MINI PROJECT

20PITE54J- Big Data for Machine Learning

Implement the Hive and Sqoop framework for the following scenarios

Online Shopping System- Identify the no of products available in the portal, segregate them according to the purpose

Instructions

- Definitely students should not be combined, everybody have to do projects individually
- All scenarios can be related to word count program
- Students themselves have to create the sample datasets according to the scenarios what they have chosen as mentioned above
- Give the created datasets as input to the Hive and Sqoop framework
- Create tables in mysql, project it by using hive and perform query in sqoop using mysql for the above scenarios
- Manually provide the result by performing in the document, it can be either pen and paper or printed document
- Follow the timespan given to complete the project
- Each student can take any one scenario from the above scenarios as their project
- Students should submit the screenshot for their work along with the report– can be included at the last of the report
- Please submit the project within the deadline

Online Shopping System

Implementation of Hive and Sqoop

1. Log in to MySQL

mysql -u root -p

password: cloudera

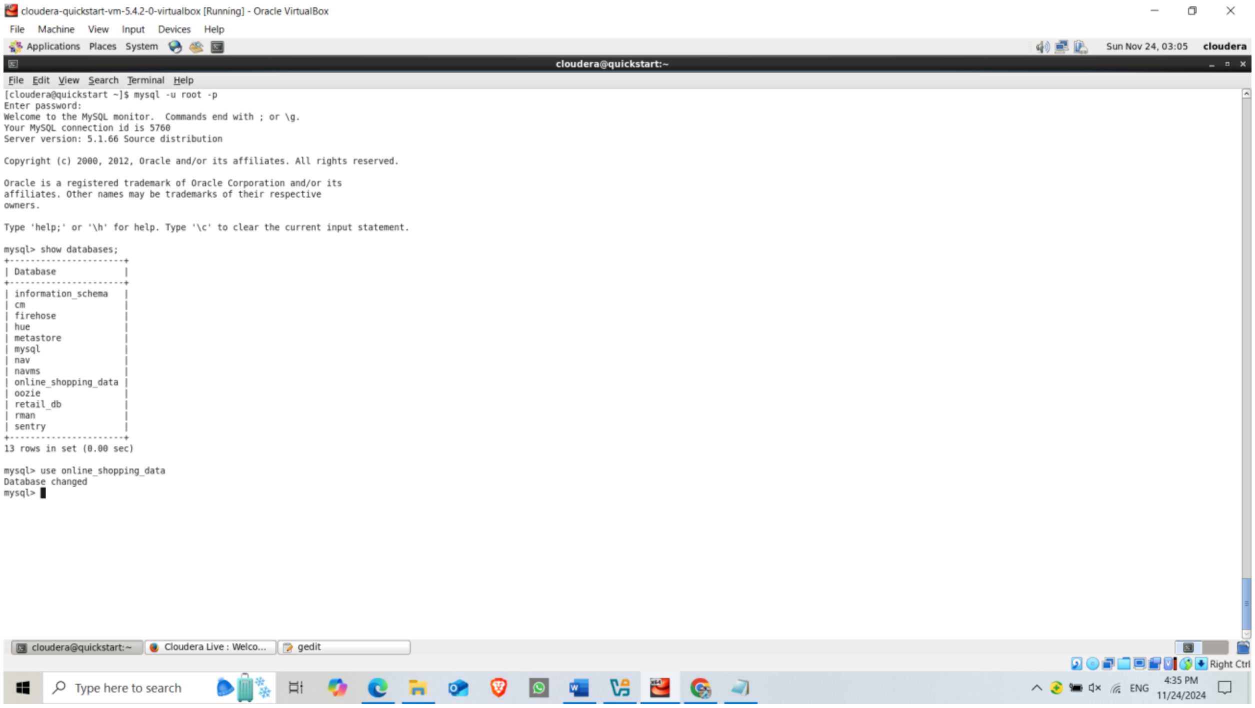
2. Create DataBase

CREATE DATABASE online_shopping_data;

Show databases;

USE online_shopping_data;

Output:



cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 03:05 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
Cloudera Live : Welco...
gedit
Right Ctrl
Type here to search
ENG
4:35 PM
11/24/2024

```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 5760
Server version: 5.1.66 Source distribution

Copyright (c) 2000, 2012, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

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

mysql> show databases;
+----------------------+
| Database             |
+----------------------+
| information_schema   |
| cm                   |
| firehose             |
| hue                  |
| metastore            |
| mysql                |
| nav                  |
| navms                |
| online_shopping_data |
| oozie                |
| retail_db            |
| rman                 |
| sentry               |
+----------------------+
13 rows in set (0.00 sec)

mysql> use online_shopping_data
Database changed
mysql> 
```

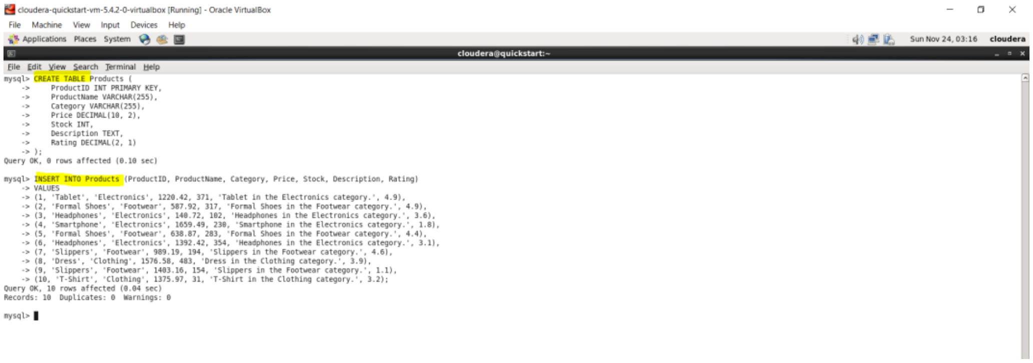
3. Create the products table and insert the records in it.

```
CREATE TABLE Products (
    ProductID INT PRIMARY KEY,
    ProductName VARCHAR(255),
    Category VARCHAR(255),
    Price DECIMAL(10, 2),
    Stock INT,
    Description TEXT,
    Rating DECIMAL(2, 1));
```

Insert the sample data:

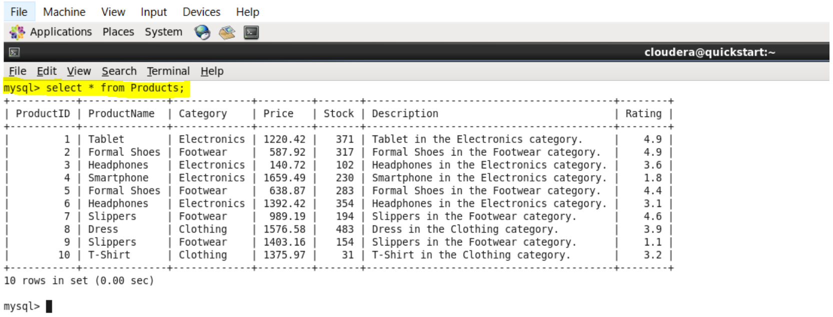
```
INSERT INTO Products (ProductID, ProductName, Category, Price, Stock, Description, Rating)
VALUES
(1, 'Tablet', 'Electronics', 1220.42, 371, 'Tablet in the Electronics category.', 4.9),
(2, 'Formal Shoes', 'Footwear', 587.92, 317, 'Formal Shoes in the Footwear category.', 4.9),
(3, 'Headphones', 'Electronics', 140.72, 102, 'Headphones in the Electronics category.', 3.6),
(4, 'Smartphone', 'Electronics', 1659.49, 230, 'Smartphone in the Electronics category.', 1.8),
(5, 'Formal Shoes', 'Footwear', 638.87, 283, 'Formal Shoes in the Footwear category.', 4.4),
(6, 'Headphones', 'Electronics', 1392.42, 354, 'Headphones in the Electronics category.', 3.1),
(7, 'Slippers', 'Footwear', 989.19, 194, 'Slippers in the Footwear category.', 4.6),
(8, 'Dress', 'Clothing', 1576.58, 483, 'Dress in the Clothing category.', 3.9),
(9, 'Slippers', 'Footwear', 1403.16, 154, 'Slippers in the Footwear category.', 1.1),
(10, 'T-Shirt', 'Clothing', 1375.97, 31, 'T-Shirt in the Clothing category.', 3.2);
```

Output:



cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 03:16 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

```
mysql> CREATE TABLE Products (
    ->     ProductID INT PRIMARY KEY,
    ->     ProductName VARCHAR(255),
    ->     Category VARCHAR(255),
    ->     Price DECIMAL(10, 2),
    ->     Stock INT,
    ->     Description TEXT,
    ->     Rating DECIMAL(2, 1)
    -> );
Query OK, 0 rows affected (0.10 sec)

mysql> INSERT INTO Products (ProductID, ProductName, Category, Price, Stock, Description, Rating)
    -> VALUES
    -> (1, 'Tablet', 'Electronics', 1220.42, 371, 'Tablet in the Electronics category.', 4.9),
    -> (2, 'Formal Shoes', 'Footwear', 587.92, 317, 'Formal Shoes in the Footwear category.', 4.9),
    -> (3, 'Headphones', 'Electronics', 140.72, 102, 'Headphones in the Electronics category.', 3.6),
    -> (4, 'Smartphone', 'Electronics', 1659.49, 230, 'Smartphone in the Electronics category.', 1.8),
    -> (5, 'Formal Shoes', 'Footwear', 638.87, 283, 'Formal Shoes in the Footwear category.', 4.4),
    -> (6, 'Headphones', 'Electronics', 1392.42, 354, 'Headphones in the Electronics category.', 3.1),
    -> (7, 'Slippers', 'Footwear', 989.19, 194, 'Slippers in the Footwear category.', 4.6),
    -> (8, 'Dress', 'Clothing', 1576.58, 483, 'Dress in the Clothing category.', 3.9),
    -> (9, 'Slippers', 'Footwear', 1403.16, 154, 'Slippers in the Footwear category.', 1.1),
    -> (10, 'T-Shirt', 'Clothing', 1375.97, 31, 'T-Shirt in the Clothing category.', 3.2);
Query OK, 10 rows affected (0.04 sec)
Records: 10  Duplicates: 0  Warnings: 0

mysql> 
```


File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

```
mysql> select * from Products;
+-----------+--------------+-------------+---------+-------+----------------------------------------+--------+
| ProductID | ProductName  | Category    | Price   | Stock | Description                            | Rating |
+-----------+--------------+-------------+---------+-------+----------------------------------------+--------+
|         1 | Tablet       | Electronics | 1220.42 |   371 | Tablet in the Electronics category.    |    4.9 |
|         2 | Formal Shoes | Footwear    |  587.92 |   317 | Formal Shoes in the Footwear category. |    4.9 |
|         3 | Headphones   | Electronics |  140.72 |   102 | Headphones in the Electronics category. |    3.6 |
|         4 | Smartphone   | Electronics | 1659.49 |   230 | Smartphone in the Electronics category. |    1.8 |
|         5 | Formal Shoes | Footwear    |  638.87 |   283 | Formal Shoes in the Footwear category. |    4.4 |
|         6 | Headphones   | Electronics | 1392.42 |   354 | Headphones in the Electronics category. |    3.1 |
|         7 | Slippers     | Footwear    |  989.19 |   194 | Slippers in the Footwear category.     |    4.6 |
|         8 | Dress        | Clothing    | 1576.58 |   483 | Dress in the Clothing category.        |    3.9 |
|         9 | Slippers     | Footwear    | 1403.16 |   154 | Slippers in the Footwear category.     |    1.1 |
|        10 | T-Shirt      | Clothing    | 1375.97 |    31 | T-Shirt in the Clothing category.      |    3.2 |
+-----------+--------------+-------------+---------+-------+----------------------------------------+--------+
10 rows in set (0.00 sec)

mysql> 
```

4. Import MySQL Table into Hive Using Sqoop

sqoop import \

--connect jdbc:mysql://quickstart:3306/online_shopping_data \

--username=root \

--password=cloudera \

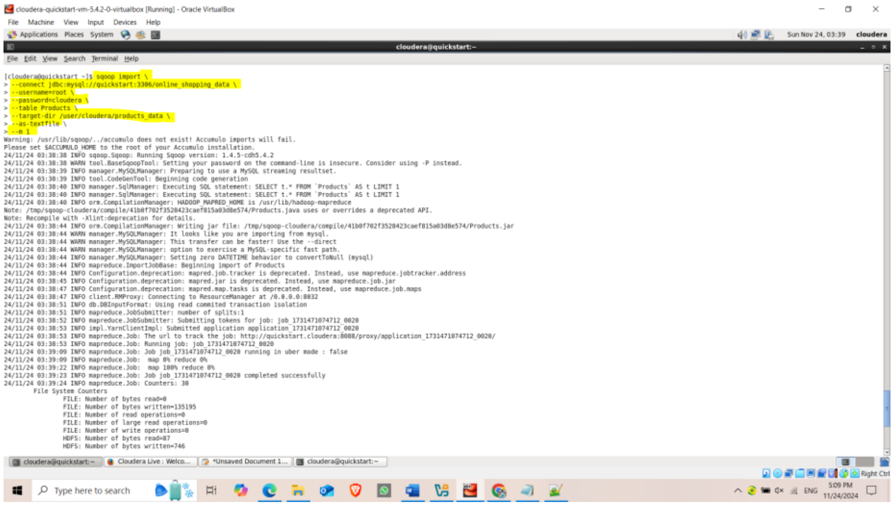
--table Products \

--target-dir /user/cloudera/products_data \

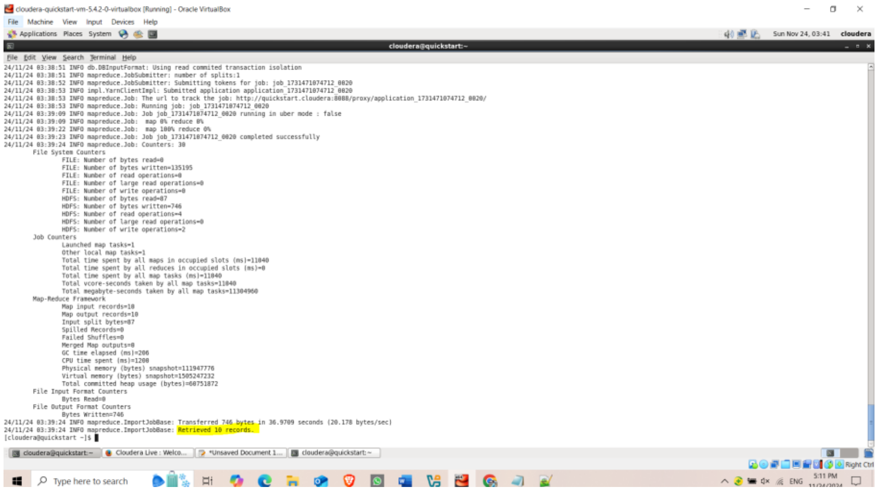
--as-textfile \

--m 1

Output:


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 03:39 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
Cloudera Live : Welco...
*Unsaved Document 1...
cloudera@quickstart:~
Type here to search
Right Ctrl
ENG
5:09 PM
11/24/2024

```
[cloudera@quickstart ~]$ sqoop import \
> --connect jdbc:mysql://quickstart:3306/online_shopping_data \
> --username=root \
> --password=cloudera \
> --table Products \
> --target-dir /user/cloudera/products_data \
> --as-textfile \
> --m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
24/11/24 03:38:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.5-cdh5.4.2
24/11/24 03:38:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
24/11/24 03:38:39 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
24/11/24 03:38:39 INFO tool.CodeGenTool: Beginning code generation
24/11/24 03:38:40 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Products` AS t LIMIT 1
24/11/24 03:38:40 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Products` AS t LIMIT 1
24/11/24 03:38:40 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/41b0f702f3528423caef815a03d8e574/Products.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
24/11/24 03:38:44 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/41b0f702f3528423caef815a03d8e574/Products.jar
24/11/24 03:38:44 WARN manager.MySQLManager: It looks like you are importing from mysql.
24/11/24 03:38:44 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
24/11/24 03:38:44 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
24/11/24 03:38:44 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
24/11/24 03:38:44 INFO mapreduce.ImportJobBase: Beginning import of Products
24/11/24 03:38:44 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
24/11/24 03:38:45 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
24/11/24 03:38:47 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
24/11/24 03:38:47 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
24/11/24 03:38:51 INFO db.DBInputFormat: Using read commited transaction isolation
24/11/24 03:38:51 INFO mapreduce.JobSubmitter: number of splits:1
24/11/24 03:38:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731471074712_0020
24/11/24 03:38:53 INFO impl.YarnClientImpl: Submitted application application_1731471074712_0020
24/11/24 03:38:53 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1731471074712_0020/
24/11/24 03:38:53 INFO mapreduce.Job: Running job: job_1731471074712_0020
24/11/24 03:39:09 INFO mapreduce.Job: Job job_1731471074712_0020 running in uber mode : false
24/11/24 03:39:09 INFO mapreduce.Job:  map 0% reduce 0%
24/11/24 03:39:22 INFO mapreduce.Job:  map 100% reduce 0%
24/11/24 03:39:23 INFO mapreduce.Job: Job job_1731471074712_0020 completed successfully
24/11/24 03:39:24 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=135195
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=746
```



```
24/11/24 03:38:51 INFO db.DBInputFormat: Using read commited transaction isolation
24/11/24 03:38:51 INFO mapreduce.JobSubmitter: number of splits:1
24/11/24 03:38:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1731471074712_0020
24/11/24 03:38:53 INFO impl.YarnClientImpl: Submitted application application_1731471074712_0020
24/11/24 03:38:53 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1731471074712_0020/
24/11/24 03:38:53 INFO mapreduce.Job: Running job: job_1731471074712_0020
24/11/24 03:39:09 INFO mapreduce.Job: Job job_1731471074712_0020 running in uber mode : false
24/11/24 03:39:09 INFO mapreduce.Job:  map 0% reduce 0%
24/11/24 03:39:22 INFO mapreduce.Job:  map 100% reduce 0%
24/11/24 03:39:23 INFO mapreduce.Job: Job job_1731471074712_0020 completed successfully
24/11/24 03:39:24 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=135195
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=746
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=11040
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=11040
                Total vcore-seconds taken by all map tasks=11040
                Total megabyte-seconds taken by all map tasks=11304960
        Map-Reduce Framework
                Map input records=10
                Map output records=10
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=206
                CPU time spent (ms)=1200
                Physical memory (bytes) snapshot=111947776
                Virtual memory (bytes) snapshot=1505247232
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=746
24/11/24 03:39:24 INFO mapreduce.ImportJobBase: Transferred 746 bytes in 36.9709 seconds (20.178 bytes/sec)
24/11/24 03:39:24 INFO mapreduce.ImportJobBase: Retrieved 10 records.
[cloudera@quickstart ~]$ 
```

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 03:41 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
Cloudera Live : Welco...
*Unsaved Document 1...
cloudera@quickstart:~
Right Ctrl
Type here to search
ENG
5:11 PM

5. Create Hive Table and load data to table.

hive

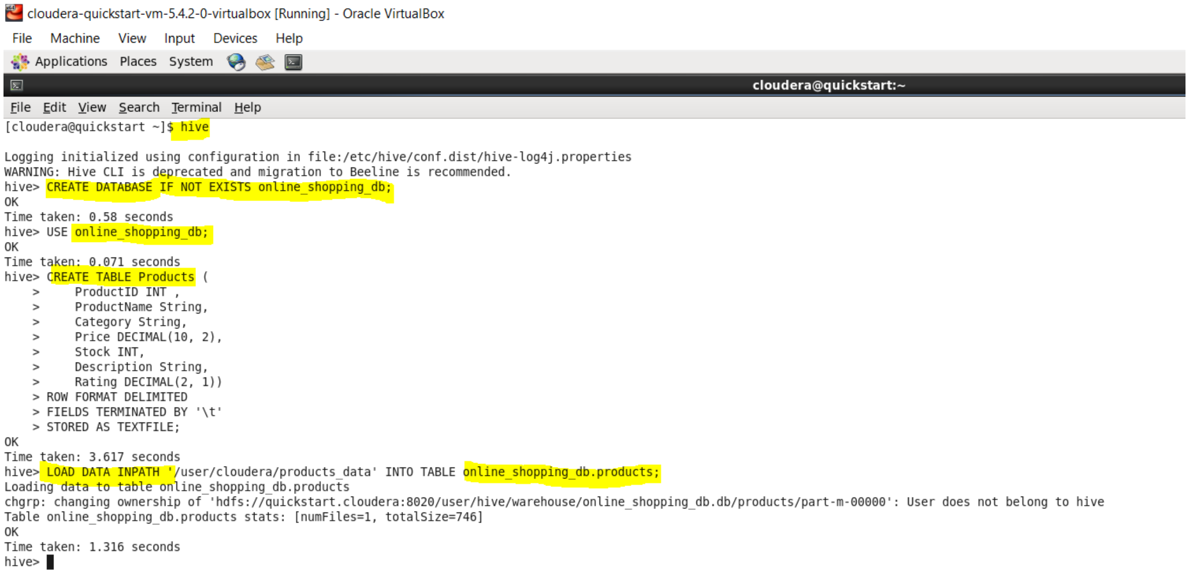
CREATE DATABASE IF NOT EXISTS online_shopping_db;

USE online_shopping_db;

```
CREATE TABLE Products (
    ProductID INT ,
    ProductName String,
    Category String,
    Price DECIMAL(10, 2),
    Stock INT,
    Description String,
    Rating DECIMAL(2, 1))
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

LOAD DATA INPATH '/user/cloudera/products_data' INTO TABLE online_shopping_db.products;

Output:


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

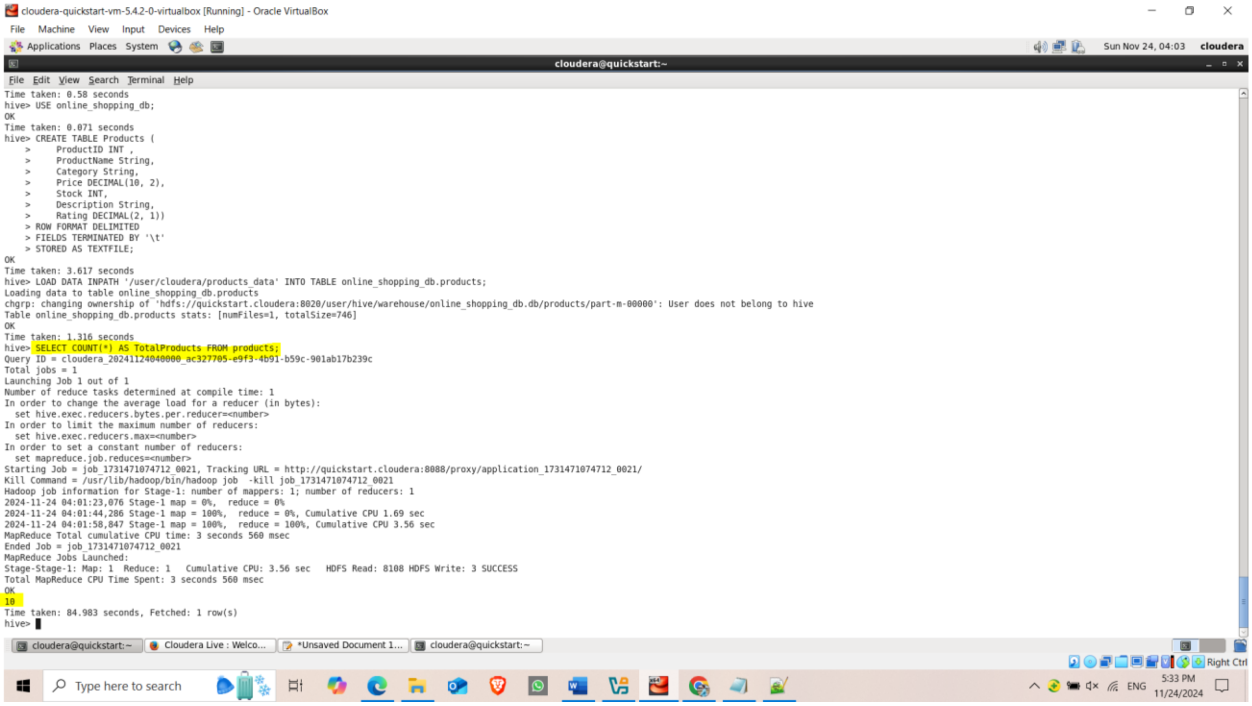
```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE DATABASE IF NOT EXISTS online_shopping_db;
OK
Time taken: 0.58 seconds
hive> USE online_shopping_db;
OK
Time taken: 0.071 seconds
hive> CREATE TABLE Products (
    >     ProductID INT ,
    >     ProductName String,
    >     Category String,
    >     Price DECIMAL(10, 2),
    >     Stock INT,
    >     Description String,
    >     Rating DECIMAL(2, 1))
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 3.617 seconds
hive> LOAD DATA INPATH '/user/cloudera/products_data' INTO TABLE online_shopping_db.products;
Loading data to table online_shopping_db.products
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/online_shopping_db.db/products/part-m-00000': User does not belong to hive
Table online_shopping_db.products stats: [numFiles=1, totalSize=746]
OK
Time taken: 1.316 seconds
hive> 
```

6. Perform Queries in Hive

✓ Total Number of Products :

SELECT COUNT(*) AS TotalProducts FROM products;

Output:



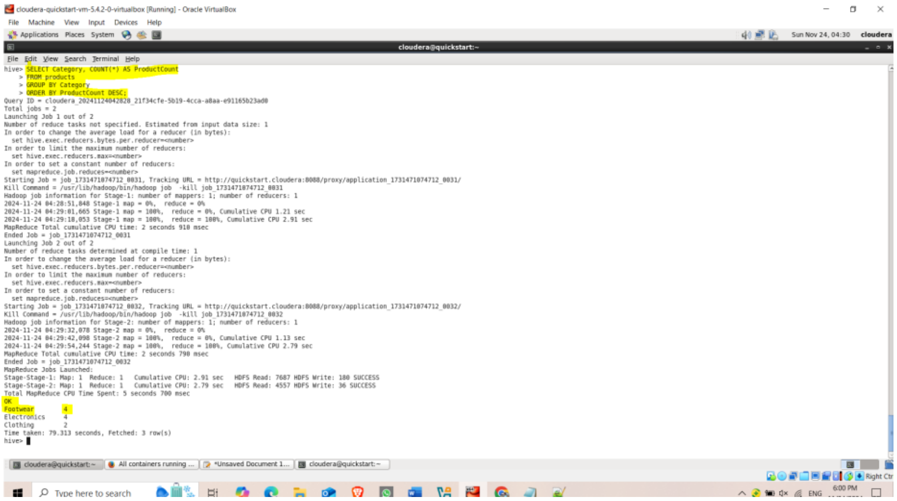
```
Time taken: 0.58 seconds
hive> USE online_shopping_db;
OK
Time taken: 0.071 seconds
hive> CREATE TABLE Products (
    >     ProductID INT ,
    >     ProductName String,
    >     Category String,
    >     Price DECIMAL(10, 2),
    >     Stock INT,
    >     Description String,
    >     Rating DECIMAL(2, 1))
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 3.617 seconds
hive> LOAD DATA INPATH '/user/cloudera/products_data' INTO TABLE online_shopping_db.products;
Loading data to table online_shopping_db.products
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/online_shopping_db.db/products/part-m-00000': User does not belong to hive
Table online_shopping_db.products stats: [numFiles=1, totalSize=746]
OK
Time taken: 1.316 seconds
hive> SELECT COUNT(*) AS TotalProducts FROM products;
Query ID = cloudera_20241124040000_ac327705-e9f3-4b91-b59c-901ab17b239c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0021, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 04:01:23,076 Stage-1 map = 0%,  reduce = 0%
2024-11-24 04:01:44,286 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.69 sec
2024-11-24 04:01:58,847 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.56 sec
MapReduce Total cumulative CPU time: 3 seconds 560 msec
Ended Job = job_1731471074712_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.56 sec   HDFS Read: 8108 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 560 msec
OK
10
Time taken: 84.983 seconds, Fetched: 1 row(s)
hive> 
```

✓ Segregate Products by Category:

SELECT Category, COUNT(*) AS ProductCount FROM products

GROUP BY Category ;

Output:



```
hive> SELECT Category, COUNT(*) AS ProductCount
    > FROM products
    > GROUP BY Category
    > ORDER BY ProductCount DESC;
Query ID = cloudera_20241124042828_21f34cfe-5b19-4cca-a8aa-e91165b23ad0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0031, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0031/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0031
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 04:28:51,848 Stage-1 map = 0%,  reduce = 0%
2024-11-24 04:29:01,665 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.21 sec
2024-11-24 04:29:10,053 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.91 sec
MapReduce Total cumulative CPU time: 2 seconds 910 msec
Ended Job = job_1731471074712_0031
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0032, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0032/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0032
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-11-24 04:29:32,078 Stage-2 map = 0%,  reduce = 0%
2024-11-24 04:29:42,098 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.13 sec
2024-11-24 04:29:54,244 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.79 sec
MapReduce Total cumulative CPU time: 2 seconds 790 msec
Ended Job = job_1731471074712_0032
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.91 sec   HDFS Read: 7687 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.79 sec   HDFS Read: 4557 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 700 msec
OK
Footwear        4
Electronics     4
Clothing        2
Time taken: 79.313 seconds, Fetched: 3 row(s)
hive> 
```

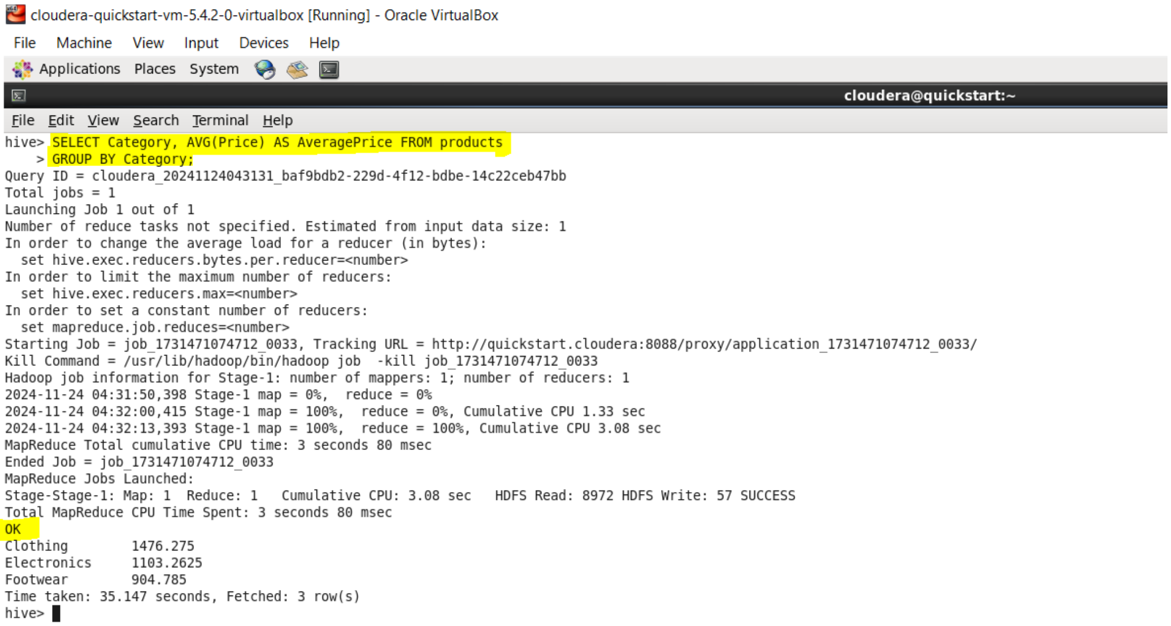
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 04:30
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
All containers running ...
*Unsaved Document 1...
cloudera@quickstart:~
Type here to search
ENG
6:00 PM

✓ Average Price by Category

SELECT Category, AVG(Price) AS AveragePrice FROM products

GROUP BY Category;

Output:



```
hive> SELECT Category, AVG(Price) AS AveragePrice FROM products
    > GROUP BY Category;
Query ID = cloudera_20241124043131_baf9bdb2-229d-4f12-bdbe-14c22ceb47bb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0033, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0033/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0033
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 04:31:50,398 Stage-1 map = 0%,  reduce = 0%
2024-11-24 04:32:00,415 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.33 sec
2024-11-24 04:32:13,393 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.08 sec
MapReduce Total cumulative CPU time: 3 seconds 80 msec
Ended Job = job_1731471074712_0033
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.08 sec   HDFS Read: 8972 HDFS Write: 57 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 80 msec
OK
Clothing        1476.275
Electronics     1103.2625
Footwear        904.785
Time taken: 35.147 seconds, Fetched: 3 row(s)
hive> 
```

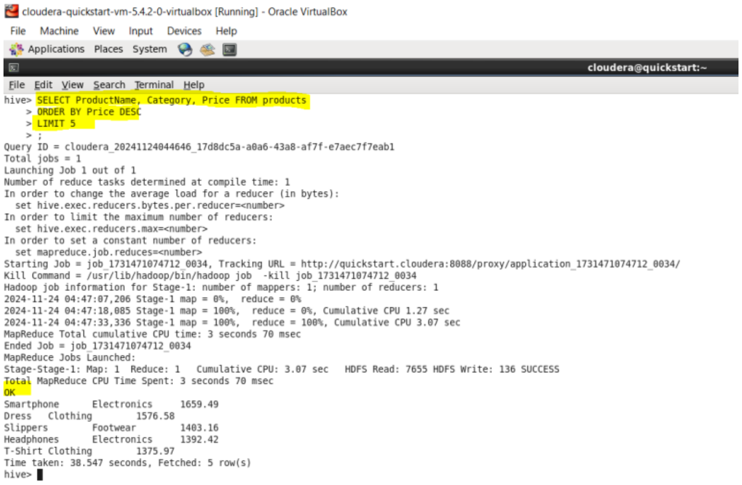
cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

- ✓ Top 5 Most Expensive Products

SELECT ProductName, Category, Price FROM products

ORDER BY Price DESC LIMIT 5;

Output


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

```
hive> SELECT ProductName, Category, Price FROM products
    > ORDER BY Price DESC
    > LIMIT 5
    > ;
Query ID = cloudera_20241124044646_17d8dc5a-a0a6-43a8-af7f-e7aec7f7eab1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0034, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0034/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0034
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 04:47:07,206 Stage-1 map = 0%,  reduce = 0%
2024-11-24 04:47:18,085 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.27 sec
2024-11-24 04:47:33,336 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.07 sec
MapReduce Total cumulative CPU time: 3 seconds 70 msec
Ended Job = job_1731471074712_0034
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.07 sec   HDFS Read: 7655 HDFS Write: 136 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 70 msec
OK
Smartphone      Electronics     1659.49
Dress   Clothing        1576.58
Slippers        Footwear        1403.16
Headphones      Electronics     1392.42
T-Shirt Clothing        1375.97
Time taken: 38.547 seconds, Fetched: 5 row(s)
hive>
```

- ✓ Count Products by Purpose with Total Stock

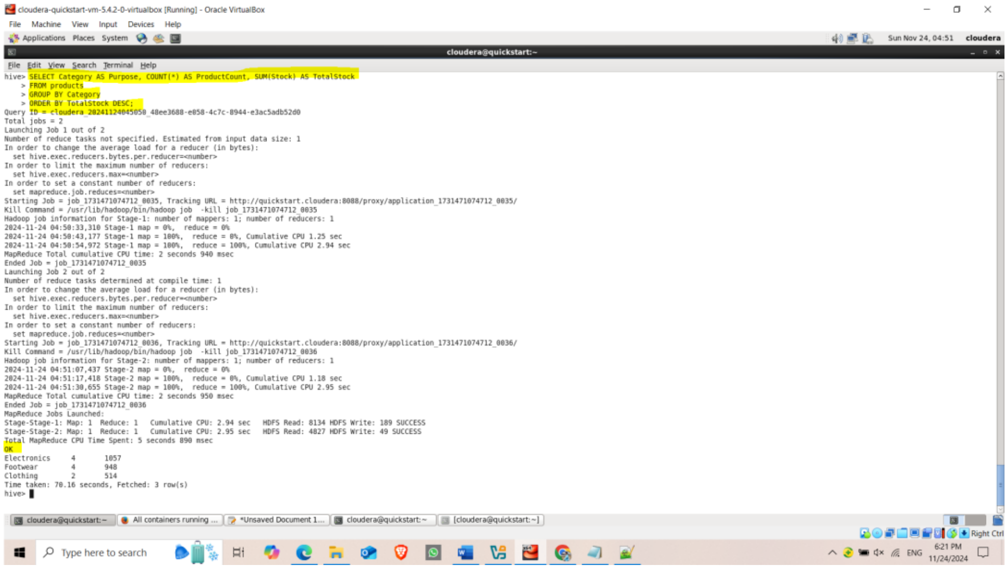
SELECT Category AS Purpose, COUNT(*) AS ProductCount, SUM(Stock) AS TotalStock

FROM products

GROUP BY Category

ORDER BY TotalStock DESC;

Output:


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 04:51
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
All containers running ...
*Unsaved Document 1...
cloudera@quickstart:~
[cloudera@quickstart:~]
Right Ctrl
Type here to search
ENG
6:21 PM
11/24/2024

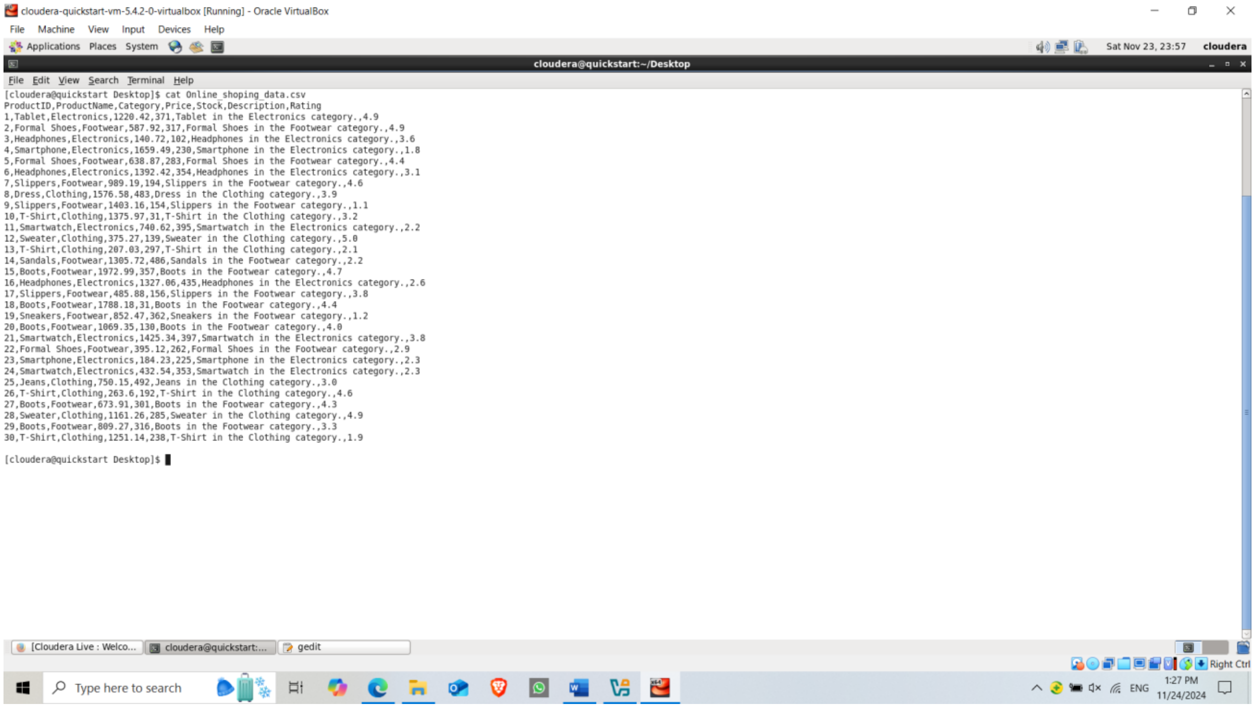
```
hive> SELECT Category AS Purpose, COUNT(*) AS ProductCount, SUM(Stock) AS TotalStock
    > FROM products
    > GROUP BY Category
    > ORDER BY TotalStock DESC;
Query ID = cloudera_20241124045050_48ee3688-e858-4c7c-8944-e3ac5adb52d0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0035, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0035/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0035
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 04:50:33,310 Stage-1 map = 0%,  reduce = 0%
2024-11-24 04:50:43,177 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.25 sec
2024-11-24 04:50:54,972 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.94 sec
MapReduce Total cumulative CPU time: 2 seconds 940 msec
Ended Job = job_1731471074712_0035
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0036, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0036/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0036
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-11-24 04:51:07,437 Stage-2 map = 0%,  reduce = 0%
2024-11-24 04:51:17,418 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.18 sec
2024-11-24 04:51:30,655 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.95 sec
MapReduce Total cumulative CPU time: 2 seconds 950 msec
Ended Job = job_1731471074712_0036
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.94 sec   HDFS Read: 8134 HDFS Write: 189 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.95 sec   HDFS Read: 4827 HDFS Write: 49 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 890 msec
OK
Electronics     4       1057
Footwear        4       948
Clothing        2       514
Time taken: 70.16 seconds, Fetched: 3 row(s)
hive>
```

Implementation of Hive

Create our Own data set for Hive and Sqoop frame work.

Dataset :

Online_shoping_data.csv



```
[cloudera@quickstart Desktop]$ cat Online_shoping_data.csv
ProductID,ProductName,Category,Price,Stock,Description,Rating
1,Tablet,Electronics,1220.42,371,Tablet in the Electronics category.,4.9
2,Formal Shoes,Footwear,587.92,317,Formal Shoes in the Footwear category.,4.9
3,Headphones,Electronics,140.72,102,Headphones in the Electronics category.,3.6
4,Smartphone,Electronics,1659.49,230,Smartphone in the Electronics category.,1.8
5,Formal Shoes,Footwear,638.87,283,Formal Shoes in the Footwear category.,4.4
6,Headphones,Electronics,1392.42,354,Headphones in the Electronics category.,3.1
7,Slippers,Footwear,989.19,194,Slippers in the Footwear category.,4.6
8,Dress,Clothing,1576.58,483,Dress in the Clothing category.,3.9
9,Slippers,Footwear,1403.16,154,Slippers in the Footwear category.,1.1
10,T-Shirt,Clothing,1375.97,31,T-Shirt in the Clothing category.,3.2
11,Smartwatch,Electronics,740.62,395,Smartwatch in the Electronics category.,2.2
12,Sweater,Clothing,375.27,139,Sweater in the Clothing category.,5.0
13,T-Shirt,Clothing,207.03,297,T-Shirt in the Clothing category.,2.1
14,Sandals,Footwear,1305.72,486,Sandals in the Footwear category.,2.2
15,Boots,Footwear,1972.99,357,Boots in the Footwear category.,4.7
16,Headphones,Electronics,1327.06,435,Headphones in the Electronics category.,2.6
17,Slippers,Footwear,485.88,156,Slippers in the Footwear category.,3.8
18,Boots,Footwear,1788.18,31,Boots in the Footwear category.,4.4
19,Sneakers,Footwear,852.47,362,Sneakers in the Footwear category.,1.2
20,Boots,Footwear,1069.35,130,Boots in the Footwear category.,4.0
21,Smartwatch,Electronics,1425.34,397,Smartwatch in the Electronics category.,3.8
22,Formal Shoes,Footwear,395.12,262,Formal Shoes in the Footwear category.,2.9
23,Smartphone,Electronics,184.23,225,Smartphone in the Electronics category.,2.3
24,Smartwatch,Electronics,432.54,353,Smartwatch in the Electronics category.,2.3
25,Jeans,Clothing,750.15,492,Jeans in the Clothing category.,3.0
26,T-Shirt,Clothing,263.6,192,T-Shirt in the Clothing category.,4.6
27,Boots,Footwear,673.91,301,Boots in the Footwear category.,4.3
28,Sweater,Clothing,1161.26,285,Sweater in the Clothing category.,4.9
29,Boots,Footwear,809.27,316,Boots in the Footwear category.,3.3
30,T-Shirt,Clothing,1251.14,238,T-Shirt in the Clothing category.,1.9

[cloudera@quickstart Desktop]$ 
```

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Nov 23, 23:57 cloudera
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[Cloudera Live : Welco...
cloudera@quickstart:...
gedit
Right Ctrl
Type here to search
ENG
1:27 PM
11/24/2024

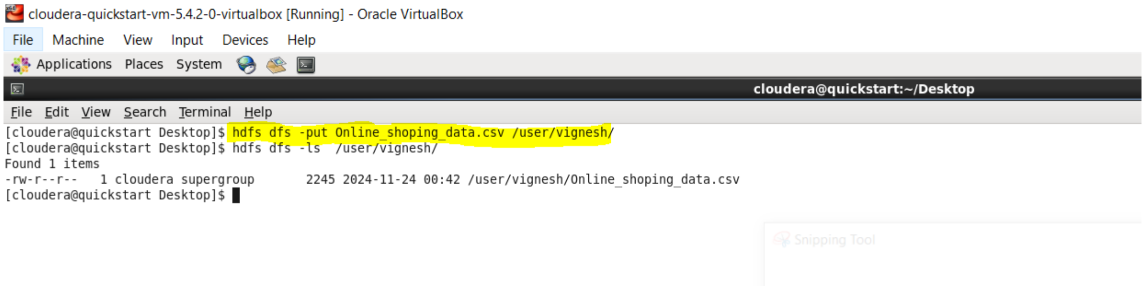
Load the data to Hadoop :

Commands

hdfs dfs -mkdir /user/vignesh

hdfs dfs -put Online_Shoping_Data.csv /user/vignesh/

hdfs dfs -ls /user/vignesh/



```
[cloudera@quickstart Desktop]$ hdfs dfs -put Online_shoping_data.csv /user/vignesh/
[cloudera@quickstart Desktop]$ hdfs dfs -ls  /user/vignesh/
Found 1 items
-rw-r--r--   1 cloudera supergroup       2245 2024-11-24 00:42 /user/vignesh/Online_shoping_data.csv
[cloudera@quickstart Desktop]$ 
```

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Snipping Tool

Implementation of Hive

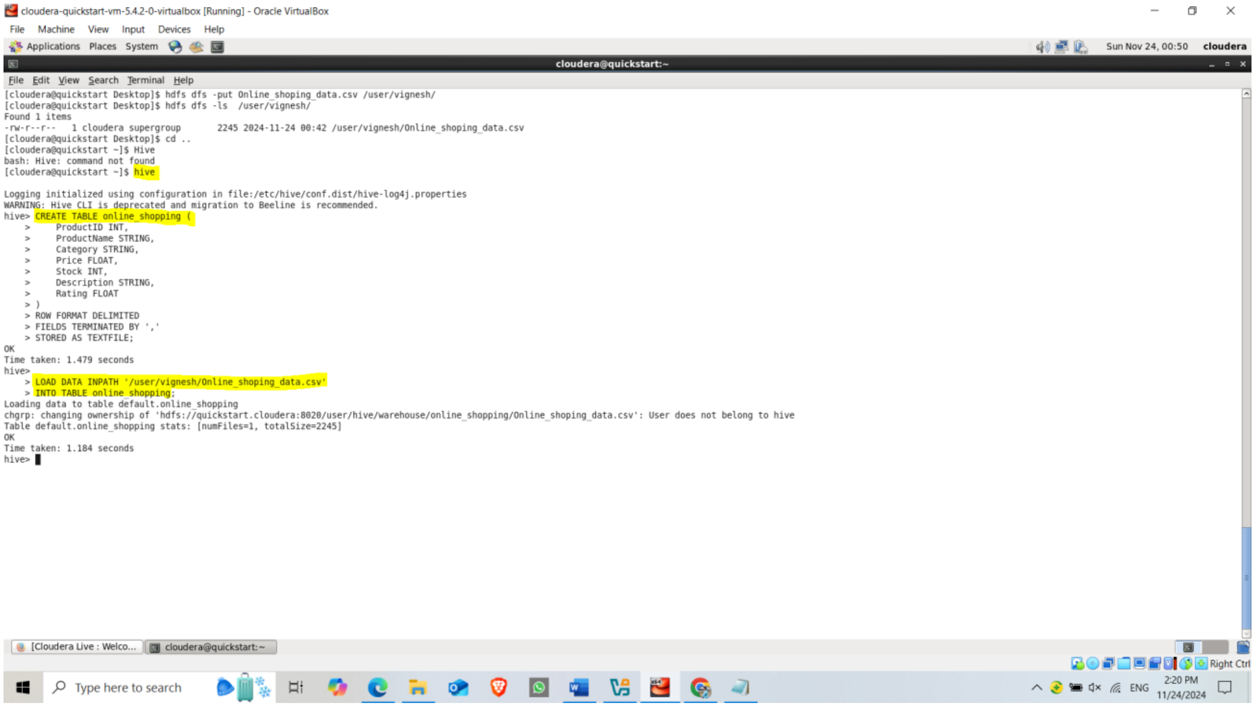
1. **Define the schema in Hive.**

```
CREATE TABLE online_shopping (
   ProductID INT,
   ProductName STRING,
   Category STRING,
   Price FLOAT,
   Stock INT,
   Description STRING,
   Rating FLOAT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

2. **Load the data into Hive**

```
LOAD DATA INPATH '/user/vignesh/Online_shoping_data.csv'
INTO TABLE online_shopping;
```

Output:


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 00:50 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
[Cloudera Live : Welco...
cloudera@quickstart:~
Right Ctrl
Type here to search
ENG
2:20 PM
11/24/2024

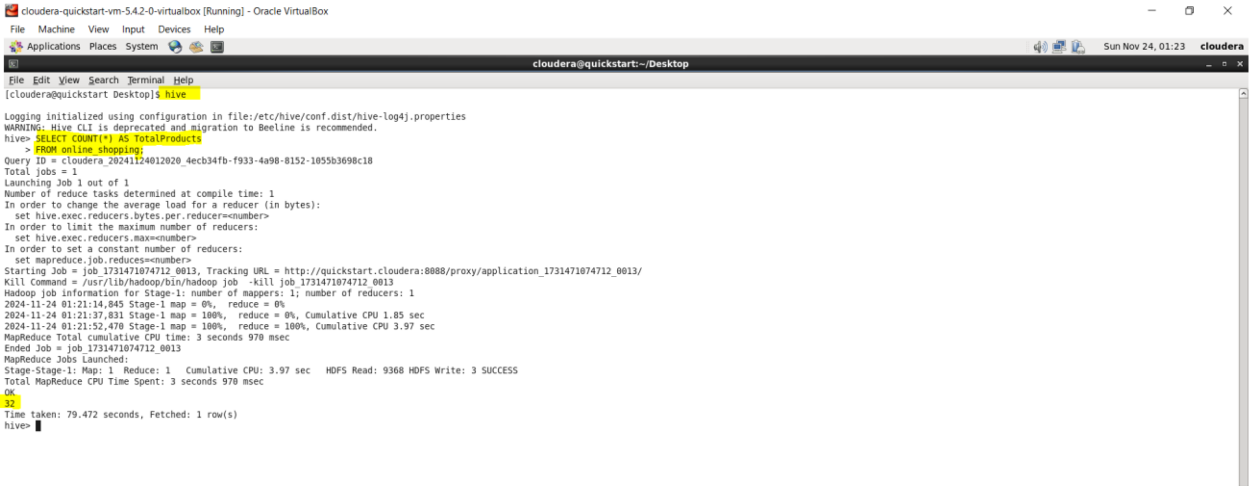
```
[cloudera@quickstart Desktop]$ hdfs dfs -put Online_shoping_data.csv /user/vignesh/
[cloudera@quickstart Desktop]$ hdfs dfs -ls  /user/vignesh/
Found 1 items
-rw-r--r--   1 cloudera supergroup       2245 2024-11-24 00:42 /user/vignesh/Online_shoping_data.csv
[cloudera@quickstart Desktop]$ cd ..
[cloudera@quickstart ~]$ Hive
bash: Hive: command not found
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE TABLE online_shopping (
    >     ProductID INT,
    >     ProductName STRING,
    >     Category STRING,
    >     Price FLOAT,
    >     Stock INT,
    >     Description STRING,
    >     Rating FLOAT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 1.479 seconds
hive>
    > LOAD DATA INPATH '/user/vignesh/Online_shoping_data.csv'
    > INTO TABLE online_shopping;
Loading data to table default.online_shopping
chgrp: changing ownership of 'hdfs://quickstart.cloudera:8020/user/hive/warehouse/online_shopping/Online_shoping_data.csv': User does not belong to hive
Table default.online_shopping stats: [numFiles=1, totalSize=2245]
OK
Time taken: 1.184 seconds
hive> 
```

3. Query to perform the operations in Hive:

✓ Total Number of Products :

SELECT COUNT(*) AS TotalProducts FROM online_shopping;

Output:


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 01:23 cloudera
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

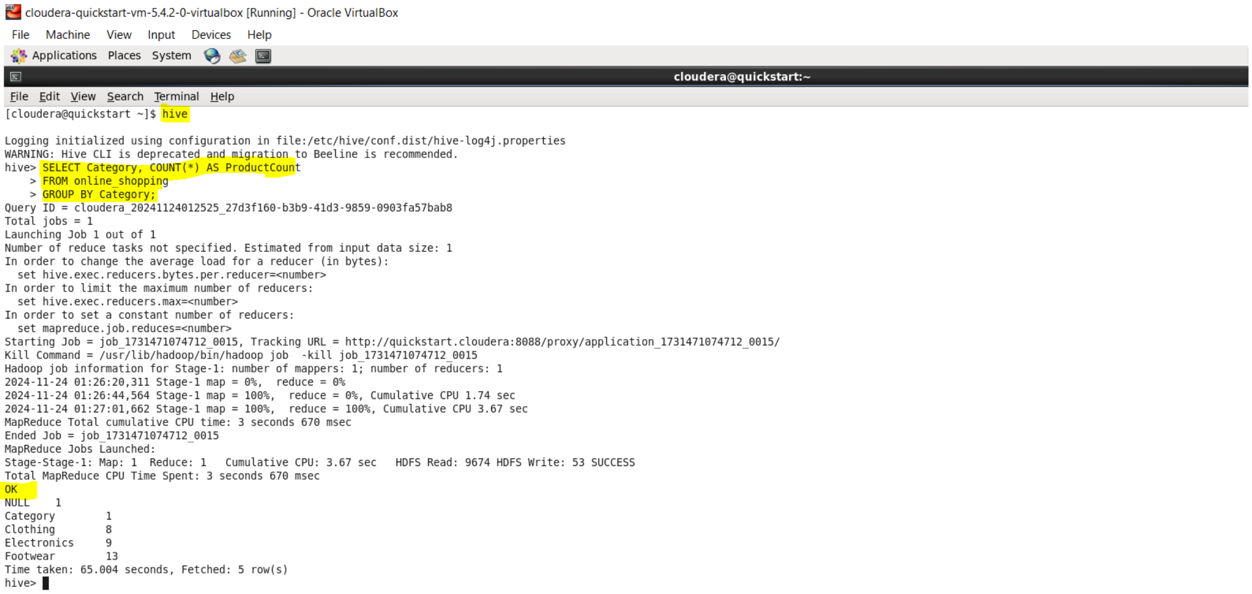
```
[cloudera@quickstart Desktop]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> SELECT COUNT(*) AS TotalProducts
    > FROM online_shopping;
Query ID = cloudera_20241124012020_4ecb34fb-f933-4a98-8152-1055b3698c18
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 01:21:14,845 Stage-1 map = 0%,  reduce = 0%
2024-11-24 01:21:37,831 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.85 sec
2024-11-24 01:21:52,470 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.97 sec
MapReduce Total cumulative CPU time: 3 seconds 970 msec
Ended Job = job_1731471074712_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.97 sec   HDFS Read: 9368 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 970 msec
OK
32
Time taken: 79.472 seconds, Fetched: 1 row(s)
hive> 
```

- ✓ Products Segregated by Category:

SELECT Category, COUNT(*) AS ProductCount FROM online_shopping

GROUP BY Category;

Output:



cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

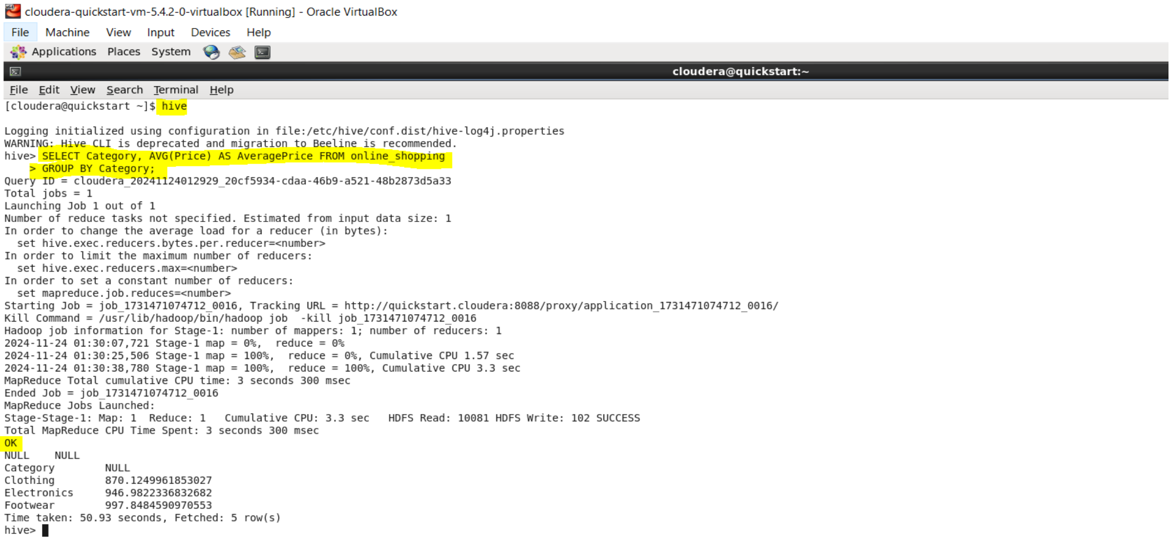
```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> SELECT Category, COUNT(*) AS ProductCount
    > FROM online_shopping
    > GROUP BY Category;
Query ID = cloudera_20241124012525_27d3f160-b3b9-41d3-9859-0903fa57bab8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 01:26:20,311 Stage-1 map = 0%,  reduce = 0%
2024-11-24 01:26:44,564 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.74 sec
2024-11-24 01:27:01,662 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.67 sec
MapReduce Total cumulative CPU time: 3 seconds 670 msec
Ended Job = job_1731471074712_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.67 sec   HDFS Read: 9674 HDFS Write: 53 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 670 msec
OK
NULL    1
Category        1
Clothing        8
Electronics     9
Footwear        13
Time taken: 65.004 seconds, Fetched: 5 row(s)
hive>
```

- ✓ Average Price by Category:

SELECT Category, AVG(Price) AS AveragePrice FROM online_shopping

GROUP BY Category;

Output:



cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> SELECT Category, AVG(Price) AS AveragePrice FROM online_shopping
    > GROUP BY Category;
Query ID = cloudera_20241124012929_20cf5934-cdaa-46b9-a521-48b2873d5a33
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 01:30:07,721 Stage-1 map = 0%,  reduce = 0%
2024-11-24 01:30:25,506 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.57 sec
2024-11-24 01:30:38,780 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.3 sec
MapReduce Total cumulative CPU time: 3 seconds 300 msec
Ended Job = job_1731471074712_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.3 sec   HDFS Read: 10081 HDFS Write: 102 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 300 msec
OK
NULL    NULL
Category        NULL
Clothing        870.1249961853027
Electronics     946.9822336832682
Footwear        997.8484590970553
Time taken: 50.93 seconds, Fetched: 5 row(s)
hive>
```

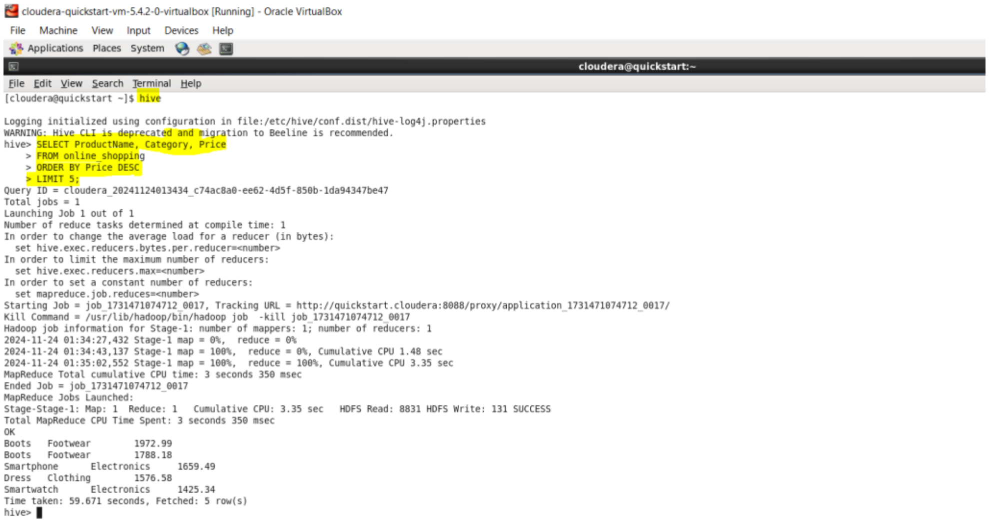
- ✓ Top 5 Most Expensive Products

SELECT ProductName, Category, Price FROM online_shopping

ORDER BY Price DESC

LIMIT 5;

Output:



cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> SELECT ProductName, Category, Price
    > FROM online_shopping
    > ORDER BY Price DESC
    > LIMIT 5;
Query ID = cloudera_20241124013434_c74ac8a0-ee62-4d5f-850b-1da94347be47
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0017, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 01:34:27,432 Stage-1 map = 0%,  reduce = 0%
2024-11-24 01:34:43,137 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.48 sec
2024-11-24 01:35:02,552 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.35 sec
MapReduce Total cumulative CPU time: 3 seconds 350 msec
Ended Job = job_1731471074712_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.35 sec   HDFS Read: 8831 HDFS Write: 131 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 350 msec
OK
Boots   Footwear        1972.99
Boots   Footwear        1788.18
Smartphone      Electronics     1659.49
Dress   Clothing        1576.58
Smartwatch      Electronics     1425.34
Time taken: 59.671 seconds, Fetched: 5 row(s)
hive> 
```

- ✓ Count Products by Purpose with Total Stock

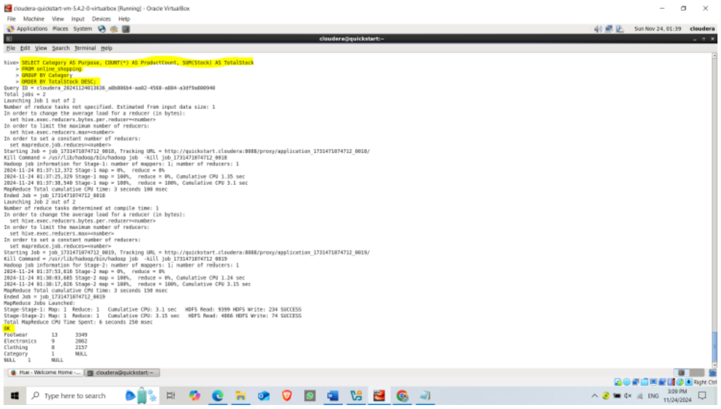
SELECT Category AS Purpose, COUNT(*) AS ProductCount, SUM(Stock) AS TotalStock

FROM online_shopping

GROUP BY Category

ORDER BY TotalStock DESC;

Output:



cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 24, 01:39
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Hue - Welcome Home -...
cloudera@quickstart:~
Type here to search
ENG
3:09 PM
11/24/2024

```
hive> SELECT Category AS Purpose, COUNT(*) AS ProductCount, SUM(Stock) AS TotalStock
    > FROM online_shopping
    > GROUP BY Category
    > ORDER BY TotalStock DESC;
Query ID = cloudera_20241124013636_a0b006b4-aa02-4568-a884-a3df9a800940
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0018, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0018
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-24 01:37:12,372 Stage-1 map = 0%,  reduce = 0%
2024-11-24 01:37:25,329 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.35 sec
2024-11-24 01:37:38,540 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
MapReduce Total cumulative CPU time: 3 seconds 100 msec
Ended Job = job_1731471074712_0018
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1731471074712_0019, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1731471074712_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1731471074712_0019
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-11-24 01:37:53,816 Stage-2 map = 0%,  reduce = 0%
2024-11-24 01:38:03,685 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.24 sec
2024-11-24 01:38:17,026 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.15 sec
MapReduce Total cumulative CPU time: 3 seconds 150 msec
Ended Job = job_1731471074712_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.1 sec   HDFS Read: 9399 HDFS Write: 234 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.15 sec   HDFS Read: 4866 HDFS Write: 74 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 250 msec
OK
Footwear        13      3349
Electronics     9       2862
Clothing        8       2157
Category        1       NULL
NULL    1       NULL
```

Result:

In this document, we demonstrated how to use Sqoop to import data from a relational database into HDFS and then use Hive to query the data for product count and segregation based on product categories (purpose). The solution provides the following insights:

- The total number of products available on the portal.
- The segregation of products into categories such as Electronics, Clothing, and Books.

This approach allows businesses to leverage the Hadoop ecosystem for scalable, efficient data processing, enabling insightful analysis of large datasets from their online shopping system.