal Joshi & Chetan Tippa

1. Creating EMR cluster

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

Amazon EMR

- EMR Studio
- EMR Serverless New

EMR on EC2

- Clusters
- Notebooks
  - Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events

EMR on EKS

- Virtual clusters

- Help
- What's new

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Create cluster | View details | Clone | Terminate

Filter: All clusters | Filter clusters ... | 15 clusters (all loaded)

Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalized instance hours
d300922	j-3OGPXIRJ574WL	Terminated User request	2022-09-30 21:43 (UTC+5:30)	1 hour, 8 minutes	12
d300922	j-3JHRWJXBAFUGZ	Terminated User request	2022-09-30 14:44 (UTC+5:30)	27 minutes	12
d300922	j-1N0QA4S59P62V	Terminated User request	2022-09-30 08:57 (UTC+5:30)	1 hour, 26 minutes	24
My cluster	j-3RB2QQ60E36TC	Terminated User request	2022-09-29 20:05 (UTC+5:30)	2 hours, 3 minutes	24
Hive_ct	j-P5ZTYOGURDOI	Terminated User request	2022-09-29 19:45 (UTC+5:30)	18 minutes	12
Hive_ct	j-VCFT2OWZ3AW8	Terminated User request	2022-09-29 19:19 (UTC+5:30)	25 minutes	12
Hive_ct	j-1SSDNCP9MNDII	Terminated User request	2022-09-28 22:07 (UTC+5:30)	1 hour, 11 minutes	12
d280922	j-27BR7NEV7QXFC	Terminated User request	2022-09-28 22:03 (UTC+5:30)	2 minutes	0
Hive_ct	j-1AA3S9XUI1BAF	Terminated User request	2022-09-28 18:24 (UTC+5:30)	2 hours, 7 minutes	24
220922	j-2HGRVZVYOQ594	Terminated User request	2022-09-22 14:22 (UTC+5:30)	1 hour, 31 minutes	16

1. Creating EMR cluster – Step 1 : Software

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Create Cluster - Advanced Options

Go to quick options

- **Step 1: Software and Steps**
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security

Software Configuration

Release emr-5.29.0

[x] Hadoop 2.8.5	[ ] Zeppelin 0.8.2	[ ] Livy 0.6.0
[ ] JupyterHub 1.0.0	[ ] Tez 0.9.2	[ ] Flink 1.9.1
[ ] Ganglia 3.7.2	[ ] HBase 1.4.10	[ ] Pig 0.17.0
[x] Hive 2.3.6	[ ] Presto 0.227	[ ] ZooKeeper 3.4.14
[ ] MXNet 1.5.1	[ ] Sqoop 1.4.7	[ ] Mahout 0.13.0
[x] Hue 4.4.0	[ ] Phoenix 4.14.3	[ ] Oozie 5.1.0
[ ] Spark 2.4.4	[ ] HCatalog 2.3.6	[ ] TensorFlow 1.14.0

Multiple master nodes (optional)

[ ] Use multiple master nodes to improve cluster availability. Learn more

AWS Glue Data Catalog settings (optional)

[ ] Use for Hive table metadata

Edit software settings

(•) Enter configuration ( ) Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Feedback | Looking for language selection? Find it in the new Unified Settings | © 2022, Amazon Internet Services Private Ltd. or its affiliates. | Privacy | Terms | Cookie preferences

1. Creating EMR cluster – Step 2 - Hardware

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Create Cluster - Advanced Options

Go to quick options

- Step 1: Software and Steps
- **Step 2: Hardware**
- Step 3: General Cluster Settings
- Step 4: Security

Hardware Configuration

Specify the networking and hardware configuration for your cluster. Request Spot instances (unused EC2 capacity) to save money.

Cluster Composition

Specify the configuration of the master, core and task nodes as an instances group or instance fleet. This choice applies to all nodes for the lifetime of the cluster. Instance fleets and instance groups cannot coexist in a cluster. see this topic.

Instance group configuration

- (●) **Uniform instance groups**
  Specify a single instance type and purchasing option for each node type.
- ( ) **Instance fleets**
  Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. Learn more

Networking

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Launch the cluster into a VPC with a public, private or shared subnet. Subnets may be associated with and AWS Outpost or AWS Local Zone.

1. Creating EMR cluster – Step 2 - Hardware

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Networking

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Launch the cluster into a VPC with a public, private or shared subnet. Subnets may be associated with and AWS Outpost or AWS Local Zone.

Launch the cluster into a VPC with a public, private, or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Network vpc-05d6fc970fa018ac9 (172.31.0.0/16) (default) Create a VPC

EC2 Subnet subnet-020fe5a1dcc69e74d | Default in us-east-1e

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	**m4.large** 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	On-demand Spot Use on-demand as max price
Core Core - 2	**m4.large** 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	2 Instances	On-demand Spot Use on-demand as max price

Feedback | Looking for language selection? Find it in the new Unified Settings | © 2022, Amazon Internet Services Private Ltd. or its affiliates. | Privacy | Terms | Cookie preferences

1. Creating EMR cluster – Step 3 – General cluster settings

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Create Cluster - Advanced Options

Go to quick options

- Step 1: Software and Steps
- Step 2: Hardware
- **Step 3: General Cluster Settings**
- Step 4: Security

General Options

Cluster name hive_cp

- [x] Logging
  - S3 folder s3://aws-logs-502224226751-us-east-1/elasticmap
- [x] Debugging
- [x] Termination protection

Tags

Key	Value (optional)
Add a key to create a tag	

Additional Options

- [ ] EMRFS consistent view

Custom AMI ID None

Feedback | Looking for language selection? Find it in the new Unified Settings | © 2022, Amazon Internet Services Private Ltd. or its affiliates. | Privacy | Terms | Cookie preferences

1. Creating EMR cluster – Step 4 – Security Options

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- **Step 4: Security**

Security Options

EC2 key pair d210922

☑ Cluster visible to all IAM users in account

Permissions

◉ Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role EMR_DefaultRole ☐ Use EMR_DefaultRole_V2

EC2 instance profile EMR_EC2_DefaultRole

Auto Scaling role EMR_AutoScaling_DefaultRole

Security Configuration

EC2 security groups

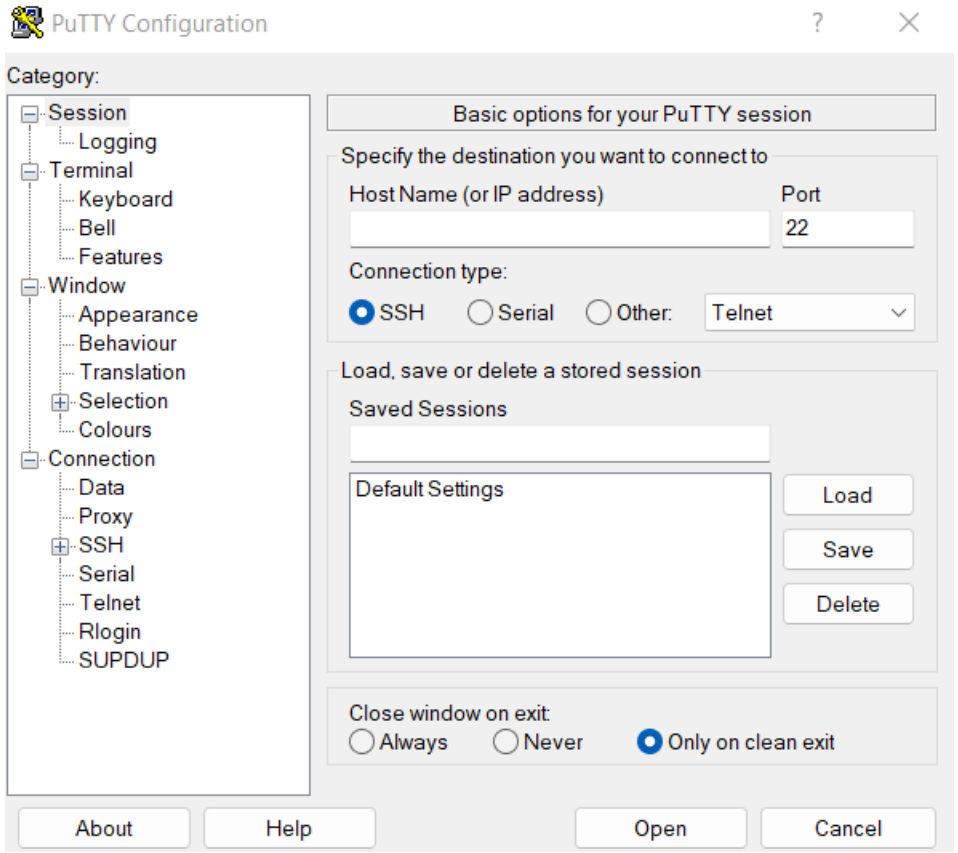
An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, EMR managed security groups and additional security groups. EMR will automatically update the rules in the EMR managed security groups in order to launch a cluster. Learn more.

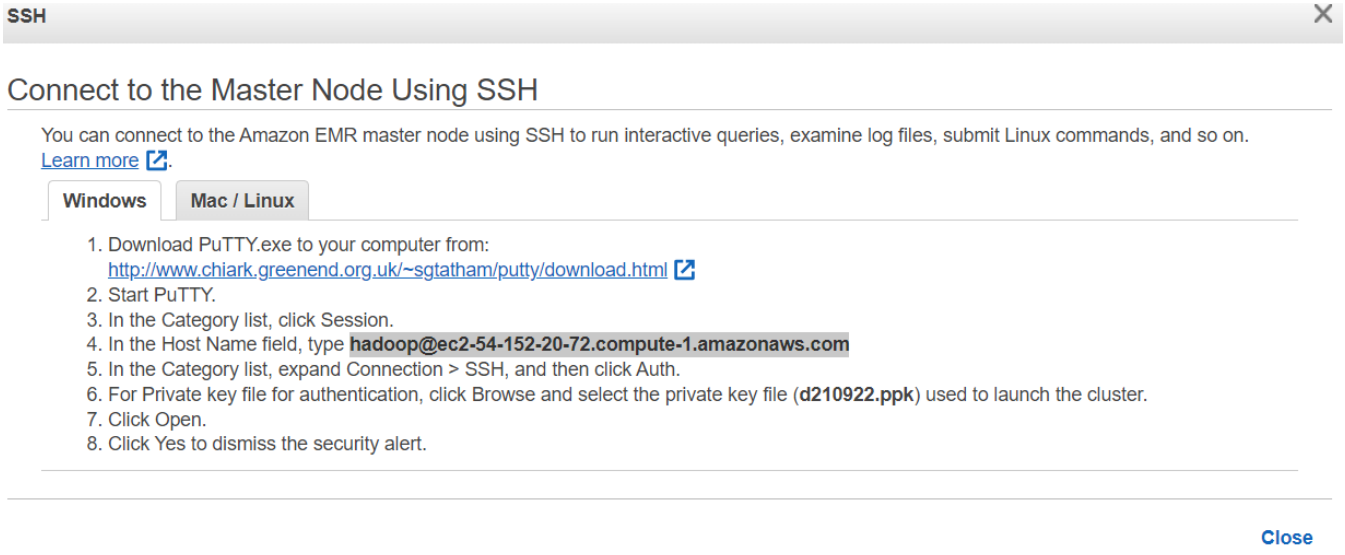
Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	Default: sg-0c9d80234e35d8ffe (ElasticMapReduc	*No security groups selected*
Core & Task	Default: sg-0d047e53fa45d2c39 (ElasticMapRedu	*No security groups selected*

Create a security group

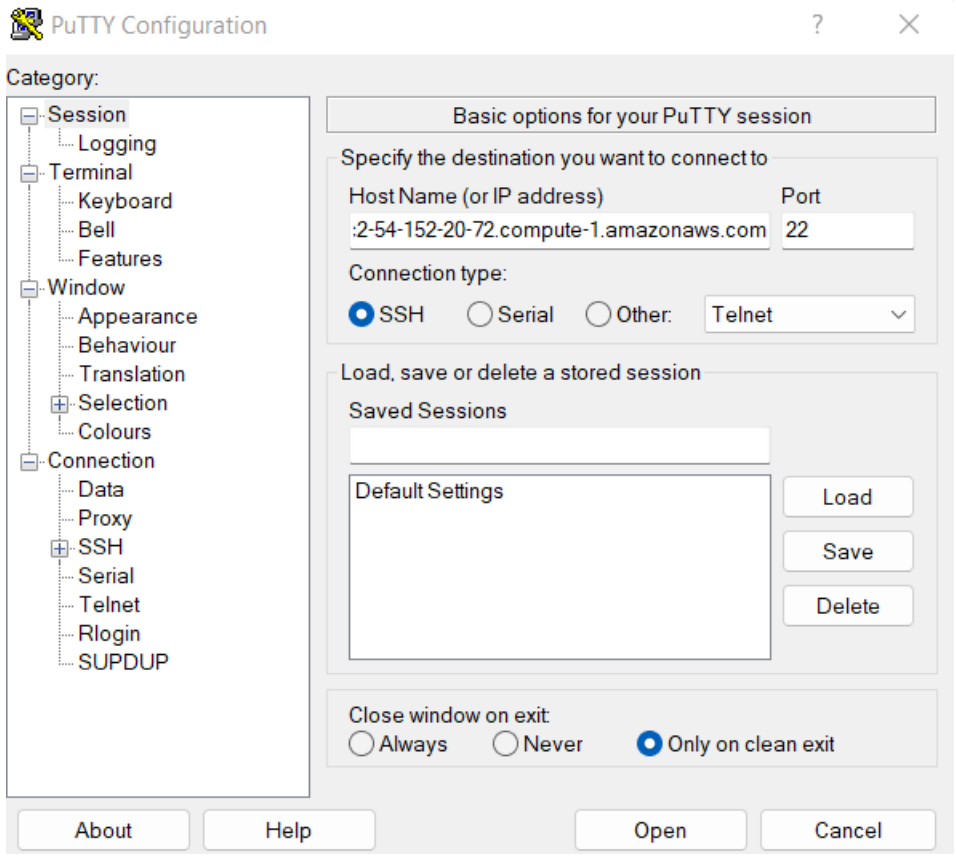
Feedback | Looking for language selection? Find it in the new Unified Settings | © 2022, Amazon Internet Services Private Ltd. or its affiliates. | Privacy | Terms | Cookie preferences

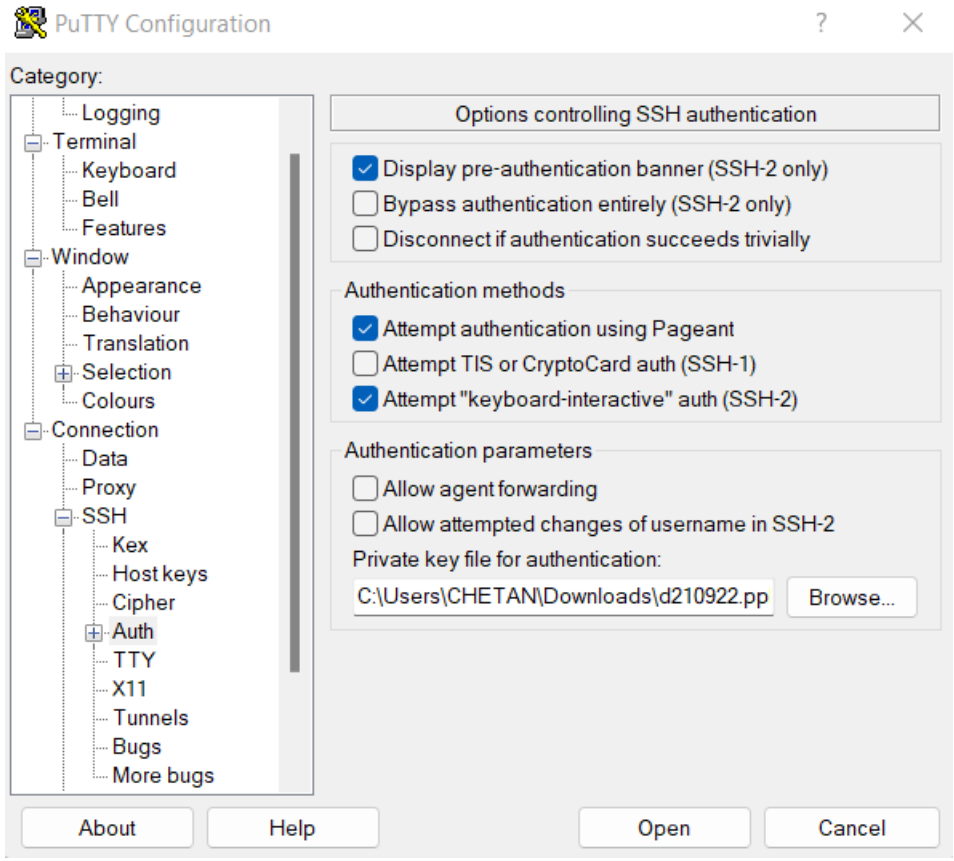
1. Creating EMR cluster – Connect to Master node using SSH





1. Creating EMR cluster – Successful creation of cluster





1. Creating EMR cluster – Successful creation of cluster

aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | N. Virginia | upgradchetanr @ 5022-2422-6751

Amazon EMR

- EMR Studio
- EMR Serverless New

EMR on EC2

- Clusters
- Notebooks
- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events

EMR on EKS

- Virtual clusters

- Help
- What's new

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Clone | Terminate | AWS CLI export

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

Cluster: hive_cp Running Running step

Summary | Application user interfaces | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Summary

ID: j-7TWH440K2GJU
Creation date: 2022-10-02 15:14 (UTC+5:30)
Elapsed time: 10 minutes
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: ec2-54-236-50-111.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-502224226751-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-1e
Subnet ID: subnet-020fe5a1dcc69e74d
Master: Running 1 m4.large
Core: Running 2 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: d210922

Feedback | Looking for language selection? Find it in the new Unified Settings | © 2022, Amazon Internet Services Private Ltd. or its affiliates. | Privacy | Terms | Cookie preferences

1. Creating EMR cluster – Successful creation of cluster

hadoop@ip-172-31-60-195:~

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

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

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
69 package(s) needed for security, out of 97 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRRR
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-172-31-60-195 ~]$
```

2. Writing command to check directories which are already present in HDFS

```
[hadoop@ip-172-31-60-195 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 08:37 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-10-02 08:40 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 08:37 /user
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 08:37 /var
```

```
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-10-02 09:50 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /user
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /var
```

3. Creating a new directory ‘case-study’ to store data file in the already present directory ‘user’

```
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-10-02 09:50 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /user
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /var
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -mkdir /user/case-study/
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x   - hadoop hadoop          0 2022-10-02 09:52 /user/case-study
drwxrwxrwx   - hadoop hadoop          0 2022-10-02 09:50 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2022-10-02 09:50 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2022-10-02 09:50 /user/hive
drwxrwxrwx   - hue    hue             0 2022-10-02 09:50 /user/hue
drwxrwxrwx   - oozie  oozie           0 2022-10-02 09:50 /user/oozie
drwxrwxrwx   - root   hadoop          0 2022-10-02 09:50 /user/root
```

[hadoop@ip-172-31-48-4 ~]$ hadoop fs -mkdir /user/case-study/
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x - hadoop hadoop 0 2022-10-02 09:52 /user/case-study
drwxrwxrwx - hadoop hadoop 0 2022-10-02 09:50 /user/hadoop
drwxr-xr-x - mapred mapred 0 2022-10-02 09:50 /user/history
drwxrwxrwx - hdfs hadoop 0 2022-10-02 09:50 /user/hive
drwxrwxrwx - hue hue 0 2022-10-02 09:50 /user/hue
drwxrwxrwx - oozie oozie 0 2022-10-02 09:50 /user/oozie
drwxrwxrwx - root hadoop 0 2022-10-02 09:50 /user/root

4. Write a command to check the creation of ‘case-study’ in ‘user’ directory.

```
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /apps
drwxrwxrwt   - hdfs hadoop          0 2022-10-02 09:50 /tmp
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /user
drwxr-xr-x   - hdfs hadoop          0 2022-10-02 09:50 /var
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -mkdir /user/case-study/
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x   - hadoop hadoop          0 2022-10-02 09:52 /user/case-study
drwxrwxrwx   - hadoop hadoop          0 2022-10-02 09:50 /user/hadoop
drwxr-xr-x   - mapred mapred          0 2022-10-02 09:50 /user/history
drwxrwxrwx   - hdfs   hadoop          0 2022-10-02 09:50 /user/hive
drwxrwxrwx   - hue    hue             0 2022-10-02 09:50 /user/hue
drwxrwxrwx   - oozie  oozie           0 2022-10-02 09:50 /user/oozie
drwxrwxrwx   - root   hadoop          0 2022-10-02 09:50 /user/root
```

[hadoop@ip-172-31-48-4 ~]$ hadoop fs -mkdir /user/case-study/
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /user/
Found 7 items
drwxr-xr-x - hadoop hadoop 0 2022-10-02 09:52 /user/case-study
drwxrwxrwx - hadoop hadoop 0 2022-10-02 09:50 /user/hadoop
drwxr-xr-x - mapred mapred 0 2022-10-02 09:50 /user/history
drwxrwxrwx - hdfs hadoop 0 2022-10-02 09:50 /user/hive
drwxrwxrwx - hue hue 0 2022-10-02 09:50 /user/hue
drwxrwxrwx - oozie oozie 0 2022-10-02 09:50 /user/oozie
drwxrwxrwx - root hadoop 0 2022-10-02 09:50 /user/root

5. Write a command to load the 1st data file '2019-Oct.csv' from S3 storage into HDFS.

hadoop@ip-172-31-48-4:~

```
[hadoop@ip-172-31-48-4 ~]$ hadoop distcp s3://e-commerce-events-ml/2019-Oct.csv /user/case-study/2019-Oct.csv
22/10/02 09:52:48 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=t
rue, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=
null, sourceFileListing=null, sourcePaths=[s3://e-commerce-events-ml/2019-Oct.csv], targetPath=/user/case-study/2019-Oct.csv, targetPathExists=false, filtersFile='null'}
22/10/02 09:52:48 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-4.ec2.internal/172.31.48.4:8032
22/10/02 09:53:00 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
22/10/02 09:53:00 INFO tools.SimpleCopyListing: Build file listing completed.
22/10/02 09:53:00 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/10/02 09:53:00 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/10/02 09:53:00 INFO tools.DistCp: Number of paths in the copy list: 1
22/10/02 09:53:00 INFO tools.DistCp: Number of paths in the copy list: 1
22/10/02 09:53:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-4.ec2.internal/172.31.48.4:8032
22/10/02 09:53:01 INFO mapreduce.JobSubmitter: number of splits:1
22/10/02 09:53:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664704288946_0001
22/10/02 09:53:02 INFO impl.YarnClientImpl: Submitted application application_1664704288946_0001
22/10/02 09:53:03 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-4.ec2.internal:20888/proxy/application_1664704288946_0001/
22/10/02 09:53:03 INFO tools.DistCp: DistCp job-id: job_1664704288946_0001
22/10/02 09:53:03 INFO mapreduce.Job: Running job: job_1664704288946_0001
22/10/02 09:53:13 INFO mapreduce.Job: Job job_1664704288946_0001 running in uber mode : false
22/10/02 09:53:13 INFO mapreduce.Job:  map 0% reduce 0%
22/10/02 09:53:31 INFO mapreduce.Job:  map 100% reduce 0%
22/10/02 09:53:34 INFO mapreduce.Job: Job job_1664704288946_0001 completed successfully
22/10/02 09:53:35 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172451
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=359
                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)=579360
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=18105
                Total vcore-milliseconds taken by all map tasks=18105
                Total megabyte-milliseconds taken by all map tasks=18539520
        Map-Reduce Framework
                Map input records=1
```

5. Write a command to load the 1st data file '2019-Oct.csv' from S3 storage into HDFS.

```
    Map-Reduce Framework
            Map input records=1
            Map output records=0
            Input split bytes=135
            Spilled Records=0
            Failed Shuffles=0
            Merged Map outputs=0
            GC time elapsed (ms)=334
            CPU time spent (ms)=19270
            Physical memory (bytes) snapshot=570200064
            Virtual memory (bytes) snapshot=3294228480
            Total committed heap usage (bytes)=502792192
    File Input Format Counters
            Bytes Read=223
    File Output Format Counters
            Bytes Written=0
    DistCp Counters
            Bytes Copied=482542278
            Bytes Expected=482542278
            Files Copied=1
```

6. Write a command to load the 2nd data file '2019-Nov.csv' from S3 storage into HDFS.

hadoop@ip-172-31-48-4:~

```
[hadoop@ip-172-31-48-4 ~]$ hadoop distcp s3://e-commerce-events-ml/2019-Nov.csv /user/case-study/2019-Nov.csv
22/10/02 09:54:43 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=t
rue, numListstatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=
null, sourceFileListing=null, sourcePaths=[s3://e-commerce-events-ml/2019-Nov.csv], targetPath=/user/case-study/2019-Nov.csv, targetPathExists=false, filtersFile='null'}
22/10/02 09:54:44 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-4.ec2.internal/172.31.48.4:8032
22/10/02 09:54:52 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
22/10/02 09:54:52 INFO tools.SimpleCopyListing: Build file listing completed.
22/10/02 09:54:52 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
22/10/02 09:54:52 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
22/10/02 09:54:52 INFO tools.DistCp: Number of paths in the copy list: 1
22/10/02 09:54:52 INFO tools.DistCp: Number of paths in the copy list: 1
22/10/02 09:54:52 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-48-4.ec2.internal/172.31.48.4:8032
22/10/02 09:54:52 INFO mapreduce.JobSubmitter: number of splits:1
22/10/02 09:54:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664704288946_0002
22/10/02 09:54:53 INFO impl.YarnClientImpl: Submitted application application_1664704288946_0002
22/10/02 09:54:53 INFO mapreduce.Job: The url to track the job: http://ip-172-31-48-4.ec2.internal:20888/proxy/application_1664704288946_0002/
22/10/02 09:54:53 INFO tools.DistCp: DistCp job-id: job_1664704288946_0002
22/10/02 09:54:53 INFO mapreduce.Job: Running job: job_1664704288946_0002
22/10/02 09:55:02 INFO mapreduce.Job: Job job_1664704288946_0002 running in uber mode : false
22/10/02 09:55:02 INFO mapreduce.Job:  map 0% reduce 0%
22/10/02 09:55:21 INFO mapreduce.Job:  map 100% reduce 0%
22/10/02 09:55:26 INFO mapreduce.Job: Job job_1664704288946_0002 completed successfully
22/10/02 09:55:27 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=172448
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=358
                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)=666752
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=20836
                Total vcore-milliseconds taken by all map tasks=20836
                Total megabyte-milliseconds taken by all map tasks=21336064
        Map-Reduce Framework
                Map input records=1
```

6. Write a command to load the 2nd data file '2019-Nov.csv' from S3 storage into HDFS.

```
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=137
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=398
                CPU time spent (ms)=20540
                Physical memory (bytes) snapshot=556814336
                Virtual memory (bytes) snapshot=3300945920
                Total committed heap usage (bytes)=468189184
        File Input Format Counters
                Bytes Read=223
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bytes Copied=545839412
                Bytes Expected=545839412
                Files Copied=1
```

7. Write a command to check the successful loading of both the data files in the created directory ‘case-study’.

```
[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /user/case-study/
Found 2 items
-rw-r--r--   1 hadoop hadoop  545839412 2022-10-02 09:55 /user/case-study/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop  482542278 2022-10-02 09:53 /user/case-study/2019-Oct.csv
```

[hadoop@ip-172-31-48-4 ~]$ hadoop fs -ls /user/case-study/
Found 2 items
-rw-r--r--   1 hadoop hadoop 545839412 2022-10-02 09:55 /user/case-study/2019-Nov.csv
-rw-r--r--   1 hadoop hadoop 482542278 2022-10-02 09:53 /user/case-study/2019-Oct.csv

8. Write a command to start the Hive system

```
[hadoop@ip-172-31-60-195 ~]$ hive



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

9. Create an external table named ‘ret’ which will hold the data for both the data files stored in directory ‘case-study’.

```
hive> create external table IF NOT EXISTS ret  (event_time timestamp, event_type string, product_id string,category_id string,category_code string, brand string,price float, user_id bigint,
 user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES("separatorChar=",", "quoteChar="\"", "escapeChar="\\") STORED AS TEXTFILE LOCATIO
N '/user/case-study/' TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.431 seconds
```

create external table IF NOT EXISTS ret (event_time timestamp, event_type string, product_id string,category_id string,category_code string, brand string,price float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES("separatorChar=",", "quoteChar="\"", "escapeChar="\\") STORED AS TEXTFILE LOCATION '/user/case-study/' TBLPROPERTIES ("skip.header.line.count"="1");

10. Write a command to enable headers in output.

```
hive> set hive.cli.print.header=true ;
```

11. Write a command in HQL to check successful creation of table & loading of both data files in the table.

```
hive> select * from ret limit 5;
OK
ret.event_time  ret.event_type  ret.product_id  ret.category_id ret.category_code      ret.brand      ret.price      ret.user_id     ret.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: 5.448 seconds, Fetched: 5 row(s)
```

hive> select * from ret limit 5;
OK
ret.event_time ret.event_type ret.product_id ret.category_id ret.category_code ret.brand ret.price ret.user_id ret.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: 5.448 seconds, Fetched: 5 row(s)

Questions & Answers by using Hive query

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

```
hive> SELECT SUM(price) as total_revenue from ret WHERE month(event_time)=10 and event_type = 'purchase';
Query ID = hadoop_20221003164436_d10efe9f-d0b9-4236-99cd-2034354afeff
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_1664813852545_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: 68.14 s
--------------------------------------------------------------------------------
OK
1211538.4299997438
Time taken: 78.941 seconds, Fetched: 1 row(s)
```

Insights : The total revenue generated due to purchases made in October is 1211538.4299997438

Questions & Answers by using Hive query

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

```
hive> SELECT SUM(price) as total_revenue from ret WHERE month(event_time) = 10 event_type ='purchase' ;
Query ID = hadoop_2022100219_bfabbf30-f565-499b-a228-ec28178d5dd7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE       STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      2          2        0        0       0       0
Reduce 2 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 68.14 s
--------------------------------------------------------------------------------
OK
total_revenue
1211538.4299997438
Time taken: 78.941 seconds, Fetched: 1 row(s)
```

Questions & Answers by using Hive query

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

```
hive> SELECT date_format(event_time,'MM') AS Month, COUNT (event_type) AS Sum_Purchase FROM ret WHERE event_type = 'purchase' GROUP BY date_format(event_time,'MM') ;
Query ID = hadoop_20221002102142_48956520-32df-47f9-b5ca-05d76ae7a7fb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0004)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 41.19 s
--------------------------------------------------------------------------------------------
OK
month   sum_purchase
10      245624
11      322417
Time taken: 41.904 seconds, Fetched: 2 row(s)
```

Insights :

*** The total sum of purchases per month is:**

10 245624

11 322417

*** The month of November is more profitable as compared to October**

Questions & Answers by using Hive query

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

```
hive> SELECT date_format(event_time,'MM') AS Month, COUNT (event_type) AS Sum_Purchase FROM ret WHERE event_type =
'purchase' GROUP BY date_format(event_time,'MM') ;
Query ID = hadoop_20221002102142_48956520-32df-47f9-b5ca-05d76ae7a7fb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE       STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      2          2        0        0       0       0
Reduce 2 ......  container     SUCCEEDED      3          3        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 41.19 s
--------------------------------------------------------------------------------
OK
month   sum_purchase
10      245624
11      322417
Time taken: 41.904 seconds, Fetched: 2 row(s)
```

Questions & Answers by using Hive query

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

```
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 ret WHERE event_type = 'purchase' AND date_format(event_time, 'MM') in ('10', '11')) SELECT Nov_Revenue, Oct_Revenue, (Nov_revenue - Oct_Revenue) AS Revenue_Dif
f FROM Monthly_Revenue ;
Query ID = hadoop_20221002110328_d4288817-e8ae-49af-b7f2-672d4c6bbb6a
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_1664704288946_0006)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 44.84 s
--------------------------------------------------------------------------------------------------
OK
nov_revenue     oct_revenue     revenue_diff
1531016.900000122       1211538.4299997438      319478.4700003781
Time taken: 53.564 seconds, Fetched: 1 row(s)
```

Insights:

- **The change in revenue generated due to purchases from October to November is 319478.4700003781**
- **The revenue generated for November is much better than that of October. It means November experienced better sale as compared to October.**

Questions & Answers by using Hive query

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

```
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 ret WHERE
event_type = 'purchase' AND date_format(event_time, 'MM') in ('10', '11')) SELECT Nov_Revenue, Oct_Revenue, (Nov_revenue -
Oct_Revenue) AS Revenue_Diff FROM Monthly_Revenue ;
Query ID = hadoop_20221002110328_d4288817-e8ae-49af-b7f2-672d4c6bbb6a
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_1664704288946_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE       STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 44.84 s
--------------------------------------------------------------------------------
OK
nov_revenue                    oct_revenue                    revenue_diff
1531016.900000122      1211538.4299997438      319478.4700003781
Time taken: 53.564 seconds, Fetched: 1 row(s)
```

Questions & Answers by using Hive query

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

```
hive> SELECT DISTINCT SPLIT (category_code,'\\.')[0] AS Category FROM ret WHERE SPLIT(category_code,'\\.')[0] <> '' ;
Query ID = hadoop_20221002103250_b2f427d0-db7d-44ee-bc46-f62733b7d3e9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      2          2        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 49.22 s
--------------------------------------------------------------------------------
OK
category
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 50.048 seconds, Fetched: 6 row(s)
```

Insights :

* **There are 6 distinct categories of products namely furniture, appliances, accessories, apparel, sport, stationery.**

Questions & Answers by using Hive query

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

```
hive> SELECT DISTINCT SPLIT (category_code,'\\.')[0] AS Category FROM ret WHERE SPLIT(category_code,'\\.')[0] <> '' ;
Query ID = hadoop_20221002103250_b2f427d0-db7d-44ee-bc46-f62733b7d3e9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0005)

--------------------------------------------------------------------------------------------
        VERTICES      MODE       STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      2          2        0        0       0       0
Reduce 2 ......  container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 49.22 s
--------------------------------------------------------------------------------------------
OK
category
furniture
appliances
accessories
apparel
sport
stationery
Time taken: 50.048 seconds, Fetched: 6 row(s)
```

Questions & Answers by using Hive query

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

```
hive> SELECT SPLIT (category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_Products FROM ret WHERE SPLIT(category_code,'\\.')[0] <> '' GROUP BY SPLIT(category_code,'\\.')[0] ORDER
BY No_of_Products DESC ;
Query ID = hadoop_20221002103656_5c40fb4a-5c56-4c96-9283-46d56f9979ec
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0005)

--------------------------------------------------------------------------------------------------
        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: 42.17 s
--------------------------------------------------------------------------------------------------
OK
category        no_of_products
appliances      61736
stationery      26722
furniture       23604
apparel 18232
accessories     12929
sport   2
Time taken: 43.012 seconds, Fetched: 6 row(s)
```

Insights :

*** Appliances have the most number of products(61736) whereas sports have the least(2).**

Questions & Answers by using Hive query

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

```
hive> SELECT SPLIT (category_code,'\\.')[0] AS Category, COUNT(product_id) AS No_of_Products FROM ret WHERE SPLIT(category_code,'\\.')[0] <> ''
GROUP BY SPLIT(category_code,'\\.')[0] ORDER BY No_of_Products DESC ;
Query ID = hadoop_20221002103656_5c40fb4a-5c56-4c96-9283-46d56f9979ec
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0005)

--------------------------------------------------------------------------------------------------
        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: 42.17 s
--------------------------------------------------------------------------------------------------
OK
category        no_of_products
appliances      61736
stationery      26722
furniture       23604
apparel         18232
accessories     12929
sport           2
Time taken: 43.012 seconds, Fetched: 6 row(s)
```

Questions & Answers by using Hive query

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

```
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 E
LSE 0 END) AS Nov_Sales FROM ret 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 Tota
l_Sales FROM Max_Sales_Brand ORDER BY Total_Sales DESC LIMIT 1;
Query ID = hadoop_20221002111206_850d442d-4921-4ce5-a835-0ef42a1c380f
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_1664704288946_0007)

--------------------------------------------------------------------------------
        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: 43.59 s
--------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.9400000003
Time taken: 53.174 seconds, Fetched: 1 row(s)
```

Insights :

*** Runail is the brand that has the maximum sales in October and November combined i.e. 148297.9400000003**

Questions & Answers by using Hive query

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

```
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 ret 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_20221002111206_850d442d-4921-4ce5-a835-0ef42a1c380f
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_1664704288946_0007)

--------------------------------------------------------------------------------------------------
        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: 43.59 s
--------------------------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.9400000003
Time taken: 53.174 seconds, Fetched: 1 row(s)
```

Questions & Answers by using Hive query

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

```
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 E
LSE 0 END) AS Nov_Sales FROM ret 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 Tota
l_Sales FROM Max_Sales_Brand ORDER BY Total_Sales DESC LIMIT 1;
Query ID = hadoop_20221002100644_fe017443-ed93-450a-982b-52cbbedb3ba9
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_1664704288946_0004)

--------------------------------------------------------------------------------------------------
        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: 44.49 s
--------------------------------------------------------------------------------------------------
OK
brand   total_sales
runail  148297.9400000003
Time taken: 57.781 seconds, Fetched: 1 row(s)
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 ret WHERE ( event_type='purchase' AND date_format(event_time, 'MM') in ('10','11') AND brand <> '') GROUP BY brand ) SELECT brand, Nov_Revenue, Oct_Revenue,
Nov_Revenue - Oct_Revenue AS Sales_Difference FROM Monthly_Revenue WHERE (Nov_Revenue - Oct_Revenue)>0 ORDER BY Sales_Difference;
Query ID = hadoop_20221002101102_a251119d-a575-467a-a304-d55dd80a7df7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0004)

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

Insights :

- **Runail is the most popular brand with an increment of total 5219. 38/- from October to November.**
- **A total of 161 brands have an increment in sales from October to November.**

Questions & Answers by using Hive query

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

hadoop@ip-172-31-48-4:~

```
--------------------------------------------------------------------------------
        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: 43.06 s
--------------------------------------------------------------------------------
OK
brand   nov_revenue     oct_revenue     sales_difference
ovale   3.1     2.54    0.56
cosima  20.929999999999993      20.23   0.6999999999999922
grace   102.61000000000001      100.92000000000002      1.6899999999999977
helloganic      3.1     0.0     3.1
skinity 12.440000000000001      8.88    3.5600000000000005
bodyton 1380.6399999999992      1376.3399999999974      4.3000000000017735
moyou   10.280000000000001      5.71    4.570000000000001
neoleor 51.7    43.41   8.290000000000006
soleo   212.5299999999998        204.2000000000003       8.329999999999501
jaguar  1110.6500000000003      1102.11 8.540000000000418
tertio  245.79999999999978      236.16000000000008      9.639999999999702
fly     27.17   17.14   10.030000000000001
rasyan  28.939999999999994      18.799999999999997      10.139999999999997
deoproce        329.1700000000001       316.84  12.330000000000098
barbie  12.39   0.0     12.39
supertan        66.51000000000002       50.37000000000001       16.140000000000008
treaclemoon     181.48999999999995      163.36999999999995      18.120000000000005
kamill  81.49000000000002       63.00999999999999       18.480000000000032
juno    21.08   0.0     21.08
veraclara       71.21000000000001       50.109999999999985      21.100000000000023
glysolid        91.58999999999997       69.72999999999998       21.86
godefroy        425.12000000000006      401.2200000000002       23.899999999999864
binacil 24.259999999999998      0.0     24.259999999999998
blixz   63.39999999999998       38.949999999999996      24.44999999999998
profepil        118.02000000000005      93.36000000000003       24.660000000000025
estelare        471.8700000000009       444.80999999999943      27.06000000000148
orly    931.0900000000003       902.3800000000005       28.70999999999981
biore   90.31   60.650000000000006      29.659999999999997
beautyblender   109.41  78.74000000000001       30.669999999999987
vilenta 231.2100000000002       197.6000000000002       33.610000000000014
mavala  446.32  409.03999999999985      37.28000000000014
likato  340.9699999999999       296.0599999999999       44.910000000000025
ladykin 170.57  125.64999999999999      44.92
foamie  80.49   35.04   45.449999999999996
elskin  307.65000000000055      251.09000000000057      56.559999999999974
balbcare        212.38000000000025      155.32999999999996      57.050000000000296
koelcia 112.75000000000003      55.5    57.25000000000003
profhenna       736.8500000000005       679.2299999999999       57.62000000000057
```

Questions & Answers by using Hive query

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

hadoop@ip-172-31-48-4:~

```
ladykin 170.57  125.64999999999999      44.92
foamie  80.49   35.04   45.449999999999996
elskin  307.65000000000055      251.09000000000057      56.559999999999974
balbcare        212.38000000000025      155.32999999999996      57.050000000000296
koelcia 112.75000000000003      55.5    57.25000000000003
profhenna       736.8500000000005       679.2299999999999       57.62000000000057
kares   59.45   0.0     59.45
marutaka-foot   109.33  49.21999999999999       60.11000000000001
dewal   61.29   0.0     61.29
inm     351.2100000000001       288.02  63.19000000000011
laboratorium    312.52  246.49999999999991      66.02000000000007
cutrin  367.62  299.36999999999995      68.25000000000006
egomania        146.04000000000002      77.47   68.57000000000002
konad   810.6700000000003       739.8299999999991       70.84000000000117
nirvel  234.32999999999984      163.03999999999996      71.28999999999988
koelf   507.2900000000002       422.72999999999985      84.56000000000034
plazan  194.01000000000002      101.37  92.64000000000001
aura    177.51  83.95   93.55999999999999
kerasys 525.2000000000002       430.90999999999985      94.2900000000003
enjoy   136.57000000000002      41.349999999999994      95.22000000000003
depilflax       2803.7799999999975      2707.069999999994       96.71000000000367
eos     152.61  54.339999999999996      98.27000000000001
carmex  243.36  145.08  98.28
batiste 874.1699999999994       772.3999999999999       101.76999999999953
osmo    762.3100000000002       645.58  116.73000000000013
dizao   945.5099999999998       819.1300000000012       126.37999999999852
igrobeauty      645.0699999999999       513.6600000000009       131.40999999999906
finish  230.38000000000008      98.38   132.00000000000009
nefertiti       366.64  233.52000000000007      133.11999999999992
elizavecca      204.3   70.53   133.77
miskin  293.07000000000005      158.04  135.03000000000006
latinoil        384.59  249.52  135.06999999999996
farmona 1843.4300000000007      1692.4599999999996      150.97000000000116
cristalinas     584.949999999999        427.6299999999999       157.31999999999914
chi     538.6100000000002       358.9400000000002       179.67000000000002
matreshka       182.67000000000002      0.0     182.67000000000002
freshbubble     502.34000000000015      318.7000000000001       183.64000000000004
mane    260.26  66.78999999999999       193.47
keen    435.62  236.35000000000005      199.26999999999995
ecocraft        241.95  41.160000000000004      200.79
fedua   263.81000000000006      52.38   211.43000000000006
provoc  1063.8200000000006      827.9900000000009       235.8299999999997
skinlite        890.4499999999979       651.9400000000002       238.50999999999772
entity  719.2599999999993       479.7100000000015       239.5499999999978
trind   542.9600000000002       298.07000000000005      244.8900000000001
protokeratin    456.79000000000013      201.25000000000003      255.5400000000001
beauugreen      768.35  511.5099999999999       256.84000000000015
bluesky 10565.529999999713      10307.239999999858      258.28999999985535
candy   799.3799999999993       534.9599999999999       264.4199999999994
```

Questions & Answers by using Hive query

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

hadoop@ip-172-31-48-4:~

```
candy   799.3799999999993       534.9599999999999       264.4199999999994
insight 1721.9600000000003      1443.7000000000012      278.2599999999991
kocostar        594.9300000000003       310.8500000000001       284.0800000000002
happyfons       1091.5900000000001      801.9200000000006       289.6699999999995
kims    632.0400000000001       330.03999999999996      302.0000000000001
shary   1176.4899999999989      871.9599999999994       304.5299999999995
nitrile 1162.679999999999       847.279999999999        315.4
lowence 567.7499999999997       242.84  324.9099999999996
jas     3657.4300000000026      3318.959999999995       338.47000000000753
ellips  606.0399999999996       245.8499999999999       360.1899999999997
lador   2471.530000000007       2083.610000000004       387.9200000000028
naomi   389.0   0.0     389.0
kiss    817.3299999999994       421.54999999999944      395.7799999999999
yu-r    673.7099999999998       271.41  402.2999999999998
sophin  1515.5200000000011      1067.8600000000001      447.660000000001
farmavita       1291.9700000000003      837.3699999999984       454.60000000000184
bioaqua 1398.1199999999997      942.8899999999996       455.23
greymy  489.49  29.21   460.28000000000003
gehwol  1557.6799999999982      1089.07 468.6099999999983
matrix  3726.7400000000007      3243.249999999999       483.4900000000016
limoni  1796.5999999999997      1308.9000000000003      487.69999999999936
s.care  913.0699999999999       412.68  500.38999999999993
coifin  1428.4899999999998      903.0000000000001       525.4899999999997
uskusi  5690.310000000005       5142.270000000017       548.0399999999881
airnails        5691.519999999996       5118.899999999939       572.6200000000572
browxenna       14916.729999999976      14331.36999999995       585.360000000026
kinetics        6945.260000000017       6334.2499999999945      611.010000000022
kosmekka        1813.37 1181.4400000000003      631.9299999999996
kaaral  5086.069999999992       4412.4299999999985      673.639999999994
refectocil      3475.580000000007       2716.180000000005       759.4000000000024
rosi    3841.560000000013       3077.0399999999927      764.5200000000204
solomeya        2685.799999999991       1899.699999999992       786.099999999999
missha  2150.2799999999984      1293.8299999999995      856.4499999999989
levissime       3085.3099999999977      2227.5000000000064      857.8099999999913
art-visage      2997.800000000011       2092.71000000001        905.090000000001
ecolab  1214.2999999999988      262.8500000000001       951.4499999999987
nagaraku        5327.680000000063       4369.740000000054       957.9400000000087
sanoto  1209.6799999999998      157.14  1052.54
markell 2834.4300000000007      1768.7499999999989      1065.6800000000019
metzger 6457.159999999988       5373.450000000006       1083.7099999999818
de.lux  2775.509999999968       1659.699999999967       1115.8100000000009
swarovski       3043.160000000003       1887.9299999999873      1155.2300000000157
beauty-free     1782.8600000000163      554.1700000000006       1228.6900000000155
zeitun  2009.63 708.6600000000004       1300.9699999999998
joico   2015.1000000000015      705.52  1309.5800000000015
severina        6120.480000000023       4775.88 1344.600000000023
irisk   46946.040000002184      45591.96000000588       1354.0799999963056
oniq    9841.650000000018       8425.41000000003        1416.239999999987
levrana 3664.099999999998       2243.560000000002       1420.5399999999959
```

Questions & Answers by using Hive query

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

```
irisk    46946.040000002184      45591.96000000588       1354.0799999963056
oniq     9841.650000000018       8425.41000000003        1416.239999999987
levrana  3664.099999999998       2243.560000000002       1420.5399999999959
roubloff         4913.769999999991       3491.360000000003       1422.4099999999885
smart    5902.140000000017       4457.260000000004       1444.8800000000128
shik     4839.720000000007       3341.2  1498.5200000000068
domix    12009.170000000022      10472.04999999994       1537.1200000000827
artex    4327.250000000017       2730.639999999998       1596.6100000000192
beautix  12222.949999999913      10493.949999999966      1728.9999999999472
milv     5642.01000000008        3904.9399999999964      1737.0700000000838
masura   33058.46999999708       31266.07999999821       1792.3899999988753
f.o.x    8577.280000000004       6624.229999999982       1953.050000000022
kapous   14093.080000000158      11927.159999999898      2165.92000000026
concept  13380.39999999993       11032.139999999925      2348.2600000000057
estel    24142.67000000022       21756.750000000342      2385.919999999878
kaypro   3268.699999999995       881.3399999999998       2387.359999999995
benovy   3259.970000000001       409.6200000000002       2850.350000000001
italwax  24799.369999999893      21940.239999999732      2859.130000000161
yoko     11707.879999999996      8756.909999999949       2950.9700000000466
haruyama         12352.91000000013       9390.689999999991       2962.2200000001394
marathon         10273.1 7280.749999999997       2992.350000000003
lovely   11939.060000000045      8704.379999999952       3234.680000000093
bpw.style        14837.440000000812      11572.150000001699      3265.289999999113
staleks  11875.61000000008       8519.730000000003       3355.8800000000774
freedecor        7671.800000000175       3421.779999999971       4250.020000000204
runail   76758.66000000098       71539.27999999933       5219.380000001649
polarus  11371.930000000018      6013.720000000003       5358.2100000000155
cosmoprofi       14536.99000000016       8322.81000000007        6214.180000000089
jessnail         33345.22999999992       26287.839999999916      7057.390000000007
strong   38671.269999999924      29196.62999999994       9474.639999999985
ingarden         33566.21000000009       23161.390000000138      10404.819999999949
lianail  16394.240000000245      5892.839999999975       10501.40000000027
uno      51039.749999998035      35302.02999999977       15737.719999998262
grattol  71472.71000000068       35445.5400000011        36027.169999999576
Time taken: 44.169 seconds, Fetched: 160 row(s)
```

Questions & Answers by using Hive query

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

```
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 ret WHERE ( event_type='purchase' AND date_format(event_time, 'MM') in ('10','11')
AND brand <> '') GROUP BY brand ) SELECT brand, Nov_Revenue, Oct_Revenue,Nov_Revenue - Oct_Revenue AS
Sales_Difference FROM Monthly_Revenue WHERE (Nov_Revenue - Oct_Revenue)>0 ORDER BY Sales_Difference;
Query ID = hadoop_20221002101102_a251119d-a575-467a-a304-d55dd80a7df7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_0004)

--------------------------------------------------------------------------------
        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: 43.06 s
--------------------------------------------------------------------------------
OK
```

Questions & Answers by using Hive query

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

brand	nov_revenue	oct_revenue	sales_difference
ovale	3.1	2.54	0.56
cosima	20.929999999999993	20.23	0.6999999999999922
grace	102.61000000000001	100.92000000000002	1.6899999999999977
helloganic	3.1	0.0	3.1
skinity	12.440000000000001	8.88	3.5600000000000005
bodyton	1380.6399999999992	1376.3399999999974	4.3000000000017735
moyou	10.280000000000001	5.71	4.570000000000001
neoleor	51.7	43.41	8.290000000000006
soleo	212.5299999999998	204.2000000000003	8.329999999999501
jaguar	1110.6500000000003	1102.11	8.540000000000418
tertio	245.79999999999978	236.16000000000008	9.639999999999702
fly	27.17	17.14	10.030000000000001
rasyan	28.939999999999994	18.799999999999997	10.139999999999997
deoproce	329.1700000000001	316.84	12.330000000000098
barbie	12.39	0.0	12.39
supertan	66.51000000000002	50.37000000000001	16.140000000000008
treaclemoon	181.48999999999995	163.36999999999995	18.120000000000005
kamill	81.49000000000002	63.00999999999999	18.480000000000032
juno	21.08	0.0	21.08
veraclara	71.21000000000001	50.109999999999985	21.100000000000023
glysolid	91.58999999999997	69.72999999999998	21.86
godefroy	425.12000000000006	401.2200000000002	23.899999999999864
binacil	24.259999999999998	0.0	24.259999999999998
blixz	63.39999999999998	38.949999999999996	24.44999999999998

Questions & Answers by using Hive query

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

profepil 118.02000000000005 93.36000000000003
24.660000000000025
estelare 471.8700000000009 444.80999999999943
27.06000000000148
orly 931.0900000000003 902.3800000000005 28.70999999999981
biore 90.31 60.650000000000006 29.659999999999997
beautyblender 109.41 78.74000000000001 30.669999999999987
vilenta 231.2100000000002 197.6000000000002 33.610000000000014
mavala 446.32 409.03999999999985 37.28000000000014
likato 340.9699999999999 296.0599999999999 44.910000000000025
ladykin 170.57 125.64999999999999 44.92
foamie 80.49 35.04 45.449999999999996
elskin 307.65000000000055 251.09000000000057 56.559999999999974
balbcare 212.38000000000025 155.32999999999996
57.050000000000296
koelcia 112.75000000000003 55.5 57.25000000000003
profhenna 736.8500000000005 679.2299999999999

Questions & Answers by using Hive query

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

kares 59.45 0.0 59.45
marutaka-foot 109.33 49.21999999999999 60.11000000000001
dewal 61.29 0.0 61.29
inm 351.2100000000001 288.02 63.19000000000011
laboratorium 312.52 246.49999999999991 66.02000000000007
cutrin 367.62 299.36999999999995 68.25000000000006
egomania 146.04000000000002 77.47 68.57000000000002
konad 810.6700000000003 739.8299999999991 70.84000000000117
nirvel 234.32999999999984 163.03999999999996 71.28999999999988
koelf 507.2900000000002 422.72999999999985 84.56000000000034
plazan 194.01000000000002 101.37 92.64000000000001
aura 177.51 83.95 93.55999999999999
kerasys 525.2000000000002 430.90999999999985 94.2900000000003
enjoy 136.57000000000002 41.349999999999994 95.22000000000003
depilflax 2803.7799999999975 2707.069999999994
96.71000000000367
eos 152.61 54.339999999999996 98.27000000000001
carmex 243.36 145.08 98.28
batiste 874.1699999999994 772.3999999999999 101.76999999999953
osmo 762.3100000000002 645.58 116.73000000000013
dizao 945.5099999999998 819.1300000000012 126.37999999999852
igrobeauty 645.0699999999999 513.6600000000009
131.40999999999906

Questions & Answers by using Hive query

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

finish 230.38000000000008 98.38 132.00000000000009
nefertiti 366.64 233.52000000000007 133.11999999999992
elizavecca 204.3 70.53 133.77
miskin 293.07000000000005 158.04 135.03000000000006
latinoil 384.59 249.52 135.06999999999996
farmona 1843.4300000000007 1692.4599999999996 150.97000000000116
cristalinas 584.949999999999 427.6299999999999 157.31999999999914
chi 538.6100000000002 358.9400000000002 179.67000000000002
matreshka 182.67000000000002 0.0 182.67000000000002
freshbubble 502.34000000000015 318.7000000000001 183.64000000000004
mane 260.26 66.78999999999999 193.47
keen 435.62 236.35000000000005 199.26999999999995
ecocraft 241.95 41.160000000000004 200.79
fedua 263.81000000000006 52.38 211.43000000000006
provoc 1063.8200000000006 827.9900000000009 235.8299999999997
skinlite 890.4499999999979 651.9400000000002 238.50999999999772
entity 719.2599999999993 479.7100000000015 239.5499999999978
trind 542.9600000000002 298.07000000000005 244.8900000000001
protokeratin 456.79000000000013 201.25000000000003 255.5400000000001
beauugreen 768.35 511.5099999999999 256.84000000000015
bluesky 10565.529999999713 10307.239999999858 258.28999999985535
candy 799.3799999999993 534.9599999999999 264.4199999999994
insight 1721.9600000000003 1443.7000000000012 278.2599999999991

Questions & Answers by using Hive query

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

kocostar 594.9300000000003 310.8500000000001
284.0800000000002
happyfons 1091.5900000000001 801.9200000000006
289.6699999999995
kims 632.0400000000001 330.03999999999996 302.0000000000001
shary 1176.4899999999989 871.9599999999994 304.5299999999995
nitrile 1162.679999999999 847.279999999999 315.4
lowence 567.7499999999997 242.84 324.9099999999996
jas 3657.4300000000026 3318.959999999995 338.47000000000753
ellips 606.0399999999996 245.8499999999999 360.1899999999997
lador 2471.530000000007 2083.610000000004 387.9200000000028
naomi 389.0 0.0 389.0
kiss 817.3299999999994 421.54999999999944 395.7799999999999
yu-r 673.7099999999998 271.41 402.2999999999998
sophin 1515.5200000000011 1067.8600000000001 447.660000000001
farmavita 1291.9700000000003 837.3699999999984
454.60000000000184
bioaqua 1398.1199999999997 942.8899999999996 455.23
greymy 489.49 29.21 460.28000000000003
gehwol 1557.6799999999982 1089.07 468.6099999999983
matrix 3726.7400000000007 3243.249999999999 483.4900000000016
limoni 1796.5999999999997 1308.9000000000003 487.69999999999936
s.care 913.0699999999999 412.68 500.38999999999993
coifin 1428.4899999999998 903.0000000000001 525.4899999999997
uskusi 5690.310000000005 5142.270000000017 548.0399999999881
airnails 5691.519999999996 5118.899999999939 572.6200000000572

Questions & Answers by using Hive query

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

browxenna	14916.729999999976	14331.36999999995	585.360000000026
kinetics	6945.260000000017	6334.2499999999945	611.010000000022
kosmekka	1813.37	1181.4400000000003	631.9299999999996
kaaral	5086.069999999992	4412.4299999999985	673.639999999994
refectocil	3475.580000000007	2716.180000000005	759.4000000000024
rosi	3841.560000000013	3077.0399999999927	764.5200000000204
solomeya	2685.799999999991	1899.699999999992	786.099999999999
missha	2150.2799999999984	1293.8299999999995	856.4499999999989
levissime	3085.3099999999977	2227.5000000000064	857.8099999999913
art-visage	2997.800000000011	2092.71000000001	905.090000000001
ecolab	1214.2999999999988	262.8500000000001	951.4499999999987
nagaraku	5327.680000000063	4369.740000000054	957.9400000000087
sanoto	1209.6799999999998	157.14	1052.54
markell	2834.4300000000007	1768.7499999999989	1065.6800000000019
metzger	6457.159999999988	5373.450000000006	1083.7099999999818
de.lux	2775.509999999968	1659.699999999967	1115.8100000000009
swarovski	3043.160000000003	1887.9299999999873	1155.2300000000157
beauty-free	1782.8600000000163	554.1700000000006	1228.6900000000155
zeitun	2009.63	708.6600000000004	1300.9699999999998
joico	2015.1000000000015	705.52	1309.5800000000015
severina	6120.480000000023	4775.88	1344.600000000023
irisk	46946.040000002184	45591.96000000588	1354.0799999963056
oniq	9841.650000000018	8425.41000000003	1416.239999999987
levrana	3664.099999999998	2243.560000000002	1420.5399999999959
roubloff	4913.769999999991	3491.360000000003	1422.4099999999885

Questions & Answers by using Hive query

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

smart	5902.140000000017	4457.260000000004	1444.8800000000128
shik	4839.720000000007	3341.2	1498.5200000000068
domix	12009.170000000022	10472.04999999994	1537.1200000000827
artex	4327.250000000017	2730.639999999998	1596.6100000000192
beautix	12222.949999999913	10493.949999999966	1728.9999999999472
milv	5642.01000000008	3904.9399999999964	1737.0700000000838
masura	33058.46999999708	31266.07999999821	1792.3899999988753
f.o.x	8577.280000000004	6624.229999999982	1953.050000000022
kapous	14093.080000000158	11927.159999999898	2165.92000000026
concept	13380.39999999993	11032.139999999925	2348.2600000000057
estel	24142.67000000022	21756.750000000342	2385.919999999878
kaypro	3268.699999999995	881.3399999999998	2387.359999999995
benovy	3259.970000000001	409.6200000000002	2850.350000000001
italwax	24799.369999999893	21940.239999999732	2859.130000000161
yoko	11707.879999999996	8756.909999999949	2950.9700000000466
haruyama	12352.91000000013	9390.689999999991	2962.2200000001394
marathon	10273.1	7280.749999999997	2992.350000000003
lovely	11939.060000000045	8704.379999999952	3234.680000000093
bpw.style	14837.440000000812	11572.150000001699	3265.289999999113
staleks	11875.61000000008	8519.730000000003	3355.8800000000774
freedecor	7671.800000000175	3421.779999999971	4250.020000000204

Questions & Answers by using Hive query

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

runail 76758.66000000098 71539.27999999933 5219.380000001649
polarus 11371.930000000018 6013.720000000003 5358.2100000000155
cosmoprofi 14536.99000000016 8322.81000000007
6214.180000000089
jessnail 33345.22999999992 26287.839999999916
7057.390000000007
strong 38671.269999999924 29196.62999999994 9474.639999999985
ingarden 33566.21000000009 23161.390000000138
10404.819999999949
lianail 16394.240000000245 5892.839999999975 10501.40000000027
uno 51039.749999998035 35302.02999999977 15737.719999998262
grattol 71472.71000000068 35445.5400000011 36027.169999999576

Questions & Answers by using Hive query

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.

```
hive> SELECT user_id, SUM(price) AS Total_Expenditure from ret WHERE event_type = 'purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10 ;
Query ID = hadoop_20221002104018_fd318a6d-97a2-44b6-9dbe-b373343bdbd9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 35.77 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: 36.451 seconds, Fetched: 10 row(s)
```

Insights : The above mentioned top 10 users should be rewarded with a Golden Customer Plan as they spend the most and this way more people will be attracted.

Questions & Answers by using Hive query

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.

```
hive> SELECT user_id, SUM(price) AS Total_Expenditure from ret WHERE event_type = 'purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10 ;
Query ID = hadoop_20221002104018_fd318a6d-97a2-44b6-9dbe-b373343bdbd9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664704288946_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
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 35.77 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: 36.451 seconds, Fetched: 10 row(s)
```

Optimized Hive query

1. To create table with partitioning & bucketing the below mentioned commands needs to be executed.

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

   Create table dyn_part_buck_ret with partition on (event_type) attribute and bucket on (price) attribute.

   create external table IF NOT EXISTS dyn_part_buck_ret (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' WITH SERDEPROPERTIES("separatorChar"=",", "quoteChar"="\"", "escapeChar"="\\") STORED AS TEXTFILE LOCATION '/user/case-study/' TBLPROPERTIES ("skip.header.line.count"="1");
   OK
   Time taken: 0.069 seconds

```
hive> create external table IF NOT EXISTS dyn_part_buck_ret (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' WITH SERDEPROPERTIES("separatorChar"=",
", "quoteChar"="\"", "escapeChar"="\\") STORED AS TEXTFILE LOCATION '/user/case-study/' TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.069 seconds
```

Optimized Hive query

2. Add data into the table ‘dyn_part_buck_ret’ from ret.

```
hive> insert into table dyn_part_buck_ret partition(event_type) select event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type from ret ;
Query ID = hadoop_20221004063837_2c51dc9a-8e6d-4e1e-8862-c735c4b3c7aa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664864134262_0005)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      2          2        0        0       0       0
Reduce 2 ......  container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 111.64 s
--------------------------------------------------------------------------------------------------
Loading data to table default.dyn_part_buck_ret partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.904 seconds
        Time taken for adding to write entity : 0.008 seconds
OK
Time taken: 119.477 seconds
```

Optimized Hive query

2. Add data into the table ‘dyn_part_buck_ret’ from ret.

```
hive> insert into table dyn_part_buck_ret partition(event_type) select event_time, product_id, category_id, category_code, brand,
price, user_id, user_session, event_type from ret ;
Query ID = hadoop_20221004063837_2c51dc9a-8e6d-4e1e-8862-c735c4b3c7aa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1664864134262_0005)

--------------------------------------------------------------------------------------------
        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: 111.64 s
--------------------------------------------------------------------------------------------
Loading data to table default.dyn_part_buck_ret partition (event_type=null)

Loaded : 4/4 partitions.
        Time taken to load dynamic partitions: 0.904 seconds
        Time taken for adding to write entity : 0.008 seconds
OK
Time taken: 119.477 seconds
```

Optimized Hive query

3. Running the optimized Hive query – Q8.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.

```
hive> SELECT user_id, SUM(price) AS Total_Expenditure from dyn_part_buck_ret WHERE event_type = 'purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10 ;
Query ID = hadoop_20221004071045_9a529626-b646-45b1-82d0-609b1ab8acb9
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_1664864134262_0007)

--------------------------------------------------------------------------------
        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: 17.16 s
--------------------------------------------------------------------------------
OK
user_id total_expenditure
557790271       2715.87
150318419       1645.97
562167663       1352.8500000000001
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3900000000008
561592095       1109.7000000000003
431950134       1097.5899999999997
566576008       1056.3600000000004
521347209       1040.9099999999996
Time taken: 26.944 seconds, Fetched: 10 row(s)
```

Optimized Hive query

3. Running the optimized Hive query – Q8.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.

```
hive> SELECT user_id, SUM(price) AS Total_Expenditure from dyn_part_buck_ret WHERE event_type = 'purchase' GROUP BY user_id ORDER BY Total_Expenditure DESC LIMIT 10 ;
Query ID = hadoop_20221004071045_9a529626-b646-45b1-82d0-609b1ab8acb9
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_1664864134262_0007)

--------------------------------------------------------------------------------------------------
        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: 17.16 s
--------------------------------------------------------------------------------------------------
OK
user_id total_expenditure
557790271       2715.87
150318419       1645.97
562167663       1352.8500000000001
531900924       1329.4500000000003
557850743       1295.4800000000002
522130011       1185.3900000000008
561592095       1109.7000000000003
431950134       1097.5899999999997
566576008       1056.3600000000004
521347209       1040.9099999999996
Time taken: 26.944 seconds, Fetched: 10 row(s)
```

Insights from the Optimized Hive Query

1. After creating an optimized table by partition on ‘event_type’ attribute and bucketing on ‘price’, we created the same query for Question 8 on this table. We found that the result is same as that was for the non-optimized table.
2. We also found that there is a significant drop in the execution time of the same query, previously it was 36.451 seconds and after it went down to 26.944 seconds with a difference of 9.507 seconds.
3. So we can conclude that with proper partitioning ad bucketing we can reduce the execution time of the query.

Terminating the EMR cluster

Amazon EMR

EMR Studio

EMR Serverless New

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Clone | Terminate | AWS CLI export

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

Cluster: hive_cp Terminating Terminated by user request

Summary | Application user interfaces | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Summary

ID: j-3MSPWGMWJPAY
Creation date: 2022-10-03 21:39 (UTC+5:30)
Elapsed time: 2 hours, 10 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-54-81-65-153.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, Hue 4.4.0, Spark 2.4.4
Log URI: s3://aws-logs-326076855193-us-east-