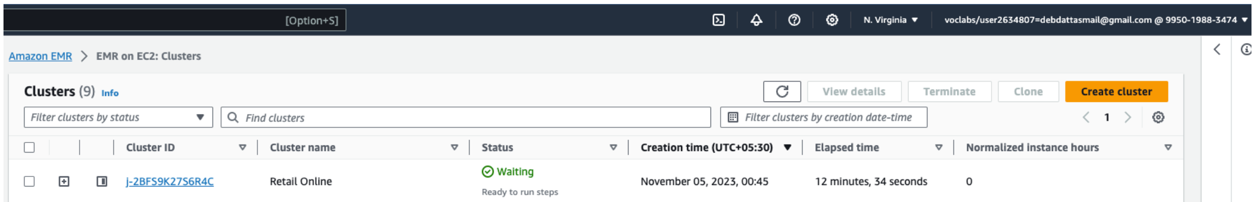
CODE LOGIC - RETAIL DATA ANALYSIS

EMR Created:


[Option+S]
N. Virginia ▼
voclabs/user2634807=debdattasmail@gmail.com @ 9950-1988-3474 ▼
Amazon EMR > EMR on EC2: Clusters
Clusters (9) Info
View details
Terminate
Clone
Create cluster
Filter clusters by status
Find clusters
Filter clusters by creation date-time
1

☐			Cluster ID	Cluster name	Status	Creation time (UTC+05:30)	Elapsed time	Normalized instance hours
☐	⊞	◫	j-2BFS9K27S6R4C	Retail Online	✓ Waiting Ready to run steps	November 05, 2023, 00:45	12 minutes, 34 seconds	0

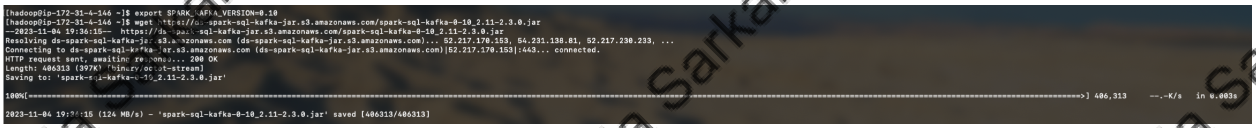
SSH into EMR:

ssh -i debskp.pem hadoop@ec2-3-83-4-3.compute-1.amazonaws.com

Setup Your EMR:

export SPARK_KAFKA_VERSION=0.10

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



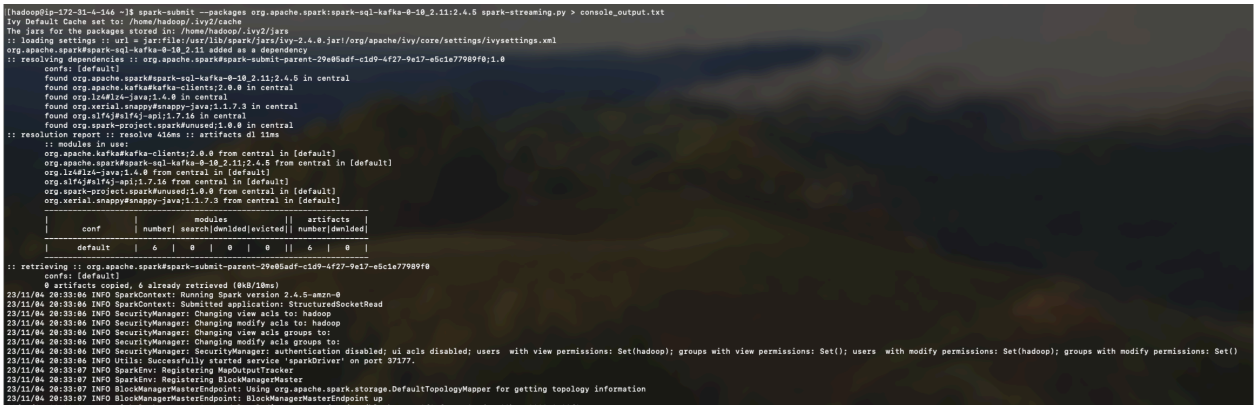
```
[hadoop@ip-172-31-4-146 ~]$ export SPARK_KAFKA_VERSION=0.10
[hadoop@ip-172-31-4-146 ~]$ wget https://ds-spark-sql-kafka-jar.s3.amazonaws.com/spark-sql-kafka-0-10_2.11-2.3.0.jar
--2023-11-04 19:36:15--  https://ds-spark-sql-kafka-jar.s3.amazonaws.com/spark-sql-kafka-0-10_2.11-2.3.0.jar
Resolving ds-spark-sql-kafka-jar.s3.amazonaws.com (ds-spark-sql-kafka-jar.s3.amazonaws.com)... 52.217.170.153, 54.231.138.81, 52.217.230.233, ...
Connecting to ds-spark-sql-kafka-jar.s3.amazonaws.com (ds-spark-sql-kafka-jar.s3.amazonaws.com)|52.217.170.153|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 406313 (397K) [binary/octet-stream]
Saving to: 'spark-sql-kafka-0-10_2.11-2.3.0.jar'

100%[==================================================================================================================================================================>] 406,313     --.-K/s   in 0.003s

2023-11-04 19:36:15 (124 MB/s) - 'spark-sql-kafka-0-10_2.11-2.3.0.jar' saved [406313/406313]
```

Command to Run to Get the Output within a File:

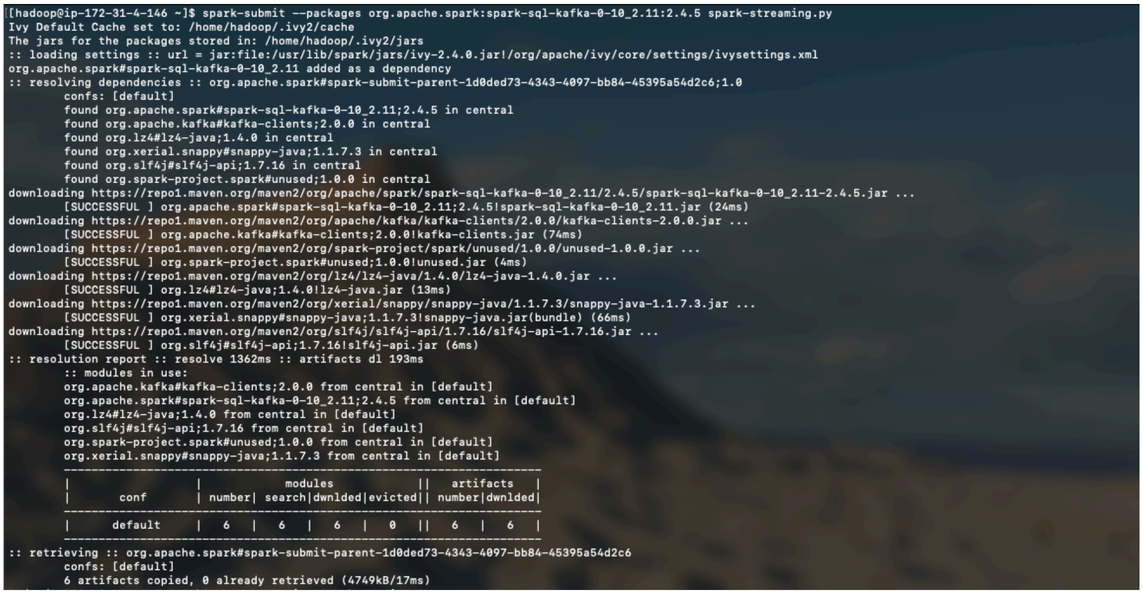
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py > console_output.txt



```
[hadoop@ip-172-31-4-146 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py > console_output.txt
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-29e05adf-c1d9-4f27-9e17-e5c1e77989f0;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 in central
        found org.apache.kafka#kafka-clients;2.0.0 in central
        found org.lz4#lz4-java;1.4.0 in central
        found org.xerial.snappy#snappy-java;1.1.7.3 in central
        found org.slf4j#slf4j-api;1.7.16 in central
        found org.spark-project.spark#unused;1.0.0 in central
:: resolution report :: resolve 416ms :: artifacts dl 11ms
        :: modules in use:
        org.apache.kafka#kafka-clients;2.0.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 from central in [default]
        org.lz4#lz4-java;1.4.0 from central in [default]
        org.slf4j#slf4j-api;1.7.16 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.7.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   6   |   0   |   0   |   0   ||   6   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-29e05adf-c1d9-4f27-9e17-e5c1e77989f0
        confs: [default]
        0 artifacts copied, 6 already retrieved (0kB/10ms)
23/11/04 20:33:06 INFO SparkContext: Running Spark version 2.4.5-amzn-0
23/11/04 20:33:06 INFO SparkContext: Submitted application: StructuredSocketRead
23/11/04 20:33:06 INFO SecurityManager: Changing view acls to: hadoop
23/11/04 20:33:06 INFO SecurityManager: Changing modify acls to: hadoop
23/11/04 20:33:06 INFO SecurityManager: Changing view acls groups to:
23/11/04 20:33:06 INFO SecurityManager: Changing modify acls groups to:
23/11/04 20:33:06 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permissions: Set(); users  with modify permissions: Set(hadoop); groups with modify permissions: Set()
23/11/04 20:33:06 INFO Utils: Successfully started service 'sparkDriver' on port 37177.
23/11/04 20:33:07 INFO SparkEnv: Registering MapOutputTracker
23/11/04 20:33:07 INFO SparkEnv: Registering BlockManagerMaster
23/11/04 20:33:07 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
23/11/04 20:33:07 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
```

Alternative Command to Run to Get the Output on Console:

spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py

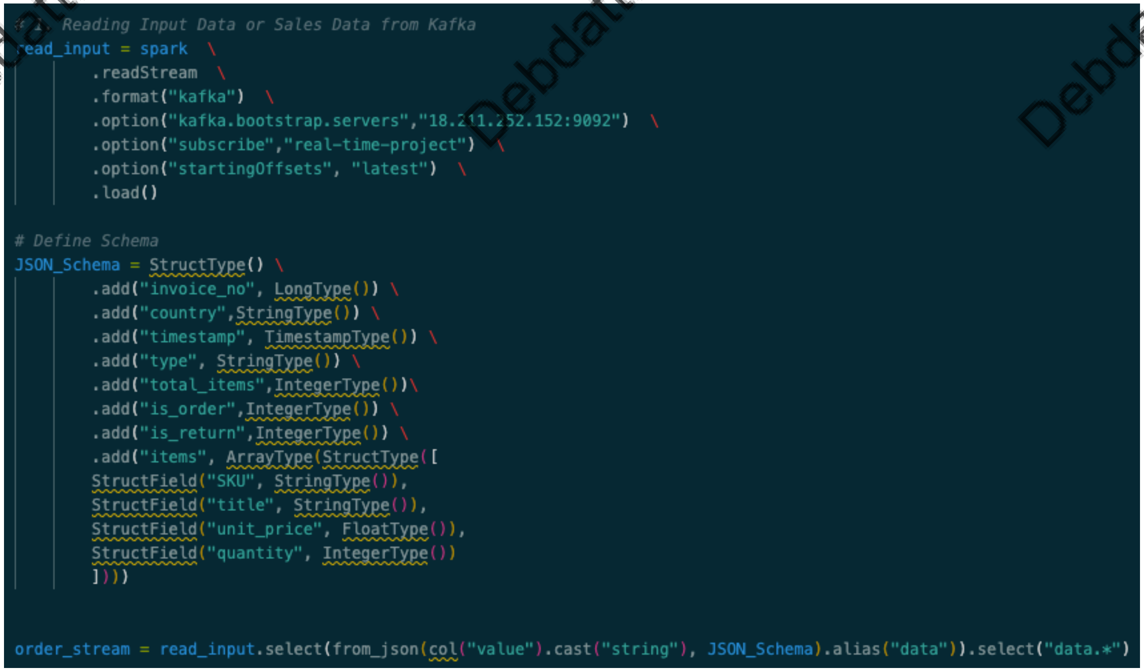


```
[hadoop@ip-172-31-4-146 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-1d0ded73-4343-4097-bb84-45395a54d2c6;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 in central
        found org.apache.kafka#kafka-clients;2.0.0 in central
        found org.lz4#lz4-java;1.4.0 in central
        found org.xerial.snappy#snappy-java;1.1.7.3 in central
        found org.slf4j#slf4j-api;1.7.16 in central
        found org.spark-project.spark#unused;1.0.0 in central
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.11/2.4.5/spark-sql-kafka-0-10_2.11-2.4.5.jar ...
        [SUCCESSFUL ] org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5!spark-sql-kafka-0-10_2.11.jar (24ms)
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/2.0.0/kafka-clients-2.0.0.jar ...
        [SUCCESSFUL ] org.apache.kafka#kafka-clients;2.0.0!kafka-clients.jar (74ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.0.0.jar ...
        [SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (4ms)
downloading https://repo1.maven.org/maven2/org/lz4/lz4-java/1.4.0/lz4-java-1.4.0.jar ...
        [SUCCESSFUL ] org.lz4#lz4-java;1.4.0!lz4-java.jar (13ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.7.3/snappy-java-1.1.7.3.jar ...
        [SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.7.3!snappy-java.jar(bundle) (66ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.16/slf4j-api-1.7.16.jar ...
        [SUCCESSFUL ] org.slf4j#slf4j-api;1.7.16!slf4j-api.jar (6ms)
:: resolution report :: resolve 1362ms :: artifacts dl 193ms
        :: modules in use:
        org.apache.kafka#kafka-clients;2.0.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 from central in [default]
        org.lz4#lz4-java;1.4.0 from central in [default]
        org.slf4j#slf4j-api;1.7.16 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.7.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   6   |   6   |   6   |   0   ||   6   |   6   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-1d0ded73-4343-4097-bb84-45395a54d2c6
        confs: [default]
        6 artifacts copied, 0 already retrieved (4749kB/17ms)
```

Code Written for the steps within spark-streaming.py file:

READING INPUT DATA FROM KAFKA

order_stream: Input Stream [Raw data]

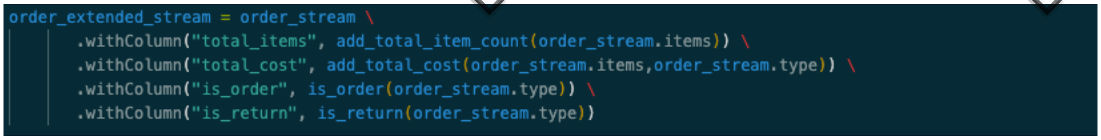


```python
# 1. Reading Input Data or Sales Data from Kafka
read_input = spark  \
        .readStream  \
        .format("kafka")  \
        .option("kafka.bootstrap.servers","18.211.252.152:9092")  \
        .option("subscribe","real-time-project")  \
        .option("startingOffsets", "latest")  \
        .load()

# Define Schema
JSON_Schema = StructType() \
        .add("invoice_no", LongType()) \
        .add("country",StringType()) \
        .add("timestamp", TimestampType()) \
        .add("type", StringType()) \
        .add("total_items",IntegerType())\
        .add("is_order",IntegerType()) \
        .add("is_return",IntegerType()) \
        .add("items", ArrayType(StructType([
        StructField("SKU", StringType()),
        StructField("title", StringType()),
        StructField("unit_price", FloatType()),
        StructField("quantity", IntegerType())
        ])))


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

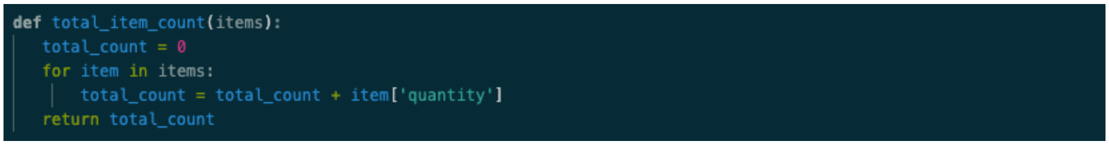
order_extended_stream: Is the code which streams the derived columns added to the raw data.



```
order_extended_stream = order_stream \
        .withColumn("total_items", add_total_item_count(order_stream.items)) \
        .withColumn("total_cost", add_total_cost(order_stream.items,order_stream.type)) \
        .withColumn("is_order", is_order(order_stream.type)) \
        .withColumn("is_return", is_return(order_stream.type))
```

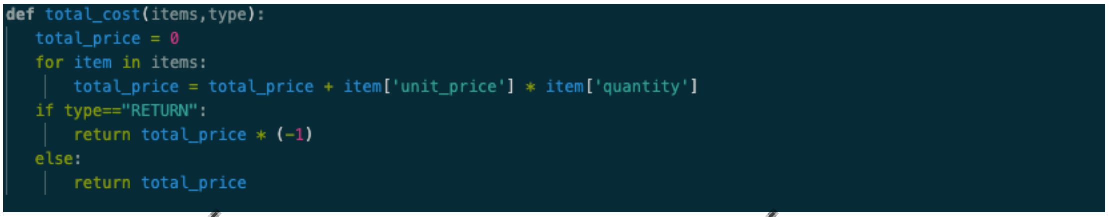
Calculating additional columns

total_item_count: Sum up quantity of items ordered for each invoice



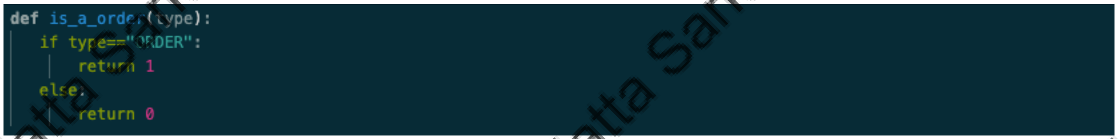
```
def total_item_count(items):
    total_count = 0
    for item in items:
        total_count = total_count + item['quantity']
    return total_count
```

total_cost: Gets the total cost using calculation → quantity * unit_price for each invoice



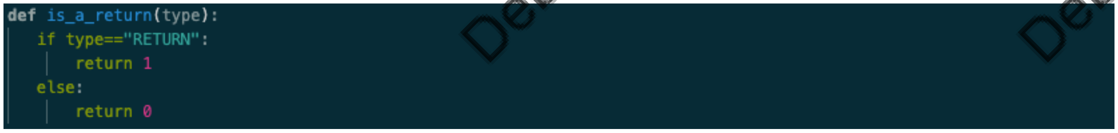
```
def total_cost(items,type):
    total_price = 0
    for item in items:
        total_price = total_price + item['unit_price'] * item['quantity']
    if type=="RETURN":
        return total_price * (-1)
    else:
        return total_price
```

is_a_order: Return 1 if type is 'ORDER' else 0



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

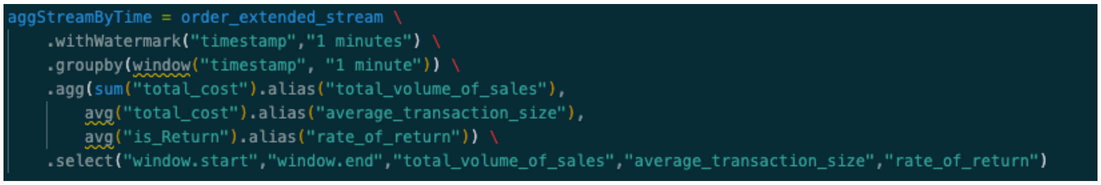
is_a_return: Return 1 if type is 'RETURN' else 0



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

Calculating Time-based KPIs

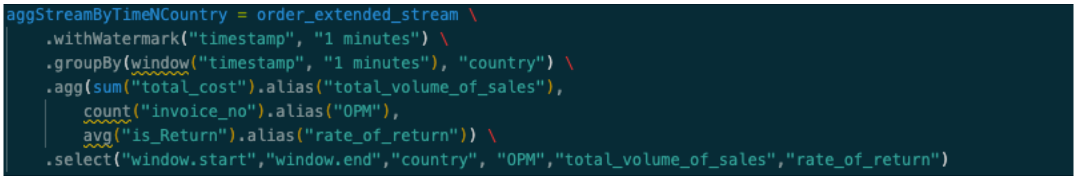
aggStreamByTime: Calculates the time-based KPIs with tumbling window of one minute on orders across the globe.



```
aggStreamByTime = order_extended_stream \
    .withWatermark("timestamp","1 minutes") \
    .groupby(window("timestamp", "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_transaction_size","rate_of_return")
```

CALCULATING TIME- AND COUNTRY-BASED KPIs

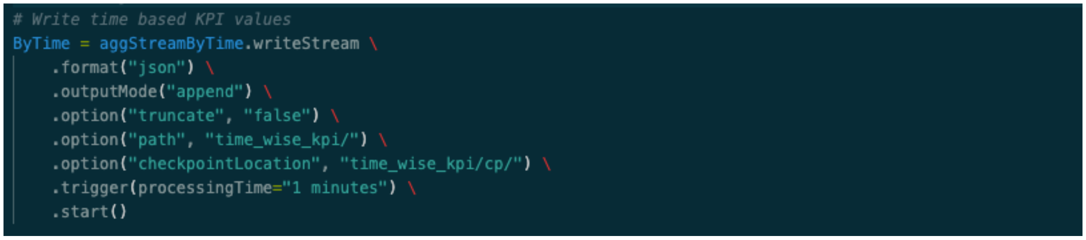
aggStreamByTimeNCountry: Calculates the time and country-based KPIs with tumbling window of one minute on orders across the globe.



```
aggStreamByTimeNCountry = order_extended_stream \
    .withWatermark("timestamp", "1 minutes") \
    .groupBy(window("timestamp", "1 minutes"), "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")
```

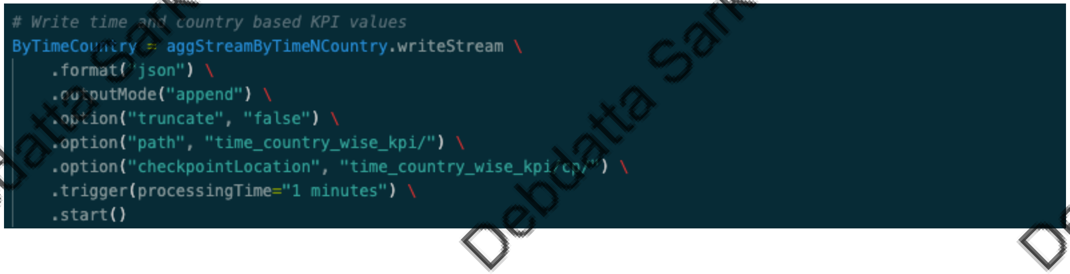
WRITING KPIs TO JSON FILES

aggStreamByTime: Gives the output into HDFS Directory named time_wise_kpi/



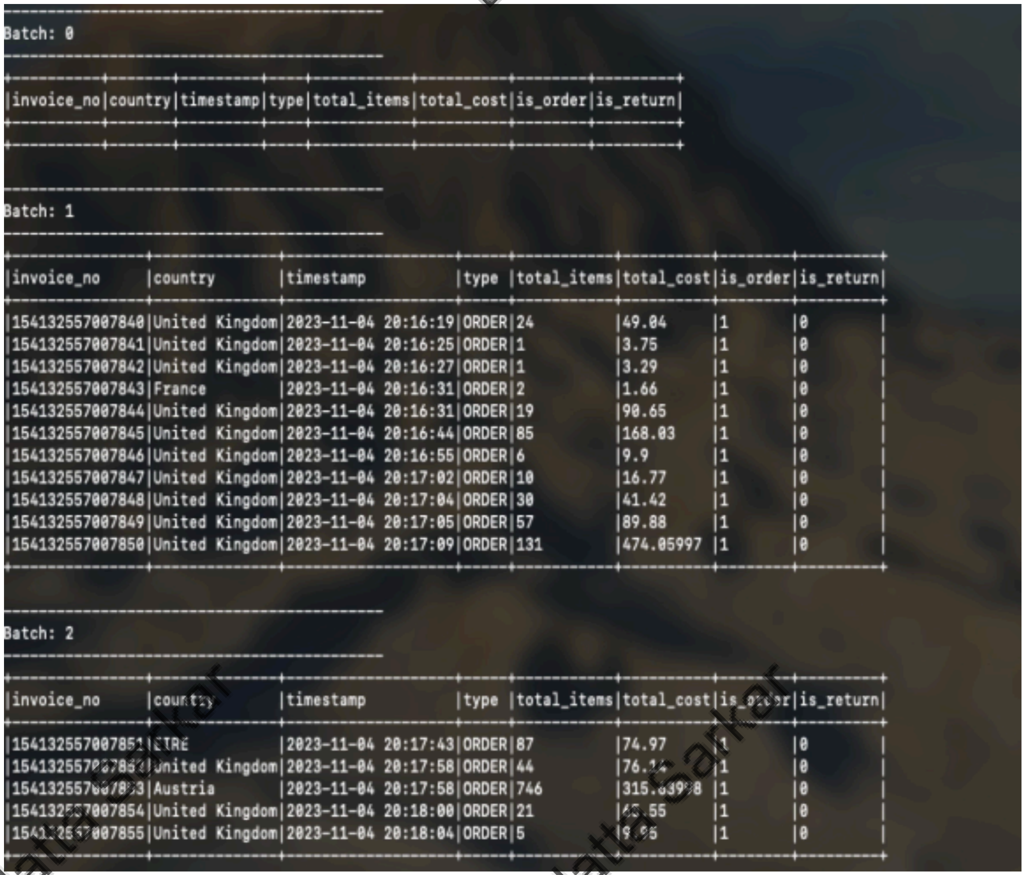
```
# Write time based KPI values
ByTime = aggStreamByTime.writeStream \
    .format("json") \
    .outputMode("append") \
    .option("truncate", "false") \
    .option("path", "time_wise_kpi/") \
    .option("checkpointLocation", "time_wise_kpi/cp/") \
    .trigger(processingTime="1 minutes") \
    .start()
```

aggStreamByTimeNCountry: Gives the output into HDFS Directory named time_country_wise_kpi/



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

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|
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|154132557007840|United Kingdom|2023-11-04 20:16:19|ORDER|24         |49.04     |1       |0        |
|154132557007841|United Kingdom|2023-11-04 20:16:25|ORDER|1          |3.75      |1       |0        |
|154132557007842|United Kingdom|2023-11-04 20:16:27|ORDER|1          |3.29      |1       |0        |
|154132557007843|France        |2023-11-04 20:16:31|ORDER|2          |1.66      |1       |0        |
|154132557007844|United Kingdom|2023-11-04 20:16:31|ORDER|19         |90.65     |1       |0        |
|154132557007845|United Kingdom|2023-11-04 20:16:44|ORDER|85         |168.03    |1       |0        |
|154132557007846|United Kingdom|2023-11-04 20:16:55|ORDER|6          |9.9       |1       |0        |
|154132557007847|United Kingdom|2023-11-04 20:17:02|ORDER|10         |16.77     |1       |0        |
|154132557007848|United Kingdom|2023-11-04 20:17:04|ORDER|30         |41.42     |1       |0        |
|154132557007849|United Kingdom|2023-11-04 20:17:05|ORDER|57         |89.88     |1       |0        |
|154132557007850|United Kingdom|2023-11-04 20:17:09|ORDER|131        |474.05997 |1       |0        |
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+

-------------------------------------------
Batch: 2
-------------------------------------------
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|invoice_no     |country       |timestamp          |type |total_items|total_cost|is_order|is_return|
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
|154132557007851|EIRE          |2023-11-04 20:17:43|ORDER|87         |74.97     |1       |0        |
|154132557007852|United Kingdom|2023-11-04 20:17:58|ORDER|44         |76.14     |1       |0        |
|154132557007853|Austria       |2023-11-04 20:17:58|ORDER|746        |315.63998 |1       |0        |
|154132557007854|United Kingdom|2023-11-04 20:18:00|ORDER|21         |45.55     |1       |0        |
|154132557007855|United Kingdom|2023-11-04 20:18:04|ORDER|5          |9.95      |1       |0        |
+---------------+--------------+-------------------+-----+-----------+----------+--------+---------+
```