In [21]:
```
## Author: Yam Jason
```

In [22]:
```
from pyspark.sql import SparkSession

spark = SparkSession\
        .builder\
        .appName("Spark")\
        .getOrCreate()
```

In [23]:
```
sc = spark.sparkContext
sc.addFile("../../de_classes/data_storage/hadoop_file_handler.py")


from hadoop_file_handler import HadoopFileHandler

handler = HadoopFileHandler()
df = handler.read_json('data/predictions/predictions3.json')
```

```
24/09/07 17:47:01 WARN SparkSession: Using an existing Spark session; only runtime S
QL configurations will take effect.
```

In [24]:
```
dataSchema = df.schema
dataSchema
```

Out[24]:
```
StructType([StructField('Review', StringType(), True), StructField('prediction', D
oubleType(), True)])
```

Create the Data Stream

In [25]:
```
streaming = spark.readStream.schema(dataSchema).option("maxFilesPerTrigger", 1)\
  .json("data/predictions/predictions3.json")
```

In [26]:
```
activityCounts = streaming.groupBy(
    "Prediction").count()
```

In [27]:
```
# Set the shuffle partitions to a small value
spark.conf.set("spark.sql.shuffle.partitions", 5)
```

In [28]:
```
activityQuery = activityCounts.writeStream.queryName("activity_counts")\
  .format("memory").outputMode("complete")\
  .start()
```

```
24/09/07 17:47:03 WARN ResolveWriteToStream: Temporary checkpoint location created w
hich is deleted normally when the query didn't fail: /tmp/temporary-774fdc71-3587-4e
a5-b03f-0d08e9584def. If it's required to delete it under any circumstances, please
set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know
deleting temp checkpoint folder is best effort.
24/09/07 17:47:03 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not suppo
rted in streaming DataFrames/Datasets and will be disabled.
```

In [29]:
```
from time import sleep
for i in range(5):
```

```
    spark.sql("SELECT * FROM activity_counts").show()
    sleep(1)
```

```
+----------+-----+
|Prediction|count|
+----------+-----+
+----------+-----+

+----------+-----+
|Prediction|count|
+----------+-----+
|       0.0|   10|
|       2.0|   94|
+----------+-----+

+----------+-----+
|Prediction|count|
+----------+-----+
|       0.0|   32|
|       2.0|  237|
+----------+-----+

+----------+-----+
|Prediction|count|
+----------+-----+
|       0.0|   56|
|       2.0|  369|
+----------+-----+

+----------+-----+
|Prediction|count|
+----------+-----+
|       0.0|   77|
|       2.0|  504|
+----------+-----+
```

In [30]:
```
spark.streams.active
```

Out[30]:
```
[<pyspark.sql.streaming.query.StreamingQuery at 0x7f6c17a6da50>]
```

In [31]:
```
activityQuery.stop()
```

In [32]:
```
from pyspark.sql.functions import expr, when

# Filter where the Prediction is 2.0
positiveFilter = streaming.withColumn("Sentiment", when(expr("Prediction = 2.0")
                                                        , "Positive"))\
  .where("Sentiment = 'Positive'")\
  .where("Prediction is not null")\
  .select("Review", "Sentiment")\
  .writeStream\
  .queryName("positive_filter")\
  .format("memory")\
  .outputMode("append")\
  .start()
```

```
# Wait for termination
#positiveFilter.awaitTermination()
```

```
24/09/07 17:47:10 WARN ResolveWriteToStream: Temporary checkpoint location created w
hich is deleted normally when the query didn't fail: /tmp/temporary-6a8c026f-248f-4c
01-bcf6-26956186b7cf. If it's required to delete it under any circumstances, please
set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know
deleting temp checkpoint folder is best effort.
24/09/07 17:47:10 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not suppo
rted in streaming DataFrames/Datasets and will be disabled.
```

In [33]:
```
# Note: this may take a while
for i in range(3):
    spark.sql("SELECT * FROM positive_filter").show()
    sleep(1)
```

```
+------+---------+
|Review|Sentiment|
+------+---------+
+------+---------+
```

```
+--------------------+---------+
|              Review|Sentiment|
+--------------------+---------+
|            all good| Positive|
|good product qual...| Positive|
|fast received pro...| Positive|
|masks are super t...| Positive|
|received package ...| Positive|
|much more comfort...| Positive|
|very fast deliver...| Positive|
|       thanks seller| Positive|
|       fast delivery| Positive|
|received good con...| Positive|
|        good product| Positive|
|ok mask ok delive...| Positive|
|received items go...| Positive|
|reasonable price ...| Positive|
|good use thin com...| Positive|
|        ok good item| Positive|
|repeat order rece...| Positive|
|all items receive...| Positive|
|fast deliver nit ...| Positive|
|i received goods ...| Positive|
+--------------------+---------+
only showing top 20 rows
```

```
+--------------------+---------+
|              Review|Sentiment|
+--------------------+---------+
|            all good| Positive|
|good product qual...| Positive|
|fast received pro...| Positive|
|masks are super t...| Positive|
|received package ...| Positive|
|much more comfort...| Positive|
|very fast deliver...| Positive|
|       thanks seller| Positive|
|       fast delivery| Positive|
|received good con...| Positive|
|        good product| Positive|
|ok mask ok delive...| Positive|
|received items go...| Positive|
|reasonable price ...| Positive|
|good use thin com...| Positive|
|        ok good item| Positive|
|repeat order rece...| Positive|
|all items receive...| Positive|
|fast deliver nit ...| Positive|
|i received goods ...| Positive|
+--------------------+---------+
```

```
only showing top 20 rows
```

In [34]:
```
positiveFilter.stop()
```

In [35]:
```
from pyspark.sql.functions import expr, when

# Filter where the Prediction is 2.0
negativeFilter = streaming.withColumn("Sentiment", when(expr("Prediction = 0.0")
                                                        , "Negative"))\
  .where("Sentiment = 'Negative'")\
  .where("Prediction is not null")\
  .select("Review", "Sentiment")\
  .writeStream\
  .queryName("negative_filter")\
  .format("memory")\
  .outputMode("append")\
  .start()

# Wait for termination
#positiveFilter.awaitTermination()
```

```
24/09/07 17:47:13 WARN ResolveWriteToStream: Temporary checkpoint location created w
hich is deleted normally when the query didn't fail: /tmp/temporary-4afe59e3-cd4c-42
d1-aa10-ca239718d00a. If it's required to delete it under any circumstances, please
set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know
deleting temp checkpoint folder is best effort.
24/09/07 17:47:13 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not suppo
rted in streaming DataFrames/Datasets and will be disabled.
```

In [36]:
```
# Note: this may take a while
for i in range(3):
    spark.sql("SELECT * FROM negative_filter").show()
    sleep(1)
```

```
+--------------------+---------+
|              Review|Sentiment|
+--------------------+---------+
|beautiful but lit...| Negative|
+--------------------+---------+
```

```
+--------------------+---------+
|              Review|Sentiment|
+--------------------+---------+
|beautiful but lit...| Negative|
|send enough stuff...| Negative|
|quality very bad ...| Negative|
| stylish black color| Negative|
|mask ply bukan pl...| Negative|
|       very thin ply| Negative|
|bad won t buy aga...| Negative|
|ordered received ...| Negative|
|don t buy even s ...| Negative|
|saturated waiting...| Negative|
|order light blue ...| Negative|
|i ordered navy bl...| Negative|
|stock received bu...| Negative|
|received good con...| Negative|
|fast delivery wra...| Negative|
|quality entered w...| Negative|
|packaging open wh...| Negative|
|received items go...| Negative|
|received good con...| Negative|
|recieved good con...| Negative|
+--------------------+---------+
only showing top 20 rows
```

```
+--------------------+---------+
|              Review|Sentiment|
+--------------------+---------+
|beautiful but lit...| Negative|
|send enough stuff...| Negative|
|quality very bad ...| Negative|
| stylish black color| Negative|
|mask ply bukan pl...| Negative|
|       very thin ply| Negative|
|bad won t buy aga...| Negative|
|ordered received ...| Negative|
|don t buy even s ...| Negative|
|saturated waiting...| Negative|
|order light blue ...| Negative|
|i ordered navy bl...| Negative|
|stock received bu...| Negative|
|received good con...| Negative|
|fast delivery wra...| Negative|
|quality entered w...| Negative|
|packaging open wh...| Negative|
|received items go...| Negative|
|received good con...| Negative|
|recieved good con...| Negative|
+--------------------+---------+
```

```
only showing top 20 rows
```

In [37]:
```
negativeFilter.stop()
```

In [38]:
```
spark.stop()
```

In [ ]:
```

```