Command Used in execution:

export SPARK_KAFKA_VERSION=0.10

spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py > "Retail_Console_Output_1"

Project Solution Approach:

1) We imported necessary Libraries :-

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

2) Defined Spark Session :-

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

3) Define the command to read the stream from given kafka topic :

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

4) Defined Schema to read the topic in defined columns:

```
retail_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())
                                                  ])))
```

5) Read the Json details in defined schema :

```
infer_schema_retail=
lines.select(from_json(col("value").cast("string"),retail_schema).alias("retail_data")).select("r
etail_data.*")
```

6) Defined the function to calculate derived columns :

```
#1) Function for Total_cost

def total_cost(type1,items):
    sum=0
    if type1.lower()=='order':
        for i in items:
```

```
            sum=sum+(i['unit_price']*i['quantity'])
        return sum
    elif type1.lower()=='return':
        for j in items:
            sum=sum+(j['unit_price']*j['quantity'])
        return sum*(-1)
#2)Function for total_items
def total_items(items):
    sum=0
    for i in items:
        sum=sum+i["quantity"]
    return sum

#3)Function for is_order and is_return
def is_order(type1):
    if type1.lower()=='order':
        return 1
    elif type1.lower()=='return':
        return 0

def is_return(type1):
    if type1.lower()=='order':
        return 0
    elif type1.lower()=='return':
        return 1
```

7) Converted Function in UDF :

```
udf_total_cost=udf(total_cost,FloatType())
udf_total_items=udf(total_items,IntegerType())
udf_is_order=udf(is_order,IntegerType())
udf_is_return=udf(is_return,IntegerType())
```

8) Added Derived columns in Existing Data Frame infer_schema_retail :

```
new_df_with_derived_columns=infer_schema_retail.withColumn("total_cost",udf_total_cost(infer_sche
ma_retail.type,infer_schema_retail.items))\
        .withColumn("total_items",udf_total_items(infer_schema_retail.items))\
        .withColumn("is_order",udf_is_order(infer_schema_retail.type))\
        .withColumn("is_return",udf_is_return(infer_schema_retail.type))
```

9) Written stream output to console with mentioned columns name and in Final command same was written to file :

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

10) Calculated Time based KPI and written the stream output in Json files :-

```
#calculating Time based KPI

time_based_kpi_df=new_df_with_derived_columns.withWatermark("timestamp","1 minutes") \
        .groupBy(window("timestamp","1 minutes")).agg(sum("total_cost").alias("Total volume of
sales"),count("invoice_no").alias("OPM"),avg("total_cost").alias("Average transaction
size"),avg("is_return").alias("Rate of return")) \
        .select("window.start","window.end","OPM","Total volume of sales","Average transaction
size","Rate of return")

#Writing Stream to Json File

time_based_kpi_df_output = time_based_kpi_df \
        .writeStream \
```

```
        .outputMode("append") \
        .format("json") \
        .option("format","append") \
        .option("truncate", "false") \
        .option("path", "time_based_kpi_1") \
        .option("checkpointLocation", "KPI_T_1") \
        .option("truncate", "False") \
        .trigger(processingTime="1 minute") \
        .start()
```

11) Calculated Time and Country based KPI and written the stream output in Json files :

```
#calculating Time and country based KPI
time_and_country_based_kpi_df=new_df_with_derived_columns.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("total_cost").alias("Average transaction
size"),avg("is_return").alias("Rate of return")) \
        .select("window.start","window.end","country","OPM","Total volume of sales","Average
transaction size","Rate of return")


#Writing Stream to Json File

time_and_country_based_kpi_df_output = time_and_country_based_kpi_df \
        .writeStream \
        .outputMode("append") \
        .format("json") \
        .option("format","append") \
        .option("truncate", "false") \
        .option("path", "time_and_country_based_kpi_1") \
        .option("checkpointLocation", "KPI_T_C_1") \
        .option("truncate", "False") \
```

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

12) Stream waited for termination :-

```
time_and_country_based_kpi_df_output.awaitTermination()
```

To Get the Output from HDFS to Local File system :-

Haoop fs -ls time_based_kpi_1

Hadoop fs –ls time_and_country_based_kpi_1

Hadoop fs -get time_based_kpi_1

Hadoop fs –get time_and_country_based_kpi_1

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|
+--------------+--------------+-------------------+----------+-----------+--------+---------+
|154132550301156|United Kingdom|2022-07-13 05:46:47|10.200001 |12         |1       |0        |
|154132550301157|United Kingdom|2022-07-13 05:46:49|44.14     |14         |1       |0        |
|154132550301158|United Kingdom|2022-07-13 05:46:50|27.599998 |10         |1       |0        |
|154132550301159|United Kingdom|2022-07-13 05:46:51|1.45      |1          |1       |0        |
|154132550301160|United Kingdom|2022-07-13 05:46:52|17.39     |10         |1       |0        |
|154132550301161|United Kingdom|2022-07-13 05:46:55|15.0      |12         |1       |0        |
|154132550301162|United Kingdom|2022-07-13 05:46:58|25.58     |4          |1       |0        |
|154132550301163|United Kingdom|2022-07-13 05:46:59|16.4      |7          |1       |0        |
|154132550301164|Spain         |2022-07-13 05:47:01|70.45     |49         |1       |0        |
|154132550301165|United Kingdom|2022-07-13 05:47:09|20.85     |6          |1       |0        |
|154132550301166|United Kingdom|2022-07-13 05:47:10|118.65    |27         |1       |0        |
|154132550301167|United Kingdom|2022-07-13 05:47:12|28.01     |11         |1       |0        |
|154132550301168|United Kingdom|2022-07-13 05:47:15|46.699997 |10         |1       |0        |
|154132550301169|United Kingdom|2022-07-13 05:47:18|26.83     |32         |1       |0        |
|154132550301170|United Kingdom|2022-07-13 05:47:29|219.3     |258        |1       |0        |
|154132550301171|United Kingdom|2022-07-13 05:47:39|40.55     |241        |1       |0        |
+--------------+--------------+-------------------+----------+-----------+--------+---------+

-------------------------------------------
Batch: 2
-------------------------------------------
+--------------+--------------+-------------------+----------+-----------+--------+---------+
|invoice_no    |country       |timestamp          |total_cost|total_items|is_order|is_return|
+--------------+--------------+-------------------+----------+-----------+--------+---------+
```

```
-------------------------------------------
Batch: 3
-------------------------------------------
+--------------+--------------+-------------------+----------+-----------+--------+---------+
|invoice_no    |country       |timestamp          |total_cost|total_items|is_order|is_return|
+--------------+--------------+-------------------+----------+-----------+--------+---------+
|154132550301183|United Kingdom|2022-07-13 05:48:41|135.9     |80         |1       |0        |
|154132550301184|United Kingdom|2022-07-13 05:48:44|70.54     |18         |1       |0        |
|154132550301185|United Kingdom|2022-07-13 05:49:00|5.9       |2          |1       |0        |
|154132550301186|United Kingdom|2022-07-13 05:49:05|-81.99    |51         |0       |1        |
|154132550301187|United Kingdom|2022-07-13 05:49:13|-165.18001|121        |0       |1        |
|154132550301188|United Kingdom|2022-07-13 05:49:13|1.17      |3          |1       |0        |
|154132550301189|United Kingdom|2022-07-13 05:49:14|4.95      |1          |1       |0        |
|154132550301190|United Kingdom|2022-07-13 05:49:15|5.04      |12         |1       |0        |
|154132550301191|United Kingdom|2022-07-13 05:49:25|3.25      |5          |1       |0        |
+--------------+--------------+-------------------+----------+-----------+--------+---------+


-------------------------------------------
Batch: 4
-------------------------------------------
+--------------+--------------+-------------------+----------+-----------+--------+---------+
|invoice_no    |country       |timestamp          |total_cost|total_items|is_order|is_return|
+--------------+--------------+-------------------+----------+-----------+--------+---------+
|154132550301192|United Kingdom|2022-07-13 05:49:42|22.1      |14         |1       |0        |
|154132550301193|United Kingdom|2022-07-13 05:49:45|33.36     |9          |1       |0        |
|154132550301194|United Kingdom|2022-07-13 05:50:00|26.26     |16         |1       |0        |
|154132550301195|United Kingdom|2022-07-13 05:50:01|46.19     |11         |1       |0        |
|154132550301196|United Kingdom|2022-07-13 05:50:03|157.39    |84         |1       |0        |
|154132550301197|United Kingdom|2022-07-13 05:50:06|6.58      |2          |1       |0        |
|154132550301198|United Kingdom|2022-07-13 05:50:10|27.04     |16         |1       |0        |
|154132550301199|United Kingdom|2022-07-13 05:50:12|16.51     |7          |1       |0        |
|154132550301200|United Kingdom|2022-07-13 05:50:29|70.840004 |16         |1       |0        |
|154132550301201|United Kingdom|2022-07-13 05:50:30|0.85      |1          |1       |0        |
|154132550301202|United Kingdom|2022-07-13 05:50:32|66.13     |52         |1       |0        |
|154132550301203|United Kingdom|2022-07-13 05:50:37|20.93     |12         |1       |0        |
+--------------+--------------+-------------------+----------+-----------+--------+---------+


-------------------------------------------
```

```
-------------------------------------------
Batch: 10
-------------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_return|
+---------------+--------------+-------------------+----------+-----------+--------+---------+
|154132550301252|United Kingdom|2022-07-13 05:55:35|169.18999 |11         |1       |0        |
|154132550301253|United Kingdom|2022-07-13 05:55:41|8.84      |4          |1       |0        |
|154132550301254|United Kingdom|2022-07-13 05:55:45|40.56     |24         |1       |0        |
|154132550301255|France        |2022-07-13 05:55:46|27.67     |3          |1       |0        |
|154132550301256|United Kingdom|2022-07-13 05:56:01|4.16      |2          |1       |0        |
|154132550301257|United Kingdom|2022-07-13 05:56:07|99.25     |25         |1       |0        |
|154132550301258|United Kingdom|2022-07-13 05:56:15|10.95     |1          |1       |0        |
|154132550301259|United Kingdom|2022-07-13 05:56:17|54.89     |18         |1       |0        |
|154132550301260|United Kingdom|2022-07-13 05:56:18|153.31    |70         |1       |0        |
|154132550301261|United Kingdom|2022-07-13 05:56:22|10.95     |1          |1       |0        |
|154132550301262|United Kingdom|2022-07-13 05:56:25|-19.43    |7          |0       |1        |
|154132550301263|United Kingdom|2022-07-13 05:56:30|16.5      |10         |1       |0        |
|154132550301264|United Kingdom|2022-07-13 05:56:31|11.53     |25         |1       |0        |
|154132550301265|United Kingdom|2022-07-13 05:56:31|1864.2    |156        |1       |0        |
|154132550301266|United Kingdom|2022-07-13 05:56:31|171.42    |59         |1       |0        |
+---------------+--------------+-------------------+----------+-----------+--------+---------+


-------------------------------------------
Batch: 11
-------------------------------------------
+---------------+--------------+-------------------+----------+-----------+--------+---------+
|invoice_no     |country       |timestamp          |total_cost|total_items|is_order|is_return|
+---------------+--------------+-------------------+----------+-----------+--------+---------+
|154132550301267|United Kingdom|2022-07-13 05:56:32|26.4      |16         |1       |0        |
|154132550301268|United Kingdom|2022-07-13 05:56:37|24.3      |21         |1       |0        |
|154132550301269|United Kingdom|2022-07-13 05:56:43|10.41     |2          |1       |0        |
|154132550301270|United Kingdom|2022-07-13 05:56:49|48.37     |38         |1       |0        |
|154132550301271|United Kingdom|2022-07-13 05:56:51|1.45      |1          |1       |0        |
|154132550301272|United Kingdom|2022-07-13 05:56:57|14.4      |7          |1       |0        |
|154132550301273|France        |2022-07-13 05:56:59|-173.23   |49         |0       |1        |
|154132550301274|United Kingdom|2022-07-13 05:56:59|37.75     |21         |1       |0        |
|154132550301275|United Kingdom|2022-07-13 05:57:04|265.4     |88         |1       |0        |
```

```
Traceback (most recent call last):
  File "/home/hadoop/scripts/spark-streaming.py", line 146, in <module>
    time_and_country_based_kpi_df_output.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
  File "/usr/lib64/python3.7/socket.py", line 589, in readinto
    return self._sock.recv_into(b)
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/context.py", line 278, in signal_handler
KeyboardInterrupt
```