Code Logic - Retail Data Analysis

In this document, the code and the overall steps taken to solve the project are described:

1. All necessary libraries are included and spark session is established

```
from pyspark.sql import
SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
import pyspark.sql.functions as F
```

2. Connected to the kafka server where streaming data is available along with topic name from where the data is read

```
lines = spark  \
        .readStream  \
        .format("kafka")  \
        .option("kafka.bootstrap.servers","18.211.252.152:9092")  \
        .option("subscribe","real-time-project")  \
        .option("failOnDataLoss","false") \
        .option("startingOffsets", "earliest")  \
        .load()
```

3. Since now the data is read from server, we now make it more readable by mapping lines of streaming data into a schema, along with data types against each column.
   Columns in raw data are :
   "country", "invoice_no", "timestamp", "type", "items"
   which is an array which is sub-divided into:
   "SKU", "title", "unit_price", "quantity" for each occurrence.
4. Unveil array items now as individual array items are used in KPI and new column derivations. Use option explode for this purpose.

```
df1 = new_df.select(col("type"),col("country"),col("invoice_no"),col("time
stamp"),explode(col("items")))
```

5. Rename columns like SKU , title, unit price & quantity which were earlier represented as col.SKU or col.title.

6. Defining UDF’s and executing them.

   i. **UDF 1**: This function accepts “type” as input and determines if the nature of request is “Order” or “Return” If Type comes as “ORDER”, then it returns a flag **is_order = 1** which says it is an order.

   ii. **UDF 2**: This function accepts “type” as input and determines if the nature of request is “Order” or “Return” If Type comes as “RETURN”, then it returns a flag **is_return = 1** which says it is a return request.

   iii. **UDF 3:** This function accepts 3 attributes in input. Unit price, quantity & type as input. If type is “Order”, it returns a value of Unit price multiplied by quantity. If type is “Return”, it returns a value of Unit price multiplied by quantity toggled by “-“ symbol.

7. Since we calculated all new columns required, we need to add them to our existing data frame. Output the data with new columns for each order for window of 1 minute.

```
df3 =
df2.withWatermark("timestamp","10
minutes") \
.groupby(window("timestamp","1
minute"),"invoice_no","country","is_Or
der","is_Return") \
.sum("Cost","quantity")
```

   Output from this step is used for calculating KPI’s in next steps.

8. **Calculation of KPI’s:**

   i. **Orders per Minute:** Calculated as count of distinct invoices as below >

```
F.approx_count_distinct("invoice_no").alias("OPM")
```

   ii. **Total volume of sales**: Calculated from UDF3 where we derived cost value of order. Use this cost value and get a sum of all order cost’s for window period to get total volume of sales:

```
agg(sum("Cost").alias("Total_sales_vol")
```

   iii. **Rate of Return:** For this KPI, we utilize is_return & is_order flags derived from UDF1 & UDF2. We take sum of is_return & is_order for window period and store them as total_Order & total_return. Below formula is used to calculate Rate of Return. Note that same formula is used twice once after grouping by country to get country specific rate of return. Second time to generate time based rate of return.

```
Final_time =
Final_time.withColumn("rate_of_return",Final_time.total_return
/(Final_time.total_Order+Final_time.total_return))
```

iv. **Average Transaction Size:** For this KPI, we utilize is_return & is_order flags derived from UDF1 & UDF2. We take sum of is_return & is_order for window period and store them as total_Order & total_return. Below formula is used to calculate average transaction size. This needs to be calculated only on time basis, not country basis.

```
Final_time.withColumn("Avg_trans_size",Final_ti
me.Total_sales_vol/(Final_time.total_Order
+Final_time.total_return))
```

9. Now that KPI calculation is done,

i. printed time based KPI to HDFS path as a JSON file using below code.

```
query_2 = Final_time.writeStream \
.outputMode("Append") \
.format("json") \
.option("format","append") \
.option("truncate", "false") \
.option("path","Timebased-KPI") \
.option("checkpointLocation", "time_KPI_json") \
.trigger(processingTime="1 minute") \
.start()
```

Here, HDFS location path is given as **Timebased-KPI** where KPI data shall be written to.

ii. Printed time and country based KPI to HDFS path as a JSON file using below code.

```
query_3 = Final_country_time.writeStream \
.outputMode("Append") \ .format("json") \
.option("format","append") \ .option("truncate", "false")
\ .option("path"," Country-and-timebased-KPI" \
.option("checkpointLocation",
"time_country_KPI_json") \
.trigger(processingTime="1 minute") \
.start()
```

Here, HDFS location path is given as **Country-and-timebased-KPI** where KPI data shall be written to.

10. Created EMR cluster and YARN configuration using following applications:

```
JupyterHub 1.1.0, Spark 2.4.5, Zeppelin 0.8.2, Livy 0.7.0
```

11. Logged in to the EMR instance and below command is executed to download Spark-SQL-Kafka jar file. This jar is used to run the Spark Streaming-Kafka codes

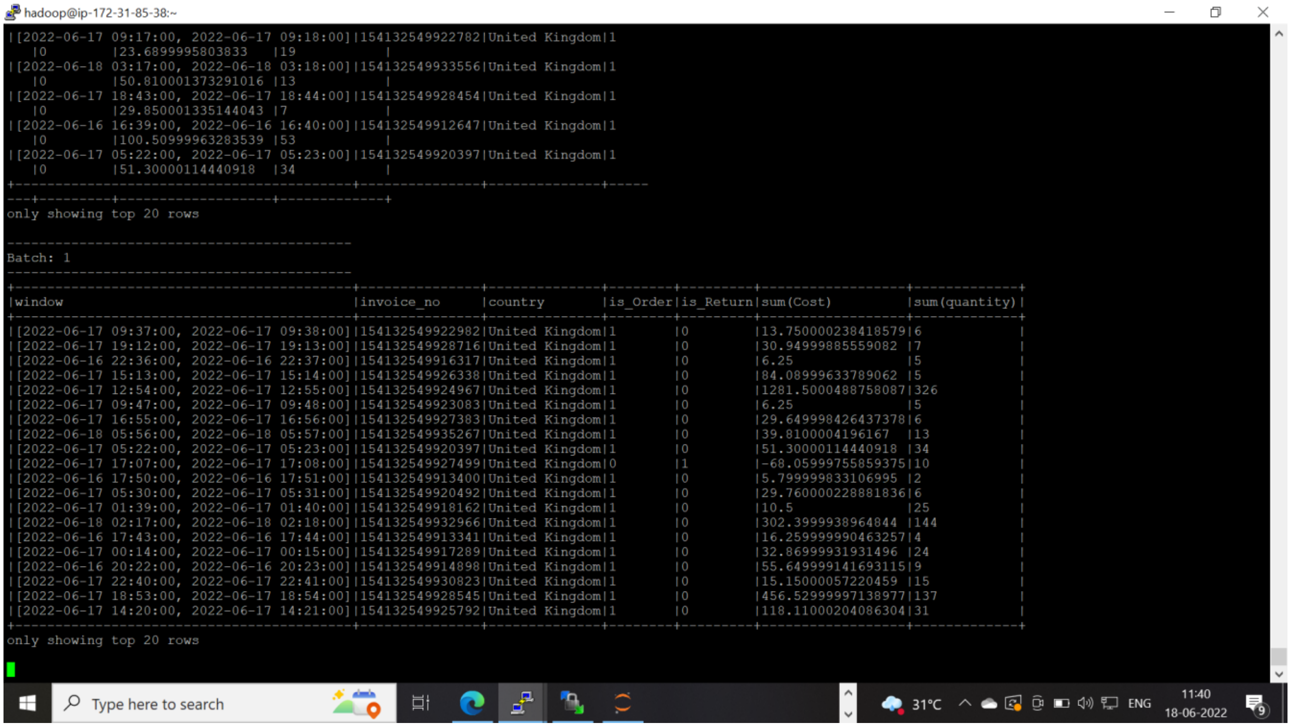
```
wget https://ds-spark-sql-kafka-jar.s3.amazonaws.com/spark-sql-kafka-0-10_2.11-2.3.0.jar
```

12. Kafka version is set using the following command:

```
export SPARK_KAFKA_VERSION=0.10
```

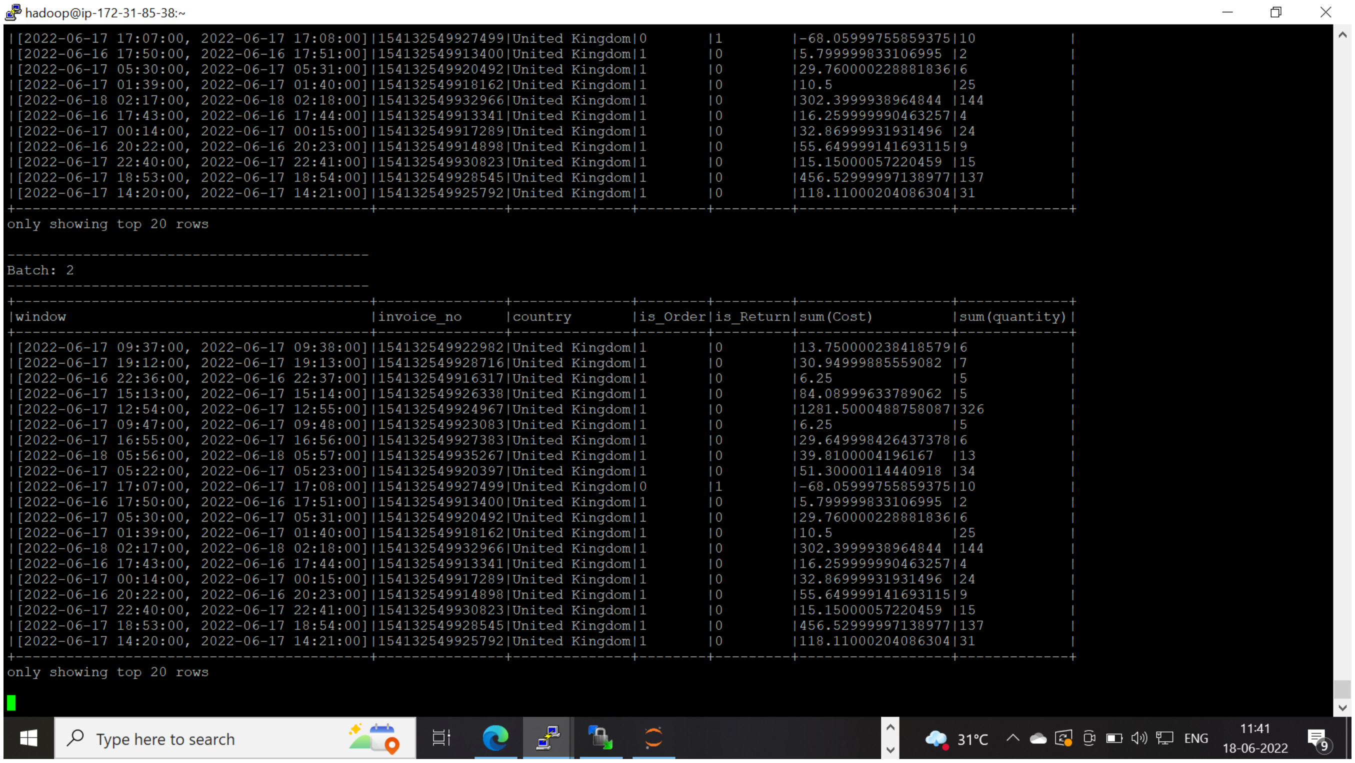
13. Submitted the spark job using command below:

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
spark-streaming.py 18.211.252.152 9092 real-time-project
```



```
|[2022-06-17 09:17:00, 2022-06-17 09:18:00]|154132549922782|United Kingdom|1
   |0        |23.6899995803833  |19           |
|[2022-06-18 03:17:00, 2022-06-18 03:18:00]|154132549933556|United Kingdom|1
   |0        |50.810001373291016 |13           |
|[2022-06-17 18:43:00, 2022-06-17 18:44:00]|154132549928454|United Kingdom|1
   |0        |29.850001335144043 |7            |
|[2022-06-16 16:39:00, 2022-06-16 16:40:00]|154132549912647|United Kingdom|1
   |0        |100.50999963283539 |53           |
|[2022-06-17 05:22:00, 2022-06-17 05:23:00]|154132549920397|United Kingdom|1
   |0        |51.30000114440918  |34           |
+------------------------------------------+---------------+--------------+-----
---+---------+-------------------+-------------+
only showing top 20 rows

-------------------------------------------
Batch: 1
-------------------------------------------
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
|window                                    |invoice_no     |country       |is_Order|is_Return|sum(Cost)         |sum(quantity)|
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
|[2022-06-17 09:37:00, 2022-06-17 09:38:00]|154132549922982|United Kingdom|1       |0        |13.750000238418579|6            |
|[2022-06-17 19:12:00, 2022-06-17 19:13:00]|154132549928716|United Kingdom|1       |0        |30.94999885559082 |7            |
|[2022-06-16 22:36:00, 2022-06-16 22:37:00]|154132549916317|United Kingdom|1       |0        |6.25              |5            |
|[2022-06-17 15:13:00, 2022-06-17 15:14:00]|154132549926338|United Kingdom|1       |0        |84.08999633789062 |5            |
|[2022-06-17 12:54:00, 2022-06-17 12:55:00]|154132549924967|United Kingdom|1       |0        |1281.5000488758087|326          |
|[2022-06-17 09:47:00, 2022-06-17 09:48:00]|154132549923083|United Kingdom|1       |0        |6.25              |5            |
|[2022-06-17 16:55:00, 2022-06-17 16:56:00]|154132549927383|United Kingdom|1       |0        |29.649998426437378|6            |
|[2022-06-18 05:56:00, 2022-06-18 05:57:00]|154132549935267|United Kingdom|1       |0        |39.8100004196167  |13           |
|[2022-06-17 05:22:00, 2022-06-17 05:23:00]|154132549920397|United Kingdom|1       |0        |51.30000114440918 |34           |
|[2022-06-17 17:07:00, 2022-06-17 17:08:00]|154132549927499|United Kingdom|0       |1        |-68.05999755859375|10           |
|[2022-06-16 17:50:00, 2022-06-16 17:51:00]|154132549913400|United Kingdom|1       |0        |5.799999833106995 |2            |
|[2022-06-17 05:30:00, 2022-06-17 05:31:00]|154132549920492|United Kingdom|1       |0        |29.760000228881836|6            |
|[2022-06-17 01:39:00, 2022-06-17 01:40:00]|154132549918162|United Kingdom|1       |0        |10.5              |25           |
|[2022-06-18 02:17:00, 2022-06-18 02:18:00]|154132549932966|United Kingdom|1       |0        |302.3999938964844 |144          |
|[2022-06-16 17:43:00, 2022-06-16 17:44:00]|154132549913341|United Kingdom|1       |0        |16.259999990463257|4            |
|[2022-06-17 00:14:00, 2022-06-17 00:15:00]|154132549917289|United Kingdom|1       |0        |32.86999931931496 |24           |
|[2022-06-16 20:22:00, 2022-06-16 20:23:00]|154132549914898|United Kingdom|1       |0        |55.649999141693115|9            |
|[2022-06-17 22:40:00, 2022-06-17 22:41:00]|154132549930823|United Kingdom|1       |0        |15.15000057220459 |15           |
|[2022-06-17 18:53:00, 2022-06-17 18:54:00]|154132549928545|United Kingdom|1       |0        |456.5299999713897 |137          |
|[2022-06-17 14:20:00, 2022-06-17 14:21:00]|154132549925792|United Kingdom|1       |0        |118.1100020408630 |31           |
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
only showing top 20 rows
```



```
|[2022-06-17 17:07:00, 2022-06-17 17:08:00]|154132549927499|United Kingdom|0       |1        |-68.05999755859375|10           |
|[2022-06-16 17:50:00, 2022-06-16 17:51:00]|154132549913400|United Kingdom|1       |0        |5.799999833106995 |2            |
|[2022-06-17 05:30:00, 2022-06-17 05:31:00]|154132549920492|United Kingdom|1       |0        |29.760000228881836|6            |
|[2022-06-17 01:39:00, 2022-06-17 01:40:00]|154132549918162|United Kingdom|1       |0        |10.5              |25           |
|[2022-06-18 02:17:00, 2022-06-18 02:18:00]|154132549932966|United Kingdom|1       |0        |302.3999938964844 |144          |
|[2022-06-16 17:43:00, 2022-06-16 17:44:00]|154132549913341|United Kingdom|1       |0        |16.259999990463257|4            |
|[2022-06-17 00:14:00, 2022-06-17 00:15:00]|154132549917289|United Kingdom|1       |0        |32.86999931931496 |24           |
|[2022-06-16 20:22:00, 2022-06-16 20:23:00]|154132549914898|United Kingdom|1       |0        |55.649999141693115|9            |
|[2022-06-17 22:40:00, 2022-06-17 22:41:00]|154132549930823|United Kingdom|1       |0        |15.15000057220459 |15           |
|[2022-06-17 18:53:00, 2022-06-17 18:54:00]|154132549928545|United Kingdom|1       |0        |456.52999997138977|137          |
|[2022-06-17 14:20:00, 2022-06-17 14:21:00]|154132549925792|United Kingdom|1       |0        |118.11000204086304|31           |
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
only showing top 20 rows

-------------------------------------------
Batch: 2
-------------------------------------------
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
|window                                    |invoice_no     |country       |is_Order|is_Return|sum(Cost)         |sum(quantity)|
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
|[2022-06-17 09:37:00, 2022-06-17 09:38:00]|154132549922982|United Kingdom|1       |0        |13.750000238418579|6            |
|[2022-06-17 19:12:00, 2022-06-17 19:13:00]|154132549928716|United Kingdom|1       |0        |30.94999885559082 |7            |
|[2022-06-16 22:36:00, 2022-06-16 22:37:00]|154132549916317|United Kingdom|1       |0        |6.25              |5            |
|[2022-06-17 15:13:00, 2022-06-17 15:14:00]|154132549926338|United Kingdom|1       |0        |84.08999633789062 |5            |
|[2022-06-17 12:54:00, 2022-06-17 12:55:00]|154132549924967|United Kingdom|1       |0        |1281.5000488758087|326          |
|[2022-06-17 09:47:00, 2022-06-17 09:48:00]|154132549923083|United Kingdom|1       |0        |6.25              |5            |
|[2022-06-17 16:55:00, 2022-06-17 16:56:00]|154132549927383|United Kingdom|1       |0        |29.649998426437378|6            |
|[2022-06-18 05:56:00, 2022-06-18 05:57:00]|154132549935267|United Kingdom|1       |0        |39.8100004196167  |13           |
|[2022-06-17 05:22:00, 2022-06-17 05:23:00]|154132549920397|United Kingdom|1       |0        |51.30000114440918 |34           |
|[2022-06-17 17:07:00, 2022-06-17 17:08:00]|154132549927499|United Kingdom|0       |1        |-68.05999755859375|10           |
|[2022-06-16 17:50:00, 2022-06-16 17:51:00]|154132549913400|United Kingdom|1       |0        |5.799999833106995 |2            |
|[2022-06-17 05:30:00, 2022-06-17 05:31:00]|154132549920492|United Kingdom|1       |0        |29.760000228881836|6            |
|[2022-06-17 01:39:00, 2022-06-17 01:40:00]|154132549918162|United Kingdom|1       |0        |10.5              |25           |
|[2022-06-18 02:17:00, 2022-06-18 02:18:00]|154132549932966|United Kingdom|1       |0        |302.3999938964844 |144          |
|[2022-06-16 17:43:00, 2022-06-16 17:44:00]|154132549913341|United Kingdom|1       |0        |16.259999990463257|4            |
|[2022-06-17 00:14:00, 2022-06-17 00:15:00]|154132549917289|United Kingdom|1       |0        |32.86999931931496 |24           |
|[2022-06-16 20:22:00, 2022-06-16 20:23:00]|154132549914898|United Kingdom|1       |0        |55.649999141693115|9            |
|[2022-06-17 22:40:00, 2022-06-17 22:41:00]|154132549930823|United Kingdom|1       |0        |15.15000057220459 |15           |
|[2022-06-17 18:53:00, 2022-06-17 18:54:00]|154132549928545|United Kingdom|1       |0        |456.52999997138977|137          |
|[2022-06-17 14:20:00, 2022-06-17 14:21:00]|154132549925792|United Kingdom|1       |0        |118.11000204086304|31           |
+------------------------------------------+---------------+--------------+--------+---------+------------------+-------------+
only showing top 20 rows
```

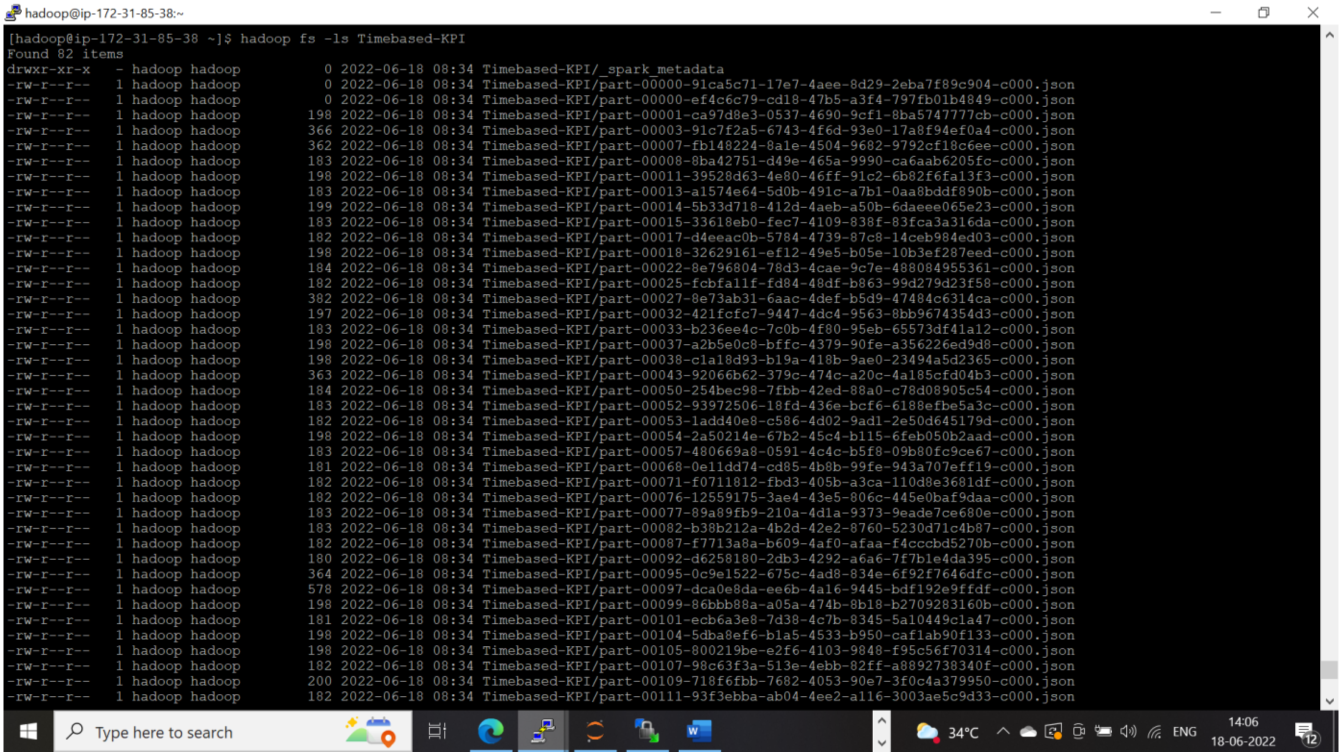
14. Checked in HDFS to make sure the KPI files were present:

```
hadoop fs -get Timebased-KPI
```

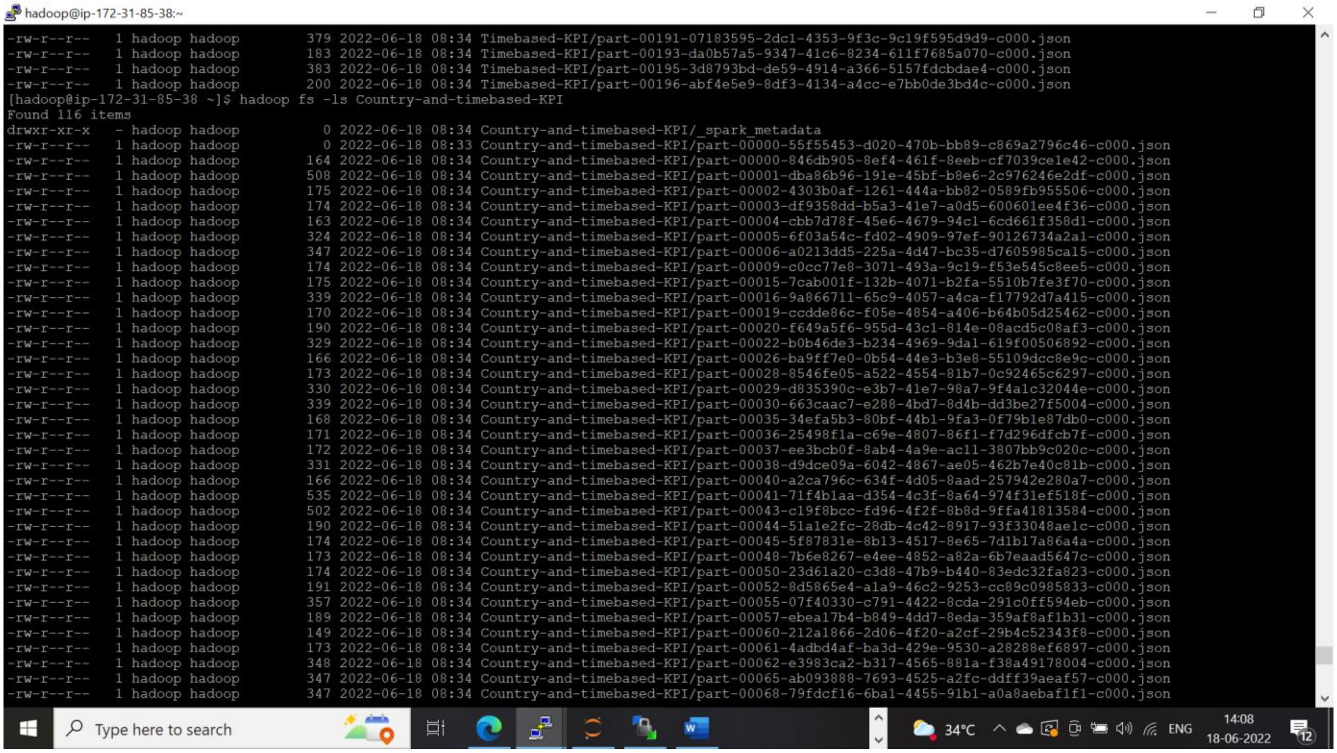
```
hadoop fs -ls Timebased-KPI
```

```
hadoop fs - get Country-and-
timebased-KPI
```

```
hadoop fs -ls Country-and-
timebased-KPI
```



```
[hadoop@ip-172-31-85-38 ~]$ hadoop fs -ls Timebased-KPI
Found 82 items
drwxr-xr-x   - hadoop hadoop          0 2022-06-18 08:34 Timebased-KPI/_spark_metadata
-rw-r--r--   1 hadoop hadoop          0 2022-06-18 08:34 Timebased-KPI/part-00000-91ca5c71-17e7-4aee-8d29-2eba7f89c904-c000.json
-rw-r--r--   1 hadoop hadoop          0 2022-06-18 08:34 Timebased-KPI/part-00000-ef4c6c79-cd18-47b5-a3f4-797fb01b4849-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00001-ca97d8e3-0537-4690-9cf1-8ba5747777cb-c000.json
-rw-r--r--   1 hadoop hadoop        366 2022-06-18 08:34 Timebased-KPI/part-00003-91c7f2a5-6743-4f6d-93e0-17a8f94ef0a4-c000.json
-rw-r--r--   1 hadoop hadoop        362 2022-06-18 08:34 Timebased-KPI/part-00007-fb148224-8a1e-4504-9682-9792cf18c6ee-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00008-8ba42751-d49e-465a-9990-ca6aab6205fc-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00011-39528d63-4e80-46ff-91c2-6b82f6fa13f3-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00013-a1574e64-5d0b-491c-a7b1-0aa8bddf890b-c000.json
-rw-r--r--   1 hadoop hadoop        199 2022-06-18 08:34 Timebased-KPI/part-00014-5b33d718-412d-4aeb-a50b-6daeee065e23-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00015-33618eb0-fec7-4109-838f-83fca3a316da-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00017-d4eeac0b-5784-4739-87c8-14ceb984ed03-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00018-32629161-ef12-49e5-b05e-10b3ef287eed-c000.json
-rw-r--r--   1 hadoop hadoop        184 2022-06-18 08:34 Timebased-KPI/part-00022-8e796804-78d3-4cae-9c7e-488084955361-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00025-fcbfa11f-fd84-48df-b863-99d279d23f58-c000.json
-rw-r--r--   1 hadoop hadoop        382 2022-06-18 08:34 Timebased-KPI/part-00027-8e73ab31-6aac-4def-b5d9-47484c6314ca-c000.json
-rw-r--r--   1 hadoop hadoop        197 2022-06-18 08:34 Timebased-KPI/part-00032-421fcfc7-9447-4dc4-9563-8bb9674354d3-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00033-b236ee4c-7c0b-4f80-95eb-65573df41a12-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00037-a2b5e0c8-bffc-4379-90fe-a356226ed9d8-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00038-c1a18d93-b19a-418b-9ae0-23494a5d2365-c000.json
-rw-r--r--   1 hadoop hadoop        363 2022-06-18 08:34 Timebased-KPI/part-00043-92066b62-379c-474c-a20c-4a185cfd04b3-c000.json
-rw-r--r--   1 hadoop hadoop        184 2022-06-18 08:34 Timebased-KPI/part-00050-254bec98-7fbb-42ed-88a0-c78d08905c54-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00052-93972506-18fd-436e-bcf6-6188efbe5a3c-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00053-1add40e8-c586-4d02-9ad1-2e50d645179d-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00054-2a50214e-67b2-45c4-b115-6feb050b2aad-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00057-480669a8-0591-4c4c-b5f8-09b80fc9ce67-c000.json
-rw-r--r--   1 hadoop hadoop        181 2022-06-18 08:34 Timebased-KPI/part-00068-0e11dd74-cd85-4b8b-99fe-943a707eff19-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00071-f0711812-fbd3-405b-a3ca-110d8e3681df-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00076-12559175-3ae4-43e5-806c-445e0baf9daa-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00077-89a89fb9-210a-4d1a-9373-9eade7ce680e-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00082-b38b212a-4b2d-42e2-8760-5230d71c4b87-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00087-f7713a8a-b609-4af0-afaa-f4cccbd5270b-c000.json
-rw-r--r--   1 hadoop hadoop        180 2022-06-18 08:34 Timebased-KPI/part-00092-d6258180-2db3-4292-a6a6-7f7b1e4da395-c000.json
-rw-r--r--   1 hadoop hadoop        364 2022-06-18 08:34 Timebased-KPI/part-00095-0c9e1522-675c-4ad8-834e-6f92f7646dfc-c000.json
-rw-r--r--   1 hadoop hadoop        578 2022-06-18 08:34 Timebased-KPI/part-00097-dca0e8da-ee6b-4a16-9445-bdf192e9ffdf-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00099-86bbb88a-a05a-474b-8b18-b2709283160b-c000.json
-rw-r--r--   1 hadoop hadoop        181 2022-06-18 08:34 Timebased-KPI/part-00101-ecb6a3e8-7d38-4c7b-8345-5a10449c1a47-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00104-5dba8ef6-b1a5-4533-b950-caf1ab90f133-c000.json
-rw-r--r--   1 hadoop hadoop        198 2022-06-18 08:34 Timebased-KPI/part-00105-800219be-e2f6-4103-9848-f95c56f70314-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00107-98c63f3a-513e-4ebb-82ff-a8892738340f-c000.json
-rw-r--r--   1 hadoop hadoop        200 2022-06-18 08:34 Timebased-KPI/part-00109-718f6fbb-7682-4053-90e7-3f0c4a379950-c000.json
-rw-r--r--   1 hadoop hadoop        182 2022-06-18 08:34 Timebased-KPI/part-00111-93f3ebba-ab04-4ee2-a116-3003ae5c9d33-c000.json
```



```
-rw-r--r--   1 hadoop hadoop        379 2022-06-18 08:34 Timebased-KPI/part-00191-07183595-2dc1-4353-9f3c-9c19f595d9d9-c000.json
-rw-r--r--   1 hadoop hadoop        183 2022-06-18 08:34 Timebased-KPI/part-00193-da0b57a5-9347-41c6-8234-611f7685a070-c000.json
-rw-r--r--   1 hadoop hadoop        383 2022-06-18 08:34 Timebased-KPI/part-00195-3d8793bd-de59-4914-a366-5157fdcbdae4-c000.json
-rw-r--r--   1 hadoop hadoop        200 2022-06-18 08:34 Timebased-KPI/part-00196-abf4e5e9-8df3-4134-a4cc-e7bb0de3bd4c-c000.json
[hadoop@ip-172-31-85-38 ~]$ hadoop fs -ls Country-and-timebased-KPI
Found 116 items
drwxr-xr-x   - hadoop hadoop          0 2022-06-18 08:34 Country-and-timebased-KPI/_spark_metadata
-rw-r--r--   1 hadoop hadoop          0 2022-06-18 08:33 Country-and-timebased-KPI/part-00000-55f55453-d020-470b-bb89-c869a2796c46-c000.json
-rw-r--r--   1 hadoop hadoop        164 2022-06-18 08:34 Country-and-timebased-KPI/part-00000-846db905-8ef4-461f-8eeb-cf7039ce1e42-c000.json
-rw-r--r--   1 hadoop hadoop        508 2022-06-18 08:34 Country-and-timebased-KPI/part-00001-dba86b96-191e-45bf-b8e6-2c976246e2df-c000.json
-rw-r--r--   1 hadoop hadoop        175 2022-06-18 08:34 Country-and-timebased-KPI/part-00002-4303b0af-1261-444a-bb82-0589fb955506-c000.json
-rw-r--r--   1 hadoop hadoop        174 2022-06-18 08:34 Country-and-timebased-KPI/part-00003-df9358dd-b5a3-41e7-a0d5-600601ee4f36-c000.json
-rw-r--r--   1 hadoop hadoop        163 2022-06-18 08:34 Country-and-timebased-KPI/part-00004-cbb7d78f-45e6-4679-94c1-6cd661f358d1-c000.json
-rw-r--r--   1 hadoop hadoop        324 2022-06-18 08:34 Country-and-timebased-KPI/part-00005-6f03a54c-fd02-4909-97ef-90126734a2a1-c000.json
-rw-r--r--   1 hadoop hadoop        347 2022-06-18 08:34 Country-and-timebased-KPI/part-00006-a0213dd5-225a-4d47-bc35-d7605985ca15-c000.json
-rw-r--r--   1 hadoop hadoop        174 2022-06-18 08:34 Country-and-timebased-KPI/part-00009-c0cc77e8-3071-493a-9c19-f53e545c8ee5-c000.json
-rw-r--r--   1 hadoop hadoop        175 2022-06-18 08:34 Country-and-timebased-KPI/part-00015-7cab001f-132b-4071-b2fa-5510b7fe3f70-c000.json
-rw-r--r--   1 hadoop hadoop        339 2022-06-18 08:34 Country-and-timebased-KPI/part-00016-9a866711-65c9-4057-a4ca-f17792d7a415-c000.json
-rw-r--r--   1 hadoop hadoop        170 2022-06-18 08:34 Country-and-timebased-KPI/part-00019-ccdde86c-f05e-4854-a406-b64b05d25462-c000.json
-rw-r--r--   1 hadoop hadoop        190 2022-06-18 08:34 Country-and-timebased-KPI/part-00020-f649a5f6-955d-43c1-814e-08acd5c08af3-c000.json
-rw-r--r--   1 hadoop hadoop        329 2022-06-18 08:34 Country-and-timebased-KPI/part-00022-b0b46de3-b234-4969-9da1-619f00506892-c000.json
-rw-r--r--   1 hadoop hadoop        166 2022-06-18 08:34 Country-and-timebased-KPI/part-00026-ba9ff7e0-0b54-44e3-b3e8-55109dcc8e9c-c000.json
-rw-r--r--   1 hadoop hadoop        173 2022-06-18 08:34 Country-and-timebased-KPI/part-00028-8546fe05-a522-4554-81b7-0c92465c6297-c000.json
-rw-r--r--   1 hadoop hadoop        330 2022-06-18 08:34 Country-and-timebased-KPI/part-00029-d835390c-e3b7-41e7-98a7-9f4a1c32044e-c000.json
-rw-r--r--   1 hadoop hadoop        339 2022-06-18 08:34 Country-and-timebased-KPI/part-00030-663caac7-e288-4bd7-8d4b-dd3be27f5004-c000.json
-rw-r--r--   1 hadoop hadoop        168 2022-06-18 08:34 Country-and-timebased-KPI/part-00035-34efa5b3-80bf-44b1-9fa3-0f79b1e87db0-c000.json
-rw-r--r--   1 hadoop hadoop        171 2022-06-18 08:34 Country-and-timebased-KPI/part-00036-25498f1a-c69e-4807-86f1-f7d296dfcb7f-c000.json
-rw-r--r--   1 hadoop hadoop        172 2022-06-18 08:34 Country-and-timebased-KPI/part-00037-ee3bcb0f-8ab4-4a9e-ac11-3807bb9c020c-c000.json
-rw-r--r--   1 hadoop hadoop        331 2022-06-18 08:34 Country-and-timebased-KPI/part-00038-d9dce09a-6042-4867-ae05-462b7e40c81b-c000.json
-rw-r--r--   1 hadoop hadoop        166 2022-06-18 08:34 Country-and-timebased-KPI/part-00040-a2ca796c-634f-4d05-8aad-257942e280a7-c000.json
-rw-r--r--   1 hadoop hadoop        535 2022-06-18 08:34 Country-and-timebased-KPI/part-00041-71f4b1aa-d354-4c3f-8a64-974f31ef518f-c000.json
-rw-r--r--   1 hadoop hadoop        502 2022-06-18 08:34 Country-and-timebased-KPI/part-00043-c19f8bcc-fd96-4f2f-8b8d-9ffa41813584-c000.json
-rw-r--r--   1 hadoop hadoop        190 2022-06-18 08:34 Country-and-timebased-KPI/part-00044-51a1e2fc-28db-4c42-8917-93f33048ae1c-c000.json
-rw-r--r--   1 hadoop hadoop        174 2022-06-18 08:34 Country-and-timebased-KPI/part-00045-5f87831e-8b13-4517-8e65-7d1b17a86a4a-c000.json
-rw-r--r--   1 hadoop hadoop        173 2022-06-18 08:34 Country-and-timebased-KPI/part-00048-7b6e8267-e4ee-4852-a82a-6b7eaad5647c-c000.json
-rw-r--r--   1 hadoop hadoop        174 2022-06-18 08:34 Country-and-timebased-KPI/part-00050-23d61a20-c3d8-47b9-b440-83edc32fa823-c000.json
-rw-r--r--   1 hadoop hadoop        191 2022-06-18 08:34 Country-and-timebased-KPI/part-00052-8d5865e4-a1a9-46c2-9253-cc89c0985833-c000.json
-rw-r--r--   1 hadoop hadoop        357 2022-06-18 08:34 Country-and-timebased-KPI/part-00055-07f40330-c791-4422-8cda-291c0ff594eb-c000.json
-rw-r--r--   1 hadoop hadoop        189 2022-06-18 08:34 Country-and-timebased-KPI/part-00057-ebea17b4-b849-4dd7-8eda-359af8af1b31-c000.json
-rw-r--r--   1 hadoop hadoop        149 2022-06-18 08:34 Country-and-timebased-KPI/part-00060-212a1866-2d06-4f20-a2cf-29b4c52343f8-c000.json
-rw-r--r--   1 hadoop hadoop        173 2022-06-18 08:34 Country-and-timebased-KPI/part-00061-4adbd4af-ba3d-429e-9530-a28288ef6897-c000.json
-rw-r--r--   1 hadoop hadoop        348 2022-06-18 08:34 Country-and-timebased-KPI/part-00062-e3983ca2-b317-4565-881a-f38a49178004-c000.json
-rw-r--r--   1 hadoop hadoop        347 2022-06-18 08:34 Country-and-timebased-KPI/part-00065-ab093888-7693-4525-a2fc-ddff39aeaf57-c000.json
-rw-r--r--   1 hadoop hadoop        347 2022-06-18 08:34 Country-and-timebased-KPI/part-00068-79fdcf16-6ba1-4455-91b1-a0a8aebaf1f1-c000.json
```

15. Thereafter WinSCP is used to establish a connection between the EMR instance and local file system to transfer all the required files into local system