**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

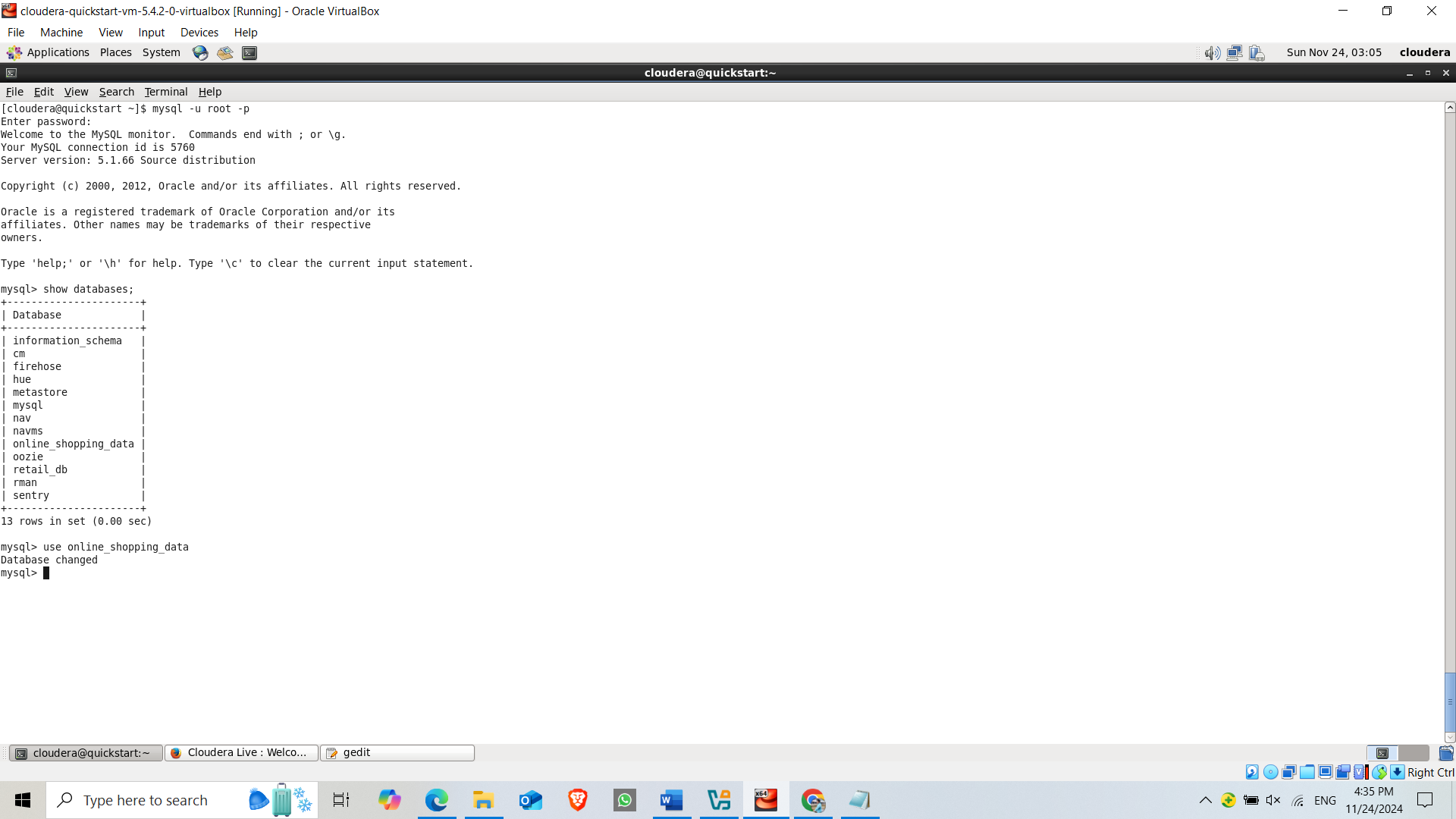
1. **Create DataBase**

CREATE DATABASE online\_shopping\_data;

Show databases;

USE online\_shopping\_data;

**Output:**



1. **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),

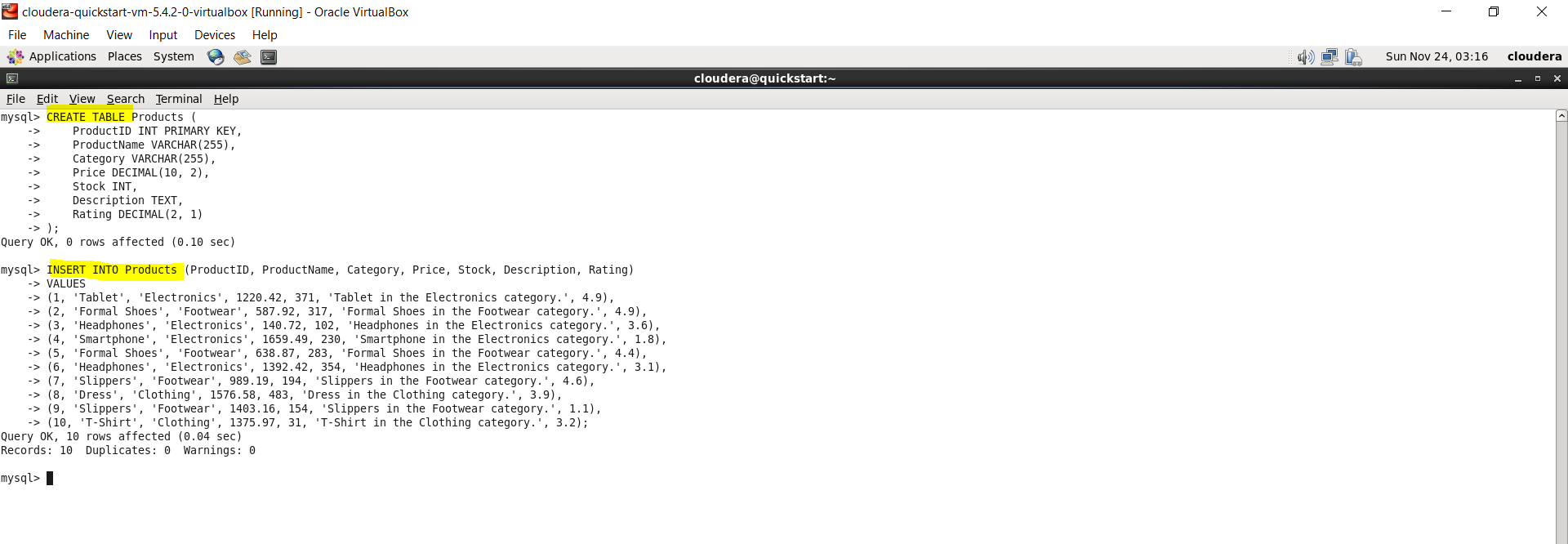
(7, 'Slippers', 'Footwear', 989.19, 194, 'Slippers in the Footwear category.', 4.6),

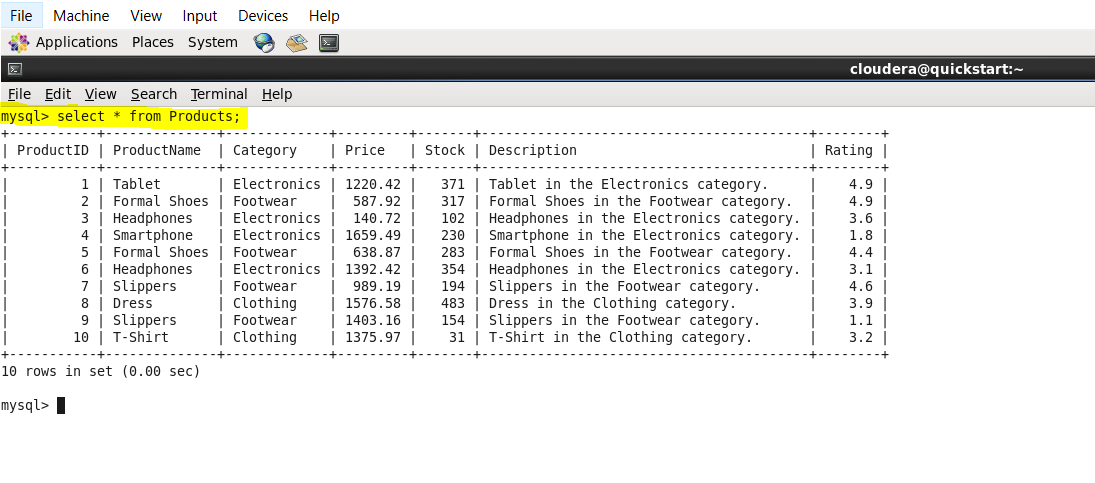
(8, 'Dress', 'Clothing', 1576.58, 483, 'Dress in the Clothing category.', 3.9),

(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:**





1. **Import MySQL Table into Hive Using Sqoop**

sqoop import \

--connect jdbc:mysql://quickstart:3306/online\_shopping\_data \

--username=root \

--password=cloudera \

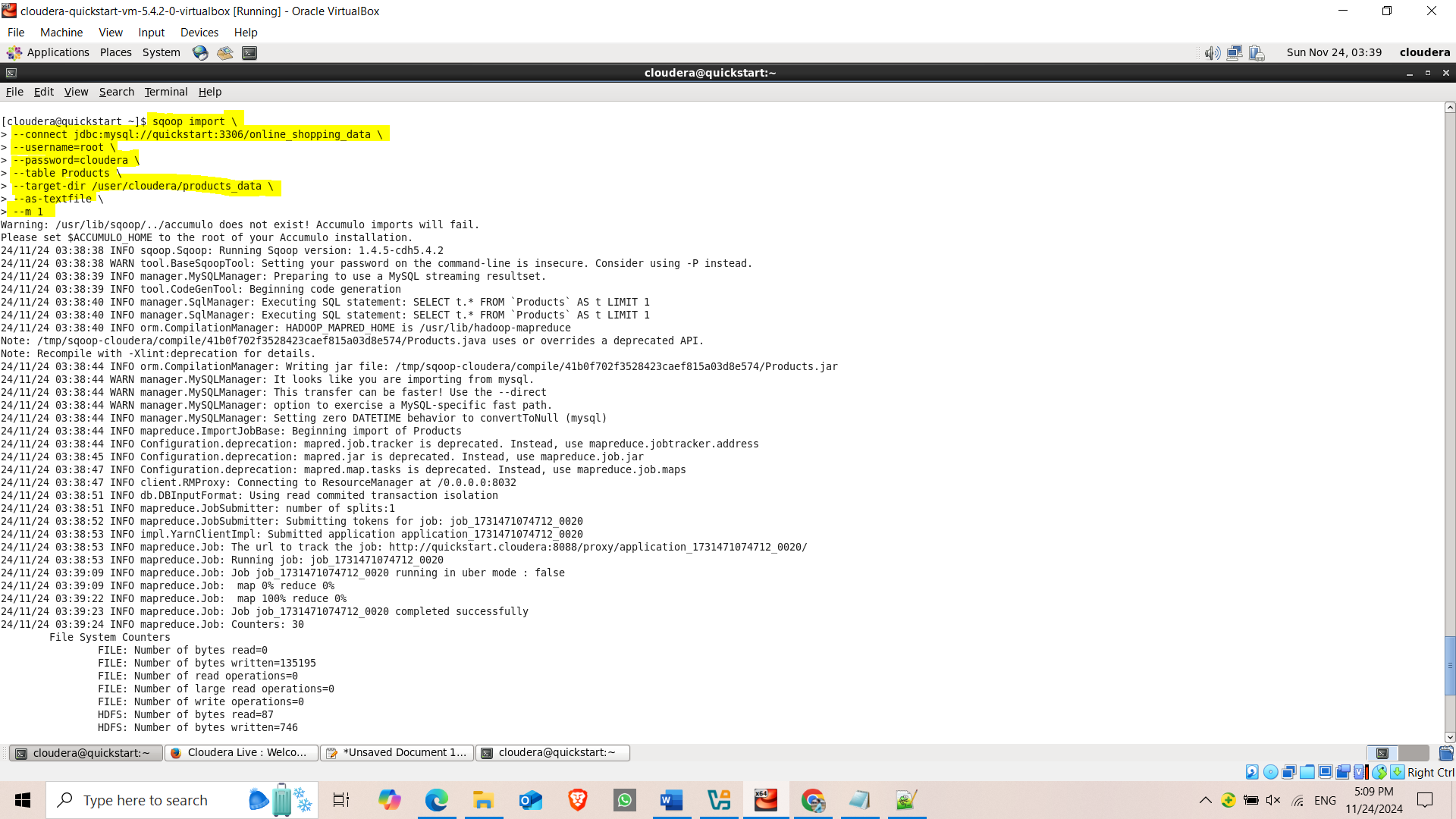
--table Products \

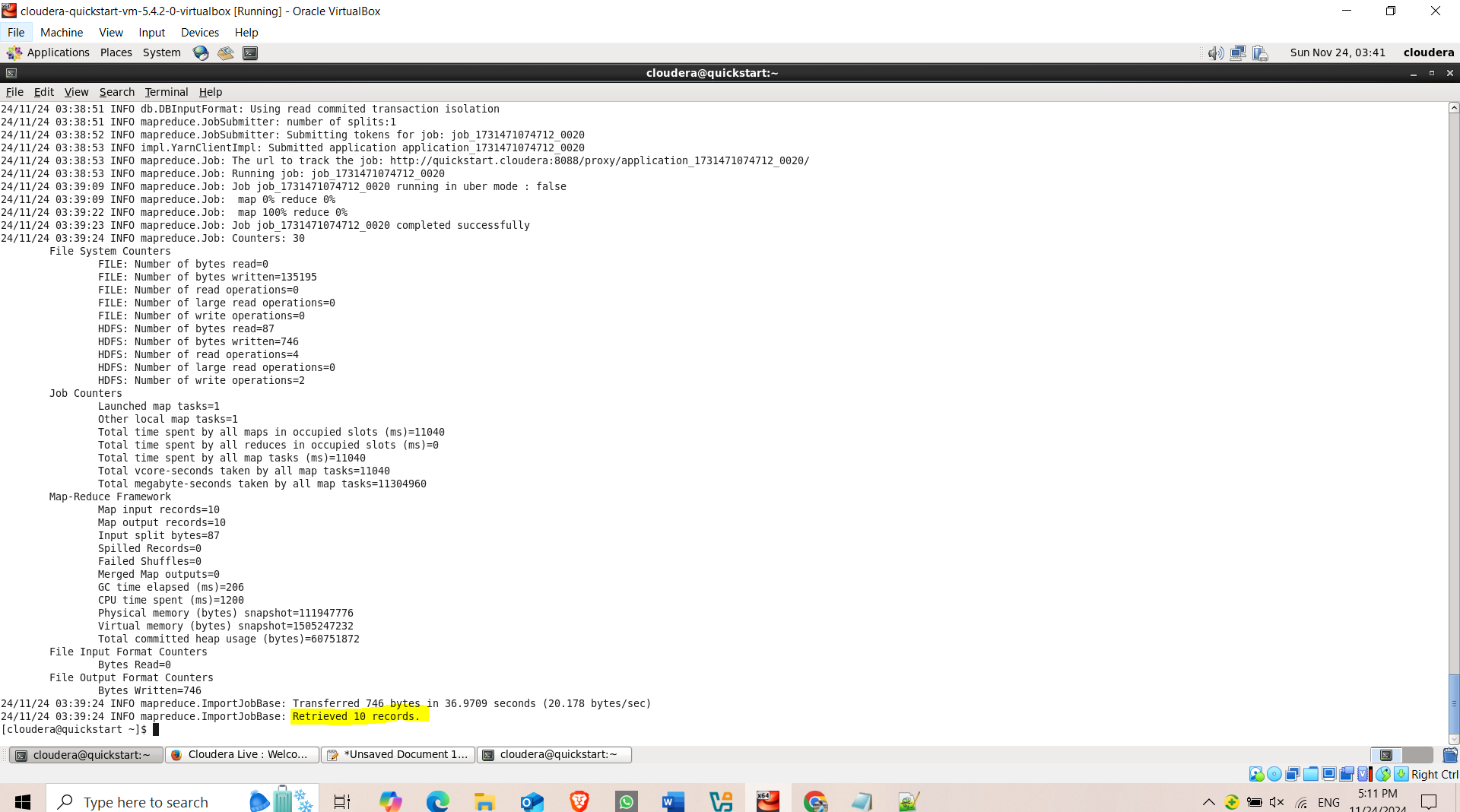
--target-dir /user/cloudera/products\_data \

--as-textfile \

--m 1

**Output:**





1. **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))

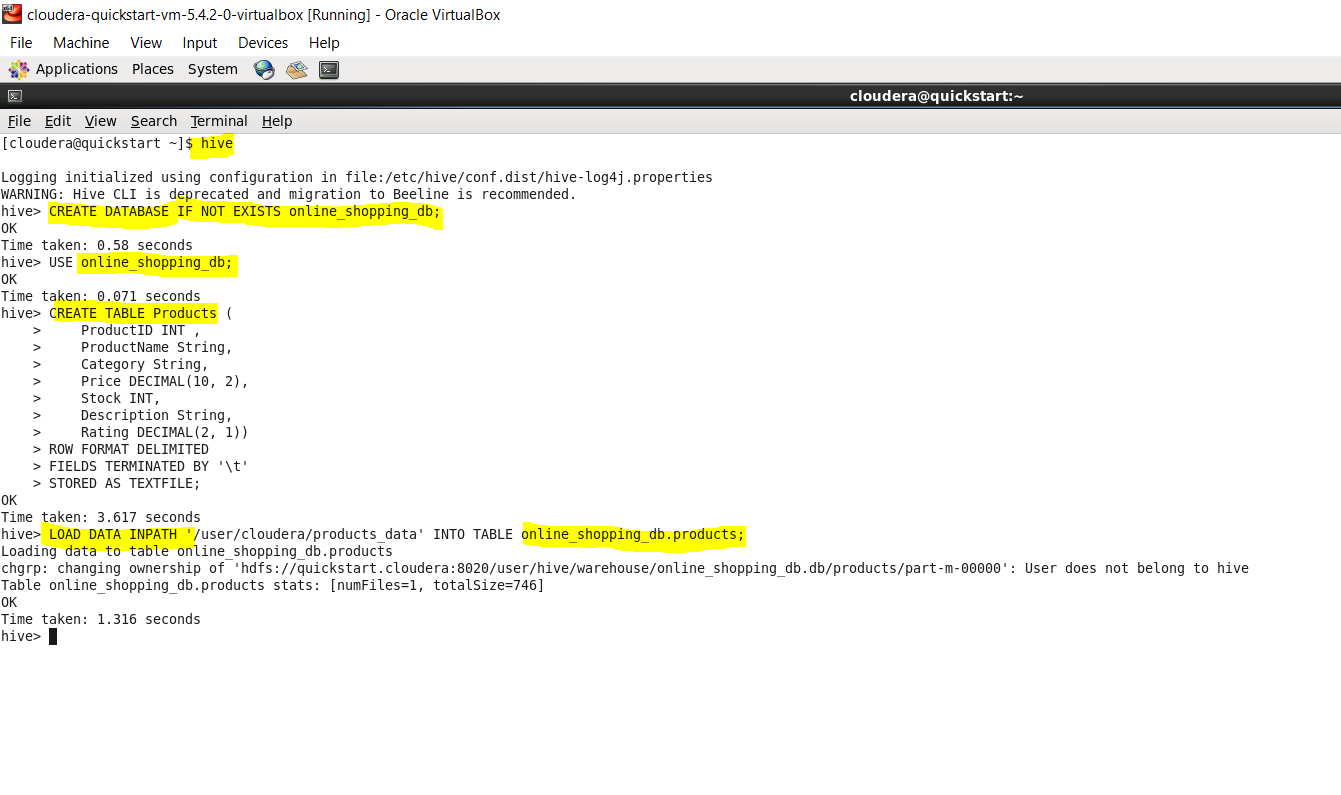
ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

LOAD DATA INPATH '/user/cloudera/products\_data' INTO TABLE online\_shopping\_db.products;

**Output:**

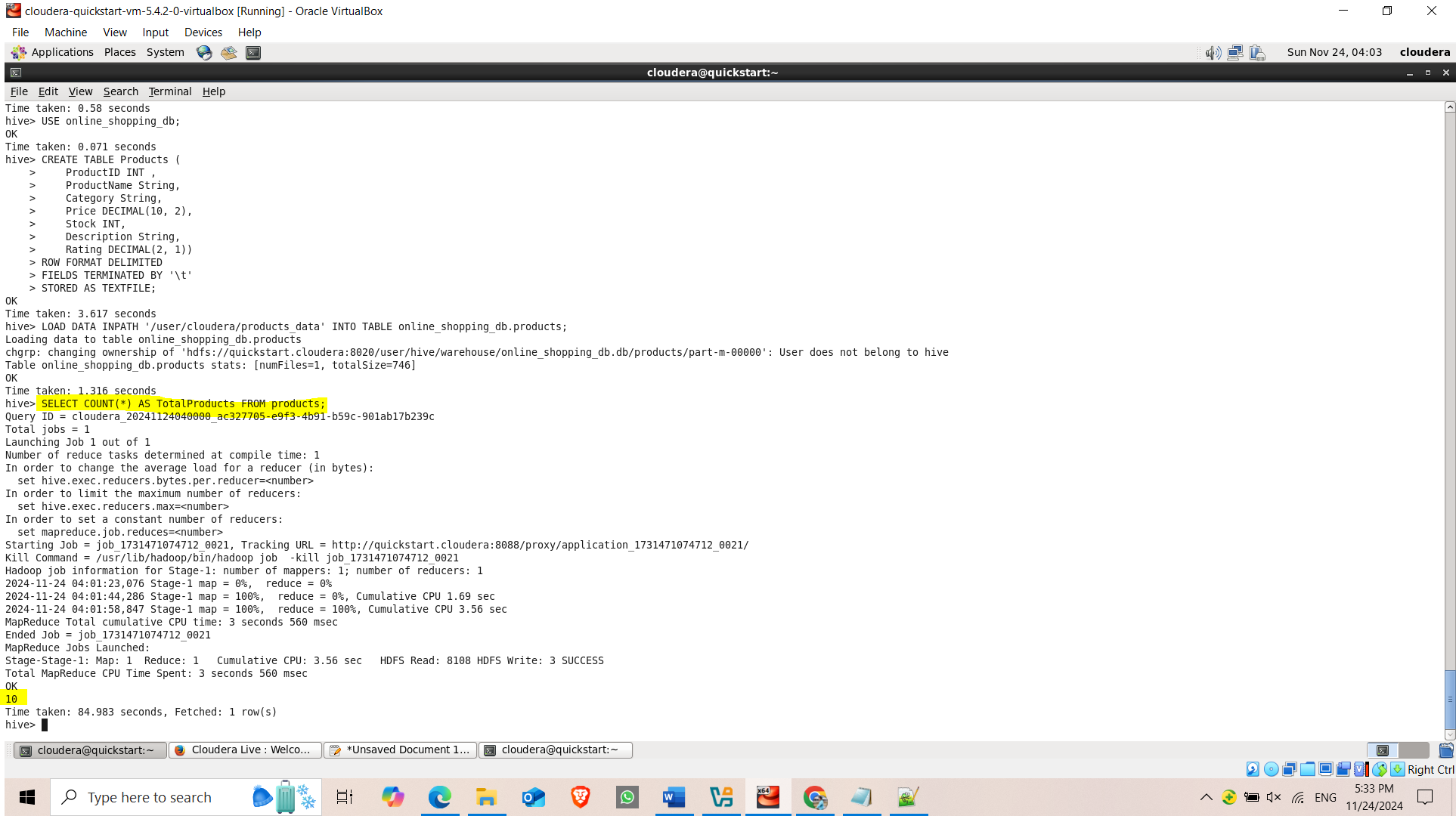


1. **Perform Queries in Hive**

* Total Number of Products :

SELECT COUNT(\*) AS TotalProducts FROM products;

**Output:**

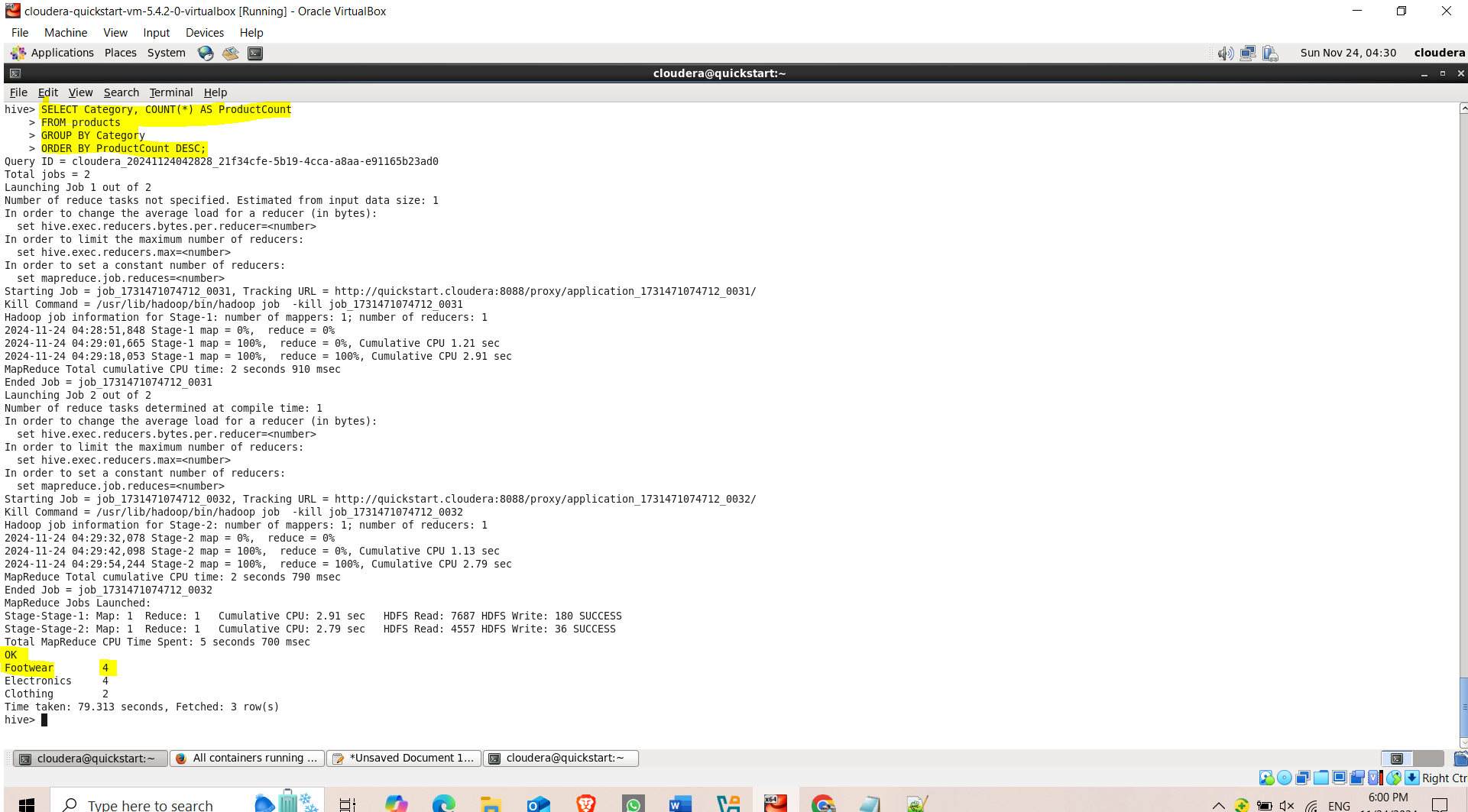


* Segregate Products by Category:

SELECT Category, COUNT(\*) AS ProductCount FROM products

GROUP BY Category ;

**Output:**

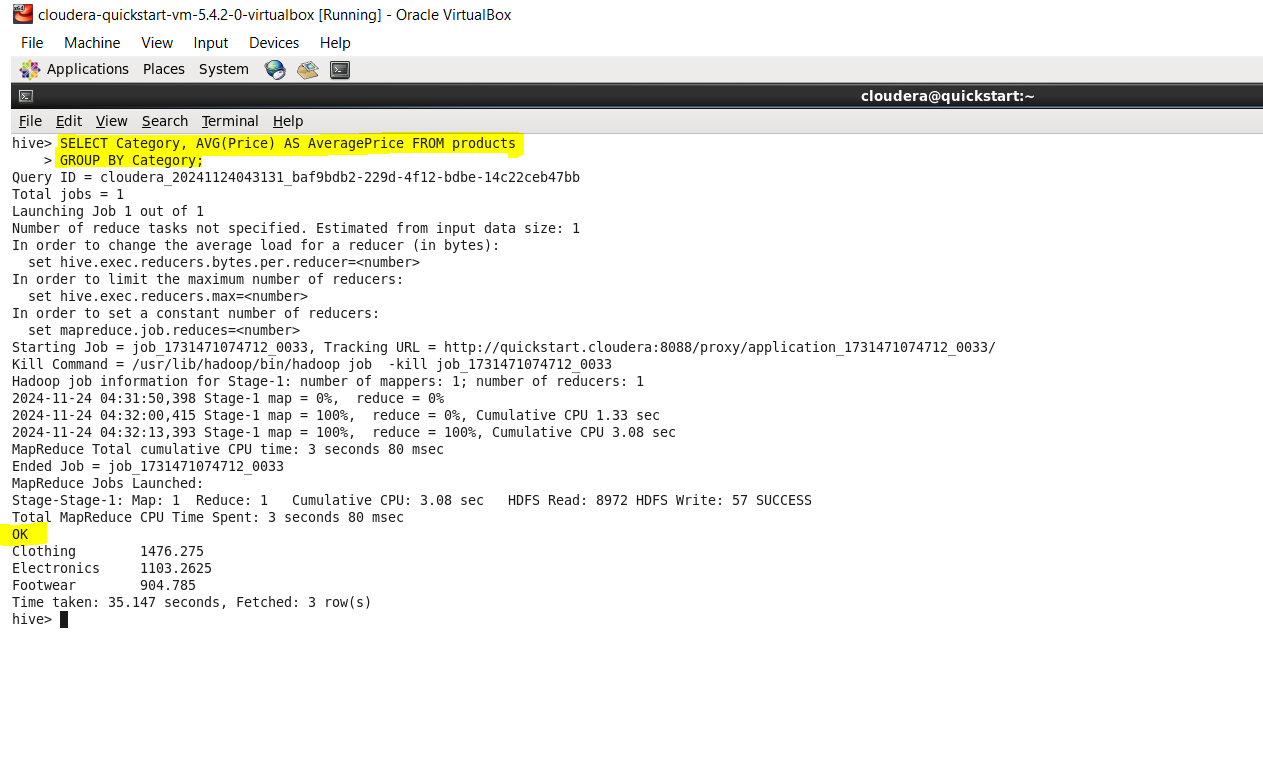


* Average Price by Category

SELECT Category, AVG(Price) AS AveragePrice FROM products

GROUP BY Category;

**Output:**

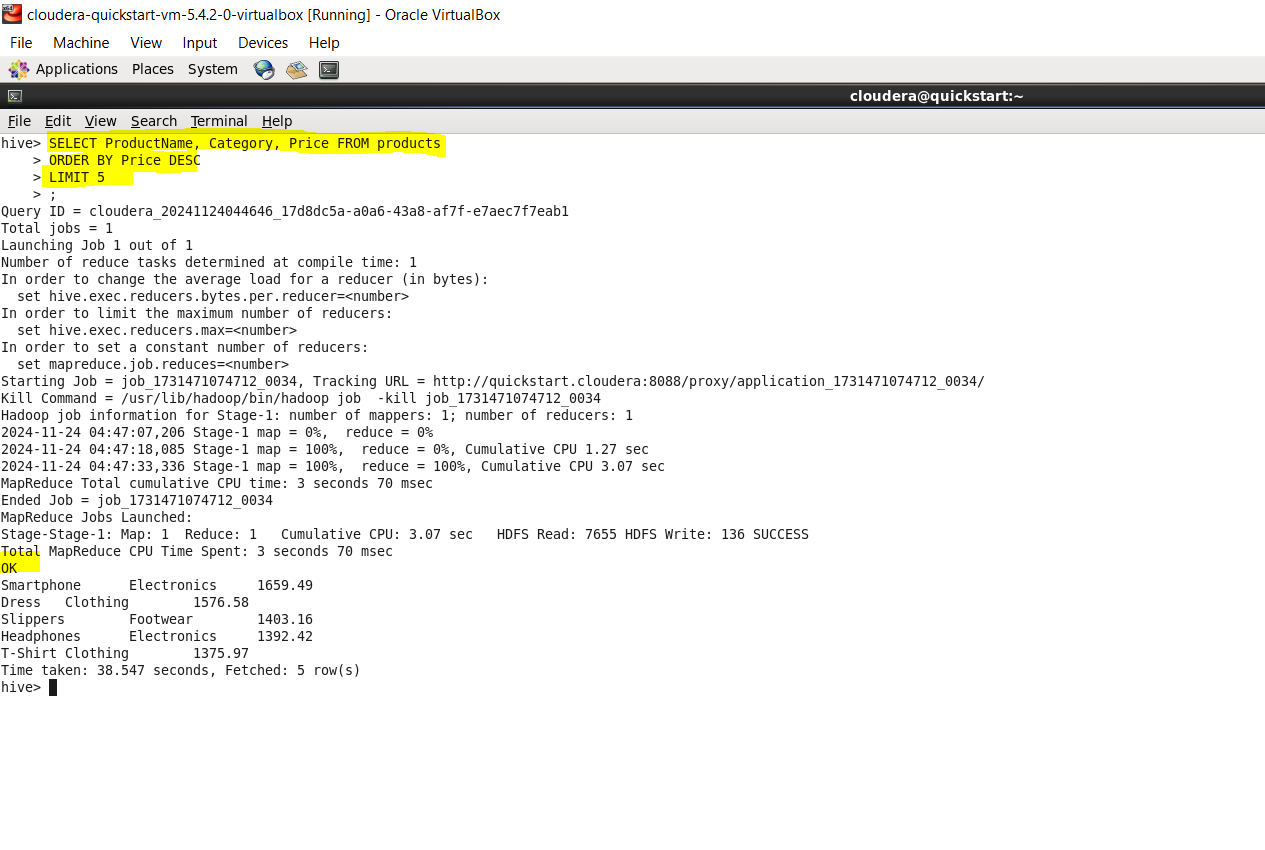


* Top 5 Most Expensive Products

SELECT ProductName, Category, Price FROM products

ORDER BY Price DESC LIMIT 5;

Output



* Count Products by Purpose with Total Stock

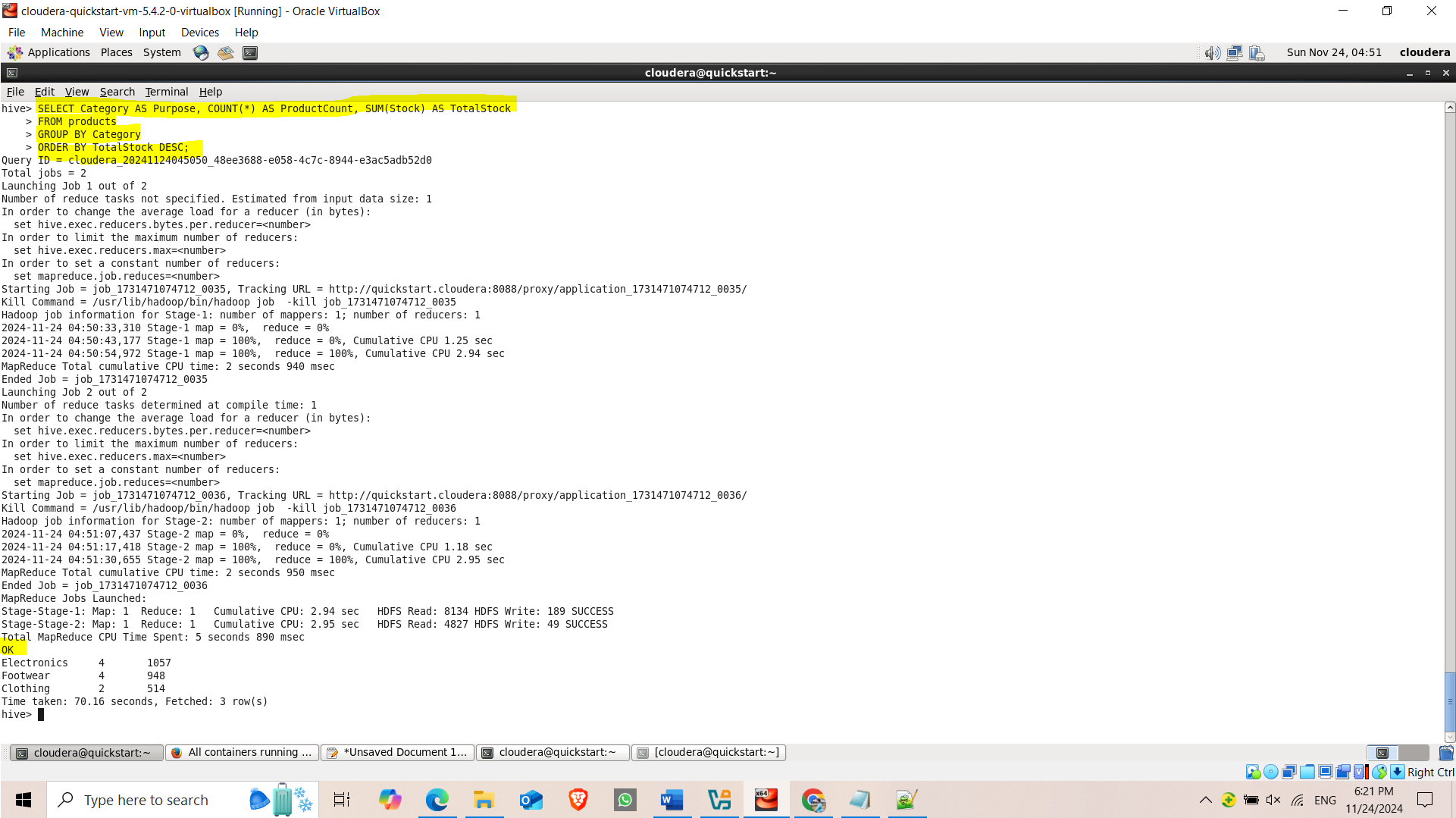
SELECT Category AS Purpose, COUNT(\*) AS ProductCount, SUM(Stock) AS TotalStock

FROM products

GROUP BY Category

ORDER BY TotalStock DESC;

**Output:**

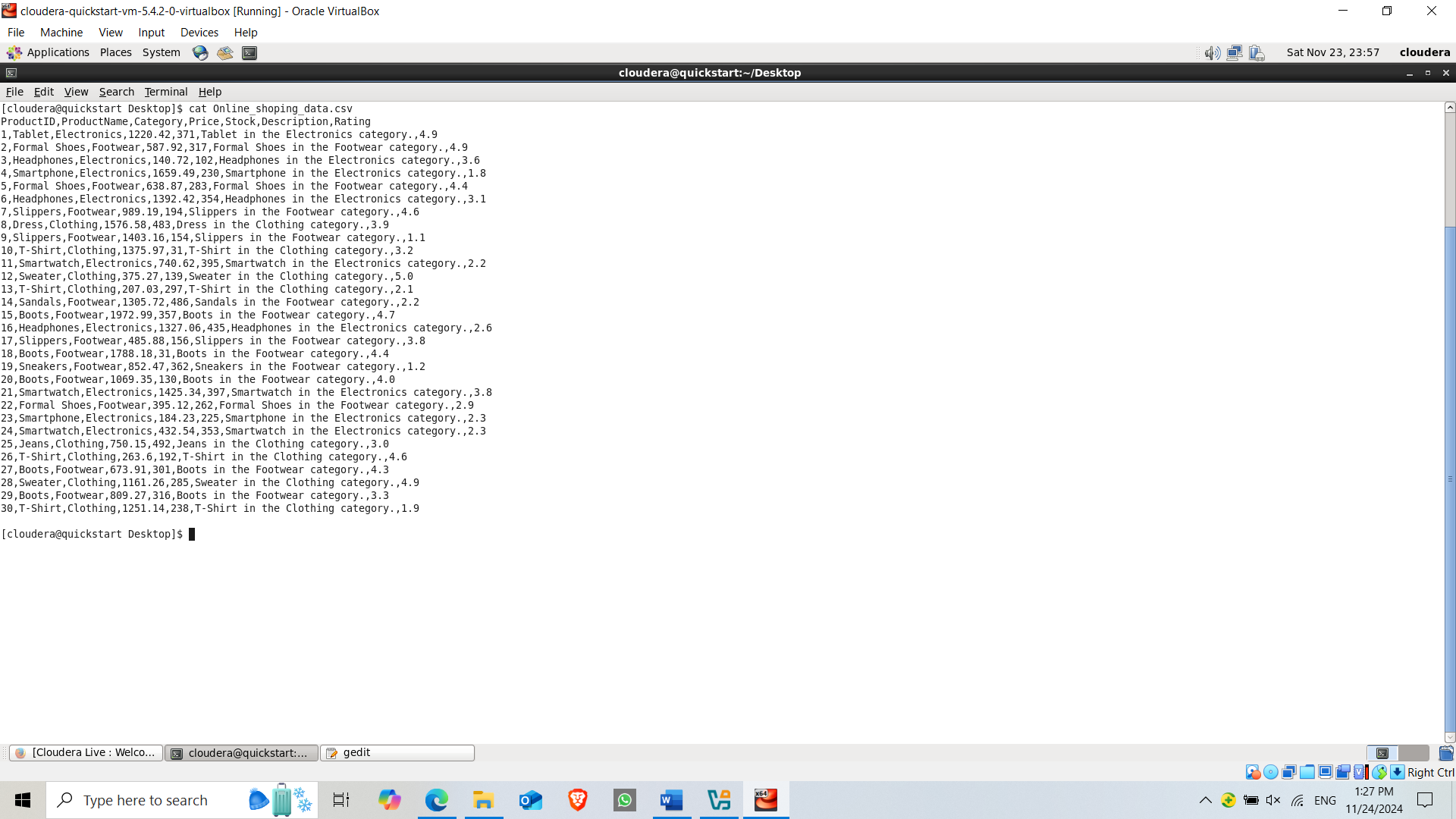


**Implementation of Hive**

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

**Dataset :**

****



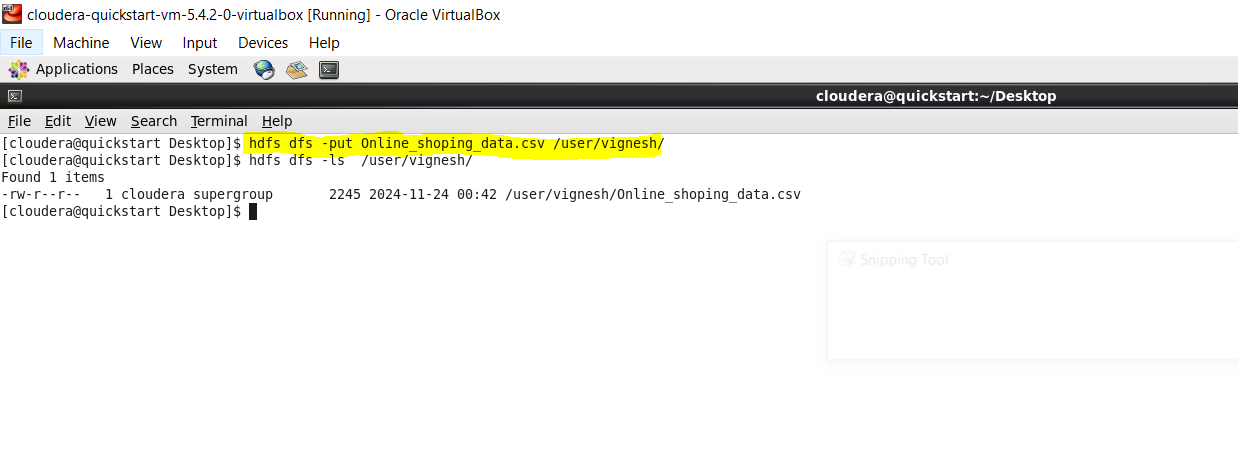
**Load the data to Hadoop :**

Commands

**hdfs dfs -mkdir /user/vignesh**

**hdfs dfs -put Online\_Shoping\_Data.csv /user/vignesh/**

**hdfs dfs -ls /user/vignesh/**



**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 ','

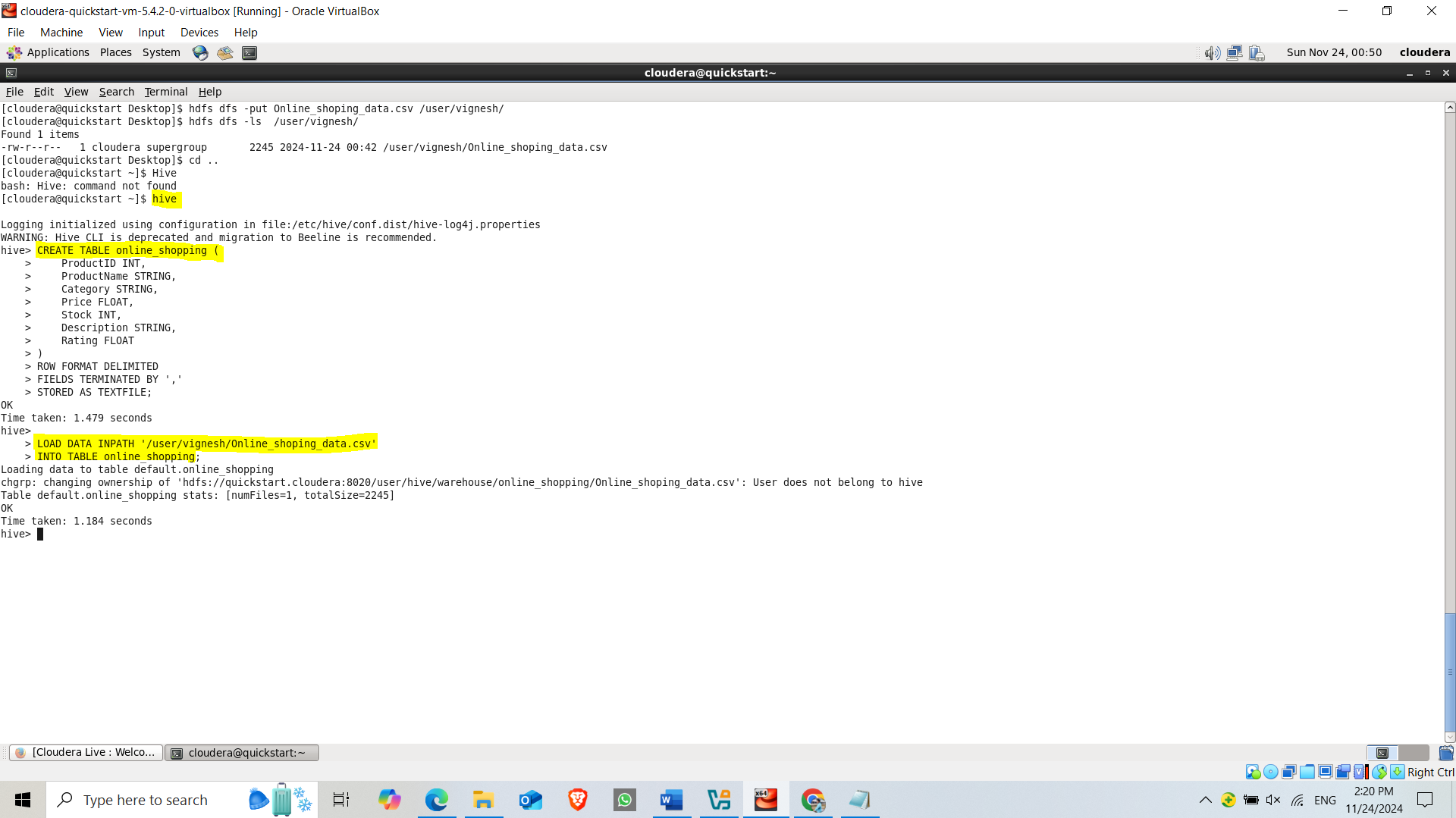
STORED AS TEXTFILE;

1. **Load the data into Hive**

LOAD DATA INPATH '/user/vignesh/Online\_shoping\_data.csv'

INTO TABLE online\_shopping;

**Output:**

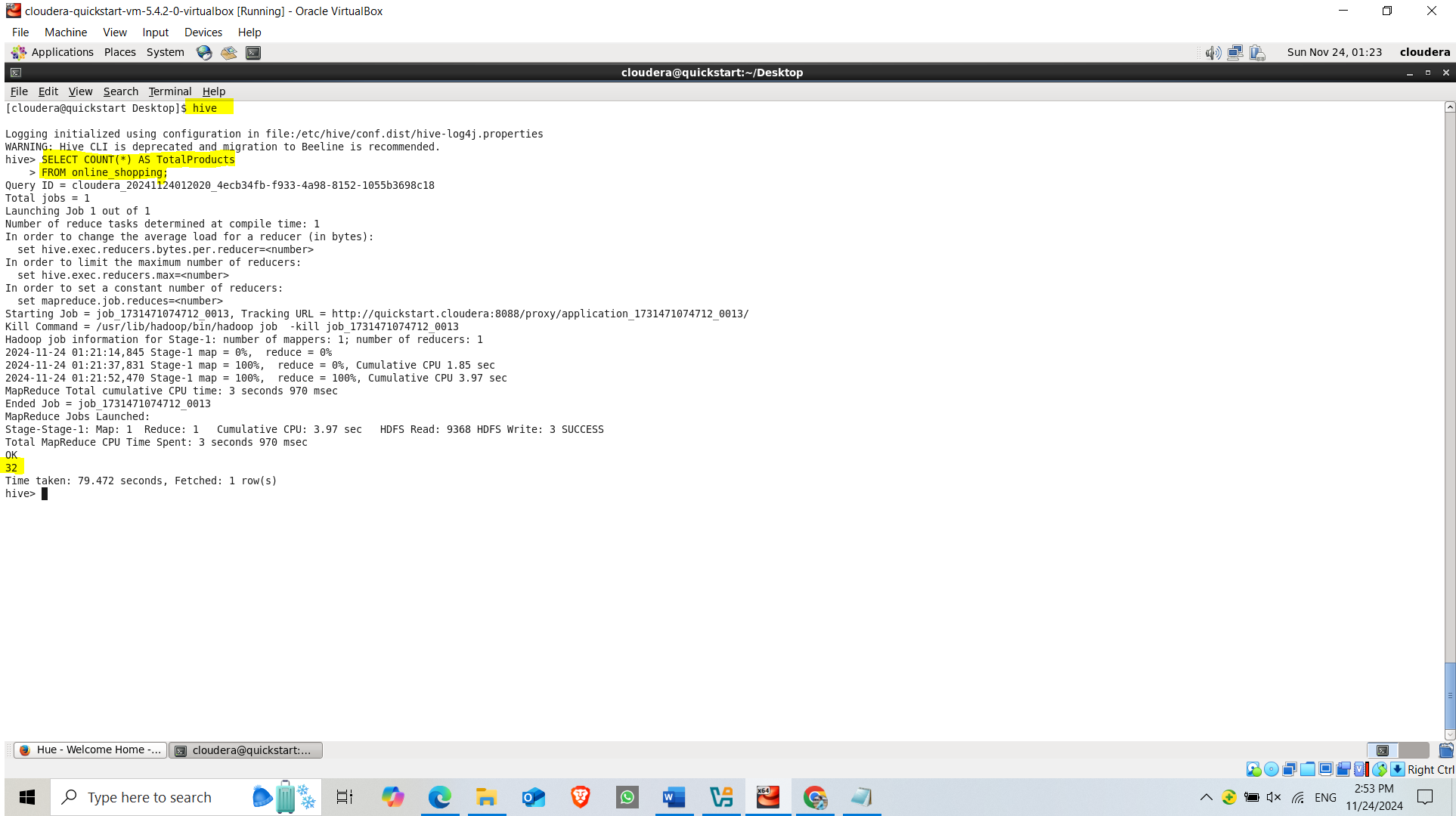


1. **Query to perform the operations in Hive:**

* Total Number of Products :

SELECT COUNT(\*) AS TotalProducts FROM online\_shopping;

**Output:**

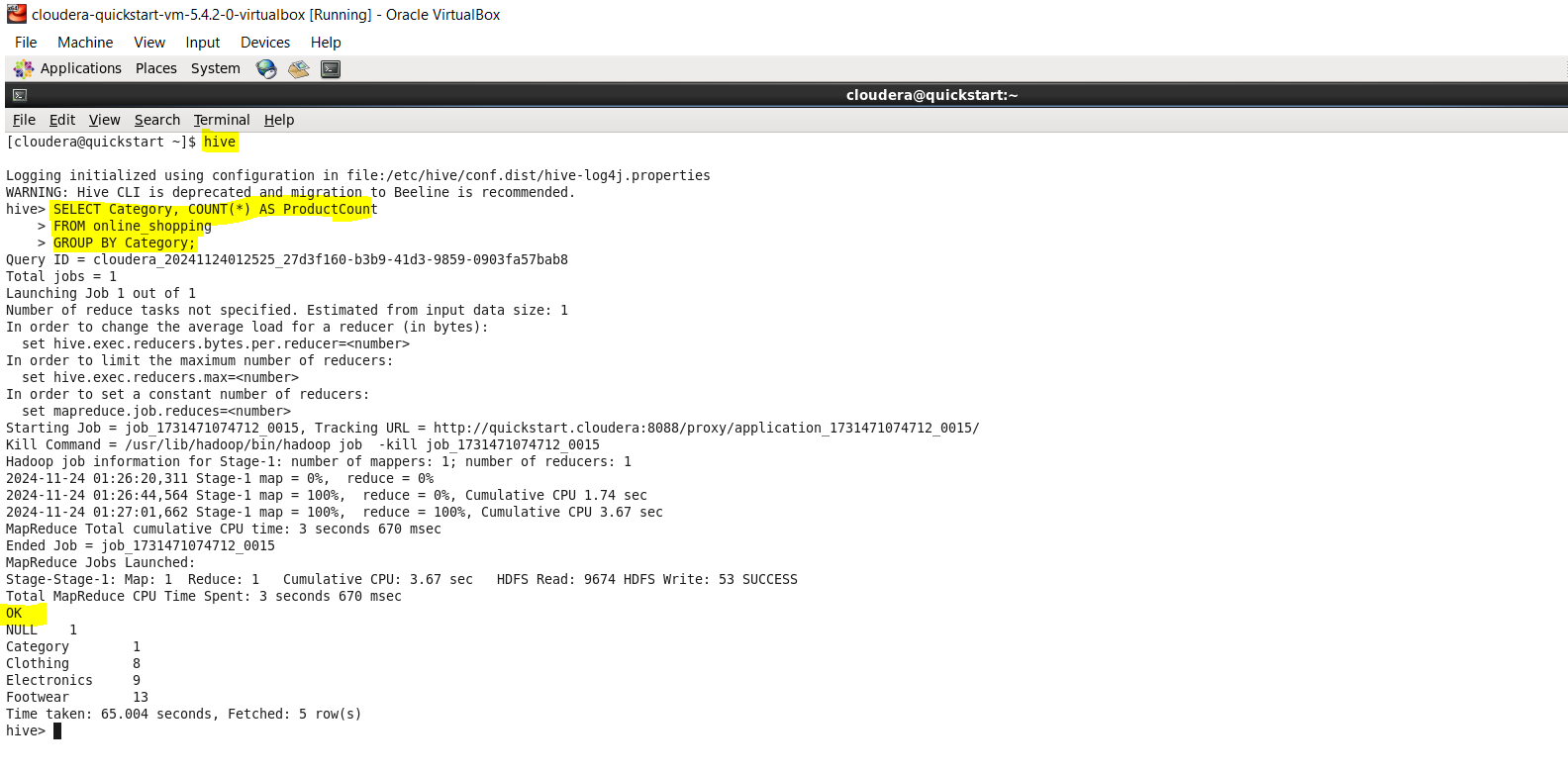


* Products Segregated by Category:

SELECT Category, COUNT(\*) AS ProductCount FROM online\_shopping

GROUP BY Category;

**Output:**

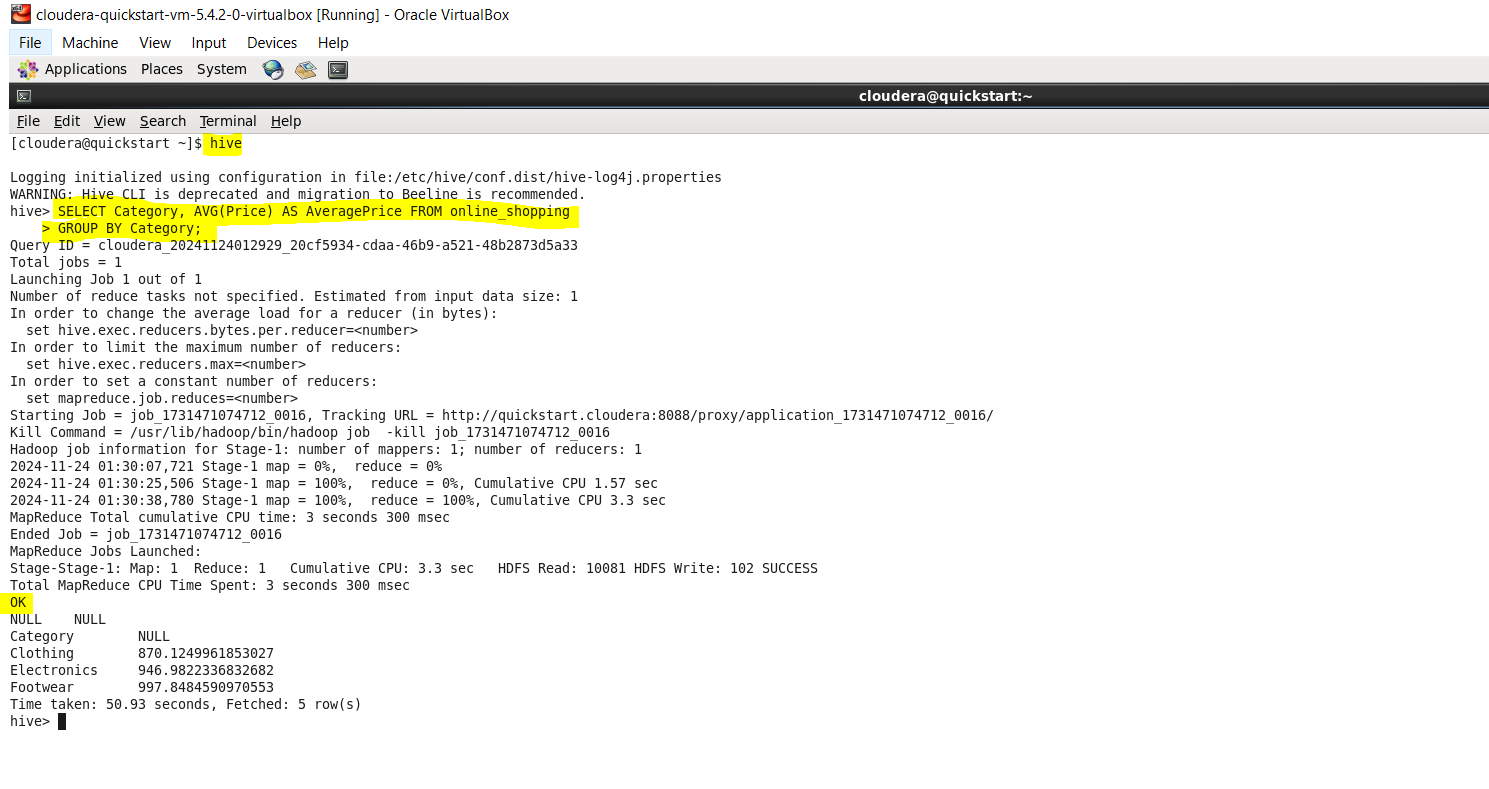


* Average Price by Category:

SELECT Category, AVG(Price) AS AveragePrice FROM online\_shopping

GROUP BY Category;

**Output:**



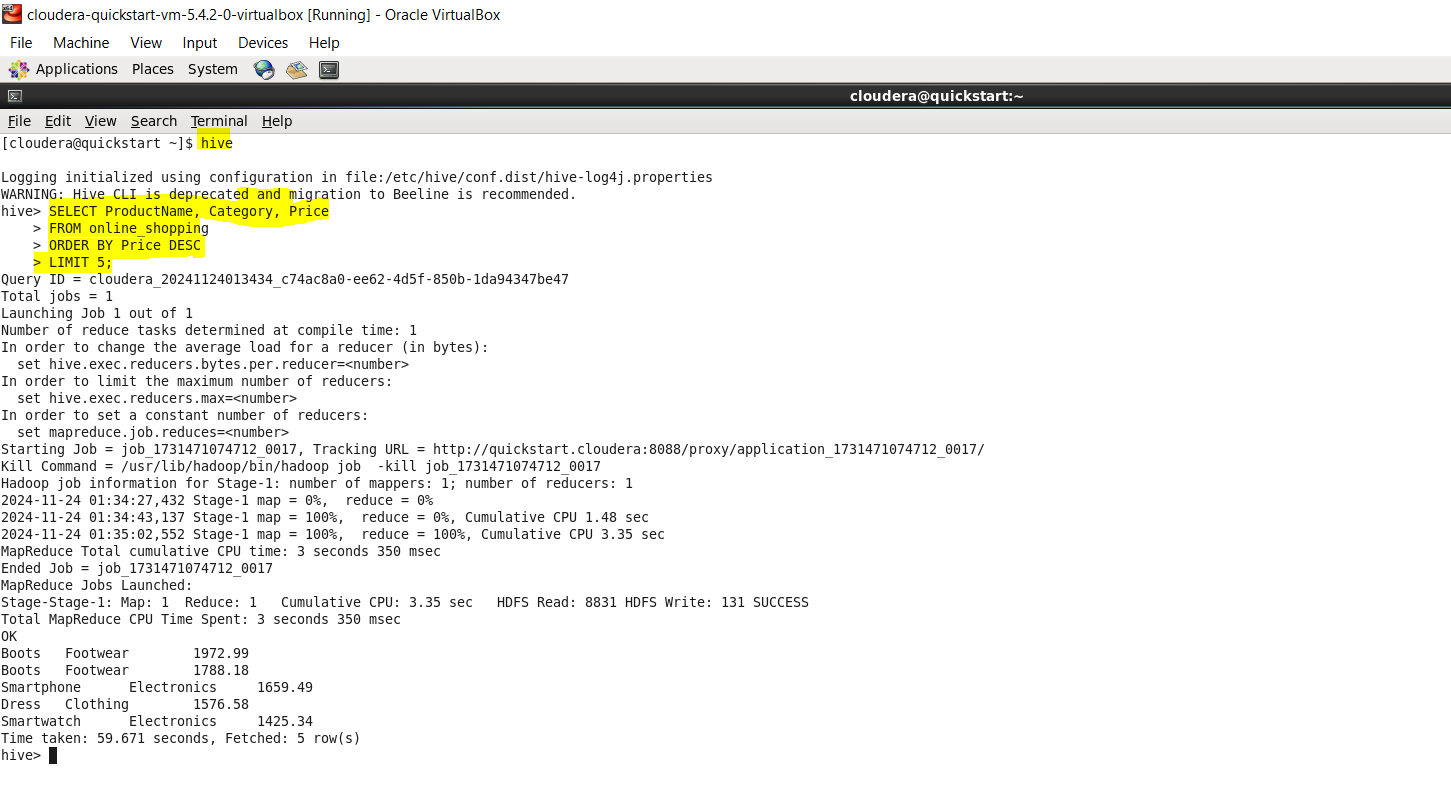
* Top 5 Most Expensive Products

SELECT ProductName, Category, Price FROM online\_shopping

ORDER BY Price DESC

LIMIT 5;

**Output:**



* Count Products by Purpose with Total Stock

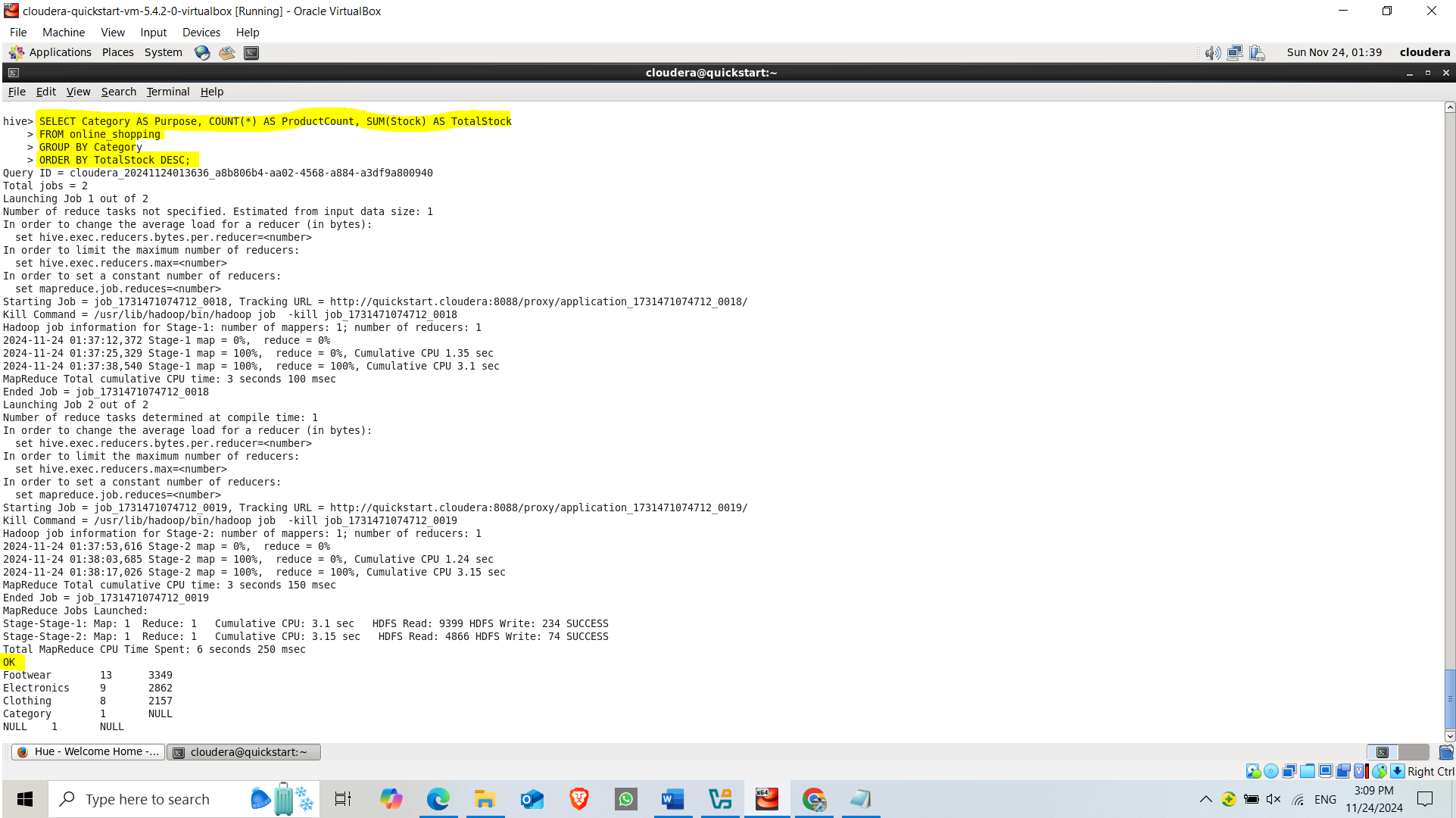
SELECT Category AS Purpose, COUNT(\*) AS ProductCount, SUM(Stock) AS TotalStock

FROM online\_shopping

GROUP BY Category

ORDER BY TotalStock DESC;

**Output:**



**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.