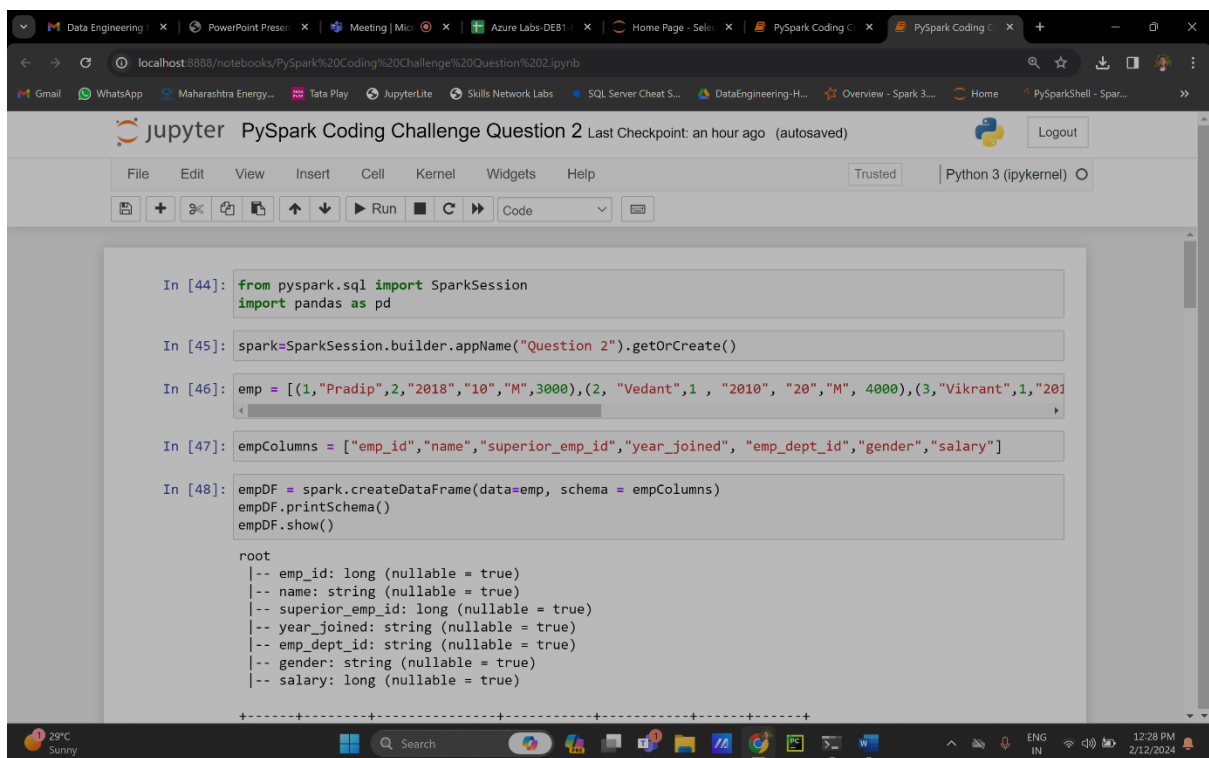


PySpark Coding Challenge Question 2

🚀 Execute PySpark -sparksql joins & Applying Functions in a Pandas DataFrame



The screenshot shows a Jupyter Notebook titled "PySpark Coding Challenge Question 2" with a last checkpoint of "an hour ago (autosaved)". The notebook is running on Python 3 (ipykernel). The code in the notebook is as follows:

```
In [44]: from pyspark.sql import SparkSession
import pandas as pd

In [45]: spark=SparkSession.builder.appName("Question 2").getOrCreate()

In [46]: emp = [(1,"Pradip",2,"2018", "10", "M", 3000), (2, "Vedant", 1, "2010", "20", "M", 4000), (3, "Vikrant", 1, "2010", "20", "M", 4000)]

In [47]: empColumns = ["emp_id", "name", "superior_emp_id", "year_joined", "emp_dept_id", "gender", "salary"]

In [48]: 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)
```

PySpark Coding Challenge Question 2 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

```
emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
-----+-----+-----+-----+-----+-----+-----+
1| Pradip| 2| 2018| 10| M| 3000|
2| Vedant| 1| 2010| 20| M| 4000|
3| Vikrant| 1| 2010| 10| M| 1000|
4| Rutuja| 2| 2005| 10| F| 2000|
5| Vinay| 2| 2010| 40| | 3000|
6| Abhishek| 2| 2010| 50| | 1000|
```

```
In [49]: dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
deptColumns = ["dept_name","dept_id"]

In [50]: deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
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|
```

PySpark Coding Challenge Question 2 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

```
In [51]: #inner join
print("Inner Join:")
inner_join=empDF.join(deptDF,emp_dept_id == deptDF.dept_id,"inner").show()

Inner Join:
emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
-----+-----+-----+-----+-----+-----+-----+-----+
1| Pradip| 2| 2018| 10| M| 3000| Finance| 10|
3| Vikrant| 1| 2010| 10| M| 1000| Finance| 10|
4| Rutuja| 2| 2005| 10| F| 2000| Finance| 10|
2| Vedant| 1| 2010| 20| M| 4000| Marketing| 20|
5| Vinay| 2| 2010| 40| | 3000| IT| 40|
```

```
In [52]: #outer join
print("Outer Join:")
outer_join=empDF.join(deptDF,emp_dept_id == deptDF.dept_id,"outer").show()

Outer Join:
emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
-----+-----+-----+-----+-----+-----+-----+-----+
1| Pradip| 2| 2018| 10| M| 3000| Finance| 10|
3| Vikrant| 1| 2010| 10| M| 1000| Finance| 10|
4| Rutuja| 2| 2005| 10| F| 2000| Finance| 10|
2| Vedant| 1| 2010| 20| M| 4000| Marketing| 20|
NULL| NULL| NULL| NULL| NULL| NULL| NULL| Sales| 30|
5| Vinay| 2| 2010| 40| | 3000| IT| 40|
6| Abhishek| 2| 2010| 50| | 1000| NULL| NULL|
```

PySpark Coding Challenge Question 2 Last Checkpoint: an hour ago (autosaved)

In [53]:

```
#Left join
print("Left Join:")
left_join=empDF.join(deptDF,emp_dept_id == deptDF.dept_id,"left").show()
```

Left Join:

emp_id	name	superior_emp_id	year_joined	emp_dept_id	gender	salary	dept_name	dept_id
1	Pradip	2	2018	10	M	3000	Finance	10
2	Vedant	1	2010	20	M	4000	Marketing	20
3	Vikrant	1	2010	10	M	1000	Finance	10
4	Rutuja	2	2005	10	F	2000	Finance	10
5	Vinay	2	2010	40		3000	IT	40
6	Abhishek	2	2010	50		1000	NULL	NULL

In [54]:

```
#right join
print("Right Join:")
right_join=empDF.join(deptDF,emp_dept_id == deptDF.dept_id,"right").show()
```

Right Join:

emp_id	name	superior_emp_id	year_joined	emp_dept_id	gender	salary	dept_name	dept_id
4	Rutuja	2	2005	10	F	2000	Finance	10
3	Vikrant	1	2010	10	M	1000	Finance	10
1	Pradip	2	2018	10	M	3000	Finance	10
2	Vedant	1	2010	20	M	4000	Marketing	20
NULL	NULL	NULL	NULL	NULL	NULL	NULL	Sales	30
5	Vinay	2	2010	40		3000	IT	40

PySpark Coding Challenge Question 2 Last Checkpoint: an hour ago (autosaved)

In [55]:

```
#Leftsemi Join
print("Leftsemi Join:")
leftsemi_join=empDF.join(deptDF,emp_dept_id == deptDF.dept_id,"leftsemi").show()
```

Leftsemi Join:

emp_id	name	superior_emp_id	year_joined	emp_dept_id	gender	salary
1	Pradip	2	2018	10	M	3000
3	Vikrant	1	2010	10	M	1000
4	Rutuja	2	2005	10	F	2000
2	Vedant	1	2010	20	M	4000
5	Vinay	2	2010	40		3000

In [56]:

```
#Left Anti Join
print("Left Anti Join:")
left_anti_join=empDF.join(deptDF,emp_dept_id == deptDF.dept_id,"leftanti").show()
```

Left Anti Join:

emp_id	name	superior_emp_id	year_joined	emp_dept_id	gender	salary
6	Abhishek	2	2010	50		1000

In [57]:

```
inner_join_pd = empDF.join(deptDF, emp_dept_id == deptDF.dept_id, "inner").toPandas()
```

localhost:8888/notebooks/PySpark%20Coding%20Challenge%20Question%202.ipynb

Jupyter PySpark Coding Challenge Question 2 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

```
In [57]: inner_join_pd = empDF.join(deptDF, empDF.emp_dept_id == deptDF.dept_id, "inner").toPandas()

In [58]: # Defining function to apply pandas dataframe
def custom_function(row):
    return f"{row['name']} - {row['dept_name']}"

# Applying custom function to pandas dataframe
inner_join_pd['employee_department'] = inner_join_pd.apply(custom_function, axis=1)

print("Inner join dataframe with applied function:")
print(inner_join_pd)
```

Inner join dataframe with applied function:

	emp_id	name	superior_emp_id	year_joined	emp_dept_id	gender	salary	
0	1	Pradip	2	2018	10	M	3000	
1	3	Vikrant	1	2010	10	M	1000	
2	4	Rutuja	2	2005	10	F	2000	
3	2	Vedant	1	2010	20	M	4000	
4	5	Vinay	2	2010	40		3000	

	dept_name	dept_id	employee_department
0	Finance	10	Pradip - Finance
1	Finance	10	Vikrant - Finance
2	Finance	10	Rutuja - Finance
3	Marketing	20	Vedant - Marketing
4	IT	40	Vinay - IT

29°C Sunny 12:30 PM 2/12/2024