Custom Learnings

Day 17

PySpark

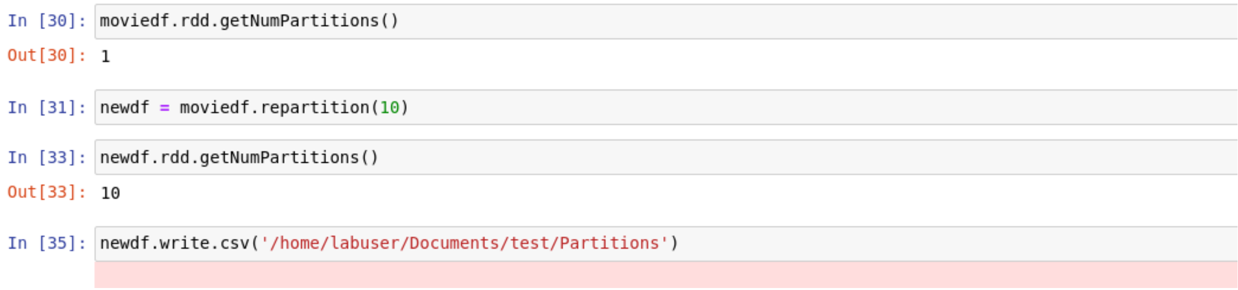
SparkContext is used for RDDs.

SparkSession is used for dataframes.

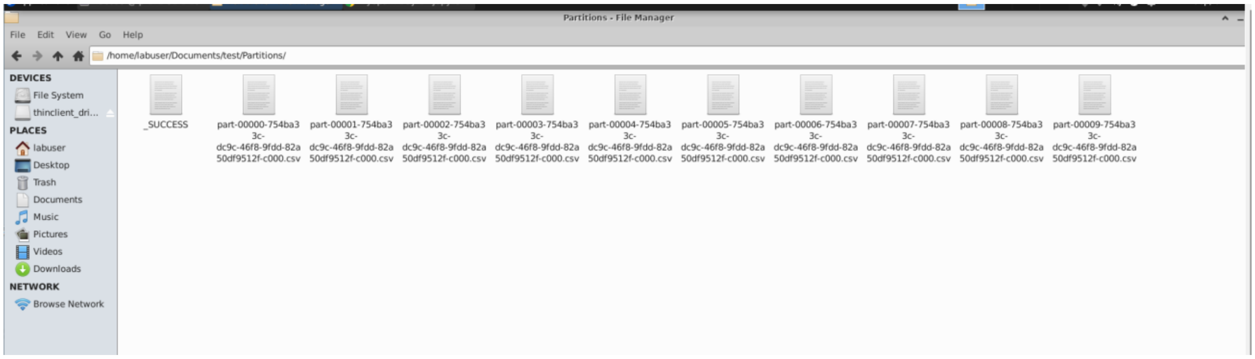
Both are entry points.

Partition size can be configured according to our need. This can be re-configured also because it is runtime process.

Partitioning and Repartitioning:

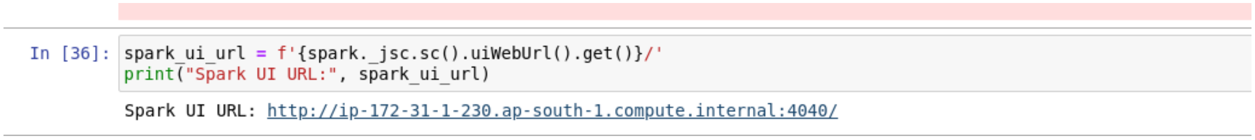


```
In [30]: moviedf.rdd.getNumPartitions()
Out[30]: 1

In [31]: newdf = moviedf.repartition(10)

In [33]: newdf.rdd.getNumPartitions()
Out[33]: 10

In [35]: newdf.write.csv('/home/labuser/Documents/test/Partitions')
```

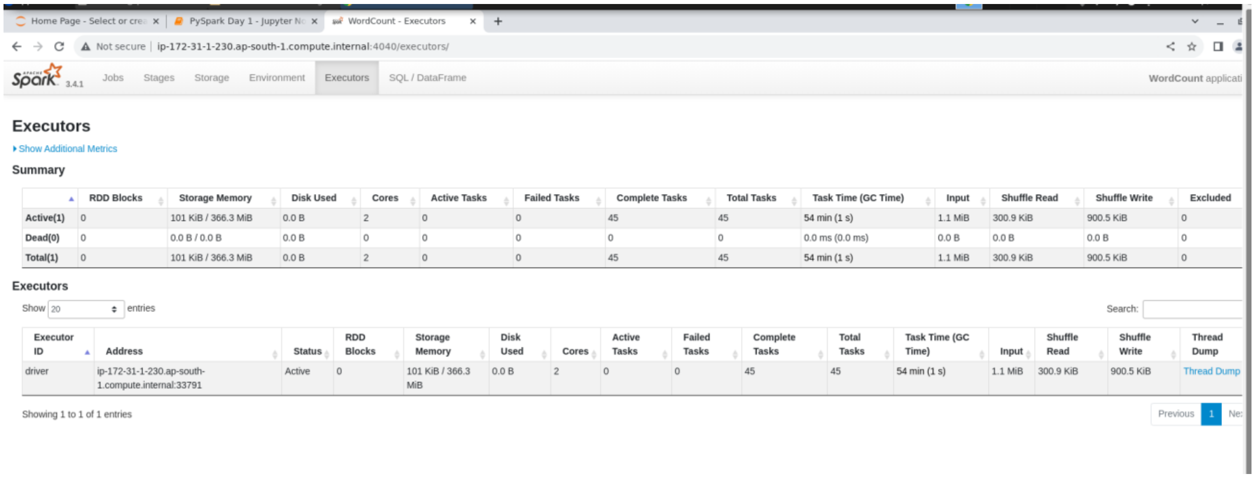


```
/home/labuser/Documents/test/Partitions/

_SUCCESS
part-00000-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00001-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00002-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00003-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00004-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00005-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00006-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00007-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00008-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
part-00009-754ba33c-dc9c-46f8-9fdd-82a50df9512f-c000.csv
```

Spark Web UI:



```
In [36]: spark_ui_url = f'{spark._jsc.sc().uiWebUrl().get()}/'
         print("Spark UI URL:", spark_ui_url)

         Spark UI URL: http://ip-172-31-1-230.ap-south-1.compute.internal:4040/
```


Home Page - Select or crea
PySpark Day 1 - Jupyter No
WordCount - Executors
Not secure | ip-172-31-1-230.ap-south-1.compute.internal:4040/executors/
Spark 3.4.1
Jobs
Stages
Storage
Environment
Executors
SQL / DataFrame
WordCount applicati

Executors

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	101 KiB / 366.3 MiB	0.0 B	2	0	0	45	45	54 min (1 s)	1.1 MiB	300.9 KiB	900.5 KiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	101 KiB / 366.3 MiB	0.0 B	2	0	0	45	45	54 min (1 s)	1.1 MiB	300.9 KiB	900.5 KiB	0

Executors

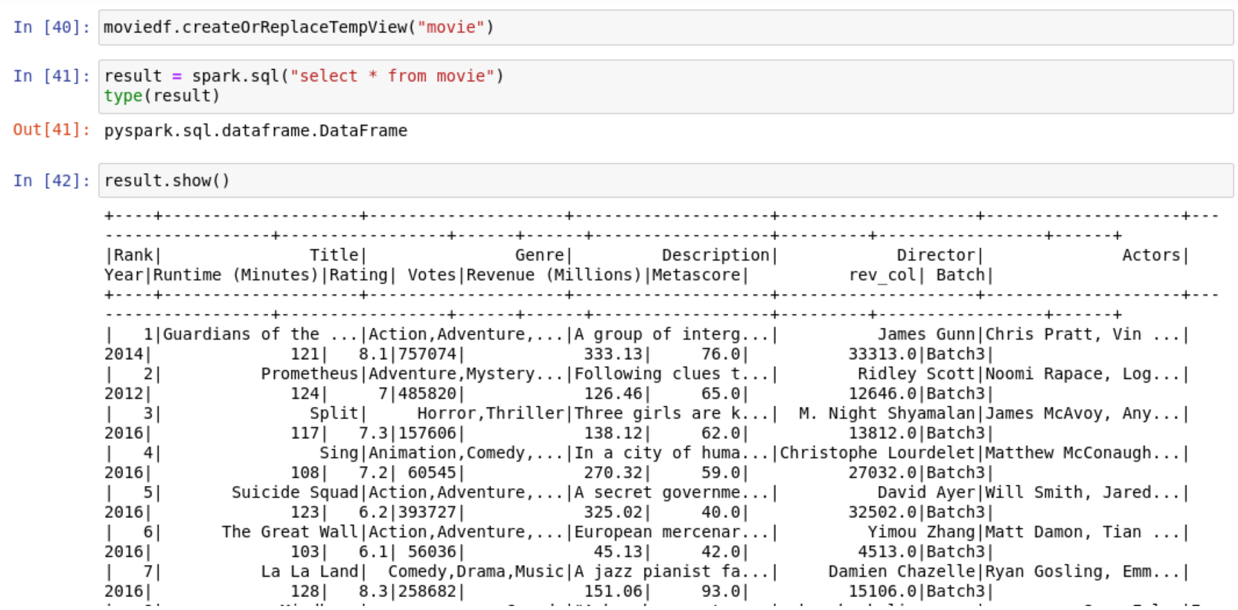
Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	ip-172-31-1-230.ap-south-1.compute.internal:33791	Active	0	101 KiB / 366.3 MiB	0.0 B	2	0	0	45	45	54 min (1 s)	1.1 MiB	300.9 KiB	900.5 KiB	Thread Dump

Showing 1 to 1 of 1 entries

PySpark SQL:



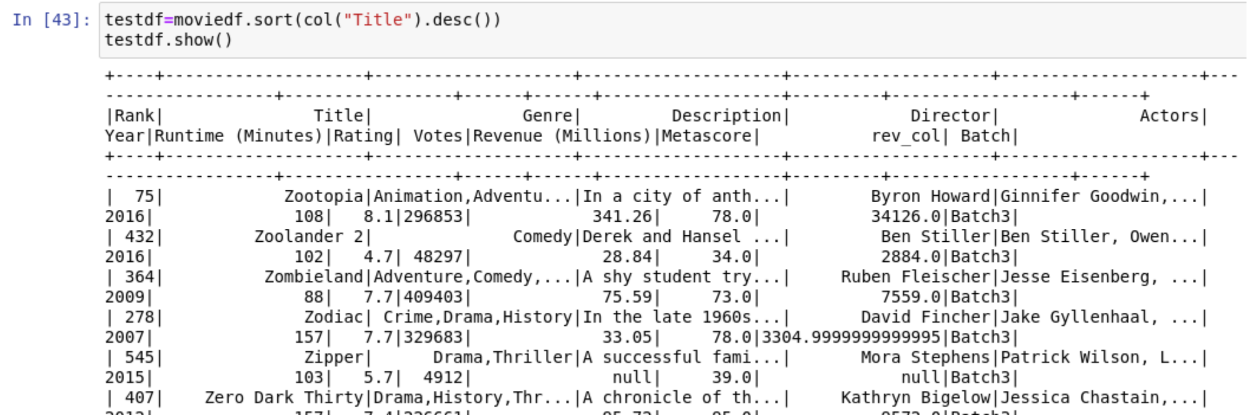
In [40]:
```
moviedf.createOrReplaceTempView("movie")
```

In [41]:
```
result = spark.sql("select * from movie")
type(result)
```

Out[41]:
```
pyspark.sql.dataframe.DataFrame
```

In [42]:
```
result.show()
```

```
+----+--------------------+--------------------+--------------------+--------------------+--------------------+---
-----------------+-----------------+------+------+------------------+---------+----------------+------+
|Rank|               Title|               Genre|         Description|            Director|              Actors|
Year|Runtime (Minutes)|Rating| Votes|Revenue (Millions)|Metascore|         rev_col| Batch|
+----+--------------------+--------------------+--------------------+--------------------+--------------------+---
-----------------+-----------------+------+------+------------------+---------+----------------+------+
|   1|Guardians of the ...|Action,Adventure,...|A group of interg...|          James Gunn|Chris Pratt, Vin ...|
2014|              121|   8.1|757074|            333.13|     76.0|         33313.0|Batch3|
|   2|          Prometheus|Adventure,Mystery...|Following clues t...|        Ridley Scott|Noomi Rapace, Log...|
2012|              124|     7|485820|            126.46|     65.0|         12646.0|Batch3|
|   3|               Split|     Horror,Thriller|Three girls are k...|  M. Night Shyamalan|James McAvoy, Any...|
2016|              117|   7.3|157606|            138.12|     62.0|         13812.0|Batch3|
|   4|                Sing|Animation,Comedy,...|In a city of huma...|Christophe Lourdelet|Matthew McConaugh...|
2016|              108|   7.2| 60545|            270.32|     59.0|         27032.0|Batch3|
|   5|       Suicide Squad|Action,Adventure,...|A secret governme...|          David Ayer|Will Smith, Jared...|
2016|              123|   6.2|393727|            325.02|     40.0|         32502.0|Batch3|
|   6|      The Great Wall|Action,Adventure,...|European mercenar...|         Yimou Zhang|Matt Damon, Tian ...|
2016|              103|   6.1| 56036|             45.13|     42.0|          4513.0|Batch3|
|   7|          La La Land|  Comedy,Drama,Music|A jazz pianist fa...|     Damien Chazelle|Ryan Gosling, Emm...|
2016|              128|   8.3|258682|            151.06|     93.0|         15106.0|Batch3|
```



In [43]:
```
testdf=moviedf.sort(col("Title").desc())
testdf.show()
```

```
+----+--------------------+--------------------+--------------------+--------------------+--------------------+---
-----------------+-----------------+------+------+------------------+---------+------------------+------+
|Rank|               Title|               Genre|         Description|            Director|              Actors|
Year|Runtime (Minutes)|Rating| Votes|Revenue (Millions)|Metascore|           rev_col| Batch|
+----+--------------------+--------------------+--------------------+--------------------+--------------------+---
-----------------+-----------------+------+------+------------------+---------+------------------+------+
|  75|            Zootopia|Animation,Adventu...|In a city of anth...|        Byron Howard|Ginnifer Goodwin,...|
2016|              108|   8.1|296853|            341.26|     78.0|           34126.0|Batch3|
| 432|         Zoolander 2|              Comedy|Derek and Hansel ...|         Ben Stiller|Ben Stiller, Owen...|
2016|              102|   4.7| 48297|             28.84|     34.0|            2884.0|Batch3|
| 364|          Zombieland|Adventure,Comedy,...|A shy student try...|     Ruben Fleischer|Jesse Eisenberg, ...|
2009|               88|   7.7|409403|             75.59|     73.0|            7559.0|Batch3|
| 278|              Zodiac| Crime,Drama,History|In the late 1960s...|       David Fincher|Jake Gyllenhaal, ...|
2007|              157|   7.7|329683|             33.05|     78.0|3304.9999999999995|Batch3|
| 545|              Zipper|      Drama,Thriller|A successful fami...|       Mora Stephens|Patrick Wilson, L...|
2015|              103|   5.7|  4912|              null|     39.0|              null|Batch3|
| 407|    Zero Dark Thirty|Drama,History,Thr...|A chronicle of th...|     Kathryn Bigelow|Jessica Chastain,...|
```

In [45]:

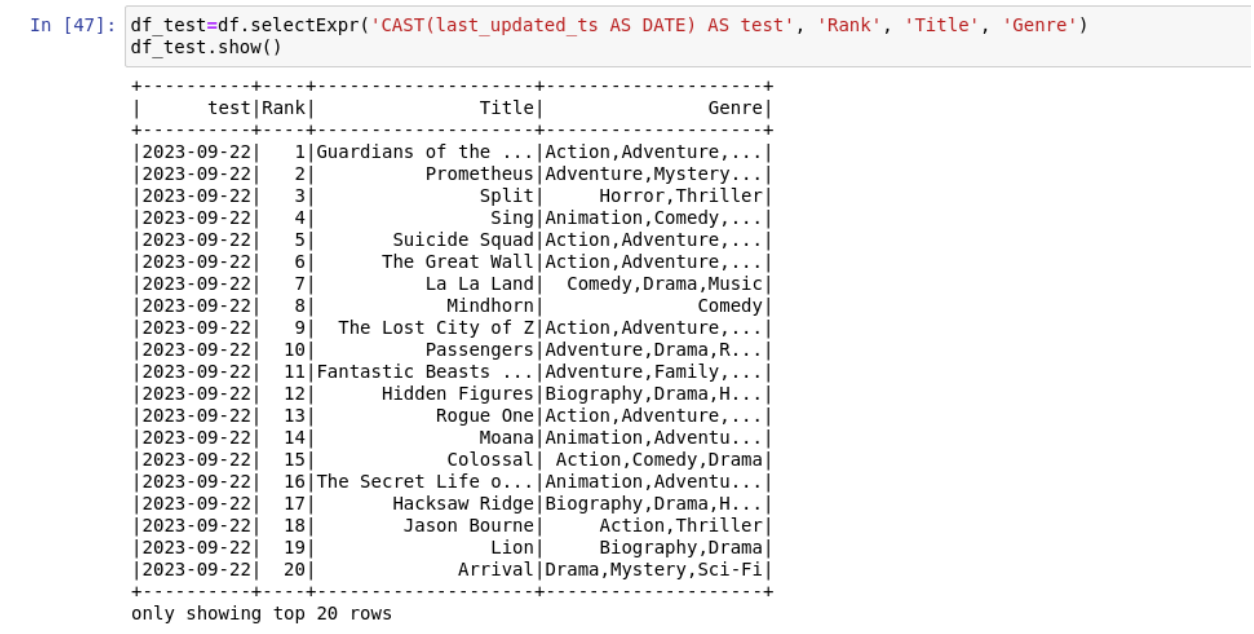
```
df=moviedf.select('Rank','Title','Genre')
df.show()
```

```
+----+--------------------+--------------------+
|Rank|               Title|               Genre|
+----+--------------------+--------------------+
|   1|Guardians of the ...|Action,Adventure,...|
|   2|          Prometheus|Adventure,Mystery...|
|   3|               Split|     Horror,Thriller|
|   4|                Sing|Animation,Comedy,...|
|   5|       Suicide Squad|Action,Adventure,...|
|   6|      The Great Wall|Action,Adventure,...|
|   7|          La La Land|  Comedy,Drama,Music|
|   8|            Mindhorn|              Comedy|
|   9|  The Lost City of Z|Action,Adventure,...|
|  10|          Passengers|Adventure,Drama,R...|
|  11|Fantastic Beasts ...|Adventure,Family,...|
|  12|      Hidden Figures|Biography,Drama,H...|
|  13|           Rogue One|Action,Adventure,...|
|  14|               Moana|Animation,Adventu...|
|  15|            Colossal| Action,Comedy,Drama|
|  16|The Secret Life o...|Animation,Adventu...|
|  17|       Hacksaw Ridge|Biography,Drama,H...|
|  18|        Jason Bourne|     Action,Thriller|
|  19|                Lion|     Biography,Drama|
|  20|             Arrival|Drama,Mystery,Sci-Fi|
+----+--------------------+--------------------+
only showing top 20 rows
```

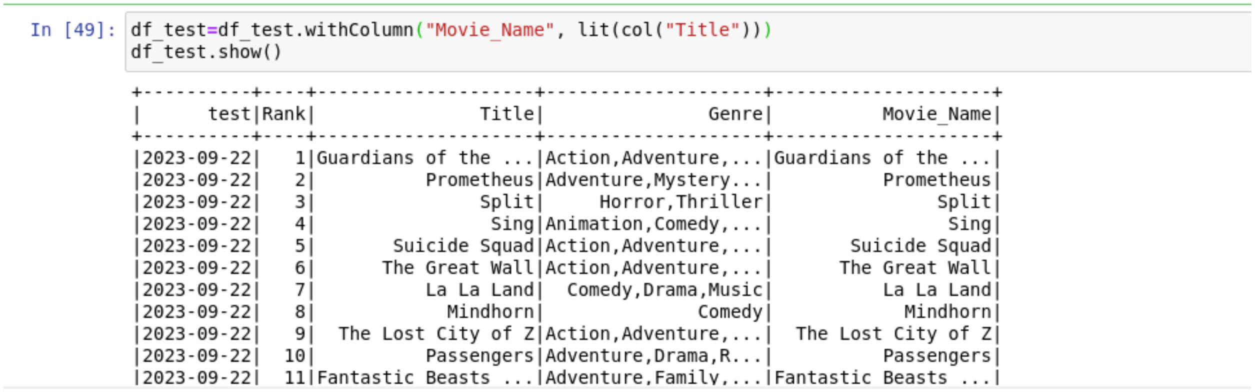
In [46]:

```
from datetime import *
df=df.withColumn("last_updated_ts", lit(datetime.now()))
df.show()
```

```
+----+--------------------+--------------------+--------------------+
|Rank|               Title|               Genre|     last_updated_ts|
+----+--------------------+--------------------+--------------------+
|   1|Guardians of the ...|Action,Adventure,...|2023-09-22 05:32:...|
|   2|          Prometheus|Adventure,Mystery...|2023-09-22 05:32:...|
|   3|               Split|     Horror,Thriller|2023-09-22 05:32:...|
|   4|                Sing|Animation,Comedy,...|2023-09-22 05:32:...|
|   5|       Suicide Squad|Action,Adventure,...|2023-09-22 05:32:...|
|   6|      The Great Wall|Action,Adventure,...|2023-09-22 05:32:...|
|   7|          La La Land|  Comedy,Drama,Music|2023-09-22 05:32:...|
|   8|            Mindhorn|              Comedy|2023-09-22 05:32:...|
|   9|  The Lost City of Z|Action,Adventure,...|2023-09-22 05:32:...|
|  10|          Passengers|Adventure,Drama,R...|2023-09-22 05:32:...|
```



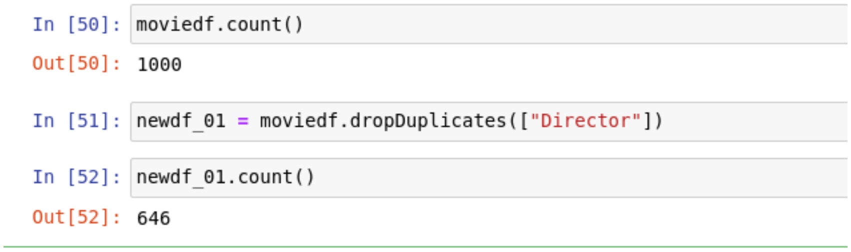
```
In [47]: df_test=df.selectExpr('CAST(last_updated_ts AS DATE) AS test', 'Rank', 'Title', 'Genre')
         df_test.show()

         +----------+----+--------------------+--------------------+
         |      test|Rank|               Title|               Genre|
         +----------+----+--------------------+--------------------+
         |2023-09-22|   1|Guardians of the ...|Action,Adventure,...|
         |2023-09-22|   2|          Prometheus|Adventure,Mystery...|
         |2023-09-22|   3|               Split|     Horror,Thriller|
         |2023-09-22|   4|                Sing|Animation,Comedy,...|
         |2023-09-22|   5|       Suicide Squad|Action,Adventure,...|
         |2023-09-22|   6|      The Great Wall|Action,Adventure,...|
         |2023-09-22|   7|          La La Land|  Comedy,Drama,Music|
         |2023-09-22|   8|            Mindhorn|              Comedy|
         |2023-09-22|   9|  The Lost City of Z|Action,Adventure,...|
         |2023-09-22|  10|          Passengers|Adventure,Drama,R...|
         |2023-09-22|  11|Fantastic Beasts ...|Adventure,Family,...|
         |2023-09-22|  12|      Hidden Figures|Biography,Drama,H...|
         |2023-09-22|  13|           Rogue One|Action,Adventure,...|
         |2023-09-22|  14|               Moana|Animation,Adventu...|
         |2023-09-22|  15|            Colossal| Action,Comedy,Drama|
         |2023-09-22|  16|The Secret Life o...|Animation,Adventu...|
         |2023-09-22|  17|       Hacksaw Ridge|Biography,Drama,H...|
         |2023-09-22|  18|         Jason Bourne|     Action,Thriller|
         |2023-09-22|  19|                Lion|     Biography,Drama|
         |2023-09-22|  20|             Arrival|Drama,Mystery,Sci-Fi|
         +----------+----+--------------------+--------------------+
         only showing top 20 rows
```



```
In [49]: df_test=df_test.withColumn("Movie_Name", lit(col("Title")))
         df_test.show()

         +----------+----+--------------------+--------------------+--------------------+
         |      test|Rank|               Title|               Genre|          Movie_Name|
         +----------+----+--------------------+--------------------+--------------------+
         |2023-09-22|   1|Guardians of the ...|Action,Adventure,...|Guardians of the ...|
         |2023-09-22|   2|          Prometheus|Adventure,Mystery...|          Prometheus|
         |2023-09-22|   3|               Split|     Horror,Thriller|               Split|
         |2023-09-22|   4|                Sing|Animation,Comedy,...|                Sing|
         |2023-09-22|   5|       Suicide Squad|Action,Adventure,...|       Suicide Squad|
         |2023-09-22|   6|      The Great Wall|Action,Adventure,...|      The Great Wall|
         |2023-09-22|   7|          La La Land|  Comedy,Drama,Music|          La La Land|
         |2023-09-22|   8|            Mindhorn|              Comedy|            Mindhorn|
         |2023-09-22|   9|  The Lost City of Z|Action,Adventure,...|  The Lost City of Z|
         |2023-09-22|  10|          Passengers|Adventure,Drama,R...|          Passengers|
         |2023-09-22|  11|Fantastic Beasts ...|Adventure,Family,...|Fantastic Beasts ...|
```

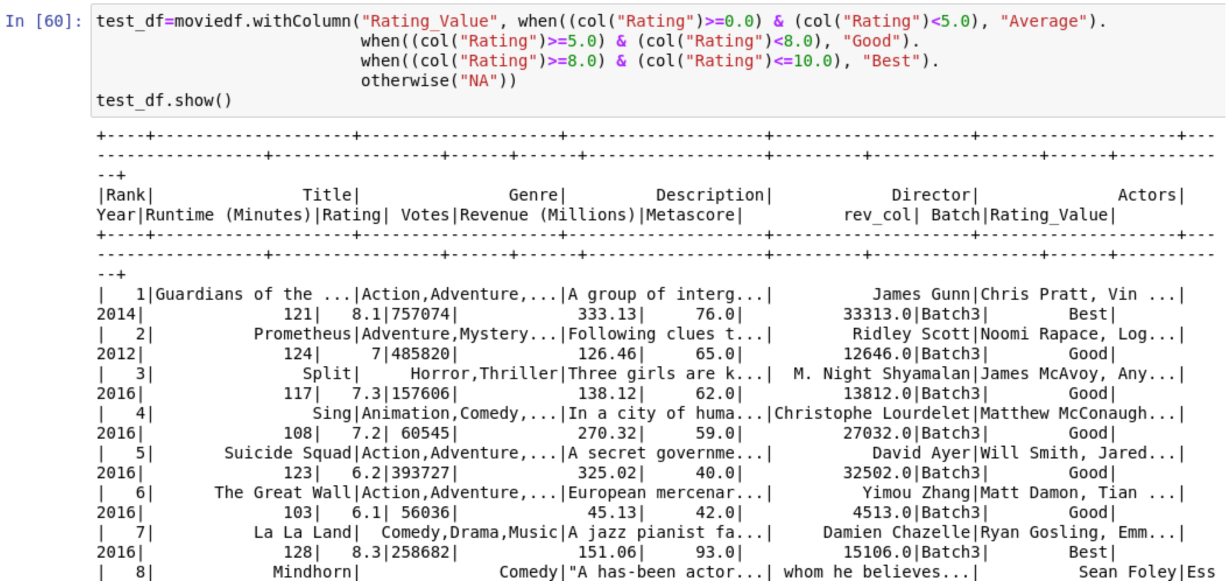
Common Transformations and Actions:

dropDuplicates():



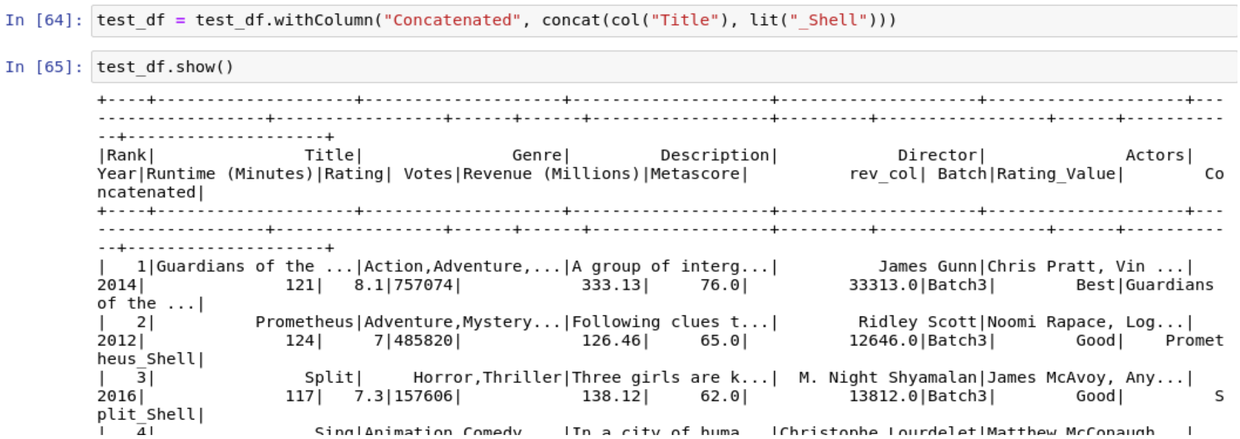
```
In [50]: moviedf.count()
Out[50]: 1000

In [51]: newdf_01 = moviedf.dropDuplicates(["Director"])

In [52]: newdf_01.count()
Out[52]: 646
```

If-else condition:



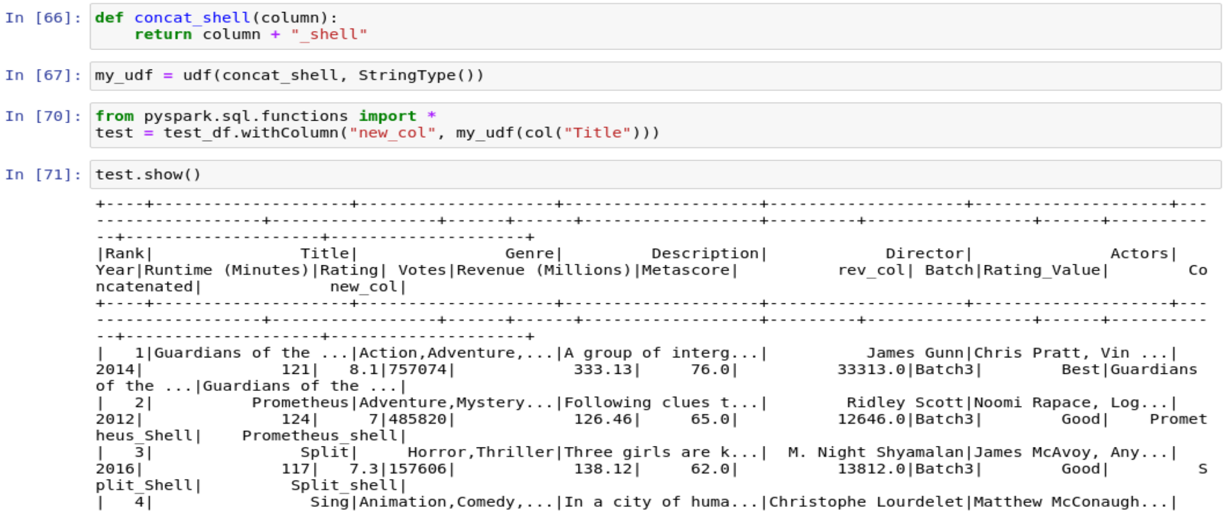
```
In [60]: test_df=moviedf.withColumn("Rating_Value", when((col("Rating")>=0.0) & (col("Rating")<5.0), "Average").
                                   when((col("Rating")>=5.0) & (col("Rating")<8.0), "Good").
                                   when((col("Rating")>=8.0) & (col("Rating")<=10.0), "Best").
                                   otherwise("NA"))
         test_df.show()

         +----+--------------------+--------------------+--------------------+--------------------+--------------------+---
         -----------------+-----------------+------+------+------------------+---------+-----------------+------+----------
         --+
         |Rank|               Title|               Genre|         Description|            Director|              Actors|
         Year|Runtime (Minutes)|Rating| Votes|Revenue (Millions)|Metascore|          rev_col| Batch|Rating_Value|
         +----+--------------------+--------------------+--------------------+--------------------+--------------------+---
         -----------------+-----------------+------+------+------------------+---------+-----------------+------+----------
         --+
         |   1|Guardians of the ...|Action,Adventure,...|A group of interg...|          James Gunn|Chris Pratt, Vin ...|
         2014|              121|   8.1|757074|            333.13|     76.0|          33313.0|Batch3|        Best|
         |   2|          Prometheus|Adventure,Mystery...|Following clues t...|        Ridley Scott|Noomi Rapace, Log...|
         2012|              124|     7|485820|            126.46|     65.0|          12646.0|Batch3|        Good|
         |   3|               Split|     Horror,Thriller|Three girls are k...|  M. Night Shyamalan|James McAvoy, Any...|
         2016|              117|   7.3|157606|            138.12|     62.0|          13812.0|Batch3|        Good|
         |   4|                Sing|Animation,Comedy,...|In a city of huma...|Christophe Lourdelet|Matthew McConaugh...|
         2016|              108|   7.2| 60545|            270.32|     59.0|          27032.0|Batch3|        Good|
         |   5|       Suicide Squad|Action,Adventure,...|A secret governme...|          David Ayer|Will Smith, Jared...|
         2016|              123|   6.2|393727|            325.02|     40.0|          32502.0|Batch3|        Good|
         |   6|      The Great Wall|Action,Adventure,...|European mercenar...|         Yimou Zhang|Matt Damon, Tian ...|
         2016|              103|   6.1| 56036|             45.13|     42.0|           4513.0|Batch3|        Good|
         |   7|          La La Land|  Comedy,Drama,Music|A jazz pianist fa...|     Damien Chazelle|Ryan Gosling, Emm...|
         2016|              128|   8.3|258682|            151.06|     93.0|          15106.0|Batch3|        Best|
         |   8|            Mindhorn|              Comedy|"A has-been actor...| whom he believes...|          Sean Foley|Ess
```

Concatenated Without Function Creation:



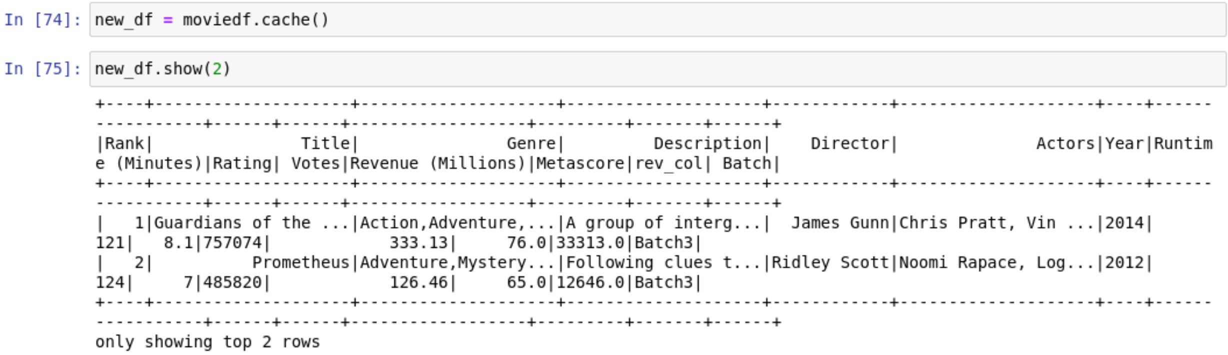
```
In [64]: test_df = test_df.withColumn("Concatenated", concat(col("Title"), lit("_Shell")))

In [65]: test_df.show()

         +----+--------------------+--------------------+--------------------+--------------------+--------------------+---
         -----------------+-----------------+------+------+------------------+---------+-----------------+------+----------
         --+--------------------+
         |Rank|               Title|               Genre|         Description|            Director|              Actors|
         Year|Runtime (Minutes)|Rating| Votes|Revenue (Millions)|Metascore|          rev_col| Batch|Rating_Value|        Co
         ncatenated|
         +----+--------------------+--------------------+--------------------+--------------------+--------------------+---
         -----------------+-----------------+------+------+------------------+---------+-----------------+------+----------
         --+--------------------+
         |   1|Guardians of the ...|Action,Adventure,...|A group of interg...|          James Gunn|Chris Pratt, Vin ...|
         2014|              121|   8.1|757074|            333.13|     76.0|          33313.0|Batch3|        Best|Guardians
         of the ...|
         |   2|          Prometheus|Adventure,Mystery...|Following clues t...|        Ridley Scott|Noomi Rapace, Log...|
         2012|              124|     7|485820|            126.46|     65.0|          12646.0|Batch3|        Good|    Promet
         heus_Shell|
         |   3|               Split|     Horror,Thriller|Three girls are k...|  M. Night Shyamalan|James McAvoy, Any...|
         2016|              117|   7.3|157606|            138.12|     62.0|          13812.0|Batch3|        Good|         S
         plit_Shell|
         |   4|                Sing|Animation,Comedy,...|In a city of huma...|Christophe Lourdelet|Matthew McConaugh...|
```

Concatenated with Function creation:



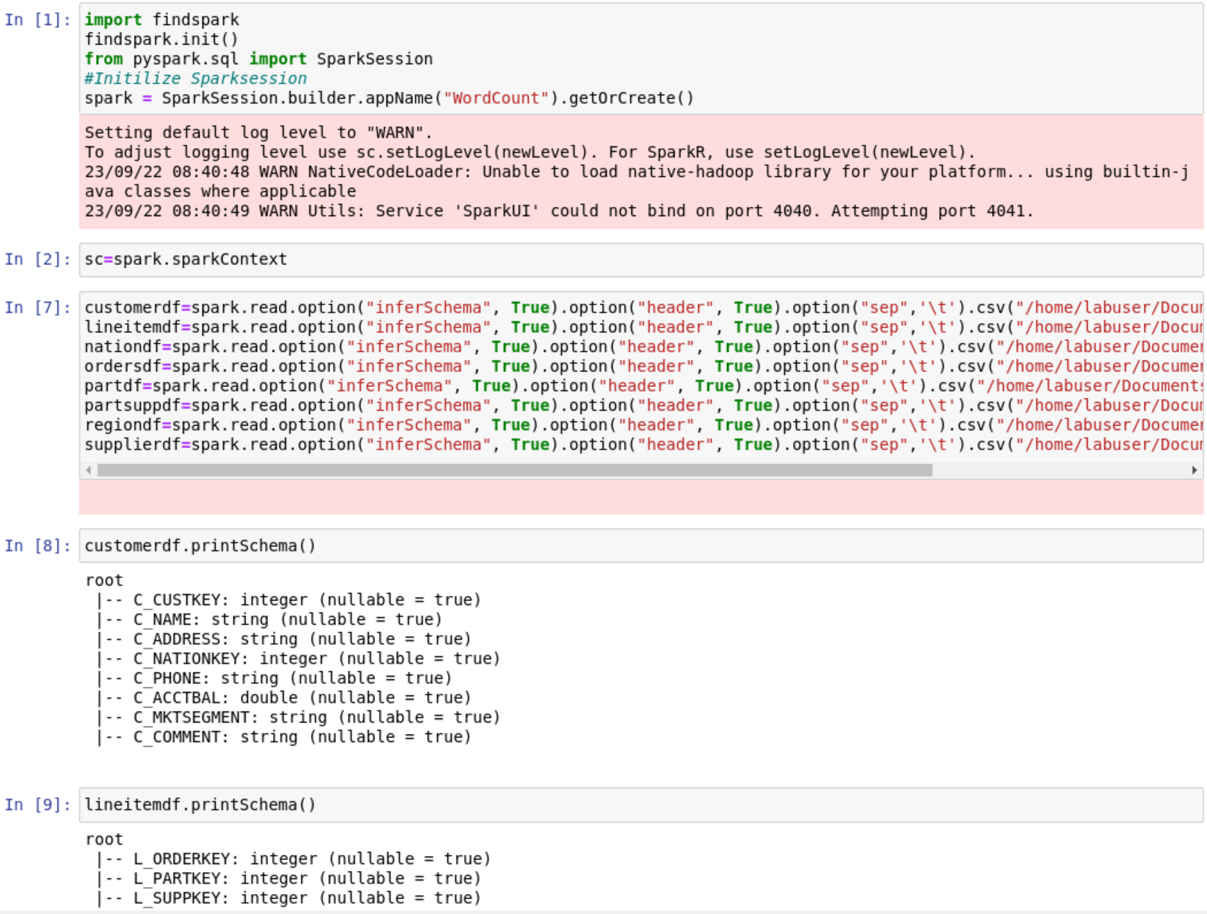
```
In [66]: def concat_shell(column):
             return column + "_shell"

In [67]: my_udf = udf(concat_shell, StringType())

In [70]: from pyspark.sql.functions import *
         test = test_df.withColumn("new_col", my_udf(col("Title")))

In [71]: test.show()

         +----+--------------------+--------------------+--------------------+--------------------+--------------------+---
         -----------------+-----------------+------+------+------------------+---------+-----------------+------+----------
         --+--------------------+--------------------+
         |Rank|               Title|               Genre|         Description|            Director|              Actors|
         Year|Runtime (Minutes)|Rating| Votes|Revenue (Millions)|Metascore|          rev_col| Batch|Rating_Value|        Co
         ncatenated|             new_col|
         +----+--------------------+--------------------+--------------------+--------------------+--------------------+---
         -----------------+-----------------+------+------+------------------+---------+-----------------+------+----------
         --+--------------------+--------------------+
         |   1|Guardians of the ...|Action,Adventure,...|A group of interg...|          James Gunn|Chris Pratt, Vin ...|
         2014|              121|   8.1|757074|            333.13|     76.0|          33313.0|Batch3|        Best|Guardians
         of the ...|Guardians of the ...|
         |   2|          Prometheus|Adventure,Mystery...|Following clues t...|        Ridley Scott|Noomi Rapace, Log...|
         2012|              124|     7|485820|            126.46|     65.0|          12646.0|Batch3|        Good|    Promet
         heus_Shell|    Prometheus_shell|
         |   3|               Split|     Horror,Thriller|Three girls are k...|  M. Night Shyamalan|James McAvoy, Any...|
         2016|              117|   7.3|157606|            138.12|     62.0|          13812.0|Batch3|        Good|         S
         plit_Shell|         Split_shell|
         |   4|                Sing|Animation,Comedy,...|In a city of huma...|Christophe Lourdelet|Matthew McConaugh...|
```

Spark Caching: To cache the data for faster fetching of data.



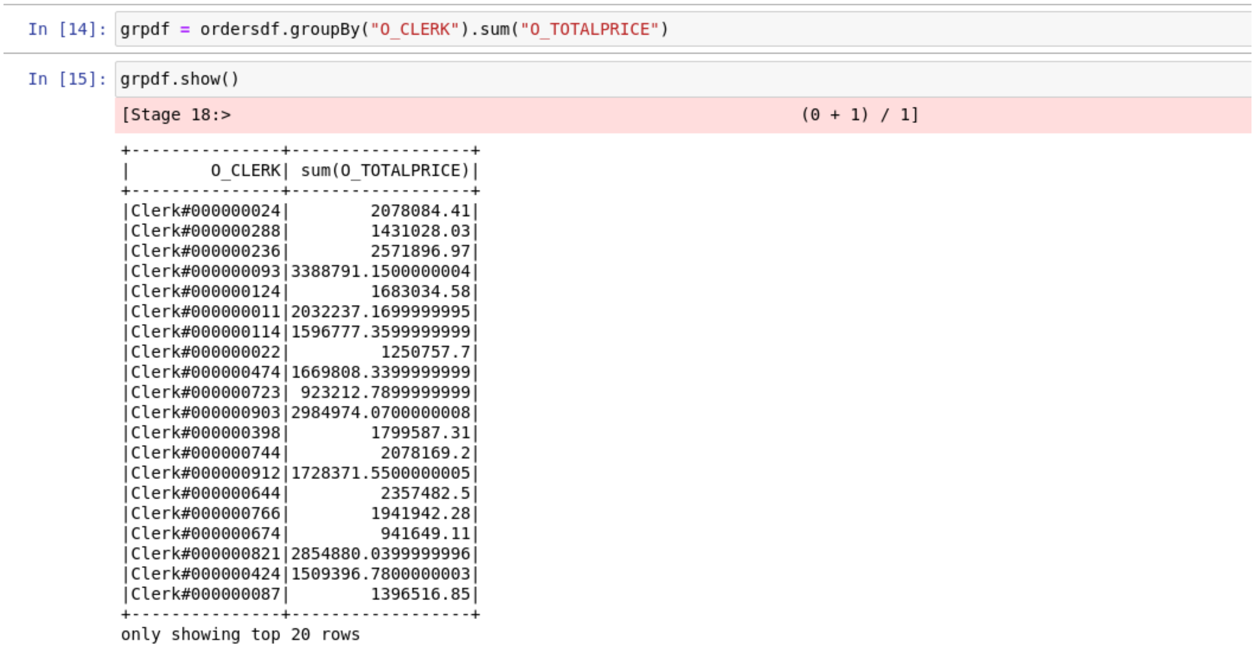
```
In [74]: new_df = moviedf.cache()

In [75]: new_df.show(2)

+----+--------------------+--------------------+--------------------+------------+--------------------+----+------
-----------+------+------+------------------+---------+-------+------+
|Rank|               Title|               Genre|         Description|    Director|              Actors|Year|Runtim
e (Minutes)|Rating| Votes|Revenue (Millions)|Metascore|rev_col| Batch|
+----+--------------------+--------------------+--------------------+------------+--------------------+----+------
-----------+------+------+------------------+---------+-------+------+
|   1|Guardians of the ...|Action,Adventure,...|A group of interg...|  James Gunn|Chris Pratt, Vin ...|2014|
121|   8.1|757074|            333.13|     76.0|33313.0|Batch3|
|   2|          Prometheus|Adventure,Mystery...|Following clues t...|Ridley Scott|Noomi Rapace, Log...|2012|
124|     7|485820|            126.46|     65.0|12646.0|Batch3|
+----+--------------------+--------------------+--------------------+------------+--------------------+----+------
-----------+------+------+------------------+---------+-------+------+
only showing top 2 rows
```

Importing data set:



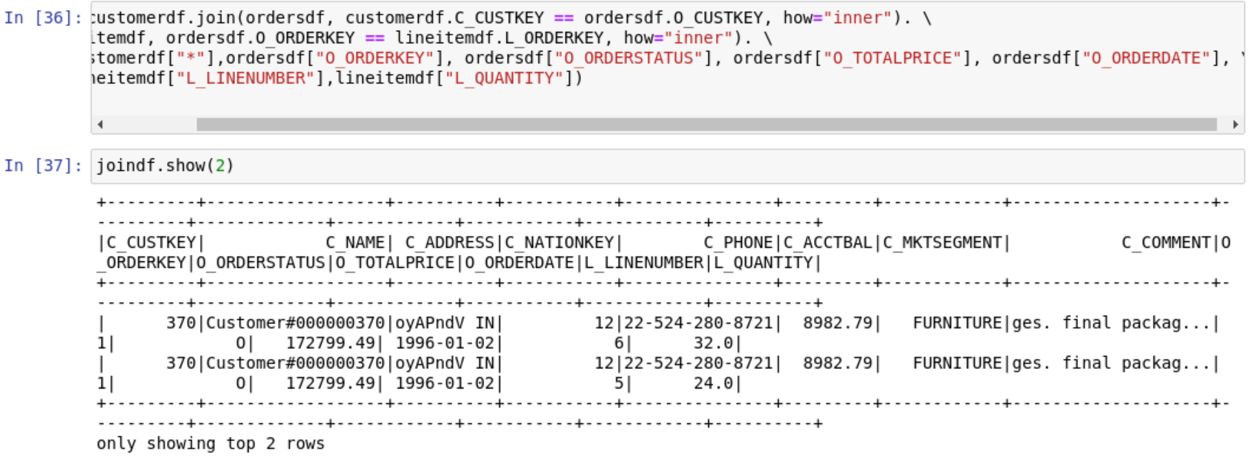
```
In [1]: import findspark
        findspark.init()
        from pyspark.sql import SparkSession
        #Initilize Sparksession
        spark = SparkSession.builder.appName("WordCount").getOrCreate()

        Setting default log level to "WARN".
        To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
        23/09/22 08:40:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-j
        ava classes where applicable
        23/09/22 08:40:49 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

In [2]: sc=spark.sparkContext

In [7]: customerdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docu
        lineitemdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docu
        nationdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docume
        ordersdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docume
        partdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Document
        partsuppdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docu
        regiondf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docume
        supplierdf=spark.read.option("inferSchema", True).option("header", True).option("sep",'\t').csv("/home/labuser/Docu

In [8]: customerdf.printSchema()

        root
         |-- C_CUSTKEY: integer (nullable = true)
         |-- C_NAME: string (nullable = true)
         |-- C_ADDRESS: string (nullable = true)
         |-- C_NATIONKEY: integer (nullable = true)
         |-- C_PHONE: string (nullable = true)
         |-- C_ACCTBAL: double (nullable = true)
         |-- C_MKTSEGMENT: string (nullable = true)
         |-- C_COMMENT: string (nullable = true)

In [9]: lineitemdf.printSchema()

        root
         |-- L_ORDERKEY: integer (nullable = true)
         |-- L_PARTKEY: integer (nullable = true)
         |-- L_SUPPKEY: integer (nullable = true)
```

Transformation:



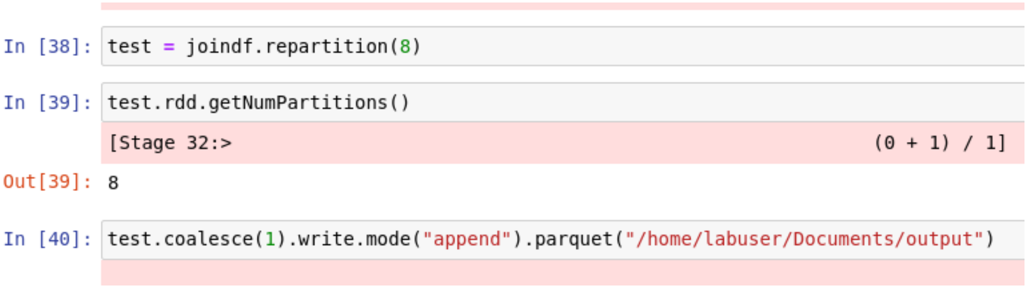
```
In [14]: grpdf = ordersdf.groupBy("O_CLERK").sum("O_TOTALPRICE")

In [15]: grpdf.show()

[Stage 18:>                                                         (0 + 1) / 1]

+---------------+------------------+
|        O_CLERK| sum(O_TOTALPRICE)|
+---------------+------------------+
|Clerk#000000024|        2078084.41|
|Clerk#000000288|        1431028.03|
|Clerk#000000236|        2571896.97|
|Clerk#000000093|3388791.1500000004|
|Clerk#000000124|        1683034.58|
|Clerk#000000011|2032237.1699999995|
|Clerk#000000114|1596777.3599999999|
|Clerk#000000022|         1250757.7|
|Clerk#000000474|1669808.3399999999|
|Clerk#000000723| 923212.7899999999|
|Clerk#000000903|2984974.0700000008|
|Clerk#000000398|        1799587.31|
|Clerk#000000744|         2078169.2|
|Clerk#000000912|1728371.5500000005|
|Clerk#000000644|         2357482.5|
|Clerk#000000766|        1941942.28|
|Clerk#000000674|         941649.11|
|Clerk#000000821|2854880.0399999996|
|Clerk#000000424|1509396.7800000003|
|Clerk#000000087|        1396516.85|
+---------------+------------------+
only showing top 20 rows
```

Joins in Spark:

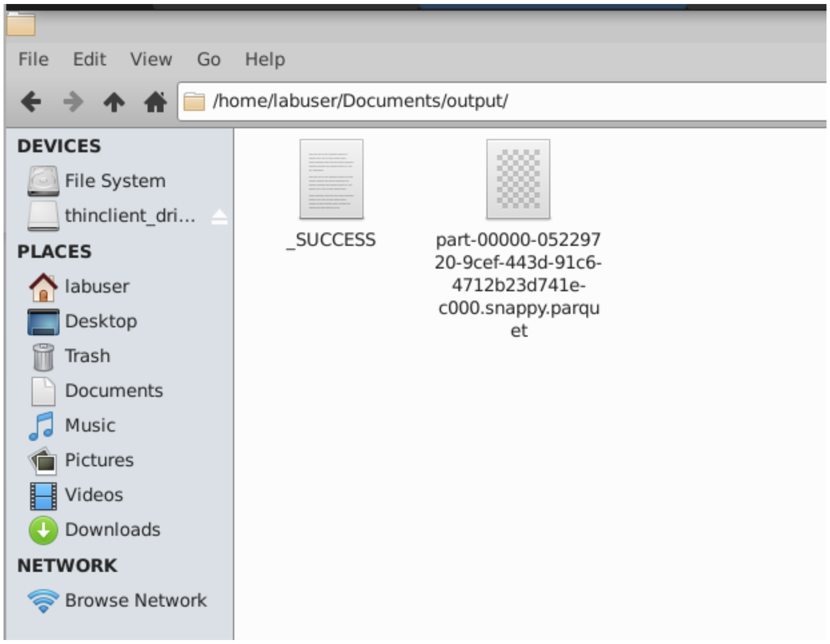


```
In [36]: :ustomerdf.join(ordersdf, customerdf.C_CUSTKEY == ordersdf.O_CUSTKEY, how="inner"). \
         .temdf, ordersdf.O_ORDERKEY == lineitemdf.L_ORDERKEY, how="inner"). \
         ;tomerdf["*"],ordersdf["O_ORDERKEY"], ordersdf["O_ORDERSTATUS"], ordersdf["O_TOTALPRICE"], ordersdf["O_ORDERDATE"], \
         ıeitemdf["L_LINENUMBER"],lineitemdf["L_QUANTITY"])

In [37]: joindf.show(2)

+---------+------------------+----------+-----------+---------------+---------+------------+--------------------+-
---------+-------------+------------+-----------+------------+----------+
|C_CUSTKEY|            C_NAME| C_ADDRESS|C_NATIONKEY|        C_PHONE|C_ACCTBAL|C_MKTSEGMENT|           C_COMMENT|O
_ORDERKEY|O_ORDERSTATUS|O_TOTALPRICE|O_ORDERDATE|L_LINENUMBER|L_QUANTITY|
+---------+------------------+----------+-----------+---------------+---------+------------+--------------------+-
---------+-------------+------------+-----------+------------+----------+
|      370|Customer#000000370|oyAPndV IN|         12|22-524-280-8721|  8982.79|   FURNITURE|ges. final packag...|
1|            0|   172799.49| 1996-01-02|           6|      32.0|
|      370|Customer#000000370|oyAPndV IN|         12|22-524-280-8721|  8982.79|   FURNITURE|ges. final packag...|
1|            0|   172799.49| 1996-01-02|           5|      24.0|
+---------+------------------+----------+-----------+---------------+---------+------------+--------------------+-
---------+-------------+------------+-----------+------------+----------+
only showing top 2 rows
```

Partitioning and Coalesce:



```
In [38]: test = joindf.repartition(8)

In [39]: test.rdd.getNumPartitions()

[Stage 32:>                                                         (0 + 1) / 1]

Out[39]: 8

In [40]: test.coalesce(1).write.mode("append").parquet("/home/labuser/Documents/output")
```



/home/labuser/Documents/output/