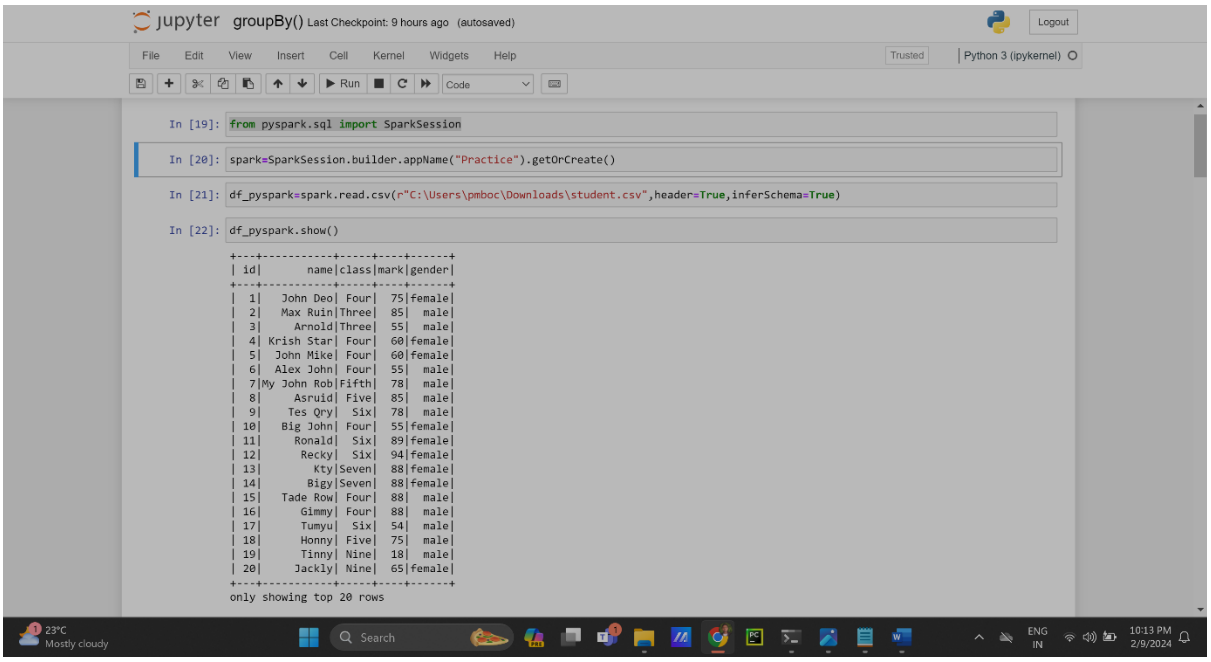
Data Engineering Batch 1 **Day 14:09/02/2024**

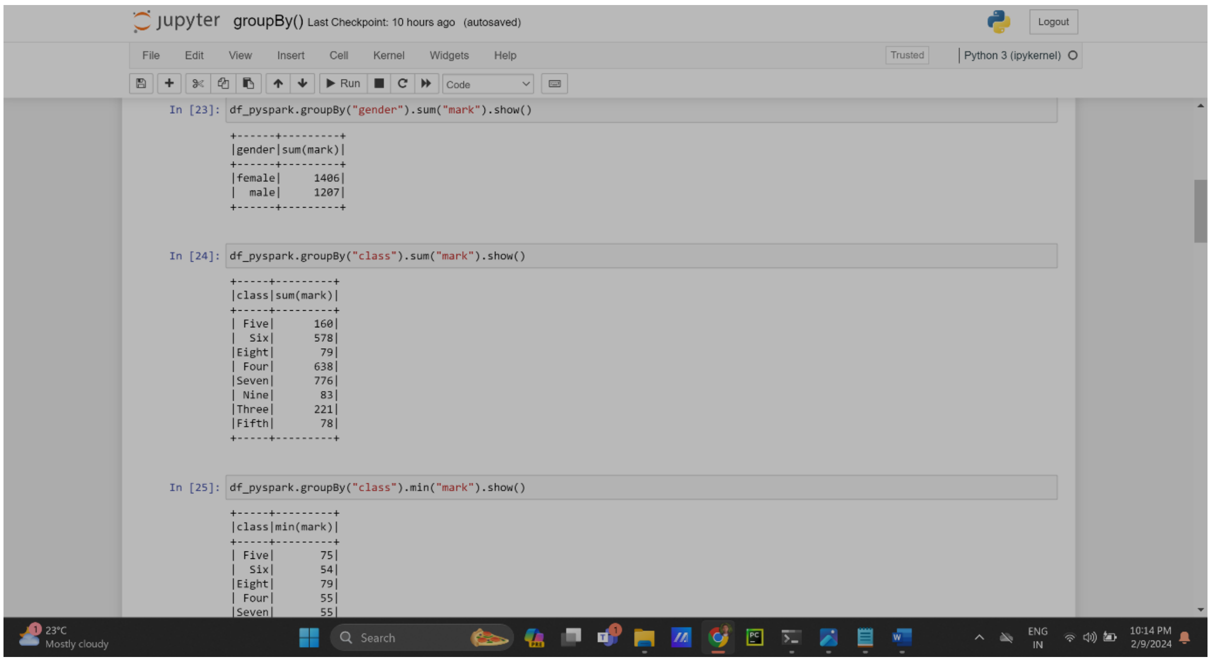
Name: Pradip Bochare

GroupBy and Aggregate function

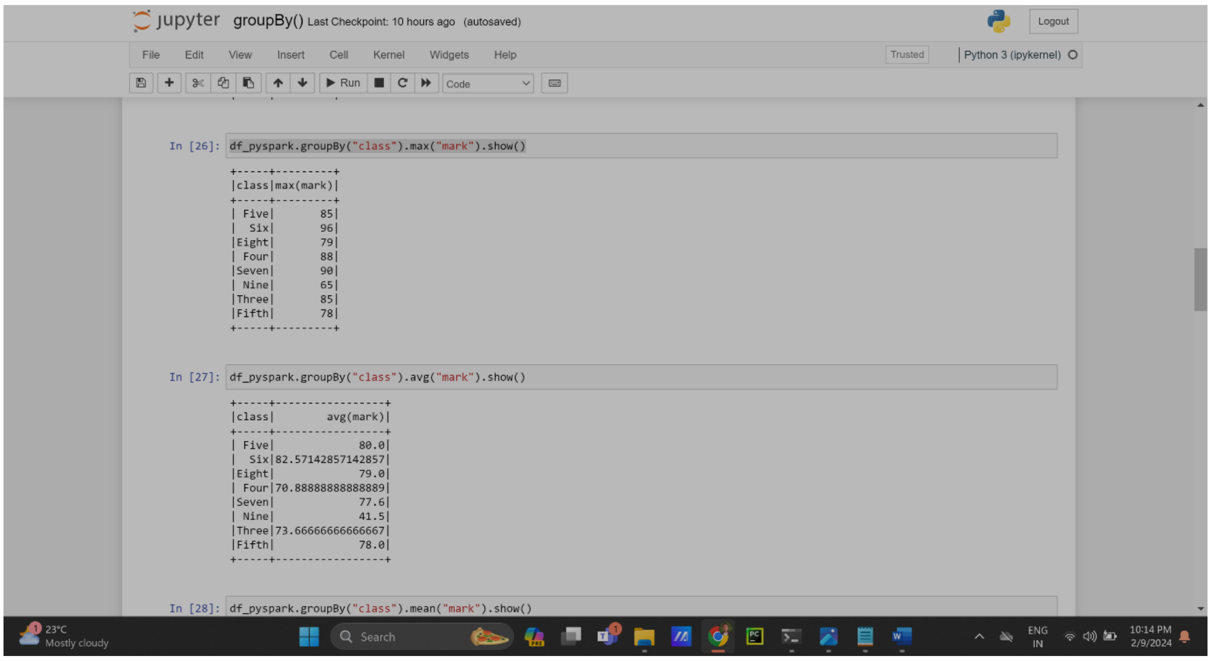
Similar to SQL GROUP BY clause, PySpark groupBy() function is used to collect the identical data into groups on DataFrame and perform count, sum, avg, min, and max functions on the grouped data.


jupyter groupBy() Last Checkpoint: 9 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:13 PM 2/9/2024

```
In [19]: from pyspark.sql import SparkSession

In [20]: spark=SparkSession.builder.appName("Practice").getOrCreate()

In [21]: df_pyspark=spark.read.csv(r"C:\Users\pmboc\Downloads\student.csv",header=True,inferSchema=True)

In [22]: df_pyspark.show()

+---+-----------+-----+----+------+
| id|       name|class|mark|gender|
+---+-----------+-----+----+------+
|  1|   John Deo| Four|  75|female|
|  2|   Max Ruin|Three|  85|  male|
|  3|     Arnold|Three|  55|  male|
|  4| Krish Star| Four|  60|female|
|  5|  John Mike| Four|  60|female|
|  6|  Alex John| Four|  55|  male|
|  7|My John Rob|Fifth|  78|  male|
|  8|     Asruid| Five|  85|  male|
|  9|    Tes Qry|  Six|  78|  male|
| 10|   Big John| Four|  55|female|
| 11|     Ronald|  Six|  89|female|
| 12|      Recky|  Six|  94|female|
| 13|        Kty|Seven|  88|female|
| 14|       Bigy|Seven|  88|female|
| 15|   Tade Row| Four|  88|  male|
| 16|      Gimmy| Four|  88|  male|
| 17|      Tumyu|  Six|  54|  male|
| 18|      Honny| Five|  75|  male|
| 19|      Tinny| Nine|  18|  male|
| 20|     Jackly| Nine|  65|female|
+---+-----------+-----+----+------+
only showing top 20 rows
```


jupyter groupBy() Last Checkpoint: 10 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:14 PM 2/9/2024

```
In [23]: df_pyspark.groupBy("gender").sum("mark").show()

+------+---------+
|gender|sum(mark)|
+------+---------+
|female|     1406|
|  male|     1207|
+------+---------+


In [24]: df_pyspark.groupBy("class").sum("mark").show()

+-----+---------+
|class|sum(mark)|
+-----+---------+
| Five|      160|
|  Six|      578|
|Eight|       79|
| Four|      638|
|Seven|      776|
| Nine|       83|
|Three|      221|
|Fifth|       78|
+-----+---------+


In [25]: df_pyspark.groupBy("class").min("mark").show()

+-----+---------+
|class|min(mark)|
+-----+---------+
| Five|       75|
|  Six|       54|
|Eight|       79|
| Four|       55|
|Seven|       55|
```


jupyter groupBy() Last Checkpoint: 10 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:14 PM 2/9/2024

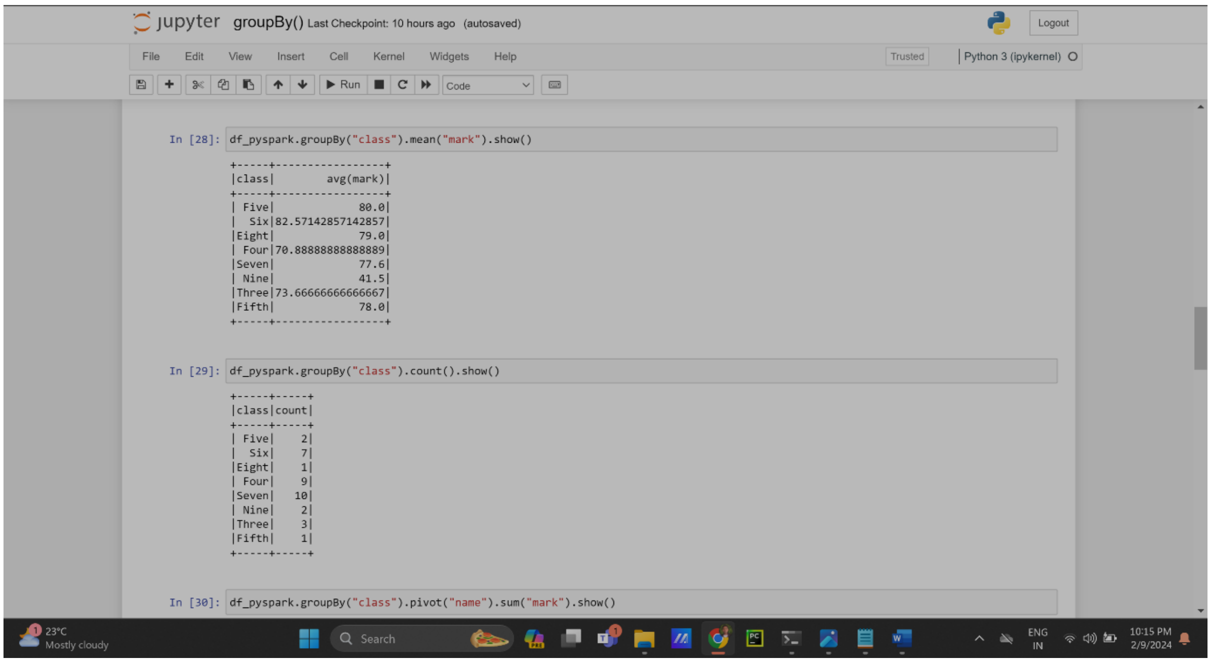
In [26]:
```
df_pyspark.groupBy("class").max("mark").show()
```

```
+-----+---------+
|class|max(mark)|
+-----+---------+
| Five|       85|
|  Six|       96|
|Eight|       79|
| Four|       88|
|Seven|       90|
| Nine|       65|
|Three|       85|
|Fifth|       78|
+-----+---------+
```

In [27]:
```
df_pyspark.groupBy("class").avg("mark").show()
```

```
+-----+-----------------+
|class|        avg(mark)|
+-----+-----------------+
| Five|             80.0|
|  Six|82.57142857142857|
|Eight|             79.0|
| Four|70.88888888888889|
|Seven|             77.6|
| Nine|             41.5|
|Three|73.66666666666667|
|Fifth|             78.0|
+-----+-----------------+
```

In [28]:
```
df_pyspark.groupBy("class").mean("mark").show()
```


jupyter groupBy() Last Checkpoint: 10 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:15 PM 2/9/2024

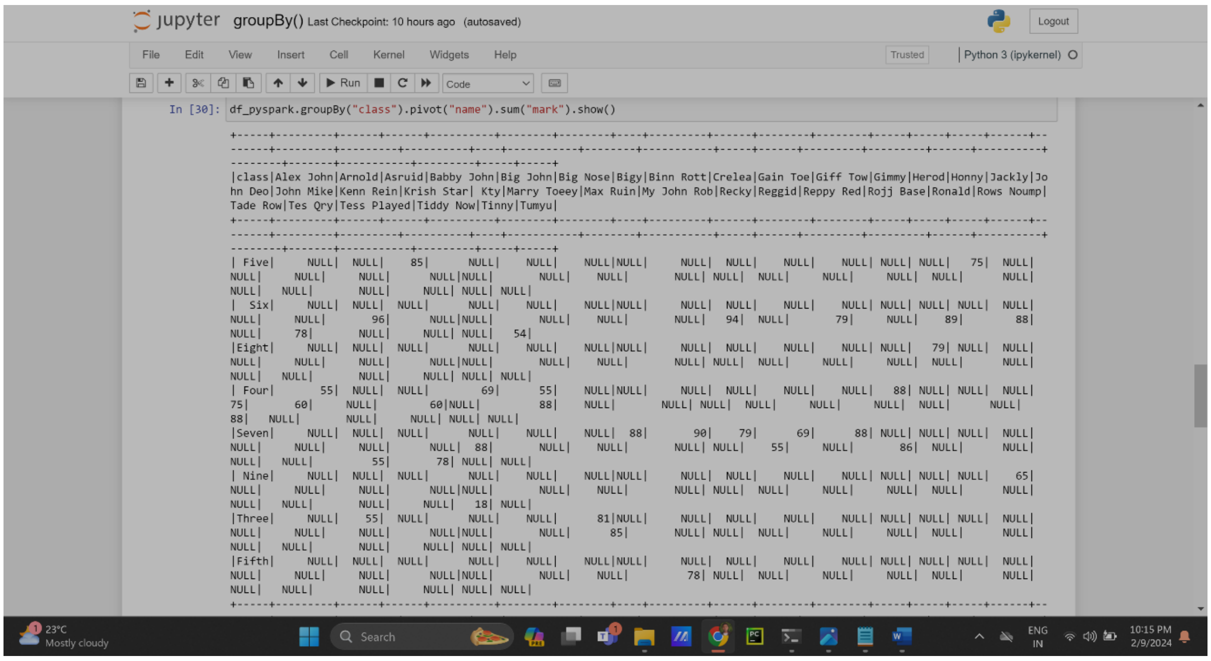
In [28]:
```
df_pyspark.groupBy("class").mean("mark").show()
```

```
+-----+-----------------+
|class|        avg(mark)|
+-----+-----------------+
| Five|             80.0|
|  Six|82.57142857142857|
|Eight|             79.0|
| Four|70.88888888888889|
|Seven|             77.6|
| Nine|             41.5|
|Three|73.66666666666667|
|Fifth|             78.0|
+-----+-----------------+
```

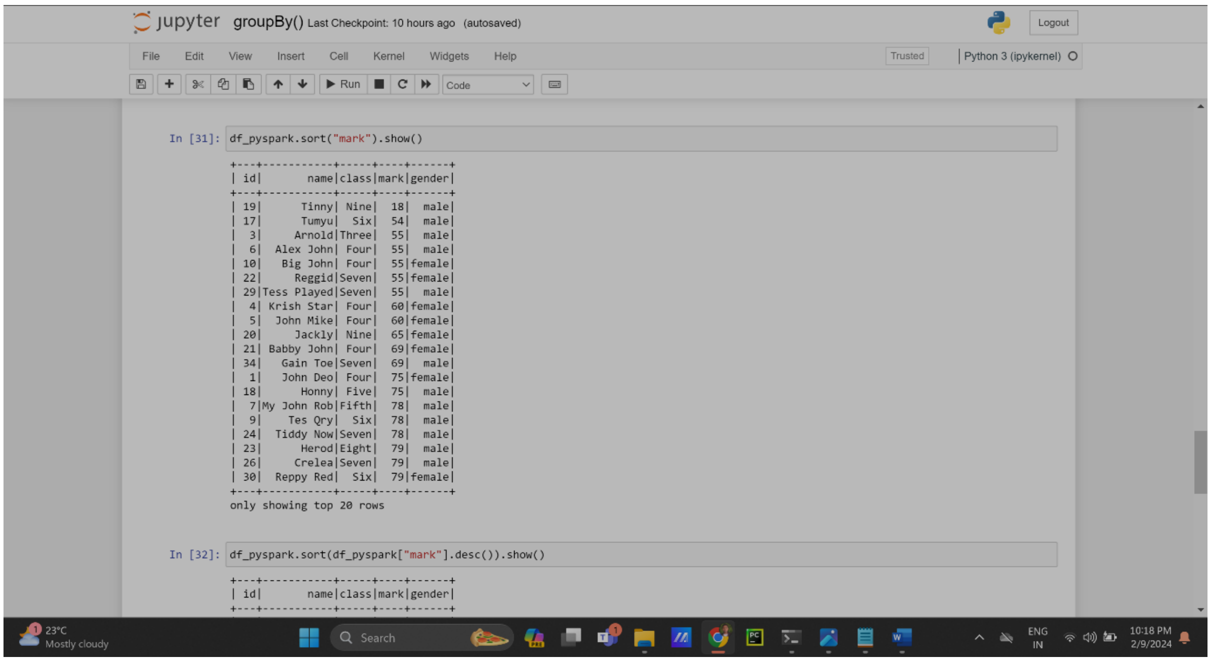
In [29]:
```
df_pyspark.groupBy("class").count().show()
```

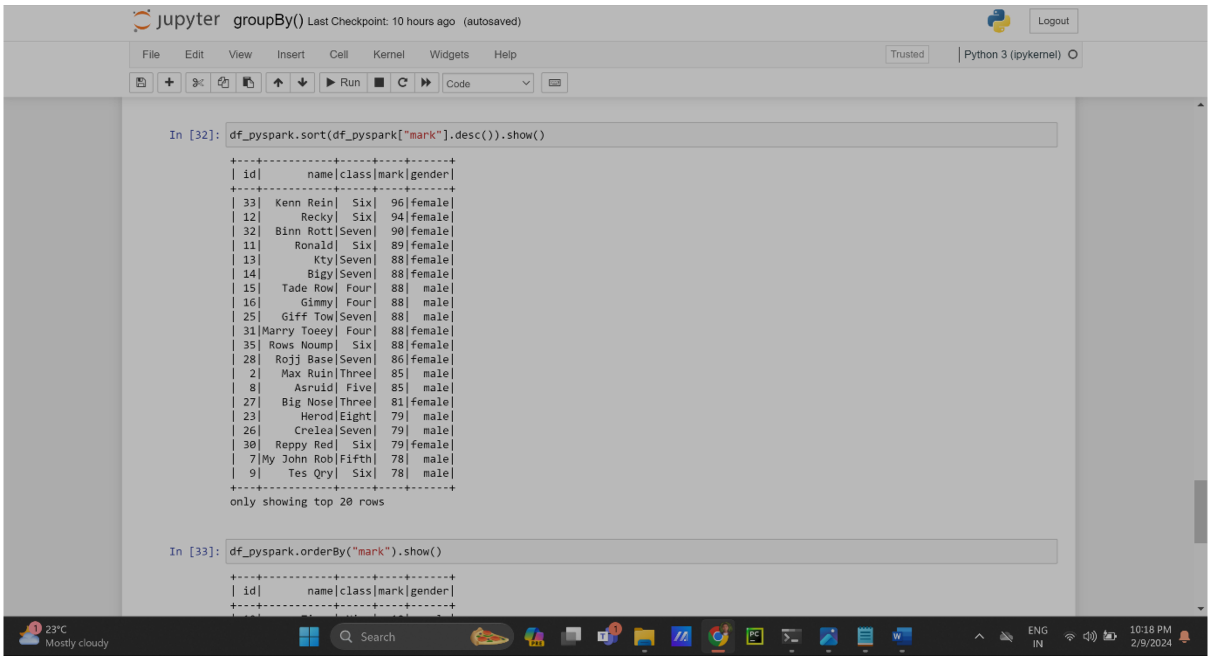
```
+-----+-----+
|class|count|
+-----+-----+
| Five|    2|
|  Six|    7|
|Eight|    1|
| Four|    9|
|Seven|   10|
| Nine|    2|
|Three|    3|
|Fifth|    1|
+-----+-----+
```

In [30]:
```
df_pyspark.groupBy("class").pivot("name").sum("mark").show()
```


```
In [30]: df_pyspark.groupBy("class").pivot("name").sum("mark").show()

+-----+---------+------+------+----------+--------+--------+----+---------+------+--------+--------+-----+-----+-----+------+--
------+---------+---------+---------+----+-----------+--------+-----+------+-----+-----+----------+---------+------+----------+
--------+-------+-----------+---------+-----+-----+
|class|Alex John|Arnold|Asruid|Babby John|Big John|Big Nose|Bigy|Binn Rott|Crelea|Gain Toe|Giff Tow|Gimmy|Herod|Honny|Jackly|Jo
hn Deo|John Mike|Kenn Rein|Krish Star| Kty|Marry Toeey|Max Ruin|My John Rob|Recky|Reggid|Reppy Red|Rojj Base|Ronald|Rows Noump|
Tade Row|Tes Qry|Tess Played|Tiddy Now|Tinny|Tumyu|
+-----+---------+------+------+----------+--------+--------+----+---------+------+--------+--------+-----+-----+-----+------+--
------+---------+---------+---------+----+-----------+--------+-----+------+-----+-----+----------+---------+------+----------+
--------+-------+-----------+---------+-----+-----+
| Five|     NULL|  NULL|    85|      NULL|    NULL|    NULL|NULL|     NULL|  NULL|    NULL|    NULL| NULL| NULL|   75|  NULL|
NULL|     NULL|      NULL|      NULL|NULL|       NULL|    NULL|       NULL| NULL|  NULL|     NULL|     NULL|  NULL|      NULL|
NULL|   NULL|       NULL|     NULL| NULL| NULL|
|  Six|     NULL|  NULL|  NULL|      NULL|    NULL|    NULL|NULL|     NULL|  NULL|    NULL|    NULL| NULL| NULL| NULL|  NULL|
NULL|     NULL|        96|      NULL|NULL|       NULL|    NULL|       NULL|   94|  NULL|       79|     NULL|    89|        88|
NULL|     78|       NULL|     NULL| NULL|   54|
|Eight|     NULL|  NULL|  NULL|      NULL|    NULL|    NULL|NULL|     NULL|  NULL|    NULL|    NULL| NULL|   79| NULL|  NULL|
NULL|     NULL|      NULL|      NULL|NULL|       NULL|    NULL|       NULL| NULL|  NULL|     NULL|     NULL|  NULL|      NULL|
NULL|   NULL|       NULL|     NULL| NULL| NULL|
| Four|       55|  NULL|  NULL|        69|      55|    NULL|NULL|     NULL|  NULL|    NULL|    NULL|   88| NULL| NULL|  NULL|
75|       60|      NULL|        60|NULL|         88|    NULL|      NULL| NULL|  NULL|     NULL|     NULL|  NULL|      NULL|
88|   NULL|       NULL|     NULL| NULL| NULL|
|Seven|     NULL|  NULL|  NULL|      NULL|    NULL|    NULL|  88|       90|    79|      69|      88| NULL| NULL| NULL|  NULL|
NULL|     NULL|      NULL|      NULL|  88|       NULL|    NULL|       NULL| NULL|    55|     NULL|       86|  NULL|      NULL|
NULL|   NULL|         55|       78| NULL| NULL|
| Nine|     NULL|  NULL|  NULL|      NULL|    NULL|    NULL|NULL|     NULL|  NULL|    NULL|    NULL| NULL| NULL| NULL|    65|
NULL|     NULL|      NULL|      NULL|NULL|       NULL|    NULL|       NULL| NULL|  NULL|     NULL|     NULL|  NULL|      NULL|
NULL|   NULL|       NULL|     NULL|   18| NULL|
|Three|     NULL|    55|  NULL|      NULL|    NULL|      81|NULL|     NULL|  NULL|    NULL|    NULL| NULL| NULL| NULL|  NULL|
NULL|     NULL|      NULL|      NULL|NULL|       NULL|      85|       NULL| NULL|  NULL|     NULL|     NULL|  NULL|      NULL|
NULL|   NULL|       NULL|     NULL| NULL| NULL|
|Fifth|     NULL|  NULL|  NULL|      NULL|    NULL|    NULL|NULL|     NULL|  NULL|    NULL|    NULL| NULL| NULL| NULL|  NULL|
NULL|     NULL|      NULL|      NULL|NULL|       NULL|    NULL|         78| NULL|  NULL|     NULL|     NULL|  NULL|      NULL|
NULL|   NULL|       NULL|     NULL| NULL| NULL|
+-----+---------+------+------+----------+--------+--------+----+---------+------+--------+--------+-----+-----+-----+------+--
```

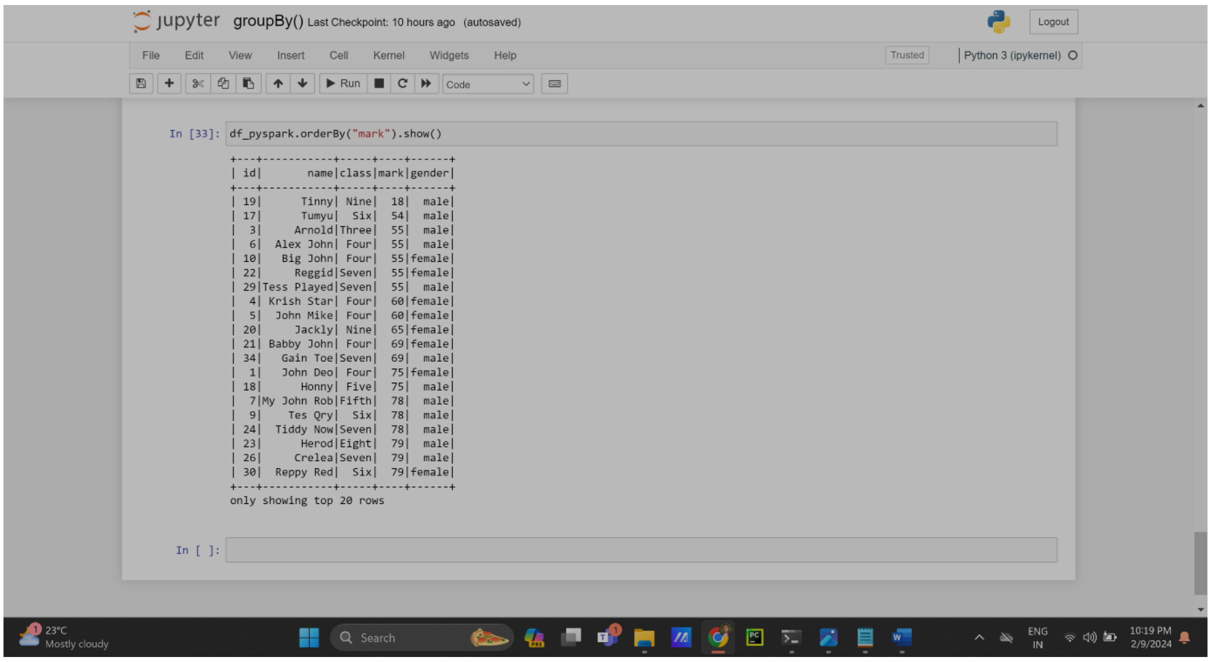
orderBy() and sort() in Pyspark DataFrame


```
In [31]: df_pyspark.sort("mark").show()

+---+-----------+-----+----+------+
| id|       name|class|mark|gender|
+---+-----------+-----+----+------+
| 19|      Tinny| Nine|  18|  male|
| 17|      Tumyu|  Six|  54|  male|
|  3|     Arnold|Three|  55|  male|
|  6|  Alex John| Four|  55|  male|
| 10|   Big John| Four|  55|female|
| 22|     Reggid|Seven|  55|female|
| 29|Tess Played|Seven|  55|  male|
|  4| Krish Star| Four|  60|female|
|  5|  John Mike| Four|  60|female|
| 20|     Jackly| Nine|  65|female|
| 21| Babby John| Four|  69|female|
| 34|   Gain Toe|Seven|  69|  male|
|  1|   John Deo| Four|  75|female|
| 18|      Honny| Five|  75|  male|
|  7|My John Rob|Fifth|  78|  male|
|  9|    Tes Qry|  Six|  78|  male|
| 24|  Tiddy Now|Seven|  78|  male|
| 23|      Herod|Eight|  79|  male|
| 26|     Crelea|Seven|  79|  male|
| 30|  Reppy Red|  Six|  79|female|
+---+-----------+-----+----+------+
only showing top 20 rows

In [32]: df_pyspark.sort(df_pyspark["mark"].desc()).show()

+---+-----------+-----+----+------+
| id|       name|class|mark|gender|
+---+-----------+-----+----+------+
```


jupyter groupBy() Last Checkpoint: 10 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [32]: df_pyspark.sort(df_pyspark["mark"].desc()).show()

+---+-----------+-----+----+------+
| id|       name|class|mark|gender|
+---+-----------+-----+----+------+
| 33|  Kenn Rein|  Six|  96|female|
| 12|      Recky|  Six|  94|female|
| 32|  Binn Rott|Seven|  90|female|
| 11|     Ronald|  Six|  89|female|
| 13|        Kty|Seven|  88|female|
| 14|       Bigy|Seven|  88|female|
| 15|   Tade Row| Four|  88|  male|
| 16|      Gimmy| Four|  88|  male|
| 25|   Giff Tow|Seven|  88|  male|
| 31|Marry Toeey| Four|  88|female|
| 35| Rows Noump|  Six|  88|female|
| 28|  Rojj Base|Seven|  86|female|
|  2|   Max Ruin|Three|  85|  male|
|  8|     Asruid| Five|  85|  male|
| 27|   Big Nose|Three|  81|female|
| 23|      Herod|Eight|  79|  male|
| 26|     Crelea|Seven|  79|  male|
| 30|  Reppy Red|  Six|  79|female|
|  7|My John Rob|Fifth|  78|  male|
|  9|    Tes Qry|  Six|  78|  male|
+---+-----------+-----+----+------+
only showing top 20 rows

In [33]: df_pyspark.orderBy("mark").show()

+---+-----------+-----+----+------+
| id|       name|class|mark|gender|
+---+-----------+-----+----+------+
```

23°C
Mostly cloudy
Search
ENG
IN
10:18 PM
2/9/2024

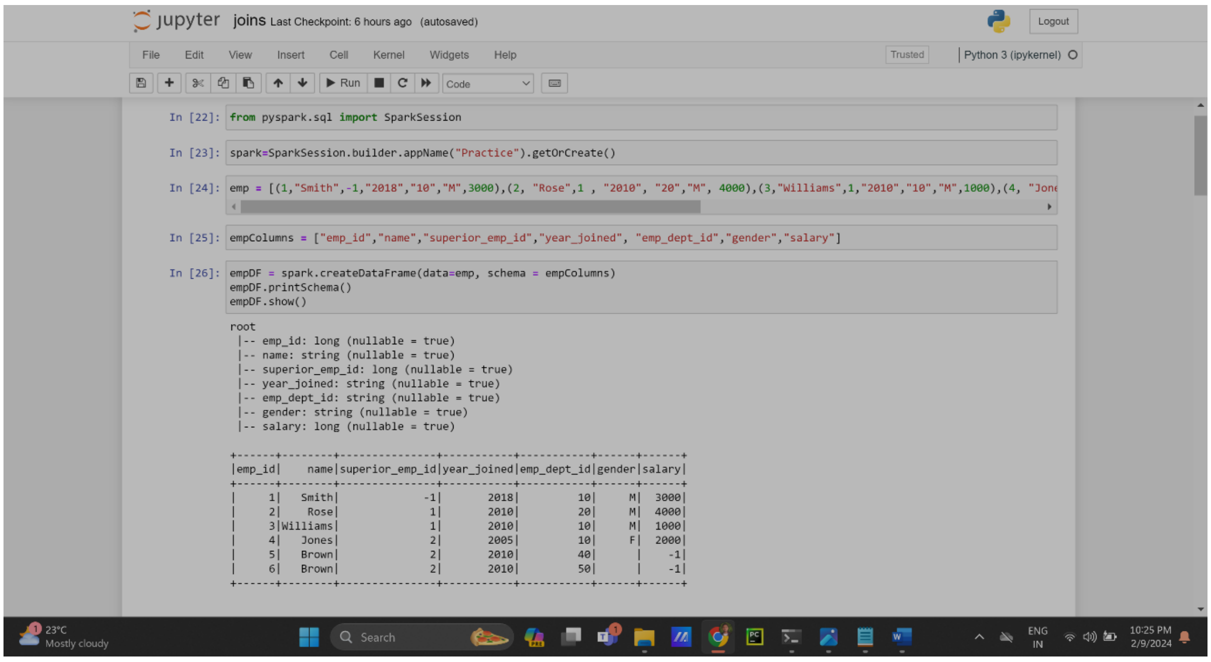

jupyter groupBy() Last Checkpoint: 10 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [33]: df_pyspark.orderBy("mark").show()

+---+-----------+-----+----+------+
| id|       name|class|mark|gender|
+---+-----------+-----+----+------+
| 19|      Tinny| Nine|  18|  male|
| 17|      Tumyu|  Six|  54|  male|
|  3|     Arnold|Three|  55|  male|
|  6|  Alex John| Four|  55|  male|
| 10|   Big John| Four|  55|female|
| 22|     Reggid|Seven|  55|female|
| 29|Tess Played|Seven|  55|  male|
|  4| Krish Star| Four|  60|female|
|  5|  John Mike| Four|  60|female|
| 20|     Jackly| Nine|  65|female|
| 21| Babby John| Four|  69|female|
| 34|   Gain Toe|Seven|  69|  male|
|  1|   John Deo| Four|  75|female|
| 18|      Honny| Five|  75|  male|
|  7|My John Rob|Fifth|  78|  male|
|  9|    Tes Qry|  Six|  78|  male|
| 24|  Tiddy Now|Seven|  78|  male|
| 23|      Herod|Eight|  79|  male|
| 26|     Crelea|Seven|  79|  male|
| 30|  Reppy Red|  Six|  79|female|
+---+-----------+-----+----+------+
only showing top 20 rows

In [ ]:
```

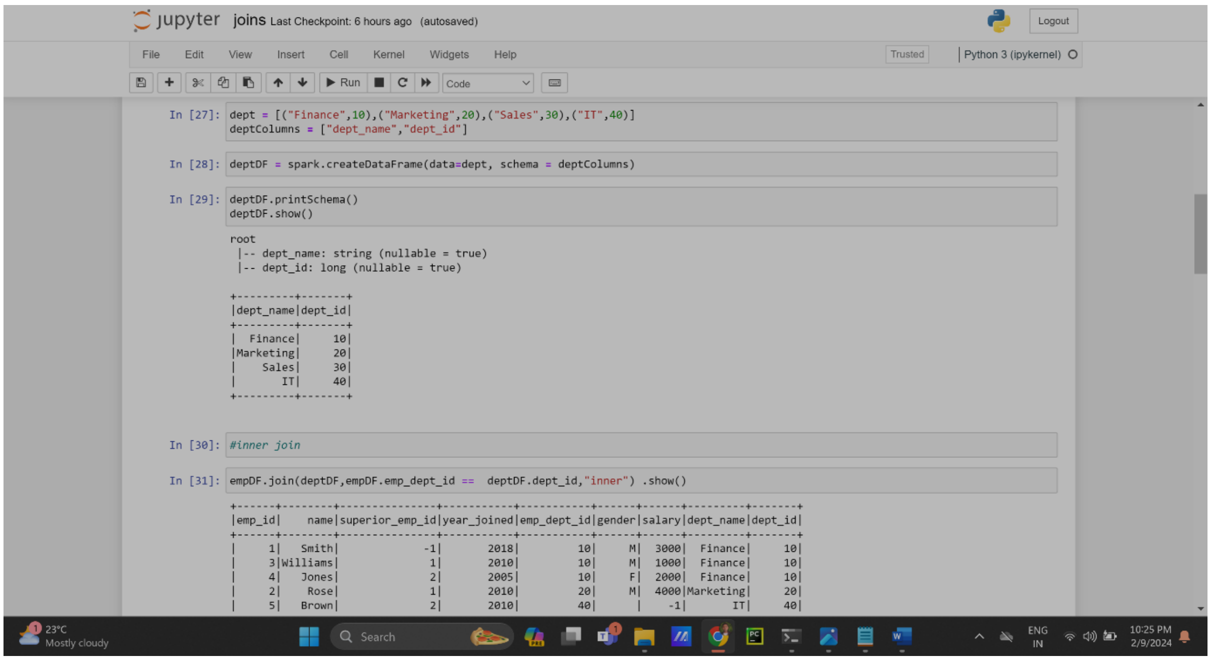
join() using pyspark

PySpark Join is used to combine two DataFrames and by chaining these you can join multiple DataFrames; it supports all basic join type operations available in traditional SQL like INNER, LEFT OUTER, RIGHT OUTER, LEFT ANTI, LEFT SEMI, CROSS, SELF join.

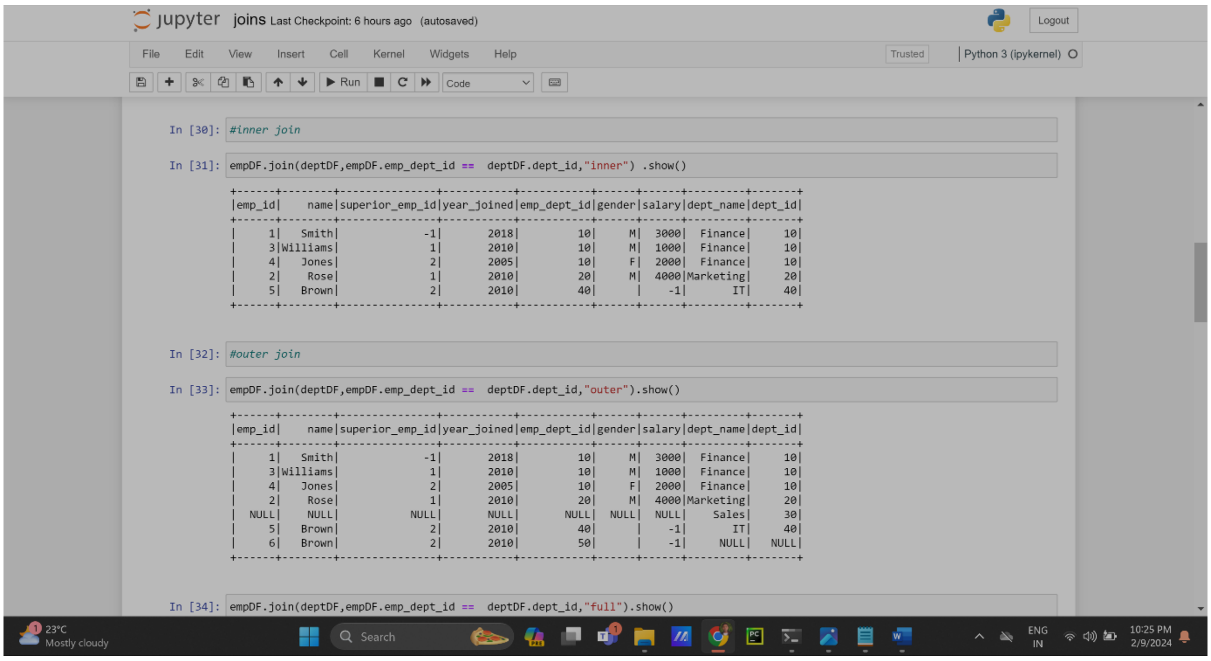
Join Type	**Description**
Inner Join	Join records when key columns are matched, and dropped when they are not matched
Outer join	Returns all rows from both datasets, where Join expression doesn't match it returns null or respective columns
Left Join/ Left outer join	Returns all rows from left dataset regardless of match found on right dataset, when Join doesn't match – it assigns null for that record
Right Join/ Right outer join	Returns all rows from Right dataset regardless of match found on left dataset, when Join doesn't match – it assigns null for that record
Left Semi Join	Returns columns from the only left dataset for the matched records in the right dataset on join expression
Left Anti Join	Returns only columns from left dataset for non-matched records


jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

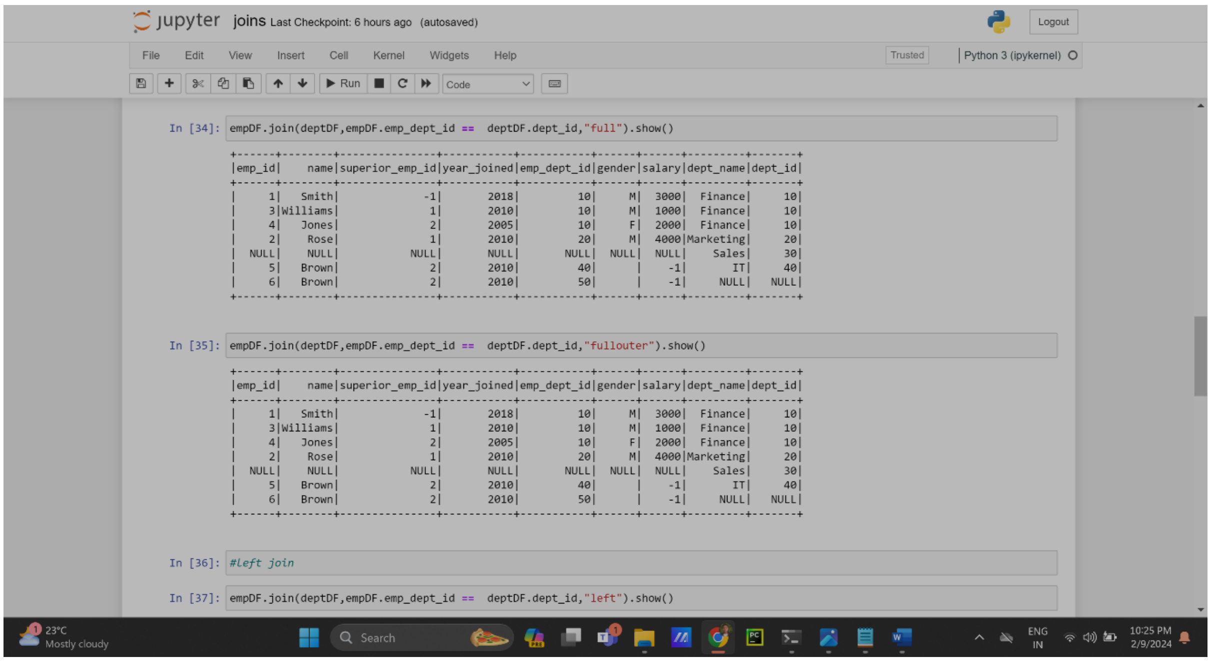
```
In [22]: from pyspark.sql import SparkSession

In [23]: spark=SparkSession.builder.appName("Practice").getOrCreate()

In [24]: emp = [(1,"Smith",-1,"2018","10","M",3000),(2, "Rose",1 , "2010", "20","M", 4000),(3,"Williams",1,"2010","10","M",1000),(4, "Jone

In [25]: empColumns = ["emp_id","name","superior_emp_id","year_joined", "emp_dept_id","gender","salary"]

In [26]: empDF = spark.createDataFrame(data=emp, schema = empColumns)
         empDF.printSchema()
         empDF.show()

root
 |-- emp_id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- superior_emp_id: long (nullable = true)
 |-- year_joined: string (nullable = true)
 |-- emp_dept_id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- salary: long (nullable = true)

+------+--------+---------------+-----------+-----------+------+------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+--------+---------------+-----------+-----------+------+------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|
|     2|    Rose|              1|       2010|         20|     M|  4000|
|     3|Williams|              1|       2010|         10|     M|  1000|
|     4|   Jones|              2|       2005|         10|     F|  2000|
|     5|   Brown|              2|       2010|         40|      |    -1|
|     6|   Brown|              2|       2010|         50|      |    -1|
+------+--------+---------------+-----------+-----------+------+------+
```


jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:25 PM 2/9/2024

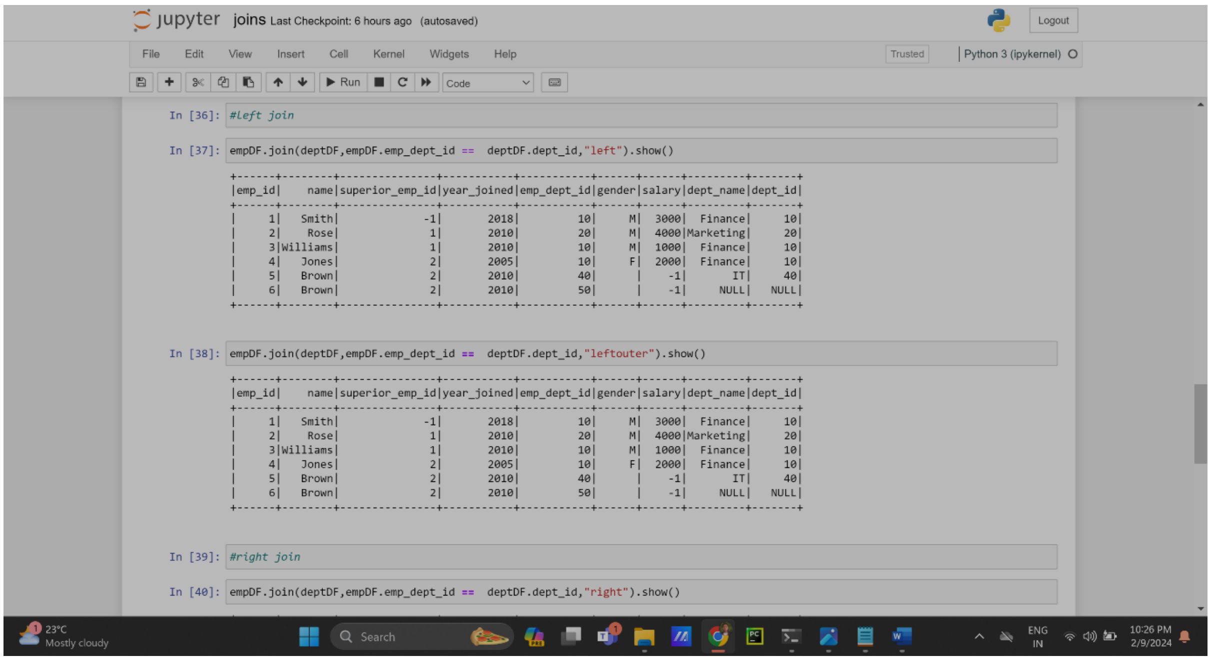
```
In [27]: dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
         deptColumns = ["dept_name","dept_id"]

In [28]: deptDF = spark.createDataFrame(data=dept, schema = deptColumns)

In [29]: deptDF.printSchema()
         deptDF.show()

         root
          |-- dept_name: string (nullable = true)
          |-- dept_id: long (nullable = true)

         +---------+-------+
         |dept_name|dept_id|
         +---------+-------+
         |  Finance|     10|
         |Marketing|     20|
         |    Sales|     30|
         |       IT|     40|
         +---------+-------+


In [30]: #inner join

In [31]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"inner") .show()

         +------+--------+---------------+-----------+-----------+------+------+---------+-------+
         |emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
         +------+--------+---------------+-----------+-----------+------+------+---------+-------+
         |     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
         |     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
         |     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
         |     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
         |     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
```


jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:25 PM 2/9/2024

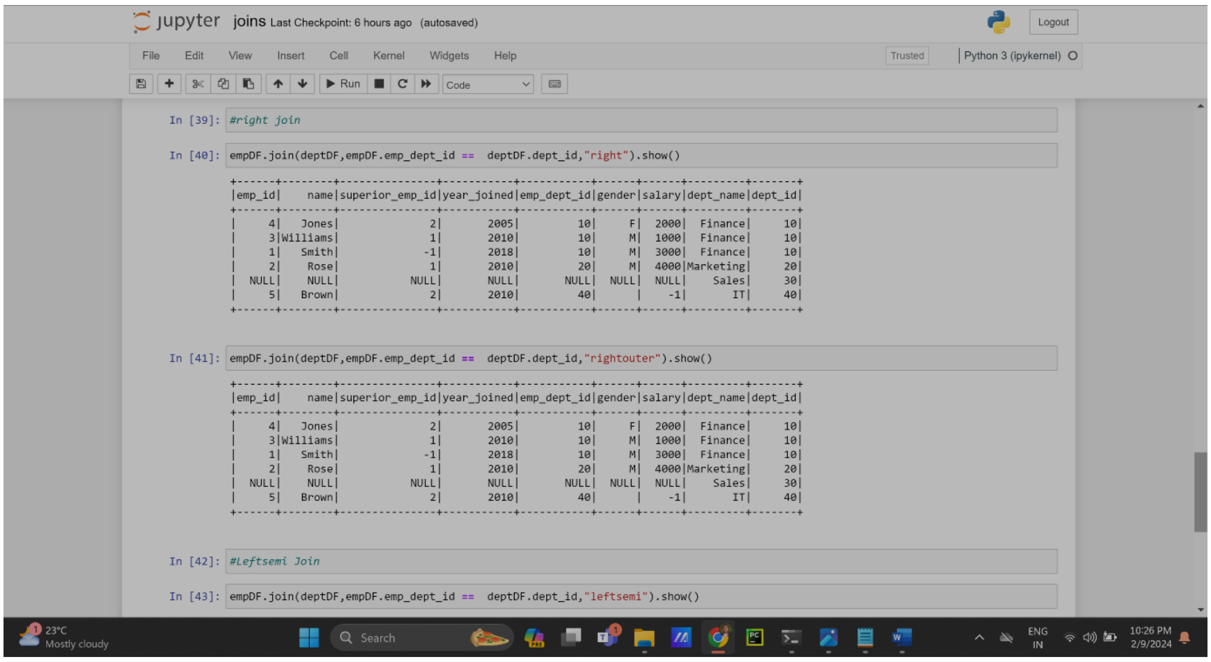
```
In [30]: #inner join

In [31]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"inner") .show()

         +------+--------+---------------+-----------+-----------+------+------+---------+-------+
         |emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
         +------+--------+---------------+-----------+-----------+------+------+---------+-------+
         |     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
         |     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
         |     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
         |     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
         |     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
         +------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [32]: #outer join

In [33]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"outer").show()

         +------+--------+---------------+-----------+-----------+------+------+---------+-------+
         |emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
         +------+--------+---------------+-----------+-----------+------+------+---------+-------+
         |     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
         |     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
         |     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
         |     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
         |  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
         |     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
         |     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
         +------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [34]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"full").show()
```


jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:25 PM 2/9/2024

```
In [34]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"full").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [35]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"fullouter").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [36]: #left join

In [37]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"left").show()
```

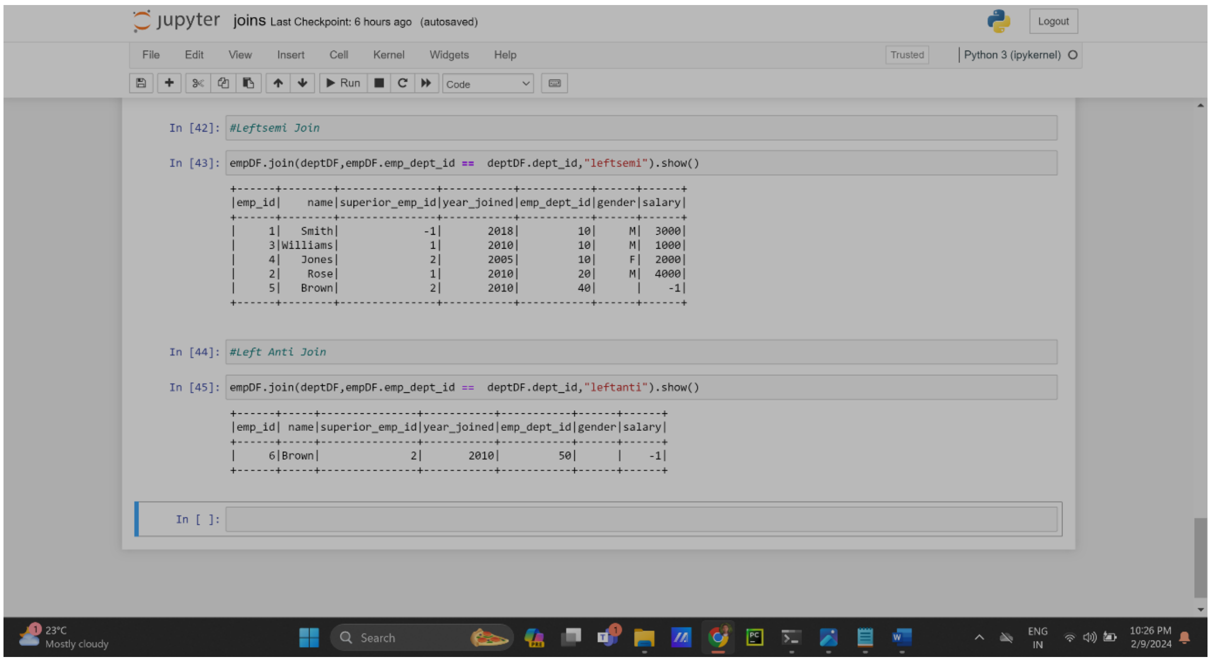

jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
23°C Mostly cloudy
Search
ENG IN
10:26 PM 2/9/2024

```
In [36]: #left join

In [37]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"left").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [38]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftouter").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [39]: #right join

In [40]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"right").show()
```


jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [39]: #right join

In [40]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"right").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [41]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"rightouter").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


In [42]: #Leftsemi Join

In [43]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftsemi").show()
```

23°C
Mostly cloudy
Search
ENG
IN
10:26 PM
2/9/2024


jupyter joins Last Checkpoint: 6 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

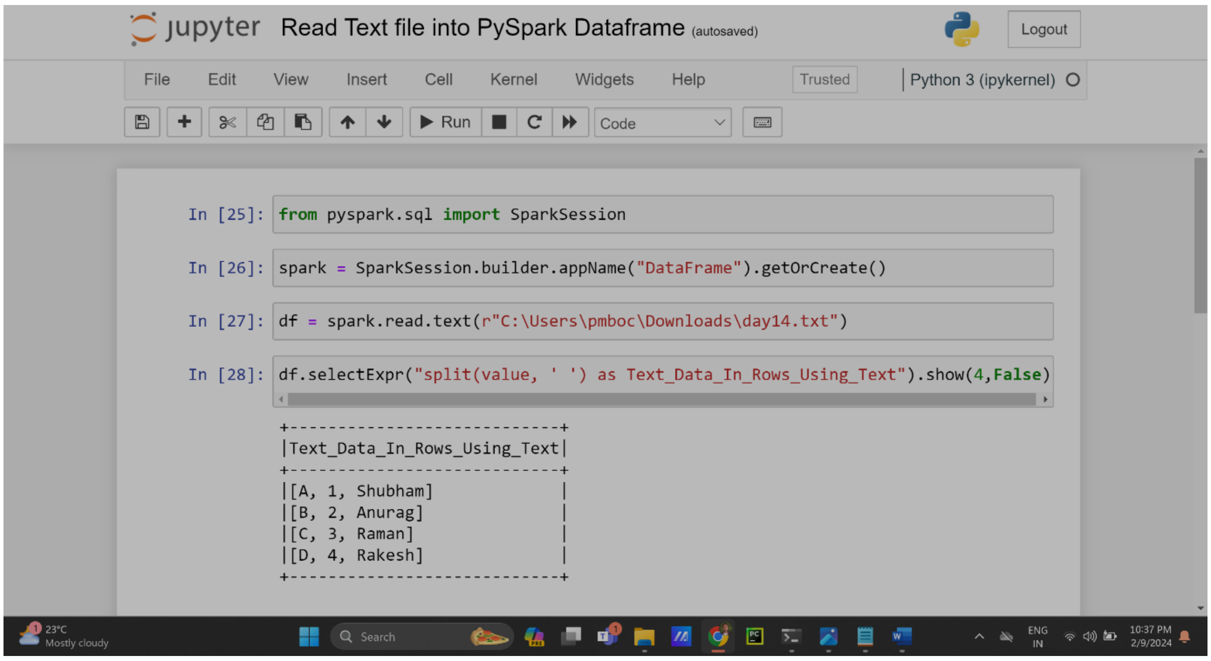
```
In [42]: #Leftsemi Join

In [43]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftsemi").show()

+------+--------+---------------+-----------+-----------+------+------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+--------+---------------+-----------+-----------+------+------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|
|     3|Williams|              1|       2010|         10|     M|  1000|
|     4|   Jones|              2|       2005|         10|     F|  2000|
|     2|    Rose|              1|       2010|         20|     M|  4000|
|     5|   Brown|              2|       2010|         40|      |    -1|
+------+--------+---------------+-----------+-----------+------+------+


In [44]: #Left Anti Join

In [45]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftanti").show()

+------+-----+---------------+-----------+-----------+------+------+
|emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+-----+---------------+-----------+-----------+------+------+
|     6|Brown|              2|       2010|         50|      |    -1|
+------+-----+---------------+-----------+-----------+------+------+

In [ ]:
```

Read Text file into PySpark Dataframe

- Using spark.read.text()
- Using spark.read.csv()
- Using spark.read.format().load()

❖ Method 1: Using spark.read.text()

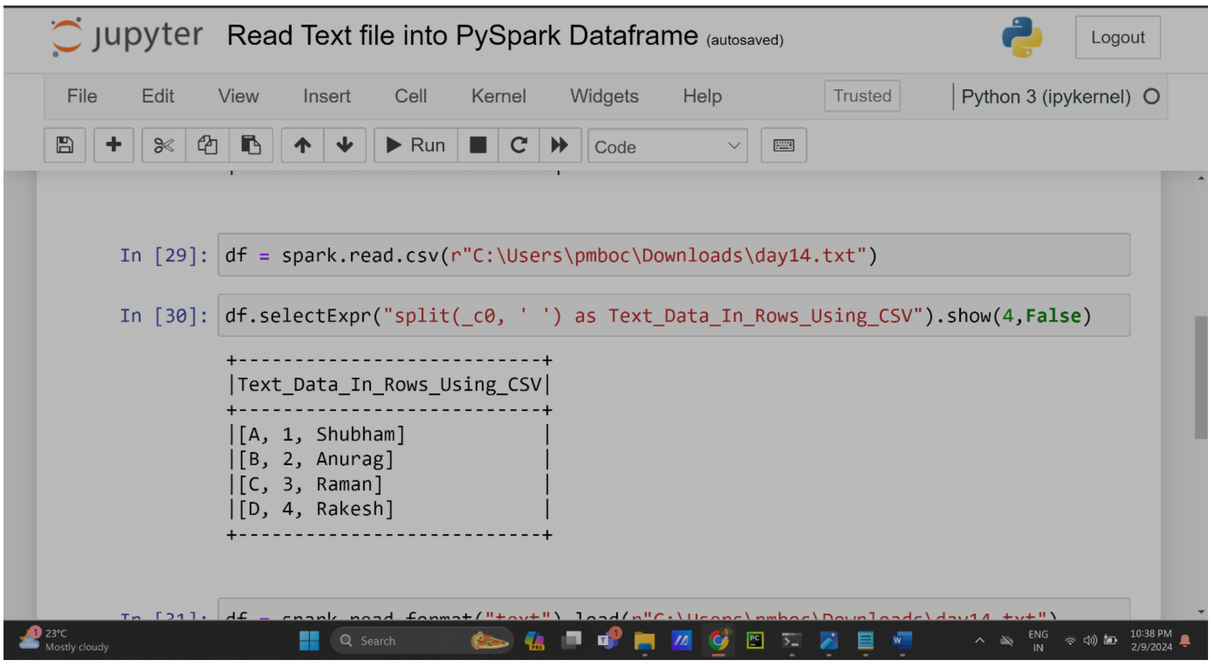
It is used to load text files into DataFrame whose schema starts with a string column. Each line in the text file is a new row in the resulting DataFrame. Using this method we can also read multiple files at a time.


jupyter
Read Text file into PySpark Dataframe (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [25]: from pyspark.sql import SparkSession

In [26]: spark = SparkSession.builder.appName("DataFrame").getOrCreate()

In [27]: df = spark.read.text(r"C:\Users\pmboc\Downloads\day14.txt")

In [28]: df.selectExpr("split(value, ' ') as Text_Data_In_Rows_Using_Text").show(4,False)

+----------------------------+
|Text_Data_In_Rows_Using_Text|
+----------------------------+
|[A, 1, Shubham]             |
|[B, 2, Anurag]              |
|[C, 3, Raman]               |
|[D, 4, Rakesh]              |
+----------------------------+
```

❖ Method 2: Using spark.read.csv()

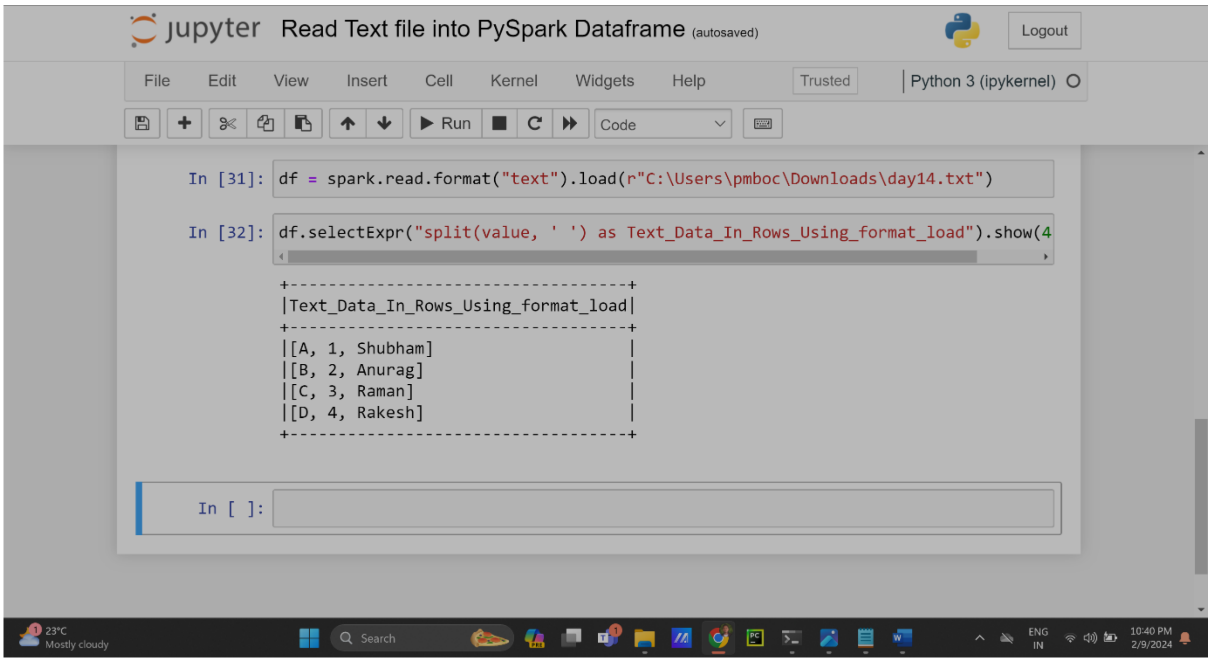
It is used to load text files into DataFrame. Using this method we will go through the input once to determine the input schema if inferSchema is enabled. To avoid going through the entire data once, disable inferSchema option or specify the schema explicitly using the schema.


jupyter Read Text file into PySpark Dataframe (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [29]: df = spark.read.csv(r"C:\Users\pmboc\Downloads\day14.txt")

In [30]: df.selectExpr("split(_c0, ' ') as Text_Data_In_Rows_Using_CSV").show(4,False)

         +---------------------------+
         |Text_Data_In_Rows_Using_CSV|
         +---------------------------+
         |[A, 1, Shubham]            |
         |[B, 2, Anurag]             |
         |[C, 3, Raman]              |
         |[D, 4, Rakesh]             |
         +---------------------------+
```

❖ Method 3: Using spark.read.format()

It is used to load text files into DataFrame. The .format() specifies the input data source format as “text”. The .load() loads data from a data source and returns DataFrame.


jupyter Read Text file into PySpark Dataframe (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [31]: df = spark.read.format("text").load(r"C:\Users\pmboc\Downloads\day14.txt")

In [32]: df.selectExpr("split(value, ' ') as Text_Data_In_Rows_Using_format_load").show(4

         +-----------------------------------+
         |Text_Data_In_Rows_Using_format_load|
         +-----------------------------------+
         |[A, 1, Shubham]                    |
         |[B, 2, Anurag]                     |
         |[C, 3, Raman]                      |
         |[D, 4, Rakesh]                     |
         +-----------------------------------+

In [ ]:
```