Akilesh K

k.akilesh123@gmail.com

Data engineering - Batch 1

Date: 09-02-24

DAY 14 - PYSPARK – selectExpr,Groupby,sort,drop,joins,union

selectExpr to split CSV

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

df = spark.read.csv("output1.txt")

df.selectExpr("split(_c0, ' ')\
as Text_Data_In_Rows_Using_CSV").show(4,False)
```

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

selectExpr to split text

```
In [53]: from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("DataFrame").getOrCreate()

df = spark.read.text("output1.txt")

df.selectExpr("split(value, ' ')as\
Text_Data_In_Rows_Using_Text").show(4,False)
```

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

Dataset from CSV

```
spark=SparkSession.builder.appName("Practice").getOrCreate()
df_pyspark= spark.read.csv("mark.csv",header=True, inferSchema=True)
df_pyspark.show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|         1|  95|Chennai|
|         2|  70|  Delhi|
|         3|  98| Mumbai|
|         4|  75|   Pune|
|         5|  89|  Kochi|
|         6|  69|Gwalior|
|         7|  52| Bhopal|
|         8|  54|Chennai|
|         9|  55|  Delhi|
|        10|  94| Mumbai|
|        11|  52|   Pune|
|        12|  90|  Kochi|
|        13|  92|Gwalior|
|        14|  55| Bhopal|
|        15|  97|   Pune|
|        16|  77|  Kochi|
|        17|  68|Gwalior|
|        18|  51| Bhopal|
|        19|  72|Chennai|
|        20|  79|  Delhi|
+----------+----+-------+
only showing top 20 rows
```

Group by sum

```
In [12]: df_pyspark.groupBy("City").sum("mark").show()
```

```
+-------+---------+
|   City|sum(mark)|
+-------+---------+
|  Kochi|     2366|
|Chennai|     2362|
| Mumbai|     2270|
|Gwalior|     2333|
|   Pune|     2958|
|  Delhi|     2440|
| Bhopal|     1836|
+-------+---------+
```

Group by average

```
In [13]: df_pyspark.groupBy("City").avg("mark").show()
```

```
+-------+-----------------+
|   City|        avg(mark)|
+-------+-----------------+
|  Kochi|          73.9375|
|Chennai|71.57575757575758|
| Mumbai|68.78787878787878|
|Gwalior|         72.90625|
|   Pune|72.14634146341463|
|  Delhi|73.93939393939394|
| Bhopal|65.57142857142857|
+-------+-----------------+
```

Group by min

```
In [14]: df_pyspark.groupBy("City").min("mark").show()
```

```
+-------+---------+
|   City|min(mark)|
+-------+---------+
|  Kochi|       45|
|Chennai|       40|
| Mumbai|       43|
|Gwalior|       40|
|   Pune|       42|
|  Delhi|       43|
| Bhopal|       40|
+-------+---------+
```

Group by max

```
In [15]: df_pyspark.groupBy("City").max("mark").show()
```

```
+-------+---------+
|   City|max(mark)|
+-------+---------+
|  Kochi|       99|
|Chennai|       98|
| Mumbai|       98|
|Gwalior|       98|
|   Pune|       99|
|  Delhi|       98|
| Bhopal|      100|
+-------+---------+
```

Group by mean

```
In [16]: df_pyspark.groupBy("City").mean("mark").show()
```

```
+-------+-----------------+
|   City|        avg(mark)|
+-------+-----------------+
|  Kochi|          73.9375|
|Chennai|71.57575757575758|
| Mumbai|68.78787878787878|
|Gwalior|         72.90625|
|   Pune|72.14634146341463|
|  Delhi|73.93939393939394|
| Bhopal|65.57142857142857|
+-------+-----------------+
```

Group by count

```
In [19]: df_pyspark.groupBy("City").count().show()
```

```
+-------+-----+
|   City|count|
+-------+-----+
|  Kochi|   32|
|Chennai|   33|
| Mumbai|   33|
|Gwalior|   32|
|   Pune|   41|
|  Delhi|   33|
| Bhopal|   28|
+-------+-----+
```

Sort by mark

```
In [27]: df_pyspark.sort("mark").show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|        96|  40|Chennai|
|       143|  40| Bhopal|
|       188|  40|Gwalior|
|       108|  41| Bhopal|
|        58|  42| Bhopal|
|       105|  42|   Pune|
|       201|  42|Chennai|
|        24|  43|  Delhi|
|        65|  43| Mumbai|
|       101|  43|Gwalior|
|       145|  43|  Delhi|
|       195|  43|  Delhi|
|        98|  44| Mumbai|
|        40|  44| Bhopal|
|       192|  44| Mumbai|
|        43|  45| Mumbai|
|        27|  45|  Kochi|
|       111|  45| Mumbai|
|       152|  45|  Kochi|
|       114|  46|Gwalior|
+----------+----+-------+
only showing top 20 rows
```

Sort by mark desc

```
In [28]: df_pyspark.sort(df_pyspark["mark"].desc()).show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|       150| 100| Bhopal|
|       148|  99|  Kochi|
|       228|  99|   Pune|
|         3|  98| Mumbai|
|        28|  98|Gwalior|
|       102|  98|Chennai|
|        61|  98|Gwalior|
|        63|  98|Chennai|
|        83|  98|Gwalior|
|       146|  98| Mumbai|
|       173|  98|  Delhi|
|        15|  97|   Pune|
|       112|  97|   Pune|
|        86|  97|  Delhi|
|       144|  97|Chennai|
```

Sort by mark and student Id

```
In [29]: df_pyspark.sort("mark","student_id").show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|        96|  40|Chennai|
|       143|  40| Bhopal|
|       188|  40|Gwalior|
|       108|  41| Bhopal|
|        58|  42| Bhopal|
|       105|  42|   Pune|
|       201|  42|Chennai|
|        24|  43|  Delhi|
|        65|  43| Mumbai|
|       101|  43|Gwalior|
```

Drop All null value

```
In [31]: df_pyspark.na.drop().show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|         1|  95|Chennai|
|         2|  70|  Delhi|
|         3|  98| Mumbai|
|         4|  75|   Pune|
|         5|  89|  Kochi|
|         6|  69|Gwalior|
|         7|  52| Bhopal|
|         8|  54|Chennai|
|         9|  55|  Delhi|
|        10|  94| Mumbai|
```

```
In [32]: df_pyspark.na.drop(how="all").show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|         1|  95|Chennai|
|         2|  70|  Delhi|
|         3|  98| Mumbai|
|         4|  75|   Pune|
|         5|  89|  Kochi|
|         6|  69|Gwalior|
|         7|  52| Bhopal|
|         8|  54|Chennai|
|         9|  55|  Delhi|
|        10|  94| Mumbai|
```

Drop any with two null value

```
In [34]: df_pyspark.na.drop(how="any",thresh=2).show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|         1|  95|Chennai|
|         2|  70|  Delhi|
|         3|  98| Mumbai|
|         4|  75|   Pune|
|         5|  89|  Kochi|
|         6|  69|Gwalior|
|         7|  52| Bhopal|
|         8|  54|Chennai|
|         9|  55|  Delhi|
|        10|  94| Mumbai|
```

```
In [35]: df_pyspark.na.drop(how="any",subset=["mark"]).show()
```

```
+----------+----+-------+
|Student_id|Mark|   City|
+----------+----+-------+
|         1|  95|Chennai|
|         2|  70|  Delhi|
|         3|  98| Mumbai|
|         4|  75|   Pune|
|         5|  89|  Kochi|
|         6|  69|Gwalior|
|         7|  52| Bhopal|
|         8|  54|Chennai|
|         9|  55|  Delhi|
|        10|  94| Mumbai|
```

Setting up dataframe

```python
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('sparkdf').getOrCreate()
data = [["1", "sravan", "company 1"],
        ["2", "ojaswi", "company 1"],
        ["3", "rohith", "company 2"],
        ["4", "sridevi", "company 1"],
        ["5", "bobby", "company 1"]]

columns = ['ID', 'NAME', 'Company']
dataframe = spark.createDataFrame(data, columns)
dataframe.show()

```

▸ (3) Spark Jobs

▸ dataframe: pyspark.sql.dataframe.DataFrame = [ID: string, NAME: string ... 1 more field]

```
+---+-------+---------+
| ID|   NAME|  Company|
+---+-------+---------+
|  1| sravan|company 1|
|  2| ojaswi|company 1|
|  3| rohith|company 2|
|  4|sridevi|company 1|
|  5|  bobby|company 1|
+---+-------+---------+
```

Add column with constant

```python
from pyspark.sql.functions import lit
dataframe.withColumn("salary", lit(34000)).show()
```

▸ (3) Spark Jobs

```
+---+-------+---------+------+
| ID|   NAME|  Company|salary|
+---+-------+---------+------+
|  1| sravan|company 1| 34000|
|  2| ojaswi|company 1| 34000|
|  3| rohith|company 2| 34000|
|  4|sridevi|company 1| 34000|
|  5|  bobby|company 1| 34000|
+---+-------+---------+------+
```

Command took 1.36 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:32:00 PM on My Cluster

Add column with respect to another column

```
dataframe.withColumn("salary", dataframe.ID*2300).show()
```

(3) Spark Jobs

```
+---+-------+---------+-------+
| ID|   NAME|  Company| salary|
+---+-------+---------+-------+
|  1| sravan|company 1| 2300.0|
|  2| ojaswi|company 1| 4600.0|
|  3| rohith|company 2| 6900.0|
|  4|sridevi|company 1| 9200.0|
|  5|  bobby|company 1|11500.0|
+---+-------+---------+-------+
```

Command took 1.37 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:32:33 PM on My Cluster

Add column by concat two other column values

Cmd 21

```
from pyspark.sql.functions import concat_ws
dataframe.withColumn("Details", concat_ws("-", "NAME", 'Company')).show()
```

(3) Spark Jobs

```
+---+-------+---------+-----------------+
| ID|   NAME|  Company|          Details|
+---+-------+---------+-----------------+
|  1| sravan|company 1| sravan-company 1|
|  2| ojaswi|company 1| ojaswi-company 1|
|  3| rohith|company 2| rohith-company 2|
|  4|sridevi|company 1|sridevi-company 1|
|  5|  bobby|company 1|  bobby-company 1|
+---+-------+---------+-----------------+
```

Command took 1.10 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:33:11 PM on My Cluster

Add column which has null value

```
from pyspark.sql.functions import concat_ws, lit
if 'salary' not in dataframe.columns:
    dataframe.withColumn("salary", lit(34000)).show()
```

(3) Spark Jobs

```
+---+-------+---------+------+
| ID|   NAME|  Company|salary|
+---+-------+---------+------+
|  1| sravan|company 1| 34000|
|  2| ojaswi|company 1| 34000|
|  3| rohith|company 2| 34000|
|  4|sridevi|company 1| 34000|
|  5|  bobby|company 1| 34000|
+---+-------+---------+------+
```

Command took 0.82 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:33:52 PM on My Cluster

Dataframe setup

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

emp = [(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)]
empColumns = ["emp_id","name","superior_emp_id","year_joined", "emp_dept_id","gender","salary"]

empDF = spark.createDataFrame(data=emp, schema = empColumns)
empDF.show()
```

(3) Spark Jobs

empDF: pyspark.sql.dataframe.DataFrame = [emp_id: long, name: string ... 5 more fields]

```
+------+--------+---------------+-----------+-----------+------+------+
|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|
+------+--------+---------------+-----------+-----------+------+------+
```

```python
dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
deptColumns = ["dept_name","dept_id"]
deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
deptDF.show()

```

(3) Spark Jobs

deptDF: pyspark.sql.dataframe.DataFrame = [dept_name: string, dept_id: long]

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

Command took 0.98 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:49:46 PM on My Cluster

Inner join

```python
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"inner") .show()
```

(3) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 4.46 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:51:14 PM on My Cluster

Outer join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"outer").show()
```

(3) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 2.36 seconds -- by kakilesh123@gmail.com at 2/9/2024, 2:57:19 PM on My Cluster

Left join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"left").show()
```

(6) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 3.03 seconds -- by kakilesh123@gmail.com at 2/9/2024, 3:04:22 PM on My Cluster

Right join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"right").show()
```

(6) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 2.06 seconds -- by kakilesh123@gmail.com at 2/9/2024, 3:04:38 PM on My Cluster

Leftsemi join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftsemi").show()
```

(3) Spark Jobs

```
+------+--------+---------------+-----------+-----------+------+------+
|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|
+------+--------+---------------+-----------+-----------+------+------+
```

Command took 2.25 seconds -- by kakilesh123@gmail.com at 2/9/2024, 3:34:03 PM on My Cluster

Leftanti join

```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftanti").show()
```

▸ (6) Spark Jobs

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

Command took 1.95 seconds -- by kakilesh123@gmail.com at 2/9/2024, 3:34:22 PM on My Cluster

Setting dataframe

```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
simpleData = [("James","Sales","NY",90000,34,10000), \
("Michael","Sales","NY",86000,56,20000), \
("Robert","Sales","CA",81000,30,23000), \
("Maria","Finance","CA",90000,24,23000) \
]
columns= ["employee_name","department","state","salary","age","bonus"]
df = spark.createDataFrame(data = simpleData, schema = columns)
df.show(truncate=False)
```

▸ (3) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [employee_name: string, department: string ... 4 more fields]

```
+-------------+----------+-----+------+---+-----+
|employee_name|department|state|salary|age|bonus|
+-------------+----------+-----+------+---+-----+
|James        |Sales     |NY   |90000 |34 |10000|
|Michael      |Sales     |NY   |86000 |56 |20000|
|Robert       |Sales     |CA   |81000 |30 |23000|
|Maria        |Finance   |CA   |90000 |24 |23000|
+-------------+----------+-----+------+---+-----+
```

Command took 1.06 seconds -- by kakilesh123@gmail.com at 2/9/2024, 4:16:15 PM on My Cluster

```python
simpleData2 = [("James","Sales","NY",90000,34,10000), \
("Maria","Finance","CA",90000,24,23000), \
("Jen","Finance","NY",79000,53,15000), \
("Jeff","Marketing","CA",80000,25,18000), \
("Kumar","Marketing","NY",91000,50,21000) \
]
columns2= ["employee_name","department","state","salary","age","bonus"]
df2 = spark.createDataFrame(data = simpleData2, schema = columns2)
df2.show(truncate=False)
```

▸ (3) Spark Jobs

▸ df2: pyspark.sql.dataframe.DataFrame = [employee_name: string, department: string ... 4 more fields]

```
+-------------+----------+-----+------+---+-----+
|employee_name|department|state|salary|age|bonus|
+-------------+----------+-----+------+---+-----+
|James        |Sales     |NY   |90000 |34 |10000|
|Maria        |Finance   |CA   |90000 |24 |23000|
|Jen          |Finance   |NY   |79000 |53 |15000|
|Jeff         |Marketing |CA   |80000 |25 |18000|
|Kumar        |Marketing |NY   |91000 |50 |21000|
+-------------+----------+-----+------+---+-----+
```

Command took 0.73 seconds -- by kakilesh123@gmail.com at 2/9/2024, 4:17:00 PM on My Cluster

Union

```python
unionDF = df.union(df2)
unionDF.show(truncate=False)
```

▸ (3) Spark Jobs

▸ unionDF: pyspark.sql.dataframe.DataFrame = [employee_name: string, department: string ... 4 more

```
+-------------+----------+-----+------+---+-----+
|employee_name|department|state|salary|age|bonus|
+-------------+----------+-----+------+---+-----+
|James        |Sales     |NY   |90000 |34 |10000|
|Michael      |Sales     |NY   |86000 |56 |20000|
|Robert       |Sales     |CA   |81000 |30 |23000|
|Maria        |Finance   |CA   |90000 |24 |23000|
|James        |Sales     |NY   |90000 |34 |10000|
|Maria        |Finance   |CA   |90000 |24 |23000|
|Jen          |Finance   |NY   |79000 |53 |15000|
|Jeff         |Marketing |CA   |80000 |25 |18000|
|Kumar        |Marketing |NY   |91000 |50 |21000|
+-------------+----------+-----+------+---+-----+
```

Command took 1.52 seconds -- by kakilesh123@gmail.com at 2/9/2024, 4:41:28 PM on My Cluster

Union distinct

```
disDF = df.union(df2).distinct()
disDF.show(truncate=False)
```

(2) Spark Jobs

disDF: pyspark.sql.dataframe.DataFrame = [employee_name: string, department: string ...

```
+-------------+----------+-----+------+---+-----+
|employee_name|department|state|salary|age|bonus|
+-------------+----------+-----+------+---+-----+
|James        |Sales     |NY   |90000 |34 |10000|
|Michael      |Sales     |NY   |86000 |56 |20000|
|Robert       |Sales     |CA   |81000 |30 |23000|
|Maria        |Finance   |CA   |90000 |24 |23000|
|Jen          |Finance   |NY   |79000 |53 |15000|
|Jeff         |Marketing |CA   |80000 |25 |18000|
|Kumar        |Marketing |NY   |91000 |50 |21000|
+-------------+----------+-----+------+---+-----+
```

Command took 2.00 seconds -- by kakilesh123@gmail.com at 2/9/2024, 4:55:48 PM on My Cluster

Union All

```
unionAllDF = df.unionAll(df2)
unionAllDF.show(truncate=False)
```

(3) Spark Jobs

unionAllDF: pyspark.sql.dataframe.DataFrame = [employee_name: string, department: string ... 4 more f

```
+-------------+----------+-----+------+---+-----+
|employee_name|department|state|salary|age|bonus|
+-------------+----------+-----+------+---+-----+
|James        |Sales     |NY   |90000 |34 |10000|
|Michael      |Sales     |NY   |86000 |56 |20000|
|Robert       |Sales     |CA   |81000 |30 |23000|
|Maria        |Finance   |CA   |90000 |24 |23000|
|James        |Sales     |NY   |90000 |34 |10000|
|Maria        |Finance   |CA   |90000 |24 |23000|
|Jen          |Finance   |NY   |79000 |53 |15000|
|Jeff         |Marketing |CA   |80000 |25 |18000|
|Kumar        |Marketing |NY   |91000 |50 |21000|
+-------------+----------+-----+------+---+-----+
```

Command took 1.14 seconds -- by kakilesh123@gmail.com at 2/9/2024, 4:58:09 PM on My Cluster