Lab 05 Programming Spark- SQL

[IT494, Big Data Processing, Autumn'23]

Instructor: PM Jat (pm_jat@daiict.ac.in)

In this lab, you do some hands on Spark-SQL.
You do this lab on https://community.cloud.databricks.com

Consider a US Sales Data from: https://data.world/dataman-udit/us-regional-sales-data

Data Information is available at: https://data.world/dataman-udit/us-regional-sales-data

Dataset here is an excel file with few sheets, you can see each sheet as a table.
The excel file should also be available in my dataset folder.

Perform following computation using Spark Data Frame API

1. Compute top-10 selling products in terms of numbers (i. e. `sum(qty)`)
2. Compute top-10 selling products in terms of value (i. e. `sum (qty*price)`)
3. Compute top-10 profit making products. Profit `= sum(qty*(price-cost))`
4. Give top-3 stores selling product product number 25
5. Give top-3 products sold in `midwest` region
6. Give region wise quantity sold for each product. Compute: Region, Product ID, Sum(Qty). Region is related to a order through sales team.
7. Compute Average monthly sale in terms of numbers at each store; that , that is on average what numbers of a product are sold on a store in a month.
8. Compute sales bifurcation of each warehouse; that total sales amount through each channel
9. Compute **average "product retention period"** (i. e. the difference between procurement date and order date) at each warehouse
10. Give Year-Month sale of all products.
    Here you actually print `'Year-Month', ProductID, sum(qty)`.
    Use Order Date for extracting Year and Month of sale. For simplicity you can read order date as string only in YYYY-MM-DD format, and extract required info accordingly.
11. Compute a **fact file** with the dimensions of "`store_id`", "`product_id`", "`month_year`". Let facts to be computed are "`quantity`" and "`amount`". Let `month_year` be represented as `YYYY-MM`.

Let us say fact file would be computed as indicated below (in SQL):

```
SELECT store_id, product_id, month_year, sum(quantity), sum(amount)
    FROM ....
    GROUP BY store_id, product_id, month_year
```

Submission Required:

A document in pdf format (all in one pdf file) that contains the following:

1. All source code is pasted here. Try having the code that is formatted and colored.
2. Link to “colab notebook” that contains source code and outputs of programs for all problems.

Using of markdown or latex for creating PDF would be better.