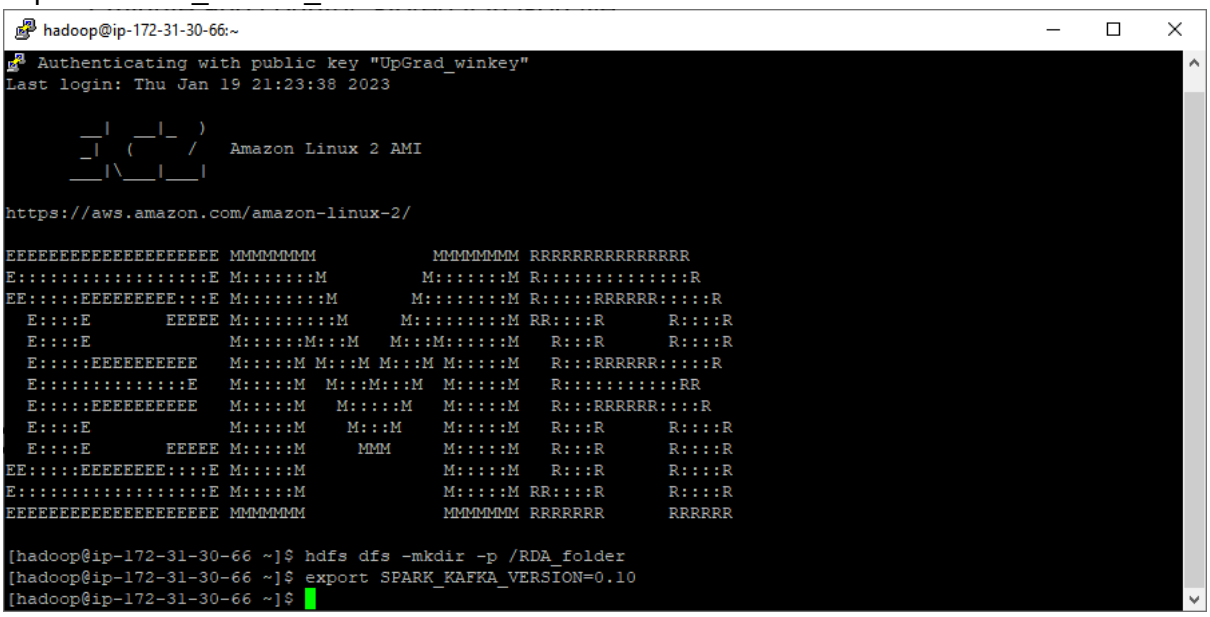
Code Logic - Retail Data Analysis

1. First, we create EMR Cluster on AWS with necessary applications for the project.
2. Started reading from Kafka topic using spark streaming.
3. Made sure to read the kafka topic from the very beginning with ".option("startingOffsets", "latest") \"
4. Defining the streamed data's schema.
5. Created dataframe from the said schema.
6. Creating new columns using user-defined functions;
   - for calculating total_items (items_TotalCount)
   - for calculating order type (is_order)
   - for calculating return type (is_return)
   - for calculating total_cost (TotalCostSum)
7. Converted all UDFs with utility function.
8. Printing data into console with 1 minute interval.
9. Calculated time based KPI with watermark, grouped by window timestamp of 1 minute, stored it in json file.
10. Calculated time-country based KPI with watermark, grouped by window timestamp of 1 minute and country, stored it in json file.
11. Kept stream open to read data infinitely.
12. Stored the console output to a file.
13. Copied the json data, spark script and console output file to local machine.

Console Commands:

1. **Creating directory:**
   hdfs dfs -mkdir -p /RDA_folder

2. **Choosing spark kafka verison:**
   export SPARK_KAFKA_VERSION=0.10



3. **Running the python script using spark submit:**
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py

```
hadoop@ip-172-31-30-66:~
23/01/19 21:27:52 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after re
ached minRegisteredResourcesRatio: 0.0
23/01/19 21:27:53 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/hive-site.xml
23/01/19 21:27:53 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.wa
rehouse.dir ('hdfs:///user/spark/warehouse').
23/01/19 21:27:53 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
23/01/19 21:27:53 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL.
23/01/19 21:27:53 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/json.
23/01/19 21:27:53 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/execution.
23/01/19 21:27:53 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/execution/json.
23/01/19 21:27:53 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /static/sql.
23/01/19 21:27:53 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+-------+---------+----------+-----------+--------+---------+
|invoice_no|country|timestamp|total_cost|total_items|is_order|is_return|
+----------+-------+---------+----------+-----------+--------+---------+
+----------+-------+---------+----------+-----------+--------+---------+
```

```
hadoop@ip-172-31-30-66:~
-------------------------------------------
Batch: 7
-------------------------------------------
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost        |total_items|is_order|is_return|
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|154132552848734|United Kingdom|2023-01-19 21:33:31|11.64             |5          |1       |0        |
|154132552848735|United Kingdom|2023-01-19 21:33:31|16.6              |4          |1       |0        |
|154132552848736|United Kingdom|2023-01-19 21:33:35|-72.62            |14         |0       |1        |
|154132552848737|United Kingdom|2023-01-19 21:33:41|97.8              |28         |1       |0        |
|154132552848738|United Kingdom|2023-01-19 21:33:41|13.95             |15         |1       |0        |
|154132552848739|United Kingdom|2023-01-19 21:33:57|-7.5              |2          |0       |1        |
|154132552848740|United Kingdom|2023-01-19 21:34:00|50.84             |21         |1       |0        |
|154132552848741|United Kingdom|2023-01-19 21:34:05|3.95              |1          |1       |0        |
|154132552848742|United Kingdom|2023-01-19 21:34:05|-14.89            |5          |0       |1        |
|154132552848743|EIRE          |2023-01-19 21:34:09|3.3               |2          |1       |0        |
|154132552848744|United Kingdom|2023-01-19 21:34:12|18.2              |12         |1       |0        |
|154132552848745|United Kingdom|2023-01-19 21:34:12|181.19            |68         |1       |0        |
|154132552848746|United Kingdom|2023-01-19 21:34:14|128.55            |59         |1       |0        |
|154132552848747|United Kingdom|2023-01-19 21:34:20|-43.60000000000001|18         |0       |1        |
|154132552848748|United Kingdom|2023-01-19 21:34:22|88.63             |39         |1       |0        |
|154132552848749|United Kingdom|2023-01-19 21:34:23|878.5099999999999 |220        |1       |0        |
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
```

4. **Storing the console output to a file:**
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py > Console-output

5. **Checking the json files:**
   hadoop fs -ls

```
hadoop@ip-172-31-30-66:~
ached minRegisteredResourcesRatio: 0.0
23/01/19 21:36:32 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/hive-site.xml
23/01/19 21:36:32 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.wa
rehouse.dir ('hdfs:///user/spark/warehouse').
23/01/19 21:36:32 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/json.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/execution.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/execution/json.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /static/sql.
23/01/19 21:36:33 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
^C[hadoop@ip-172-31-30-66 ~]$ ls
Console-output  spark-streaming.py
[hadoop@ip-172-31-30-66 ~]$ hadoop fs -ls
Found 5 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:38 .sparkStaging
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:28 time-country-kpi
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:38 time-country-wise-kpi
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:28 time-kpi
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:38 time-wise-kpi
[hadoop@ip-172-31-30-66 ~]$
```

6. **Creating requested folders:**
   mkdir time-wise-kpi
   mkdir time-country-wise-kpi

```
hadoop@ip-172-31-30-66:~
23/01/19 21:36:32 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.wa
rehouse.dir ('hdfs:///user/spark/warehouse').
23/01/19 21:36:32 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/json.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/execution.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /SQL/execution/json.
23/01/19 21:36:32 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to
 /static/sql.
23/01/19 21:36:33 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
^C[hadoop@ip-172-31-30-66 ~]$ ls
Console-output  spark-streaming.py
[hadoop@ip-172-31-30-66 ~]$ hadoop fs -ls
Found 5 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:38 .sparkStaging
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:28 time-country-kpi
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:38 time-country-wise-kpi
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:28 time-kpi
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-01-19 21:38 time-wise-kpi
[hadoop@ip-172-31-30-66 ~]$ mkdir time-wise-kpi
[hadoop@ip-172-31-30-66 ~]$ mkdir time-country-wise-kpi
[hadoop@ip-172-31-30-66 ~]$
```

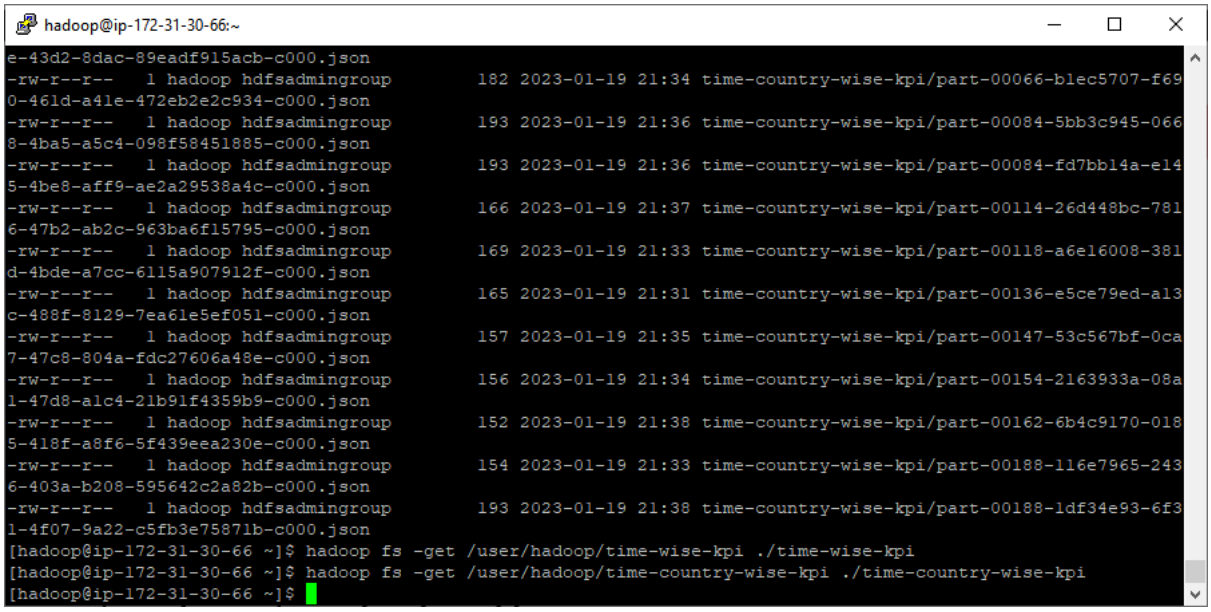
7. **Inspecting json file:**
hadoop fs -ls time-wise-kpi

```
hadoop@ip-172-31-30-66:~
ca4-13e871280ddb-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-01-19 21:34 time-wise-kpi/part-00000-a96713bd-b818-4173-8
bd1-ef1b791ce171-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-01-19 21:30 time-wise-kpi/part-00000-beace5f4-3adf-4026-8
993-53e945dc07f6-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-01-19 21:31 time-wise-kpi/part-00000-d0a97878-2422-480c-8
eld-9ed2f7fde919-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup          0 2023-01-19 21:35 time-wise-kpi/part-00000-de11cfc5-4fcd-4896-9
572-873a7ca38f15-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        173 2023-01-19 21:33 time-wise-kpi/part-00021-fa675467-1170-4acc-a
06a-4cc668cdf169-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        175 2023-01-19 21:34 time-wise-kpi/part-00030-16a8b07c-09bf-4637-b
451-2a66b6e06ad1-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        172 2023-01-19 21:37 time-wise-kpi/part-00036-08a0edc4-cd57-4bf9-b
f7f-4e806f280c26-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        196 2023-01-19 21:32 time-wise-kpi/part-00049-ba50dfe6-228c-43a4-9
838-83a3fd8c03fc-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        212 2023-01-19 21:38 time-wise-kpi/part-00074-1fc3b8b3-00f3-439a-8
1d2-6c1f4ad63d7c-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        211 2023-01-19 21:36 time-wise-kpi/part-00127-e6545901-85bd-4fed-9
998-6cf5e9d342cc-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        172 2023-01-19 21:31 time-wise-kpi/part-00146-dca514b8-a180-4112-a
aaa-7a8c2b097847-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        197 2023-01-19 21:35 time-wise-kpi/part-00147-a7a02626-564d-440f-a
e98-3fe8e49840ce-c000.json
[hadoop@ip-172-31-30-66 ~]$
```

8. **Inspecting json file:**
hadoop fs -ls time-country-wise-kpi

```
hadoop@ip-172-31-30-66:~
2-4ca8-8946-eb72941dacdc-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        166 2023-01-19 21:33 time-country-wise-kpi/part-00049-1289fc2f-3db
e-43d2-8dac-89eadf915acb-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        182 2023-01-19 21:34 time-country-wise-kpi/part-00066-b1ec5707-f69
0-461d-a41e-472eb2e2c934-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        193 2023-01-19 21:36 time-country-wise-kpi/part-00084-5bb3c945-066
8-4ba5-a5c4-098f58451885-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        193 2023-01-19 21:36 time-country-wise-kpi/part-00084-fd7bb14a-e14
5-4be8-aff9-ae2a29538a4c-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        166 2023-01-19 21:37 time-country-wise-kpi/part-00114-26d448bc-781
6-47b2-ab2c-963ba6f15795-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        169 2023-01-19 21:33 time-country-wise-kpi/part-00118-a6e16008-381
d-4bde-a7cc-6115a907912f-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        165 2023-01-19 21:31 time-country-wise-kpi/part-00136-e5ce79ed-a13
c-488f-8129-7ea61e5ef051-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        157 2023-01-19 21:35 time-country-wise-kpi/part-00147-53c567bf-0ca
7-47c8-804a-fdc27606a48e-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        156 2023-01-19 21:34 time-country-wise-kpi/part-00154-2163933a-08a
1-47d8-a1c4-21b91f4359b9-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        152 2023-01-19 21:38 time-country-wise-kpi/part-00162-6b4c9170-018
5-418f-a8f6-5f439eea230e-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        154 2023-01-19 21:33 time-country-wise-kpi/part-00188-116e7965-243
6-403a-b208-595642c2a82b-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup        193 2023-01-19 21:38 time-country-wise-kpi/part-00188-1df34e93-6f3
1-4f07-9a22-c5fb3e75871b-c000.json
[hadoop@ip-172-31-30-66 ~]$
```

9. **Transfer of data from HDFS to Local file system:**
   hadoop fs -get /user/hadoop/time-wise-kpi ./time-wise-kpi
   hadoop fs -get /user/hadoop/time-country-wise-kpi ./time-country-wise-kpi



10. **Using WinSCP, copying above json data, console output file and spark-streaming.py to my local machine.**