Code Logic - Retail Data Analysis

- Downloaded Jar file "spark-sql-kafka-0-10_2.11-2.3.0.jar"
- Before submitting the spark2-submit job we need to run command: "export SPARK_KAFKA_VERSION=0.10", which asks spark to use **Kafka** version **0.10** to run the spark job.
- Create a python code file.
  - Initial steps to setup all environment, importing all required modules and initialize spark session.
  - Gathered bootstrap server details and topic name from where the data needs to be read.
  - Now using the Kafka bootstrap server details provided, we are reading the data and store the data to a variable.
  - Using the from_json method we are converting the read json data into a data frame using a json schema created based on the how the data is arranged.
  - We are adding few columns to data frame created with name and usage as:
    - "Total_Amount" – which shows value of total amount on invoice.
    - "Total_Quantity" – which shows total items count per invoice.
    - "Is_Order" – If its order shows as 1 else 0.
    - "Is_Return" – If its return shows as 1 else 0.

    These new columns were added, so that they could help us to calculate KPI's.
  - Created two user defined functions to calculate "Total_Amount" and "Total_Quantity".
    - Methods Names: total_Invoice_amount_method and total_items_per_invoice_method.
    - Defining those Methods as UDF with name:
      - total_Invoice_amount and total_items_per_invoice respectively.
  - Now final dataframe is created. We use this data frame to show summarized batch output received for a min to the console and also to calculate all the KPI's.
  - Also calculating the four KPI's with two distribution one is based on Time, another based on time and country based.
  - For Time based KPI we are calculating on 1 min tumbling window KPI's like Orders per min, Total sale Volume, Average transaction size and rate of return.
  - For Time and Country based KPI we are calculating on 1 min tumbling window KPI's like Orders per min, Total sale Volume and rate of return grouped by countries.
  - All these calculated Time KPI and Time and Country based KPI are written to a json file.
  - These files can be further process into Hive for data analysis and for visualization tools for further insights.
- Once coding part is complete, we need to submit the spark2 job on console.
- Command used to submit spark Job:
  - spark2-submit --jars "spark-sql-kafka-0-10_2.11-2.3.0.jar" spark-streaming.py > Console_output
  - KPI's data is written in the HDFS server.
- ConsoleOutput.txt show all the batch outputs.