

HIVE CASE STUDY

By Deepthi Yalavarthi

KEY-PAIR CREATION:

Creating Key Pair with ppk file format directly to use with putty

The screenshot shows the 'Create key pair' wizard on the AWS EC2 console. The 'Name' field is set to 'HiveCaseStudy-KeyPair'. The 'File format' section has 'ppk' selected, which is described as 'For use with PuTTY'. There are no tags added. The 'Create key pair' button is at the bottom right.

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
HiveCaseStudy-KeyPair

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.

Cancel [Create key pair](#)

The screenshot shows the 'Key pairs' page on the AWS EC2 console. A green success message 'Successfully created key pair' is displayed. The table lists two key pairs: 'HiveCaseStudy-KeyPair' and 'Srinath_292848376449_2021-05-05...'. The first key pair is selected. The left sidebar shows the EC2 dashboard and various instance-related options.

Key pairs (1/2)

Name	Fingerprint	ID
HiveCaseStudy-KeyPair	3d:f6:c2:e8:ef:13:4f:b4:7d:7e:50:f6:b7:...	key-0969bfbe7145186c4
Srinath_292848376449_2021-05-05...	d4:dd:6f:0d:1cc:05:a:f6:4e:6d:23:00...	key-0fc8249ff897623c5

New EC2 Experience
Tell us what you think

EC2 Dashboard [New](#)

Events

Tags

Limits

Instances

- Instances [New](#)
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances [New](#)
- Dedicated Hosts
- Scheduled Instances
- Capacity Reservations

Images

Feedback English (US) ▾

© 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Cookie preferences](#)

HiveCaseStudy-Ke....ppk

HiveCaseStudy-KeyPair successfully created and downloaded (refer above screenshot)

S3 BUCKET:

To Store the data – Click on “Create Bucket”

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with 'Amazon S3' selected. Under 'Buckets', it lists 'Access Points', 'Object Lambda Access Points', 'Batch Operations', and 'Access analyzer for S3'. Below that is a section for 'Block Public Access settings for this account'. Under 'Storage Lens', it lists 'Dashboards' and 'AWS Organizations settings'. At the bottom of the sidebar is a 'Feature spotlight' section with a blue badge showing '3'. The main area is titled 'Amazon S3' and contains an 'Account snapshot' box with a link to 'View Storage Lens dashboard'. Below it is a 'Buckets (0)' section with a table header for 'Name', 'AWS Region', 'Access', and 'Creation date'. A message says 'No buckets' and 'You don't have any buckets.' with a 'Create bucket' button.

The screenshot shows the 'Create bucket' configuration page. The title is 'Create bucket' and it says 'Buckets are containers for data stored in S3. [Learn more](#)'. The 'General configuration' section has a 'Bucket name' field containing 'hivecasestudy-deepthi-srinath'. A note below it says 'Bucket name must be unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)'. It also has an 'AWS Region' dropdown set to 'US East (N. Virginia) us-east-1'. Below that is a section for 'Copy settings from existing bucket - optional' with a note 'Only the bucket settings in the following configuration are copied.' and a 'Choose bucket' button. The bottom of the page has a standard Windows taskbar with icons for File Explorer, Task View, Start, Edge, Google Chrome, WhatsApp, Microsoft Edge, and Word.

Creating “**hivecasestudy-deepthi-srinath**” with all default options

s3.console.aws.amazon.com/s3/bucket/create?region=us-east-1

No tags associated with this bucket.

Add tag

Default encryption
Automatically encrypt new objects stored in this bucket. [Learn more](#)

Server-side encryption
 Disable
 Enable

► Advanced settings

After creating the bucket you can upload files and folders to the bucket, and configure additional bucket settings.

Cancel **Create bucket**

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use Cookie preferences 10:18 AM 5/31/2021

Bucket Successfully created

← → ⌂ s3.console.aws.amazon.com/s3/home?region=us-east-1

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

Storage Lens

- Dashboards
- AWS Organizations settings

Feature spotlight 3

Successfully created bucket "hivecasestudy-deepthi-srinath"
To upload files and folders, or to configure additional bucket settings choose [View details](#).

View details X

Amazon S3

► Account snapshot

Storage lens provides visibility into storage usage and activity trends. [Learn more](#)

View Storage Lens dashboard

Buckets (1)

Buckets are containers for data stored in S3. [Learn more](#)

Find buckets by name

Name	AWS Region	Access	Creation date
hivecasestudy-deepthi-srinath	US East (N. Virginia) us-east-1	Bucket and objects not public	May 31, 2021, 10:19:53 (UTC+03:00)

Copy ARN Empty Delete Create bucket

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use Cookie preferences 10:20 AM 5/31/2021

Successfully uploaded the 2019 October and 2019 November csv files to S3 bucket

The screenshot shows the AWS S3 console interface. At the top, a green banner displays a success message: "Upload succeeded" with a link to "View details below". Below this, a summary table shows the destination as "s3://hivecasestudy-deepthi-srinath", with "Succeeded" status for "2 files, 980.7 MB (100.00%)" and "Failed" status for "0 files, 0 B (0%)". A navigation bar at the bottom includes tabs for "Files and folders" (selected) and "Configuration".

Destination	Succeeded	Failed
s3://hivecasestudy-deepthi-srinath	2 files, 980.7 MB (100.00%)	0 files, 0 B (0%)

Files and folders (2 Total, 980.7 MB)

Name	Folder	Type	Size	Status	Error
2019-Nov.csv	-	application/vnd.ms-excel	520.6 MB	✓ Succeeded	-
2019-Oct.csv	-	application/vnd.ms-excel	460.2 MB	✓ Succeeded	-

The screenshot shows the AWS S3 console interface. The left sidebar shows the path: "Amazon S3 > hivecasestudy-deepthi-srinath". The main area is titled "hivecasestudy-deepthi-srinath" and has a tab bar with "Objects" selected. Below the title, a section for "Objects (2)" is shown. It contains a table of objects with columns: Name, Type, Last modified, Size, and Storage class. The objects listed are "2019-Nov.csv" and "2019-Oct.csv", both of which are CSV files with sizes of 520.6 MB and 460.2 MB respectively, and are stored in the Standard storage class.

Name	Type	Last modified	Size	Storage class
2019-Nov.csv	csv	May 31, 2021, 10:22:04 (UTC+03:00)	520.6 MB	Standard
2019-Oct.csv	csv	May 31, 2021, 10:22:04 (UTC+03:00)	460.2 MB	Standard

EMR CLUSTER CREATION:

Click on “Create cluster” button to create the EMR cluster

The screenshot shows the Amazon EMR service page. On the left, there's a sidebar with links for EMR Studio, EMR on EC2, Clusters (which is selected and highlighted in orange), Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, Events, EMR on EKS, and Virtual clusters. Below that is a Help section with "What's new". The main content area has a title "Welcome to Amazon Elastic MapReduce". It includes a brief description of what EMR is, a message saying "You do not appear to have any clusters. Create one now:", and a prominent blue "Create cluster" button. To the right, there's an "Additional Information" section with links to "More about Elastic MapReduce", "EMR overview", "FAQs", "Pricing", and "More Help Using Elastic MapReduce" which lists "Forum", "Documentation", "Developer Guide", "API Reference", "EMR on GitHub", and "Help portal". At the bottom, there's a standard Windows taskbar with icons for File Explorer, Task View, Start, Taskbar settings, and a search bar.

Creating cluster with advanced options

The screenshot shows the "Create Cluster - Advanced Options" page. On the left, there's a sidebar with links for Step 1: Software and Steps (selected and highlighted in orange), Step 2: Hardware, Step 3: General Cluster Settings, and Step 4: Security. The main content area starts with a "Software Configuration" section where the "Release" dropdown is set to "emr-5.29.0". There are two columns of checkboxes for various software components: Hadoop 2.8.5, JupyterHub 1.0.0, Ganglia 3.7.2, Hive 2.3.6, MXNet 1.5.1, Hue 4.4.0, Spark 2.4.4 in the first column, and Zeppelin 0.8.2, Tez 0.9.2, HBase 1.4.10, Presto 0.227, Sqoop 1.4.7, Phoenix 4.14.3, HCatalog 2.3.6 in the second column. Some checkboxes are checked (e.g., Hadoop, Hive, Hue, Tez). Below this is a "Multiple master nodes (optional)" section with a checkbox for "Use multiple master nodes to improve cluster availability." The next section is "AWS Glue Data Catalog settings (optional)" with a checkbox for "Use for Hive table metadata". There's also an "Edit software settings" section with radio buttons for "Enter configuration" (selected) and "Load JSON from S3", and a text input field containing "classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]". At the bottom, there's a standard Windows taskbar with icons for File Explorer, Task View, Start, Taskbar settings, and a search bar.

Software and Steps page: Changed the Release from emr-5.33.0 to “**emr-5.29.0**”

Hardware page: Changing the Master and Core nodes from m5.xlarge to “m4.large”

The screenshot shows the 'Cluster Nodes and Instances' section of the AWS EMR console. It displays two node types: 'Master' and 'Core'. Both are currently configured with 'm4.large' instances. The 'Master' section shows 1 instance with 'On-demand' purchasing option selected. The 'Core' section also shows 1 instance with 'On-demand' selected. There are buttons to 'Add configuration settings' and 'Add task instance group'.

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	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
Core Core - 2	m4.large 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price

+ Add task instance group

Total core and task units: 1 Total units

Cluster scaling

Adjust the number of Amazon EC2 instances available to an EMR cluster via EMR-managed scaling or a custom automatic scaling policy. [Learn](#)

General Cluster Settings page: Giving the name to cluster “HiveCaseStudy”

The screenshot shows the 'Create Cluster - Advanced Options' page. The 'Step 3: General Cluster Settings' tab is active. The 'Cluster name' field is set to 'HiveCaseStudy'. Under 'General Options', there are checkboxes for 'Logging' (selected), 'S3 folder s3://aws-logs-292848376449-us-east-1/elasticmapre...'), 'Debugging', and 'Termination protection'. The 'Tags' section has a placeholder 'Add a key to create a tag'. The 'Additional Options' section includes a checkbox for 'EMRFS consistent view' and a dropdown for 'Custom AMI ID' set to 'None'. The 'Bootstrap Actions' section is partially visible at the bottom.

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Create Cluster - Advanced Options [Go to quick options](#)

General Options

Cluster name: HiveCaseStudy

Logging [i](#)
S3 folder s3://aws-logs-292848376449-us-east-1/elasticmapre...
 Debugging [i](#)
 Termination protection [i](#)

Tags [i](#)

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

Additional Options

EMRFS consistent view [i](#)

Custom AMI ID: None

Bootstrap Actions

Security page: changing the EC2 key pair option to our created key pair - "HiveCaseStudy-KeyPair"

The screenshot shows the 'Create Cluster - Advanced Options' page. On the left, a sidebar lists steps: Step 1: Software and Steps, Step 2: Hardware, Step 3: General Cluster Settings, and Step 4: Security (which is selected). The main area is titled 'Security Options'. It shows the 'EC2 key pair' dropdown set to 'HiveCaseStudy-KeyPair'. Below it, there's a checkbox for 'Cluster visible to all IAM users in account' which is checked. Under 'Permissions', there are three tabs: 'Default' (selected), 'Custom', and 'IAM roles'. The 'Default' tab has a note about automatically creating roles if none are present. Below these are sections for 'EMR role', 'EC2 instance profile', and 'Auto Scaling role', each with a dropdown menu. At the bottom right are 'Cancel', 'Previous', and 'Create cluster' buttons.



Click on “Create Cluster” button

The screenshot shows the 'Cluster: HiveCaseStudy Starting' page. On the left, a sidebar for 'Amazon EMR' lists 'EMR Studio', 'EMR on EC2' (with 'Clusters' selected), 'Notebooks', 'Git repositories', 'Security configurations', 'Block public access', 'VPC subnets', 'Events', and 'EMR on EKS' (with 'Virtual clusters' listed). The main area shows the cluster status as 'Starting'. It has tabs for 'Summary', 'Application user interfaces', 'Monitoring', 'Hardware', 'Configurations', 'Events', 'Steps', and 'Bootstrap actions'. The 'Summary' tab is active, displaying information like ID: j-2KPLYT0TEQYJB, Creation date: 2021-05-31 11:04 (UTC+3), Elapsed time: 1 second, and termination protection set to 'On'. The 'Configuration details' tab shows 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-292848376449-us-east-1/elasticmapreduce/...', and EMRFS consistent view set to 'Disabled'. The 'Application user interfaces' tab shows persistent user interfaces. At the bottom, there are 'Feedback', language 'English (US)', and other navigation links.

Cluster is ready with status “Waiting”

EMR – AWS Console

console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details;j-2KPLYT0TEQYJB

Services ▾

Search for services, features, marketplace products, and docs [Alt+S]

upgradsrinathdron @ 2928-4837-6449 N. Virginia Support

Clone Terminate AWS CLI export

Cluster: HiveCaseStudy Waiting Cluster ready after last step completed.

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-2KPLYT0TEQYJB
Creation date: 2021-05-31 11:04 (UTC+3)
Elapsed time: 50 minutes
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: ec2-34-207-251-55.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-292848376449-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use Cookie preferences

Type here to search

12:08 PM 5/31/2021

EMR – AWS Console

console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details;j-2KPLYT0TEQYJB

Services ▾

Search for services, features, marketplace products, and docs [Alt+S]

upgradsrinathdron @ 2928-4837-6449 N. Virginia Support

Master public DNS: ec2-34-207-251-55.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-292848376449-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

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

Network and hardware

Availability zone: us-east-1
Subnet ID: subnet-0044494d
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: HiveCaseStudy-KeyPair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0252bb3a17e45518 (ElasticMapReduce-master)
Security groups for Core & sg-0b1b25cd0be723751 (ElasticMapReduce-Task: slave)

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use Cookie preferences

Type here to search

12:09 PM 5/31/2021

Both the Master and Core nodes are running

EMR – AWS Console

EC2 Management Console

console.aws.amazon.com/ec2/home?region=us-east-1#ModifyInboundSecurityGroupRules:securityGroupId=sg-0252bbb3a17e45518

All ICMP - IPv4

ICMP

All

Custom

sg-0252bbb3a17e45518

sg-0b1b25cd0be723751

SSH

TCP

22

Anywhere

0.0.0.0/0

::/0

Add rule

⚠ NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Cancel

Preview changes

Save rules

Then save the SSH rule to the inbound rules

CONNECT TO MASTER NODE:

Open the putty and enter the Host Name as hadoop@ec2-34-207-251-55.compute-1.amazonaws.com and navigate to Connection > SSH > Auth then browse and select the private key, which we creating initially.

EMR – AWS Console

Services

Search for services, features, marketplace products, and docs

Feedback English (US)

Clusters

Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Clone

Terminate

AWS CLI export

Cluster Details Open Cluster Details

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine logs, and more.

Windows Mac / Linux

1. Download PuTTY.exe to your computer from: <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>

2. Start PuTTY.

3. In the Category list, click Session.

4. In the Host Name field, type hadoop@ec2-34-207-251-55.compute-1.amazonaws.com

5. In the Category list, expand Connection > SSH, and then click Auth.

6. For Private key file for authentication, click Browse and select the private key file (HiveCases.pem) cluster.

7. Click Open.

8. Click Yes to dismiss the security alert.

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

PuTTY Configuration

Category: Session

Basic options for your PuTTY session

Host Name (or IP address): hadoop@ec2-34-207-251-55.compute-1.amazonaws.com

Port: 22

Connection type: SSH

Window

Appearance

Behaviour

Translation

Selection

Colours

Connection

Data

Proxy

SSH

Serial

Telnet

>Login

SUPDUP

Load, save or delete a stored session

Saved Sessions

Default Settings

Load

Save

Delete

About

Help

Open

Cancel

Close window on exit:

Always

Never

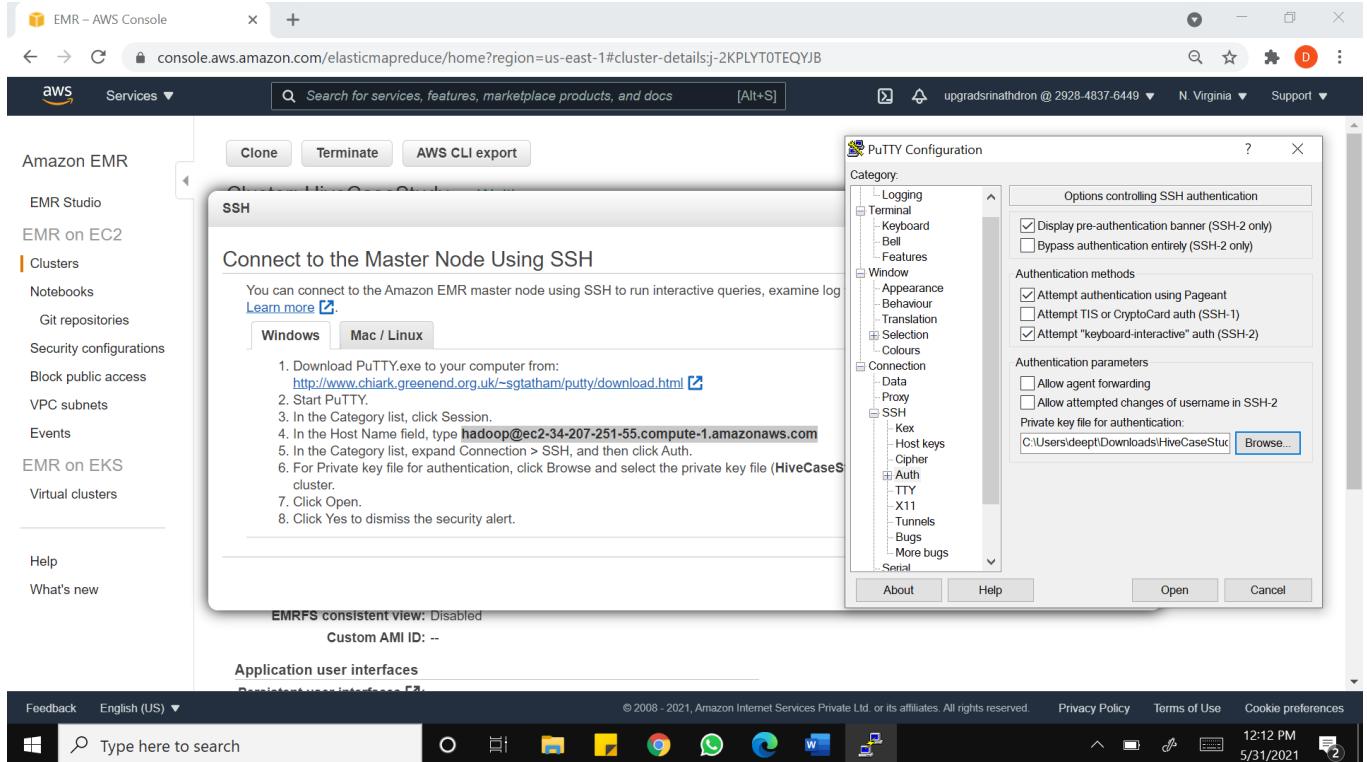
Only on clean exit

Feedback English (US)

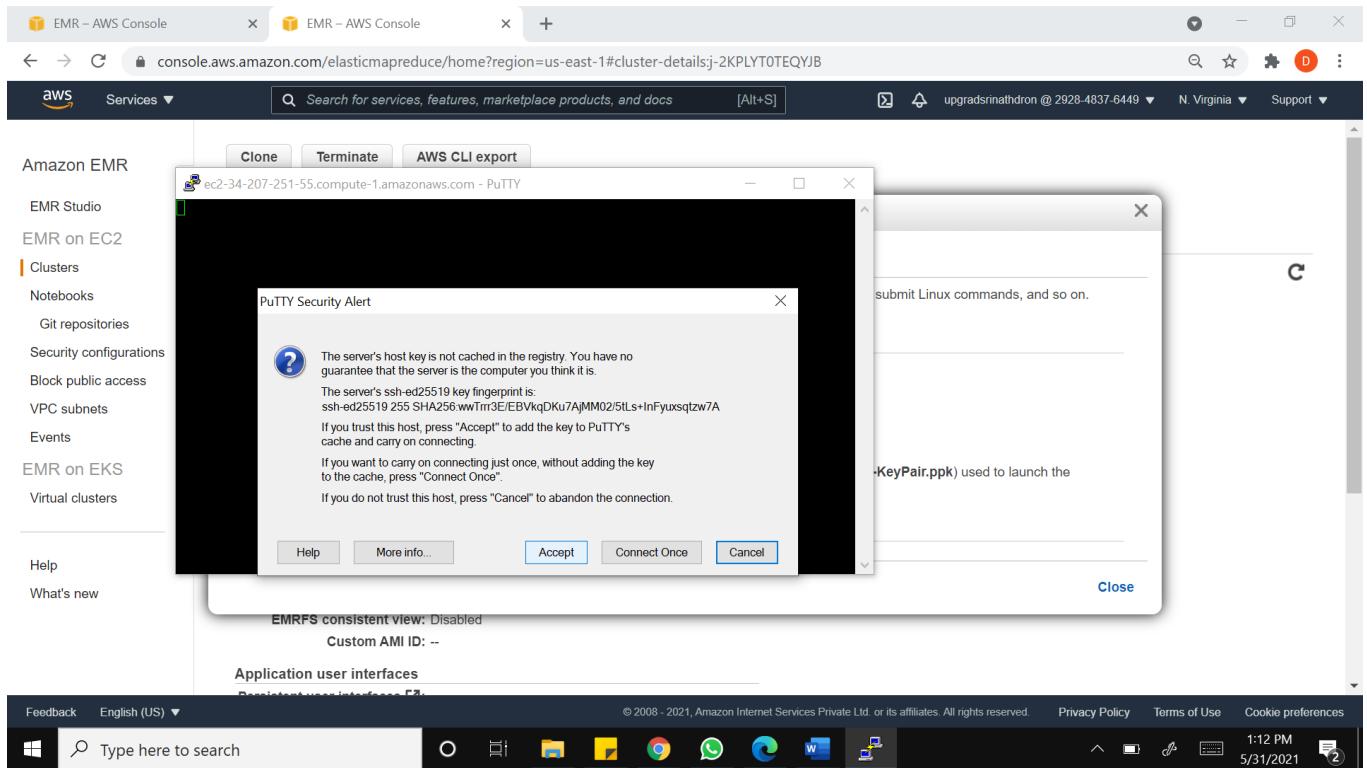
Privacy Policy Terms of Use Cookie preferences

Type here to search

12:10 PM 5/31/2021



Click on “open” and then Accept the connection



EMR CLI is launched

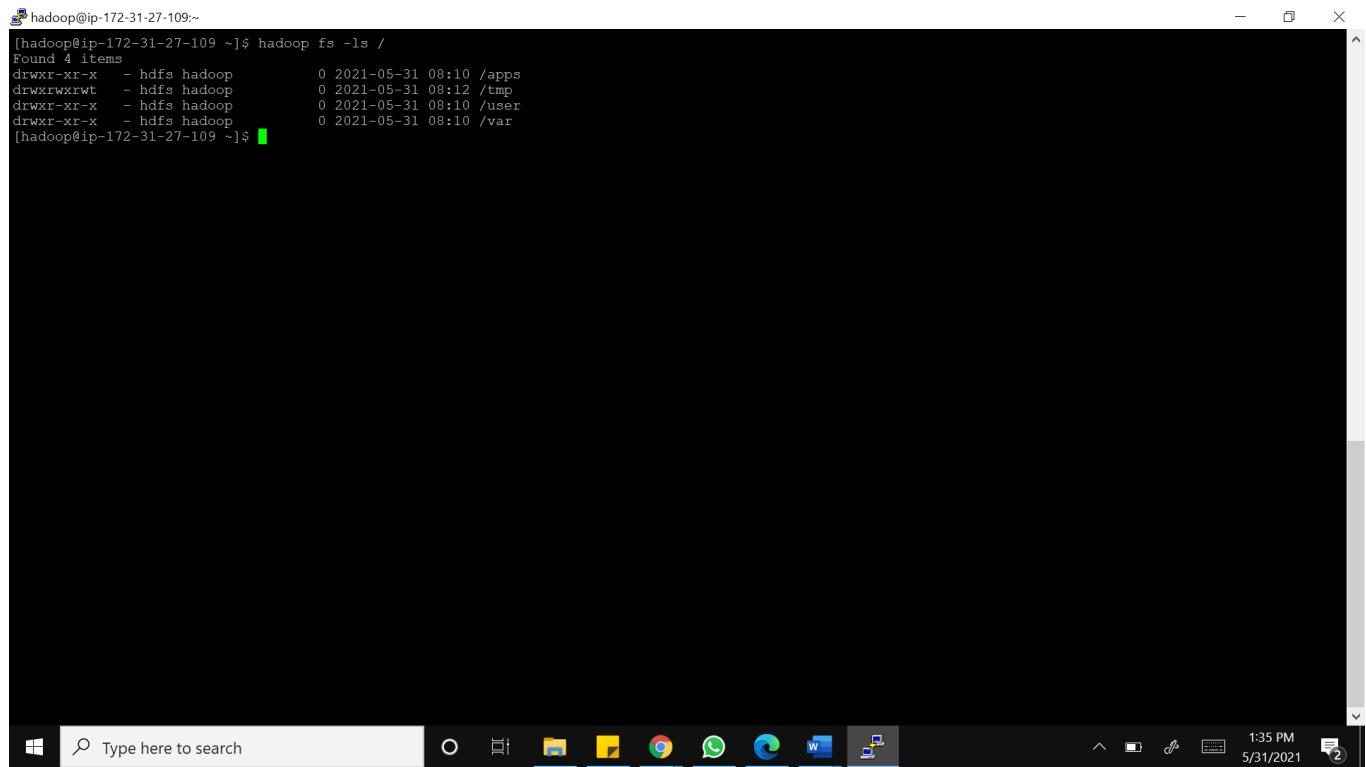
Verifying the services that are running on Hadoop cluster with command “**`sudo initctl list`**”

```
E:::::E     EEEEEE M:::::M    MMM    M:::::M R::::R    R::::R
EE:::::EEEEEEEEE::::E M:::::M    M:::::M R::::R    R::::R
E::::::::::::E M:::::M    M:::::M RR:::::R    R:::::R
EEEEEEEEEEEEEEEEEEE MMMMMRM    MMMMMRM RRRRRR    RRRRRR

[hadoop@ip-172-31-27-109 ~]$ sudo initctl list
rc stop/waiting
tty (/dev/tty3) start/running, process 4926
tty (/dev/tty2) start/running, process 4924
tty (/dev/tty1) start/running, process 4920
tty (/dev/tty6) start/running, process 4939
tty (/dev/tty5) start/running, process 4937
tty (/dev/tty4) start/running, process 4935
update-moto stop/waiting
hive-server2 start/running, process 14408
hadoop-mapreduce-hiveserver start/running, process 12780
hadoop-yarn-timelineserver start/running, process 12087
plymouth-shutdown stop/waiting
whisper-server stop/waiting
control-alt-delete stop/waiting
hive-hcatalog-server start/running, process 15168
rcS-emergency stop/waiting
kexec-disable stop/waiting
quit-plymouth stop/waiting
hadoop-kms start/running, process 7548
hadoop-yarn-resourcemanager start/running, process 12196
rcS stop/waiting
hue start/running, process 15434
prefdm stop/waiting
hadoop-httpfs start/running, process 7832
init-system-dbus stop/waiting
print-image-id stop/waiting
oozie start/running, process 16176
elastic-network-interfaces stop/waiting
splash-manager stop/waiting
hadoop-yarn-proxyserver start/running, process 11988
start-ttys stop/waiting
ranger-kms-server stop/waiting
amazon-ssm-agent start/running, process 3860
hadoop-hdfs-namenode start/running, process 7979
rcS-sulogin stop/waiting
scout-server stop/waiting
serial (ttyS0) start/running, process 4918
[hadoop@ip-172-31-27-109 ~]$
```

We can see that Hive services are running

Verifying the Hadoop file system with command “**hadoop fs -ls /**”



```
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -ls /
Found 4 items
drwxr-xr-x  - hdfs hadoop      0 2021-05-31 08:10 /apps
drwxrwxrwt  - hdfs hadoop      0 2021-05-31 08:12 /tmp
drwxr-xr-x  - hdfs hadoop      0 2021-05-31 08:10 /user
drwxr-xr-x  - hdfs hadoop      0 2021-05-31 08:10 /var
[hadoop@ip-172-31-27-109 ~]$
```

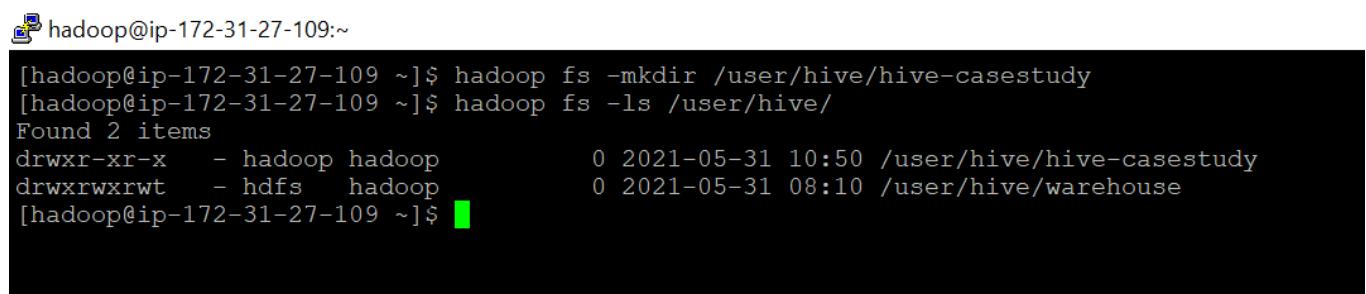
All the above are inbuilt directories in HDFS.

CREATING A NEW DIRECTORY FOR HIVE CASE STUDY:

Creating a new directory under user>hive for Hive case study to store the data files and directory name creating is “**hive-casestudy**” and verifying whether the new directory is listed in Hadoop file system>user>hive

```
hadoop fs -mkdir /user/hive/hive-casestudy
```

```
hadoop fs -ls /user/hive/
```

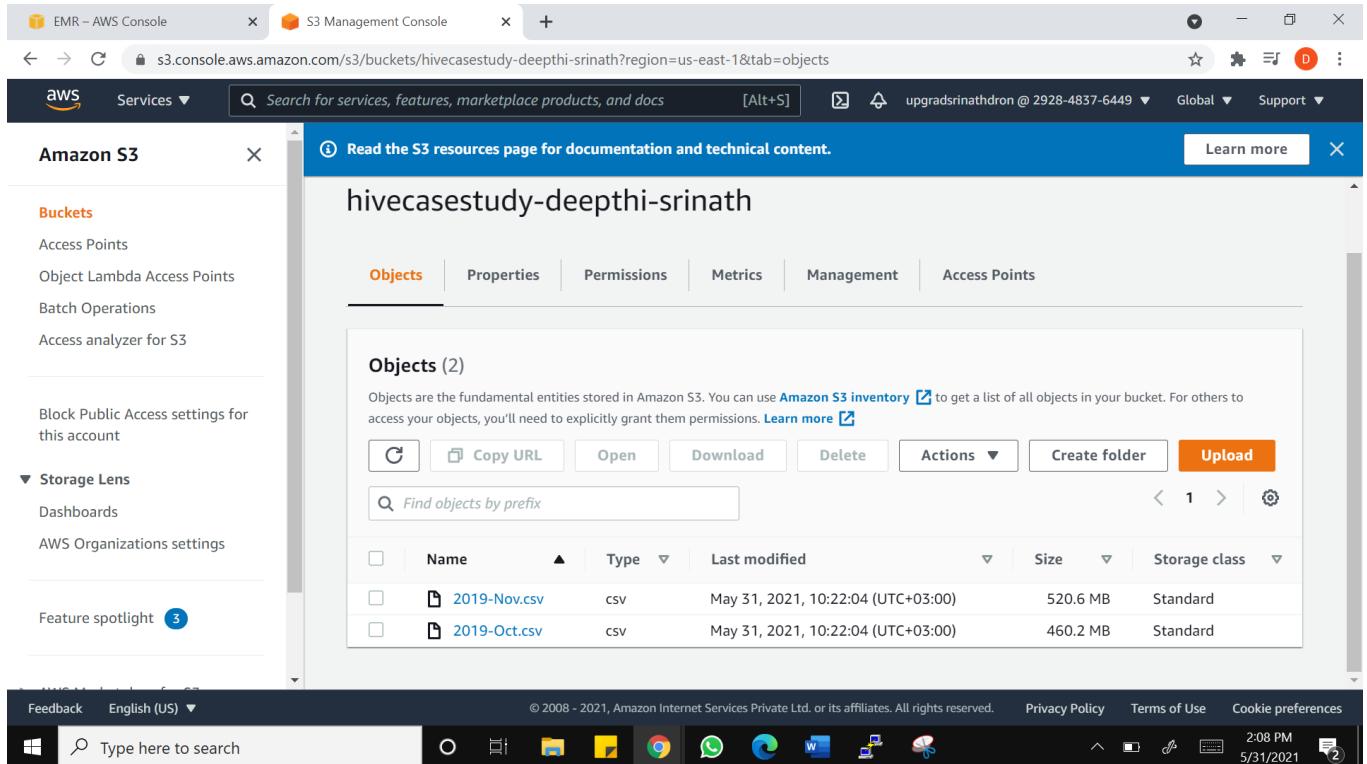


```
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -mkdir /user/hive/hive-casestudy
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -ls /user/hive/
Found 2 items
drwxr-xr-x  - hadoop hadoop      0 2021-05-31 10:50 /user/hive/hive-casestudy
drwxrwxrwt  - hdfs   hadoop      0 2021-05-31 08:10 /user/hive/warehouse
[hadoop@ip-172-31-27-109 ~]$
```

New directory is successfully created

LOADING THE DATA FROM S3 BUCKET to HDFS:

Copying the file path from S3



The screenshot shows the AWS S3 Management Console interface. The left sidebar has sections for Buckets, Access Points, Object Lambda Access Points, Batch Operations, and Access analyzer for S3. Under Storage Lens, there are Dashboards and AWS Organizations settings. A Feature spotlight section is also present. The main area displays the 'hivecasestudy-deepthi-srinath' bucket. The 'Objects' tab is selected, showing two objects: '2019-Nov.csv' and '2019-Oct.csv'. Both are CSV files, last modified on May 31, 2021, at 10:22:04 UTC+03:00, with sizes of 520.6 MB and 460.2 MB respectively, and stored in Standard storage class. Action buttons include Copy URL, Open, Download, Delete, Actions, Create folder, and Upload. A search bar at the top right says 'Search for services, features, marketplace products, and docs'.

Distributed copy command is using to copy the data from S3 to HDFS –

For 2019 October:

hadoop distcp s3n://hivecasestudy-deepthi-srinath/2019-Oct.csv /user/hive/hive-casestudy/2019-Oct.csv

For 2019 November:

hadoop distcp s3n://hivecasestudy-deepthi-srinath/2019-Nov.csv /user/hive/hive-casestudy/2019-Nov.csv

Below are the screenshots for copying October 2019 and November 2019 data individually

October 2019:

```
[hadoop@ip-172-31-27-109 ~]$ hadoop distcp s3n://hivecasestudy-deepthi-srinath/2019-Oct.csv /user/hive/hive-casestudy/2019-Oct.csv
21/05/31 11:20:23 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListStatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preservesStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://hivecasestudy-deepthi-srinath/2019-Oct.csv], targetPath=/user/hive/hive-casestudy/2019-Oct.csv, targetPathExists=false, filtersFile='null'}
21/05/31 11:20:24 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-109.ec2.internal/172.31.27.109:8032
21/05/31 11:20:28 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/05/31 11:20:28 INFO tools.SimpleCopyListing: Build file listing completed.
21/05/31 11:20:28 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/05/31 11:20:28 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/05/31 11:20:28 INFO tools.DistCp: Number of paths in the copy list: 1
21/05/31 11:20:28 INFO tools.DistCp: Number of paths in the copy list: 1
21/05/31 11:20:28 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-109.ec2.internal/172.31.27.109:8032
21/05/31 11:20:28 INFO mapreduce.JobSubmitter: number of splits:1
21/05/31 11:20:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622448704713_0001
21/05/31 11:20:29 INFO impl.YarnClientImpl: Submitted application application_1622448704713_0001
21/05/31 11:20:30 INFO mapreduce.Job: The url to track the job: http://ip-172-31-27-109.ec2.internal:20888/proxy/application_1622448704713_0001/
21/05/31 11:20:30 INFO tools.DistCp: DistCp job-id: job_1622448704713_0001
21/05/31 11:20:30 INFO mapreduce.Job: Running job: job_1622448704713_0001
21/05/31 11:20:40 INFO mapreduce.Job: Job job_1622448704713_0001 running in uber mode : false
21/05/31 11:20:40 INFO mapreduce.Job: map 0% reduce 0%
21/05/31 11:20:57 INFO mapreduce.Job: map 100% reduce 0%
21/05/31 11:20:58 INFO mapreduce.Job: Job job_1622448704713_0001 completed successfully
21/05/31 11:20:58 INFO mapreduce.Job: Counters: 38
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=172509
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=369
    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
    S3N: Number of bytes read=482542278
    S3N: Number of bytes written=0
    S3N: Number of read operations=0
    S3N: Number of large read operations=0
    S3N: 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)=511072
[Windows Taskbar] Type here to search [Icons] 2:21 PM 5/31/2021
```

```
[hadoop@ip-172-31-27-109 ~]
  FILE: Number of bytes read=0
  FILE: Number of bytes written=172509
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=369
  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
  S3N: Number of bytes read=482542278
  S3N: Number of bytes written=0
  S3N: Number of read operations=0
  S3N: Number of large read operations=0
  S3N: 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)=511072
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=15971
  Total vcore-milliseconds taken by all map tasks=15971
  Total megabyte-milliseconds taken by all map tasks=16354304
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)=267
  CPU time spent (ms)=18710
  Physical memory (bytes) snapshot=570552320
  Virtual memory (bytes) snapshot=3296665600
  Total committed heap usage (bytes)=504365056
File Input Format Counters
  Bytes Read=233
File Output Format Counters
  Bytes Written=0
DistCp Counters
  Bytes Copied=482542278
  Bytes Expected=482542278
  Files Copied=1
[hadoop@ip-172-31-27-109 ~]$
```

November 2019

```
[hadoop@ip-172-31-27-109:~]$ hadoop distcp s3n://hivecasestudy-deepthi-srinath/2019-Nov.csv /user/hive/hive-casestudy/2019-Nov.csv
21/05/31 11:25:55 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, skipCRC=false, blocking=true, numListStatusThreads=0, maxMaps=20, mapBandwidth=100, sslConfigurationFile='null', copyStrategy='uniformsize', preserveStatus=[], preserveRawXattrs=false, atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3n://hivecasestudy-deepthi-srinath/2019-Nov.csv], targetPath=/user/hive/hive-casestudy/2019-Nov.csv, targetPathExists=false, filtersFile='null'}
21/05/31 11:25:55 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-109.ec2.internal/172.31.27.109:8032
21/05/31 11:25:59 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
21/05/31 11:25:59 INFO tools.SimpleCopyListing: Build file listing completed.
21/05/31 11:25:59 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
21/05/31 11:25:59 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
21/05/31 11:26:00 INFO tools.DistCp: Number of paths in the copy list: 1
21/05/31 11:26:00 INFO tools.DistCp: Number of paths in the copy list: 1
21/05/31 11:26:00 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-27-109.ec2.internal/172.31.27.109:8032
21/05/31 11:26:00 INFO mapreduce.JobSubmission: number of splits:1
21/05/31 11:26:00 INFO mapreduce.JobSubmission: Submitting tokens for job: job_1622448704713_0002
21/05/31 11:26:00 INFO impl.YarnClientImpl: Submitted application application_1622448704713_0002
21/05/31 11:26:00 INFO mapreduce.Job: The url to track the job: http://ip-172-31-27-109.ec2.internal:20888/proxy/application_1622448704713_0002/
21/05/31 11:26:00 INFO tools.DistCp: DistCp job-id: job_1622448704713_0002
21/05/31 11:26:00 INFO mapreduce.Job: Running job: job_1622448704713_0002
21/05/31 11:26:09 INFO mapreduce.Job: Job job_1622448704713_0002 running in uber mode : false
21/05/31 11:26:09 INFO mapreduce.Job: map 0% reduce 0%
21/05/31 11:26:27 INFO mapreduce.Job: map 100% reduce 0%
21/05/31 11:26:29 INFO mapreduce.Job: Job job_1622448704713_0002 completed successfully
21/05/31 11:26:29 INFO mapreduce.Job: Counters: 38
    File System Counters
        FILE: Number of bytes read=0
        FILE: Number of bytes written=172509
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=369
        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
        S3N: Number of bytes read=545839412
        S3N: Number of bytes written=0
        S3N: Number of read operations=0
        S3N: Number of large read operations=0
        S3N: 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)=558624
```

```
hadoop@ip-172-31-27-109:~$ 
FILE: Number of bytes read=0
FILE: Number of bytes written=172509
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=369
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
S3N: Number of bytes read=545839412
S3N: Number of bytes written=0
S3N: Number of read operations=0
S3N: Number of large read operations=0
S3N: 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)=558624
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=17457
Total voore-milliseconds taken by all map tasks=17457
Total megabyte-milliseconds taken by all map tasks=17875968
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)=357
CPU time spent (ms)=20140
Physical memory (bytes) snapshot=520134656
Virtual memory (bytes) snapshot=3301638144
Total committed heap usage (bytes)=438304768
File Input Format Counters
Bytes Read=233
File Output Format Counters
Bytes Written=0
DistCp Counters
Bytes Copied=545839412
Bytes Expected=545839412
Files Copied=1
[hadoop@ip-172-31-27-109 ~]$ 
```

Verifying whether the data is successfully copied into HDFS from S3 buckets

Command: **hadoop fs -ls /user/hive/hive-casestudy**

```
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -ls /user/hive/hive-casestudy
Found 2 items
-rw-r--r-- 1 hadoop hadoop 545839412 2021-05-31 11:26 /user/hive/hive-casestudy/2019-Nov.csv
-rw-r--r-- 1 hadoop hadoop 482542278 2021-05-31 11:20 /user/hive/hive-casestudy/2019-Oct.csv
[hadoop@ip-172-31-27-109 ~]$
```

Inspecting the table data to know which columns are available before creating the hive table with command “**hadoop fs -cat /user/hive/hive-casestudy/2019-Oct.csv |head**” and “**hadoop fs -cat /user/hive/hive-casestudy/2019-Nov.csv |head**”

```
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -cat /user/hive/hive-casestudy/2019-Oct.csv |head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-10-01 00:00:00 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC,cart,5773203,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC,cart,5881589,2151191071051219817,,lovely,13.48,429681830,49e8a843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC,cart,5723490,1487580005134238553,,runail,2.62,463240011,26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC,cart,5881449,14875800013522845895,,lovely,0.56,429681830,49e8a843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC,cart,5857269,1487580005134238553,,runail,2.62,430174032,73deale7-664e-43f4-8b30-d32b9d5af04f
2019-10-01 00:00:19 UTC,cart,5739055,1487580008246412266,,kapous,4.75,377667011,81326ac6-daa4-4f0a-b488-fd0956a78733
2019-10-01 00:00:24 UTC,cart,5825598,1487580009445982239,,,0.56,467916806,2f5b5546-b8cb-9ee7-7ecd-84276f8ef486
2019-10-01 00:00:25 UTC,cart,5698989,1487580006317032337,,,1.27,385985999,d30965e8-1101-44ab-b45d-cc1bb9fae694
cat: Unable to write to output stream.
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -cat /user/hive/hive-casestudy/2019-Nov.csv |head
event_time,event_type,product_id,category_id,category_code,brand,price,user_id,user_session
2019-11-01 00:00:02 UTC,view,5802432,1487580009286598681,,,0.32,562076640,09fafdfc-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC,cart,5844397,1487580006317032337,,,2.38,553329724,2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:10 UTC,view,5837166,1783999064103190764,,pnb,22.22,556138645,57ed222e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC,cart,5876812,14875800010100293687,,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-alcc-af0575a34ffb
2019-11-01 00:00:24 UTC,remove_from_cart,5826182,1487580007483048900,,,3.33,553329724,2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:25 UTC,view,5856189,1487580009026551821,,runail,15.71,562076640,09fafdfc-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:32 UTC,view,5837835,1933472286753424063,,,3.49,514649199,432a4e95-375c-4b40-bd36-0fc039e77580
2019-11-01 00:00:34 UTC,remove_from_cart,5870838,1487580007675986893,,milv,0.79,429913900,2f0bff3c-252f-4fe6-afcd-5d8a6a92839a
cat: Unable to write to output stream.
[hadoop@ip-172-31-27-109 ~]$
```

Both the tables are having same columns of data

Moving to hive:

```
[hadoop@ip-172-31-27-109 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

CREATING AN EXTERNAL TABLE IN HIVE:

```
CREATE EXTERNAL TABLE IF NOT EXISTS retailsstore (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/user/hive/hive-casestudy' tblproperties("skip.header.line.count"="1");
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS retailsstore (event_time timestamp, event_type string, product_id string, category_id string, category_code string, brand string, price float, user_id bigint, user_session string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS textfile LOCATION '/user/hive/hive-casestudy' tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.085 seconds
```

Below command is used to set the display the header columns

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

APPLYING OPTIMIZATION TECHNIQUES - PARTITIONING AND BUCKETING:

Below commands are to enable the dynamic partitioning and bucketing

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

```
hive> set hive.exec.dynamic.partition = true;
```

```
hive> 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> █
```

Creating an optimized table by applying partitioning on “event_type” and bucketing on “price”

```
CREATE TABLE IF NOT EXISTS dynpart_buck_retailsstore(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 10 BUCKETS ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/user/hive/hive-casestudy' tblproperties('skip.header.line.count' = '1');
```

```

hive> CREATE TABLE IF NOT EXISTS dynpart_buck_retailsstore(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 10 BUCKETS
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS TEXTFILE
    > LOCATION '/user/hive/hive-casestudy'
    > tblproperties('skip.header.line.count' = '1');
OK
Time taken: 0.061 seconds
hive>

```

Verifying the created table

```

hive> show tables;
OK
tab_name
dynpart_buck_retailsstore
retailsstore
Time taken: 0.021 seconds, Fetched: 2 row(s)
hive>

```

INSERTING THE DATA INTO NEWLY CREATED OPTIMIZED TABLE (dynpart buck retailsstore) FROM EXISTING TABLE(retailsstore):

INSERT INTO TABLE dynpart_buck_retailsstore PARTITION (event_type) SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type FROM retailsstore;

```

hive> INSERT INTO TABLE dynpart_buck_retailsstore
    > PARTITION (event_type)
    > SELECT event_time, product_id, category_id, category_code, brand, price, user_id, user_session, event_type
    > FROM retailsstore;
Query ID = hadoop_20210531133455_970f6238-6993-4be5-a751-93e4bad61d78
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_1622448704713_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: 168.08 s
-----
Loading data to table default.dynpart_buck_retailsstore partition (event_type=null)
Loaded : 4/4 partitions.
          Time taken to load dynamic partitions: 0.41 seconds
          Time taken for adding to write entity : 0.007 seconds
OK
event_time      product_id      category_id      category_code      brand      price      user_id user_session      event_type
Time taken: 177.824 seconds
hive>

```

Output: Based on the above results, it partitioned into 4

Verifying the partitioned in the Hadoop file system

Randomly verifying the partitioned data in hadoop

```
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -ls /user/hive/hive-casestudy/event_type=cart
Found 10 items
-rwxr-xr-x  1 hadoop hadoop  27512579 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=cart/000000_0
-rwxr-xr-x  1 hadoop hadoop  32190447 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=cart/000001_0
-rwxr-xr-x  1 hadoop hadoop  33302805 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=cart/000002_0
-rwxr-xr-x  1 hadoop hadoop  32602023 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=cart/000003_0
-rwxr-xr-x  1 hadoop hadoop  34104132 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=cart/000004_0
-rwxr-xr-x  1 hadoop hadoop  325383513 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=cart/000005_0
-rwxr-xr-x  1 hadoop hadoop  39257340 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=cart/000006_0
-rwxr-xr-x  1 hadoop hadoop  24825787 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=cart/000007_0
-rwxr-xr-x  1 hadoop hadoop  28504487 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=cart/000008_0
-rwxr-xr-x  1 hadoop hadoop  35410315 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=cart/000009_0
[hadoop@ip-172-31-27-109 ~]$
```

```
[hadoop@ip-172-31-27-109 ~]$ hadoop fs -ls /user/hive/hive-casestudy/event_type=purchase
Found 10 items
-rw-r-xr-x  1 hadoop hadoop  6241877 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=purchase/00000_0
-rw-r-xr-x  1 hadoop hadoop  7235640 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=purchase/00001_0
-rw-r-xr-x  1 hadoop hadoop  7231471 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=purchase/00002_0
-rw-r-xr-x  1 hadoop hadoop  7526313 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=purchase/00003_0
-rw-r-xr-x  1 hadoop hadoop  7227979 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=purchase/00004_0
-rw-r-xr-x  1 hadoop hadoop  7310389 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=purchase/00005_0
-rw-r-xr-x  1 hadoop hadoop  8915123 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=purchase/00006_0
-rw-r-xr-x  1 hadoop hadoop  5366094 2021-05-31 13:37 /user/hive/hive-casestudy/event_type=purchase/00007_0
-rw-r-xr-x  1 hadoop hadoop  6469070 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=purchase/00008_0
-rw-r-xr-x  1 hadoop hadoop  8004214 2021-05-31 13:36 /user/hive/hive-casestudy/event_type=purchase/00009_0
[hadoop@ip-172-31-27-109 ~]$
```

VERIFYING THE PERFORMANCE OF BOTH THE TABLES – BEFORE AND AFTER OPTIMIZED TECHNIQUES:

select * from retailsstore limit 5;

```
hive> set hive.cli.print.header = true;
hive> select * from retailsstore limit 5;
OK
retailsstore.event_time    retailsstore.event_type    retailsstore.product_id    retailsstore.category_id    retailsstore.category_code    retailsstore.brand    r
retailsstore.price    retailsstore.user_id    retailsstore.user_session
2019-11-01 00:00:02 UTC view    5802432 1487580009286598681    0.32    562076640    09fafd6c-6c99-46b1-834f-33527f4de241
2019-11-01 00:00:09 UTC cart    5844397 1487580006317032337    2.38    553329724    2067216c-31b5-455d-alcc-af0575a34ffb
2019-11-01 00:00:10 UTC view    5837166 1783999064103190764    pnb    22.22    556138645    57ed22e-a54a-4907-9944-5a875c2d7f4f
2019-11-01 00:00:11 UTC cart    5876812 14875800010100293687    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-alcc-af0575a34ffb
Time taken: 0.212 seconds, Fetched: 5 row(s)
hive> [REDACTED]
```

Time taken to retrieve first 5 rows of data **before optimization** is **0.212 seconds (above)**

select * from dynpart_buck_retailsstore limit 5;

```
hive> set hive.cli.print.header = true;
hive> select * from dynpart_buck_retailsstore limit 5;
OK
dynpart_buck_retailsstore.event_time    dynpart_buck_retailsstore.product_id    dynpart_buck_retailsstore.category_id    dynpart_buck_retailsstore.category_code    dynpart_buck_retailsstore.brand    dynpart_buck_retailsstore.price    dynpart_buck_retailsstore.user_id    dynpart_buck_retailsstore.user_session    dynpart_buck_retailsstore.event_type
2019-10-08 09:19:19 UTC 89350 1487580011652186237    runail    1.27    232701853    3f1469f5-d926-44ce-a9f6-dff5ae276c9c    cart
2019-10-10 05:29:47 UTC 5866208 1487580013841613016    concept    3.16    493381333    535bb657-08f4-4021-ac66-b340178f7a37    cart
2019-10-08 12:25:50 UTC 5821183 1487580007717929935    1.27    546703849    3daf4d64-5ffa-46cc-827b-59760eb819b    cart
2019-10-10 08:19:06 UTC 5848901 1487580007675986893    bpw.style    1.27    439370683    9aeb4d9a-1bed-4f42-b12d-88be1148d3a9    cart
2019-10-09 18:32:50 UTC 5869152 1487580005268456287    cosmoprofi    7.94    558533352    cfde0f74-8705-4a2f-ba83-a5b99581c294    cart
Time taken: 0.185 seconds, Fetched: 5 row(s)
hive> [REDACTED]
```

Time taken to retrieve the first 5 rows of data **after optimization** is **0.185 seconds (above screenshot)**

ANSWERING GIVEN QUESTIONS:

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

Base table:

SELECT SUM(price) AS tot_revenue_oct FROM retailsstore WHERE MONTH(event_time) = '10' AND event_type = 'purchase';

```
hive> SELECT SUM(price) AS tot_revenue_oct FROM retailsstore WHERE MONTH(event_time) = '10' AND event_type = 'purchase';
Query ID = hadoop_20210531123157_a77aa0ec-fc06-4a02-a121-1222590c3741
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622448704713_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: 69.72 s  

-----  

OK
tot_revenue_oct
1211538.4299997438
Time taken: 70.448 seconds, Fetched: 1 row(s)
hive> [REDACTED]
```

Time taken is **70.448 seconds**

Optimized table:

```
SELECT SUM(price) AS tot_revenue_oct FROM dynpart_buck_retailsstore WHERE MONTH(event_time) = 10 AND event_type = 'purchase';
```

```
hive> SELECT SUM(price) AS tot_revenue_oct FROM dynpart_buck_retailsstore WHERE MONTH(event_time) = 10 AND event_type = 'purchase';
Query ID = hadoop_20210531143211_c0a7491d-6e53-457c-a681-b3373f3e7e72
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_1622448704713_0006)

-----  
 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  
-----  
VERTICES: 02/02  [======>>] 100%  ELAPSED TIME: 24.67 s  
-----  
OK  
tot_revenue_oct  
1211532.4500002791  
Time taken: 33.181 seconds, Fetched: 1 row(s)
hive> █
```

Time taken with optimized table is **33.181 seconds**

Insights:

1. The total revenue generated based on Purchase made in the month of October is 1,211,538.43 /-
2. Non-optimized table query took the execution time of 70.448 seconds whereas optimized table query took execution time of 33.181 seconds. We can see there is a significant drop in the execution time of the same query.
3. Hence, optimized table gives better performance in execution time.

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

Base Query:

```
SELECT MONTH(event_time) AS month, COUNT(event_type) AS sum_of_purchases FROM retailstore WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
```

```
hive> SELECT MONTH(event_time) AS month, COUNT(event_type) AS sum_of_purchases FROM retailstore WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
Query ID = hadoop_20210531145005_ab89e20b-5fd2-4441-b3ff-d5d693fe41ff
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622448704713_0007)

-----  
 VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 1 ..... container  SUCCEEDED   5       5       0       0       0       0  
Reducer 2 ..... container  SUCCEEDED   3       3       0       0       0       0  
-----  
VERTICES: 02/02  [======>>] 100%  ELAPSED TIME: 105.34 s  
-----  
OK  
month  sum_of_purchases  
10     245624  
11     322417  
Time taken: 105.951 seconds, Fetched: 2 row(s)
hive> █
```

Time taken is **105.951 seconds**

Optimized table:

```
SELECT MONTH(event_time) AS month, COUNT(event_type) AS sum_of_purchases FROM dynpart_buck_retailsstore WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
```

```
hive> SELECT MONTH(event_time) AS month, COUNT(event_type) AS sum_of_purchases FROM dynpart_buck_retailsstore WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
Query ID = hadoop_20210531145441_1abba22a-a2cb-4940-97ae-a0f41b94a166
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622448704713_0007)

-----  
 VERTICES    MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
Map 1 ..... container  SUCCEEDED   3       3       0       0       0       0       0  
Reducer 2 ..... container  SUCCEEDED   1       1       0       0       0       0       0  
-----  
VERTICES: 02/02  [=====>>] 100% ELAPSED TIME: 23.67 s  
-----  
OK  
month  sum_of_purchases  
10     245619  
11     322412  
Time taken: 24.2 seconds, Fetched: 2 row(s)
hive>
```

Time taken is 24.2 seconds

Insights:

1. Sum of purchases made in the month of October is 245624 and in the month of November 322417, which means number of purchases are increased in November month
2. Non-optimized table query took the execution time of 105.951 seconds whereas optimized table query took execution time of 24.2 seconds. We can see there is a significant drop in the execution time of the same query.
3. Hence, with proper partitioning and bucketing on table we can reduce execution time.

Using Optimized table from below questions onwards:

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

```
SELECT (SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END) - SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END)) AS change_in_rev FROM dynpart_buck_retailsstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11');
```

```
hive> SELECT (SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END) - SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END)) AS change_in_rev FROM dynpart_buck_retailsstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11');
Query ID = hadoop_20210531150845_8a0d07ab-9a3c-41c0-9cla-642b97f55ba2
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_1622448704713_0008)

-----  
 VERTICES    MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
Map 1 ..... container  SUCCEEDED   3       3       0       0       0       0       0  
Reducer 2 ..... container  SUCCEEDED   1       1       0       0       0       0       0  
-----  
VERTICES: 02/02  [=====>>] 100% ELAPSED TIME: 25.95 s  
-----  
OK  
change_in_rev  
319437.789997565  
Time taken: 34.921 seconds, Fetched: 1 row(s)
```

Insights:

1. Time taken to execute the query is 34.921 seconds
2. Revenue increased in November by 319437.789 from October

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

```
SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category FROM dynpart_buck_retailsstore  
WHERE category_code != '';
```

```
hive> SELECT DISTINCT SPLIT(category_code,'\\.')[0] AS Category  
> FROM dynpart_buck_retailsstore  
> WHERE category_code != '';  
Query ID = hadoop_20210531154008_5259e965-64a6-4244-b581-9d099494c969  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1622448704713_0008)  
  
-----  
 VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  
-----  
Map 1 ..... container    SUCCEEDED      6          6          0          0          0          0          0  
Reducer 2 ..... container    SUCCEEDED      5          5          0          0          0          0          0  
-----  
VERTICES: 02/02  [=====>>] 100%  ELAPSED TIME: 65.71 s  
-----  
OK  
category  
furniture  
appliances  
accessories  
apparel  
sport  
stationery  
Time taken: 66.472 seconds, Fetched: 6 row(s)  
hive> █
```

Insights:

1. Time taken to execute the query is 66.472 seconds
2. Total we got 6 distinct categories are – furniture, appliances, accessories, apparel, sport, stationary.

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

```
SELECT SPLIT(category_code,'\\.')[0] AS Category, COUNT(product_id) AS num_of_prod  
FROM dynpart_buck_retailsstore  
WHERE category_code != ''  
GROUP BY SPLIT(category_code,'\\.')[0]  
ORDER BY num_of_prod DESC;
```

```

hive> SELECT SPLIT(category_code,'\\.') [0] AS Category, COUNT(product_id) AS num_of_prod
  > FROM dynpart_buck_retailsstore
  > WHERE category_code != ''
  > GROUP BY SPLIT(category_code,'\\.') [0]
  > ORDER BY num_of_prod DESC;
Query ID = hadoop_20210531154545_6b1e9ab2-632e-48df-a837-caf8d76d0406
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622448704713_0008)

-----  

      VERTICES    MODE     STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED   6        6        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: 66.49 s  

-----  

OK  

category      num_of_prod  

appliances    61736  

stationery    26722  

furniture     23604  

apparel       18232  

accessories   12928  

sport         2  

Time taken: 67.111 seconds, Fetched: 6 row(s)
hive> █

```

Insights:

1. Time taken to execute the query is 67.111 seconds
2. Appliances are having highest number of products available with 61736 compared to other categories.
3. Stationary and Furniture categories are almost equally registered with available ranges from 23000 to 27000.
4. Sports category is least available with 2 products.

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

WITH tot_sales AS(

```

SELECT brand, (SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END) +
SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END)) AS total_sales FROM
dynpart_buck_retailsstore WHERE event_type = 'purchase' AND MONTH(event_time) in
('10','11') AND brand != " GROUP BY brand)

```

```

SELECT brand, total_sales FROM tot_sales ORDER BY total_sales DESC LIMIT 1;
```

```

hive> WITH tot_sales AS(
    > SELECT brand, (SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END) + SUM(CASE WHEN MONTH(event_time)=11 THEN PRICE ELSE 0 END)) AS total_sales
    > FROM dynpart_buck_retailsstore
    > WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11') AND brand != ''
    > GROUP BY brand
    > SELECT brand, total_sales
    > FROM tot_sales
    > ORDER BY total_sales DESC
    > LIMIT 1;
Query ID = hadoop_20210531155143_f7faldf7-3002-4826-bf02-b56a4e73f8fb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1622448704713_0008)

-----  

      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

-----  

Map 1 ..... container  SUCCEEDED   3       3       0       0       0       0       0  

Reducer 2 ..... container  SUCCEEDED   1       1       0       0       0       0       0  

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

-----  

VERTICES: 03/03  [=====>>] 100%  ELAPSED TIME: 27.25 s  

-----  

OK  

brand  total_sales  

runail  148292.46000001638  

Time taken: 27.83 seconds, Fetched: 1 row(s)
hive> █

```

Insights:

1. Runail is the brand that has the highest sales in total of both the months October and November
2. It seems that Runail brand has high popularity among cosmetic lovers and bringing in more
3. products related to Runail brand could help in increasing their profit.

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

WITH brand_sales AS(

```

SELECT brand, SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END) AS
Oct_sales, SUM(CASE WHEN MONTH(event_time)=11 THEN PRICE ELSE 0 END) AS Nov_sales
FROM dynpart_buck_retailsstore WHERE event_type = 'purchase' AND MONTH(event_time) in
('10','11') AND brand != '' GROUP BY brand)

```

SELECT brand, Oct_sales, Nov_sales, Nov_sales-Oct_sales AS sale_diff

FROM brand_sales

WHERE Nov_sales-Oct_sales > 0

ORDER BY sale_diff DESC;

```

hadoop@ip-172-31-27-109:~ 
hive> WITH brand_sales AS(
>   SELECT brand, SUM(CASE WHEN MONTH(event_time)=10 THEN price ELSE 0 END) AS Oct_sales, SUM(CASE WHEN MONTH(event_time)=11 THEN price ELSE 0 END) AS Nov_sales
>   FROM dynpart_buck_retailsstore WHERE event_type = 'purchase' AND MONTH(event_time) in ('10','11') AND brand != ''
>   GROUP BY brand
>   SELECT brand, Oct_sales, Nov_sales, Nov_sales-Oct_sales AS sale_diff
>   FROM brand_sales
>   WHERE Nov_sales-Oct_sales > 0
>   ORDER BY sale_diff DESC;
Query ID = hadoop_20210531160056_02ac390e-0104-4c15-a911-b7eb04393532
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_1622448704713_0009)

-----  

      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: 27.52 s  

-----  

OK  

brand oct_sales nov_sales sale diff  

grattol 35445.540000000205 71472.7100000044 36027.1700000042  

uno 35302.0300000019 51039.7500000047 15737.72000000285  

lianail 5892.84000000024 16394.240000000544 10501.4000000052  

ingarden 23161.3900000002 33566.210000000625 10404.82000000425  

strong 29196.630000000005 38671.26999999975 9474.6399999997  

jesssnail 26287.84000000084 33345.2299999992 7057.389999999839  

cosmoprofi 8322.80999999901 14536.98999999909 6214.18000000008  

polarus 6013.71999999998 11371.93000000004 5358.210000000055  

runail 71537.77000001163 76754.69000000473 5216.919999993101  

freedecor 3421.780000000097 7671.80000000062 4250.020000000052  

staleks 8519.73000000014 11875.61000000015 3355.88000000001  

bpw.style 11572.15000000046 14837.440000000377 3265.289999999917  

lovely 8704.37999999985 11939.05999999967 3234.67999999982  

marathon 7280.74999999996 10273.09999999999 2992.350000000002  

haruyama 9390.69000000088 12352.91000000145 2962.220000000576  

yoko 8756.90999999983 11707.87999999976 2950.96999999992  

italwax 21940.23999999968 24799.369999999864 2859.130000000183  

benovy 409.6199999999999 3255.970000000002 2850.350000000002  

kaypro 881.3399999999995 3268.6999999999985 2387.3599999999985

```



hadoop@ip-172-31-27-109:~

kaypro	881.339999999999	3268.699999999985	2387.359999999988
estel	21756.750000000025	24142.670000000027	2385.920000000002
concept	11032.139999999978	13380.399999999978	2348.26
kapous	11927.159999999738	14093.079999999984	2165.9200000001256
f.o.x	6624.230000000007	8577.279999999979	1953.049999999972
masura	31266.079999999318	33058.46999999992	1792.3900000006033
milv	3904.940000000026	5642.0100000001285	1737.0700000001025
beautix	10493.94999999997	12222.95	1729.000000000031
artex	2730.640000000002	4327.24999999996	1596.609999999942
domix	10472.04999999992	12009.169999999851	1537.1199999999317
shik	3341.200000000035	4839.720000000001	1498.519999999977
smart	4457.259999999875	5902.13999999998	1444.88000000001
roublöff	3491.3600000000006	4913.770000000003	1422.4100000000026
levrana	2243.560000000002	3664.10000000004	1420.5400000000018
oniq	8425.40999999947	9841.6499999998	1416.240000000325
irisk	45591.95999999969	46946.03999999916	1354.079999994633
severina	4775.87999999955	6120.47999999956	1344.600000000013
joico	705.52 2015.100000000006	1309.580000000006	
zeitun	708.660000000003	2009.63	1300.969999999998
beauty-free	554.170000000014	1782.859999999983	1228.689999999969
swarovski	1887.929999999898	3043.159999999794	1155.229999999896
de.lux	1659.699999999784	2775.50999999973	1115.809999999945
metzger	5373.45000000001	6457.160000000005	1083.709999999955
markell	1768.749999999998	2834.42999999994	1065.679999999942
sanoto	157.14 1209.679999999998	1052.54	
nagaraku	4369.740000000042	5327.6800000000285	957.939999999869
ecolab	262.85 1214.300000000009	951.450000000008	
art-visage	2092.709999999978	2997.800000000056	905.090000000079
levissime	2227.500000000064	3085.309999999854	857.80999999979
missha	1293.83 2150.279999999997	856.44999999998	
solomeya	1899.700000000012	2685.79999999996	786.099999999949
rosi	3077.040000000002	3841.560000000001	764.519999999991
refectocil	2716.180000000003	3475.580000000036	759.4000000000005
kaaral	4412.42999999999	5086.07	673.640000000003
kosmekka	1181.439999999996	1813.37	631.930000000003
kinetics	6334.250000000006	6945.25999999998	611.00999999992
browxenna	14331.370000000057	14916.73000000007	585.3600000000133
airnails	5118.900000000015	5691.520000000021	572.620000000063
uskusi	5142.2699999998	5690.310000000031	548.0400000000509
coifin	903.000000000002	1428.48999999998	525.489999999996
s.care	412.68 913.07 500.390000000004		
limoni	1308.900000000005	1796.600000000004	487.699999999998
matrix	3243.250000000001	3726.740000000016	483.490000000007
gehwoł	1089.070000000002	1557.68	468.609999999999

 hadoop@ip-172-31-27-109:~

greymy	29.21	489.49	460.28000000000003	
bioqua	942.8900000000001		1398.1200000000001	455.23
farmavita	837.37	1291.97	454.6	
sophin	1067.8600000000006		1515.520000000002	447.6600000000145
yu-r	271.41	673.709999999999	402.299999999999	
kiss	421.55	817.330000000004	395.7800000000037	
naomi	0.0	389.0	389.0	
lador	2083.610000000002		2471.5300000000016	387.919999999996
ellips	245.8499999999997		606.04	360.19
jas	3318.960000000002		3657.4300000000026	338.4700000000007
lowence	242.8399999999997		567.749999999999	324.909999999999
nitrile	847.279999999999		1162.679999999999	315.4
shary	871.959999999994		1176.489999999996	304.5300000000002
kims	330.04	632.0400000000001	302.0000000000006	
happyfons	801.9200000000004		1091.5900000000008	289.6700000000004
kocostar	310.849999999999		594.929999999998	284.079999999999
insight	1443.7000000000005		1721.960000000001	278.2600000000045
candy	534.959999999999		799.379999999994	264.419999999995
bluesky	10307.240000000156		10565.52999999784	258.28999999628
beauugreen	511.51000000000016		768.349999999999	256.8399999999975
protokeratin	201.25	456.79	255.5400000000002	
trind	298.0700000000005		542.96	244.89
entity	479.71000000000157		719.259999999991	239.549999999975
skinlite	651.939999999997		890.449999999998	238.5100000000001
provoc	827.9900000000004		1063.820000000024	235.830000000021
fedua	52.38	263.81	211.43	
ecocraft	41.16000000000004		241.949999999996	200.789999999996
keen	236.35	435.619999999995	199.269999999995	
mane	66.789999999999		260.26	193.47
freshbubble	318.699999999999		502.339999999975	183.639999999987
matreshka	0.0	182.6700000000004	182.6700000000004	
chi	358.9400000000001		538.6100000000001	179.6700000000002
cristalinas	427.629999999999		584.949999999999	157.3200000000005
farmona	1692.460000000003		1843.429999999998	150.969999999957
latinoil	249.52	384.5900000000015	135.0700000000014	
miskin	158.0400000000002		293.069999999994	135.029999999992
elizavecca	70.53	204.300000000004	133.770000000004	
nefertiti	233.5200000000007		366.64	133.119999999992
finish	98.38	230.38	132.0	
igrobeauty	513.660000000005		645.070000000006	131.4100000000008
dizao	819.130000000003		945.510000000014	126.3800000000102
osmo	645.58	762.31	116.729999999999	
batiste	772.4000000000001		874.169999999998	101.769999999975
carmex	145.0800000000004		243.36	98.279999999997

 hadoop@ip-172-31-27-109:~

eos	54.33999999999996	152.61	98.27000000000001	
depilflax	2707.069999999956	2803.77999999998		96.71000000000231
enjoy	41.35	136.5700000000002	95.2200000000003	
kerasys	430.91000000000014	525.2	94.2899999999999	
aura	83.95	177.5099999999996	93.5599999999996	
plazan	101.3699999999999	194.0100000000005	92.64000000000006	
koelf	422.7299999999996	507.2899999999985	84.5599999999989	
nirvel	163.04	234.3299999999987	71.2899999999988	
konad	739.829999999997	810.669999999992	70.8399999999946	
egomania	77.47	146.0400000000002	68.57000000000002	
cutrin	299.37	367.62	68.25	
laboratorium	246.5	312.52	66.0199999999998	
inm	288.019999999999	351.2100000000001	63.19000000000017	
dewal	0.0	61.2899999999999	61.2899999999999	
marutaka-foot	49.22	109.3300000000001	60.11000000000014	
kares	0.0	59.45	59.45	
profhenna	679.2300000000002	736.849999999999	57.6199999999966	
koelcia	55.5	112.75	57.25	
balbcare	155.3299999999996	212.3799999999997	57.05000000000001	
elskin	251.0900000000001	307.65000000000015	56.56000000000006	
foamie	35.04	80.49	45.4499999999996	
ladykin	125.6499999999999	170.57	44.92	
likato	296.0599999999983	340.9699999999997	44.91000000000014	
mavala	409.0400000000001	446.3200000000001	37.28000000000003	
vilenta	197.5999999999997	231.2100000000004	33.61000000000007	
beautyblender	78.74000000000001	109.4099999999998	30.66999999999973	
biore	60.65000000000006	90.31	29.65999999999997	
orly	902.3800000000002	931.0900000000004	28.71000000000015	
estelare	444.8100000000005	471.87000000000023	27.059999999999718	
profepil	93.3600000000001	118.02000000000001	24.659999999999997	
blixz	38.95	63.4	24.44999999999996	
binacil	0.0	24.2599999999998	24.25999999999998	
godefroy	401.22	425.12	23.89999999999977	
glysolid	69.7299999999998	91.5899999999999	21.86000000000014	
veraclara	50.11000000000001	71.21000000000001	21.1	
juno	0.0	21.08	21.08	
kamill	63.01000000000005	81.49000000000001	18.48000000000004	
treaclemoon	163.3700000000003	181.49000000000004	18.12000000000005	
supertan	50.37000000000001	66.51000000000002	16.14000000000008	
barbie	0.0	12.39	12.39	
deoproce	316.8400000000003	329.1700000000001	12.33000000000041	
rasyan	18.79999999999997	28.9399999999998	10.14	
fly	17.14	27.1699999999998	10.02999999999998	
tertio	236.16	245.8	9.64000000000015	

jaguar	1102.110000000004	1110.650000000003	8.53999999999964
soleo	204.2	212.529999999998	8.32999999999814
neoleor	43.41	51.7	8.29000000000006
moyou	5.71	10.28000000000001	4.57000000000001
bodyton	1376.339999999983	1380.63999999999	4.30000000000637
skinity	8.88	12.44000000000001	3.560000000000005
helloganic	0.0	3.1	3.1
grace	100.9200000000002	102.6099999999999	1.689999999999693
cosima	20.23	20.9299999999993	0.699999999999922
ovale	2.54	3.1	0.56

Time taken: 36.317 seconds, Fetched: 160 row(s)

hive> 

Insights:

1. Here are some 160 brands with increment in the selling from October to November.
2. 'Grattol' brand has the highest total increment i.e., 36,027 /- and 'Ovale' seems to have the least increment of 0.56 /- from October to November.
3. Among all these brands lists, '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.
4. This implies that 'Runail' is the best and popular brand among all other brands within people.

8. Your company wants to reward the top 10 users of its websites with a golden customer plan. Write a query to generate a list of top 10 users who spend the most.

SELECT user_id, SUM(price) AS tot_amt_spend FROM dynpart_buck_retailsstore

WHERE event_type = 'purchase'

GROUP BY user_id

ORDER BY tot_amt_spend DESC

LIMIT 10;

```
hive> SELECT user_id, SUM(price) AS tot_amt_spend FROM dynpart_buck_retailsstore
    > WHERE event_type = 'purchase'
    > GROUP BY user_id
    > ORDER BY tot_amt_spend DESC
    > LIMIT 10;
Query ID = hadoop_20210531161943_b6c52ee0-47d4-4a07-8fa9-a78e316b7fab
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_1622448704713_0010)
```

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: 26.69 s

OK

user_id	tot_amt_spend
557790271	2715.869999999995
150318419	1645.969999999998
562167663	1352.850000000001
531900924	1329.449999999996
557850743	1295.480000000007
522130011	1185.389999999999
561592095	1109.700000000003
431950134	1097.589999999997
566576008	1056.359999999997
521347209	1040.910000000003

Time taken: 34.875 seconds, Fetched: 10 row(s)

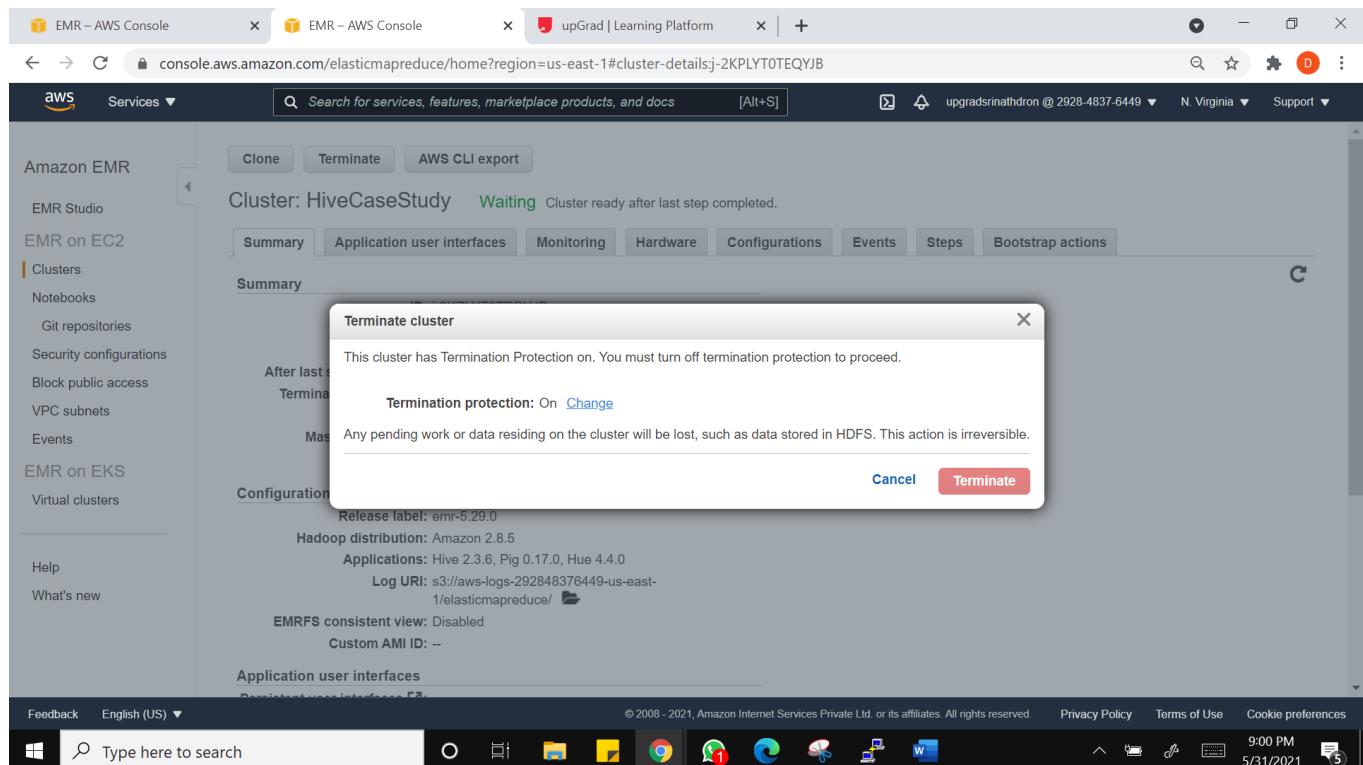
hive> █

Insights:

1. Here is the list of the top 10 users or buyers who have spent the most and could be rewarded with a Golden Customer plan to attract more people in the coming future.
2. With the Optimized table the execution time reduced with proper partitioning and bucketing.
3. Time taken to execute this query on optimized table is 34.875 seconds.

TERMINATION PROCESS:

After completing our analysis, we should terminate the EMR cluster



Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Cluster: HiveCaseStudy Waiting Cluster ready after last step completed.

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Terminate cluster

This cluster has Termination Protection on. You must turn off termination protection to proceed.

Termination protection: On Off  

Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.

Cancel Terminate

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-292848376449-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use Cookie preferences

9:00 PM 5/31/2021

Amazon EMR

EMR Studio

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Create cluster View details Clone Terminate

Filter: All clusters Filter clusters ... 1 cluster (all loaded) C

Name	ID	Status	Creation time (UTC+3)	Elapsed time	Normalized instance hours
HiveCaseStudy	j-2KPLYT0TEQYJB	Terminating User request	2021-05-31 11:04 (UTC+3)	9 hours, 56 minutes	80

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use Cookie preferences

9:01 PM 5/31/2021

Cluster terminated!!