Business Report

Welcome to the assignment for the third session. You're well on your way to obtain the Wizeline Certification for Big Data Engineering with Spark!

If you have any feedback about our courses, email us at academy@wizeline.com or use the Academy Slack channel.

Problem Description

Alimazon's Board of Directors has issued a requirement to the Data Analysis Departament for a full report to have more insights on products and consumers. The list of specifications for this report is as follows:

Report Summary

The report should be executed as one application and should include performance optimizations as the use of `cache` or `persist` methods, multiple executions and maintenance of the report for future customizations are expected.

Best Selling Products

Top 10 selling products by day of week by gross sales and by orders.

For each day of the week we'll have two different sets of files:

The first set of files should be called "`top10-products-by-gross-sales-[day of the week as number]`" (for example: "`top10-products-by-gross-sales-1`") and will contain the 10 best selling products for each day of the week in the whole dataset. The files should contain the next columns:

- `product_id`: with the id of the product.
- `gross_sales`: with the total sum of all sales for the given product and for the corresponding day of the week.

The second set of files should be "`top10-products-by-orders-[day of the week as number]`" with the columns:

- `product_id`: with the id of the product.
- `orders_count`: with the total number of orders of the product for the corresponding day of the week.

Best Customers

Top 10 customers by month by orders and spending.

For each month in the dataset we'll have two different sets of files:

The first set of files should be called "`top10-customers-by-gross-spending-[month as number]`" and will contain the 10 best customers for each month in the whole dataset. The files should contain the next columns:

- `client_id`: with the id of the customer..
- `gross_spending`: with the total sum of all sales placed by the given customer during the corresponding month period.

The second set of files should be "`top10-customers-by-orders-[month as number]`" with the columns:

- `client_id`: with the id of the customer.
- `orders_count`: with the total number of orders placed by the customer during the corresponding month period.

NOTE: When saving your files, replace **`[day of the week as number]`** with numbers 1 to 7, or **`[month as number]`** with numbers 1 to 12.

Input Dataset

You can find the input dataset at:

```
gs://de-training-input/alimazon/200000/client-orders/
```

Expected Output

All files generated from this report should use the csv format with headers included. You can write your output dataset to your assigned output bucket following this format:

```
gs://de-training-output-<student-name>/assignment-3/<file-name>
```

Using the first set of files as example, the content of the files should look like:

```
product_id,gross-sales
B00BP16R2S,51604.96
B0010ZJF1E,39928.0
B00TX32SJG,34319.36
B005N40PYO,31141.46
```

```
B00P1NDZC6,29734.53
B000JLV4FA,29246.28
B00BIO5A9I,27527.68
B00DQB6QH6,27450.9
B006R34D52,27197.25
0892814888,26872.32
```

NOTE: the headers are different for each set of files; the total gross sales and ordering is NOT the actual answer for the first set of files.

Hint: For the report you can use the functions available on the Spark libraries:
Spark: https://spark.apache.org/docs/2.2.1/api/scala/index.html#org.apache.spark.sql.functions$
PySpark: http://spark.apache.org/docs/2.2.1/api/python/_modules/pyspark/sql/functions.html