Asmita Porwal Data Engineering Batch-1 12/02/2024

# **Coding Challenge -3 Question-1**

### **Execute Manipulating**

```
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

RDD Transformation
```

# **GroupeBy DataFrames & Aggregations**

#### SUM

### **Min and Max**

### **Count and Sum two columns**

## Aggregations using agg sum

### And pivot element

# **Droping**

```
In [51]: df_pyspark.na.drop(how="all").show()
             | Alex | 62 | 80 | 20 | |
| Brad | 45 | 56 | 19 |
| Joey | 85 | 98 | 21 |
| Abhi | 54 | 79 | 20 |
            |Brad|
|Joey|
            |Abhi|
In [52]: df_pyspark.na.drop(how="any",thresh=2).show()
            |Name|M1 Score|M2 Score|age|
            Alex
                       62 |
45 |
                                       56 19
             |Brad|
                       85 |
54 |
            Abhil
                                       79 20
  In [53]: df_pyspark.na.drop(how="any",subset=["M2 Score"]).show()
               |Name|M1 Score|M2 Score|age|
              |Alex|
|Brad|
                                        80 | 20 |
56 | 19 |
                                        98 | 21 |
79 | 20 |
```

## **Sorting**

# Sort by age in ascending order and descending order

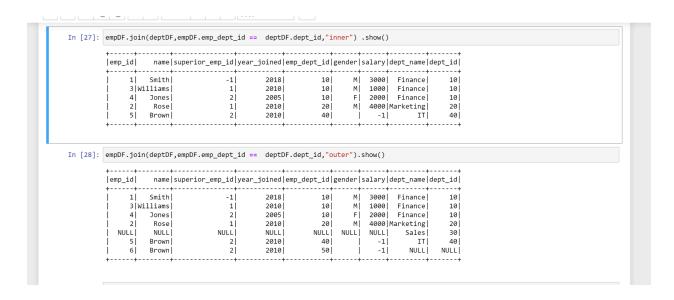
```
In [54]: df_pyspark.sort("age").show()
          |Name|M1 Score|M2 Score|age|
                                56| 19|
          Brad
          |Abhi|
                      541
                                79 | 20 |
          |Joey|
In [55]: df_pyspark.sort(df_pyspark["age"].desc()).show()
          |Name|M1 Score|M2 Score|age|
                           98 | 21 |
80 | 20 |
          |Joey|
|Alex|
                      85 l
          İΔhhiİ
                                79 20
          |Brad|
                                56 19
```

# Sorting 2 columns at a time

# **Joining**

```
In [25]: empDF.show()
             |emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
                    1 Smith
                                                                  2018
                                                                                                     3000
                     3 Williams
                                                                                                M 1000 F 2000
                                                                  2010
                                                                                     101
                     5
                           Brown
                                                                  2010
                                                                                     40
                                                                                                       -1
In [26]:
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()
              |-- dept_name: string (nullable = true)
|-- dept_id: long (nullable = true)
             |dept_name|dept_id|
              Finance
                                  10
              |Marketing|
                                  201
```

### Inner join and outer join



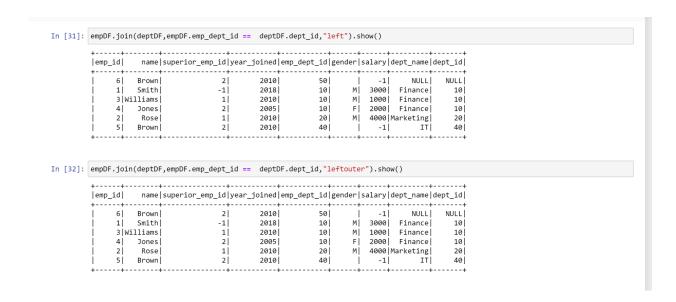
#### **Full Join**

```
In [29]: empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"full").show()
       2018
           3|Williams|
                                    2010
                                              10
                                                    Μİ
                                                      1000
                                                            Financel
                                                                     101
                                    2005
                                            10 |
20 |
                                                      2000
                                                            Finance
              Jones
                                                                     10|
                                    2010
                                                       4000 | Marketing |
                                                                     20
         NULLI
               NULLI
                           NULLI
                                   NULL
                                            NULLI NULLI
                                                      NULL
                                                             Sales
                                                                     30 l
                                    2010
                                              40
                                                               IΤ
                                                                     40
              Brown
```

#### Full outer join

```
In [30]: 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
                                             2018
                                                                     30001
              3|Williams|
                                    11
                                             2010
                                                          10
                                                                     1000
                                                                            Financel
                                                                                        10
             4| Jones|
                                             2005
                                                          10
                                                                     2000
                                                                            Finance
                                                                                        10
                                                                     4000|Marketing|
                                             2010
                                                          20
           NULL
                  NULL
                                  NULL|
                                             NULL
                                                        NULL| NULL|
                                                                     NULL
                                                                              Sales
                                                                                        301
                                                                                IT
                                             2010
                                                                                        40
             5
                  Brown
                                                          40
                                                                       -1
```

#### Left and left outer



#### Right and Right outer

```
In [33]: empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"right").show()
                2|
1|
-1|
NULL|
                                                                               41
                                Jones
                         3|Williams|
                                                                                                                                                             10
                                              -1|
NULL|
1|
2|
                                                                                           1| Smith|
NULL| NULL|
                                                                                2018|
NULL|
                                                                                                                                                            10
                         2 |
                                   Rosel
                                                                                2010
                                                                                                                                                            201
                         5
                                                                                                                                                            40
                                                                                2010
                                 Brown
In [34]: empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"rightouter").show()
                |\verb|emp_id|| \qquad |\verb|name|| superior_emp_id|| year_joined|| emp_dept_id|| gender|| salary|| dept_name|| dept_id||

        4
        Jones
        2
        2005
        10
        F
        2000
        Finance

        3
        Williams
        1
        2010
        10
        M
        1000
        Finance

        1
        Smith
        -1
        2018
        10
        M
        3000
        Finance

        NULL
        NULL
        NULL
        NULL
        NULL
        NULL
        NULL
        Sales

        2
        Rose
        1
        2010
        20
        M
        4000
        Marketing

        5
        Brown
        2
        2010
        40
        -1
        IT

                                                                                                                                                            10
                                                                                                                                                            10
                                                                                                                                                         20
```

#### Left semi and left anti

```
In [35]: 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|
                                                 10|
                                                        M| 3000|
                                       2018
            11
                Smithl
            3|Williams|
                                                        M 1000
                                       2010
                                                  10| F| 2001,
20| M| 4000|
-1| -1|
              Jones
                                       2005
                                       2010
                Rose
                Brown
In [36]: 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|
```