

# HIVE

# CASE STUDY

SUBMITTED BY

**APOORVA KULKARNI**

**KEERTHI P**

**DS37**

## CREATING AN EMR CLUSTER

New Tab | Subscription Details | Nuvepro | EMR - AWS Console

https://us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#quick-create:

Search for services, features, blogs, docs, and more [Alt+S]

### Create Cluster - Quick Options [Go to advanced options](#)

#### General Configuration

Cluster name:

☒ Logging ⓘ

S3 folder:  ⓘ

Launch mode: ☒ Cluster ⓘ ☐ Step execution ⓘ

#### Software configuration

Release:  ⓘ

Applications: ☒ Core Hadoop: Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

☐ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Phoenix 4.14.3, and ZooKeeper 3.4.14

☐ Presto: Presto 0.267 with Hadoop 2.10.1 HDFS and Hive 2.3.9 Metastore

☐ Spark: Spark 2.4.8 on Hadoop 2.10.1 YARN and Zeppelin 0.10.0

☐ Use AWS Glue Data Catalog for table metadata ⓘ

#### Hardware configuration

Instance type:  ⓘ The selected instance type adds 32 GiB of GP2 EBS

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

27°C Light rain ENG IN 11:29 AM 7/5/2022

New Tab | Subscription Details | Nuvepro | EMR - AWS Console

https://us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-3LDBYEGBM16EJ

Search for services, features, blogs, docs, and more [Alt+S]

### Amazon EMR

Clone Terminate AWS CLI export

#### Cluster: hivecasestudy Starting

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

##### Summary

ID: j-3LDBYEGBM16EJ

Creation date: 2022-07-05 11:30 (UTC+5:30)

Elapsed time: 4 seconds

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All / Edit](#)

Master public DNS: --

##### Configuration details

Release label: emr-5.36.0

Hadoop distribution: Amazon 2.10.1

Applications: Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2

Log URI: s3://aws-logs-427018142170-us-east-1/elasticmapreduce/ ⓘ

EMRFS consistent view: Disabled

Custom AMI ID: --

Amazon Linux Release: 2.0.20220426.0 [Learn more](#) ⓘ

##### Application user interfaces

Persistent user interfaces ⓘ: --

On-cluster user -- interfaces ⓘ

##### Network and hardware

Availability zone: --

Subnet ID: [subnet-b1106f90](#) ⓘ

Master: Provisioning 1 m4.large

Core: Provisioning 2 m4.large

Task: --

Cluster scaling: Not enabled

Auto-termination: Not enabled

##### Security and access

Key name: custeremrkey

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

27°C Light rain ENG IN 11:30 AM 7/5/2022

## ESTABLISHING CONECTION USING PUTTY

Amazon EMR

Cluster: **hivecasestudy** Starting

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

**SSH**

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on.

[Learn more](#)

Windows Mac / Linux

1. Download PuTTY.exe to your computer from:  
<http://www.chiark.greenend.org.uk/~sglatham/putty/download.html>
2. Start PuTTY.
3. In the Category list, click Session.
4. In the Host Name field, type `hadoop@ec2-44-204-88-179.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 (`custeremrkey.ppk`) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.

[Close](#)

Security and access

Key name: custeremrkey

Amazon EMR

Cluster: **hivecasestudy** Starting

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

**PuTTY Configuration**

Category: Session, Logging, Terminal, Keyboard, Bell, Features, Appearance, Behaviour, Translation, Selection, Colours, Connection, Data, Proxy, SSH, Serial, Telnet, Rlogin, SFTP/DUP

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address) `hadoop@ec2-44-204-88-179.compute-1.amazonaws.com` Port `22`

Connection type: ☒ SSH ☐ Serial ☐ Other: `Telnet`

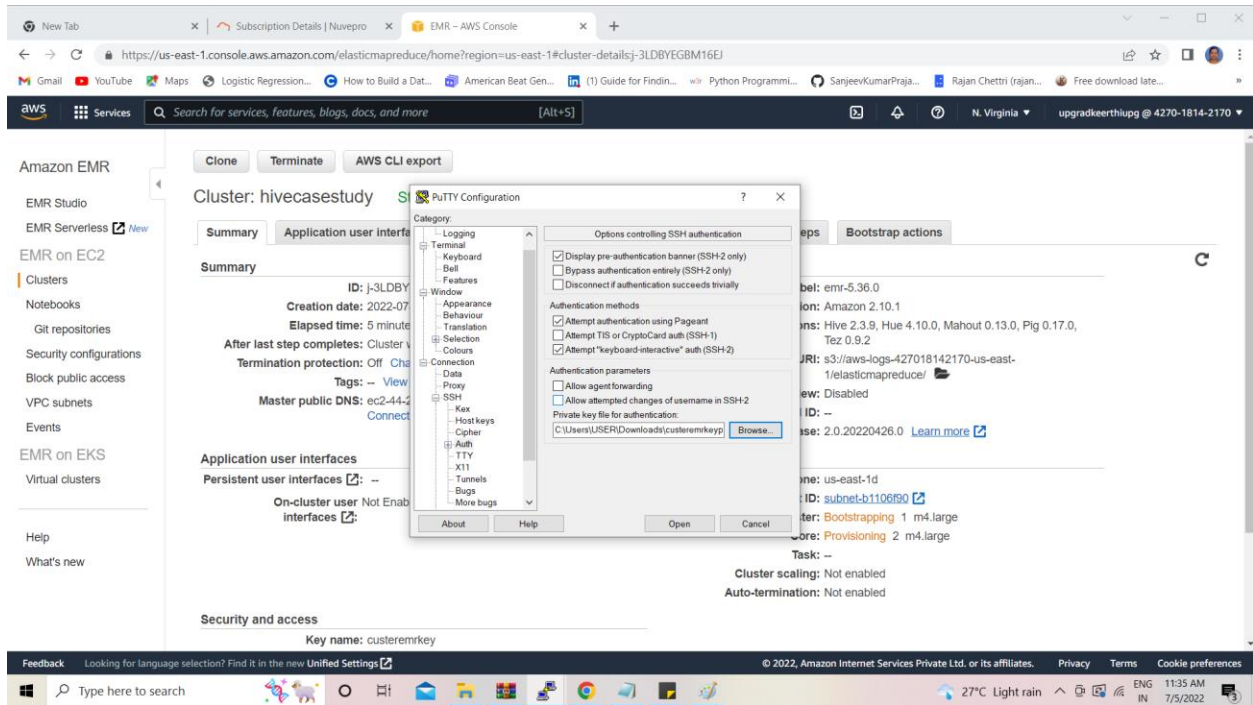
Load, save or delete a stored session

Saved Sessions

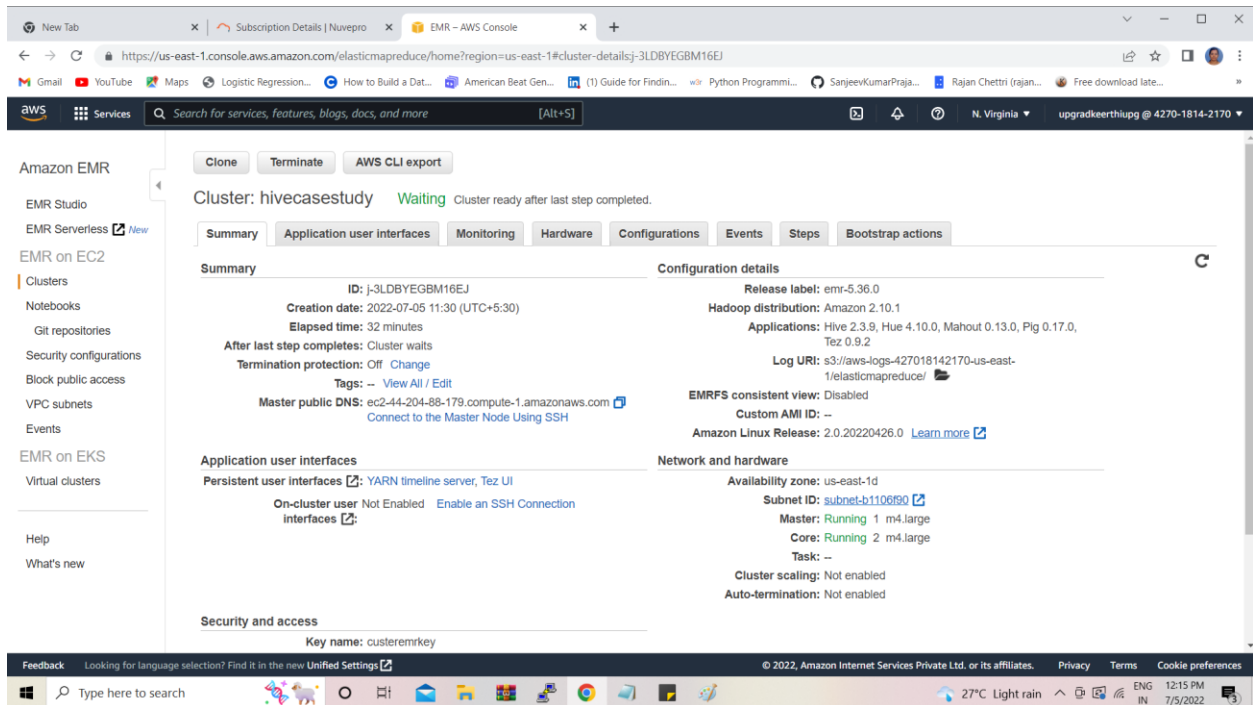
Default Settings [Load](#) [Save](#) [Delete](#)

Close window on exit: ☐ Always ☐ Never ☒ Only on clean exit

[About](#) [Help](#) [Open](#) [Cancel](#)



## CONNECTION ESTABLISHED SUCCESSFULLY AND RUNNING



## EMR CLUSTER

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

  _ | _ | _ )
  _ | ( _ - /
 _ _ | \ _ _ | _ _ |

Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R
EE::::EEEEEEEE::::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 MMMMMMMM MMMMMMMM RRRRRRR RRRRRR
```

## DOWNLOADING THE INPUT FILES FROM S3 TO THE LOCAL SYSTEM

wget https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv

wget <https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv>

```
[hadoop@ip-172-31-2-56 ~]$ wget https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv
--2022-07-04 10:46:09-- https://e-commerce-events-ml.s3.amazonaws.com/2019-Oct.csv
Resolving e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)... 52.217.90.92
Connecting to e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)|52.217.90.92|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 482542278 (460M) [text/csv]
Saving to: '2019-Oct.csv'

100%[=====>] 482,542,278 68.1MB/s in 6.8s

2022-07-04 10:46:16 (67.3 MB/s) - '2019-Oct.csv' saved [482542278/482542278]

[hadoop@ip-172-31-2-56 ~]$ wget https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv
--2022-07-04 10:46:36-- https://e-commerce-events-ml.s3.amazonaws.com/2019-Nov.csv
Resolving e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)... 54.231.194.57
Connecting to e-commerce-events-ml.s3.amazonaws.com (e-commerce-events-ml.s3.amazonaws.com)|54.231.194.57|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 545839412 (521M) [text/csv]
Saving to: '2019-Nov.csv'

100%[=====>] 545,839,412 68.0MB/s in 7.6s

2022-07-04 10:46:44 (68.1 MB/s) - '2019-Nov.csv' saved [545839412/545839412]
```

### CREATING A DIRECTORY

```
hadoop fs -mkdir /ecom/
```

### CHECKING IF DIRECTORY IS CREATED

```
hadoop fs -ls /
```

### INSERTING FILES FROM LOCAL SYSTEM TO HDFS

```
hadoop fs -put ./2019-Oct.csv /ecom/
```

```
hadoop fs -put ./2019-Nov.csv /ecom/
```

```
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -mkdir /ecom
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-07-04 10:48 /apps
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-07-04 10:49 /ecom
drwxrwxrwt - hdfs hdfsadmingroup 0 2022-07-04 10:48 /tmp
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-07-04 10:48 /user
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-07-04 10:48 /var
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -put ./2019-Oct.csv /ecom/
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -put ./2019-Nov.csv /ecom/
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-07-04 10:48 /apps
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-07-04 10:51 /ecom
drwxrwxrwt - hdfs hdfsadmingroup 0 2022-07-04 10:48 /tmp
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-07-04 10:48 /user
drwxr-xr-x - hdfs hdfsadmingroup 0 2022-07-04 10:48 /var
```

### CHECKING THE PATH IF FILES ARE LOADED CORRECTLY

```
hadoop fs -ls /ecom
```

### LOGGING INTO HIVE

```
hive
```

```
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -ls /ecom
Found 2 items
-rw-r--r-- 1 hadoop hdfsadmingroup 545839412 2022-07-04 10:51 /ecom/2019-Nov.
csv
-rw-r--r-- 1 hadoop hdfsadmingroup 482542278 2022-07-04 10:50 /ecom/2019-Oct.
csv
[hadoop@ip-172-31-2-56 ~]$ hive
Hive Session ID = b6ef9fe4-89e7-435f-ac1a-2d54cf3c288b

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

## CREATING DATABASE IN HIVE

create database if not exists ecom;

use ecom;

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> create database if not exists ecom;
OK
Time taken: 1.494 seconds
hive> use ecom;
OK
Time taken: 0.097 seconds
```

## CREATING A STATIC EXTERNAL TABLE

set hive.cli.print.header=true;

create External table if not exists ecom\_tabl(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 '/ecom' TBLPROPERTIES("skip.header.line.count"="1");

```
hive>
> use ecom;
OK
Time taken: 1.008 seconds
hive>
> ;
hive> create External table if not exists ecom_tabl(event_time timestamp,event_t
ype 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.Open
CSVSerde'
> WITH SERDEPROPERTIES ("separatorChar"=",","quoteChar"="\","escapeChar"="\
\\")stored as textfile LOCATION '/ecom' TBLPROPERTIES("skip.header.line.count"="
1");
OK
Time taken: 0.206 seconds
```



## CHECKING IF THE TABLE IS CREATED AND ITS PROPERTIES

desc ecom\_tabl;

```
hive> desc ecom_tabl;
OK
event_time          string          from deserializer
event_type           string          from deserializer
product_id           string          from deserializer
category_id          string          from deserializer
category_code        string          from deserializer
brand                string          from deserializer
price                string          from deserializer
user_id              string          from deserializer
user_session         string          from deserializer
Time taken: 0.513 seconds, Fetched: 9 row(s)
```

## CHECKING IF THE DATA IS LOADED FROM HDFS TO HIVE TABLE

SELECT \* FROM ecom\_tabl limit 5;

```
hive>
> set hive.cli.print.header=true;
hive> select * from ecom_tabl limit 5;
OK
ecom_tabl.event_time  ecom_tabl.event_type  ecom_tabl.product_id  ecom_tabl.category_id  ecom_tabl.category_code  ecom_tabl.brand  ecom_tabl.price  ecom_tabl.user_id  ecom_tabl.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 2
2.22 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: 0.191 seconds, Fetched: 5 row(s)
```

## DYNAMIC PARTITION AND BUCKETING

set hive.exec.dynamic.partition=true;

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

set hive.enforce.bucketing=true;



## CREATE EXTERNAL TABLE WITH DYNAMIC PARTITIONING ON EVENT\_TYPE, BUCKET OF 8

create External table if not exists dyn\_ecom\_tab(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 (user\_id) into 8 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

stored as textfile;

```
hive> set hive.exec.dynamic.partition=true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.enforce.bucketing=true;
hive> create External table if not exists dyn_ecom_tab(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 (user_id) into 8 buckets ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > stored as textfile;
OK
Time taken: 0.084 seconds
```

desc dyn\_ecom\_tab;

```
hive> desc dyn_ecom_tab;
OK
col_name      data_type      comment
event_time    string         from deserializer
product_id    string         from deserializer
category_id   string         from deserializer
category_code string         from deserializer
brand         string         from deserializer
price         string         from deserializer
user_id       string         from deserializer
user_session  string         from deserializer
event_type    string

# Partition Information
# col_name      data_type      comment
event_type     string
Time taken: 0.261 seconds, Fetched: 13 row(s)
```

## LOADING DATA INTO DYNAMIC TABLE

insert into dyn\_ecom\_tab partition (event\_type) select event\_time, product\_id, category\_id, category\_code, brand, price, user\_id, user\_session, event\_type from ecom\_tabl;

```
hive> insert into dyn_ecom_tab partition (event_type) select event_time, product
_id, category_id, category_code, brand, price, user_id, user_session, event_type
from ecom_tabl
>
> ;
Query ID = hadoop_20220704111144_83c444a9-2787-4e52-90ca-3ac01c462ade
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1656931749140
_0003)
```

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FA
ILED	KILLED							
Map 1		container	INITED	4	0	0	4	
		container	INITED	2	0	0	2	
	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FA
ILED	KILLED							
-----2	[>>-----]			0%	ELAPSED TIME: 0.03 s			
Map 1		container	INITED	4	0	0	4	
		container	INITED	2	0	0	2	
	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FA
ILED	KILLED							

```
hadoop@ip-172-31-2-56:~$
VERTICES  MODE  STATUS TOTAL COMPLETED RUNNING PENDING FA
ILED KILLED
-----2 [ >>-----] 0% ELAPSED TIME: 15.29 s
Map 1 container RUNNING 4 0 2 2
container INITED 2 0 0 2
VERTICES  MODE  STATUS TOTAL COMPLETED RUNNING PENDING PENDING FA
ILED KILLED
-----2 [ >>-----] 0% ELAPSED TIME: 15.80 s
Map 1 container RUNNING 4 0 2 2
container INITED 2 0 0 2
VERTICES  MODE  STATUS TOTAL COMPLETED RUNNING PENDING PENDING FA
ILED KILLED
-----2 [ >>-----] 0% ELAPSED TIME: 16.32 s
Map 1 container RUNNING 4 0 2 2
container INITED 2 0 0 2
VERTICES  MODE  STATUS TOTAL COMPLETED RUNNING PENDING PENDING FA
ILED KILLED
-----2 [ >>-----] 0% ELAPSED TIME: 16.83 s
Map 1 container RUNNING 4 0 2 2
container INITED 2 0 0 2
VERTICES  MODE  STATUS TOTAL COMPLETED RUNNING PENDING PENDING FA
ILED KILLED
-----2 [ >>-----] 0% ELAPSED TIME: 17.33 s
Map 1 container RUNNING 4 0 2 2
VERTICES  MODE  STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 4 4 0 0 0 0
Reducer 2 ..... container SUCCEEDED 2 2 0 0 0 0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 167.10 s
Loading data to table ecom.dyn_ecom_tab partition (event_type=null)
Loaded : 4/4 partitions.
Time taken to load dynamic partitions: 0.569 seconds
Time taken for adding to write entity : 0.009 seconds
OK
event_time product_id category_id category_code brand price user_id user_session event_type
Time taken: 178.286 seconds
hive>
```

### CHECKING IF DATA IS INSERTED

select \* from dyn\_ecom\_tab limit 7;

```
hive>
> select * from dyn_ecom_tab limit 7;
OK
dyn_ecom_tab.event_time dyn_ecom_tab.product_id dyn_ecom_tab.category_id      d
yn_ecom_tab.category_code      dyn_ecom_tab.brand      dyn_ecom_tab.price      d
yn_ecom_tab.user_id      dyn_ecom_tab.user_session      dyn_ecom_tab.event_type
2019-10-30 14:40:25 UTC 5847870 1487580006317032337      1.90 4
23582414      bc8d2c60-cd55-4380-90f5-12ca42d952b0      cart
2019-10-28 16:09:54 UTC 5724608 1487580005427839846      irisk 2.48 5
47930959      1289a494-650f-46aa-8083-a2bb3cb25c09      cart
2019-10-27 16:27:36 UTC 5878151 1487580005268456287      cosmoprofi 7
.14 444041100      bldef3f7-ecdf-479e-b686-ca14ecee729e      cart
2019-10-31 12:30:18 UTC 5844309 1487580006317032337      1.57 5
66113200      bb92950d-462d-473d-8ea3-989e84456003      cart
2019-10-31 08:16:19 UTC 5696152 1487580005134238553      runail 2.38 4
97964732      63600e1e-044e-4885-a517-87fa9b643316      cart
2019-10-27 16:27:33 UTC 6977 1487580006895846315      runail 5.13 4
44041100      bldef3f7-ecdf-479e-b686-ca14ecee729e      cart
2019-10-28 16:10:03 UTC 5892522 1487580006317032337      4.76 4
76419375      56e86b11-c310-4fba-b300-852c88b253b0      cart
Time taken: 0.32 seconds, Fetched: 7 row(s)
```

### CHECKING FOR PARTITIONS CREATED IN HIVE

show partitions dyn\_ecom\_tab;

```
hive> show partitions dyn_ecom_tab;
OK
partition
event_type=cart
event_type=purchase
event_type=remove_from_cart
event_type=view
Time taken: 0.084 seconds, Fetched: 4 row(s)
```

### CHECKING PARTITIONS CREATED IN HADOOP

hadoop fs -ls /user/hive/warehouse/ecom.db/dyn\_ecom\_tab

```
[hadoop@ip-172-31-2-56 ~]$ hadoop fs -ls /user/hive/warehouse/ecom.db/dyn_ecom_tab
Found 4 items
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-07-04 11:14 /user/hive/warehouse/ecom.db/dyn_ecom_tab/event_type=cart
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-07-04 11:14 /user/hive/warehouse/ecom.db/dyn_ecom_tab/event_type=purchase
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-07-04 11:14 /user/hive/warehouse/ecom.db/dyn_ecom_tab/event_type=remove_from_cart
drwxr-xr-x - hadoop hdfsadmingroup 0 2022-07-04 11:14 /user/hive/warehouse/ecom.db/dyn_ecom_tab/event_type=view
[hadoop@ip-172-31-2-56 ~]$
```

## CHECKING TIME DIFFERENCE BETWEEN STATIC AND DYNAMIC TABLES

SELECT SUM(price) AS NOV\_REVENUE FROM dyn\_ecom\_tab WHERE month(event\_time) = 11 AND event\_type = 'purchase';

```
hive>
> SELECT SUM(price) AS NOV_REVENUE FROM dyn_ecom_tab WHERE month(event_time) = 11 AND event_type = 'purchase';
Query ID = hadoop_20220705063938_77aab1cc-6708-4fc7-8671-f6bd4c870156
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    8        8        0        0        0        0
Reducer 2 ..... container  SUCCEEDED    1        1        0        0        0        0
-----
VERTICES: 02/02  [=====>>] 100%  ELAPSED TIME: 28.12 s
-----
OK
nov_revenue
1531016.900000001
Time taken: 28.752 seconds, Fetched: 1 row(s)
```

Time taken for dynamic table to execute the query and display the result is 28.752 seconds

SELECT SUM(price) AS NOV\_REVENUE FROM ecom\_tab1 WHERE month(event\_time) = 11 AND event\_type = 'purchase';

```
hive> SELECT SUM(price) AS NOV_REVENUE FROM ecom_tab1 WHERE month(event_time) = 11 AND event_type = 'purchase';
Query ID = hadoop_20220705064051_c5e99c8d-2395-4464-b776-a1358fbbb05f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
      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: 39.66 s
-----
OK
nov_revenue
1531016.900000122
Time taken: 40.23 seconds, Fetched: 1 row(s)
```

Time taken for static table to execute the query and display the result is 40.23 seconds.

## QUERIES

1. SELECT SUM(price) AS OCT\_REVENUE FROM dyn\_ecom\_tab WHERE month(event\_time) = 10 AND event\_type = 'purchase';

OCT\_REVENUE

1211538.430000025

```
hive> SELECT SUM(price) AS OCT_REVENUE FROM dyn_ecom_tab WHERE month(event_time) = 10 AND event_type = 'purchase';
Query ID = hadoop_20220705063019_4031a1fe-26fc-4bf6-8fc8-606a7fb71a05
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_1657001276041_0002)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   8         8         0         0         0         0
Reducer 2 ..... container  SUCCEEDED   1         1         0         0         0         0
-----
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 29.05 s
-----
OK
oct_revenue
1211538.4300000053
Time taken: 37.531 seconds, Fetched: 1 row(s)
```

2. SELECT MONTH(event\_time) AS MONTH, count(product\_id) AS PURCHASE\_PER\_MONTH FROM dyn\_ecom\_tab WHERE event\_type = 'purchase' GROUP BY MONTH(event\_time);

MONTH PURCHASE\_PER\_MONTH

10          245624

11          322417

```
hive> SELECT MONTH(event_time) AS MONTH, count(product_id) AS PURCHASE_PER_MONTH FROM dyn_ecom_tab WHERE event_type = 'purchase' GROUP BY MONTH(event_time);
Query ID = hadoop_20220705063108_20bdaed6-cc5c-494f-ab63-995f18970442
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   8         8         0         0         0         0
Reducer 2 ..... container  SUCCEEDED   2         2         0         0         0         0
-----
VERTICES: 02/02  [=====>>>] 100%  ELAPSED TIME: 29.06 s
-----
OK
month  purchase_per_month
10     245624
11     322417
Time taken: 29.794 seconds, Fetched: 2 row(s)
```

3. WITH MONTHLY\_SALE AS ( select round ( sum (cAse when date\_format (event\_time, 'MM') = 10 then price else 0 end),2) AS Oct\_Sales,

round (sum (cAse when date\_format (event\_time, 'MM') =11 then price else 0 end),2) AS Nov\_Sales  
from dyn\_ecom\_tab WHERE event\_type = 'purchase' AND date\_format (event\_time, 'MM')

in ('10', '11') ) SELECT Oct\_Sales, Nov\_Sales, (Nov\_Sales - Oct\_Sales) AS DIFFERENCE from  
MONTHLY\_SALE;

Oct_sales	Nov_sales	DIFFERENCE
1211538.43	1531016.9	319478.47

```
hive> WITH MONTHLY_SALE AS ( select round ( sum (cAse when date_format (event_time, 'MM') = 10 then price else 0 end),2) AS Oct_Sales,
> round (sum (cAse when date_format (event_time, 'MM') =11 then price else 0 end),2) AS Nov_Sales from dyn_ecom_tab WHERE event_type = 'purchase' AND date_format (event_time, 'MM')
> in ('10', '11') ) SELECT Oct_Sales, Nov_Sales, (Nov_Sales - Oct_Sales) AS DIFFERENCE from MONTHLY_SALE;
Query ID = hadoop_20220705063149_ebf76a90-7929-48bd-b3db-c14a60274227
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   8         8         0         0         0         0
Reducer 2 ..... container  SUCCEEDED   1         1         0         0         0         0
-----
VERTICES: 02/02  [======>>>] 100% ELAPSED TIME: 39.61 s
-----
OK
oct_sales      nov_sales      difference
1211538.43     1531016.9     319478.47
Time taken: 40.548 seconds, Fetched: 1 row(s)
```

4. SELECT DISTINCT SPLIT(category\_code,'\\\.')[0] AS Prdct\_Category FROM dyn\_ecom\_tab WHERE  
SPLIT(category\_code,'\\\.')[0] IS NOT NULL;

Prdct\_Category

accessories

apparel

appliances

furniture

sport

stationery

```

hive> SELECT DISTINCT SPLIT(category_code,'\\\.')[0] AS Prdct_Category FROM dyn_ecom_tab WHERE SPLIT(category_code,'\\\.')[0] IS NOT NULL;
Query ID = hadoop_20220705063238_4297819e-5612-4622-93a2-40678c4288c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   16         16         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    1          1         0         0         0         0
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 59.47 s
-----
OK
prdct_category

accessories
apparel
appliances
furniture
sport
stationery
Time taken: 60.282 seconds, Fetched: 7 row(s)
hive> SELECT SPLIT(category_code,'\\\.')[0] AS Prdct_Category, COUNT(product_id) as Tot_Prdcts FROM dyn_ecom_tab GROUP BY SPLIT(category_code,'\\\.')[0];

```

5. SELECT SPLIT(category\_code,'\\\.')[0] AS Prdct\_Category, COUNT(product\_id) as Tot\_Prdcts FROM dyn\_ecom\_tab GROUP BY SPLIT(category\_code,'\\\.')[0];

Prdct_Category	Tot_Prdcts
accessories	12929
apparel	18232
appliances	61736
furniture	23604
sport	2
stationery	26722

```

hive> SELECT SPLIT(category_code,'\\\.')[0] AS Prdct_Category, COUNT(product_id) as Tot_Prdcts FROM dyn_ecom_tab GROUP BY SPLIT(category_code,'\\\.')[0];
Query ID = hadoop_20220705063346_32f98394-2735-4018-89b2-be4499a07ff9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   16         16         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    1          1         0         0         0         0
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 57.52 s
-----
OK
prdct_category  tot_prdcts
8594895
accessories    12929
apparel 18232
appliances    61736
furniture     23604
sport         2
stationery    26722
Time taken: 58.147 seconds, Fetched: 7 row(s)

```



6. SELECT brand, ROUND(price) AS MAX\_SALES FROM dyn\_ecom\_tab WHERE brand IS NOT NULL AND event\_type = 'purchase' GROUP BY brand ORDER BY MAX\_SALES DESC LIMIT 3;

brand MAX\_SALES

runail 148297.94

```
hive> SELECT brand, ROUND(SUM(price)) AS MAX_SALES FROM dyn_ecom_tab WHERE brand IS NOT NULL AND event_type = 'purchase' GROUP BY brand ORDER BY MAX_SALES DESC LIMIT 3;
Query ID = hadoop_20220705063606_52ee675e-ff27-45f1-b7f3-624d98e414f4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)

-----
VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   8         8         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: 27.75 s
-----
OK
brand  max sales
      1094188.0
runail 148298.0
grattol 106918.0
```

7. WITH MONTHLY\_SALES AS ( SELECT brand, round ( SUM (case when date\_format (event\_time, 'MM') = 10

then price else 0 end),2) AS Sales\_Oct, round (SUM (case when date\_format

(event\_time, 'MM') =11 then price else 0 end),2) AS Sales\_Nov FROM dyn\_ecom\_tab WHERE event\_type = 'purchase' AND date\_format (event\_time, 'MM')

in ('10', '11') GROUP BY brand ) SELECT brand, Sales\_Oct, Sales\_Nov, (Sales\_Nov - Sales\_Oct)

AS MONTHLY\_SALESDIFF FROM MONTHLY\_SALES WHERE (Sales\_Nov-Sales\_Oct) > 0 ORDER BY MONTHLY\_SALESDIFF DESC;

A total of 161 brands have increased its sales from October to November.

```
hive> WITH MONTHLY_SALES AS ( SELECT brand, round ( SUM (case when date_format (event_time, 'MM') = 10
> then price else 0 end),2) AS Sales_Oct, round (SUM (case when date_format
> (event_time, 'MM') =11 then price else 0 end),2) AS Sales_Nov FROM dyn_ecom_tab WHERE event_type = 'purchase' AND date_format (event_time, 'MM')
> in ('10', '11') GROUP BY brand ) SELECT brand, Sales_Oct, Sales_Nov, (Sales_Nov - Sales_Oct)
> AS MONTHLY_SALESDIFF FROM MONTHLY_SALES WHERE (Sales_Nov-Sales_Oct) > 0 ORDER BY MONTHLY_SALESDIFF DESC;
```

Query ID = hadoop\_20220705063642\_f8f41f57-eb77-425f-b8d1-fbb51a186e0a

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1657001276041\_0002)

	VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1 .....	container	SUCCEEDED	8	8	0	0	0	0	0
Reducer 2 .....	container	SUCCEEDED	2	2	0	0	0	0	0
Reducer 3 .....	container	SUCCEEDED	1	1	0	0	0	0	0

VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 37.29 s

OK

brand	sales oct	sales nov	monthly_salesdiff
474679.06	619509.24	144830.18	
grattol 35445.54	71472.71	36027.1700000000006	
uno 35302.03	51039.75	15737.7200000000001	
lianail 5892.84	16394.24	10501.4000000000001	
ingarden 23161.39	33566.21	10404.82	
strong 29196.63	38671.27	9474.6399999999996	
jessnail 26287.84	33345.23	7057.3900000000003	
cosmoprofi 8322.81	14536.99	6214.18	
polarus 6013.72	11371.93	5358.21	
runail 71539.28	76758.66	5219.3800000000005	
freedecor 3421.78	7671.8	4250.02	
staleks 8519.73	11875.61	3355.8800000000001	
bpw.style 11572.15	14837.44	3265.2900000000001	
lovely 8704.38	11939.06	3234.6800000000003	
marathon 7280.75	10273.1	2992.3500000000004	
haruyama 9390.69	12352.91	2962.2199999999993	
yoko 8756.91	11707.88	2950.9699999999993	
italwax 21940.24	24799.37	2859.12999999999974	
benovy 409.62	3259.97	2850.35	
kaypro 881.34	3268.7	2387.3599999999997	
estel 21756.75	24142.67	2385.9199999999993	
concept 11032.14	13380.4	2348.26	

kares 0.0	59.45	59.45
profhenna	679.23	736.85
koelcia 55.5	112.75	57.25
balbcare	155.33	212.38
elskin 251.09	307.65	56.559999999999974
foamie 35.04	80.49	45.449999999999996
ladykin 125.65	170.57	44.919999999999999
likato 296.06	340.97	44.910000000000025
mavala 409.04	446.32	37.279999999999997
vilenta 197.6	231.21	33.610000000000014
beautyblender	78.74	109.41
biore 60.65	90.31	29.660000000000004
orly 902.38	931.09	28.710000000000036
estelare	444.81	471.87
profepil	93.36	118.02
blixz 38.95	63.4	24.449999999999996
binacil 0.0	24.26	24.26
godefroy	401.22	425.12
glysolid	69.73	91.59
veraclara	50.11	71.21
juno 0.0	21.08	21.08
kamill 63.01	81.49	18.479999999999997
treaclemoon	163.37	181.49
supertan	50.37	66.51
barbie 0.0	12.39	12.39
deoproce	316.84	329.17
rasyan 18.8	28.94	10.14
fly 17.14	27.17	10.030000000000001
tertio 236.16	245.8	9.640000000000015
jaguar 1102.11	1110.65	8.540000000000191
soleo 204.2	212.53	8.330000000000013
neoleor 43.41	51.7	8.290000000000006
moyou 5.71	10.28	4.569999999999999
bodyton 1376.34	1380.64	4.300000000000182
skinity 8.88	12.44	3.5599999999999987
helloganic	0.0	3.1
grace 100.92	102.61	1.6899999999999977
cosima 20.23	20.93	0.6999999999999993
ovale 2.54	3.1	0.56

Time taken: 37.865 seconds, Fetched: 161 row(s)

8. SELECT user\_id, round(SUM(price),2) AS TOT\_PURCHASE FROM dyn\_ecom\_tab WHERE event\_type = 'purchase' GROUP BY user\_id ORDER BY TOT\_PURCHASE DESC LIMIT 10;

user_id	TOT_PURCHASE
557790271	2715.87
150318419	1645.97
562167663	1352.85
531900924	1329.45
557850743	1295.48
522130011	1185.39
561592095	1109.7
431950134	1097.59
566576008	1056.36
521347209	1040.91

```
hive> SELECT user_id, round(SUM(price),2) AS TOT_PURCHASE FROM dyn_ecom_tab WHERE event_type = 'purchase' GROUP BY user_id ORDER BY TOT_PURCHASE DESC LIMIT 10;
Query ID = hadoop_20220705063739_a0c20df1-4287-405f-9d43-76323a1b14e2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1657001276041_0002)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1 .....	container	SUCCEEDED	8	8	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: 29.56 s
OK
user_id tot_purchase
557790271      2715.87
150318419      1645.97
562167663      1352.85
531900924      1329.45
557850743      1295.48
522130011      1185.39
561592095      1109.7
431950134      1097.59
566576008      1056.36
521347209      1040.91
Time taken: 30.235 seconds, Fetched: 10 row(s)
```

## DELETING THE TABLES AND DATABASE IN HIVE

DROP TABLE dyn\_ecom\_tab;

DROP TABLE ecom\_tabl;

DROP DATABASE ecom;

```
hive> DROP TABLE dyn_ecom_tab;
OK
Time taken: 0.241 seconds
hive> DROP TABLE ecom_tabl;
OK
Time taken: 0.106 seconds
```

```
hive> DROP DATABASE ecom;
OK
Time taken: 0.111 seconds
hive> █
```

## TERMINATING THE CLUSTER

The screenshot displays the AWS Management Console interface for an Amazon EMR cluster. The cluster is in the 'Terminating' state, as indicated by the orange label and the text 'Terminated by user request'. The console shows various tabs for the cluster, including Summary, Application user interfaces, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. The Summary tab is selected, displaying the following information:

- Cluster:** keeemrcluster
- ID:** j-1SM8OKZ82J4V4
- Release label:** emr-6.6.0
- Creation date:** 2022-07-04 16:11 (UTC+5:30)
- Hadoop distribution:** Amazon 3.2.1
- Elapsed time:** 2 hours, 21 minutes
- Applications:** Hive 3.1.2, Hue 4.10.0
- After last step completes:** Cluster waits
- Log URI:** s3://aws-logs-427018142170-us-east-1/elasticmapreduce/
- Termination protection:** Off
- EMRFS consistent view:** Disabled
- Tags:** --
- Custom AMI ID:** --
- Master public DNS:** ec2-3-238-52-213.compute-1.amazonaws.com
- Amazon Linux Release:** 2.0.20220426.0

The console also shows the 'Application user interfaces' section, which includes 'YARN timeline server' and 'Tez UI'. The 'Network and hardware' section displays the 'Availability zone' as us-east-1c, the 'Subnet ID' as subnet-6af6870c, and the 'Master' node as 'Terminated 1 m4.large'. The 'Core' node is also 'Terminated 1 m4.large'. The 'Task' node is '--'. The 'Cluster scaling' is 'Not enabled' and 'Auto-termination' is 'Not enabled'.