Code Logic - Retail Data Analysis

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

SETUP

Setting Environment before running the Spark Job:

```
export SPARK_KAFKA_VERSION=0.10
```

Downloaded the Jar file using:

https://search.maven.org/artifact/org.apache.spark/spark-sql-kafka-0-10_2.11/2.4.6/jar

Commands used to run the Spark Submit job:

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.6 spark-streaming.py
```

UTILITY FUNCTIONS

total_item_count: to sum up quantity of items ordered for each invoice total_cost: get total cost using→quantity * unit_price for each invoice is_a_order: return 1 if type is ORDER else 0
is_a_return: return 1 if type is RETURN else 0

STREAMS

order_stream: Input Stream [Raw data]
order_extended_stream – is the stream with the derived columns added to the raw data.

agg_time : Calculated the time-based KPIs with tumbling window of one minute on orders across the globe.

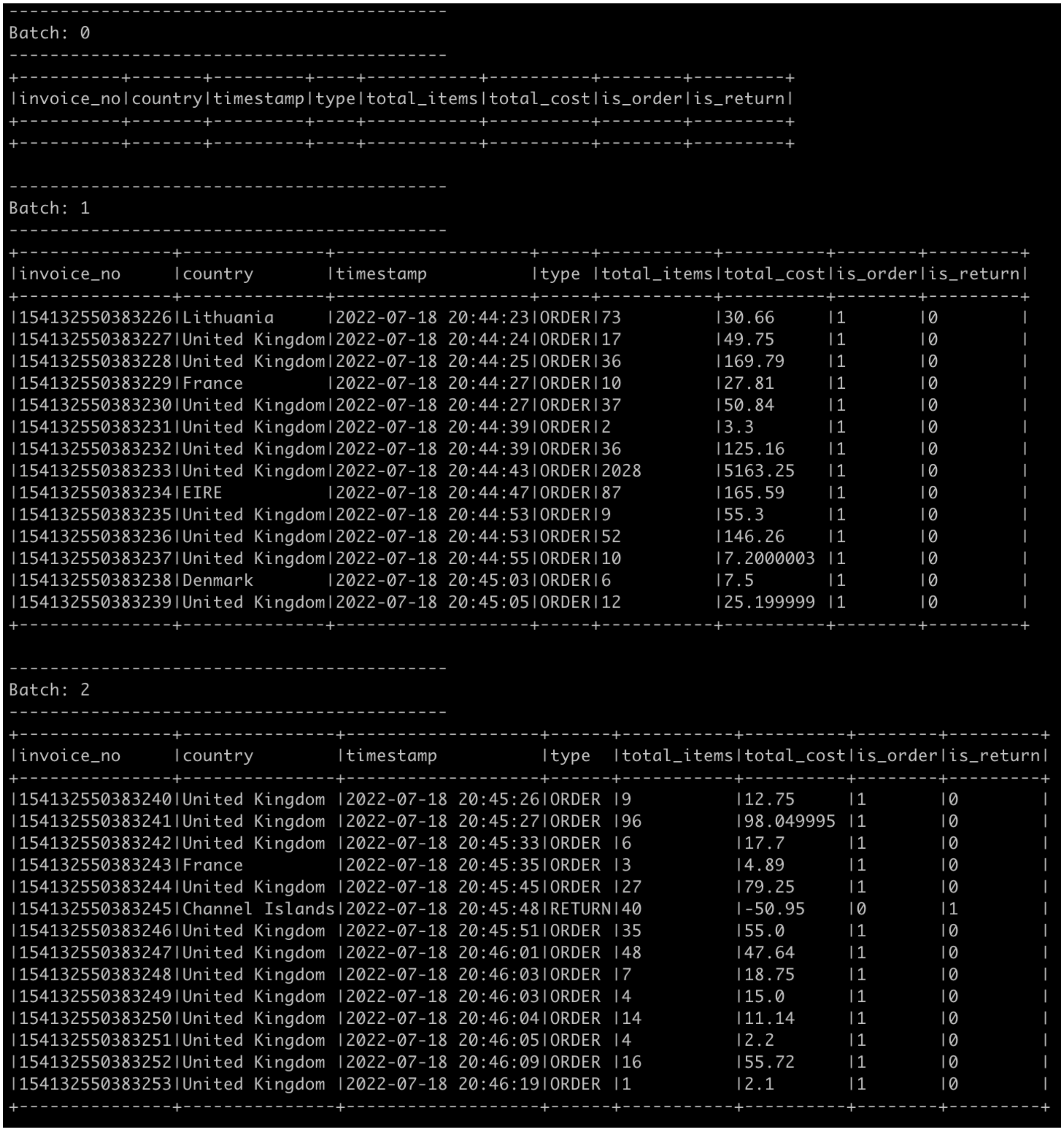
agg_time_country: Calculated the time and country-based KPIs with tumbling window of one minute on orders across the globe.

Output HDFS Dir for agg_time: timeKPI/
Output HDFS Dir for agg_time_country: time_countryKPI/

Coping HDFS Files

```
hadoop fs -get /user/hadoop/timeKPI/* /home/hadoop/timeKPI/
hadoop fs -get /user/hadoop/countryKPI/* /home/hadoop/countryKPI/
```

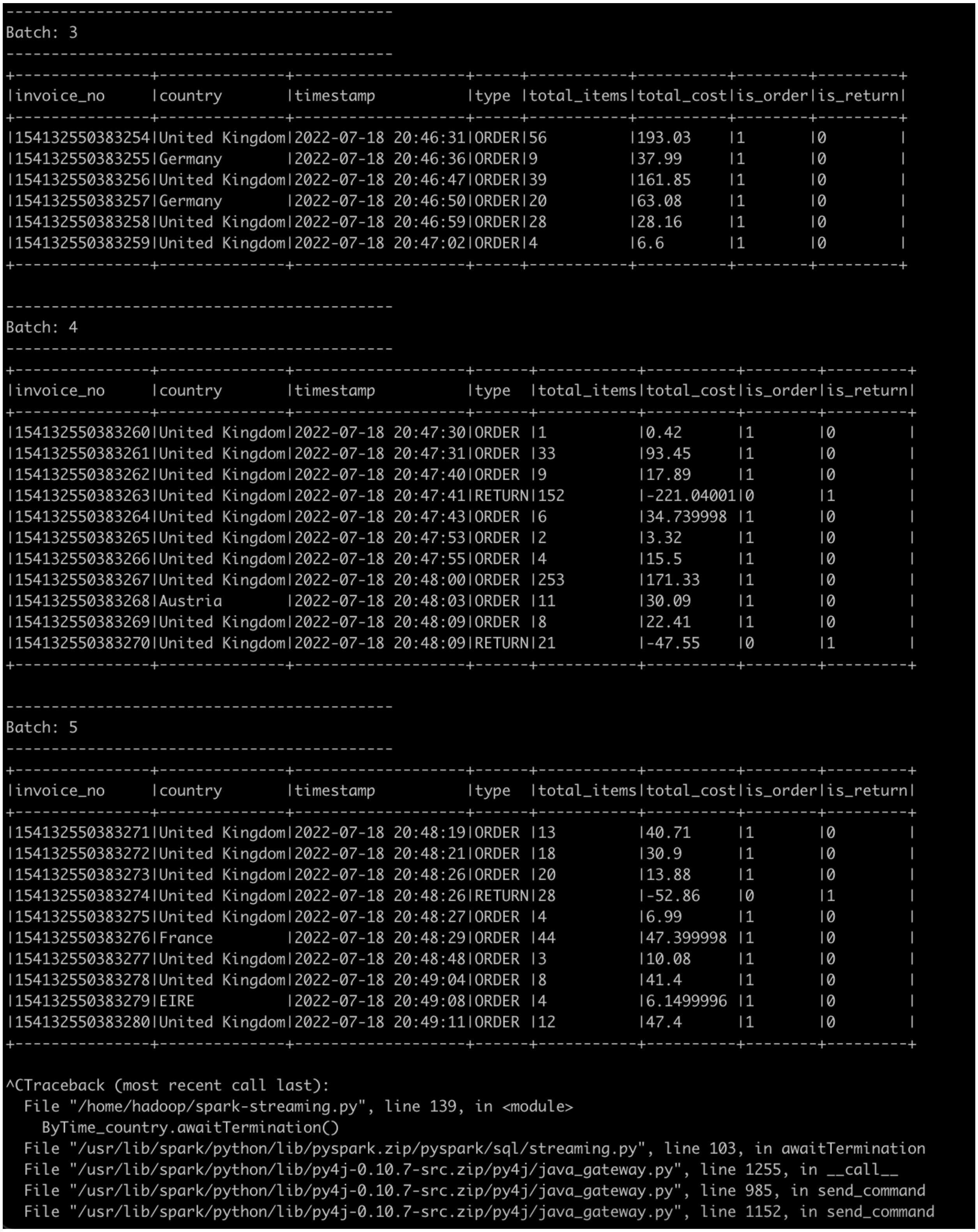
Console Output



```
-------------------------------------------
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|
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|154132550383226|Lithuania     |2022-07-18 20:44:23|ORDER|73         |30.66     |1       |0        |
|154132550383227|United Kingdom|2022-07-18 20:44:24|ORDER|17         |49.75     |1       |0        |
|154132550383228|United Kingdom|2022-07-18 20:44:25|ORDER|36         |169.79    |1       |0        |
|154132550383229|France        |2022-07-18 20:44:27|ORDER|10         |27.81     |1       |0        |
|154132550383230|United Kingdom|2022-07-18 20:44:27|ORDER|37         |50.84     |1       |0        |
|154132550383231|United Kingdom|2022-07-18 20:44:39|ORDER|2          |3.3       |1       |0        |
|154132550383232|United Kingdom|2022-07-18 20:44:39|ORDER|36         |125.16    |1       |0        |
|154132550383233|United Kingdom|2022-07-18 20:44:43|ORDER|2028       |5163.25   |1       |0        |
|154132550383234|EIRE          |2022-07-18 20:44:47|ORDER|87         |165.59    |1       |0        |
|154132550383235|United Kingdom|2022-07-18 20:44:53|ORDER|9          |55.3      |1       |0        |
|154132550383236|United Kingdom|2022-07-18 20:44:53|ORDER|52         |146.26    |1       |0        |
|154132550383237|United Kingdom|2022-07-18 20:44:55|ORDER|10         |7.2000003 |1       |0        |
|154132550383238|Denmark       |2022-07-18 20:45:03|ORDER|6          |7.5       |1       |0        |
|154132550383239|United Kingdom|2022-07-18 20:45:05|ORDER|12         |25.199999 |1       |0        |
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+

-------------------------------------------
Batch: 2
-------------------------------------------
+---------------+---------------+-------------------+------+-----------+----------+--------+---------+
|invoice_no     |country        |timestamp          |type  |total_items|total_cost|is_order|is_return|
+---------------+---------------+-------------------+------+-----------+----------+--------+---------+
|154132550383240|United Kingdom |2022-07-18 20:45:26|ORDER |9          |12.75     |1       |0        |
|154132550383241|United Kingdom |2022-07-18 20:45:27|ORDER |96         |98.049995 |1       |0        |
|154132550383242|United Kingdom |2022-07-18 20:45:33|ORDER |6          |17.7      |1       |0        |
|154132550383243|France         |2022-07-18 20:45:35|ORDER |3          |4.89      |1       |0        |
|154132550383244|United Kingdom |2022-07-18 20:45:45|ORDER |27         |79.25     |1       |0        |
|154132550383245|Channel Islands|2022-07-18 20:45:48|RETURN|40         |-50.95    |0       |1        |
|154132550383246|United Kingdom |2022-07-18 20:45:51|ORDER |35         |55.0      |1       |0        |
|154132550383247|United Kingdom |2022-07-18 20:46:01|ORDER |48         |47.64     |1       |0        |
|154132550383248|United Kingdom |2022-07-18 20:46:03|ORDER |7          |18.75     |1       |0        |
|154132550383249|United Kingdom |2022-07-18 20:46:03|ORDER |4          |15.0      |1       |0        |
|154132550383250|United Kingdom |2022-07-18 20:46:04|ORDER |14         |11.14     |1       |0        |
|154132550383251|United Kingdom |2022-07-18 20:46:05|ORDER |4          |2.2       |1       |0        |
|154132550383252|United Kingdom |2022-07-18 20:46:09|ORDER |16         |55.72     |1       |0        |
|154132550383253|United Kingdom |2022-07-18 20:46:19|ORDER |1          |2.1       |1       |0        |
+---------------+---------------+-------------------+------+-----------+----------+--------+---------+
```



```
-------------------------------------------
Batch: 3
-------------------------------------------
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|invoice_no     |country       |timestamp          |type |total_items|total_cost|is_order|is_return|
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|154132550383254|United Kingdom|2022-07-18 20:46:31|ORDER|56         |193.03    |1       |0        |
|154132550383255|Germany       |2022-07-18 20:46:36|ORDER|9          |37.99     |1       |0        |
|154132550383256|United Kingdom|2022-07-18 20:46:47|ORDER|39         |161.85    |1       |0        |
|154132550383257|Germany       |2022-07-18 20:46:50|ORDER|20         |63.08     |1       |0        |
|154132550383258|United Kingdom|2022-07-18 20:46:59|ORDER|28         |28.16     |1       |0        |
|154132550383259|United Kingdom|2022-07-18 20:47:02|ORDER|4          |6.6       |1       |0        |
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+

-------------------------------------------
Batch: 4
-------------------------------------------
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|invoice_no     |country       |timestamp          |type  |total_items|total_cost|is_order|is_return|
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|154132550383260|United Kingdom|2022-07-18 20:47:30|ORDER |1          |0.42      |1       |0        |
|154132550383261|United Kingdom|2022-07-18 20:47:31|ORDER |33         |93.45     |1       |0        |
|154132550383262|United Kingdom|2022-07-18 20:47:40|ORDER |9          |17.89     |1       |0        |
|154132550383263|United Kingdom|2022-07-18 20:47:41|RETURN|152        |-221.04001|0       |1        |
|154132550383264|United Kingdom|2022-07-18 20:47:43|ORDER |6          |34.739998 |1       |0        |
|154132550383265|United Kingdom|2022-07-18 20:47:53|ORDER |2          |3.32      |1       |0        |
|154132550383266|United Kingdom|2022-07-18 20:47:55|ORDER |4          |15.5      |1       |0        |
|154132550383267|United Kingdom|2022-07-18 20:48:00|ORDER |253        |171.33    |1       |0        |
|154132550383268|Austria       |2022-07-18 20:48:03|ORDER |11         |30.09     |1       |0        |
|154132550383269|United Kingdom|2022-07-18 20:48:09|ORDER |8          |22.41     |1       |0        |
|154132550383270|United Kingdom|2022-07-18 20:48:09|RETURN|21         |-47.55    |0       |1        |
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+

-------------------------------------------
Batch: 5
-------------------------------------------
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|invoice_no     |country       |timestamp          |type  |total_items|total_cost|is_order|is_return|
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+
|154132550383271|United Kingdom|2022-07-18 20:48:19|ORDER |13         |40.71     |1       |0        |
|154132550383272|United Kingdom|2022-07-18 20:48:21|ORDER |18         |30.9      |1       |0        |
|154132550383273|United Kingdom|2022-07-18 20:48:26|ORDER |20         |13.88     |1       |0        |
|154132550383274|United Kingdom|2022-07-18 20:48:26|RETURN|28         |-52.86    |0       |1        |
|154132550383275|United Kingdom|2022-07-18 20:48:27|ORDER |4          |6.99      |1       |0        |
|154132550383276|France        |2022-07-18 20:48:29|ORDER |44         |47.399998 |1       |0        |
|154132550383277|United Kingdom|2022-07-18 20:48:48|ORDER |3          |10.08     |1       |0        |
|154132550383278|United Kingdom|2022-07-18 20:49:04|ORDER |8          |41.4      |1       |0        |
|154132550383279|EIRE          |2022-07-18 20:49:08|ORDER |4          |6.1499996 |1       |0        |
|154132550383280|United Kingdom|2022-07-18 20:49:11|ORDER |12         |47.4      |1       |0        |
+---------------+--------------+-------------------+------+-----------+----------+--------+---------+

^CTraceback (most recent call last):
  File "/home/hadoop/spark-streaming.py", line 139, in <module>
    ByTime_country.awaitTermination()
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/streaming.py", line 103, in awaitTermination
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1255, in __call__
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 985, in send_command
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1152, in send_command
```