Code Logic - Retail Data Analysis

PySpark Code Logic

Step 1: Importing the required modules - *System dependencies for CDH* and declaring PySpark environment variables

 # importing the modules - System dependencies for CDH

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
from pyspark.sql.functions import from_json
from pyspark.sql.window import Window
```

Step 2: Creating Initial Spark Session

 # Initializing Spark Session

```
spark = SparkSession \
    .builder \
    .appName("RetailDataAnalysis") \
    .getOrCreate()
spark.sparkContext.setLogLevel('ERROR')
```

Step 3: Reading the Input from Kafka

 # Read Input from kafka

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

Step 4: *Defining the schema for incoming data JSON file data*

 # Defining the Schema

```
jsonSchema = 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", DoubleType()),
        StructField("quantity", IntegerType())
])))
```

Step 5: *Creating dataframe from incoming data*

 # Creating dataframe from input data after applying the schema

```
orderStream = orderRaw.select(from_json(col("value").cast("string"),
jsonSchema).alias("data")).select("data.*")
```

Step 6: *Defining UDF functions (total_tems, total_cost, is_order, is_return), conversion to UDF types and calculating columns*

 # UDF for calculating total_items

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

```python
# UDF for calculating order type

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


# UDF for calculating return type

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


# UDF for calculating total_cost

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


# Converting to UDF's with the utility functions

isorder = udf(is_order, IntegerType())
isreturn = udf(is_return, IntegerType())
totalcount = udf(items_TotalCount, IntegerType())
totalcost = udf(TotalCostSum, DoubleType())
```

 # Calculating columns(total_cost, total_items, is_order, is_return)

```
order_stream = orderStream \
        .withColumn("total_cost", totalcost(orderStream.items, orderStream.type)) \
        .withColumn("total_items", totalcount(orderStream.items)) \
        .withColumn("is_order", isorder(orderStream.type)) \
        .withColumn("is_return", isreturn(orderStream.type))
```

Step 7: Writing intermediate dataset to console with 1 Minute Interval

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

Step 8:*Calculating time-based KPIs using with Watermark and groupBy*

 # Calculating time based KPIs

```
timeBasedKPIs = order_stream \
        .withWatermark("timestamp", "1 minute") \
        .groupby(window("timestamp", "1 minute", "1 minute")) \
        .agg(count("invoice_no").alias("OPM"),
                sum("total_cost").alias("total_sales_volume"),
                avg("total_cost").alias("average_transaction_size"),
                avg("is_return").alias("rate_of_return")) \
        .select("window", "OPM", "total_sales_volume", "average_transaction_size",
        "rate_of_return")
```

 # write stream for time based KPIs

```
timeBasedKPIsOutput = timeBasedKPIs \
        .writeStream \
        .outputMode("append") \
        .format("json") \
        .option("format","append") \
        .option("truncate", "false") \
        .option("path", "time-wise-kpi") \
        .option("checkpointLocation", "time-kpi") \
        .trigger(processingTime="1 minute") \
        .start()
```

Step 9: *Calculating time-based and country-based KPIs using withWatermark and groupBy*

 # Calculating time-based and country-based KPIs

```
timeAndCountryBasedKPIs = order_stream \
        .withWatermark("timestamp", "1 minute") \
        .groupby(window("timestamp", "1 minute", "1 minute"), "country") \
        .agg(count("invoice_no").alias("OPM"),
                sum("total_cost").alias("total_sales_volume"),
                avg("is_return").alias("rate_of_return")) \
        .select("window", "country", "OPM", "total_sales_volume", "rate_of_return")
```

 # write stream for time and country based KPIs

```
timeAndCountryBasedKPIsOutput = timeAndCountryBasedKPIs \
        .writeStream \
        .outputMode("Append") \
        .format("json") \
        .option("format","append") \
        .option("truncate", "false") \
        .option("path", "time-country-wise-kpi") \
        .option("checkpointLocation","time-country-kpi") \
```

```
.trigger(processingTime="1 minute") \
.start()
```

Step 10: Waiting for the termination of stream infinitely

 # Waiting infinitely to read the data

timeAndCountryBasedKPIsOutput.awaitTermination()

Console Commands and Analysis

Spark Submit Command and too Console-output file

Spark2-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py
18.211.252.152 9092 real-time-project > Console-output

Mkdir command to make directory

mkdir time-wise-kpi
mkdir time-country-wise-kpi

command to check json files generated

hadoop fs -ls

hadoop fs -ls time-wise-kpi
hadoop fs -ls time-country-wise-kpi

hadoop fs -cat time-wise-kpi/part-00198-21224b93-1ac7-4748-806b-d39e02d0b82a-c000/
hadoop fs -cat time-country-wise-kpi/part-00185-ffa2a99d-0618-41c2-abab-cb73865996b7-
c000/

Transfer of data from HDFS to Local using

- hadoop fs -get /user/hadoop/time-wise-kpi ./time-wise-kpi
- hadoop fs -get /user/hadoop/time-country-wise-kpi ./time-country-wise-kpi
- WinSCP used to copy the files to local machine

Output Screens
Summarized Console Output

```
[hadoop@ip-172-31-86-58 ~]$ clear
[hadoop@ip-172-31-86-58 ~]$ cat Console-output.txt
----------------------------------------
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_retu
+---------------+--------------+-------------------+----------+-----------+--------+------
|154132554154207|France        |2023-04-16 15:41:01|1.25      |1          |1       |0
|154132554154208|United Kingdom|2023-04-16 15:41:08|105.9     |36         |1       |0
|154132554154209|United Kingdom|2023-04-16 15:41:12|84.6      |32         |1       |0
|154132554154210|United Kingdom|2023-04-16 15:41:15|34.15     |19         |1       |0
|154132554154211|Germany       |2023-04-16 15:41:28|37.41     |33         |1       |0
|154132554154212|Spain         |2023-04-16 15:41:34|-165.0    |100        |0       |1
|154132554154213|United Kingdom|2023-04-16 15:41:38|19.17     |8          |1       |0
|154132554154214|United Kingdom|2023-04-16 15:41:45|73.48     |55         |1       |0
+---------------+--------------+-------------------+----------+-----------+--------+------


----------------------------------------
Batch: 2
----------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+------
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_retu
+---------------+--------------+-------------------+----------+-----------+--------+------
|154132554154215|United Kingdom|2023-04-16 15:41:58|258.96    |24         |1       |0
|154132554154216|United Kingdom|2023-04-16 15:42:03|100.78    |31         |1       |0
|154132554154217|United Kingdom|2023-04-16 15:42:05|1.25      |1          |1       |0
|154132554154218|United Kingdom|2023-04-16 15:42:10|114.16    |79         |1       |0
|154132554154219|United Kingdom|2023-04-16 15:42:16|12.86     |26         |1       |0
```

```
Batch: 3
-------------------------------------------
+-------------+--------------+-----------------+----------+----------+--------+-----
|invoice_no   |country       |timestamp        |total_cost|total_items|is_order|is_ret
+-------------+--------------+-----------------+----------+----------+--------+-----
|154132554154228|United Kingdom|2023-04-16 15:43:05|187.76    |16        |1       |0
|154132554154229|United Kingdom|2023-04-16 15:43:13|5.3999996 |30        |1       |0
|154132554154230|United Kingdom|2023-04-16 15:43:15|386.66    |117       |1       |0
|154132554154231|United Kingdom|2023-04-16 15:43:33|54.04     |20        |1       |0
+-------------+--------------+-----------------+----------+----------+--------+-----


-------------------------------------------
Batch: 4
-------------------------------------------
+-------------+--------------+-----------------+----------+----------+--------+-----
|invoice_no   |country       |timestamp        |total_cost|total_items|is_order|is_ret
+-------------+--------------+-----------------+----------+----------+--------+-----
|154132554154232|United Kingdom|2023-04-16 15:43:51|-15.38    |14        |0       |1
|154132554154233|United Kingdom|2023-04-16 15:43:51|182.06    |84        |1       |0
|154132554154234|United Kingdom|2023-04-16 15:44:03|36.55     |11        |1       |0
|154132554154235|United Kingdom|2023-04-16 15:44:07|55.13     |19        |1       |0
|154132554154236|United Kingdom|2023-04-16 15:44:11|45.7      |38        |1       |0
|154132554154237|United Kingdom|2023-04-16 15:44:12|4.2       |2         |1       |0
|154132554154238|Malta         |2023-04-16 15:44:12|-118.08   |48        |0       |1
|154132554154239|United Kingdom|2023-04-16 15:44:17|19.87     |15        |1       |0
|154132554154240|United Kingdom|2023-04-16 15:44:20|81.36     |44        |1       |0
|154132554154241|Germany       |2023-04-16 15:44:21|302.13    |151       |1       |0
|154132554154242|EIRE          |2023-04-16 15:44:23|266.16    |97        |1       |0
|154132554154243|United Kingdom|2023-04-16 15:44:28|18.27     |13        |1       |0
|154132554154244|EIRE          |2023-04-16 15:44:33|33.47     |13        |1       |0
|154132554154245|Japan         |2023-04-16 15:44:34|9.9       |6         |1       |0
+-------------+--------------+-----------------+----------+----------+--------+-----
```

```
----------------------------------------
Batch: 5
----------------------------------------
+--------------+--------------+-------------------+----------+-----------+--------+------
|invoice_no    |country       |timestamp          |total_cost|total_items|is_order|is_ret
+--------------+--------------+-------------------+----------+-----------+--------+------
|154132554154246|United Kingdom|2023-04-16 15:44:56|38.9      |15         |1       |0
|154132554154247|United Kingdom|2023-04-16 15:44:57|19.8      |12         |1       |0
|154132554154248|EIRE          |2023-04-16 15:45:00|147.14    |73         |1       |0
|154132554154249|United Kingdom|2023-04-16 15:45:04|55.559998 |24         |1       |0
|154132554154250|United Kingdom|2023-04-16 15:45:04|36.98     |14         |1       |0
|154132554154251|United Kingdom|2023-04-16 15:45:10|-513.73   |592        |0       |1
|154132554154252|United Kingdom|2023-04-16 15:45:10|72.82     |29         |1       |0
|154132554154253|United Kingdom|2023-04-16 15:45:16|135.52    |87         |1       |0
|154132554154254|United Kingdom|2023-04-16 15:45:38|23.16     |36         |1       |0
|154132554154255|EIRE          |2023-04-16 15:45:42|120.29    |97         |1       |0
|154132554154256|United Kingdom|2023-04-16 15:45:43|196.68999 |89         |1       |0
+--------------+--------------+-------------------+----------+-----------+--------+------


----------------------------------------
Batch: 6
----------------------------------------
+--------------+--------------+-------------------+----------+-----------+--------+------
|invoice_no    |country       |timestamp          |total_cost|total_items|is_order|is_ret
+--------------+--------------+-------------------+----------+-----------+--------+------
|154132554154257|United Kingdom|2023-04-16 15:45:49|283.78    |172        |1       |0
|154132554154258|United Kingdom|2023-04-16 15:45:53|663.0     |100        |1       |0
|154132554154259|United Kingdom|2023-04-16 15:45:55|67.380005 |18         |1       |0
|154132554154260|Germany       |2023-04-16 15:45:59|32.32     |73         |1       |0
|154132554154261|EIRE          |2023-04-16 15:46:01|24.9      |14         |1       |0
|154132554154262|United Kingdom|2023-04-16 15:46:03|16.199999 |15         |1       |0
|154132554154263|United Kingdom|2023-04-16 15:46:05|115.65    |53         |1       |0
|154132554154264|United Kingdom|2023-04-16 15:46:05|63.78     |26         |1       |0
|154132554154265|United Kingdom|2023-04-16 15:46:11|9.95      |1          |1       |0
|154132554154266|United Kingdom|2023-04-16 15:46:12|2.46      |1          |1       |0
|154132554154267|United Kingdom|2023-04-16 15:46:18|178.34999 |65         |1       |0
|154132554154268|United Kingdom|2023-04-16 15:46:28|29.5      |10         |1       |0
|154132554154269|United Kingdom|2023-04-16 15:46:31|165.84999 |101        |1       |0
|154132554154270|United Kingdom|2023-04-16 15:46:40|15.0      |12         |1       |0
|154132554154271|United Kingdom|2023-04-16 15:46:41|8.3       |2          |1       |0
|154132554154272|United Kingdom|2023-04-16 15:46:47|20.74     |8          |1       |0
+--------------+--------------+-------------------+----------+-----------+--------+------
```

```
----------------------------------------
Batch: 7
----------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+-----
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_ret
+---------------+--------------+-------------------+----------+-----------+--------+-----
|154132554154273|United Kingdom|2023-04-16 15:46:52|227.66    |75         |1       |0
|154132554154274|United Kingdom|2023-04-16 15:47:06|-31.3     |6          |0       |1
|154132554154275|United Kingdom|2023-04-16 15:47:06|270.74    |150        |1       |0
|154132554154276|United Kingdom|2023-04-16 15:47:46|12.719999 |12         |1       |0
+---------------+--------------+-------------------+----------+-----------+--------+-----


----------------------------------------
Batch: 8
----------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+-----
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_ret
+---------------+--------------+-------------------+----------+-----------+--------+-----
|154132554154277|United Kingdom|2023-04-16 15:47:49|122.020004|50         |1       |0
|154132554154278|United Kingdom|2023-04-16 15:47:57|10.04     |6          |1       |0
|154132554154279|United Kingdom|2023-04-16 15:47:57|6.6       |4          |1       |0
|154132554154280|United Kingdom|2023-04-16 15:47:59|77.35     |25         |1       |0
|154132554154281|United Kingdom|2023-04-16 15:48:08|85.399994 |86         |1       |0
|154132554154282|United Kingdom|2023-04-16 15:48:21|-20.8     |10         |0       |1
|154132554154283|United Kingdom|2023-04-16 15:48:27|159.17    |74         |1       |0
|154132554154284|United Kingdom|2023-04-16 15:48:27|17.42     |4          |1       |0
|154132554154285|United Kingdom|2023-04-16 15:48:35|16.89     |11         |1       |0
+---------------+--------------+-------------------+----------+-----------+--------+-----


----------------------------------------
Batch: 9
----------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+-----
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_ret
+---------------+--------------+-------------------+----------+-----------+--------+-----
|154132554154286|United Kingdom|2023-04-16 15:48:46|24.710001 |13         |1       |0
|154132554154287|United Kingdom|2023-04-16 15:48:48|18.72     |76         |1       |0
|154132554154288|United Kingdom|2023-04-16 15:48:53|36.82     |7          |1       |0
|154132554154289|EIRE          |2023-04-16 15:48:53|17.55     |25         |1       |0
|154132554154290|United Kingdom|2023-04-16 15:48:58|21.75     |7          |1       |0
|154132554154291|United Kingdom|2023-04-16 15:49:04|19.05     |8          |1       |0
|154132554154292|United Kingdom|2023-04-16 15:49:07|65.52     |10         |1       |0
```

```
Batch: 9
-------------------------------------------
+--------------+--------------+------------------+----------+-----------+--------+-----
|invoice_no    |country       |timestamp         |total_cost|total_items|is_order|is_ret
+--------------+--------------+------------------+----------+-----------+--------+-----
|154132554154286|United Kingdom|2023-04-16 15:48:46|24.710001 |13         |1       |0
|154132554154287|United Kingdom|2023-04-16 15:48:48|18.72     |76         |1       |0
|154132554154288|United Kingdom|2023-04-16 15:48:53|36.82     |7          |1       |0
|154132554154289|EIRE          |2023-04-16 15:48:53|17.55     |25         |1       |0
|154132554154290|United Kingdom|2023-04-16 15:48:58|21.75     |7          |1       |0
|154132554154291|United Kingdom|2023-04-16 15:49:04|19.05     |8          |1       |0
|154132554154292|United Kingdom|2023-04-16 15:49:07|65.52     |10         |1       |0
|154132554154293|United Kingdom|2023-04-16 15:49:12|118.71    |80         |1       |0
|154132554154294|United Kingdom|2023-04-16 15:49:16|153.66    |34         |1       |0
|154132554154295|United Kingdom|2023-04-16 15:49:17|71.8      |20         |1       |0
|154132554154296|United Kingdom|2023-04-16 15:49:18|21.58     |12         |1       |0
|154132554154297|United Kingdom|2023-04-16 15:49:21|56.48     |44         |1       |0
|154132554154298|United Kingdom|2023-04-16 15:49:27|50.32     |16         |1       |0
|154132554154299|United Kingdom|2023-04-16 15:49:40|49.84     |4          |1       |0
|154132554154300|United Kingdom|2023-04-16 15:49:44|73.32     |59         |1       |0
+--------------+--------------+------------------+----------+-----------+--------+-----


-------------------------------------------
Batch: 10
-------------------------------------------
+--------------+--------------+------------------+----------+-----------+--------+-----
|invoice_no    |country       |timestamp         |total_cost|total_items|is_order|is_ret
+--------------+--------------+------------------+----------+-----------+--------+-----
|154132554154301|United Kingdom|2023-04-16 15:49:53|214.37    |45         |1       |0
|154132554154302|Denmark       |2023-04-16 15:50:02|74.48     |101        |1       |0
|154132554154303|United Kingdom|2023-04-16 15:50:07|6.2999997 |15         |1       |0
|154132554154304|United Kingdom|2023-04-16 15:50:08|17.0      |20         |1       |0
|154132554154305|United Kingdom|2023-04-16 15:50:15|0.83      |1          |1       |0
|154132554154306|United Kingdom|2023-04-16 15:50:18|34.95     |30         |1       |0
|154132554154307|United Kingdom|2023-04-16 15:50:24|-6.76     |4          |0       |1
|154132554154308|United Kingdom|2023-04-16 15:50:25|23.5      |20         |1       |0
|154132554154309|United Kingdom|2023-04-16 15:50:30|14.78     |10         |1       |0
|154132554154310|United Kingdom|2023-04-16 15:50:43|72.409996 |36         |1       |0
+--------------+--------------+------------------+----------+-----------+--------+-----

[hadoop@ip-172-31-86-58 ~]$
```

```
login as: hadoop
Authenticating with public key "imported-openssh-key"


      __|  __|_  )
      _|  (     /   Amazon Linux 2 AMI
     ___|\___|___|

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


EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRRR
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:::RRRRR::::R
  E:::::::::::::::E    M::::::M  M:::M:::M  M::::::M   R::::::::::::RR
  E::::::EEEEEEEEEE    M::::::M   M:::::M   M::::::M   R:::RRRRR::::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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-86-58 ~]$ export SPARK_KAFKA_VERSION=0.10
[hadoop@ip-172-31-86-58 ~]$ vi spark-streaming.py
[hadoop@ip-172-31-86-58 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10
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/c
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-4a933d7e-7b33-49a8-864d-
        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
        [SUCCESSFUL ] org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5!spark-sql-kafka-0-
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/2.0.0/kafka-cli
        [SUCCESSFUL ] org.apache.kafka#kafka-clients;2.0.0!kafka-clients.jar (70ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.
        [SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (2ms)
```

```
23/04/16 15:41:12 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-86-
-172-31-86-58.ec2.internal, 45959, None)
23/04/16 15:41:12 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver,
23/04/16 15:41:12 INFO BlockManager: external shuffle service port = 7337
23/04/16 15:41:12 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-
23/04/16 15:41:12 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster regi
23/04/16 15:41:12 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.a
23/04/16 15:41:13 INFO EventLoggingListener: Logging events to hdfs:/var/log/spark/apps/a
23/04/16 15:41:13 INFO Utils: Using initial executors = 50, max of spark.dynamicAllocatio
xecutor.instances
23/04/16 15:41:13 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduli
23/04/16 15:41:14 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/h
23/04/16 15:41:14 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the
23/04/16 15:41:14 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
23/04/16 15:41:14 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.a
23/04/16 15:41:14 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.a
23/04/16 15:41:14 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.a
23/04/16 15:41:14 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.a
23/04/16 15:41:14 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.a
23/04/16 15:41:15 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
^C[hadoop@ip-172-31-86-58 ~]$ hadoop fs -get time_*
[hadoop@ip-172-31-86-58 ~]$ ls
Console-output.txt  spark-streaming.py  time_country_KPI  time_KPI
[hadoop@ip-172-31-86-58 ~]$ vi script.py
```

```
  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
EEEEEEEEEEEEEEEEEEEE MMMMMMM                  MMMMMMM RRRRRRR       RRRRRR

[hadoop@ip-172-31-86-58 ~]$ hadoop fs -ls time_KPI
Found 12 items
drwxr-xr-x   - hadoop hadoop          0 2023-04-16 15:46 time_KPI/_spark_metadat
a
drwxr-xr-x   - hadoop hadoop          0 2023-04-16 15:41 time_KPI/checkpoint
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:42 time_KPI/part-00000-381
a80fc-f474-45ab-8ac6-672f4b331f7b-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:46 time_KPI/part-00000-3cf
2e262-74eb-4579-a328-ae54ff8d0f9a-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:44 time_KPI/part-00000-4c6
096cf-6024-4ebb-85bc-60a8b464988c-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:41 time_KPI/part-00000-948
19987-d55e-4afe-b1f4-2848a2336cad-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:47 time_KPI/part-00000-c70
d3a6a-b95c-4509-81da-133e92f1e2f2-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:45 time_KPI/part-00000-dc1
ba828-1962-447a-8d83-37f633ae9bb8-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:43 time_KPI/part-00000-fbb
7fd1f-3dd4-4619-b428-af14f0971788-c000.json
-rw-r--r--   1 hadoop hadoop        210 2023-04-16 15:47 time_KPI/part-00029-e0a
51611-f73f-4714-abd9-395006d7bc02-c000.json
-rw-r--r--   1 hadoop hadoop        194 2023-04-16 15:46 time_KPI/part-00090-27e
886c3-f8b6-469f-a63b-2ab97c6cd7fc-c000.json
-rw-r--r--   1 hadoop hadoop        209 2023-04-16 15:45 time_KPI/part-00186-870
b299e-953e-4978-8cf4-3330efbc65f8-c000.json
[hadoop@ip-172-31-86-58 ~]$ ^C
[hadoop@ip-172-31-86-58 ~]$ hadoop fs -ls time_KPI/part-00186-870
-8cf4-3330efbc65f8-c000.json
ls: `time_KPI/part-00186-870': No such file or directory
ls: `b299e-953e-4978-8cf4-3330efbc65f8-c000.json': No such file or directory
[hadoop@ip-172-31-86-58 ~]$ hadoop fs -cat time_KPI/part-00186-870
8-8cf4-3330efbc65f8-c000.json
cat: `time_KPI/part-00186-870': No such file or directory
cat: `b299e-953e-4978-8cf4-3330efbc65f8-c000.json': No such file or directory
[hadoop@ip-172-31-86-58 ~]$ hadoop fs -cat time_KPI/part-00186-870b299e-953e-4978-8cf4-33:
{"window":{"start":"2023-04-16T15:41:00.000Z","end":"2023-04-16T15:42:00.000Z"},"OPM":9,"
91110695732964,"rate_of_return":0.1111111111111111}
[hadoop@ip-172-31-86-58 ~]$ 
```

```
login as: hadoop
Authenticating with public key "imported-openssh-key"
Last login: Sun Apr 16 15:41:11 2023 from 103.220.17.33

      __|  __|_  )
      _|  (     /   Amazon Linux 2 AMI
     ___|\___|___|

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


EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRRR
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:::RRRRR:::::R
  E::::::::::::::::E    M::::::M  M:::M:::M  M::::::M   R:::::::::::RR
  E:::::EEEEEEEEEE    M::::::M   M:::::M   M::::::M   R:::RRRRR::::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
EEEEEEEEEEEEEEEEEEEE MMMMMM             MMMMMM RRRRRR      RRRRRR


[hadoop@ip-172-31-86-58 ~]$ hadoop fs -ls time_KPI
Found 12 items
drwxr-xr-x   - hadoop hadoop          0 2023-04-16 15:46 time_KPI/_spark_metadat
a
drwxr-xr-x   - hadoop hadoop          0 2023-04-16 15:41 time_KPI/checkpoint
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:42 time_KPI/part-00000-381
a80fc-f474-45ab-8ac6-672f4b331f7b-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:46 time_KPI/part-00000-3cf
2e262-74eb-4579-a328-ae54ff8d0f9a-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:44 time_KPI/part-00000-4c6
096cf-6024-4ebb-85bc-60a8b464988c-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:41 time_KPI/part-00000-948
19987-d55e-4afe-b1f4-2848a2336cad-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:47 time_KPI/part-00000-c70
d3a6a-b95c-4509-81da-133e92f1e2f2-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:45 time_KPI/part-00000-dc1
ba828-1962-447a-8d83-37f633ae9bb8-c000.json
-rw-r--r--   1 hadoop hadoop          0 2023-04-16 15:43 time_KPI/part-00000-fbb
7fd1f-3dd4-4619-b428-af14f0971788-c000.json
-rw-r--r--   1 hadoop hadoop        210 2023-04-16 15:47 time_KPI/part-00029-e0a
```

```
whl (246kB)
    100% |████████████████████████████████| 256kB 4.9MB/s
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
[hadoop@ip-172-31-86-58 ~]$ python script.py
```

{"items": [{"SKU": "23120", "title": "PACK OF 6 SMALL FRUIT STRAWS", "unit_price": 0.42,
ING CASES", "unit_price": 0.83, "quantity": 2}, {"SKU": "21714", "title": "CITRONELLA CAN
, "title": "DOLLY GIRL LUNCH BOX", "unit_price": 1.95, "quantity": 64}, {"SKU": "23176",
88}, {"SKU": "21718", "title": "RED METAL BEACH SPADE ", "unit_price": 1.25, "quantity":
unit_price": 9.13, "quantity": 4}, {"SKU": "22334", "title": "DINOSAUR PARTY BAG + STICKE
: "HEART OF WICKER SMALL", "unit_price": 1.65, "quantity": 12}], "type": "ORDER", "country
023-04-16 16:17:23"}
{"items": [{"SKU": "22853", "title": "CAT BOWL VINTAGE CREAM", "unit_price": 3.25, "quant
, "unit_price": 1.95, "quantity": 1}, {"SKU": "37467", "title": "PIG MUG IN TWO COLOUR DE
": "PLASTERS IN TIN STRONGMAN", "unit_price": 1.65, "quantity": 2}, {"SKU": "22750", "tit
 1}], "type": "ORDER", "country": "United Kingdom", "invoice_no": 154132554154566, "times
{"items": [{"SKU": "20869", "title": "GOLD FABRIC MIRROR", "unit_price": 0.83, "quantity"
_price": 1.63, "quantity": 2}], "type": "ORDER", "country": "United Kingdom", "invoice_no"
{"items": [{"SKU": "84879", "title": "ASSORTED COLOUR BIRD ORNAMENT", "unit_price": 1.45,
ice": 101.92, "quantity": 1}, {"SKU": "20723", "title": "STRAWBERRY CHARLOTTE BAG", "unit
BAG SUKI DESIGN", "unit_price": 2.46, "quantity": 100}, {"SKU": "22383", "title": "LUNCH
210", "title": "SET OF 72 RETROSPOT PAPER DOILIES", "unit_price": 1.45, "quantity": 1},
5, "quantity": 2}], "type": "ORDER", "country": "United Kingdom", "invoice_no": 154132554
{"items": [{"SKU": "22457", "title": "NATURAL SLATE HEART CHALKBOARD ", "unit_price": 2.9
COME IN", "unit_price": 7.08, "quantity": 10}], "type": "ORDER", "country": "United Kingd
1"}
{"items": [{"SKU": "23075", "title": "DOUBLE CERAMIC PARLOUR HOOK", "unit_price": 4.15, "
ce_no": 154132554154570, "timestamp": "2023-04-16 16:17:51"}
{"items": [{"SKU": "23100", "title": "SILVER BELLS TABLE DECORATION", "unit_price": 1.25,
, "unit_price": 12.5, "quantity": 4}, {"SKU": "21156", "title": "RETROSPOT CHILDRENS APRO
SMALL POPCORN HOLDER", "unit_price": 0.85, "quantity": 4}, {"SKU": "47591D", "title": "PI
{"SKU": "23190", "title": "BUNDLE OF 3 SCHOOL EXERCISE BOOKS", "unit_price": 1.65, "quant
it_price": 7.46, "quantity": 1}, {"SKU": "23266", "title": "SET OF 3 WOODEN STOCKING DECO
e": "3 HEARTS HANGING DECORATION RUSTIC", "unit_price": 5.79, "quantity": 8}], "type": "O
timestamp": "2023-04-16 16:17:56"}
{"items": [{"SKU": "84519B", "title": "CARROT CHARLIE+LOLA COASTER SET", "unit_price": 1.
CLIPS", "unit_price": 3.81, "quantity": 1}], "type": "ORDER", "country": "United Kingdom",
{"items": [{"SKU": "21195", "title": "PINK HONEYCOMB PAPER BALL ", "unit_price": 1.69, "
CEBOY", "unit_price": 0.65, "quantity": 3}], "type": "ORDER", "country": "United Kingdom",
{"items": [{"SKU": "84879", "title": "ASSORTED COLOUR BIRD ORNAMENT", "unit_price": 3.12,
, "unit_price": 1.65, "quantity": 20}], "type": "ORDER", "country": "United Kingdom", "inv
{"items": [{"SKU": "22029", "title": "SPACEBOY BIRTHDAY CARD", "unit_price": 0.42, "quant
t_price": 2.95, "quantity": 1}], "type": "ORDER", "country": "Germany", "invoice_no": 154