High Level Spark

```
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 3.7 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488490 sha256=f1e0a2c9872a099fb4823a675d75e461ad426f8d0707fa1950fa033c5245e90e
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
[ ] from pyspark.sql import SparkSession

    spark = SparkSession.builder.getOrCreate()

    read_file = spark.read.format("csv") \
                .option("header", "true") \
                .load("/content/drive/MyDrive/Colab Notebooks/BIGData(File_Classroom)/HighLevel-Spark/*.csv")

    read_file.printSchema()
```

```
from pyspark.sql.functions import count

read_file.createOrReplaceTempView("tempTable")

sqlDF1 = spark.sql("SELECT * FROM tempTable WHERE status_type = 'photo'")

result_df = sqlDF1.select(count("status_published").alias("total_published"))

result_df.show()
```

```
+---------------+
|total_published|
+---------------+
|             22|
+---------------+
```

```
[ ] from pyspark.sql.functions import countDistinct

    sqlDF2 = spark.sql("SELECT * FROM tempTable ")

    result_df2 = sqlDF2.select(countDistinct('status_type'))
    result_df2.show()
```

```
+---------------------------+
|count(DISTINCT status_type)|
+---------------------------+
|                          2|
+---------------------------+
```

```
[ ] from pyspark.sql.functions import first , last

    result_df3 = sqlDF2.select(first('status_published'),last('status_published'))
    result_df3.show()
```

```
+-----------------------+----------------------+
|first(status_published)|last(status_published)|
+-----------------------+----------------------+
|         4/22/2018 6:00|        3/23/2018 7:09|
+-----------------------+----------------------+
```

```
[ ] from pyspark.sql.functions import min , max
    from pyspark.sql.types import *

    sqlDF2 = sqlDF2.withColumn('num_reactions', sqlDF2['num_reactions'].cast(IntegerType()))

    result_Max_min = sqlDF2.select(min('num_reactions'),max('num_reactions'))
    result_Max_min.show()
```

```
+------------------+------------------+
|min(num_reactions)|max(num_reactions)|
+------------------+------------------+
|                18|               529|
+------------------+------------------+
```

```
from pyspark.sql.functions import sum , sumDistinct

result_SUM = sqlDF2.select(sum('num_reactions'))
result_SUM.show()
result_sumDistinct = sqlDF2.select(sumDistinct('num_reactions'))
result_sumDistinct.show()
```

```
+------------------+
|sum(num_reactions)|
+------------------+
|              8382|
+------------------+

/usr/local/lib/python3.10/dist-packages/pyspark/sql/functions.py:988: FutureWarning: Deprecated in 3.2, use sum_distinct instead.
  warnings.warn("Deprecated in 3.2, use sum_distinct instead.", FutureWarning)
+--------------------------+
|sum(DISTINCT num_reactions)|
+--------------------------+
|                      5557|
+--------------------------+
```

```
[ ] from pyspark.sql.functions import avg

    result_avg = sqlDF2.select(avg('num_reactions'))
    result_avg.show()
```

```
+------------------+
|avg(num_reactions)|
+------------------+
|220.57894736842104|
+------------------+
```

```
print(sqlDF1)
print(sqlDF2)

join_column = sqlDF1['status_id'] == sqlDF2['status_id']

sqlDF1.join(sqlDF2 , join_column).show()

joinType = 'inner'
sqlDF1.join(sqlDF2,join_column,joinType).show()
```

```
DataFrame[Table: string, status_id: string, status_type: string, status_published: string, num_reactions: string, num_comments: string]
DataFrame[Table: string, status_id: string, status_type: string, status_published: string, num_reactions: int, num_comments: string]
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
|Table|           status_id|status_type|status_published|num_reactions|num_comments|Table|           status_id|status_type|status_published|num_reactions|num_comments|
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|  FB3|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|  FB2|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           6|  FB3|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           6|  FB2|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           6|
|  FB2|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           1|  FB3|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           0|
|  FB2|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           1|  FB2|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           1|
|  FB2|246675545449582_1...|      photo|  3/22/2018 1:25|          152|           2|  FB2|246675545449582_1...|      photo|  3/22/2018 1:25|          152|           2|
|  FB2|246675545449582_1...|      photo|  3/21/2018 8:40|          234|          15|  FB2|246675545449582_1...|      photo|  3/21/2018 8:40|          234|          15|
|  FB2|246675545449582_1...|      photo|  3/21/2018 7:46|          227|           7|  FB2|246675545449582_1...|      photo|  3/21/2018 7:46|          227|           7|
|  FB2|246675545449582_1...|      photo|  3/20/2018 1:54|           98|           0|  FB2|246675545449582_1...|      photo|  3/20/2018 1:54|           98|           0|
|  FB2|246675545449582_1...|      photo|  3/20/2018 0:15|          102|           0|  FB2|246675545449582_1...|      photo|  3/20/2018 0:15|          102|           0|
|  FB2|246675545449582_1...|      photo|  3/12/2018 5:51|          145|           9|  FB2|246675545449582_1...|      photo|  3/12/2018 5:51|          145|           9|
|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
only showing top 20 rows
```


```
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
|Table|           status_id|status_type|status_published|num_reactions|num_comments|Table|           status_id|status_type|status_published|num_reactions|num_comments|
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|  FB3|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|  FB2|246675545449582_1...|      photo|  4/18/2018 3:22|          213|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           6|  FB3|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           0|
|  FB2|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           6|  FB2|246675545449582_1...|      photo|  4/18/2018 2:14|          217|           6|
|  FB2|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           1|  FB3|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           0|
|  FB2|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           1|  FB2|246675545449582_1...|      photo|  4/17/2018 3:33|          203|           1|
|  FB2|246675545449582_1...|      photo|  3/22/2018 1:25|          152|           2|  FB2|246675545449582_1...|      photo|  3/22/2018 1:25|          152|           2|
|  FB2|246675545449582_1...|      photo|  3/21/2018 8:40|          234|          15|  FB2|246675545449582_1...|      photo|  3/21/2018 8:40|          234|          15|
|  FB2|246675545449582_1...|      photo|  3/21/2018 7:46|          227|           7|  FB2|246675545449582_1...|      photo|  3/21/2018 7:46|          227|           7|
|  FB2|246675545449582_1...|      photo|  3/20/2018 1:54|           98|           0|  FB2|246675545449582_1...|      photo|  3/20/2018 1:54|           98|           0|
|  FB2|246675545449582_1...|      photo|  3/20/2018 0:15|          102|           0|  FB2|246675545449582_1...|      photo|  3/20/2018 0:15|          102|           0|
|  FB2|246675545449582_1...|      photo|  3/12/2018 5:51|          145|           9|  FB2|246675545449582_1...|      photo|  3/12/2018 5:51|          145|           9|
|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB3|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|  FB2|246675545449582_1...|      photo| 4/21/2018 22:45|          150|           0|
|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
|  FB3|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|  FB2|246675545449582_1...|      photo|  4/21/2018 2:29|          111|           0|
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
only showing top 20 rows
```


```
joinType2 = 'outer'
sqlDF1.join(sqlDF2,join_column,joinType2).show()
```

```
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
|Table|           status_id|status_type|status_published|num_reactions|num_comments|Table|           status_id|status_type|status_published|num_reactions|num_comments|
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
|  FB2|246675545449582_1...|      photo|  3/12/2018 5:51|          145|           9|  FB2|246675545449582_1...|      photo|  3/12/2018 5:51|          145|           9|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB2|246675545449582_1...|      video|  3/17/2018 7:47|           90|          78|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB2|246675545449582_1...|      video|  3/17/2018 8:07|           75|          36|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB2|246675545449582_1...|      video| 3/19/2018 22:34|           37|           0|
|  FB2|246675545449582_1...|      photo|  3/20/2018 0:15|          102|           0|  FB2|246675545449582_1...|      photo|  3/20/2018 0:15|          102|           0|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB2|246675545449582_1...|      video|  3/20/2018 1:28|           18|           0|
|  FB2|246675545449582_1...|      photo|  3/20/2018 1:54|           98|           0|  FB2|246675545449582_1...|      photo|  3/20/2018 1:54|           98|           0|
|  FB2|246675545449582_1...|      photo|  3/21/2018 7:46|          227|           7|  FB2|246675545449582_1...|      photo|  3/21/2018 7:46|          227|           7|
|  FB2|246675545449582_1...|      photo|  3/21/2018 8:40|          234|          15|  FB2|246675545449582_1...|      photo|  3/21/2018 8:40|          234|          15|
|  FB2|246675545449582_1...|      photo|  3/22/2018 1:25|          152|           2|  FB2|246675545449582_1...|      photo|  3/22/2018 1:25|          152|           2|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB3|246675545449582_1...|      video|  3/23/2018 7:09|          221|          36|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB3|246675545449582_1...|      video|  3/26/2018 8:28|          150|          47|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB3|246675545449582_1...|      video|  3/30/2018 8:28|          135|          79|
| NULL|                NULL|       NULL|            NULL|         NULL|        NULL|  FB3|246675545449582_1...|      video|   4/1/2018 5:16|          332|          30|
|  FB3|246675545449582_1...|      photo|   4/5/2018 9:23|          346|           0|  FB3|246675545449582_1...|      photo|   4/5/2018 9:23|          346|           0|
|  FB3|246675545449582_1...|      photo|   4/8/2018 2:23|          209|           0|  FB3|246675545449582_1...|      photo|   4/8/2018 2:23|          209|           0|
|  FB3|246675545449582_1...|      photo|   4/8/2018 5:10|          313|           2|  FB3|246675545449582_1...|      photo|   4/8/2018 5:10|          313|           2|
|  FB3|246675545449582_1...|      photo|   4/9/2018 2:06|          222|           0|  FB3|246675545449582_1...|      photo|   4/9/2018 2:06|          222|           0|
|  FB3|246675545449582_1...|      photo|  4/10/2018 1:01|          210|           3|  FB3|246675545449582_1...|      photo|  4/10/2018 1:01|          210|           3|
|  FB3|246675545449582_1...|      photo|  4/11/2018 4:53|          170|           1|  FB3|246675545449582_1...|      photo|  4/11/2018 4:53|          170|           1|
+-----+--------------------+-----------+----------------+-------------+------------+-----+--------------------+-----------+----------------+-------------+------------+
only showing top 20 rows
```

```
joinTypeLeft = 'left_outer' # Left outer join
sqlDF1.join(sqlDF2,join_column,joinTypeLeft).show()
joinTypeRight = 'Right_outer' # Right outer join
sqlDF1.join(sqlDF2,join_column,joinTypeRight).show()
```

```
+-----+--------------------+----------+----------------+------------+------------+-----+--------------------+----------+----------------+------------+------------+
|Table|           status_id|status_type|status_published|num_reactions|num_comments|Table|           status_id|status_type|status_published|num_reactions|num_comments|
+-----+--------------------+----------+----------------+------------+------------+-----+--------------------+----------+----------------+------------+------------+
|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|  FB3|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|
|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|
|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|  FB3|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|
|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|
|  FB2|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|  FB3|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|
|  FB2|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|  FB2|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|
|  FB2|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           6|  FB3|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           0|
|  FB2|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           6|  FB2|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           6|
|  FB2|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           1|  FB3|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           0|
|  FB2|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           1|  FB2|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           1|
|  FB2|246675545449582_1...|     photo|  3/22/2018 1:25|         152|           2|  FB2|246675545449582_1...|     photo|  3/22/2018 1:25|         152|           2|
|  FB2|246675545449582_1...|     photo|  3/21/2018 8:40|         234|          15|  FB2|246675545449582_1...|     photo|  3/21/2018 8:40|         234|          15|
|  FB2|246675545449582_1...|     photo|  3/21/2018 7:46|         227|           7|  FB2|246675545449582_1...|     photo|  3/21/2018 7:46|         227|           7|
|  FB2|246675545449582_1...|     photo|  3/20/2018 1:54|          98|           0|  FB2|246675545449582_1...|     photo|  3/20/2018 1:54|          98|           0|
|  FB2|246675545449582_1...|     photo|  3/20/2018 0:15|         102|           0|  FB2|246675545449582_1...|     photo|  3/20/2018 0:15|         102|           0|
|  FB2|246675545449582_1...|     photo|  3/12/2018 5:51|         145|           9|  FB2|246675545449582_1...|     photo|  3/12/2018 5:51|         145|           9|
|  FB3|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|  FB3|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|
|  FB3|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|
|  FB3|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|  FB3|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|
|  FB3|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|
+-----+--------------------+----------+----------------+------------+------------+-----+--------------------+----------+----------------+------------+------------+
only showing top 20 rows
```

```
+-----+--------------------+----------+----------------+------------+------------+-----+--------------------+----------+----------------+------------+------------+
|Table|           status_id|status_type|status_published|num_reactions|num_comments|Table|           status_id|status_type|status_published|num_reactions|num_comments|
+-----+--------------------+----------+----------------+------------+------------+-----+--------------------+----------+----------------+------------+------------+
| NULL|                NULL|      NULL|            NULL|        NULL|        NULL|  FB2|246675545449582_1...|     video|  4/22/2018 6:00|         529|         512|
|  FB3|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|
|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|  FB2|246675545449582_1...|     photo| 4/21/2018 22:45|         150|           0|
| NULL|                NULL|      NULL|            NULL|        NULL|        NULL|  FB2|246675545449582_1...|     video|  4/21/2018 6:17|         227|         236|
|  FB3|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|
|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|  FB2|246675545449582_1...|     photo|  4/21/2018 2:29|         111|           0|
|  FB3|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|  FB2|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|
|  FB2|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|  FB2|246675545449582_1...|     photo|  4/18/2018 3:22|         213|           0|
|  FB3|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           0|  FB2|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           6|
|  FB2|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           6|  FB2|246675545449582_1...|     photo|  4/18/2018 2:14|         217|           6|
| NULL|                NULL|      NULL|            NULL|        NULL|        NULL|  FB2|246675545449582_1...|     video|  4/18/2018 0:24|         503|         614|
| NULL|                NULL|      NULL|            NULL|        NULL|        NULL|  FB2|246675545449582_1...|     video|  4/17/2018 7:42|         295|         453|
|  FB3|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           0|  FB2|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           1|
|  FB2|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           1|  FB2|246675545449582_1...|     photo|  4/17/2018 3:33|         203|           1|
|  FB2|246675545449582_1...|     photo|  3/22/2018 1:25|         152|           2|  FB2|246675545449582_1...|     photo|  3/22/2018 1:25|         152|           2|
|  FB2|246675545449582_1...|     photo|  3/21/2018 8:40|         234|          15|  FB2|246675545449582_1...|     photo|  3/21/2018 8:40|         234|          15|
|  FB2|246675545449582_1...|     photo|  3/21/2018 7:46|         227|           7|  FB2|246675545449582_1...|     photo|  3/21/2018 7:46|         227|           7|
|  FB2|246675545449582_1...|     photo|  3/20/2018 1:54|          98|           0|  FB2|246675545449582_1...|     photo|  3/20/2018 1:54|          98|           0|
| NULL|                NULL|      NULL|            NULL|        NULL|        NULL|  FB2|246675545449582_1...|     video|  3/20/2018 1:28|          18|           0|
|  FB2|246675545449582_1...|     photo|  3/20/2018 0:15|         102|           0|  FB2|246675545449582_1...|     photo|  3/20/2018 0:15|         102|           0|
+-----+--------------------+----------+----------------+------------+------------+-----+--------------------+----------+----------------+------------+------------+
only showing top 20 rows
```

Lowlevel – Spark

```
[ ] !pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 4.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=c4e9f820adcfaf3318686137409f7c8c52e57b93c1f6dd452c50f8aae93eaaef
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

```
[ ] from pyspark.sql import SparkSession

    spark = SparkSession.builder.getOrCreate()
    print(spark)
```

```
<pyspark.sql.session.SparkSession object at 0x7b707b0f3010>
```

```
[ ] from pyspark.sql import SparkSession
    spark = SparkSession.builder.getOrCreate()
    alphabet = ['a', 'b','c', 'd','e', 'f', 'g', 'h']
    rdd = spark.sparkContext.parallelize(alphabet, 4)
    print("Number of partitions: " + str(rdd.getNumPartitions()))
    rdd2 = spark.sparkContext.wholeTextFiles("fb_live_thailand.csv", 5)
    print("Number of partitions: " + str(rdd2.getNumPartitions()))
```

```
Number of partitions: 4
Number of partitions: 1
```

```python
from pyspark.sql import SparkSession


spark  = SparkSession \
    .builder \
    .getOrCreate()

rdd = spark.sparkContext.textFile('fb_live_thailand.csv', 5)
print("Number of partitions: " + str(rdd.getNumPartitions()))

count_distinct = rdd.distinct().count()
print("Number of  distinct records: ", count_distinct)

filter_rdd = rdd.filter(lambda x: x.split(','[1] == 'link'.collect))
print(filter_rdd)

flatmap = rdd.flatMap(lambda x: x.split(','))
pair = flatmap.map(lambda x: (x,1))
take_pair = pair.take(200)
for f in take_pair:
    if str(f[0]) == 'photo' or str(f[0]) == 'video':
        print(str(f[0]), str(f[1]))

sort_data = pair.sortByKey().collect()
for f in sort_data:
    print(str(f[0]), str(f[1]))

sort_data = pair.sortBy(lambda x:x, False, 5).collect()
for f in sort_data:
    print(str(f[0]), str(f[1]))

reduce_key = pair.reduceByKey(lambda x,y: x+y)
print(reduce_key.take(10))

# tranform
alphabet = [('a',1), ('b',2), ('c',3), ('a',1), ('b',2)]
rdd = spark.sparkContext.parallelize(alphabet, 4)
```

```python
#1
def tolist(x):
    return [x]
def append(x, y):
    x.append(y)
    return x
def extend(x, y):
    x.extend(y)
    return x
combine = sorted(rdd.combineByKey(tolist, append, extend).take(10))
print(combine)

#2
zero_val = (0,0)
par_agg = lambda x,y: (x[0] + y, x[1] + 1)
allpar_agg = lambda x,y: (x[0] + y[0], x[1] + y[1])
agg = rdd.aggregateByKey(zero_val, par_agg, allpar_agg).take(10)
print(agg)

#3
from operator import add
fold = sorted(rdd.foldByKey(0, add).collect())
print(fold)

#4
group1 = sorted(rdd.groupByKey().mapValues(len).take(10))
print(group1)
group2 = sorted(rdd.groupByKey().mapValues(list).take(10))
print(group2)
```

```python
# alphabet1
alphabet1 = [('a',1), ('b',2), ('c',3)]
rdd1 = spark.sparkContext.parallelize(alphabet1)

alphabet2 = [('a',1), ('b',2), ('a',1), ('b',2)]
rdd2 = spark.sparkContext.parallelize(alphabet2)
joinRDD = rdd1.join(rdd2).collect()
print(joinRDD)
left = rdd1.leftOuterJoin(rdd2).collect()
print(left)
right = rdd1.rightOuterJoin(rdd2).collect()
print(right)
count = rdd.countByKey()
print(count)
count_val = rdd.countByValue()
print(count_val)
col_asmap = rdd.collectAsMap()
print(col_asmap)
look = rdd.lookup('a')
print(look)
first = rdd.first()
print(first)
max = rdd.max()
print(max)
min = rdd.min()
print(min)
```

Result Lowlevel – Spark

```
0 1
0 1
0 1
0 1
0 1
0 1
0 1
0 1
0 1
[('num_comments', 1), ('246675545449582_1649696485147474', 1), ('4/21/2018 22:45', 1), ('246675545449582_1648730588577397', 1), ('236', 18), ('5', 761), ('4/18/2018 0:24', 1), ('10', 438), ('246675545449582_1644824665634656', 1), (
[('a', [1, 1]), ('b', [2, 2]), ('c', [3])]
[('b', (4, 2)), ('c', (3, 1)), ('a', (2, 2))]
[('a', 2), ('b', 4), ('c', 3)]
[('a', 2), ('b', 2), ('c', 1)]
[('a', [1, 1]), ('b', [2, 2]), ('c', [3])]
[('b', (2, 2)), ('b', (2, 2)), ('a', (1, 1)), ('a', (1, 1))]
[('b', (2, 2)), ('b', (2, 2)), ('c', (3, None)), ('a', (1, 1)), ('a', (1, 1))]
[('b', (2, 2)), ('b', (2, 2)), ('a', (1, 1)), ('a', (1, 1))]
defaultdict(<class 'int'>, {'a': 2, 'b': 2, 'c': 1})
defaultdict(<class 'int'>, {('a', 1): 2, ('b', 2): 2, ('c', 3): 1})
{'a': 1, 'b': 2, 'c': 3}
[1, 1]
('a', 1)
('c', 3)
('a', 1)
```

Streaming DataSources

```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import split, col, current_timestamp
from pyspark.sql.types import StructType, StructField, StringType

spark = SparkSession.builder \
    .appName("WatermarkExample") \
    .getOrCreate()

file_schema = StructType([
    StructField("status_id", StringType(), True),
    StructField("status_type", StringType(), True),
    StructField("status_published", StringType(), True),
    StructField("num_reactions", StringType(), True),
    StructField("num_comments", StringType(), True),
    StructField("num_shares", StringType(), True),
    StructField("num_likes", StringType(), True),
    StructField("num_loves", StringType(), True),
    StructField("num_wows", StringType(), True),
    StructField("num_hahas", StringType(), True),
    StructField("num_sads", StringType(), True),
    StructField("num_angrys", StringType(), True)
])

lines = spark \
    .readStream \
    .format("csv") \
    .option("maxFilesPerTrigger", 1) \
    .option("header", True) \
    .schema(file_schema) \
    .load("File_dataset")

words = lines \
    .withColumn("date", split(col("status_published"), " ").getItem(0)) \
    .withColumn("timestamp", current_timestamp()) \
    .withWatermark("timestamp", "10 seconds")

wordCounts = words.groupBy('date', 'status_type','timestamp').count()
# Save File + path Folder
wordCounts.writeStream \
.format("csv")\
.option("path", "Test")\
.trigger(processingTime='5 seconds') \
.option("checkpointLocation", "Test") \
.outputMode("append") \
.option("truncate", False) \
.start().awaitTermination()

query = words.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()

query.awaitTermination()
```