Code Logic - Retail Data Analysis

Setup:

1. Setting up spark environment before running the spark job

   export SPARK_KAFKA_VERSION=0.10

2. Run the spark job to execute spark-streaming.py script
   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py

   org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 - Spark Kafka Sql Jar version spark-streaming.py - Script to be executed

Code Logic:

spark-streaming.py script executes sequentially with following steps:

➔ Import the necessary modules

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.functions import *
```

➔ Initialising Spark Session

```
spark = SparkSession \
        .builder \
        .appName("Retail_Project") \
        .getOrCreate()
```

➔ Setting the log level to be ERROR

```
spark.sparkContext.setLogLevel('ERROR')
```

➔ Reading data from Kafka

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

Format - Kafka
Bootstrap Server - 18.211.252.152:9092

Subscribe to Topic - real-time-project
Starting Offsets – Earliest

➔ Construct schema in a format provided as a input:

Input

```
{
"invoice_no": 154132541653705,
"country": "United Kingdom",
"timestamp": "2020-09-18 10:55:23",
"type": "ORDER",
"items": [
        { "SKU": "21485",
        "title": "RETROSPOT HEART HOT WATER BOTTLE",
        "unit_price": 4.95,
        "quantity": 6
        },
        { "SKU": "23499",
        "title": "SET 12 VINTAGE DOILY CHALK",
        "unit_price": 0.42,
        "quantity": 2
        }
        ]
        }
```

Schema

```
schema = StructType() \
        .add("invoice_no", LongType()) \
        .add("country",StringType()) \
        .add("timestamp", TimestampType()) \
        .add("type", StringType()) \
        .add("items", ArrayType(StructType([
        StructField("SKU", StringType()),
        StructField("title", StringType()),
        StructField("unit_price", FloatType()),
        StructField("quantity", IntegerType())
        ])))
```

➔ Read the input data with respect to schema provided

```
order_stream = raw_data.select(from_json(col("value").cast("string"),
schema).alias("data")).select("data.*")
```

➔ Add the following utility functions

is_order - A user defined function to return whether it's type is ORDER or not

```
def is_order(type):
    if type=="ORDER":
        return 1
    else:
        return 0
```

is_return - A user defined function to return whether it's type is RETURN or not

```
def is_return(type):
    if type=="RETURN":
        return 1
    else:
        return 0
```

total_items_count - A user defined function to return the total count of items ordered.

```
def total_items_count(items):
    total_count = 0
    for item in items:
        total_count = total_count + item['quantity']
    return total_count
```

calculate_total_cost - A user defined function to return the total cost of items ordered.

```
def calculate_total_cost(items,type):
    total_price = 0
    for item in items:
        total_price = total_price + item['unit_price'] * item['quantity']
    if type=="RETURN":
        return total_price * -1
    else:
        return total_price
```

```
# Define the UDFs with the utility functions
flag_order = udf(is_order, IntegerType()) f
lag_return = udf(is_return, IntegerType())
calculate_item = udf(total_items_count, IntegerType())
calculate_cost = udf(calculate_total_cost, FloatType())
```

➔ Append additional columns to the order_stream with the user defined functions

```
order_stream_extended = order_stream \
.withColumn("total_items", calculate_item(order_stream.items)) \
.withColumn("total_cost",calculate_cost(order_stream.items,order_stream.type))\
.withColumn("is_order", flag_order(order_stream.type)) \
.withColumn("is_return", flag_return(order_stream.type))
```

➔ Select columns required to be logged as console. Every 1 minute the new data gets processed in a stream.

```
order_table_console = order_stream_extended \
        .select("invoice_no", "country",
"timestamp","type","total_items","total_cost","is_order","is_return") \
        .writeStream \
        .outputMode("append") \
        .format("console") \
        .option("truncate", "false") \
        .trigger(processingTime="1 minute") \
        .start()
```

➔ Calculate Time and Country based KPIs

```
# Calculate time based KPIs
agg_time = order_stream_extended \
        .withWatermark("timestamp","1 minute") \
        .groupby(window("timestamp", "1 minute",”1 minute”)) \
        .agg(sum("total_cost").alias("total_volume_of_sales"),
        avg("total_cost").alias("average_transaction_size"),
        avg("is_return").alias("rate_of_return")) \
        .select("window.start","window.end","total_volume_of_sales","average_transactio
        n_size", "rate_of_return")

# Calculate time and country based KPIs
agg_time_country = order_stream_extended \
        .withWatermark("timestamp", "1 minute") \
        .groupBy(window("timestamp", "1 minute",”1 minute”), "country") \
        .agg(sum("total_cost").alias("total_volume_of_sales"),
        count("invoice_no").alias("OPM"), avg("is_return").alias("rate_of_return")) \
        .select("window.start","window.end","country",
        "OPM","total_volume_of_sales","rate_of_return")
```

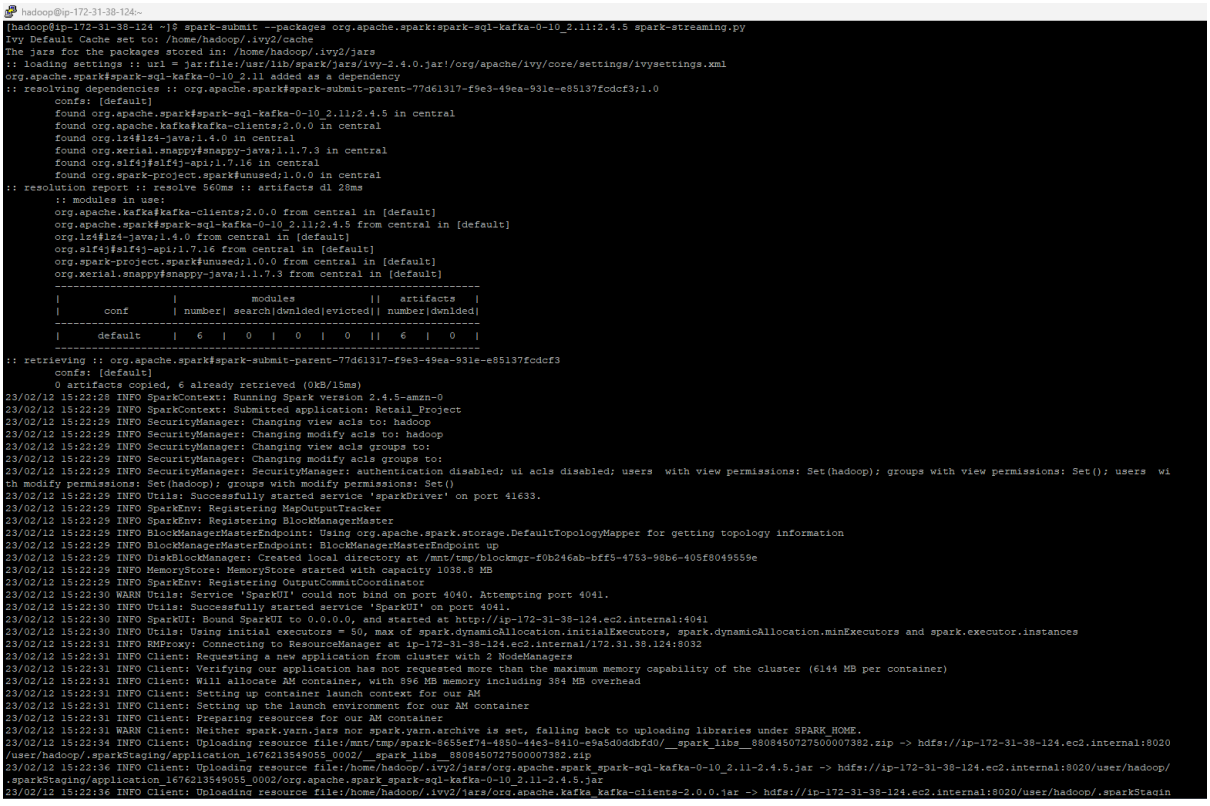
➔ Logging Time and Country based KPIs

```
        # Write time based KPI values
        console_by_time = agg_time.writeStream \
                .format("json") \
                .outputMode("append") \
                .option("truncate", "false") \
                .option("path", "time_based_kpi/") \
                .option("checkpointLocation", "time_based_kpi/cp/") \
                .trigger(processingTime="1 minute") \
```

```
                .start()

        # Write time and country based KPI values
            console_by_country = agg_time_country.writeStream \
                .format("json") \
                .outputMode("append") \
                .option("truncate", "false") \
                .option("path", "time_country_based_kpi/") \
                .option("checkpointLocation", "time_country_based_kpi/cp/") \
                .trigger(processingTime="1 minute") \
                .start()
```

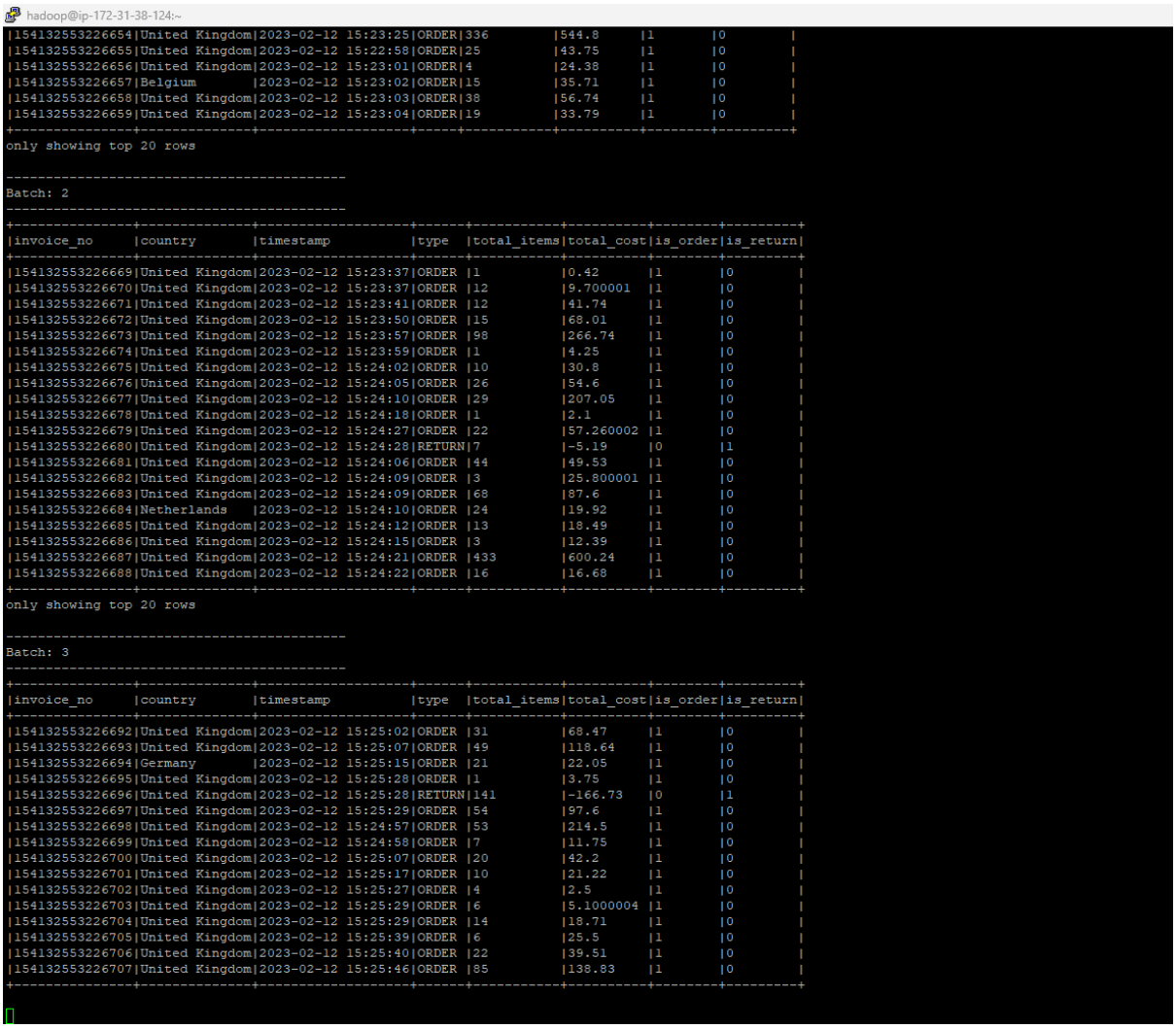
➔ Await Termination execution for writestream queries

```
order_table_console.awaitTermination()
    console_by_time.awaitTermination()
    console_by_country.awaitTermination()
```



```
hadoop@ip-172-31-38-124:~
23/02/12 15:22:36 INFO Client: Uploading resource file:/home/hadoop/.ivy2/jars/org.apache.kafka_kafka-clients-2.0.0.jar -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStagin
g/application_1676213549055_0002/org.apache.kafka_kafka-clients-2.0.0.jar
23/02/12 15:22:36 INFO Client: Uploading resource file:/home/hadoop/.ivy2/jars/org.spark-project.spark_unused-1.0.0.jar -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStagin
g/application_1676213549055_0002/org.spark-project.spark_unused-1.0.0.jar
23/02/12 15:22:36 INFO Client: Uploading resource file:/home/hadoop/.ivy2/jars/org.lz4_lz4-java-1.4.0.jar -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStaging/application_
1676213549055_0002/org.lz4_lz4-java-1.4.0.jar
23/02/12 15:22:37 INFO Client: Uploading resource file:/home/hadoop/.ivy2/jars/org.xerial.snappy_snappy-java-1.1.7.3.jar -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStagi
ng/application_1676213549055_0002/org.xerial.snappy_snappy-java-1.1.7.3.jar
23/02/12 15:22:37 INFO Client: Uploading resource file:/home/hadoop/.ivy2/jars/org.slf4j_slf4j-api-1.7.16.jar -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStaging/applicat
ion_1676213549055_0002/org.slf4j_slf4j-api-1.7.16.jar
23/02/12 15:22:37 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStaging/application_1676213549055
_0002/pyspark.zip
23/02/12 15:22:38 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.7-src.zip -> hdfs://ip-172-31-38-124.ec2.internal:8020/user/hadoop/.sparkStaging/application_16762
13549055_0002/py4j-0.10.7-src.zip
23/02/12 15:22:38 WARN Client: Same path resource file:///home/hadoop/.ivy2/jars/org.apache.spark_spark-sql-kafka-0-10_2.11-2.4.5.jar added multiple times to distributed cache.
23/02/12 15:22:38 WARN Client: Same path resource file:///home/hadoop/.ivy2/jars/org.apache.kafka_kafka-clients-2.0.0.jar added multiple times to distributed cache.
23/02/12 15:22:38 WARN Client: Same path resource file:///home/hadoop/.ivy2/jars/org.spark-project.spark_unused-1.0.0.jar added multiple times to distributed cache.
23/02/12 15:22:38 WARN Client: Same path resource file:///home/hadoop/.ivy2/jars/org.lz4_lz4-java-1.4.0.jar added multiple times to distributed cache.
23/02/12 15:22:38 WARN Client: Same path resource file:///home/hadoop/.ivy2/jars/org.xerial.snappy_snappy-java-1.1.7.3.jar added multiple times to distributed cache.
23/02/12 15:22:38 WARN Client: Same path resource file:///home/hadoop/.ivy2/jars/org.slf4j_slf4j-api-1.7.16.jar added multiple times to distributed cache.
23/02/12 15:22:38 INFO Client: Uploading resource file:/mnt/tmp/spark-8655ef74-4850-44e3-8410-e9a5d0ddbfd0/__spark_conf__3439102147057535606.zip -> hdfs://ip-172-31-38-124.ec2.internal:8020
/user/hadoop/.sparkStaging/application_1676213549055_0002/__spark_conf__.zip
23/02/12 15:22:39 INFO SecurityManager: Changing view acls to: hadoop
23/02/12 15:22:39 INFO SecurityManager: Changing modify acls to: hadoop
23/02/12 15:22:39 INFO SecurityManager: Changing view acls groups to:
23/02/12 15:22:39 INFO SecurityManager: Changing modify acls groups to:
23/02/12 15:22:39 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permissions: Set(); users  wi
th modify permissions: Set(hadoop); groups with modify permissions: Set()
23/02/12 15:22:40 INFO Client: Submitting application application_1676213549055_0002 to ResourceManager
23/02/12 15:22:40 INFO YarnClientImpl: Submitted application application_1676213549055_0002
23/02/12 15:22:41 INFO SchedulerExtensionServices: Starting Yarn extension services with app application_1676213549055_0002 and attemptId None
23/02/12 15:22:42 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:42 INFO Client:
         client token: N/A
         diagnostics: AM container is launched, waiting for AM container to Register with RM
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: default
         start time: 1676215360942
         final status: UNDEFINED
         tracking URL: http://ip-172-31-38-124.ec2.internal:20888/proxy/application_1676213549055_0002/
         user: hadoop
23/02/12 15:22:43 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:44 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:45 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:46 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:47 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:48 INFO Client: Application report for application_1676213549055_0002 (state: ACCEPTED)
23/02/12 15:22:48 INFO YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> ip-172-31-38-124.ec2.internal, PROXY_URI_
BASES -> http://ip-172-31-38-124.ec2.internal:20888/proxy/application_1676213549055_0002), /proxy/application_1676213549055_0002
23/02/12 15:22:49 INFO Client: Application report for application_1676213549055_0002 (state: RUNNING)
23/02/12 15:22:49 INFO Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 172.31.47.91
         ApplicationMaster RPC port: -1
         queue: default
         start time: 1676215360942
         final status: UNDEFINED
         tracking URL: http://ip-172-31-38-124.ec2.internal:20888/proxy/application_1676213549055_0002/
         user: hadoop
23/02/12 15:22:49 INFO YarnClientSchedulerBackend: Application application_1676213549055_0002 has started running.
23/02/12 15:22:49 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 42561.
```

```
hadoop@ip-172-31-38-124:~
23/02/12 15:22:49 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 42561.
23/02/12 15:22:49 INFO NettyBlockTransferService: Server created on ip-172-31-38-124.ec2.internal:42561
23/02/12 15:22:49 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
23/02/12 15:22:49 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
23/02/12 15:22:49 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-38-124.ec2.internal, 42561, None)
23/02/12 15:22:49 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-38-124.ec2.internal:42561 with 1038.8 MB RAM, BlockManagerId(driver, ip-172-31-38-124.ec2.internal, 42
561, None)
23/02/12 15:22:49 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-38-124.ec2.internal, 42561, None)
23/02/12 15:22:49 INFO BlockManager: external shuffle service port = 7337
23/02/12 15:22:49 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-38-124.ec2.internal, 42561, None)
23/02/12 15:22:49 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /metrics/json.
23/02/12 15:22:49 INFO EventLoggingListener: Logging events to hdfs:/var/log/spark/apps/application_1676213549055_0002
23/02/12 15:22:49 INFO Utils: Using initial executors = 50, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances
23/02/12 15:22:50 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
23/02/12 15:22:50 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/hive-site.xml
23/02/12 15:22:50 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('hdfs:///user/spark/warehouse').
23/02/12 15:22:50 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
23/02/12 15:22:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
23/02/12 15:22:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
23/02/12 15:22:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.
23/02/12 15:22:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
23/02/12 15:22:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
23/02/12 15:22:52 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+-------+---------+----+-----------+----------+--------+---------+
|invoice_no|country|timestamp|type|total_items|total_cost|is_order|is_return|
+----------+-------+---------+----+-----------+----------+--------+---------+
+----------+-------+---------+----+-----------+----------+--------+---------+

-------------------------------------------
Batch: 1
-------------------------------------------
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|invoice_no    |country       |timestamp          |type |total_items|total_cost|is_order|is_return|
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|1541325532266640|United Kingdom|2023-02-12 15:22:33|ORDER|68         |132.79    |1       |0        |
|1541325532266641|United Kingdom|2023-02-12 15:22:36|ORDER|4          |15.0      |1       |0        |
|1541325532266642|Germany       |2023-02-12 15:22:37|ORDER|4          |19.8      |1       |0        |
|1541325532266643|United Kingdom|2023-02-12 15:22:45|ORDER|34         |24.58     |1       |0        |
|1541325532266644|Portugal      |2023-02-12 15:22:49|ORDER|8          |26.859999 |1       |0        |
|1541325532266645|United Kingdom|2023-02-12 15:22:51|ORDER|3          |14.849999 |1       |0        |
|1541325532266646|United Kingdom|2023-02-12 15:22:59|ORDER|51         |108.45    |1       |0        |
|1541325532266647|United Kingdom|2023-02-12 15:23:01|ORDER|24         |39.649998 |1       |0        |
|1541325532266648|United Kingdom|2023-02-12 15:23:04|ORDER|193        |372.45    |1       |0        |
|1541325532266649|United Kingdom|2023-02-12 15:23:12|ORDER|37         |59.95     |1       |0        |
|1541325532266650|United Kingdom|2023-02-12 15:23:13|ORDER|30         |37.81     |1       |0        |
|1541325532266651|United Kingdom|2023-02-12 15:23:15|ORDER|10         |36.3      |1       |0        |
|1541325532266652|United Kingdom|2023-02-12 15:23:23|ORDER|31         |76.97     |1       |0        |
|1541325532266653|United Kingdom|2023-02-12 15:23:24|ORDER|2          |14.4      |1       |0        |
|1541325532266654|United Kingdom|2023-02-12 15:23:25|ORDER|336        |544.8     |1       |0        |
|1541325532266655|United Kingdom|2023-02-12 15:22:58|ORDER|25         |43.75     |1       |0        |
|1541325532266656|United Kingdom|2023-02-12 15:23:01|ORDER|4          |24.38     |1       |0        |
|1541325532266657|Belgium       |2023-02-12 15:23:02|ORDER|15         |35.71     |1       |0        |
|1541325532266658|United Kingdom|2023-02-12 15:23:03|ORDER|38         |56.74     |1       |0        |
|1541325532266659|United Kingdom|2023-02-12 15:23:04|ORDER|19         |33.79     |1       |0        |
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
only showing top 20 rows

-------------------------------------------
Batch: 2
-------------------------------------------
```



hadoop@ip-172-31-38-124:~

```
|154132553226706|United Kingdom|2023-02-12 15:25:40|ORDER |22         |39.51     |1       |0        |
|154132553226707|United Kingdom|2023-02-12 15:25:46|ORDER |85         |138.83    |1       |0        |
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+


-------------------------------------------
Batch: 4
-------------------------------------------
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|invoice_no     |country       |timestamp          |type  |total_items|total_cost|is_order|is_return|
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|154132553226708|United Kingdom|2023-02-12 15:25:29|ORDER |30         |89.44     |1       |0        |
|154132553226709|United Kingdom|2023-02-12 15:25:34|ORDER |12         |38.5      |1       |0        |
|154132553226710|United Kingdom|2023-02-12 15:25:39|ORDER |2          |1.7       |1       |0        |
|154132553226711|United Kingdom|2023-02-12 15:25:46|ORDER |2          |5.9       |1       |0        |
|154132553226712|Channel Islands|2023-02-12 15:25:50|RETURN|31         |-55.71    |0       |1        |
|154132553226713|United Kingdom|2023-02-12 15:25:52|ORDER |2          |1.48      |1       |0        |
|154132553226714|United Kingdom|2023-02-12 15:25:55|ORDER |40         |124.8     |1       |0        |
|154132553226715|United Kingdom|2023-02-12 15:26:04|ORDER |126        |118.98    |1       |0        |
|154132553226716|United Kingdom|2023-02-12 15:26:07|ORDER |200        |416.0     |1       |0        |
|154132553226717|United Kingdom|2023-02-12 15:26:09|RETURN|5          |-55.65    |0       |1        |
|154132553226718|United Kingdom|2023-02-12 15:26:10|ORDER |15         |20.85     |1       |0        |
|154132553226719|United Kingdom|2023-02-12 15:26:14|ORDER |57         |30.75     |1       |0        |
|154132553226720|United Kingdom|2023-02-12 15:26:18|ORDER |5          |24.75     |1       |0        |
|154132553226721|United Kingdom|2023-02-12 15:26:26|ORDER |42         |97.82     |1       |0        |
|154132553226722|United Kingdom|2023-02-12 15:26:27|ORDER |1          |0.19      |1       |0        |
|154132553226723|United Kingdom|2023-02-12 15:25:54|ORDER |3          |4.15      |1       |0        |
|154132553226724|Belgium       |2023-02-12 15:26:04|ORDER |10         |79.259995 |1       |0        |
|154132553226725|United Kingdom|2023-02-12 15:26:05|ORDER |4          |18.599998 |1       |0        |
|154132553226726|United Kingdom|2023-02-12 15:26:09|ORDER |86         |107.32    |1       |0        |
|154132553226727|United Kingdom|2023-02-12 15:26:13|ORDER |21         |46.879997 |1       |0        |
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
only showing top 20 rows

-------------------------------------------
Batch: 5
-------------------------------------------
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|invoice_no     |country       |timestamp          |type  |total_items|total_cost|is_order|is_return|
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|154132553226738|United Kingdom|2023-02-12 15:26:28|ORDER |1          |1.65      |1       |0        |
|154132553226739|United Kingdom|2023-02-12 15:26:29|ORDER |34         |66.46     |1       |0        |
|154132553226740|United Kingdom|2023-02-12 15:26:42|ORDER |24         |106.299995|1       |0        |
|154132553226741|United Kingdom|2023-02-12 15:26:51|ORDER |30         |28.29     |1       |0        |
|154132553226742|United Kingdom|2023-02-12 15:26:51|ORDER |12         |15.0      |1       |0        |
|154132553226743|United Kingdom|2023-02-12 15:26:52|ORDER |11         |41.379997 |1       |0        |
|154132553226744|France        |2023-02-12 15:27:06|ORDER |25         |42.87     |1       |0        |
|154132553226745|United Kingdom|2023-02-12 15:27:06|ORDER |34         |62.66     |1       |0        |
|154132553226746|United Kingdom|2023-02-12 15:27:09|ORDER |29         |100.2     |1       |0        |
|154132553226747|United Kingdom|2023-02-12 15:27:13|ORDER |11         |36.83     |1       |0        |
|154132553226748|United Kingdom|2023-02-12 15:27:21|RETURN|153        |-249.53   |0       |1        |
|154132553226749|United Kingdom|2023-02-12 15:26:59|ORDER |1          |2.1       |1       |0        |
|154132553226750|United Kingdom|2023-02-12 15:27:00|RETURN|57         |-134.24   |0       |1        |
|154132553226751|United Kingdom|2023-02-12 15:27:05|ORDER |31         |31.47     |1       |0        |
|154132553226752|United Kingdom|2023-02-12 15:27:07|ORDER |2          |16.94     |1       |0        |
|154132553226753|United Kingdom|2023-02-12 15:27:10|ORDER |6          |12.48     |1       |0        |
|154132553226754|United Kingdom|2023-02-12 15:27:13|ORDER |10         |29.5      |1       |0        |
|154132553226755|United Kingdom|2023-02-12 15:27:29|ORDER |2          |4.2       |1       |0        |
|154132553226756|United Kingdom|2023-02-12 15:27:29|ORDER |64         |62.34     |1       |0        |
|154132553226757|United Kingdom|2023-02-12 15:27:30|ORDER |3          |8.75      |1       |0        |
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
only showing top 20 rows
```

```
hadoop@ip-172-31-38-124:~
-------------------------------------------
Batch: 6
-------------------------------------------
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|invoice_no    |country       |timestamp          |type |total_items|total_cost|is_order|is_return|
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|154132553226761|United Kingdom|2023-02-12 15:27:31|ORDER|2          |9.9       |1       |0        |
|154132553226762|United Kingdom|2023-02-12 15:27:32|ORDER|34         |105.54    |1       |0        |
|154132553226763|United Kingdom|2023-02-12 15:27:41|ORDER|13         |51.95     |1       |0        |
|154132553226764|United Kingdom|2023-02-12 15:27:48|ORDER|103        |221.73999 |1       |0        |
|154132553226765|United Kingdom|2023-02-12 15:27:54|ORDER|19         |86.58     |1       |0        |
|154132553226766|United Kingdom|2023-02-12 15:28:05|ORDER|76         |154.70999 |1       |0        |
|154132553226767|United Kingdom|2023-02-12 15:28:09|ORDER|27         |44.5      |1       |0        |
|154132553226768|United Kingdom|2023-02-12 15:28:12|ORDER|3          |4.3500004 |1       |0        |
|154132553226769|United Kingdom|2023-02-12 15:28:24|ORDER|131        |126.49    |1       |0        |
|154132553226770|United Kingdom|2023-02-12 15:27:55|ORDER|38         |53.1      |1       |0        |
|154132553226771|United Kingdom|2023-02-12 15:27:57|ORDER|12         |45.0      |1       |0        |
|154132553226772|United Kingdom|2023-02-12 15:28:00|ORDER|9          |75.74     |1       |0        |
|154132553226773|United Kingdom|2023-02-12 15:28:05|ORDER|7          |22.21     |1       |0        |
|154132553226774|United Kingdom|2023-02-12 15:28:07|ORDER|12         |22.68     |1       |0        |
|154132553226775|United Kingdom|2023-02-12 15:28:10|ORDER|6          |9.1       |1       |0        |
|154132553226776|United Kingdom|2023-02-12 15:28:17|ORDER|1          |2.95      |1       |0        |
|154132553226777|United Kingdom|2023-02-12 15:28:18|ORDER|32         |36.78     |1       |0        |
|154132553226778|Iceland       |2023-02-12 15:28:23|ORDER|28         |36.16     |1       |0        |
|154132553226779|United Kingdom|2023-02-12 15:28:29|ORDER|8          |24.119999 |1       |0        |
|154132553226780|United Kingdom|2023-02-12 15:28:30|ORDER|9          |23.8      |1       |0        |
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
only showing top 20 rows

-------------------------------------------
Batch: 7
-------------------------------------------
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|invoice_no    |country       |timestamp          |type |total_items|total_cost|is_order|is_return|
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|154132553226784|United Kingdom|2023-02-12 15:28:26|ORDER|17         |58.07     |1       |0        |
|154132553226785|United Kingdom|2023-02-12 15:28:31|ORDER|12         |25.9      |1       |0        |
|154132553226786|United Kingdom|2023-02-12 15:28:38|ORDER|11         |43.71     |1       |0        |
|154132553226787|United Kingdom|2023-02-12 15:28:39|ORDER|15         |25.449999 |1       |0        |
|154132553226788|United Kingdom|2023-02-12 15:28:48|ORDER|29         |55.92     |1       |0        |
|154132553226789|United Kingdom|2023-02-12 15:28:53|ORDER|13         |21.1      |1       |0        |
|154132553226790|United Kingdom|2023-02-12 15:28:56|ORDER|12         |19.8      |1       |0        |
|154132553226791|United Kingdom|2023-02-12 15:28:57|ORDER|22         |52.23     |1       |0        |
|154132553226792|United Kingdom|2023-02-12 15:28:57|ORDER|6          |12.48     |1       |0        |
|154132553226793|United Kingdom|2023-02-12 15:28:58|ORDER|60         |75.0      |1       |0        |
|154132553226794|United Kingdom|2023-02-12 15:29:01|ORDER|2          |6.58      |1       |0        |
|154132553226795|United Kingdom|2023-02-12 15:29:01|ORDER|2          |6.58      |1       |0        |
|154132553226796|United Kingdom|2023-02-12 15:29:01|ORDER|14         |35.54     |1       |0        |
|154132553226797|United Kingdom|2023-02-12 15:29:07|ORDER|12         |24.48     |1       |0        |
|154132553226798|United Kingdom|2023-02-12 15:29:10|ORDER|15         |40.32     |1       |0        |
|154132553226799|United Kingdom|2023-02-12 15:29:24|ORDER|10         |22.74     |1       |0        |
|154132553226800|United Kingdom|2023-02-12 15:29:02|ORDER|4          |10.33     |1       |0        |
|154132553226801|United Kingdom|2023-02-12 15:29:14|ORDER|33         |72.229996 |1       |0        |
|154132553226802|United Kingdom|2023-02-12 15:29:19|ORDER|6          |10.700001 |1       |0        |
|154132553226803|United Kingdom|2023-02-12 15:29:22|ORDER|36         |23.4      |1       |0        |
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
only showing top 20 rows

-------------------------------------------
Batch: 8
-------------------------------------------
+--------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|invoice_no    |country       |timestamp          |type |total_items|total_cost|is_order|is_return|
```

I checked HDFS to make sure the KPI files were present.
hadoop fs -ls /user/ec2-user/

```
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -ls user/ec2-user/
Found 4 items
drwxr-xr-x   - hadoop hadoop          0 2023-02-12 15:39 user/ec2-user/country_kpi
drwxr-xr-x   - hadoop hadoop          0 2023-02-12 15:23 user/ec2-user/country_kpi_checkpoints
drwxr-xr-x   - hadoop hadoop          0 2023-02-12 15:39 user/ec2-user/time_kpi
drwxr-xr-x   - hadoop hadoop          0 2023-02-12 15:23 user/ec2-user/time_kpi_checkpoints
[hadoop@ip-172-31-38-124 ~]$
```

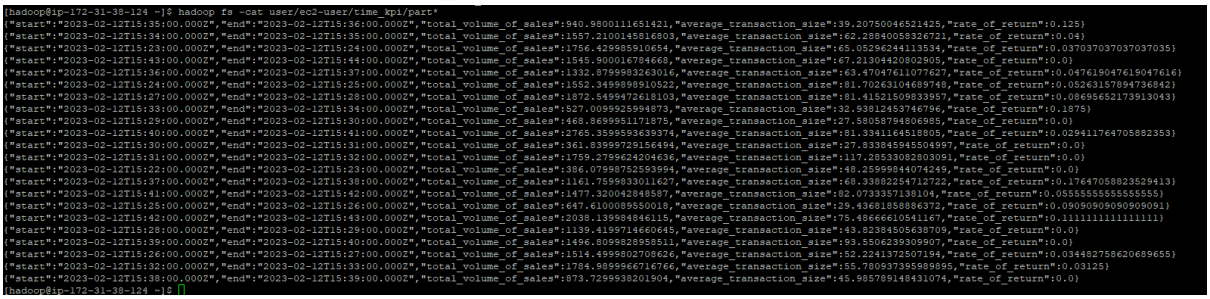
I also checked the folders to see the JSON files:
hadoop fs -ls /user/ec2-user/time_kpi

```
hadoop@ip-172-31-38-124:~
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R::::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R::::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R::::RRRRRR:::R
  E::::E             M:::::M    M:::M    M:::::M   R::::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R::::R      R::::R
EE:::::EEEEEEEE:::E M:::::M             M:::::M   R::::R      R::::R
E::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-38-124 ~]$
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -ls /user/ec2-user
ls: `/user/ec2-user': No such file or directory
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -ls user/ec2-user/time_kpi
Found 31 items
drwxr-xr-x   - hadoop hadoop          0 2023-02-12 15:38 user/ec2-user/time_kpi/_spark_metadata
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:23 user/ec2-user/time_kpi/part-00000-05a586be-9f7f-4a83-9c60-84244c9ace01-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:32 user/ec2-user/time_kpi/part-00000-47033002-cfad-4487-a58f-b21ceeeeca30-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:26 user/ec2-user/time_kpi/part-00000-4e641a15-0b1c-46df-a594-45ee4a61fcd9-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:29 user/ec2-user/time_kpi/part-00000-6e3f240f-7abb-4394-9874-c0806f45fab7-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:30 user/ec2-user/time_kpi/part-00000-80224fbe-b58e-4edb-adf0-5c9f5ca52995-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:36 user/ec2-user/time_kpi/part-00000-8589c6a0-bae9-4517-9b65-5803b4b30dde-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:34 user/ec2-user/time_kpi/part-00000-8dfc240b-6c2a-468d-a0d1-6708cea9b4cb-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:37 user/ec2-user/time_kpi/part-00000-a13d119e-97ed-4d5d-8606-9d2b77968799-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:38 user/ec2-user/time_kpi/part-00000-a4c12585-e385-484d-b9c8-19f5a77ec8b3-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:27 user/ec2-user/time_kpi/part-00000-ab228e89-76d2-4803-a032-3377d2deafae-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:35 user/ec2-user/time_kpi/part-00000-b141c47d-1fb2-4982-8bc4-be682c192d1e-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:23 user/ec2-user/time_kpi/part-00000-b586b962-a15d-4304-bb46-3ff2263fd259-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:31 user/ec2-user/time_kpi/part-00000-bb520eff-b665-427e-bb1b-cecf2afb1c9-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:28 user/ec2-user/time_kpi/part-00000-ce9ae5f9-619c-4ae2-ae46-97d43338d4f3-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:25 user/ec2-user/time_kpi/part-00000-cfb3657a-95ac-459d-9831-4476045cc6de-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:33 user/ec2-user/time_kpi/part-00000-d807ef62-d73d-4975-be1f-0d9b285f0960-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:24 user/ec2-user/time_kpi/part-00000-f478b3ab-3c76-4958-ae1c-260beba0b8c7-c000.json
-rw-r--r--   1 hadoop hadoop        180 2023-02-12 15:38 user/ec2-user/time_kpi/part-00035-d68124f4-00a3-4f69-baf9-21ac7fd08c3e-c000.json
-rw-r--r--   1 hadoop hadoop        195 2023-02-12 15:27 user/ec2-user/time_kpi/part-00040-590b39cf-47c8-4c2f-b9a8-ca3596054f60-c000.json
-rw-r--r--   1 hadoop hadoop        195 2023-02-12 15:28 user/ec2-user/time_kpi/part-00075-3199e47a-b22e-43f1-b9d7-7851f4da8310-c000.json
-rw-r--r--   1 hadoop hadoop        195 2023-02-12 15:31 user/ec2-user/time_kpi/part-00079-4f41f4e8-07df-488f-b530-b99c9c09a46e-c000.json
-rw-r--r--   1 hadoop hadoop        181 2023-02-12 15:37 user/ec2-user/time_kpi/part-00083-320445a9-3c3d-4400-92aa-6e4dd4797d52-c000.json
-rw-r--r--   1 hadoop hadoop        178 2023-02-12 15:33 user/ec2-user/time_kpi/part-00083-ec66501d-9cdd-41fd-93bc-2aad199aa662-c000.json
-rw-r--r--   1 hadoop hadoop        180 2023-02-12 15:34 user/ec2-user/time_kpi/part-00091-a6d94638-ec49-497b-89b4-6c1aee5b65ac-c000.json
-rw-r--r--   1 hadoop hadoop        180 2023-02-12 15:35 user/ec2-user/time_kpi/part-00095-c85d2d9d-419c-4fc4-9445-98fc96be4afb-c000.json
-rw-r--r--   1 hadoop hadoop        179 2023-02-12 15:26 user/ec2-user/time_kpi/part-00097-c35f5cac-907a-4b73-845b-173d3132b80f-c000.json
-rw-r--r--   1 hadoop hadoop        194 2023-02-12 15:29 user/ec2-user/time_kpi/part-00128-5a92f5c2-b96f-445a-9796-804abd6effc4-c000.json
-rw-r--r--   1 hadoop hadoop        179 2023-02-12 15:32 user/ec2-user/time_kpi/part-00137-64dc6288-8ff9-4b87-9049-9133cb355ef5-c000.json
-rw-r--r--   1 hadoop hadoop        195 2023-02-12 15:30 user/ec2-user/time_kpi/part-00154-dc47be3b-ddd8-4080-9cd9-27dd95e2cf89-c000.json
-rw-r--r--   1 hadoop hadoop        184 2023-02-12 15:36 user/ec2-user/time_kpi/part-00172-643610d2-c337-4faa-8f58-91a2a55717dd-c000.json
```

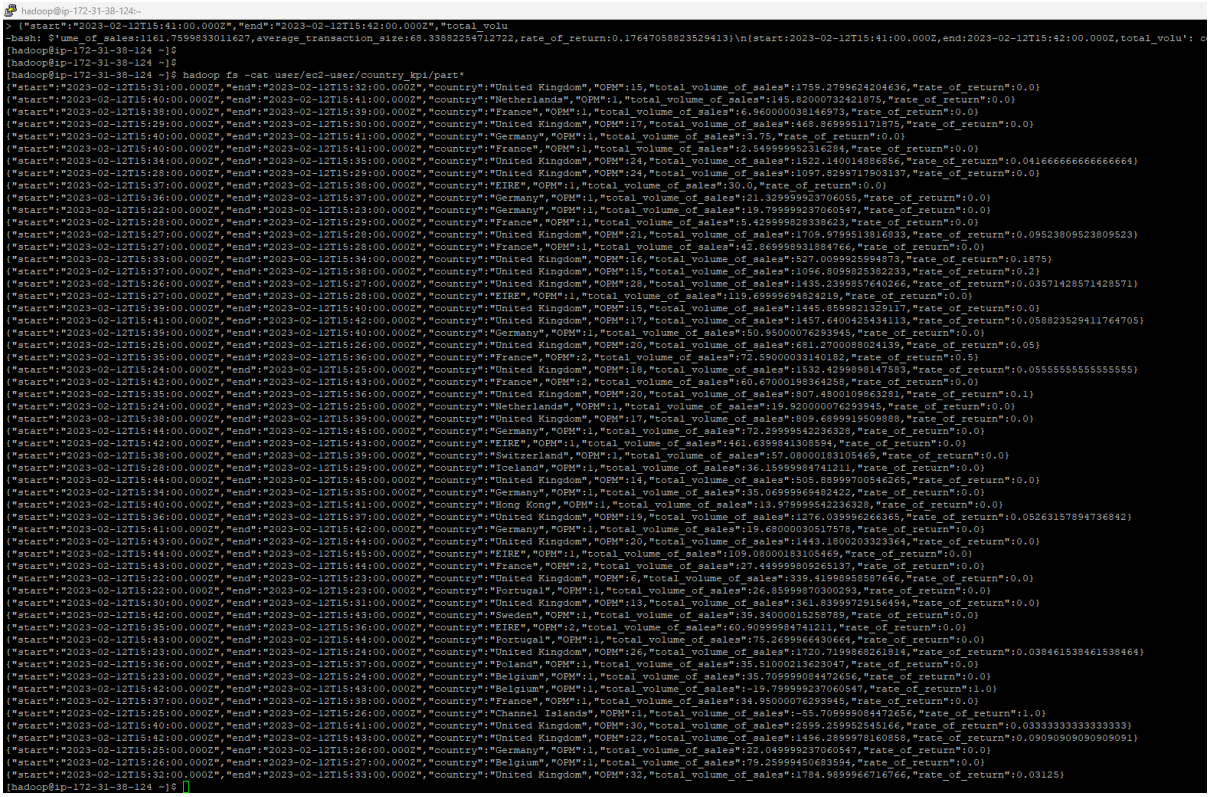
hadoop fs -ls user/ec2-user/country_kpi

```
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -ls user/ec2-user/country_kpi/
Found 77 items
drwxr-xr-x   - hadoop hadoop          0 2023-02-12 15:46 user/ec2-user/country_kpi/_spark_metadata
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:41 user/ec2-user/country_kpi/part-00000-18a0fa1b-8053-4b37-8e35-2409fa2898be-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:37 user/ec2-user/country_kpi/part-00000-3de021e1-67b6-4d94-bed3-417fcc05bb36-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:28 user/ec2-user/country_kpi/part-00000-3e8d4fad-fbd5-438b-b43f-636d74c379b3-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:39 user/ec2-user/country_kpi/part-00000-44015789-1d8f-4eaf-8287-cb5b3145ba02-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:34 user/ec2-user/country_kpi/part-00000-4c63f62d-c7bd-40e9-ab6b-b8ebb2d2e609-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:36 user/ec2-user/country_kpi/part-00000-59da5a7d-f3c3-4401-b2f3-3284987afafb-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:32 user/ec2-user/country_kpi/part-00000-5a24da11-b57e-4f0b-91d0-a54139725f11-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:26 user/ec2-user/country_kpi/part-00000-5bf069c9-3a99-4b24-b2f0-96c6ae990c4a-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:30 user/ec2-user/country_kpi/part-00000-65e3f390-aace-4fa8-9a5f-75b6aa169b01-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:31 user/ec2-user/country_kpi/part-00000-69d9aaf0-10e7-4823-b6e9-911982670fcb-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:24 user/ec2-user/country_kpi/part-00000-6ef620a8-1aaf-4d99-bbe4-a15a4f1c2c43-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:29 user/ec2-user/country_kpi/part-00000-77323b57-40fd-4886-8dcd-1e8c0044a2ca-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:33 user/ec2-user/country_kpi/part-00000-89d9adef-8a15-4e18-afa5-f71a255d2cad-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:38 user/ec2-user/country_kpi/part-00000-8f874f85-3b07-45ac-b518-96cb8de8e28b-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:25 user/ec2-user/country_kpi/part-00000-96574c5f-5973-44e5-b8f2-391540f5f73e-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:45 user/ec2-user/country_kpi/part-00000-9b894905-ff1d-46ff-89dd-8bf59585eebf-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:27 user/ec2-user/country_kpi/part-00000-9d5ee69e-d228-4cbf-872a-fcaa71741d79-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:23 user/ec2-user/country_kpi/part-00000-9f1ee3c9-c2d1-461b-907c-d2aae061fdec-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:46 user/ec2-user/country_kpi/part-00000-a6879129-f322-473c-b58c-b4b03f9c3633-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:44 user/ec2-user/country_kpi/part-00000-a79870c1-aacd-4f45-ac7b-c9dbe2f5d37c-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:40 user/ec2-user/country_kpi/part-00000-bf60e3e7-676c-4e79-aa46-b08fa0061c12-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:23 user/ec2-user/country_kpi/part-00000-c917cb8a-f3ac-4360-b270-a0fd9080d6da-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:42 user/ec2-user/country_kpi/part-00000-cd7beadl-dd55-4c45-84b6-d57cf0ae9559-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:43 user/ec2-user/country_kpi/part-00000-ddf9c09b-8432-4f9d-bede-ee09e8032db4-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-02-12 15:35 user/ec2-user/country_kpi/part-00000-ec5035e6-9bbc-4020-8787-e7bb24e7d3d5-c000.json
-rw-r--r--   1 hadoop hadoop        170 2023-02-12 15:35 user/ec2-user/country_kpi/part-00003-a38c394f-8bdb-4060-a163-e0aad322f9f3-c000.json
-rw-r--r--   1 hadoop hadoop        166 2023-02-12 15:44 user/ec2-user/country_kpi/part-00006-cb314002-6227-434b-b649-3c080737086f-c000.json
-rw-r--r--   1 hadoop hadoop        160 2023-02-12 15:42 user/ec2-user/country_kpi/part-00011-f1a4c6fe-d346-4f4b-902e-1287c90d00df-c000.json
-rw-r--r--   1 hadoop hadoop        169 2023-02-12 15:33 user/ec2-user/country_kpi/part-00013-5338c89b-c6c3-4275-9ec3-6875e5429fc9-c000.json
-rw-r--r--   1 hadoop hadoop        148 2023-02-12 15:44 user/ec2-user/country_kpi/part-00014-6ad17eac-0f7c-4f91-82a4-78c205cdf437-c000.json
-rw-r--r--   1 hadoop hadoop        160 2023-02-12 15:44 user/ec2-user/country_kpi/part-00016-d7c59cb9-f150-40f3-b187-c7a6eccb1ca7-c000.json
-rw-r--r--   1 hadoop hadoop        186 2023-02-12 15:38 user/ec2-user/country_kpi/part-00017-87636ad7-0602-42d2-aea5-e6d842be9c78-c000.json
-rw-r--r--   1 hadoop hadoop        170 2023-02-12 15:32 user/ec2-user/country_kpi/part-00021-1375fcad-1523-4084-905e-ade57cc99872-c000.json
-rw-r--r--   1 hadoop hadoop        145 2023-02-12 15:41 user/ec2-user/country_kpi/part-00021-78127fe8-a7f7-4fdc-8bba-27184fb5cdbe-c000.json
-rw-r--r--   1 hadoop hadoop        162 2023-02-12 15:40 user/ec2-user/country_kpi/part-00024-dc63d6ec-b3ed-42a5-8b3e-36c37f978093-c000.json
-rw-r--r--   1 hadoop hadoop        162 2023-02-12 15:26 user/ec2-user/country_kpi/part-00028-645653c1-6b39-4f9a-9c46-3b4eec4c9150-c000.json
-rw-r--r--   1 hadoop hadoop        160 2023-02-12 15:32 user/ec2-user/country_kpi/part-00036-65a3e74b-5456-4873-9bd9-4c4216b876cf-c000.json
-rw-r--r--   1 hadoop hadoop        186 2023-02-12 15:31 user/ec2-user/country_kpi/part-00044-577cad0e-8ba2-400b-9ec0-309c8d0f59cd-c000.json
-rw-r--r--   1 hadoop hadoop        161 2023-02-12 15:31 user/ec2-user/country_kpi/part-00045-74a59b8f-2bbc-419a-9ab7-2592210a362f-c000.json
-rw-r--r--   1 hadoop hadoop        172 2023-02-12 15:37 user/ec2-user/country_kpi/part-00046-c261c26a-2b68-49d0-ad87-ba90d5a177f5-c000.json
-rw-r--r--   1 hadoop hadoop        170 2023-02-12 15:41 user/ec2-user/country_kpi/part-00052-cc47d543-27d7-41f6-8be8-5b80438e663d-c000.json
-rw-r--r--   1 hadoop hadoop        186 2023-02-12 15:30 user/ec2-user/country_kpi/part-00054-476c9cdb-749b-4c56-bee6-017310218ead-c000.json
-rw-r--r--   1 hadoop hadoop        159 2023-02-12 15:31 user/ec2-user/country_kpi/part-00056-42db9452-5b8b-4b2f-aa99-323b3ea7a480-c000.json
-rw-r--r--   1 hadoop hadoop        170 2023-02-12 15:43 user/ec2-user/country_kpi/part-00064-e94c3737-f6f2-4992-82b1-7bf7d594afe0-c000.json
-rw-r--r--   1 hadoop hadoop        187 2023-02-12 15:45 user/ec2-user/country_kpi/part-00069-831597e9-73fd-463f-bc6a-9cac76f681ae-c000.json
-rw-r--r--   1 hadoop hadoop        161 2023-02-12 15:43 user/ec2-user/country_kpi/part-00071-ebf91ad4-4879-4c56-8154-e8d4df5206f3-c000.json
```

And used ‘cat’ command to take a look at the data:
hadoop fs -cat user/ec2-user/time_kpi/part*



hadoop fs -cat user/ec2-user/country_kpi/part*



Transfer of files from EMR Instance on AWS to my system, using WinSCP :
First, I needed to transfer the JSON files from HDFS into the the EC2 system I created directories for time-based and then time-and-country-based KPIs as ec2-user.
Using the ‘get’ command I copied the contents of the output folders into the EC2 system.
mkdir timebased-KPI
hadoop fs -get user/ec2-user/time_kpi /home/hadoop/timebased-KPI

```
[hadoop@ip-172-31-38-124 ~]$ mkdir timebased-KPI
[hadoop@ip-172-31-38-124 ~]$ ls
read_from_kafka.py  spark-streaming.py  timebased-KPI
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -get user/ec2-user/time_kpi /home/ec2-user/timebased-KPI
get: mkdir `/home/ec2-user/timebased-KPI': Input/output error
[hadoop@ip-172-31-38-124 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -get user/ec2-user/time_kpi /home/hadoop/timebased-KPI
[hadoop@ip-172-31-38-124 ~]$
```

mkdir country-and-timebased-KPI

hadoop fs -get user/ec2-user/country_kpi /home/hadoop/country-and-timebased-KPI

```
[hadoop@ip-172-31-38-124 ~]$ mkdir country-and-timebased-KPI
[hadoop@ip-172-31-38-124 ~]$ hadoop fs -get user/ec2-user/country_kpi /home/hadoop/country-and-timebased-KPI
[hadoop@ip-172-31-38-124 ~]$
```

Thereafter I used WinSCP to establish a connection between the EC2 instance and my local file system to transfer all the required files into my system.