Code Logic - Retail Data Analysis

Project Objective:

To ingest data from a centralized Kafka server in real-time and process it using PySpark to calculate various KPIs (Key Performance Indicators) for an e-commerce company, RetailCorp Inc.

Key tasks performed in PySpark:

1. Reading the data from the Kafka server (connecting to server on IP: 18.211.252.152 through port 9092)
2. Calculating derived columns (total_cost, total_count, is_return, is_order)
3. Calculating time-based KPIs and country-and-time based KPIs
4. Storing the above calculated KPIs into separate JSON files for further analysis.

Solving the project:

Step 1: Start the EC2 instance and when it is up and running, create a “spark-streaming.py” python file in the terminal using the command

```
1. vi spark-streaming.py
```

Step 2: Enter into insert mode by clicking “I” and write the PySpark code to achieve all the key tasks and calculate the KPIs. Once done, click on “:wq” to save and exit the editor.

Step 3: Run the following command to enable Spark and Kafka integration.

```
1. export SPARK_KAFKA_VERSION=0.10
```

Step 4: Run the Spark Submit command by providing the “--jars” argument and our python file name and write this console output into a separate file. Command for the same is given below.

```
1. spark2-submit --jars spark-sql-kafka-0-10_2.11-2.3.0.jar spark-streaming.py > Console-
   output.txt
```

This completes the overview of the project approach and step-wise description of how it is completed.