Code Logic - Retail Data Analysis

1. We import the required libraries from pyspark.sql package.
2. We define the schema of the JSON data received from the Kafka stream.
3. We define the udf function to –
    a. Get the total cost i.e the total amount spent in the order
    b. Get the number of items purchased in any order
    c. Is the order type an Order type
    d. Is the order type a Return type
4. We define the spark session
5. The data is read in the form of key and value from Kafka
6. We use the function from_json() to convert data to JSON format.
7. We create the Kafkadf dataframe with all the required fields and columns and use UDF function to calculate other fields.
8. We write the stream output to the console.
9. With a watermark of 1 minute and tumbling window of 1 minute, we create a stream to calculate the time basis KPI. We perform various aggregation function to calculate the KPI.
10. With a watermark of 1 minute and tumbling window of 1 minute, we create a stream to calculate the country time basis KPI.We perform various aggregation function to calculate the KPI.
11. We wait for the termination of the stream from kafka.