Code Logic - Retail Data Analysis

Step 1: Import all the required libraries and establish a spark session.

Step 2: Connect to the kafka server and read the data.

Step 3: Create a Schema for the data. Columns in the data are "country", "invoice_no", "timestamp", "type", "items" which is an array which again divided into "SKU", "title", "unit_price", "quantity".

Step 4: Select the required columns and store into a new dataframe.

orderStreamDF =
masterDF.select(from_json(col("value"),customSchema).alias("data")).select("data.*")

Step 5: Rename these columns: SKU , title, unit price & quantity.

Step 6: Define UDFs and execute them.

• UDF 1: This function accepts “type” as input and determines if the nature of request is “Order” or “Return” If Type comes as “RETURN”, then it returns a flag is_return = 1 which says it is a return.

• UDF 2: This function accepts “type” as input and determines if the nature of request is “Order” or “Return” If Type comes as “ORDER”, then it returns a flag is_order = 1 which says it is an order.

• UDF 3: This function accepts 2 attributes in input ; Items & type as input. If type is “Order”, it returns a value of Unit price multiplied by quantity. If type is “Return”, it returns a value of Unit price multiplied by quantity toggled by “-“ symbol.

Step 7 : Calculate the Time Based and Country&Time Bases KPIs.(line 92-124)

Step 8 : Store these KPIs to Json files. Step 9 : Run the Spark submit command along with the package and python file name and also write the console output to a separate file :

spark2-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.0 spark-streaming.py > console_test