Code Logic - Retail Data Analysis

In this document, you will describe the code and the overall steps taken to solve the project.

1. Need to give dependencies for program to run , which we have set environment by giving variable such has pyspark_python ,java_home,spark_home,pylib .
2. We need to import library for our code to run.
3. Then we need to define function which is used in UDF (user define function). Such has total value of single order by multiplying unit price to quantity of product ,total number of items per order by adding item in quantity,is_order and is_return
4. Then validate command line arguments and initializing park session
5. We need to read the input from Kafka servers.
6. So need to provide a scheme for the coming data and file format .
7. We need do preprocessing of order by adding UDF from function defined previously and add the columns to table and print summarized input in the console per minute .
8. Now to calculate the KPI based on the summarized input .
9. Calculating the time based KPI by giving watermark and window of 1 min .
10. Calculating the county based KPI for per minute.
11. Saving this KPI in json file format in hadoop .
12. After this we need to extract the json file from hadoop using WInSCP .
13. We need store the summarized input table in the console output file.