Assignment

Question 1 :

Perform the below tasks:

1.1. Create a managed spark table and load data to it from the given csv file.

(path : /public/trendytech/groceries.csv)

1.2. Create an external spark table with the same data.

(path : /public/trendytech/groceries.csv)

1.3. Verify that the data has been successfully loaded into both the tables.

(hint : Print the first 10 records of both the tables)

1.4. Drop the managed and external tables and see the differences.

1.5. Perform all the above tasks with the given json file.

(path : /public/trendytech/orders_wh.json/part-00000-68544d18-9a34-443f-bf0e-1dd8103ff94e-c000.json)

Question 2 :

Use the dataset given in HDFS (path : /public/trendytech/retail_db/products) :

ProductID, Category, ProductName, Description(here no data is given), Price, ImageURL

1,2,Quest Q64 10 FT. x 10 FT. Slant Leg Instant U,,59.98,http://images.acmesports.sports/Quest+Q64+10+FT.+x+10+FT.+Slant+Leg+Instant+Up+Canopy

Write the spark program using Dataframes and spark SQL for the below tasks:

2.1. Find the total number of products in the given dataset.

2.2. Find the number of unique categories of products in the given dataset.

2.3. Find the top 5 most expensive products based on their price, along with their product name, category, and image URL.

2.4. Find the number of products in each category that have a price greater than $100. Display the results in a tabular format that shows the category name and the number of products that satisfy the condition.

2.5. What are the product names and prices of products that have a price greater than $200 and belong to category 5?

Question 3 :

Use the dataset given in HDFS(path: /public/trendytech/retail_db/customers)

cust_id,cust_fname,cust_lname,cust_email,cust_password,cust_street,cust_city,cust_state,cust_zipcode

1,Richard,Hernandez,XXXXXXXXX,XXXXXXXXX,6303 Heather Plaza,Brownsville,TX,78521

Write the spark program using Dataframes and spark SQL for the below tasks:

3.1. Find the total number of customers in each state.

3.2. Find the top 5 most common last names among the customers.

3.3. Check whether there are any customers whose zip codes are not valid (i.e., not equal to 5 digits).

3.4. Count the number of customers who have valid zip codes.

3.5. Find the number of customers from each city in the state of California(CA).