Code Logic - Retail Data Analysis

Assignment Steps

- Download the spark-sql-kafka JAR file to ensure that the spark streaming script runs as expected.
  - wget https://ds-spark-sql-kafka-jar.s3.amazonaws.com/spark-sql-kafka-0-10_2.11-2.3.0.jar
- Set the following environment variables:
  - export HADOOP_USER_NAME=hdfs
  - export SPARK_KAFKA_VERSION=0.10
- Run the spark2-submit command to run the Python script:
  - spark2-submit --jars spark-sql-kafka-0-10_2.11-2.3.0.jar spark-streaming.py
- Check for the output files in HDFS.

Code Walkthrough

1. Import all required libraries and set the environment variables.
2. Define functions to calculate new fields and convert them to UDFs.
3. Create a new Spark Session.
4. Read the data from Kafka Server.
5. Create a new schema for the JSON data read from Kafka server.
6. Add the calculated fields to the schema.
7. Calculate time-based and time & country based KPIs.
8. Write the data read by the stream to console.
9. Write the KPIs into separate JSON files for time-based and time & country based KPIs.

Calculations made

- “is_order” - The “order” field converted into binary field.
- “is_return” - The “return” field converted into binary field.
- “total_sale_volume” – Total transaction value of all orders in a time frame.
- “total_returns” – Total number of returned orders in a time frame.
- “total_orders” – Total number of orders placed in a time frame.
- “total_items” – Total number of items in an invoice
- “total_cost” – Total cost of the order aggregated by invoice number.

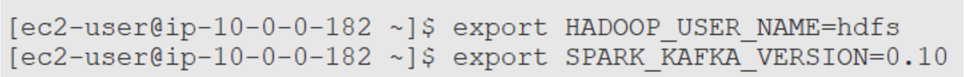
- “rate_of_return” – Percentage of returns compared to total number of invoices.
- “average_transaction_size” – Amount spent on average on each order.
- “OPM” – Total number of orders received in a minute.

KPIs

1. Total volume of sales – “total_sale_volume”
2. Orders Per Minute – “OPM”
3. Rate of Return – “rate_of_return”
4. Average Transaction Size – “average_transaction_size”

Screenshots of commands and results

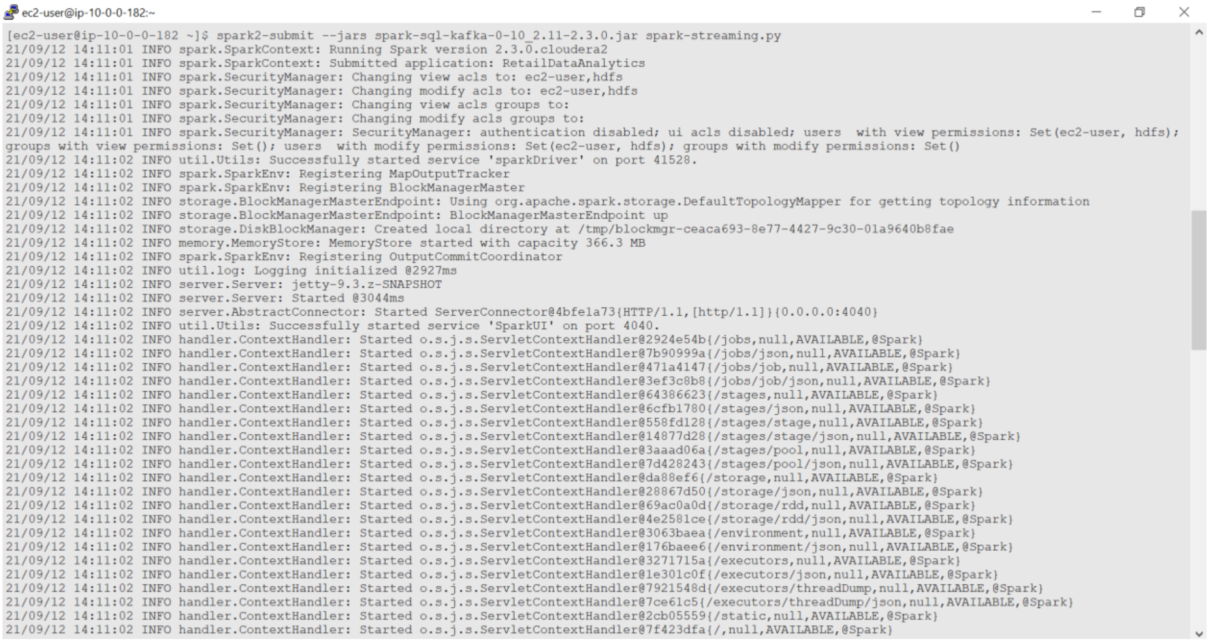
Set variables



```
[ec2-user@ip-10-0-0-182 ~]$ export HADOOP_USER_NAME=hdfs
[ec2-user@ip-10-0-0-182 ~]$ export SPARK_KAFKA_VERSION=0.10
```

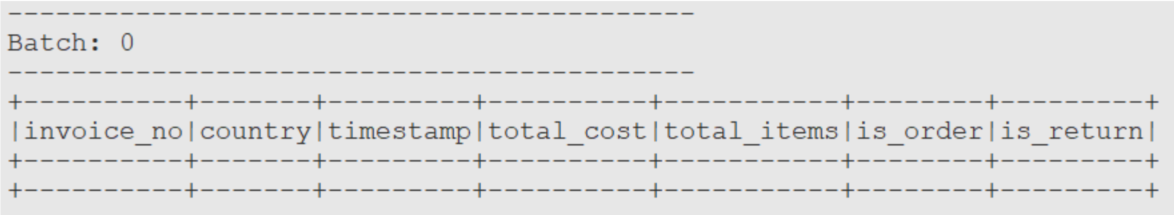
Run the Python script

spark2-submit --jars spark-sql-kafka-0-10_2.11-2.3.0.jar spark-streaming.py

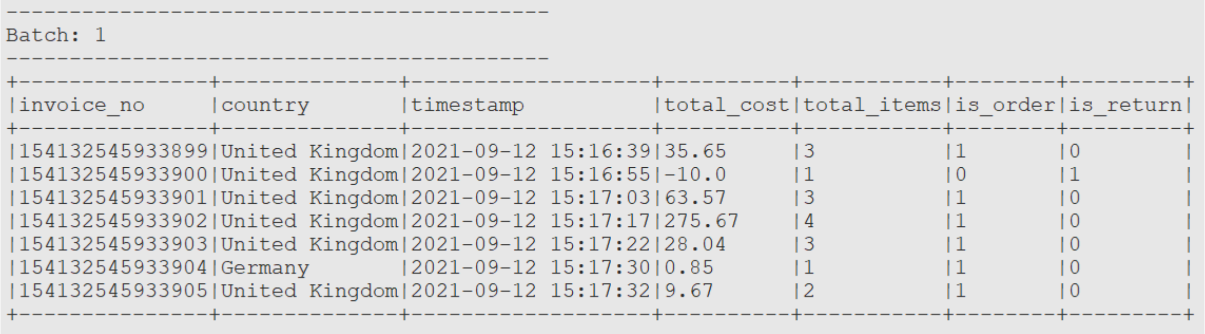


```
[ec2-user@ip-10-0-0-182 ~]$ spark2-submit --jars spark-sql-kafka-0-10_2.11-2.3.0.jar spark-streaming.py
21/09/12 14:11:01 INFO spark.SparkContext: Running Spark version 2.3.0.cloudera2
21/09/12 14:11:01 INFO spark.SparkContext: Submitted application: RetailDataAnalytics
21/09/12 14:11:01 INFO spark.SecurityManager: Changing view acls to: ec2-user,hdfs
21/09/12 14:11:01 INFO spark.SecurityManager: Changing modify acls to: ec2-user,hdfs
21/09/12 14:11:01 INFO spark.SecurityManager: Changing view acls groups to:
21/09/12 14:11:01 INFO spark.SecurityManager: Changing modify acls groups to:
21/09/12 14:11:01 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ec2-user, hdfs);
groups with view permissions: Set(); users  with modify permissions: Set(ec2-user, hdfs); groups with modify permissions: Set()
21/09/12 14:11:02 INFO util.Utils: Successfully started service 'sparkDriver' on port 41528.
21/09/12 14:11:02 INFO spark.SparkEnv: Registering MapOutputTracker
21/09/12 14:11:02 INFO spark.SparkEnv: Registering BlockManagerMaster
21/09/12 14:11:02 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
21/09/12 14:11:02 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
21/09/12 14:11:02 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-ceaca693-8e77-4427-9c30-01a9640b8fae
21/09/12 14:11:02 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB
21/09/12 14:11:02 INFO spark.SparkEnv: Registering OutputCommitCoordinator
21/09/12 14:11:02 INFO util.log: Logging initialized @2927ms
21/09/12 14:11:02 INFO server.Server: jetty-9.3.z-SNAPSHOT
21/09/12 14:11:02 INFO server.Server: Started @3044ms
21/09/12 14:11:02 INFO server.AbstractConnector: Started ServerConnector@4bfe1a73{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
21/09/12 14:11:02 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2924e54b{/jobs,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7b90999a{/jobs/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@471a4147{/jobs/job,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3ef3c8b8{/jobs/job/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@64386623{/stages,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6cfb1780{/stages/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@558fd128{/stages/stage,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@14877d28{/stages/stage/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3aaad06a{/stages/pool,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7d428243{/stages/pool/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@da88ef6{/storage,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@28867d50{/storage/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@69ac0a0d{/storage/rdd,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4e2581ce{/storage/rdd/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3063baea{/environment,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@176baee6{/environment/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3271715a{/executors,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1e301c0f{/executors/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7921548d{/executors/threadDump,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7ce61c5{/executors/threadDump/json,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2cb05559{/static,null,AVAILABLE,@Spark}
21/09/12 14:11:02 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7f423dfa{/,null,AVAILABLE,@Spark}
```

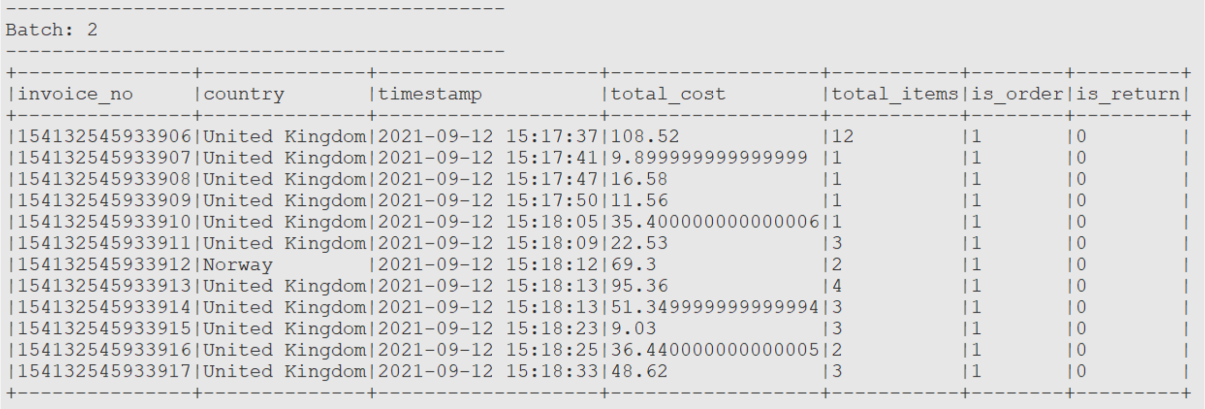
Console output



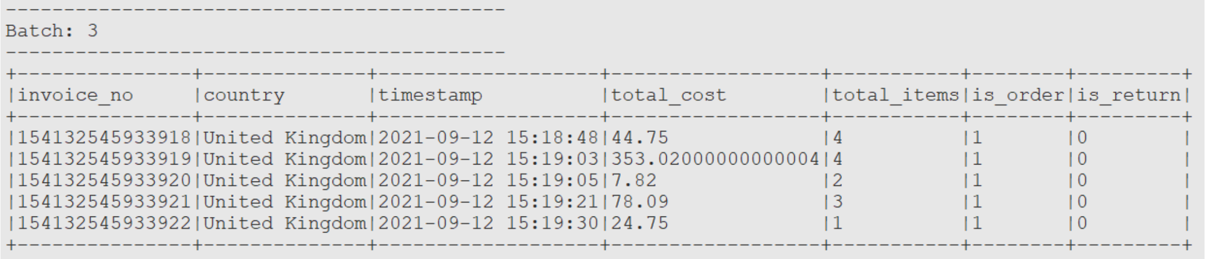
```
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+-------+---------+----------+-----------+--------+---------+
|invoice_no|country|timestamp|total_cost|total_items|is_order|is_return|
+----------+-------+---------+----------+-----------+--------+---------+
+----------+-------+---------+----------+-----------+--------+---------+
```



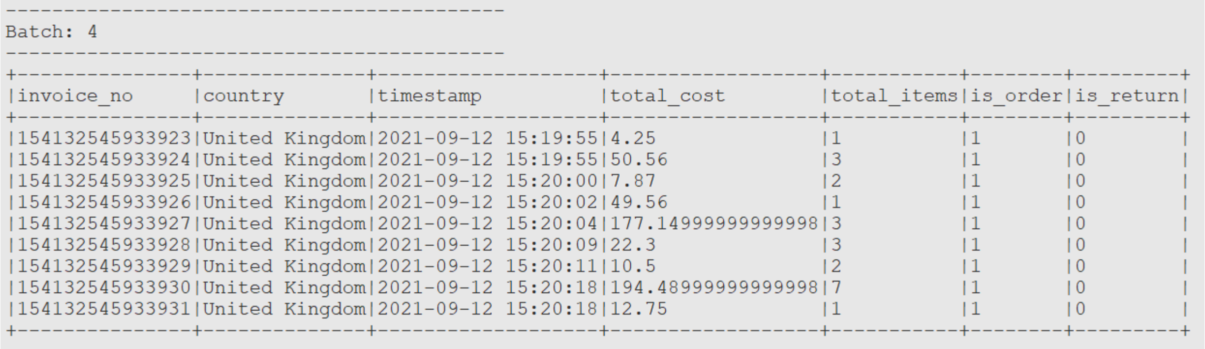
```
-------------------------------------------
Batch: 1
-------------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_return|
+---------------+--------------+-------------------+----------+-----------+--------+---------+
|154132545933899|United Kingdom|2021-09-12 15:16:39|35.65     |3          |1       |0        |
|154132545933900|United Kingdom|2021-09-12 15:16:55|-10.0     |1          |0       |1        |
|154132545933901|United Kingdom|2021-09-12 15:17:03|63.57     |3          |1       |0        |
|154132545933902|United Kingdom|2021-09-12 15:17:17|275.67    |4          |1       |0        |
|154132545933903|United Kingdom|2021-09-12 15:17:22|28.04     |3          |1       |0        |
|154132545933904|Germany       |2021-09-12 15:17:30|0.85      |1          |1       |0        |
|154132545933905|United Kingdom|2021-09-12 15:17:32|9.67      |2          |1       |0        |
+---------------+--------------+-------------------+----------+-----------+--------+---------+
```



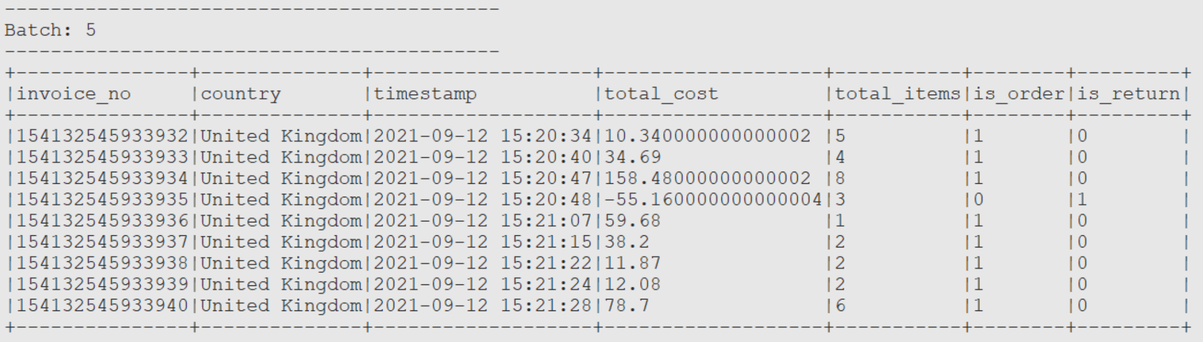
```
-------------------------------------------
Batch: 2
-------------------------------------------
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost        |total_items|is_order|is_return|
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|154132545933906|United Kingdom|2021-09-12 15:17:37|108.52            |12         |1       |0        |
|154132545933907|United Kingdom|2021-09-12 15:17:41|9.899999999999999 |1          |1       |0        |
|154132545933908|United Kingdom|2021-09-12 15:17:47|16.58             |1          |1       |0        |
|154132545933909|United Kingdom|2021-09-12 15:17:50|11.56             |1          |1       |0        |
|154132545933910|United Kingdom|2021-09-12 15:18:05|35.400000000000006|1          |1       |0        |
|154132545933911|United Kingdom|2021-09-12 15:18:09|22.53             |3          |1       |0        |
|154132545933912|Norway        |2021-09-12 15:18:12|69.3              |2          |1       |0        |
|154132545933913|United Kingdom|2021-09-12 15:18:13|95.36             |4          |1       |0        |
|154132545933914|United Kingdom|2021-09-12 15:18:13|51.349999999999994|3          |1       |0        |
|154132545933915|United Kingdom|2021-09-12 15:18:23|9.03              |3          |1       |0        |
|154132545933916|United Kingdom|2021-09-12 15:18:25|36.440000000000005|2          |1       |0        |
|154132545933917|United Kingdom|2021-09-12 15:18:33|48.62             |3          |1       |0        |
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
```



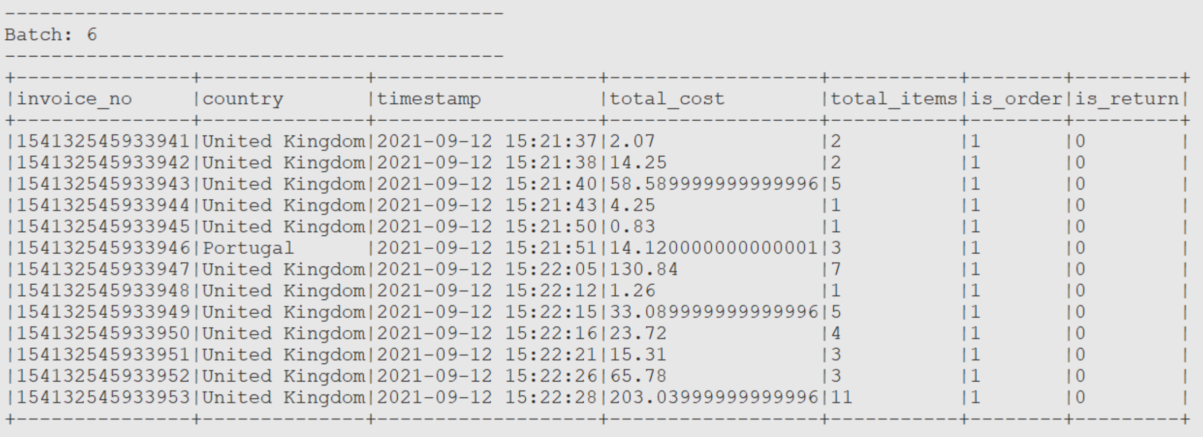
```
-------------------------------------------
Batch: 3
-------------------------------------------
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost        |total_items|is_order|is_return|
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|154132545933918|United Kingdom|2021-09-12 15:18:48|44.75             |4          |1       |0        |
|154132545933919|United Kingdom|2021-09-12 15:19:03|353.02000000000004|4          |1       |0        |
|154132545933920|United Kingdom|2021-09-12 15:19:05|7.82              |2          |1       |0        |
|154132545933921|United Kingdom|2021-09-12 15:19:21|78.09             |3          |1       |0        |
|154132545933922|United Kingdom|2021-09-12 15:19:30|24.75             |1          |1       |0        |
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
```



```
-------------------------------------------
Batch: 4
-------------------------------------------
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost        |total_items|is_order|is_return|
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|154132545933923|United Kingdom|2021-09-12 15:19:55|4.25              |1          |1       |0        |
|154132545933924|United Kingdom|2021-09-12 15:19:55|50.56             |3          |1       |0        |
|154132545933925|United Kingdom|2021-09-12 15:20:00|7.87              |2          |1       |0        |
|154132545933926|United Kingdom|2021-09-12 15:20:02|49.56             |1          |1       |0        |
|154132545933927|United Kingdom|2021-09-12 15:20:04|177.14999999999998|3          |1       |0        |
|154132545933928|United Kingdom|2021-09-12 15:20:09|22.3              |3          |1       |0        |
|154132545933929|United Kingdom|2021-09-12 15:20:11|10.5              |2          |1       |0        |
|154132545933930|United Kingdom|2021-09-12 15:20:18|194.48999999999998|7          |1       |0        |
|154132545933931|United Kingdom|2021-09-12 15:20:18|12.75             |1          |1       |0        |
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
```



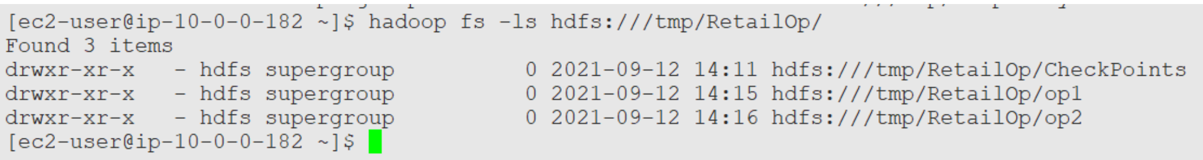
```
-------------------------------------------
Batch: 5
-------------------------------------------
+---------------+--------------+-------------------+-------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost         |total_items|is_order|is_return|
+---------------+--------------+-------------------+-------------------+-----------+--------+---------+
|154132545933932|United Kingdom|2021-09-12 15:20:34|10.340000000000002 |5          |1       |0        |
|154132545933933|United Kingdom|2021-09-12 15:20:40|34.69              |4          |1       |0        |
|154132545933934|United Kingdom|2021-09-12 15:20:47|158.48000000000002 |8          |1       |0        |
|154132545933935|United Kingdom|2021-09-12 15:20:48|-55.160000000000004|3          |0       |1        |
|154132545933936|United Kingdom|2021-09-12 15:21:07|59.68              |1          |1       |0        |
|154132545933937|United Kingdom|2021-09-12 15:21:15|38.2               |2          |1       |0        |
|154132545933938|United Kingdom|2021-09-12 15:21:22|11.87              |2          |1       |0        |
|154132545933939|United Kingdom|2021-09-12 15:21:24|12.08              |2          |1       |0        |
|154132545933940|United Kingdom|2021-09-12 15:21:28|78.7               |6          |1       |0        |
+---------------+--------------+-------------------+-------------------+-----------+--------+---------+
```



```
-------------------------------------------
Batch: 6
-------------------------------------------
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost        |total_items|is_order|is_return|
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
|154132545933941|United Kingdom|2021-09-12 15:21:37|2.07              |2          |1       |0        |
|154132545933942|United Kingdom|2021-09-12 15:21:38|14.25             |2          |1       |0        |
|154132545933943|United Kingdom|2021-09-12 15:21:40|58.589999999999996|5          |1       |0        |
|154132545933944|United Kingdom|2021-09-12 15:21:43|4.25              |1          |1       |0        |
|154132545933945|United Kingdom|2021-09-12 15:21:50|0.83              |1          |1       |0        |
|154132545933946|Portugal      |2021-09-12 15:21:51|14.120000000000001|3          |1       |0        |
|154132545933947|United Kingdom|2021-09-12 15:22:05|130.84            |7          |1       |0        |
|154132545933948|United Kingdom|2021-09-12 15:22:12|1.26              |1          |1       |0        |
|154132545933949|United Kingdom|2021-09-12 15:22:15|33.089999999999996|5          |1       |0        |
|154132545933950|United Kingdom|2021-09-12 15:22:16|23.72             |4          |1       |0        |
|154132545933951|United Kingdom|2021-09-12 15:22:21|15.31             |3          |1       |0        |
|154132545933952|United Kingdom|2021-09-12 15:22:26|65.78             |3          |1       |0        |
|154132545933953|United Kingdom|2021-09-12 15:22:28|203.03999999999996|11         |1       |0        |
+---------------+--------------+-------------------+------------------+-----------+--------+---------+
```

Results written to HDFS directory

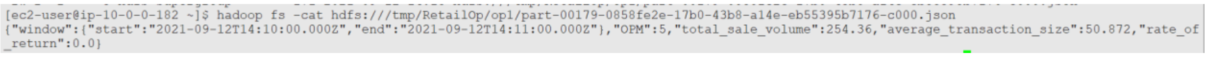
hadoop fs -ls hdfs:///tmp/RetailOp/



```
[ec2-user@ip-10-0-0-182 ~]$ hadoop fs -ls hdfs:///tmp/RetailOp/
Found 3 items
drwxr-xr-x   - hdfs supergroup          0 2021-09-12 14:11 hdfs:///tmp/RetailOp/CheckPoints
drwxr-xr-x   - hdfs supergroup          0 2021-09-12 14:15 hdfs:///tmp/RetailOp/op1
drwxr-xr-x   - hdfs supergroup          0 2021-09-12 14:16 hdfs:///tmp/RetailOp/op2
[ec2-user@ip-10-0-0-182 ~]$ █
```

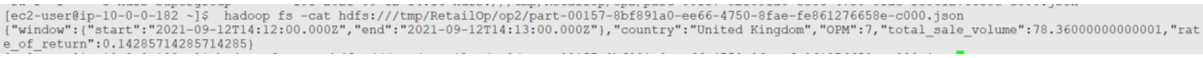
Sample results

hadoop fs -cat hdfs:///tmp/RetailOp/op1/part-00179-0858fe2e-17b0-43b8-a14e-eb55395b7176-c000.json



```
[ec2-user@ip-10-0-0-182 ~]$ hadoop fs -cat hdfs:///tmp/RetailOp/op1/part-00179-0858fe2e-17b0-43b8-a14e-eb55395b7176-c000.json
{"window":{"start":"2021-09-12T14:10:00.000Z","end":"2021-09-12T14:11:00.000Z"},"OPM":5,"total_sale_volume":254.36,"average_transaction_size":50.872,"rate_of
_return":0.0}
```

hadoop fs -cat hdfs:///tmp/RetailOp/op2/part-00157-8bf891a0-ee66-4750-8fae-fe861276658e-c000.json



```
[ec2-user@ip-10-0-0-182 ~]$  hadoop fs -cat hdfs:///tmp/RetailOp/op2/part-00157-8bf891a0-ee66-4750-8fae-fe861276658e-c000.json
{"window":{"start":"2021-09-12T14:12:00.000Z","end":"2021-09-12T14:13:00.000Z"},"country":"United Kingdom","OPM":7,"total_sale_volume":78.36000000000001,"rat
e_of_return":0.14285714285714285}
```