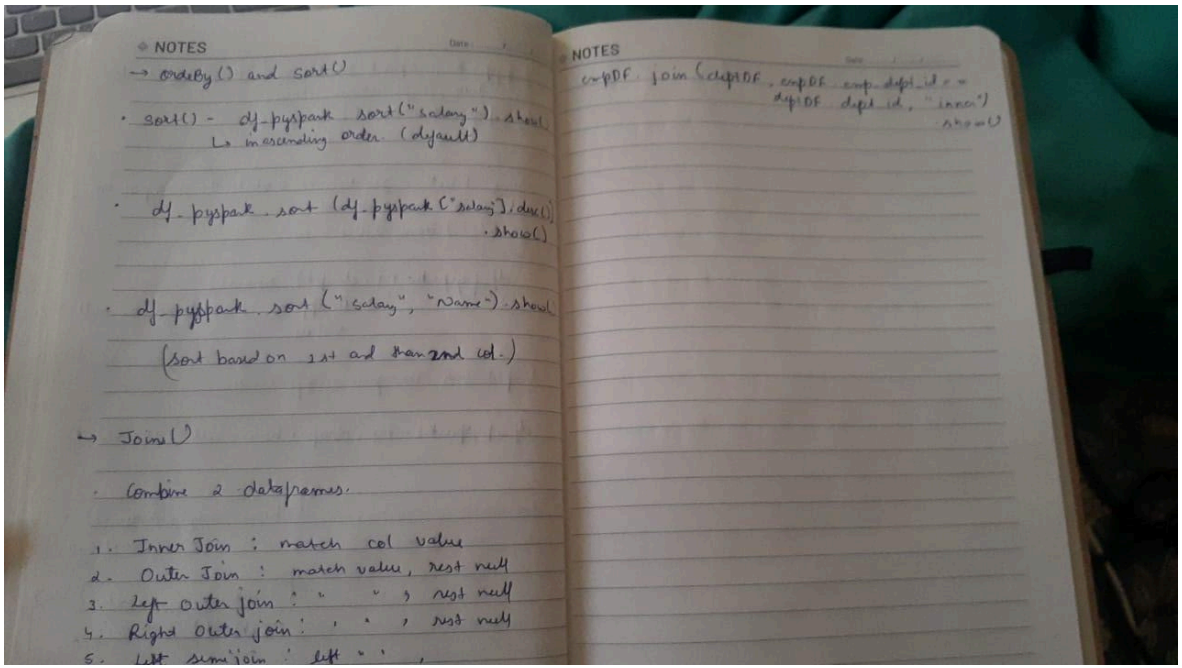
Asmita Porwal
Batch-1
Day-14
9/2/2024
Data engineering

Assignment-14

Hand written notes during the session :


NOTES

→ orderBy() and sort()

- sort() - df_pyspark.sort("salary").show()
  ↳ in ascending order (default)

- df_pyspark.sort(df_pyspark["salary"].desc()).show()

- df_pyspark.sort("salary", "Name").show()
  (sort based on 1st and then 2nd col.)

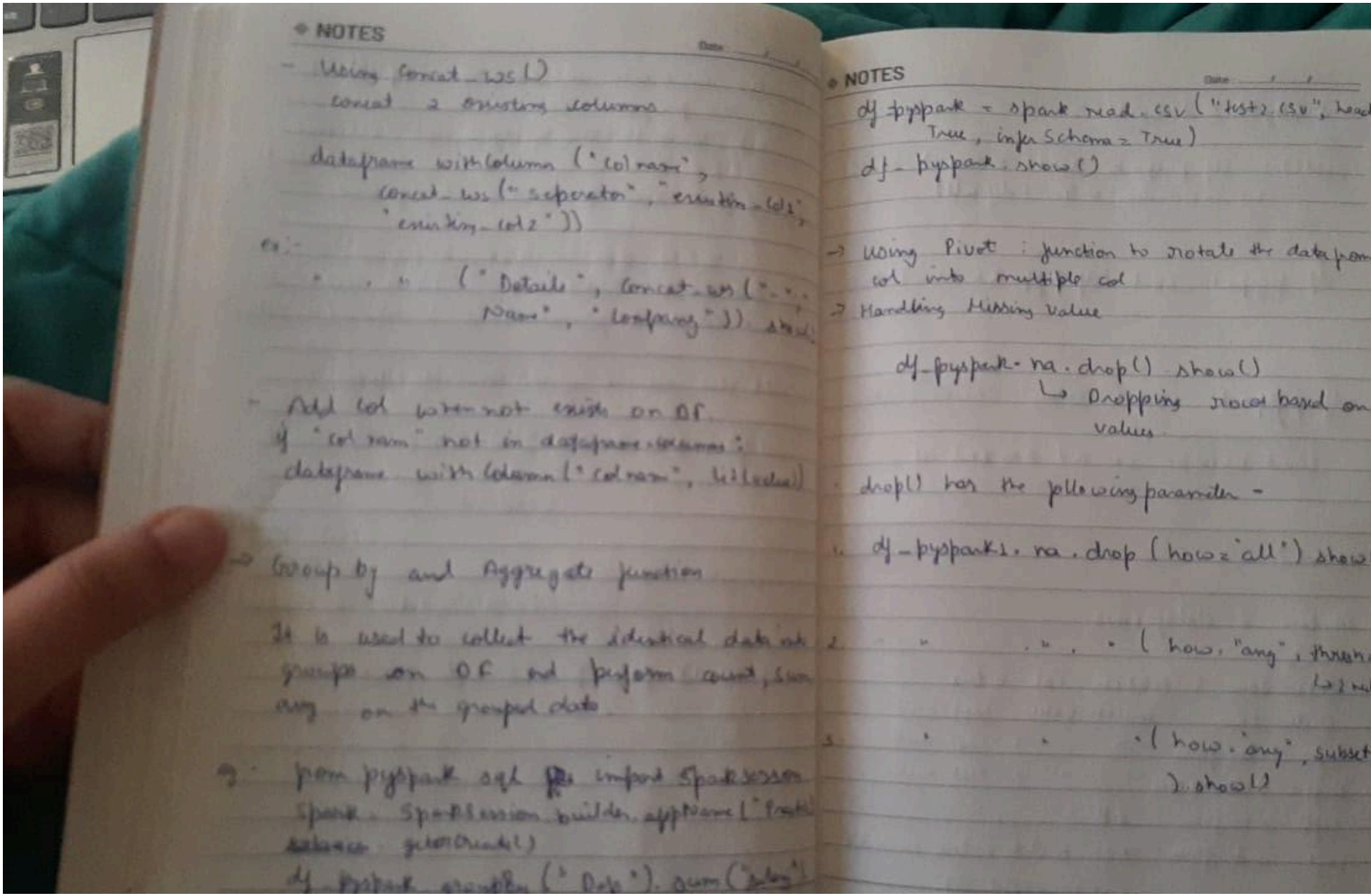
→ Joins()

- Combine 2 dataframes.

1. Inner Join : match col value
2. Outer Join : match value, rest null
3. Left Outer join : " " , rest null
4. Right Outer join : " " , rest null
5. Left semi join : left " " ,

NOTES

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



NOTES

- Using Concat_ws()
  Concat 2 existing columns

  dataframe.withColumn("col_name",
  concat_ws("seperator", "existing_col1",
  "existing_col2"))

ex:-

" " " ("Details", Concat_ws("-",
Name", "Location")).show()

- Add col when not exist on DF
if "col_name" not in dataframe.columns:
dataframe.withColumn("col_name", lit(value))

→ Group by and Aggregate function

It is used to collect the identical data into groups on DF and perform count, sum, avg on the grouped data

from pyspark.sql import SparkSession
Spark = SparkSession.builder.appName("Practice").getOrCreate()
df_pyspark.groupBy("Date").sum("Sales").show()

NOTES

df_pyspark = spark.read.csv("test2.csv", header=True, inferSchema=True)
df_pyspark.show()

→ Using Pivot : function to rotate the dataframe col into multiple col

→ Handling Missing value

df_pyspark.na.drop().show()
↳ Dropping rows based on values

drop() has the following parameter -

1. df_pyspark1.na.drop(how="all").show()

2. " " " " (how="any", threshold=2)

3. " " " (how="any", subset=[])
.show()

NOTES Date : / /

patterns using partitions

- Processing JSON and CSV data with pyspark
- ETL (Extract, Transform, Load) with pyspark

- spark.read.format("text").load(path=None, format=None, schema=None, **options)
  - parquet, orc, avro are file formats apart from csv, textfile we have
  - paths, format (default parquet), schema, options
  - return dataframe
- spark.read.format("text").load("output.txt") it converts text file into data frame
  - when it runs on the pyspark notebook, it stores as RDD

eg:- from pyspark.sql import SparkSession (it imports spark session of library from pyspark.sql module)

NOTES Date : / /

(inorder to create a session of spark to run the program, initialization starts from "SparkSession.builder.getOrCreate()" command

df = spark.read.format("text").load('outp

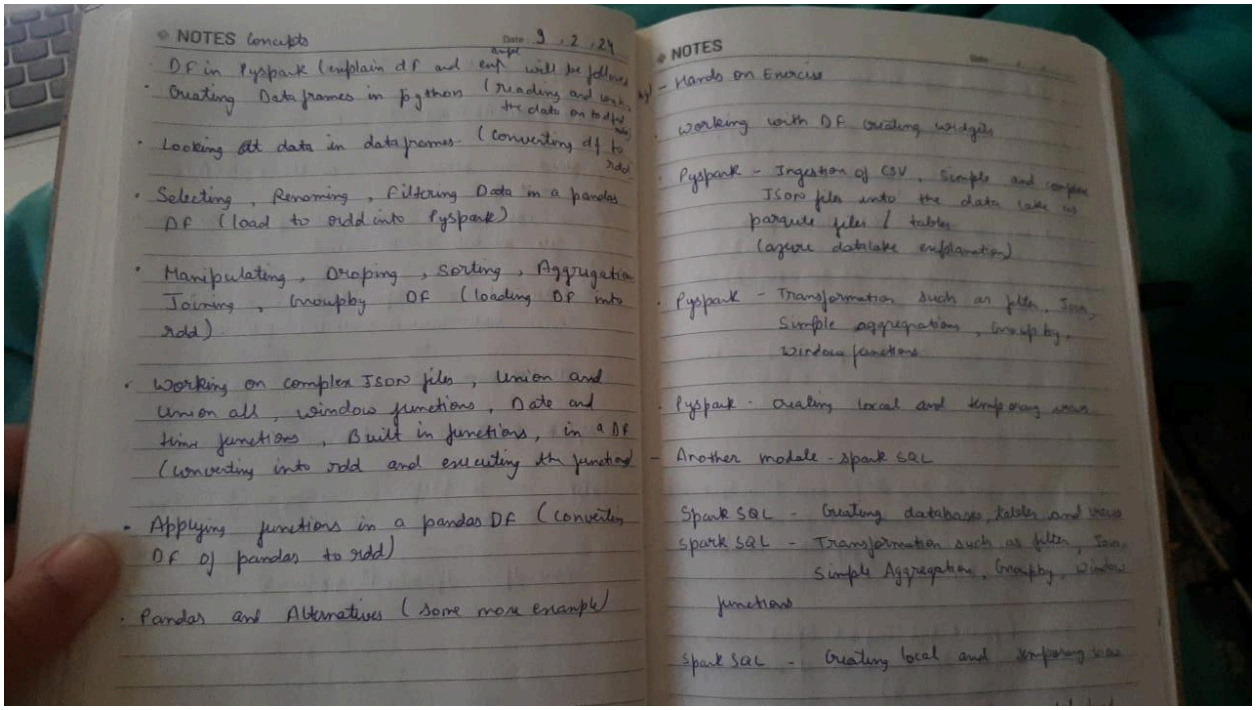
- df.selectExpr("split(value, ' ') as Text_D In_Rows_Using_format_load").show(4

→ Add new column with constant value

dataframe.withColumn("col name", lit(valu literal value Val ar

eg:- .withColumn('salary', lit(3400))

import lit from pyspark.sql.function



Groupby , joins , aggregate function

In [1]:
```
from pyspark.sql import SparkSession
```

In [2]:
```
import pyspark
import findspark
findspark.init()
```

In [3]:
```
from pyspark import SparkContext
sc = SparkContext("local", "RDD Transformation")
sc
```

Out[3]:

SparkContext

Spark UI

Version
`v3.5.0`
Master
`local`
AppName
`RDD Transformation`

In [6]:
```
spark =SparkSession.builder.appName("Practice").getOrCreate()
```

In [8]:
```
df_pyspark= spark.read.csv("E:\downloads\Marks_data.csv",header=True,inferSchema=True)
```

In [8]:
```
df_pyspark= spark.read.csv("E:\downloads\Marks_data.csv",header=True,inferSchema=True)
df_pyspark.show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
|Joey|      85|      98| 21|
|Abhi|      54|      79| 20|
+----+--------+--------+---+
```

In [9]:
```
df_pyspark.groupBy("age").sum("M2 Score").show()
```

```
+---+-------------+
|age|sum(M2 Score)|
+---+-------------+
| 20|          159|
| 19|           56|
| 21|           98|
+---+-------------+
```

In [10]:
```
df_pyspark.groupBy("age").min("M2 Score").show()
```

```
+---+-------------+
|age|min(M2 Score)|
+---+-------------+
```

```
|age|min(M2 Score)|
+---+-------------+
| 20|           79|
| 19|           56|
| 21|           98|
+---+-------------+
```

In [11]:
```
df_pyspark.groupBy("age").max("M2 Score").show()
```

```
+---+-------------+
|age|max(M2 Score)|
+---+-------------+
| 20|           80|
| 19|           56|
| 21|           98|
+---+-------------+
```

In [12]:
```
df_pyspark.groupBy("age").mean("M2 Score").show()
```

```
+---+-------------+
|age|avg(M2 Score)|
+---+-------------+
| 20|         79.5|
| 19|         56.0|
| 21|         98.0|
+---+-------------+
```

In [13]:
```
df_pyspark.groupBy("age").avg("M2 Score").show()
```

```
+---+-------------+
|age|avg(M2 Score)|
+---+-------------+
| 20|         79.5|
| 19|         56.0|
| 21|         98.0|
+---+-------------+
```

In [14]:
```
df_pyspark.groupBy("age").count().show()
```

```
+---+-----+
|age|count|
+---+-----+
| 20|    2|
| 19|    1|
| 21|    1|
+---+-----+
```

In [15]:
```
df_pyspark.groupBy("Name","age").sum("M2 Score").show()
```

```
+----+---+-------------+
|Name|age|sum(M2 Score)|
+----+---+-------------+
|Alex| 20|           80|
|Brad| 19|           56|
```

In [16]:
```
df_pyspark.groupBy("age").agg(({"M2 Score":"sum"})).show()
```

```
+---+-------------+
|age|sum(M2 Score)|
+---+-------------+
| 20|          159|
| 19|           56|
| 21|           98|
+---+-------------+
```

In [17]:
```
df_pyspark.agg(({"M1 Score":"sum"})).show()
```

```
+-------------+
|sum(M1 Score)|
+-------------+
|          246|
+-------------+
```

In [18]:
```
df_pyspark.groupBy("age").pivot("Name").sum("M2 Score").show()
```

```
+---+----+----+----+----+
|age|Abhi|Alex|Brad|Joey|
+---+----+----+----+----+
| 20|  79|  80|NULL|NULL|
| 19|NULL|NULL|  56|NULL|
| 21|NULL|NULL|NULL|  98|
+---+----+----+----+----+
```

In [19]:
```
df_pyspark.na.drop(how="all").show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
|Joey|      85|      98| 21|
|Abhi|      54|      79| 20|
+----+--------+--------+---+
```

In [20]:
```
df_pyspark.na.drop(how="any",thresh=2).show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
|Joey|      85|      98| 21|
|Abhi|      54|      79| 20|
+----+--------+--------+---+
```

In [21]:
```
df_pyspark.na.drop(how="any",subset=["M2 Score"]).show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
```


```
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
|Joey|      85|      98| 21|
|Abhi|      54|      79| 20|
+----+--------+--------+---+
```


In [22]:
```
df_pyspark.sort("age").show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Brad|      45|      56| 19|
|Alex|      62|      80| 20|
|Abhi|      54|      79| 20|
|Joey|      85|      98| 21|
+----+--------+--------+---+
```

In [23]:
```
df_pyspark.sort(df_pyspark["age"].desc()).show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Joey|      85|      98| 21|
|Alex|      62|      80| 20|
|Abhi|      54|      79| 20|
|Brad|      45|      56| 19|
+----+--------+--------+---+
```

In [24]:
```
df_pyspark.sort("age","Name").show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Brad|      45|      56| 19|
|Abhi|      54|      79| 20|
|Alex|      62|      80| 20|
|Joey|      85|      98| 21|
+----+--------+--------+---+
```

In [25]:
```
df_pyspark.orderBy("age").show()
```

```
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
+----+--------+--------+---+
|Brad|      45|      56| 19|
|Alex|      62|      80| 20|
|Abhi|      54|      79| 20|
|Joey|      85|      98| 21|
+----+--------+--------+---+
```

In [26]:
```
emp = [(1,"Smith",-1,"2018","10","M",3000),(2, "Rose",1 , "2010", "20","M", 4000),(3,"Williams",1,"2010","10","M",1000),(4, "Jone
empColumns = ["emp_id","name","superior_emp_id","year_joined", "emp_dept_id","gender","salary"]
```

In [27]:
```
empDF = spark.createDataFrame(data=emp, schema = empColumns)
empDF.printSchema()
```

```
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)
```

In [28]:
```
empDF.show()
```

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

In [29]:
```
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()
```

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

In [31]:
```
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 [32]:
```
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 [33]:
```
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 [34]:
```
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|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
|     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 [35]:
```
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|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
|     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 [36]:
```
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|
|  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

In [37]:
```
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|
```

In [38]:
```
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 [39]:
```
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|
+------+-----+---------------+-----------+-----------+------+------+
```