Project

Set up a shopping cart database and transfer data between MySQL and Hive using Sqoop.

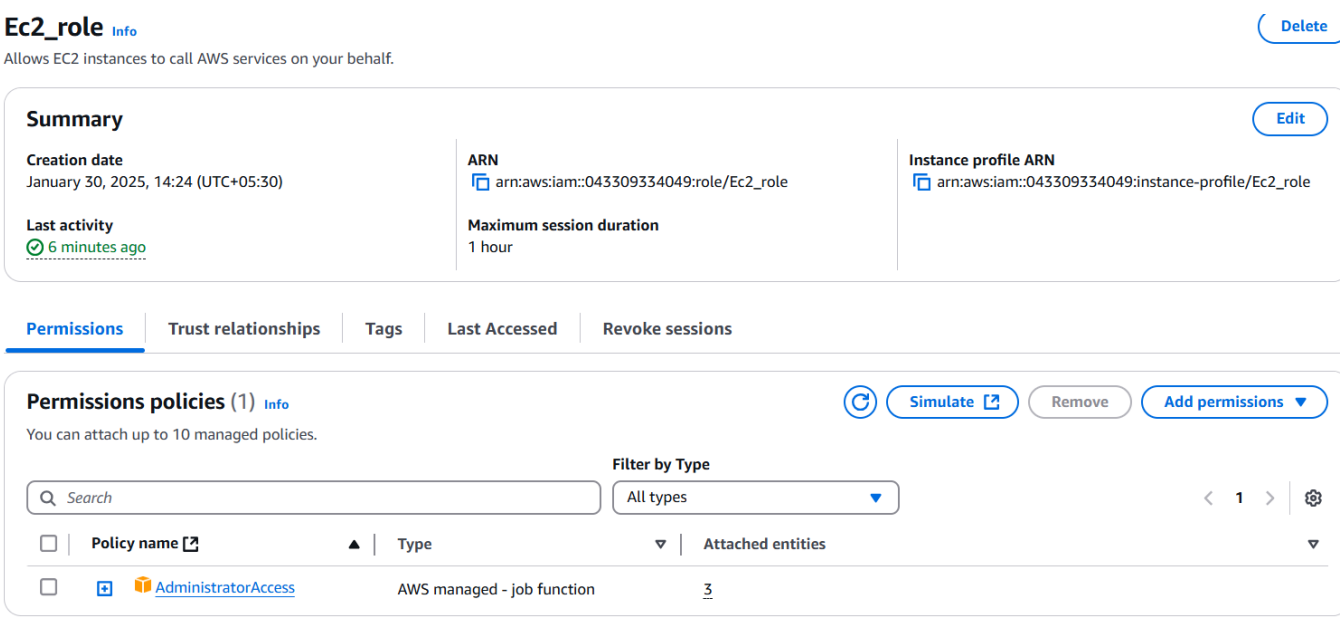
Step 1:

create two Iam roles

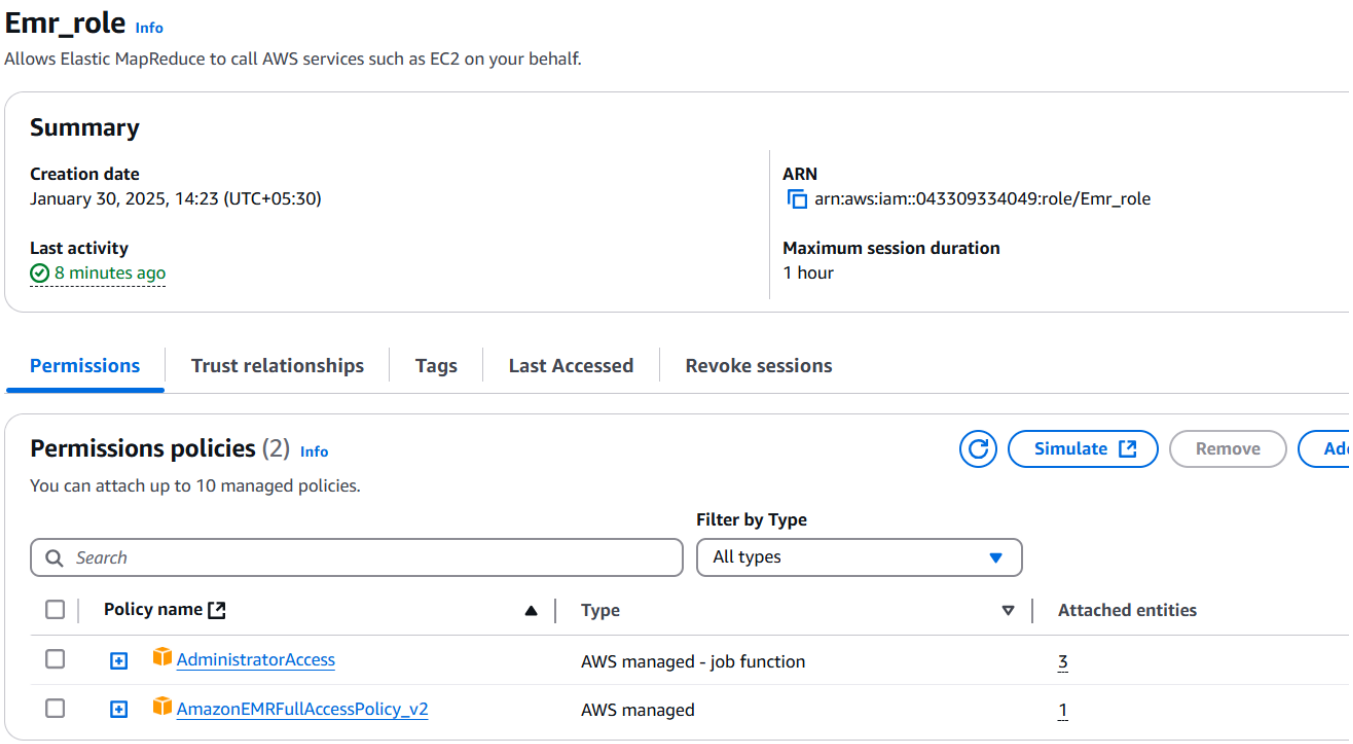
These roles are used in Emr cluster for authentication and authorization of service in aws.

1. *Ec2-Role*
2. *EMR-Role*

Add only administration access



Add both Administration Access and AmazonEMRFullAccessPolicy_v2

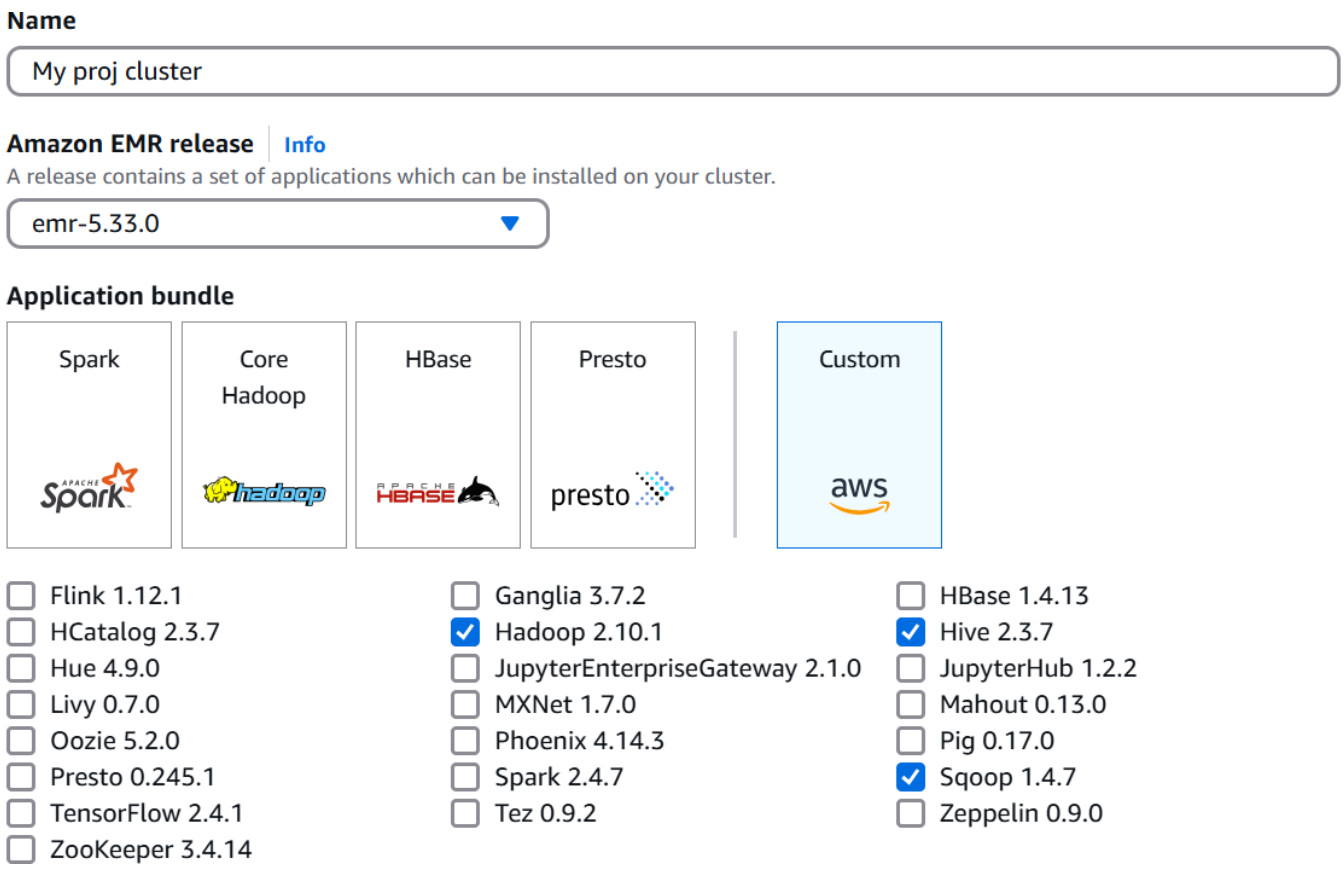


Step 2:

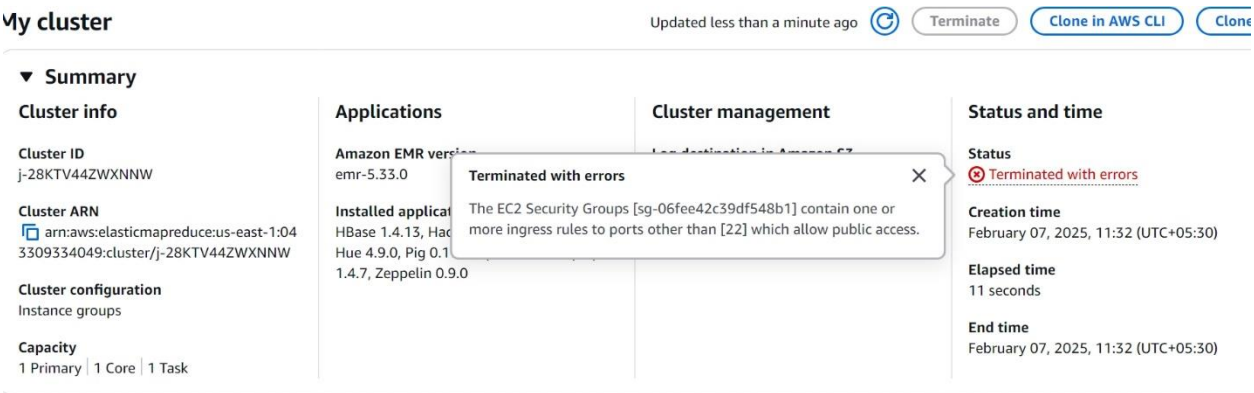
create an Emr cluster

With version 5.33.0

And select the required technologies



While creating EMR cluster sometimes cluster gets terminated instantly due to duplicate inbound rule in security groups



Summary

Cluster info

Cluster ID
j-G6U7K3MXODXW

Cluster ARN
arn:aws:elasticmapreduce:ap-southeast-1:396608798635:cluster/j-G6U7K3MXODXW

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version
emr-5.33.0

Installed applications
HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Pig 0.17.0, Spark 2.4.7, Sqoop 1.4.7, Zeppelin 0.9.0

Cluster management

Log destination in Amazon S3
aws-logs-396608798635-ap-southeast-1/elasticmapreduce

Persistent application UIs
Spark History Server
YARN timeline server
Tez UI

Primary node public DNS
ec2-13-215-60-115.ap-southeast-1.compute.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
Waiting

Creation time
February 11, 2025, 11:49 (UTC+05:30)

Elapsed time
14 minutes, 25 seconds

Instances (1/5) Info

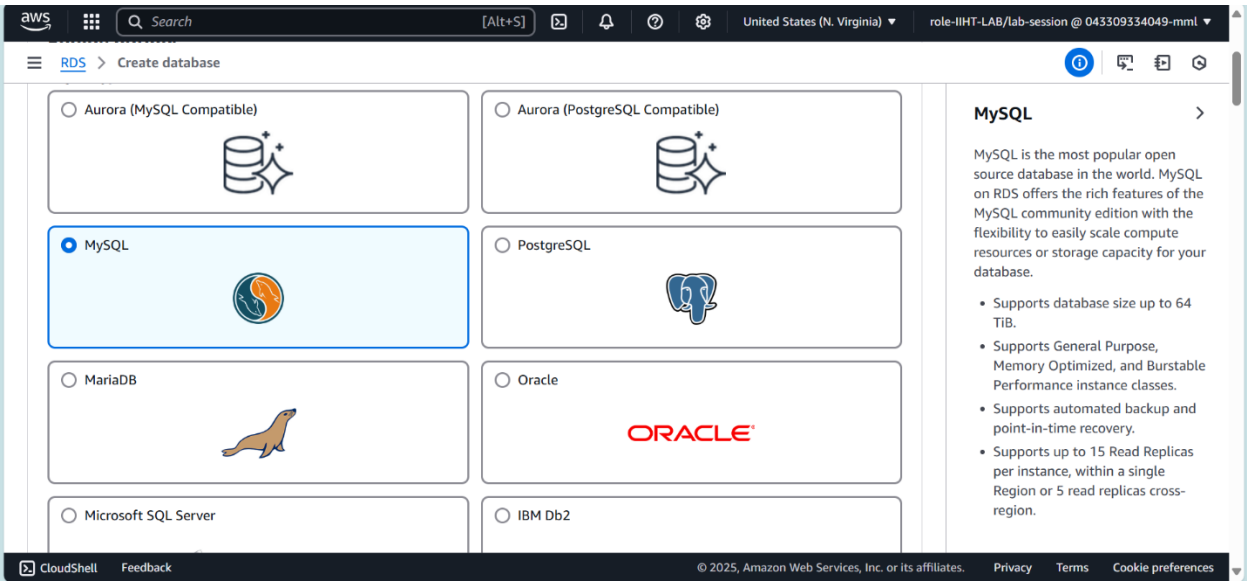
Last updated 1 minute ago | Connect | Instance state | Actions | Launch instances

Find Instance by attribute or tag (case-sensitive) | All states | 1

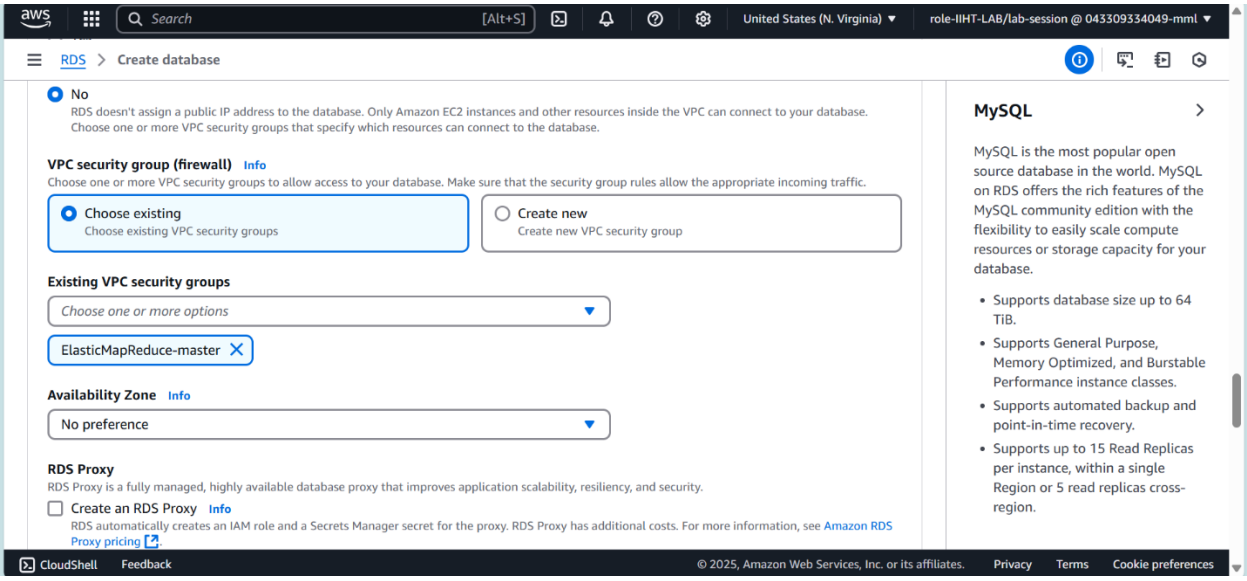
ic IPv4 DNS	Public IPv4 ...	Elastic IP	IPv6 IPs	Monitoring	Security group name
13-212-162-3.ap-s...	13.212.162.3	–	–	disabled	ElasticMapReduce-slave
13-214-39-185.ap-...	13.214.39.185	–	–	disabled	ElasticMapReduce-slave
13-215-60-115.ap-...	13.215.60.115	–	–	disabled	ElasticMapReduce-master
	–	–	–	disabled	–

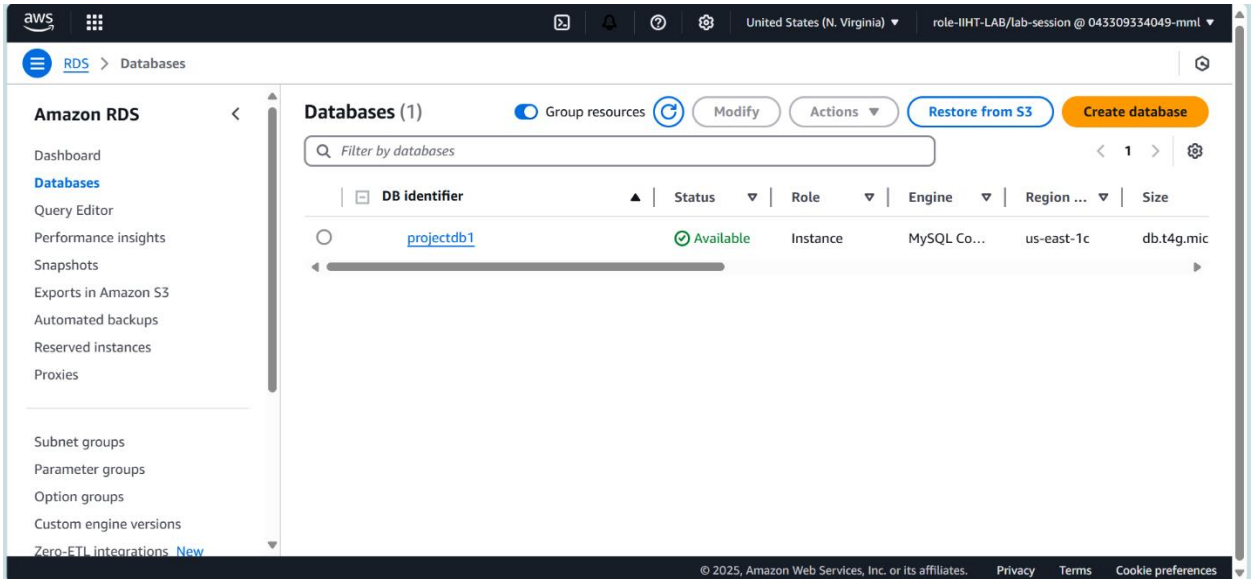
Step 3:

Create a database in RDS name projectdb1



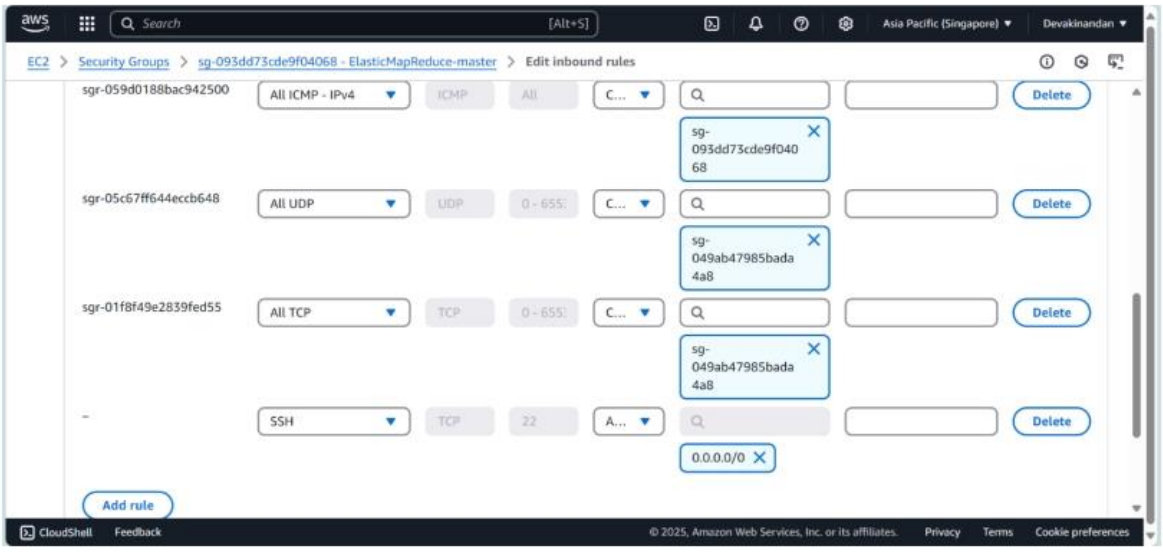
Choose VPC security group as Emr master.





Step 4:

Add ssh inbound rule to emr-master security groups



Step 5:

Open power shell and connect to emr using:

ssh -i key-path hadoop@emr-dns

```
PS C:\Users\2376779> ssh -i "C:\Users\2376779\Downloads\mynewkey.pem" hadoop@ec2-34-207-213-50.compute-1.amazonaws.com
The authenticity of host 'ec2-34-207-213-50.compute-1.amazonaws.com (34.207.213.50)' can't be established.
ED25519 key fingerprint is SHA256:U3oUFEzdZruOt+Fz+KXna5dzjRew+iQZbGQ67fIr4o4.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-34-207-213-50.compute-1.amazonaws.com' (ED25519) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
81 package(s) needed for security, out of 130 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE::::::EEEEEEEEE::E M::::::::M       M::::::::M R::::::RRRRRR::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR:::R      R::::R
  E::::E             M::::::M::::M   M::::M::::::M   R:::R      R::::R
  E::::::EEEEEEEEEE   M::::M M::::M M::::M M::::M   R:::RRRRRR::::R
  E::::::::::::::::E  M::::M  M::::M:::M  M::::M   R::::::::::RR
  E::::::EEEEEEEEEE   M::::M   M::::M   M::::M   R:::RRRRRR::::R
  E::::E             M::::M    M:::M    M::::M   R:::R      R::::R
  E::::E       EEEEE M::::M     MMM     M::::M   R:::R      R::::R
EE::::::EEEEEEEE::::E M::::M             M::::M   R:::R      R::::R
E::::::::::::::::::E M::::M             M::::M RR:::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMM             MMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-33-13 ~]$
```

Add mysql/aurora inbound rule to master security group.

Connect to mysql using mysql -h rds endpoint -u admin -p

```
Select hadoop@ip-172-31-33-13:~
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
81 package(s) needed for security, out of 130 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE::::::EEEEEEEEE::E M::::::::M       M::::::::M R::::::RRRRRR::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR:::R      R::::R
  E::::E             M::::::M::::M   M::::M::::::M   R:::R      R::::R
  E::::::EEEEEEEEEE   M::::M M::::M M::::M M::::M   R:::RRRRRR::::R
  E::::::::::::::::E  M::::M  M::::M:::M  M::::M   R::::::::::RR
  E::::::EEEEEEEEEE   M::::M   M::::M   M::::M   R:::RRRRRR::::R
  E::::E             M::::M    M:::M    M::::M   R:::R      R::::R
  E::::E       EEEEE M::::M     MMM     M::::M   R:::R      R::::R
EE::::::EEEEEEEE::::E M::::M             M::::M   R:::R      R::::R
E::::::::::::::::::E M::::M             M::::M RR:::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMM             MMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-33-13 ~]$ mysql -h projectdb.ct8ayu8km6es.us-east-1.rds.amazonaws.com -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 30
Server version: 8.0.40 Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]>
```

Insert data into mysql table

```
Select hadoop@ip-172-31-34-170:~

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sample             |
| sys                |
+--------------------+
5 rows in set (0.00 sec)

MySQL [(none)]> use sample;
Database changed
MySQL [sample]> create table cart_info(id int,cust_name varchar(20),prod_name varchar(20),price int,quantity int);
Query OK, 0 rows affected (0.04 sec)

MySQL [sample]> insert into cart_info(1,"nandu","soya sticks",10,4),(2,"anil","lays",20,3),(3,"vijay","soya sticks",10,4),(4,"harshith","kurkure",20,2);
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '1,"nandu","soya s
ticks",10,4),(2,"anil","lays",20,3),(3,"vijay","soya sticks",10' at line 1
MySQL [sample]> insert into cart_info values(1,"nandu","soya sticks",10,4),(2,"anil","lays",20,3),(3,"vijay","soya sticks",10,4),(4,"harshith","kurkure",20,2);
Query OK, 4 rows affected (0.00 sec)
Records: 4  Duplicates: 0  Warnings: 0

MySQL [sample]> select * from cart_info;
+------+-----------+------------+-------+----------+
| id   | cust_name | prod_name  | price | quantity |
+------+-----------+------------+-------+----------+
|    1 | nandu     | soya sticks |    10 |        4 |
|    2 | anil      | lays       |    20 |        3 |
|    3 | vijay     | soya sticks |    10 |        4 |
|    4 | harshith  | kurkure    |    20 |        2 |
+------+-----------+------------+-------+----------+
4 rows in set (0.00 sec)
```

And run Sqoop eval for checking jdbc connection

Sqoop eval --connect jdbc:mysql://projectdb1.cz6qqaw02gv7.ap-southeast-1.rds.amazonaws.com/sample

--query "select count(*) from sample.cart_info"

--username admin -P

```
Select hadoop@ip-172-31-46-158:~
MySQL [projectdatabase]> select * from cart_inf;
+------+-----------+------------+-------+----------+
| id   | cust_name | prod_name  | price | quantity |
+------+-----------+------------+-------+----------+
|    1 | nandu     | soya sticks |    10 |        4 |
|    2 | anil      | lays       |    20 |        2 |
|    3 | harshith  | kurkure    |    20 |        3 |
|    4 | vijay     | soya sticks |    10 |        2 |
+------+-----------+------------+-------+----------+
4 rows in set (0.00 sec)

MySQL [projectdatabase]>
[1]+  Stopped                 mysql -h projectdatabase.ct8ayu8km6es.us-east-1.rds.amazonaws.com -u admin -p
[hadoop@ip-172-31-46-158 ~]$ ACCULUMO_HOME='/var/lib/accumulo'
[hadoop@ip-172-31-46-158 ~]$ export ACCULULO_HOME
[hadoop@ip-172-31-46-158 ~]$ ACCUMULO_HOME='/var/lib/accumulo'
[hadoop@ip-172-31-46-158 ~]$ export ACCUMULO_HOME
[hadoop@ip-172-31-46-158 ~]$ sqoop eval --connect jdbc:mysql://projectdatabase.ct8ayu8km6es.us-east-1.rds.amazonaws.com/projectdatabase --query "select count(*) from projec
tdatabase.cart_inf" --username admin -P
Warning: /var/lib/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
25/02/07 05:32:22 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
25/02/07 05:33:08 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
------------------------
| count(*)             |
------------------------
| 4                    |
------------------------
[hadoop@ip-172-31-46-158 ~]$
```

Step 6 :

Create a database in hive

Using: create database hivedb;

Then Import data into hive using Sqoop import commands

Sqoop import-all-tables -m 1

--connect "jdbc:mysql://projectdb1.ct8ayu8km6es.us-east-1.rds.amazonaws.com/sample"

--username admin -P

--hive-database hivedb --create-hive-table

--hive-import --compression-codec=snappy --hive-overwrite

```
Select hadoop@ip-172-31-34-170:~
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
25/02/07 07:00:03 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
25/02/07 07:00:07 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
------------------------
| count(*)             |
------------------------
| 4                    |
------------------------
[hadoop@ip-172-31-34-170 ~]$ sqoop import-all-tables -m 1 --connect "jdbc:mysql://projectdb1.ct8ayu8km6es.us-east-1.rds.amazonaws.com/sample" --username admin -P --hive-dat
abase hivedb --create-hive-table --hive-import --compression-codec=snappy --hive-overwrite
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/redshift/jdbc/redshift-jdbc42-1.2.37.1061.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
25/02/07 07:05:33 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
25/02/07 07:05:40 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
25/02/07 07:05:40 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
25/02/07 07:05:40 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
25/02/07 07:05:41 INFO tool.CodeGenTool: Beginning code generation
25/02/07 07:05:41 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `cart_info` AS t LIMIT 1
25/02/07 07:05:41 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `cart_info` AS t LIMIT 1
25/02/07 07:05:41 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
/tmp/sqoop-hadoop/compile/3886a2005f3fdc2d96b061b61d7d44cc/cart_info.java:37: warning: Can't initialize javac processor due to (most likely) a class loader problem: java.la
ng.NoClassDefFoundError: com/sun/tools/javac/processing/JavacProcessingEnvironment
public class cart_info extends SqoopRecord  implements DBWritable, Writable {
       ^
        at lombok.javac.apt.LombokProcessor.getJavacProcessingEnvironment(LombokProcessor.java:411)
        at lombok.javac.apt.LombokProcessor.init(LombokProcessor.java:91)
        at lombok.core.AnnotationProcessor$JavacDescriptor.want(AnnotationProcessor.java:124)
        at lombok.core.AnnotationProcessor.init(AnnotationProcessor.java:177)
```

```
Select hadoop@ip-172-31-34-170:~
25/02/07 07:06:07 INFO ql.Driver: Concurrency mode is disabled, not creating a lock manager
25/02/07 07:06:07 INFO ql.Driver: Executing command(queryId=hadoop_20250207070607_9f82ac22-c9f1-43a7-9600-ac3e49a6193e):
LOAD DATA INPATH 'hdfs://ip-172-31-34-170.ec2.internal:8020/user/hadoop/cart_info' OVERWRITE INTO TABLE `hivedb`.`cart_info`
25/02/07 07:06:07 INFO ql.Driver: Starting task [Stage-0:MOVE] in serial mode
25/02/07 07:06:07 INFO hive.metastore: Closed a connection to metastore, current connections: 0
Loading data to table hivedb.cart_info
25/02/07 07:06:07 INFO exec.Task: Loading data to table hivedb.cart_info from hdfs://ip-172-31-34-170.ec2.internal:8020/user/hadoop/cart_info
25/02/07 07:06:07 INFO hive.metastore: Trying to connect to metastore with URI thrift://ip-172-31-34-170.ec2.internal:9083
25/02/07 07:06:07 INFO hive.metastore: Opened a connection to metastore, current connections: 1
25/02/07 07:06:07 INFO hive.metastore: Connected to metastore.
25/02/07 07:06:07 INFO common.FileUtils: Creating directory if it doesn't exist: hdfs://ip-172-31-34-170.ec2.internal:8020/user/hive/warehouse/hivedb.db/cart_info
25/02/07 07:06:07 INFO ql.Driver: Starting task [Stage-1:STATS] in serial mode
25/02/07 07:06:07 INFO exec.StatsTask: Executing stats task
25/02/07 07:06:07 INFO hive.metastore: Closed a connection to metastore, current connections: 0
25/02/07 07:06:07 INFO hive.metastore: Trying to connect to metastore with URI thrift://ip-172-31-34-170.ec2.internal:9083
25/02/07 07:06:07 INFO hive.metastore: Opened a connection to metastore, current connections: 1
25/02/07 07:06:07 INFO hive.metastore: Connected to metastore.
25/02/07 07:06:07 INFO hive.metastore: Closed a connection to metastore, current connections: 0
25/02/07 07:06:07 INFO hive.metastore: Trying to connect to metastore with URI thrift://ip-172-31-34-170.ec2.internal:9083
25/02/07 07:06:07 INFO hive.metastore: Opened a connection to metastore, current connections: 1
25/02/07 07:06:07 INFO hive.metastore: Connected to metastore.
25/02/07 07:06:07 INFO exec.StatsTask: Table hivedb.cart_info stats: [numFiles=1, numRows=0, totalSize=87, rawDataSize=0]
25/02/07 07:06:07 INFO ql.Driver: Completed executing command(queryId=hadoop_20250207070607_9f82ac22-c9f1-43a7-9600-ac3e49a6193e); Time taken: 0.475 seconds
OK
25/02/07 07:06:07 INFO ql.Driver: OK
Time taken: 0.661 seconds
25/02/07 07:06:07 INFO CliDriver: Time taken: 0.661 seconds
25/02/07 07:06:07 INFO conf.HiveConf: Using the default value passed in for log id: a64c3a1c-094c-4723-a0aa-a7266dfc6587
25/02/07 07:06:07 INFO session.SessionState: Resetting thread name to  main
25/02/07 07:06:07 INFO conf.HiveConf: Using the default value passed in for log id: a64c3a1c-094c-4723-a0aa-a7266dfc6587
25/02/07 07:06:07 INFO tez.TezSessionPoolManager: Closing tez session if not default: sessionId=a64c3a1c-094c-4723-a0aa-a7266dfc6587, queueName=null, user=hadoop, doAs=true
, isOpen=false, isDefault=false
25/02/07 07:06:07 INFO client.TezClient: Shutting down Tez Session, sessionName=HIVE-a64c3a1c-094c-4723-a0aa-a7266dfc6587, applicationId=application_1738910851147_0003
25/02/07 07:06:07 INFO client.TezClient: Could not connect to AM, killing session via YARN, sessionName=HIVE-a64c3a1c-094c-4723-a0aa-a7266dfc6587, applicationId=application
_1738910851147_0003
25/02/07 07:06:07 INFO session.SessionState: Deleted directory: /tmp/hive/hadoop/a64c3a1c-094c-4723-a0aa-a7266dfc6587 on fs with scheme hdfs
25/02/07 07:06:07 INFO session.SessionState: Deleted directory: /tmp/hadoop/a64c3a1c-094c-4723-a0aa-a7266dfc6587 on fs with scheme file
25/02/07 07:06:07 INFO hive.metastore: Closed a connection to metastore, current connections: 0
25/02/07 07:06:07 INFO hive.HiveImport: Hive import complete.
[hadoop@ip-172-31-34-170 ~]$
```

Step 7:

Check whether it is imported or not into hive.

```
Select hadoop@ip-172-31-34-170:~
OK
25/02/07 07:06:07 INFO ql.Driver: OK
Time taken: 0.661 seconds
25/02/07 07:06:07 INFO CliDriver: Time taken: 0.661 seconds
25/02/07 07:06:07 INFO conf.HiveConf: Using the default value passed in for log id: a64c3a1c-094c-4723-a0aa-a7266dfc6587
25/02/07 07:06:07 INFO session.SessionState: Resetting thread name to  main
25/02/07 07:06:07 INFO conf.HiveConf: Using the default value passed in for log id: a64c3a1c-094c-4723-a0aa-a7266dfc6587
25/02/07 07:06:07 INFO tez.TezSessionPoolManager: Closing tez session if not default: sessionId=a64c3a1c-094c-4723-a0aa-a7266dfc6587, queueName=null, user=hadoop, doAs=true
, isOpen=false, isDefault=false
25/02/07 07:06:07 INFO client.TezClient: Shutting down Tez Session, sessionName=HIVE-a64c3a1c-094c-4723-a0aa-a7266dfc6587, applicationId=application_1738910851147_0003
25/02/07 07:06:07 INFO client.TezClient: Could not connect to AM, killing session via YARN, sessionName=HIVE-a64c3a1c-094c-4723-a0aa-a7266dfc6587, applicationId=application
_1738910851147_0003
25/02/07 07:06:07 INFO session.SessionState: Deleted directory: /tmp/hive/hadoop/a64c3a1c-094c-4723-a0aa-a7266dfc6587 on fs with scheme hdfs
25/02/07 07:06:07 INFO session.SessionState: Deleted directory: /tmp/hadoop/a64c3a1c-094c-4723-a0aa-a7266dfc6587 on fs with scheme file
25/02/07 07:06:07 INFO hive.metastore: Closed a connection to metastore, current connections: 0
25/02/07 07:06:07 INFO hive.HiveImport: Hive import complete.
[hadoop@ip-172-31-34-170 ~]$ hive;

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases;
OK
default
hivedb
Time taken: 0.596 seconds, Fetched: 2 row(s)
hive> use hivedb;
OK
Time taken: 0.032 seconds
hive> show tables;
OK
cart_info
Time taken: 0.041 seconds, Fetched: 1 row(s)
hive> select * from cart_info;
OK
1       nandu   soya sticks     10      4
2       anil    lays    20      3
3       vijay   soya sticks     10      4
4       harshith        kurkure 20      2
Time taken: 2.853 seconds, Fetched: 4 row(s)
hive>
```

Step 8:

Now create a bucket and dump the data

```
Select hadoop@ip-172-31-34-170:~
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:244)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:158)
FAILED: ParseException line 5:0 mismatched input 'price' expecting ) near 'string' in create table statement
hive> CREATE TABLE bucketed_shoping_info (
    >     order_id INT,
    >     customer_name String,
    > product_name string,
    > price int,
    > quantity int
    > )
    > CLUSTERED BY (order_id) INTO 4 BUCKETS;
OK
Time taken: 0.422 seconds
hive> insert into bucketed_shoping_info (select * from cart_info);
Query ID = hadoop_20250207072255_3f2e5073-922d-4fd2-a682-af20bf9ffd37
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738910851147_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.53 s
--------------------------------------------------------------------------------
Loading data to table hivedb.bucketed_shoping_info
OK
Time taken: 6.51 seconds
hive>
```

Step 9:

Analyzing the data:

1.product name and the no of quantities sold.

```
Select hadoop@ip-172-31-34-170:~
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.56 s
--------------------------------------------------------------------------------
OK
1
1
2
Time taken: 4.389 seconds, Fetched: 3 row(s)
hive> select product_name,count(sum(quantity)) from cart_info group by prod_name;
FAILED: SemanticException [Error 10128]: Line 1:26 Not yet supported place for UDAF 'sum'
hive> select product_name,sum(quantity) from cart_info group by prod_name;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'product_name'
hive> select prod_name,sum(quantity) from cart_info group by prod_name;
Query ID = hadoop_20250207072634_b4070218-2cf1-45e9-aa7f-96eedc4decbf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738910851147_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.53 s
--------------------------------------------------------------------------------
OK
kurkure 2
lays    3
soya sticks     8
Time taken: 3.992 seconds, Fetched: 3 row(s)
hive>
```

2.Find the average price of products sold:

```
Select hadoop@ip-172-31-47-29:~
    > ORDER BY total_sales DESC
    > LIMIT 3;
Query ID = hadoop_20250207120652_6e9d391d-159a-4bf6-b509-1aa52aacd663
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738920108837_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 4.24 s
--------------------------------------------------------------------------------
OK
anil    60
harshith        40
vijay   40
Time taken: 7.286 seconds, Fetched: 3 row(s)
hive> SELECT AVG(price) AS average_price
    > FROM cart_info;
Query ID = hadoop_20250207120908_b76b0d3b-6cfc-4e2a-aaf8-9f36156ef4e1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738920108837_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.60 s
--------------------------------------------------------------------------------
OK
15.0
Time taken: 4.691 seconds, Fetched: 1 row(s)
hive>
```

3.Identify the top 3 customers by total sales amount:

```
Select hadoop@ip-172-31-47-29:~
  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 MMMMMM             MMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-47-29 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> use hivedb
    > ;
OK
Time taken: 0.406 seconds
hive> SELECT cust_name, SUM(price * quantity) AS total_sales
    > FROM cart_info
    > GROUP BY cust_name
    > ORDER BY total_sales DESC
    > LIMIT 3;
Query ID = hadoop_20250207120652_6e9d391d-159a-4bf6-b509-1aa52aacd663
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738920108837_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 4.24 s
--------------------------------------------------------------------------------
OK
anil    60
harshith        40
vijay   40
Time taken: 7.286 seconds, Fetched: 3 row(s)
hive> SELECT AVG(price) AS average_price
```

4.List products that have been sold more than once:

```
Select hadoop@ip-172-31-47-29:~
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738920108837_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.60 s
--------------------------------------------------------------------------------
OK
15.0
Time taken: 4.691 seconds, Fetched: 1 row(s)
hive> SELECT AVG(price) AS average_price
    > ;
FAILED: SemanticException [Error 10004]: Line 1:11 Invalid table alias or column reference 'price': (possible column names are: )
hive> SELECT prod_name, COUNT(*) AS times_sold
    > FROM cart_info
    > GROUP BY prod_name
    > HAVING times_sold > 1;
Query ID = hadoop_20250207121027_bb15ffef-18e2-4424-8891-31f710fd1bf0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1738920108837_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 3.67 s
--------------------------------------------------------------------------------
OK
soya sticks     2
Time taken: 4.144 seconds, Fetched: 1 row(s)
hive>
```