22520467_Lab3

April 13, 2025

[ ]:
```
!pip install findspark
```

```
Collecting findspark
  Downloading findspark-2.0.1-py2.py3-none-any.whl.metadata (352 bytes)
Downloading findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Installing collected packages: findspark
Successfully installed findspark-2.0.1
```

[ ]:
```
mount_point = "/content/drive"
from google.colab import drive
drive.mount(mount_point)
```

```
Mounted at /content/drive
```

[ ]:
```
import findspark
findspark.init()
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import from_unixtime
from datetime import datetime
from pyspark.sql.window import Window
from pyspark.sql.functions import col, date_format, min, max, count, sum, avg,␣
  ↪count_distinct, month, to_date, dense_rank, row_number, regexp_extract,␣
  ↪explode, split, first, from_unixtime, year, desc, collect_list, collect_set,␣
  ↪array_intersect, size, round, sum as spark_sum, when, coalesce, lit
```

[ ]:
```
spark = SparkSession.builder.appName("Lab 3").getOrCreate()
```

[ ]:
```
customer_list_path ="/content/drive/MyDrive/ds200/Lab3/data/Customer_List.csv"
order_items_path ="/content/drive/MyDrive/ds200/Lab3/data/Order_Items.csv"
order_reviews_path ="/content/drive/MyDrive/ds200/Lab3/data/Order_Reviews.csv"
orders_path ="/content/drive/MyDrive/ds200/Lab3/data/Orders.csv"
products_path ="/content/drive/MyDrive/ds200/Lab3/data/Products.csv"
```

1. Hãy đọc dữ liệu từ các file csv, sử dụng tự suy ra kiểu dữ liệu cho mỗi cột.

[ ]:
```
customer_list_df = spark.read.csv(customer_list_path, header=True,␣
  ↪inferSchema=True, sep=";")
```

```
order_items_df = spark.read.csv(order_items_path, header=True, inferSchema=True, sep=";")
order_reviews_df = spark.read.csv(order_reviews_path, header=True, inferSchema=True, sep=";")
orders_df = spark.read.csv(orders_path, header=True, inferSchema=True, sep=";")
products_df = spark.read.csv(products_path, header=True, inferSchema=True, sep=";")
```

[ ]:

```
print("------------------------------------------------------------------------")
print("customer_list_df")
customer_list_df.printSchema()
print("------------------------------------------------------------------------")
print("order_items_df")
order_items_df.printSchema()
print("------------------------------------------------------------------------")
print("order_reviews_df")
order_reviews_df.printSchema()
print("------------------------------------------------------------------------")
print("orders_df")
orders_df.printSchema()
print("------------------------------------------------------------------------")
print("products_df")
products_df.printSchema()
print("------------------------------------------------------------------------")
```

```
------------------------------------------------------------------------
customer_list_df
root
 |-- Customer_Trx_ID: string (nullable = true)
 |-- Subscriber_ID: string (nullable = true)
 |-- Subscribe_Date: date (nullable = true)
 |-- First_Order_Date: date (nullable = true)
 |-- Customer_Postal_Code: string (nullable = true)
 |-- Customer_City: string (nullable = true)
 |-- Customer_Country: string (nullable = true)
 |-- Customer_Country_Code: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Gender: string (nullable = true)

------------------------------------------------------------------------
order_items_df
root
 |-- Order_ID: string (nullable = true)
 |-- Order_Item_ID: integer (nullable = true)
 |-- Product_ID: string (nullable = true)
 |-- Seller_ID: string (nullable = true)
 |-- Shipping_Limit_Date: timestamp (nullable = true)
 |-- Price: double (nullable = true)
 |-- Freight_Value: double (nullable = true)

-----------------------------------------------------------------------
order_reviews_df
root
 |-- Review_ID: string (nullable = true)
 |-- Order_ID: string (nullable = true)
 |-- Review_Score: string (nullable = true)
 |-- Review_Comment_Title_En: string (nullable = true)
 |-- Review_Comment_Message_En: string (nullable = true)
 |-- Review_Creation_Date: string (nullable = true)
 |-- Review_Answer_Timestamp: timestamp (nullable = true)

-----------------------------------------------------------------------
orders_df
root
 |-- Order_ID: string (nullable = true)
 |-- Customer_Trx_ID: string (nullable = true)
 |-- Order_Status: string (nullable = true)
 |-- Order_Purchase_Timestamp: timestamp (nullable = true)
 |-- Order_Approved_At: timestamp (nullable = true)
 |-- Order_Delivered_Carrier_Date: timestamp (nullable = true)
 |-- Order_Delivered_Customer_Date: timestamp (nullable = true)
 |-- Order_Estimated_Delivery_Date: timestamp (nullable = true)

-----------------------------------------------------------------------
products_df
root
 |-- Product_ID: string (nullable = true)
 |-- Product_Category_Name: string (nullable = true)
 |-- Product_Weight_Gr: integer (nullable = true)
 |-- Product_Length_Cm: integer (nullable = true)
 |-- Product_Height_Cm: integer (nullable = true)
 |-- Product_Width_Cm: integer (nullable = true)

-----------------------------------------------------------------------
```

2. Thống kê tổng số đơn hàng, số lượng khách hàng và người bán.

[ ]:

```python
total_orders = orders_df.count()
total_customers = customer_list_df.distinct().count()
total_sellers = order_items_df.select("Seller_ID").distinct().count()

print("Tổng số đơn hàng:", total_orders)
print("Tổng số khách hàng:", total_customers)
print("Tổng số người bán:", total_sellers)
```

```
Tổng số đơn hàng: 99441
Tổng số khách hàng: 102727
```

```
Tổng số người bán: 3095
```

3. Phân tích số lượng đơn hàng theo quốc gia, sắp xếp theo thứ tự giảm dần.

[ ]:
```
total_orders_groupby_country_df = orders_df.join(customer_list_df.
  ↪select("Customer_Trx_ID", "Customer_Country"), on="Customer_Trx_ID",␣
  ↪how="left")
total_orders_groupby_country = total_orders_groupby_country_df.
  ↪groupBy("Customer_Country").agg(count("*").alias("Total_Orders")).
  ↪orderBy(desc("Total_Orders"))
total_orders_groupby_country.show(n=50, truncate=False)
```

```
+----------------+------------+
|Customer_Country|Total_Orders|
+----------------+------------+
|Germany         |41754       |
|France          |12848       |
|Netherlands     |11629       |
|Belgium         |5464        |
|Austria         |5043        |
|Switzerland     |3640        |
|United Kingdom  |3382        |
|Poland          |2139        |
|Czechia         |2034        |
|Italy           |2025        |
|Spain           |1651        |
|Portugal        |1336        |
|Sweden          |975         |
|Denmark         |905         |
|Serbia          |746         |
|Norway          |716         |
|Slovakia        |534         |
|Slovenia        |495         |
|Turkey          |485         |
|Greece          |412         |
|Lithuania       |351         |
|Latvia          |280         |
|Croatia         |254         |
|Estonia         |148         |
|Finland         |81          |
|Luxembourg      |68          |
|Andorra         |46          |
+----------------+------------+
```

4. Phân tích số lượng đơn hàng nhóm theo năm, tháng đặt hàng (Hiển thị theo năm tăng dần, tháng giảm dần).

[ ]:

```
orders_with_year_month_df = orders_df.withColumn("Year",␣
 ↪year("Order_Purchase_Timestamp")).withColumn("Month",␣
 ↪month("Order_Purchase_Timestamp")).select("Order_ID", "Year", "Month")
orders_grouped = orders_with_year_month_df.groupBy("Year", "Month").
 ↪agg(count("Order_ID").alias("Total_Orders"))
orders_sorted = orders_grouped.orderBy("Year", desc("Month"))
orders_sorted.show(n=50, truncate=False)
```

```
+----+-----+------------+
|Year|Month|Total_Orders|
+----+-----+------------+
|2022|12   |1           |
|2022|10   |324         |
|2022|9    |4           |
|2023|12   |5673        |
|2023|11   |7544        |
|2023|10   |4631        |
|2023|9    |4285        |
|2023|8    |4331        |
|2023|7    |4026        |
|2023|6    |3245        |
|2023|5    |3700        |
|2023|4    |2404        |
|2023|3    |2682        |
|2023|2    |1780        |
|2023|1    |800         |
|2024|10   |4           |
|2024|9    |16          |
|2024|8    |6512        |
|2024|7    |6292        |
|2024|6    |6167        |
|2024|5    |6873        |
|2024|4    |6939        |
|2024|3    |7211        |
|2024|2    |6728        |
|2024|1    |7269        |
+----+-----+------------+
```

5. Thống kê điểm đánh giá trung bình, số lượng đánh giá theo từng mức (ví dụ: 1 đến 5).

[ ]:

```
reviews_cleaned = order_reviews_df.withColumn("Review_Score_Int",␣
 ↪col("Review_Score").cast("int"))
reviews_filtered = reviews_cleaned.filter((col("Review_Score_Int").isNotNull())␣
 ↪& (col("Review_Score_Int") >= 1) & (col("Review_Score_Int") <= 5))
overall_avg = reviews_filtered.agg(avg("Review_Score_Int")).first()[0]
```

```
review_stats = reviews_filtered.groupBy("Review_Score_Int").agg(count("*").
 ↪alias("Total_Reviews")).orderBy("Review_Score_Int").collect()

print(f"- Average Review Score: {overall_avg}")
print("- Thống kê review_score:")
for row in review_stats:
    print(f"  Rating {row['Review_Score_Int']} : {row['Total_Reviews']}")
```

```
- Average Review Score: 4.0864214950162765
- Thống kê review_score:
  Rating 1 : 11424
  Rating 2 : 3151
  Rating 3 : 8179
  Rating 4 : 19141
  Rating 5 : 57328
```

6. Tính doanh thu (giá sản phẩm + phí vận chuyển) trong năm 2024 và nhóm theo danh mục sản phẩm.

[ ]:

```
order_items_with_date = order_items_df.join(orders_df.select("Order_ID",␣
 ↪"Order_Purchase_Timestamp"), on="Order_ID", how="inner")
order_items_2024 = order_items_with_date.
 ↪filter(year("Order_Purchase_Timestamp") == 2024)
items_with_category = order_items_2024.join(products_df.select("Product_ID",␣
 ↪"Product_Category_Name"), on="Product_ID", how="left")
items_with_revenue = items_with_category.withColumn("Revenue", col("Price") +␣
 ↪col("Freight_Value"))
revenue_by_category = items_with_revenue.groupBy("Product_Category_Name").
 ↪agg(spark_sum("Revenue").alias("Total_Revenue")).
 ↪orderBy(col("Total_Revenue").desc())
revenue_by_category.show(n=100, truncate=False)
```

```
+----------------------------------------+------------------+
|Product_Category_Name                   |Total_Revenue     |
+----------------------------------------+------------------+
|Health_Beauty                           |885191.119999997  |
|Watches_Gifts                           |771986.750000001  |
|Bed_Bath_Table                          |650794.700000002  |
|Sports_Leisure                          |621999.3399999994 |
|Computers_Accessories                   |594771.0400000002 |
|Housewares                              |491576.9600000012 |
|Furniture_Decor                         |476466.1300000007 |
|Auto                                    |404210.5700000002 |
|Baby                                    |299052.5599999998 |
|Cool_Stuff                              |273910.0500000001 |
|Garden_Tools                            |259068.31999999983|
|Telephony                               |217452.1299999995 |
|Perfumery                               |204562.53999999992|
|Toys                                   |200634.07000000007|
|Office_Furniture                       |181745.7300000001 |
|Stationery                             |164743.84999999986|
|Pet_Shop                               |152804.94000000012|
|Construction_Tools_Construction        |141187.33999999985|
|Electronics                            |134265.4499999999 |
|Musical_Instruments                    |121476.30999999997|
|Small_Appliances                       |97210.18          |
|Home_Appliances_2                      |96291.60000000002 |
|#N/A                                   |90849.48000000004 |
|Fashion_Bags_Accessories               |85091.27999999997 |
|Luggage_Accessories                    |77374.65000000001 |
|Consoles_Games                         |72967.49999999997 |
|Home_Construction                      |69357.77000000002 |
|Computers                              |67860.8           |
|Home_Appliances                        |66460.23999999998 |
|Small_Appliances_Home_Oven_And_Coffee  |49297.59000000001 |
|Agro_Industry_And_Commerce             |47262.18000000002 |
|Construction_Tools_Lights              |43867.84000000001 |
|Furniture_Living_Room                  |42774.14000000003 |
|Industry_Commerce_And_Business         |37332.84000000001 |
|Audio                                  |36258.83          |
|Construction_Tools_Safety              |35409.3           |
|Kitchen_Dining_Laundry_Garden_Furniture|34951.24999999999 |
|Books_General_Interest                 |32252.79999999999 |
|Fixed_Telephony                        |30497.86          |
|Air_Conditioning                       |28889.119999999988|
|Food                                   |25709.750000000007|
|Home_Confort                           |24186.810000000005|
|Signaling_And_Security                 |22413.09          |
|Drinks                                 |21834.060000000012|
|Costruction_Tools_Garden               |20492.689999999995|
|Art                                    |17947.630000000005|
|Books_Technical                        |16570.430000000004|
|Furniture_Bedroom                      |13992.429999999997|
|Costruction_Tools_Tools                |13448.200000000003|
|Fashion_Shoes                          |10118.42          |
|Food_Drink                             |10117.999999999998|
|Market_Place                           |9364.009999999998 |
|Christmas_Supplies                     |8078.5            |
|Cine_Photo                             |7253.71           |
|Music                                  |4994.579999999999 |
|Books_Imported                         |4090.710000000001 |
|Party_Supplies                         |3278.4100000000008|
|Fashion_Underwear_Beach                |3116.21           |
|Fashion_Male_Clothing                  |2810.4100000000003|
|Furniture_Mattress_And_Upholstery      |2688.6399999999994|
|Arts_And_Craftmanship                  |2008.9200000000003|
|Diapers_And_Hygiene                    |1890.58           |
|Dvds_Blu_Ray                           |1674.8500000000001|
|Tablets_Printing_Image                 |1541.8300000000002|
|Flowers                                |1216.6399999999999|
|La_Cuisine                             |890.8499999999999 |
|Fashio_Female_Clothing                 |783.37            |
|Fashion_Sport                          |431.19            |
|Home_Comfort_2                         |283.90999999999997|
|Fashion_Childrens_Clothes              |194.04            |
|Cds_Dvds_Musicals                      |117.58            |
+---------------------------------------+------------------+
```

[ ]: