

Code Logic - Retail Data Analysis

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

Step-1: Import the required libraries/modules and set-up PySpark environment

```
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: Initialize SparkSession

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

Notes: The above steps(#1 and #2) are being utilized to include all required libraries followed by initialization of Spark session.

Step-3: Reading input data from Kafka server

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

Notes: Step#3 is being utilized to connect Kafka server by leveraging the topic name provided to read the input data.

Step-4: Define the schema for incoming data

```
define_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", DoubleType()),  
        StructField("quantity", IntegerType())]))))
```

Notes: Step#4 is being utilized to define the schema for incoming data.

Step-5: Create dataframe from the input data

```
order_df = raw_stream_data.select(from_json(col("value").cast("string"),  
define_schema).alias("data")).select("data.*")
```

Step-6: Define user-defined functions(UDF's)

➤ UDF to calculate total_items

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

➤ UDF to calculate order type

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

➤ **UDF to calculate return type**

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

➤ **UDF to calculate total_cost**

```
def total_cost_sum(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
```

➤ **Convert UDF's with utility functions**

```
totalcount = udf(total_items, IntegerType())
isorder = udf(is_order, IntegerType())
isreturn = udf(is_return, IntegerType())
totalcost = udf(total_cost_sum, DoubleType())
```

➤ **Calculating columns(total_cost,total_items,is_order,is_return)**

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

Notes: The above steps(#5 and #6) are being utilized to create dataframe from the input data followed by defining user defined functions(UDF's) to calculate total_cost, total_items, is_order and is_return columns.

Step-7: Write intermediate dataset to the console with one-minute interval

```
output_to_console = order_stream_data \
    .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()
```

Notes: Step#7 helps to write intermediate dataset to the console with one-minute interval as per requirement.

Step-8: Calculate time-based KPIs

```
time_based_KPI = order_stream_data \
    .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")
```

Step-9: Write time based KPI to JSON files

```
time_based_KPI_output_files = time_based_KPI \
    .writeStream \
    .outputMode("Append") \
    .format("json") \
    .option("format", "append") \
    .option("truncate", "false") \
    .option("path", "timebasedKPI/") \
    .option("checkpointLocation", "timebasedKPI/checkpoint/") \
    .option("truncate", "False") \
    .trigger(processingTime="1 minute") \
    .start()
```

Notes: The above steps(#8 and #9) are being leveraged to calculate time based KPIs & write to JSON files.

Step-10: Calculate time and country-based KPIs

```
time_and_country_based_KPI = order_stream_data \  
    .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")
```

Step-11: Write time and country-based KPI to JSON files

```
time_and_country_based_KPI_output = time_and_country_based_KPI \  
    .writeStream \  
    .outputMode("Append") \  
    .format("json") \  
    .option("format", "append") \  
    .option("truncate", "false") \  
    .option("path", "timecountrybasedKPI/") \  
    .option("checkpointLocation", "timecountrybasedKPI/checkpoint/") \  
    .trigger(processingTime="1 minute") \  
    .start()
```

Notes: The above steps(#10 and #11) are being leveraged to calculate time and country based KPIs & write to JSON files.

Step-12: Waiting for termination

```
time_and_country_based_KPI_output.awaitTermination()
```

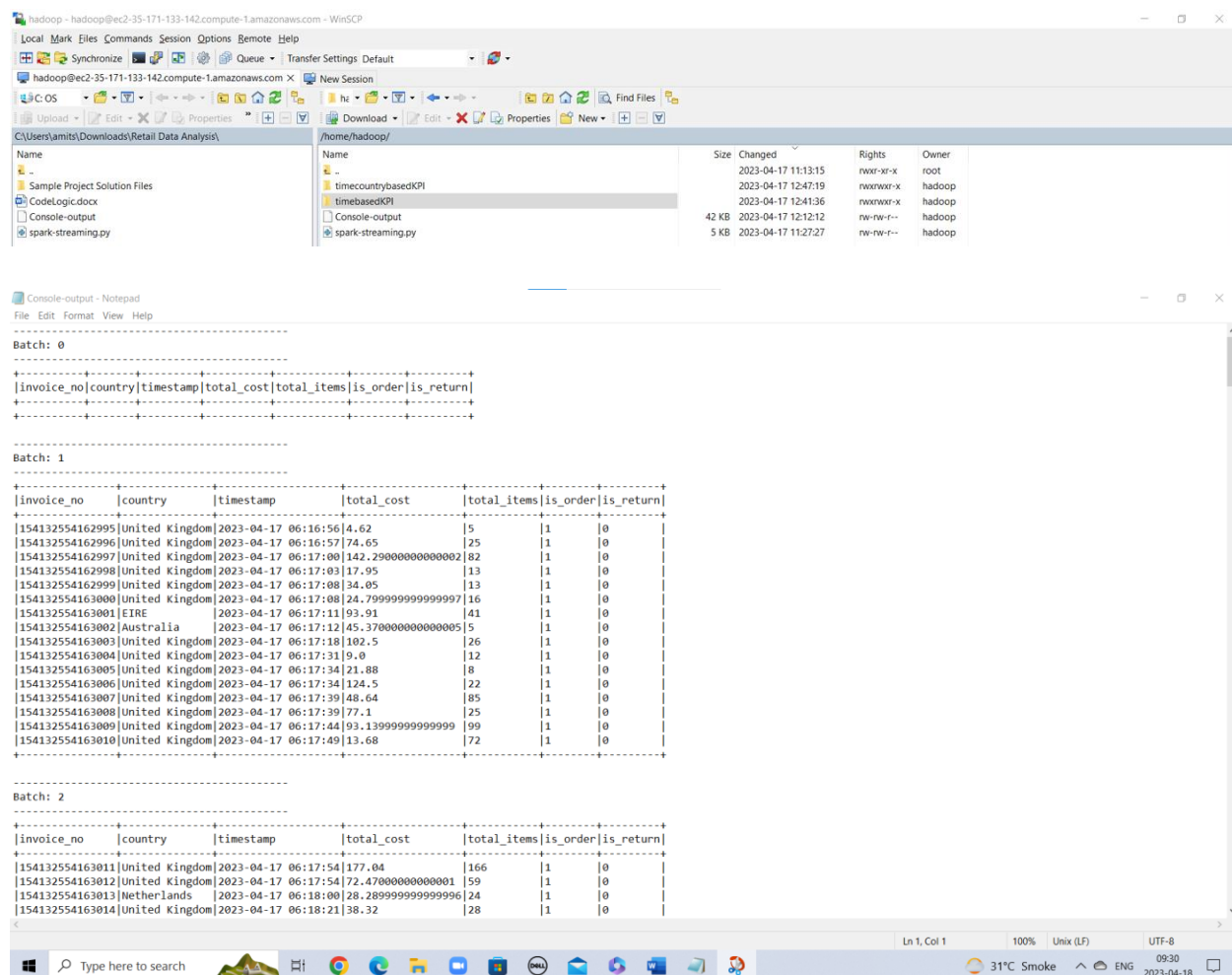
Notes: The above step(#12) waits for the termination signal from the user.

❖ Snapshot of console commands and validate JSON files generated:

1. Spark Submit Command to generate console output

export SPARK_KAFKA_VERSION=0.10 (use this export command to setup the environment before executing the below command)

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



Batch: 0

invoice_no	country	timestamp	total_cost	total_items	is_order	is_return
154132554162995	United Kingdom	2023-04-17 06:16:56	4.62	5	1	0
154132554162996	United Kingdom	2023-04-17 06:16:57	74.65	25	1	0
154132554162997	United Kingdom	2023-04-17 06:17:00	142.29000000000002	82	1	0
154132554162998	United Kingdom	2023-04-17 06:17:03	17.95	13	1	0
154132554162999	United Kingdom	2023-04-17 06:17:08	34.05	13	1	0
154132554163000	United Kingdom	2023-04-17 06:17:08	24.799999999999997	16	1	0
154132554163001	EIRE	2023-04-17 06:17:11	93.91	41	1	0
154132554163002	Australia	2023-04-17 06:17:12	45.370000000000005	5	1	0
154132554163003	United Kingdom	2023-04-17 06:17:18	102.5	26	1	0
154132554163004	United Kingdom	2023-04-17 06:17:31	9.0	12	1	0
154132554163005	United Kingdom	2023-04-17 06:17:34	21.88	8	1	0
154132554163006	United Kingdom	2023-04-17 06:17:34	124.5	22	1	0
154132554163007	United Kingdom	2023-04-17 06:17:39	48.64	85	1	0
154132554163008	United Kingdom	2023-04-17 06:17:39	77.1	25	1	0
154132554163009	United Kingdom	2023-04-17 06:17:44	93.13999999999999	99	1	0
154132554163010	United Kingdom	2023-04-17 06:17:49	13.68	72	1	0

Batch: 1

invoice_no	country	timestamp	total_cost	total_items	is_order	is_return
154132554163011	United Kingdom	2023-04-17 06:17:54	177.04	166	1	0
154132554163012	United Kingdom	2023-04-17 06:17:54	72.47000000000001	59	1	0
154132554163013	Netherlands	2023-04-17 06:18:00	28.289999999999996	24	1	0
154132554163014	United Kingdom	2023-04-17 06:18:21	38.32	28	1	0

Batch: 2

invoice_no	country	timestamp	total_cost	total_items	is_order	is_return
154132554163011	United Kingdom	2023-04-17 06:17:54	177.04	166	1	0
154132554163012	United Kingdom	2023-04-17 06:17:54	72.47000000000001	59	1	0
154132554163013	Netherlands	2023-04-17 06:18:00	28.289999999999996	24	1	0
154132554163014	United Kingdom	2023-04-17 06:18:21	38.32	28	1	0

2. Validate the JSON files generated

hadoop fs -ls

```
hadoop@ip-172-31-2-14:~$ hadoop fs -ls
Found 3 items
drwxr-xr-x - hadoop hdfsadmin group 0 2023-04-17 06:16 .sparkStaging
drwxr-xr-x - hadoop hdfsadmin group 0 2023-04-17 06:34 timebasedKPI
drwxr-xr-x - hadoop hdfsadmin group 0 2023-04-17 06:34 timecountrybasedKPI
```

hadoop fs -ls timebasedKPI

```
hadoop@ip-172-31-2-14:~$ hadoop fs -ls timebasedKPI
Found 67 items
drwxr-xr-x - hadoop hdfsadmin group 0 2023-04-17 06:34 timebasedKPI/spark metadata
drwxr-xr-x - hadoop hdfsadmin group 0 2023-04-17 05:58 timebasedKPI/checkpoint
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:27 timebasedKPI/part-00000-0371b96b-2d11-4fc2-a621-0b4dd83f040c-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:08 timebasedKPI/part-00000-0460cc89-b3f2-4642-93bb-ecae6f7149b3-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:18 timebasedKPI/part-00000-129864b3-7b49-4a45-aec9-d41adcfc4d8-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:09 timebasedKPI/part-00000-1521def0-7bd4-4066-alda-4b490f64351-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:02 timebasedKPI/part-00000-1f6ae782-efef-4b94-a1b6-03549631f983-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:10 timebasedKPI/part-00000-1af7c5d0-6e1a-4a46-b3ad-7bf2015970a4-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:29 timebasedKPI/part-00000-247f64e3-a83f-4248-8491-76a581b64390-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:16 timebasedKPI/part-00000-2567c7cf-b891-4e22-91e5-186ddde21243-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:01 timebasedKPI/part-00000-2ed1c4a4-93f2-4d4b-95b1-f8659b40b9b0-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 05:58 timebasedKPI/part-00000-38b94d5e-e889-49d0-aad7-9f6320bc8b69-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:00 timebasedKPI/part-00000-362a53e5-3026-4b9a-b4a6-32c01542df45-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:19 timebasedKPI/part-00000-4cf1b88d-b37d-45ff-bdd3-cad978d395c6-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:26 timebasedKPI/part-00000-58600a9c-d392-40c8-823d-ea7da166827b-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:17 timebasedKPI/part-00000-5c27e9a1-e05d-4ac8-ba68-e5361aebaf13-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:24 timebasedKPI/part-00000-60302c40-fc4a-41a8-b6de-c24e24cb24a-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:20 timebasedKPI/part-00000-6791d56f-2cac-4bd0-b98a-cc601bd9c34c-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:22 timebasedKPI/part-00000-6b0eb9a6-f757-4575-9b2e-bc589a041a01-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:32 timebasedKPI/part-00000-6c247a36-d52e-4869-b30b-2813ea34e009-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:30 timebasedKPI/part-00000-6ef47fcf-9482-4a6c-86e3-24d1c43871-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:28 timebasedKPI/part-00000-78ce0a46-a058-4101-8af0-38661d53f970-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:33 timebasedKPI/part-00000-7cc103f6-2a2c-4d87-bcc4-701e0254f099-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:07 timebasedKPI/part-00000-8cd7a733-5cdd-4c6b-9926-b60285b3ab1d-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:25 timebasedKPI/part-00000-8ec2d4d4-8de4-4cbb-9f34-b0f618a35653-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:05 timebasedKPI/part-00000-957c8465-603e-4903-bf05-88e19903289e-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:31 timebasedKPI/part-00000-a32e4b91-f70b-48b2-8f6a-4b9d611da46a-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:11 timebasedKPI/part-00000-a41590e9-a097-4593-9fcd-f6ec14c05936-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:34 timebasedKPI/part-00000-b7f70c6d-aad7-4da2-8eac-3c124acd3cc2-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:03 timebasedKPI/part-00000-c2976691-65bb-42f7-aad4-fd5d8ae28e-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:06 timebasedKPI/part-00000-d35103ff-605e-47ac-84d4-5430e6c99189-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:21 timebasedKPI/part-00000-d4687890-29e8-41ce-bc2e-c9a39522e23f-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:26 timebasedKPI/part-00000-e5762f66-707a-42fa-806a-d1f2b43bfdbf-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 0 2023-04-17 06:23 timebasedKPI/part-00000-f66236d4-f236-43a3-9f62-662ba397afc2-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 196 2023-04-17 06:31 timebasedKPI/part-00004-6b49c83e-a19b-4edd-82ee-44f6f1f20fdd-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 182 2023-04-17 06:22 timebasedKPI/part-00018-067b010e-f1b4-4c7a-8f80-bcb1259bcb3fb-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 212 2023-04-17 06:17 timebasedKPI/part-00021-ade4d815-b590-435d-89de-a90f867653ec-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 195 2023-04-17 06:17 timebasedKPI/part-00027-17fd270f-9393-4ed2-b345-65b213da3bb2-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 213 2023-04-17 06:26 timebasedKPI/part-00032-b497d10b-df56-4f49-9f0f-85efaf824d33-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 184 2023-04-17 06:23 timebasedKPI/part-00032-f2ea5a05-fc20-4726-abbf-5907a41361c3-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 195 2023-04-17 06:17 timebasedKPI/part-00035-0a337249-b56c-4eee-a59e-8e3a57b7d28f-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 185 2023-04-17 06:03 timebasedKPI/part-00042-eeda4b29-efcd-45f4-acea-4c1c83178dd0-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 176 2023-04-17 06:27 timebasedKPI/part-00043-07da3424-9128-4034-929e-3471ab6dd3de-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 210 2023-04-17 06:26 timebasedKPI/part-00045-58c4f895-9ae9-40ed-0b10-4d1f75db1eaf-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 212 2023-04-17 06:04 timebasedKPI/part-00045-d12e9627-f825-4c57-928e-aac1ce796c2c-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 173 2023-04-17 06:33 timebasedKPI/part-00054-bfc0fa15-2b3a-4ecc-94d8-6003c9c01b43-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 196 2023-04-17 06:17 timebasedKPI/part-00057-df1323c5-606d-47a5-ad13-7c1cb5885a8-c000.json
-rw-r--r-- 1 hadoop hdfsadmin group 174 2023-04-17 06:21 timebasedKPI/part-00058-369c3b68-cdab-44d0-8f04-3ac07100f908-c000.json
```


hadoop fs -cat timebasedKPI/part-00196-1aac3ace-a19d-4cbd-86e3-e5afc73a92d1-c000.json

```
hadoop@ip-172-31-2-14:~$ hadoop fs -cat timebasedKPI/part-00196-1aac3ace-a19d-4cbd-86e3-e5afc73a92d1-c000.json
{"window":{"start":"2023-04-17T06:06:00.000Z","end":"2023-04-17T06:07:00.000Z"},"OPM":8,"total_sales_volume":421.25000000000006,"average_transaction_size":52.65625000000001,"rate_of_return":0.0}
hadoop@ip-172-31-2-14:~$
```

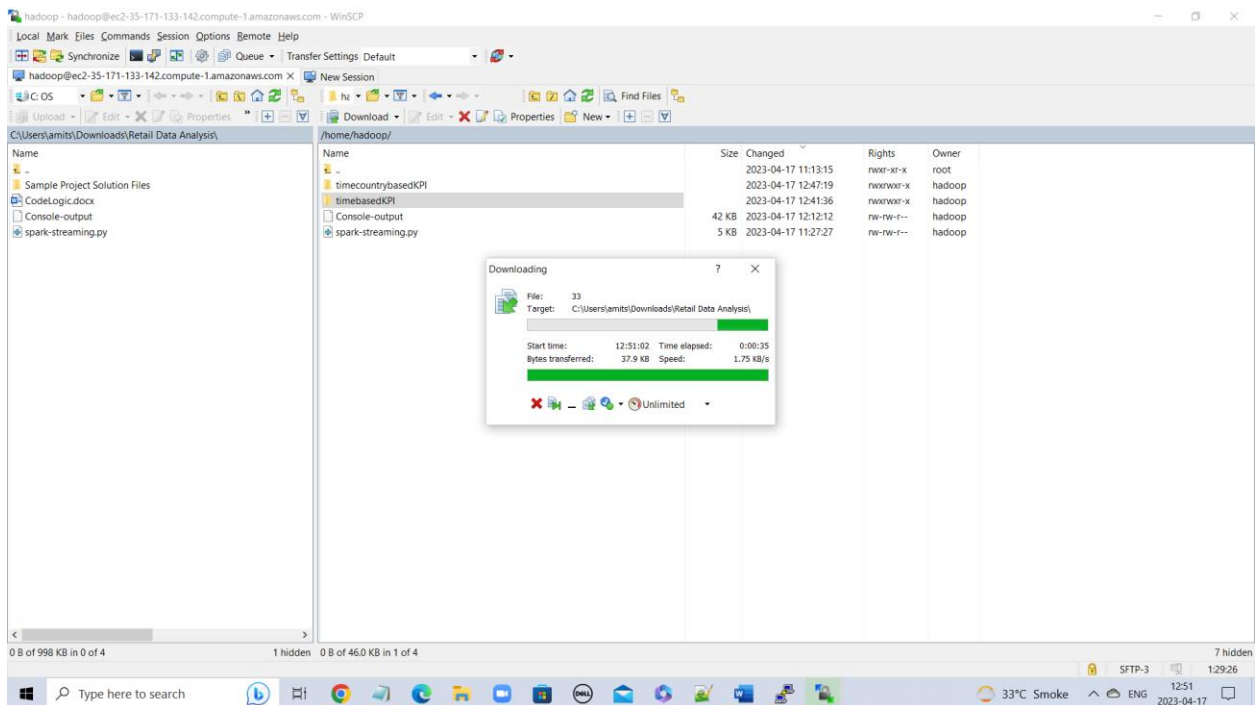
hadoop fs -ls timecountrybasedKPI

```
hadoop@ip-172-31-2-14:~$ hadoop fs -ls timecountrybasedKPI
Found 95 items
drwxr-xr-x - hadoop hdfsadmin/group 0 2023-04-17 06:35 timecountrybasedKPI/ spark_metadata
drwxr-xr-x - hadoop hdfsadmin/group 0 2023-04-17 05:58 timecountrybasedKPI/checkpoint
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:16 timecountrybasedKPI/part-00000-10c326f7-5913-48d7-9f84-8810c5e4f9f5-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:26 timecountrybasedKPI/part-00000-288f335d-6613-42de-966b-e08c410811f3-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:30 timecountrybasedKPI/part-00000-2c8da391-867b-437f-ac22-5b269e90c5be-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:01 timecountrybasedKPI/part-00000-34e180b1-8f90-437e-a155-4c13d316922b-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:25 timecountrybasedKPI/part-00000-36fb9253-b5d2-4c5e-a9a2-bcfcead6878d-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:09 timecountrybasedKPI/part-00000-375db358-b7da-4f01-85a8-a3d0211978f8-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:24 timecountrybasedKPI/part-00000-396e243d-3e26-4de8-900a-19df8a48d83e-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:17 timecountrybasedKPI/part-00000-3a310c0c-59cd-4b2c-9246-24d8fe2031ff-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:34 timecountrybasedKPI/part-00000-3a41e6d0-c773-4403-b754-1ede99593da9-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:05 timecountrybasedKPI/part-00000-404109b9-2f24-40c3-9576-b4fa44be0c4d-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:04 timecountrybasedKPI/part-00000-47ac24a4-0626-4fa2-9cff-53526abfc8b6-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:06 timecountrybasedKPI/part-00000-55d930d-fad7-487a-b3d3-0ae63c5c8b48-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:28 timecountrybasedKPI/part-00000-581111eb-a44b-4253-93b0-a20fd84357f4-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:19 timecountrybasedKPI/part-00000-5e236d80-7da1-4197-8e0a-346d1d2e0e96-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:10 timecountrybasedKPI/part-00000-6535984c-41ab-4714-a0d3-f3f72061a56f-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:29 timecountrybasedKPI/part-00000-7bfc3d65-5b07-400c-9673-a7e93bfa5d9a-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:18 timecountrybasedKPI/part-00000-8d89bbae-afa0-4a10-9356-4638e1bc5ed8-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:07 timecountrybasedKPI/part-00000-945f9178-e873-491f-b467-f6ecf8027a49-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:27 timecountrybasedKPI/part-00000-95d7d0fe-bb27-4b79-9191-0bab0c6068dc-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:03 timecountrybasedKPI/part-00000-a9d3bfe3-2f44-46af-a669-34d4dc986275-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:33 timecountrybasedKPI/part-00000-af395b6f-b375-4103-b315-77da185b9d9e-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:10 timecountrybasedKPI/part-00000-ba4ec69f-6b3e-4d41-92d6-251b19e3a44c-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:02 timecountrybasedKPI/part-00000-c2fd959c-a4b9-4633-a2a3-99fdd8c69d7b-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:23 timecountrybasedKPI/part-00000-c5975a23-423e-4f17-a4d5-d61ff4b7b0b1-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 05:58 timecountrybasedKPI/part-00000-cae7afac-446a-44d8-855a-b39ea5206f36-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:35 timecountrybasedKPI/part-00000-ce145dcf-b909-49c3-afe6-524637ed45f0-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:22 timecountrybasedKPI/part-00000-cf3f7276-4dbf-4ecd-885e-0c2c0442f1d4-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:20 timecountrybasedKPI/part-00000-d27e513a-257e-4b30-9606-70ff4f81d8f57-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:11 timecountrybasedKPI/part-00000-da8ddcf-836e-423b-9252-c2abf6a400fe-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:32 timecountrybasedKPI/part-00000-e3a8d027-81ba-4dab-8910-5155a7c784ce-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:08 timecountrybasedKPI/part-00000-e42700a4-8de0-48fe-a8b6-7869d65f61d5-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:21 timecountrybasedKPI/part-00000-f6176d29-801f-45fd-8c93-25a75a6d6e53-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 0 2023-04-17 06:31 timecountrybasedKPI/part-00000-f81d6c48-e847-45fa-b244-30c4b21a30c5-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 157 2023-04-17 06:16 timecountrybasedKPI/part-00009-e23bb640-5f7b-4652-be2a-4d222b485bdf-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 156 2023-04-17 06:11 timecountrybasedKPI/part-00012-e5691524-a897-4f24-9747-f2a888cbb67-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 165 2023-04-17 06:16 timecountrybasedKPI/part-00016-e05582f6-1899-4857-a2be-9901973cbb7a-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 164 2023-04-17 06:17 timecountrybasedKPI/part-00019-b56a8faf-7b72-4da1-a83a-9f31868dc000-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 177 2023-04-17 06:31 timecountrybasedKPI/part-00025-f073ef24-ae15-4f05-b207-3db5358594a5-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 177 2023-04-17 06:20 timecountrybasedKPI/part-00027-882a2d0a-c493-4260-a2e2-de7ad1ce0a9c-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 178 2023-04-17 06:29 timecountrybasedKPI/part-00030-e772dd2b-17d5-46ef-926b-bd4cca170abd-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 178 2023-04-17 06:11 timecountrybasedKPI/part-00042-fcb92813-627b-4b7e-9f53-8e29634e4599-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 169 2023-04-17 06:03 timecountrybasedKPI/part-00045-2455521d-7bfb-49c4-a982-9e0805479351-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 178 2023-04-17 06:02 timecountrybasedKPI/part-00046-c08b635a-d314-456e-96f3-53912e16f372-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 154 2023-04-17 06:21 timecountrybasedKPI/part-00048-c80d3d4d-3315-43fa-8105-4b6069135ee8-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 167 2023-04-17 06:35 timecountrybasedKPI/part-00050-b9b1977c-abf7-4c8e-865c-782c95699370-c000.json
-rw-r--r-- 1 hadoop hdfsadmin/group 156 2023-04-17 06:29 timecountrybasedKPI/part-00051-847e1974-1526-4d55-fdbb-0a4c8a9c0e04-c000.json
```


hadoop fs -cat timecountrybasedKPI/part-00199-cd734dc9-4e37-4d63-9edc-43bc6b82f35a-c000.json

```
hadoop@ip-172-31-2-14~$ hadoop fs -cat timecountrybasedKPI/part-00199-cd734dc9-4e37-4d63-9edc-43bc6b82f35a-c000.json
{"window":{"start":"2023-04-17T06:00:00.000Z","end":"2023-04-17T06:01:00.000Z"},"country":"Switzerland","OPM":1,"total_sales_volume":16.5,"rate_of_return":0.0}
```

3. Transfer of files generated from HDFS to Local system



The screenshot shows the WinSCP interface with the local file system on the left and the remote HDFS file system on the right. A 'Downloading' dialog box is open, showing the progress of downloading file 33. The dialog box includes a progress bar, start time, bytes transferred, and speed.

Name	Size	Changed	Rights	Owner
timecountrybasedKPI		2023-04-17 11:13:15	rw-rw-r--	root
timebasedKPI		2023-04-17 12:47:19	rw-rw-r--	hadoop
Console-output	42 KB	2023-04-17 12:12:12	rw-rw-r--	hadoop
spark-streaming.py	5 KB	2023-04-17 11:27:27	rw-rw-r--	hadoop

Downloading dialog box details:

- File: 33
- Target: C:\Users\amits\Downloads\Retail Data Analysis\
- Start time: 12:51:02
- Time elapsed: 0:00:35
- Bytes transferred: 37.9 KB
- Speed: 1.75 KB/s