

Code Logic - Retail Data Analysis

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

The code is used for streaming data processing using PySpark.

It reads data from a Kafka topic and performs some calculations to produce various key performance indicators (KPIs) for sales data. It then writes the KPIs to HDFS in JSON format.

Here is a high-level overview of the code:

- First, the necessary PySpark libraries are imported, and a SparkSession is created.
- The code reads streaming data from a Kafka topic using `spark.readStream` and the options provided in the `.format`, `.option`, and `.load` methods. The data is then parsed using a pre-defined schema and the `from_json` function, which converts a JSON string to a struct or array of structs.
- Several utility methods are defined to calculate various metrics such as total cost, total items, and flags to denote whether an order is new or a return.
- These utility methods are then converted to user-defined functions (UDFs) using `udf`.
- The `withColumn` method is used to add new columns to the DataFrame containing the streaming data. These columns are computed using the UDFs defined earlier.
- The resulting DataFrame is then written to the console using `writeStream`, with the output mode set to append and the trigger set to every 1 minute.
- Two more DataFrames are created by grouping and aggregating the data by time and country, respectively. These DataFrames are used to compute various time-based and time-and-country-based KPIs.
- These KPIs are then written to HDFS in JSON format using `writeStream`.

Note that the code also includes some additional options such as `checkpointLocation` and `watermark`. These are used to ensure fault tolerance and to handle late data in the stream, respectively.

Steps taken to solve the project:


First, I create an EC2 instance and add inbound rules to the security group to allow access to the necessary ports (2181, 9092, 9000, 8080, 8888). Then, I create an EMR instance.

Next, I write the necessary codes to start the ZooKeeper and Kafka servers.


Then, I write code to create a producer and consumer in Kafka to receive streaming data.

Finally, I submit a Spark job using the spark-submit command in EMR to process the streaming data with the Python file (spark-streaming.py).


Screenshots of hadoop commands to get the json files.

 hadoop@ip-172-31-17-82:~

```
[hadoop@ip-172-31-17-82 ~]$ hadoop fs -ls
Found 5 items
drwxr-xr-x - hadoop hadoop 0 2023-03-20 11:00 .sparkStaging
drwxr-xr-x - hadoop hadoop 0 2023-03-20 11:22 country_kpi
drwxr-xr-x - hadoop hadoop 0 2023-03-20 10:56 country_kpi_checkpoints
drwxr-xr-x - hadoop hadoop 0 2023-03-20 11:22 time_kpi
drwxr-xr-x - hadoop hadoop 0 2023-03-20 10:56 time_kpi_checkpoints
[hadoop@ip-172-31-17-82 ~]$
```

 hadoop@ip-172-31-17-82:~

```
[hadoop@ip-172-31-17-82 ~]$ hadoop fs -cat time_kpi/part-00120-e83b02e0-6c96-47f0-9e53-963ce9f8c75f-c000.json
{"window":{"start":"2023-03-19T13:12:00.000Z","end":"2023-03-19T13:13:00.000Z"},"OPM":9,"total_sale_volume":134.4099973142147,"average_transaction_size":14.934444146023
857,"rate_of_return":0.1111111111111111}
{"window":{"start":"2023-03-19T21:58:00.000Z","end":"2023-03-19T21:59:00.000Z"},"OPM":12,"total_sale_volume":193.8300004005432,"average_transaction_size":16.15250003337
86,"rate_of_return":0.0}
{"window":{"start":"2023-03-19T18:22:00.000Z","end":"2023-03-19T18:23:00.000Z"},"OPM":5,"total_sale_volume":47.049999713097705,"average_transaction_size":9.40999942779
54,"rate_of_return":0.0}
{"window":{"start":"2023-03-19T15:40:00.000Z","end":"2023-03-19T15:41:00.000Z"},"OPM":10,"total_sale_volume":98.39999902248383,"average_transaction_size":9.839999902248
383,"rate_of_return":0.0}
[hadoop@ip-172-31-17-82 ~]$
```

 hadoop@ip-172-31-17-82:~

```
[hadoop@ip-172-31-17-82 ~]$ hadoop fs -cat country_kpi/part-00070-fd323d18-357a-458c-89a1-3dc26b9af3c4-c000.json
{"window":{"start":"2023-03-20T06:51:00.000Z","end":"2023-03-20T06:52:00.000Z"},"country":"EIRE","OPM":1,"total_sale_volume":8.5,"rate_of_return":0.0}
{"window":{"start":"2023-03-20T09:51:00.000Z","end":"2023-03-20T09:52:00.000Z"},"country":"United Kingdom","OPM":12,"total_sale_volume":168.2099964618683,"rate_of_retur
n":0.0}
{"window":{"start":"2023-03-20T04:02:00.000Z","end":"2023-03-20T04:03:00.000Z"},"country":"United Kingdom","OPM":13,"total_sale_volume":271.3799958229065,"rate_of_retur
n":0.07692307692307693}
{"window":{"start":"2023-03-19T18:57:00.000Z","end":"2023-03-19T18:58:00.000Z"},"country":"United Kingdom","OPM":11,"total_sale_volume":130.74999752640724,"rate_of_retu
rn":0.09090909090909091}
{"window":{"start":"2023-03-20T06:53:00.000Z","end":"2023-03-20T06:54:00.000Z"},"country":"United Kingdom","OPM":3,"total_sale_volume":186.53999280929565,"rate_of_retur
n":0.0}
{"window":{"start":"2023-03-20T07:52:00.000Z","end":"2023-03-20T07:53:00.000Z"},"country":"United Kingdom","OPM":12,"total_sale_volume":90.39999628067017,"rate_of_retur
n":0.16666666666666666}
{"window":{"start":"2023-03-19T20:45:00.000Z","end":"2023-03-19T20:46:00.000Z"},"country":"Austria","OPM":1,"total_sale_volume":10.5,"rate_of_return":0.0}
{"window":{"start":"2023-03-20T02:05:00.000Z","end":"2023-03-20T02:06:00.000Z"},"country":"United Kingdom","OPM":7,"total_sale_volume":84.21999835968018,"rate_of_return
":0.14285714285714285}
{"window":{"start":"2023-03-20T06:06:00.000Z","end":"2023-03-20T06:07:00.000Z"},"country":"Channel Islands","OPM":1,"total_sale_volume":15.0,"rate_of_return":0.0}
{"window":{"start":"2023-03-20T00:51:00.000Z","end":"2023-03-20T00:52:00.000Z"},"country":"Australia","OPM":1,"total_sale_volume":2.370000123977661,"rate_of_return":0.0}
{"window":{"start":"2023-03-19T18:34:00.000Z","end":"2023-03-19T18:35:00.000Z"},"country":"United Kingdom","OPM":9,"total_sale_volume":46.699997156858444,"rate_of_retur
n":0.1111111111111111}
{"window":{"start":"2023-03-19T17:55:00.000Z","end":"2023-03-19T17:56:00.000Z"},"country":"France","OPM":1,"total_sale_volume":12.75,"rate_of_return":0.0}
{"window":{"start":"2023-03-19T15:54:00.000Z","end":"2023-03-19T15:55:00.000Z"},"country":"France","OPM":1,"total_sale_volume":25.0,"rate_of_return":0.0}
{"window":{"start":"2023-03-19T15:37:00.000Z","end":"2023-03-19T15:38:00.000Z"},"country":"United Kingdom","OPM":8,"total_sale_volume":211.12999844551086,"rate_of_retur
n":0.0}
[hadoop@ip-172-31-17-82 ~]$
```

```

-rw-r--r-- 1 hadoop hadoop 1201 2023-03-20 10:57 time_kpi/part-00199-daab4727-030e-4b06-a42a-571883c71134-c000.json
[hadoop@ip-172-31-17-82 ~]$ clear
[hadoop@ip-172-31-17-82 ~]$ hadoop fs -ls time_kpi
Found 255 items
drwxr-xr-x - hadoop hadoop 0 2023-03-20 11:25 time_kpi/_spark_metadata
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:24 time_kpi/part-00000-07fb091f-cec1-4f8e-912c-72681bc6ca5a-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:05 time_kpi/part-00000-0c085b46-0d5f-4d49-9d7a-764f1c792255-c000.json
-rw-r--r-- 1 hadoop hadoop 974 2023-03-20 10:57 time_kpi/part-00000-27355df2-042c-4648-9b4a-5263433372a3-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:03 time_kpi/part-00000-2bc1d093-2acd-4599-a5a5-222c0b97711d-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:07 time_kpi/part-00000-30116b7e-8937-4f6b-a6bd-al24f4541f09-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:22 time_kpi/part-00000-32828f16-015d-4ddc-a741-a96bcd37f8f9-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:01 time_kpi/part-00000-3d376300-8d55-4f5a-bdc7-3a7dceed5df4-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 10:58 time_kpi/part-00000-41fbd47e-78fb-4a09-8a97-4b6ef95a634c-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:10 time_kpi/part-00000-4df47904-f4b2-43bd-bcc7-a199af05cce8-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:12 time_kpi/part-00000-4fd6a452-426f-4fef-9595-38d8ffa6e65e-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:23 time_kpi/part-00000-5e0cle81-d460-4853-bcb0-721f1b8a2a65-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:16 time_kpi/part-00000-650ba780-f061-4068-bb48-e6f916c2e8f8-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:18 time_kpi/part-00000-77b51b5a-e631-4223-a096-eeb27a6befbd-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:15 time_kpi/part-00000-79958alb-33a2-4b56-ad26-721f0055b313-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:14 time_kpi/part-00000-a28353c3-f86d-4b6b-b935-a2add9f9eaf-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:13 time_kpi/part-00000-a36c3da2-d7bc-41fd-85b0-550b2ef503da-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:20 time_kpi/part-00000-a8d6bc81-f70b-4cf9-9119-cf0ea6a5ecf0-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:17 time_kpi/part-00000-acl332e1-935f-4ca4-97a4-ea5576e4d697-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:02 time_kpi/part-00000-ace6a758-f20f-45ae-9d57-ace90e2768de-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:21 time_kpi/part-00000-b362f2cc-e51c-4ecc-a6e5-abeb37b82f34-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:25 time_kpi/part-00000-bf33fcd2-67cd-4aa1-9dac-a7be18e96ade-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:04 time_kpi/part-00000-c2b07268-959b-4166-988e-2e18392902fd-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:19 time_kpi/part-00000-e153f40e-aea9-4a8d-a8bb-f3c5b90a89f5-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:06 time_kpi/part-00000-ebb0c5ec-7e75-4547-bbf4-5b579b35fcc6-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:09 time_kpi/part-00000-f23e0d7c-lbe0-4f09-95d3-f0f41438af0e-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:08 time_kpi/part-00000-f60d19ff-a5e8-4af0-84ae-f426d32c63c9-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:00 time_kpi/part-00000-f7fd0d76-26c8-4d2b-a95c-e7cdb358f35c-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:11 time_kpi/part-00000-fa610398-961f-49c7-ae30-2d2223b0e30-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 10:56 time_kpi/part-00000-fe48fb23-7bdc-4e84-862b-862b5fbae9cc-c000.json
-rw-r--r-- 1 hadoop hadoop 617 2023-03-20 10:57 time_kpi/part-00001-dcl4c4f1-6c84-4460-8979-1c9d85982ecc-c000.json
-rw-r--r-- 1 hadoop hadoop 1802 2023-03-20 10:57 time_kpi/part-00002-5a2f881a-6c2d-4d1a-abcf-1034010a030f-c000.json
-rw-r--r-- 1 hadoop hadoop 194 2023-03-20 11:06 time_kpi/part-00002-6e4c169f-9154-49b3-83bb-f0026d2fdcc0-c000.json
-rw-r--r-- 1 hadoop hadoop 195 2023-03-20 11:24 time_kpi/part-00002-ce9a05ca-5567-4eca-887d-916feb42151f-c000.json
-rw-r--r-- 1 hadoop hadoop 812 2023-03-20 10:57 time_kpi/part-00003-13bc6bc3-2aec-4df6-ae6b-0917239487f2-c000.json
-rw-r--r-- 1 hadoop hadoop 601 2023-03-20 10:57 time_kpi/part-00004-0600d96f-dcb1-430c-82a9-67bd7b3ee913-c000.json

```

```

-rw-r--r-- 1 hadoop hadoop 1201 2023-03-20 10:57 time_kpi/part-00199-daab4727-030e-4b06-a42a-571883c71134-c000.json
[hadoop@ip-172-31-17-82 ~]$ hadoop fs -ls country_kpi
Found 268 items
drwxr-xr-x - hadoop hadoop 0 2023-03-20 11:25 country_kpi/_spark_metadata
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:22 country_kpi/part-00000-03f5c99b-c53a-4b87-a192-d1887c956107-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:16 country_kpi/part-00000-24b88f74-ab52-4aca-ae59-01a4e3a3077e-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:01 country_kpi/part-00000-25e16b21-db83-4ee6-a799-5a4a84c10424-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:04 country_kpi/part-00000-2a240c15-d670-4a7c-aa0d-f16a6ded258f-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:02 country_kpi/part-00000-35e03911-c5aa-48ab-88f6-9026c671ae47-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:06 country_kpi/part-00000-3d8e96e3-a840-4c08-8260-b0d57cf4c326-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 10:58 country_kpi/part-00000-3dd21ee3-a7c5-4970-89c9-026dae71f911-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:03 country_kpi/part-00000-3e1b70ef-3617-4145-b34b-50dfb78af2a2-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:09 country_kpi/part-00000-4703f514-a1db-4d7f-8fb6-4cdeb196f302-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:23 country_kpi/part-00000-47793aed-407a-4f7c-906e-7e414d5e897b-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:21 country_kpi/part-00000-4d8dc727-6049-4fe2-8fb8-a68f43b4b420-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:14 country_kpi/part-00000-4da2b0dd-e4fd-4c42-93b6-b03366df68aa-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:05 country_kpi/part-00000-5011d3f7-07c7-4ded-a88f-75e2f71f86b6-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:12 country_kpi/part-00000-5670352a-6e90-4790-bf0e-c42e2b2baecd-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:25 country_kpi/part-00000-56d8a92f-205f-4065-8aa2-4f89f5041124-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:13 country_kpi/part-00000-6195a804-c6b2-4ca8-88c9-8e2e7ccc8cee-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:08 country_kpi/part-00000-627c5bb7-71dc-4d4d-bb12-b8227d340f74-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:17 country_kpi/part-00000-72c28164-0bde-4c64-b766-226839bfd365-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:19 country_kpi/part-00000-87394004-4624-4c0e-9db5-2a59f69ff946-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:11 country_kpi/part-00000-a57de797-2e82-41f8-8a7e-4ccca3fcbf7-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:00 country_kpi/part-00000-b1600ded-d9e6-4b1d-bd02-def2202fdbeb-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:20 country_kpi/part-00000-b4afdd2d-1162-4de4-a93a-34f9e635950a-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:07 country_kpi/part-00000-b926d899-4ac1-4c28-ba7c-b2bb610810dd-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 10:56 country_kpi/part-00000-bc7f5062-d627-40c3-a1e1-d150e894d3d5-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2023-03-20 11:10 country_kpi/part-00000-d6ff54ab-1126-4c6a-81de-alb0a31a6342-c000.json

```