Assignment – Day 14

-Sarthak Niranjan Kulkarni (Maverick)

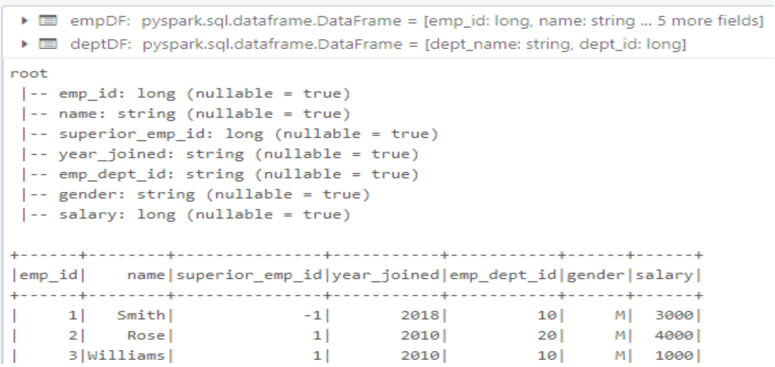
- sarthakkul2311@gmail.com - (+91) 93256 02791

21/11/2024 (Thursday)

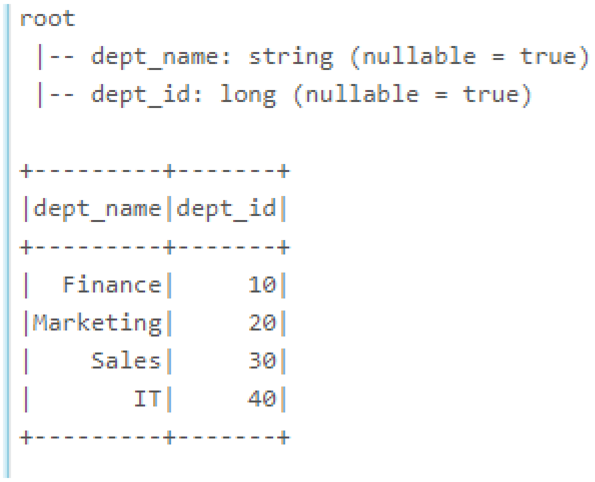
Joins in Spark Practice:-

1. **Creating and displaying PySpark DataFrames with employee and department data:-**

```
from pyspark.sql import SparkSession
# Initialize SparkSession
spark = SparkSession.builder \
.appName("example") \
.getOrCreate()
# Data
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,
"Sarthak", 2, "2010","23","",-1)]
empColumns = ["emp_id","name","superior_emp_id","year_joined",
"emp_dept_id","gender","salary"]
empDF = spark.createDataFrame(data=emp, schema = empColumns)
empDF.printSchema()
empDF.show()
dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
deptColumns = ["dept_name","dept_id"]
deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
deptDF.printSchema()
deptDF.show()
```



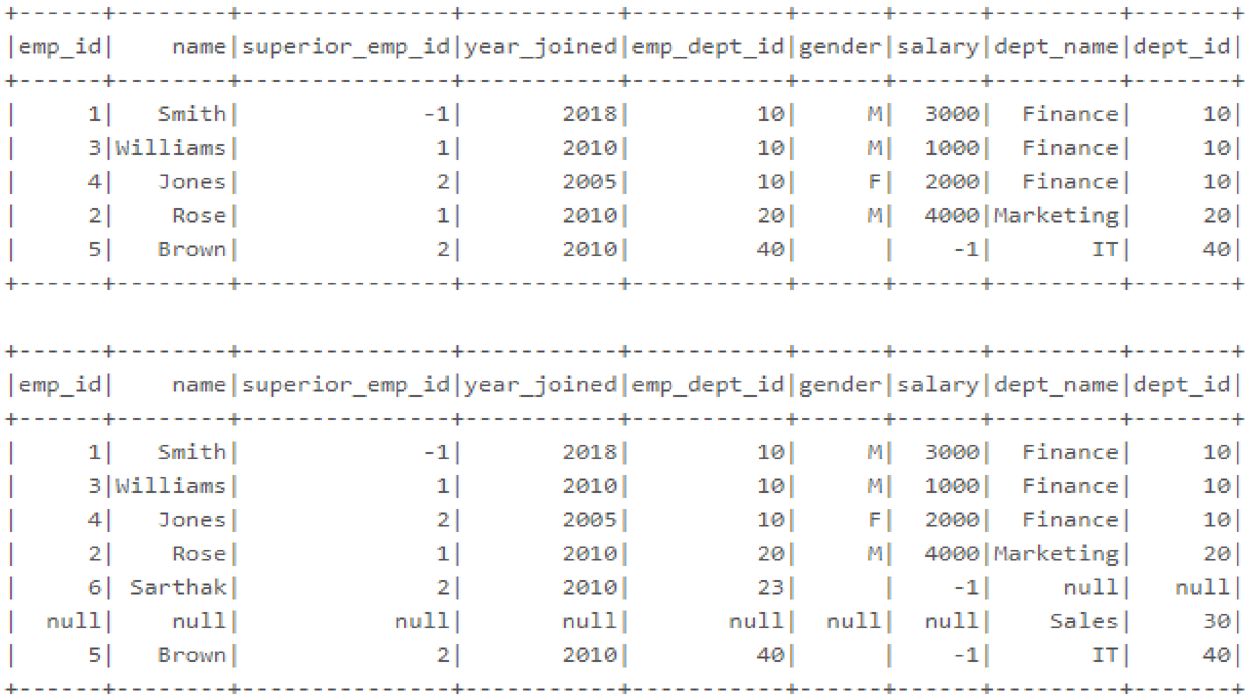
```
▸ empDF:  pyspark.sql.dataframe.DataFrame = [emp_id: long, name: string ... 5 more fields]
▸ deptDF:  pyspark.sql.dataframe.DataFrame = [dept_name: string, dept_id: long]
root
 |-- emp_id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- superior_emp_id: long (nullable = true)
 |-- 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|
```



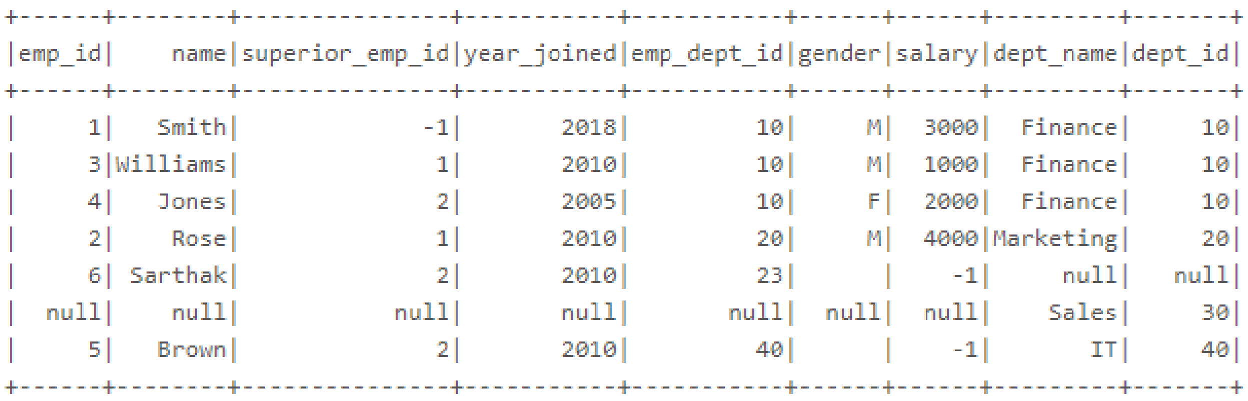
```
root
 |-- dept_name: string (nullable = true)
 |-- dept_id: long (nullable = true)

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

2. **Performing inner, outer, and full joins between employee and department DataFrames in PySpark.**

```
#Inner join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"inner").show()
#outer join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"outer").show()
#full join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"full").show()
```



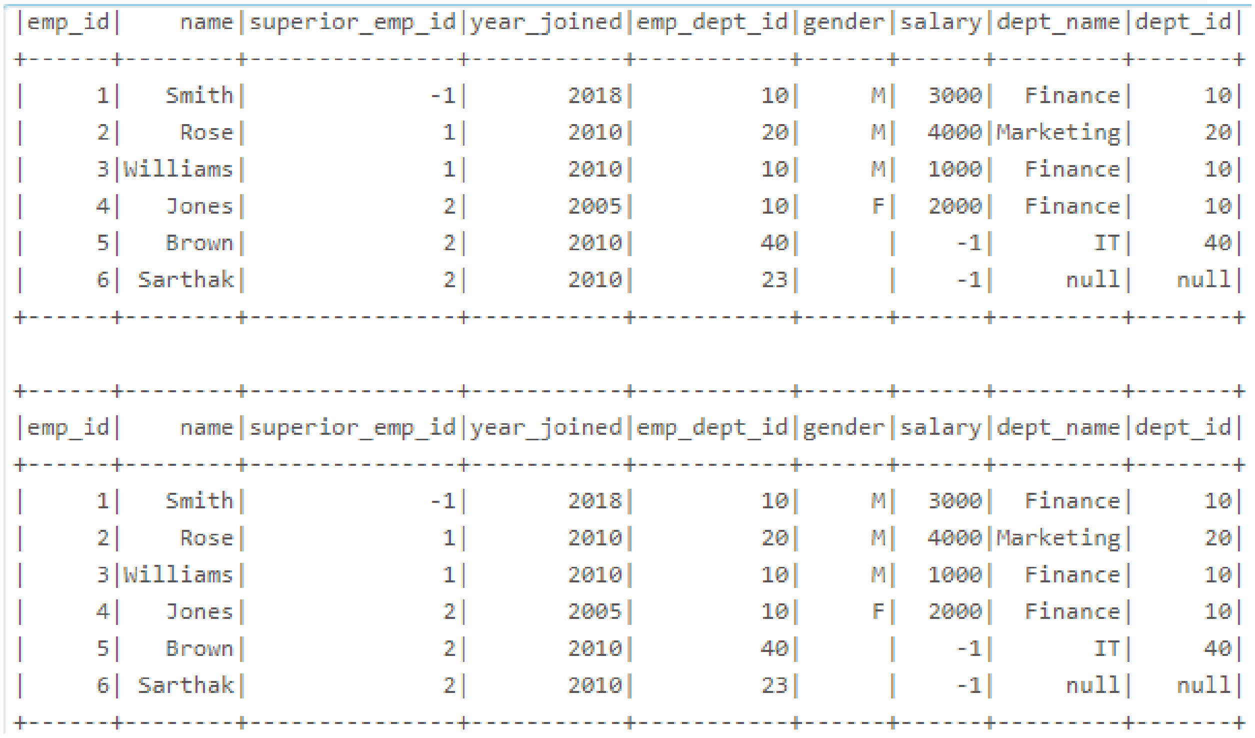
```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -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|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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|
|     6| Sarthak|              2|       2010|         23|      |    -1|     null|   null|
|  null|    null|           null|       null|       null|  null|  null|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```



```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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|
|     6| Sarthak|              2|       2010|         23|      |    -1|     null|   null|
|  null|    null|           null|       null|       null|  null|  null|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

3. Performing left and left outer joins between employee and department DataFrames in PySpark.

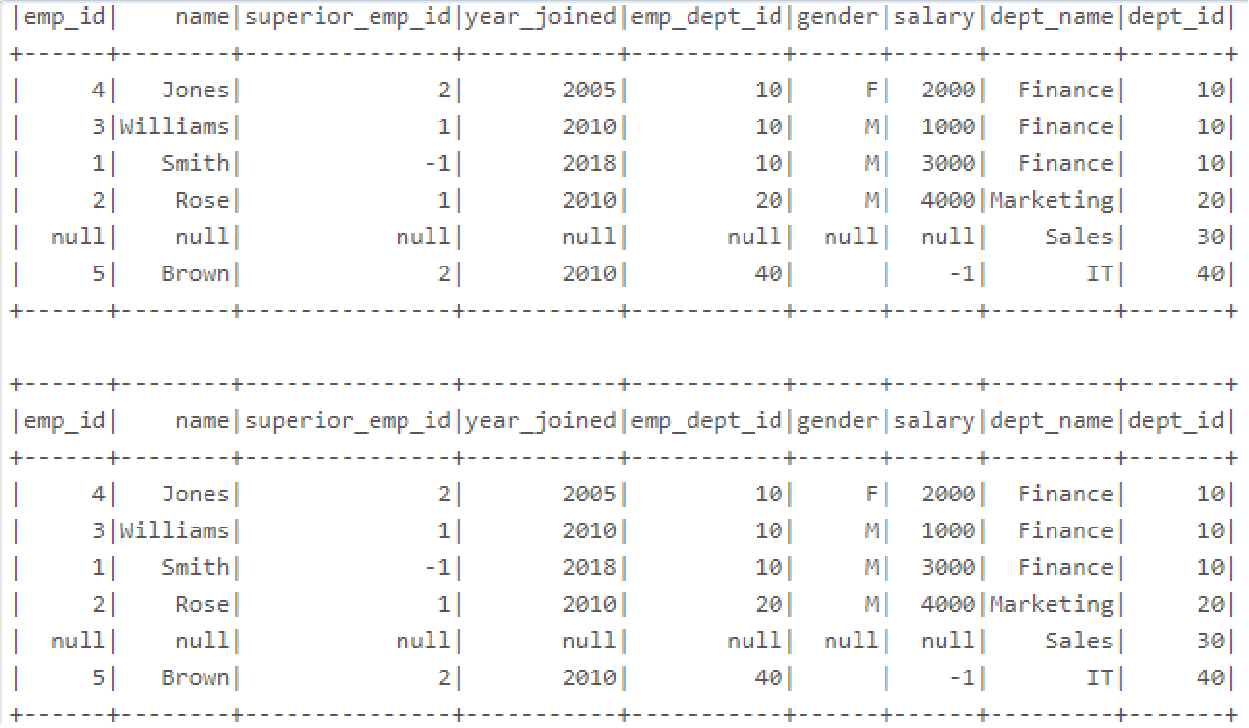
```
#Left join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"left").show()
#Left Outer join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"leftouter").show()
```



```
|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| Sarthak|              2|       2010|         23|      |    -1|     null|   null|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|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| Sarthak|              2|       2010|         23|      |    -1|     null|   null|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

4. **Performing right and right outer joins between employee and department DataFrames in PySpark.**

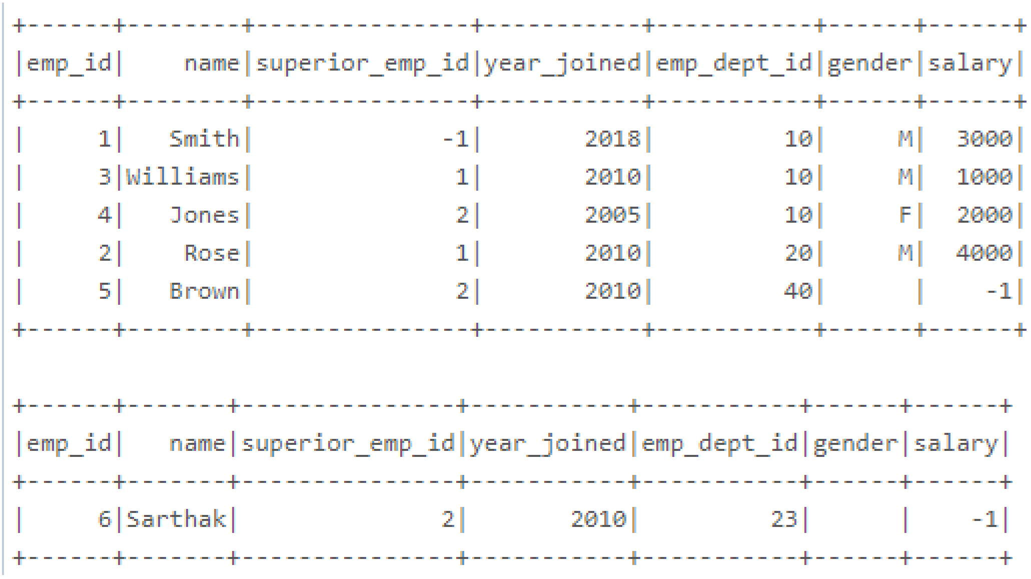
```
#right join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"right").show()
#right outer join
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,
"rightouter").show()
```



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


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

5. **Performing left semi and left anti joins between employee and department DataFrames in PySpark.**

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

#left anti
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|
+------+--------+---------------+-----------+-----------+------+------+
|     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|
+------+--------+---------------+-----------+-----------+------+------+

+------+-------+---------------+-----------+-----------+------+------+
|emp_id|   name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+-------+---------------+-----------+-----------+------+------+
|     6|Sarthak|              2|       2010|         23|      |    -1|
+------+-------+---------------+-----------+-----------+------+------+
```

Joins in Spark Summary:-

The above codes demonstrates the creation of two PySpark DataFrames: empDF containing employee data and deptDF containing department data. It showcases various types of joins to combine the two DataFrames based on the common key emp_dept_id in empDF and dept_id in deptDF.

1. **Inner Join** returns rows where there is a match in both DataFrames.
2. **Outer Join** (or Full Join) includes all rows from both DataFrames, with null values for non-matching rows.
3. **Left and Right Joins** (and their outer variants) return all rows from one DataFrame and matching rows (if any) from the other.

Additionally, **Left Semi Join** filters rows in empDF that have a match in deptDF, while **Left Anti Join** returns rows in empDF that do not match with deptDF.