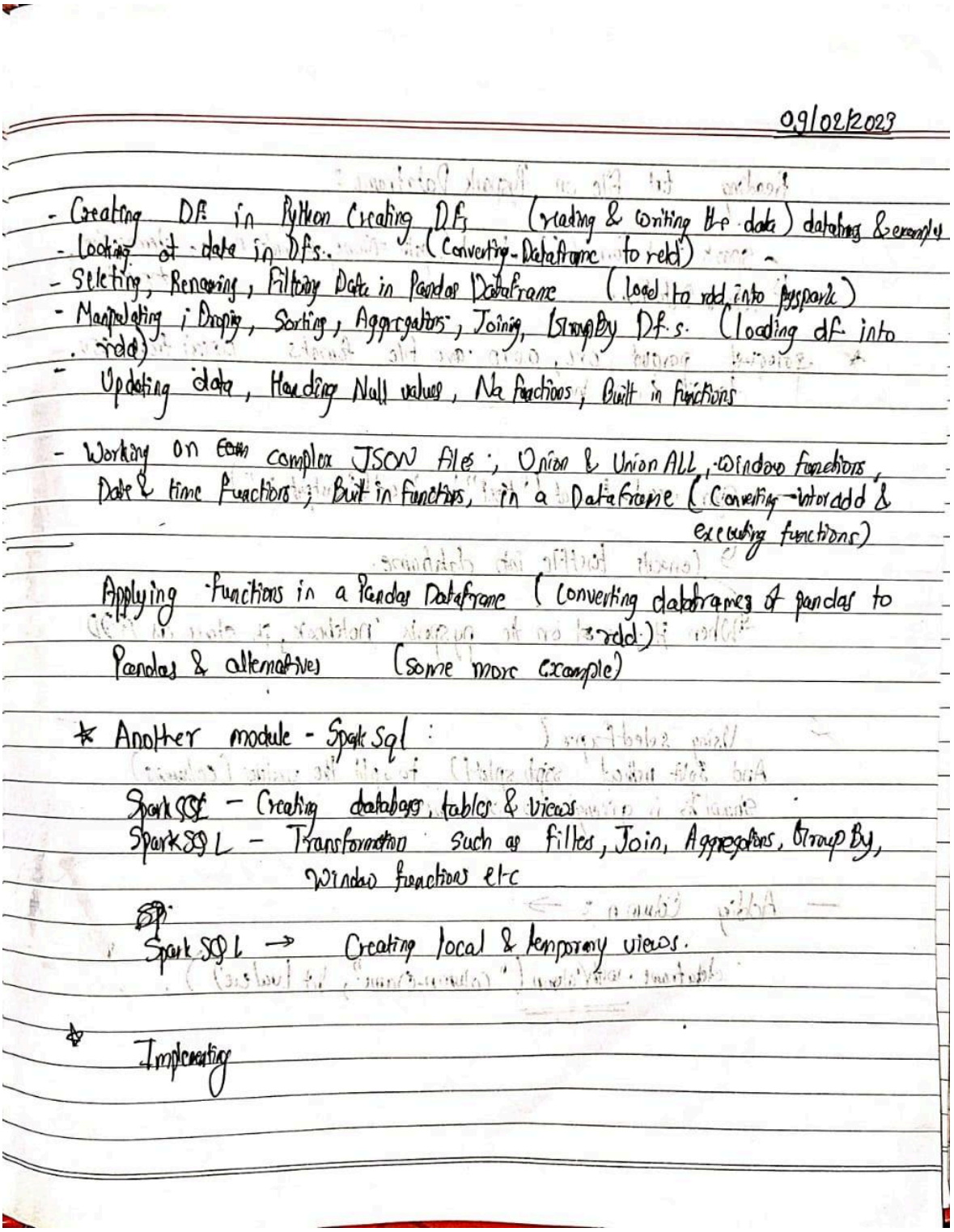
Name : Parth Nandedkar
Date : 09 Feb 2024
Topics : PySpark
Batch : Data Engineering Batch-1

Handwritten Notes :



Hands-on-Exercise :→ Working with Dataframes Creating tables

* Pyspark - Ingestion of CSV, simple & complex JSON files into the data lake as parquet files / tables.

Pyspark - Transformation as such as filter, join, simple aggregation, Groupby, Window functions

Pyspark - Creating local and temporary views Hands-on.

Reading txt file on Pyspark Dataframe:

- spark.read.format().load(path=None, format=None, schema=None, **options)
  - Returns Dataframe

* parquet parqud, orc, avro are file formats apart from csv, textfile we have

- spark.read.format("text").load("output.tx")
  - Converts textfile into dataframe.
  - When it runs on the pyspark notebook, it stores as RDD

- Using selectExpr(
And Split method split() to split the value (columns)
Should a is giving no. of rows,

- Adding Column: →

dataframe.withColumn("column-name", lit(value));

- Concatenating :

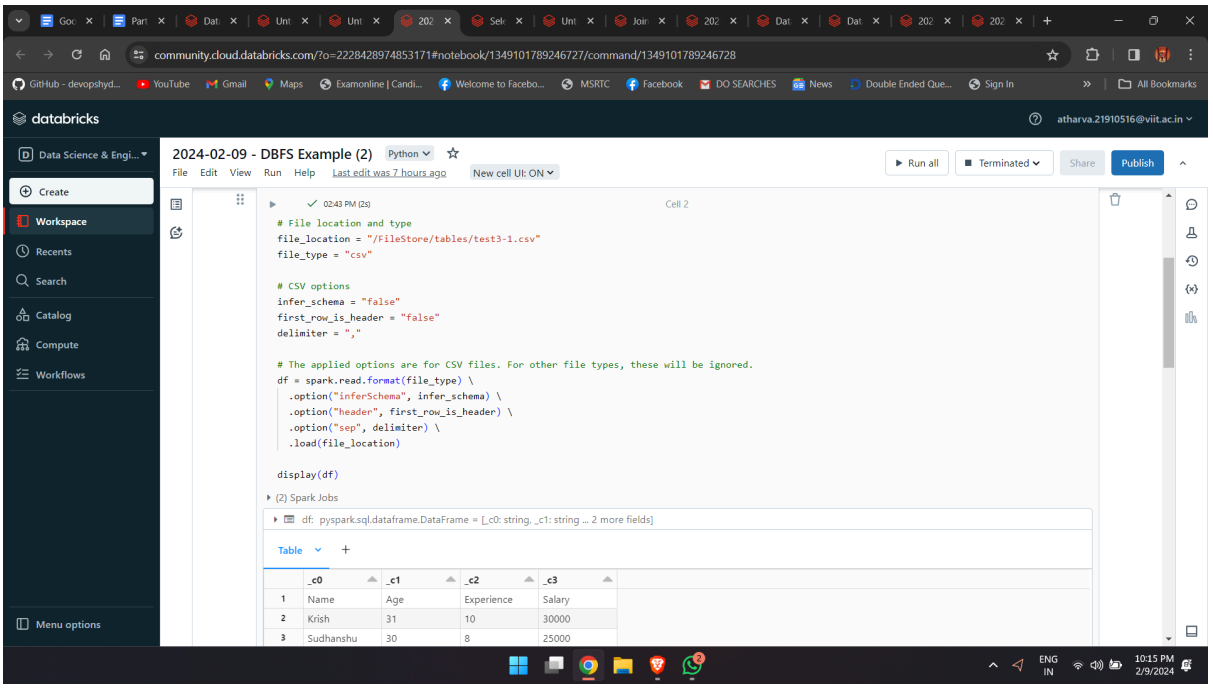
dataframe.withColumn("Column_name", concat_ws("separator", "existing1", "existing2"))

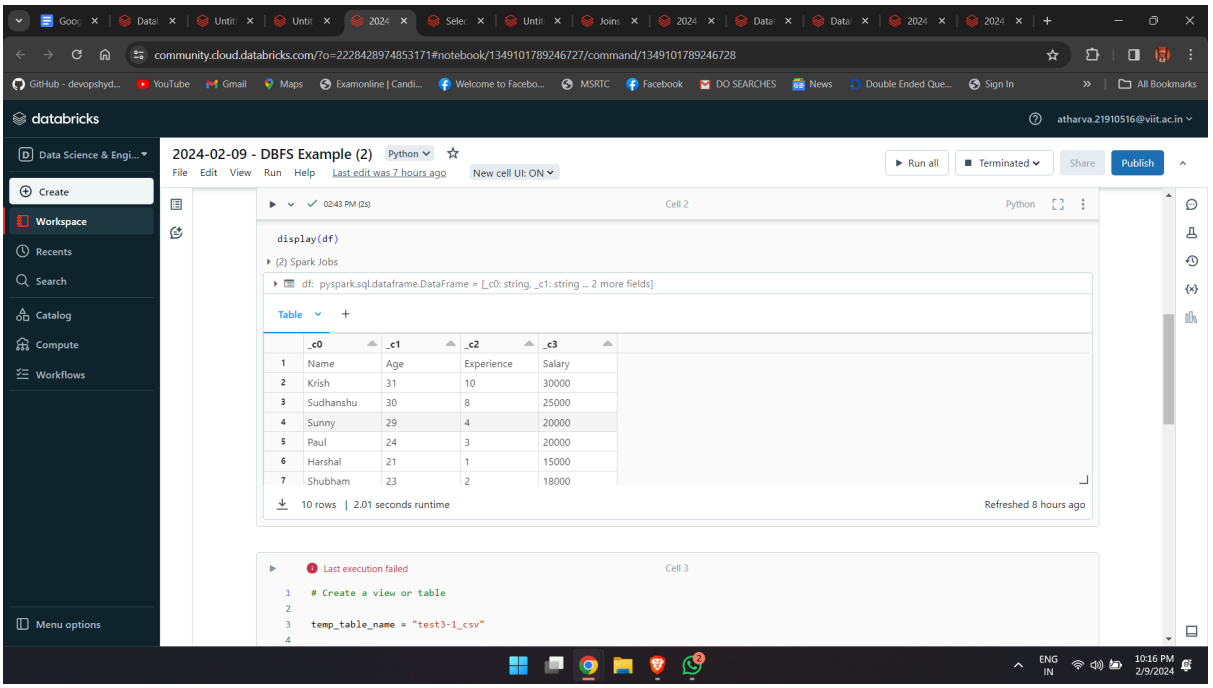
- Group by Aggregations :

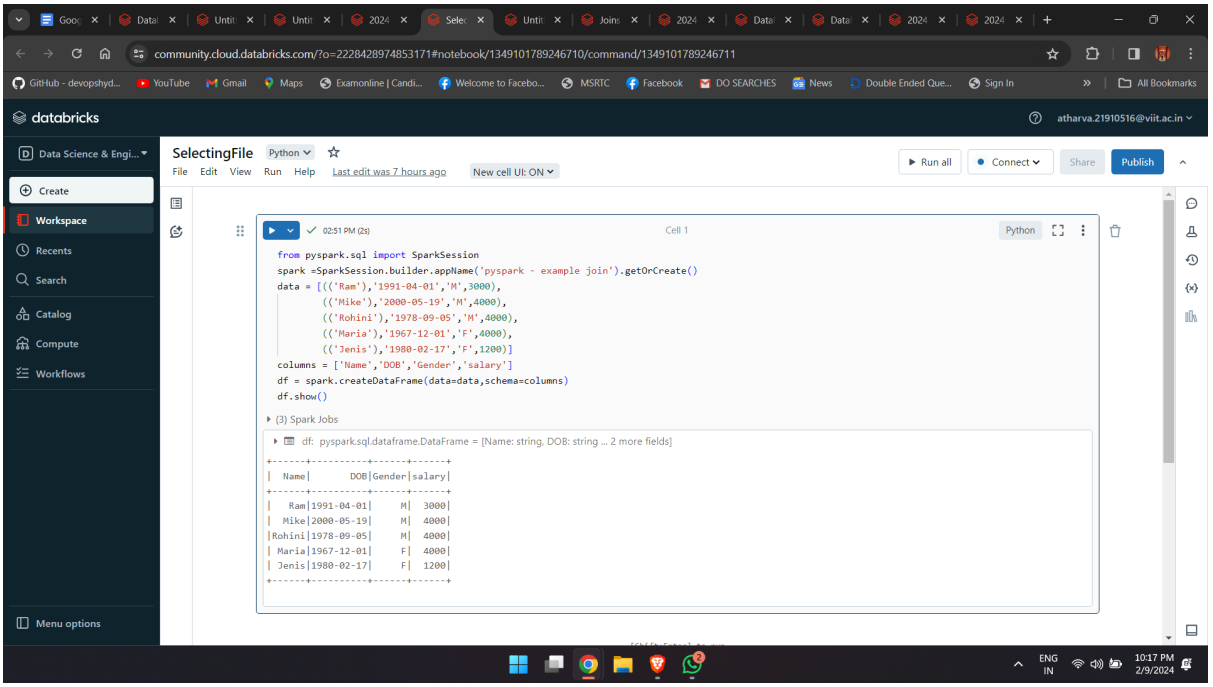
- Afternoon Session :→

Joins →
LEFT
RIGHT
Outer Join
Left semi Join
Right semi Join
Inner Join.
Left anti join.
Right anti join.

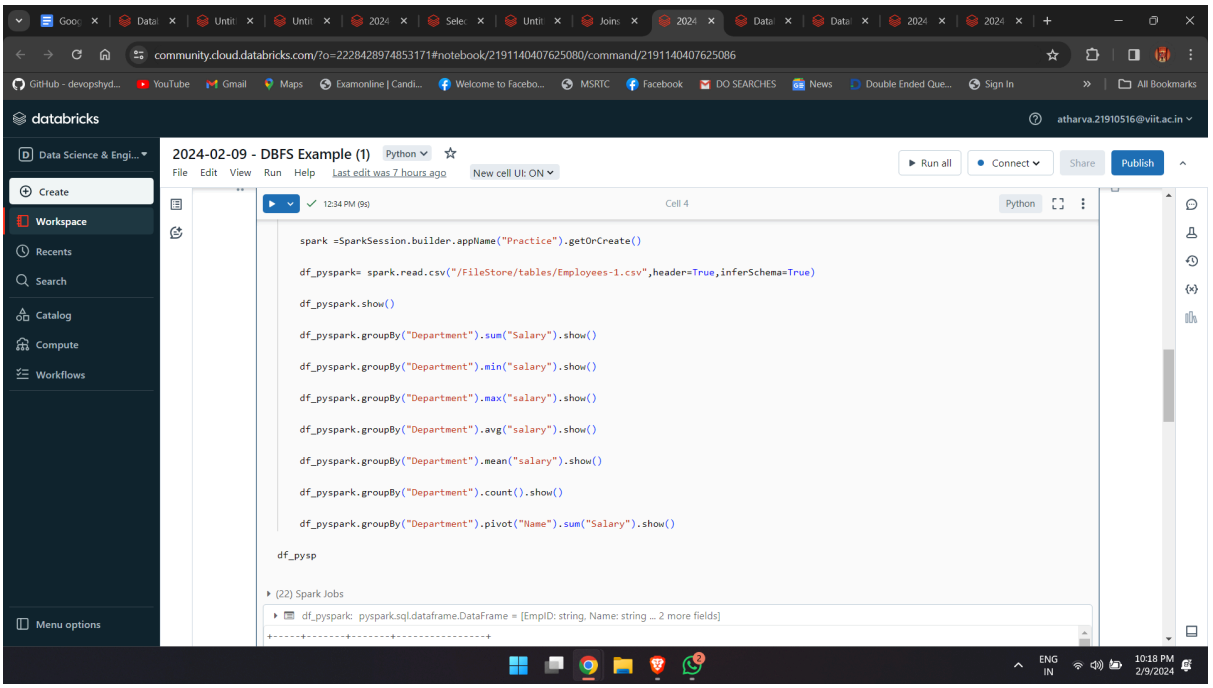
Select from CSV file :

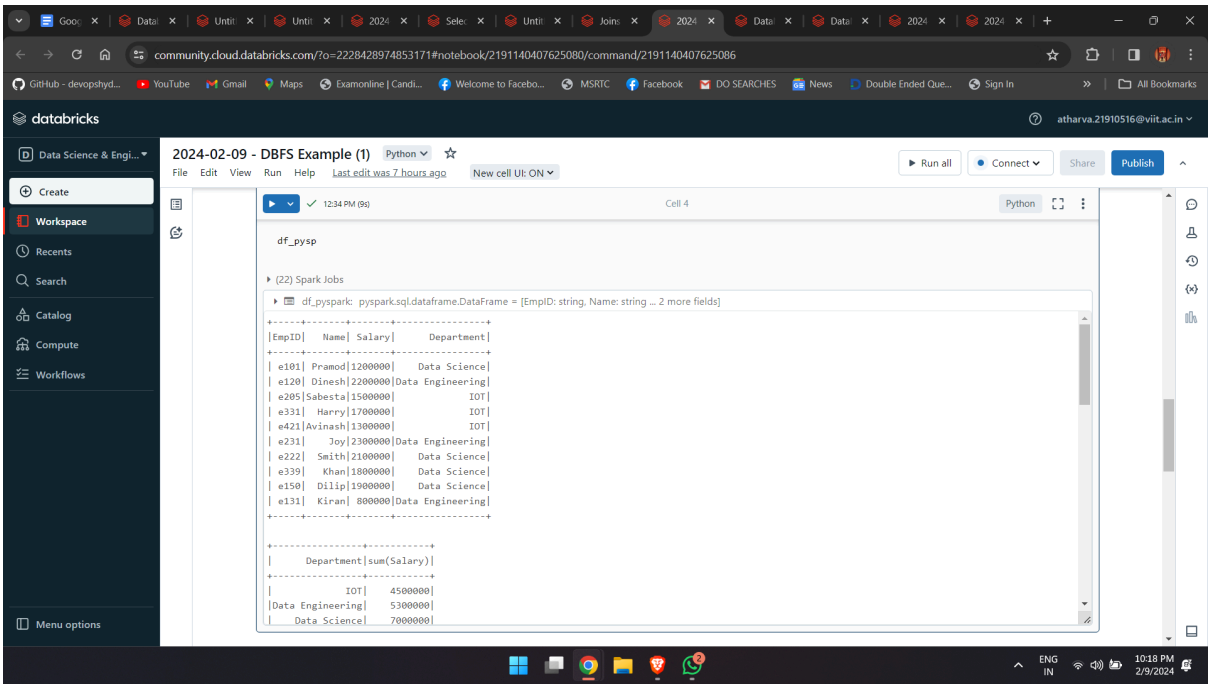


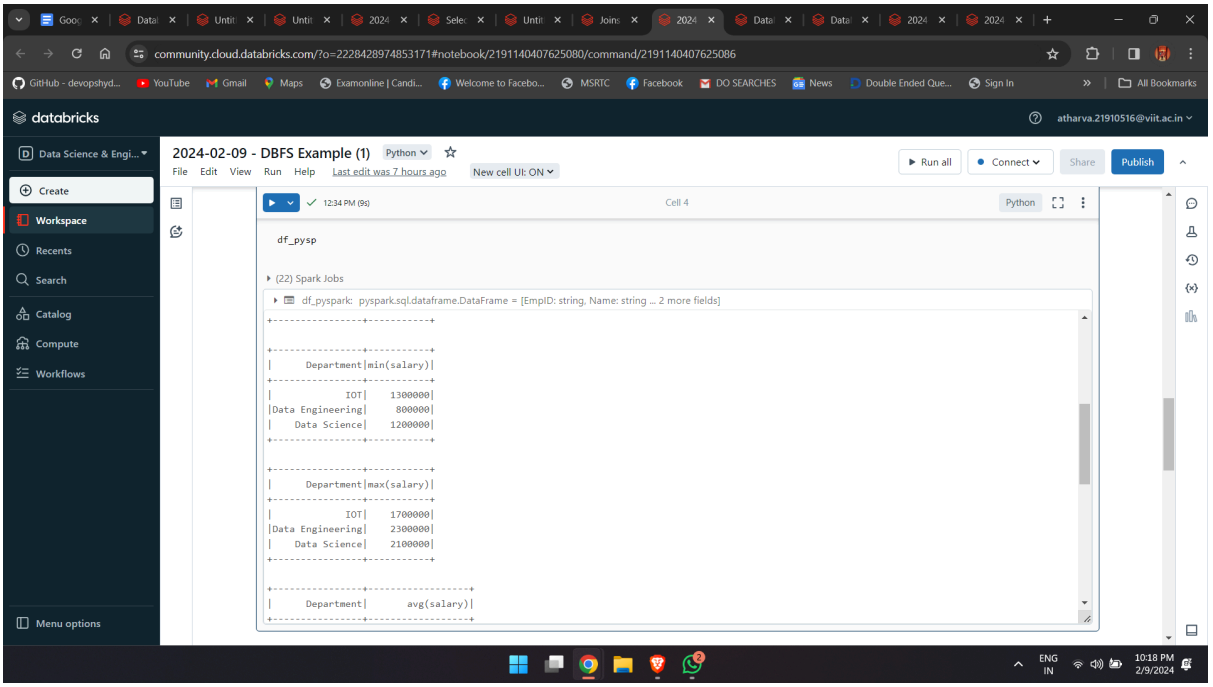


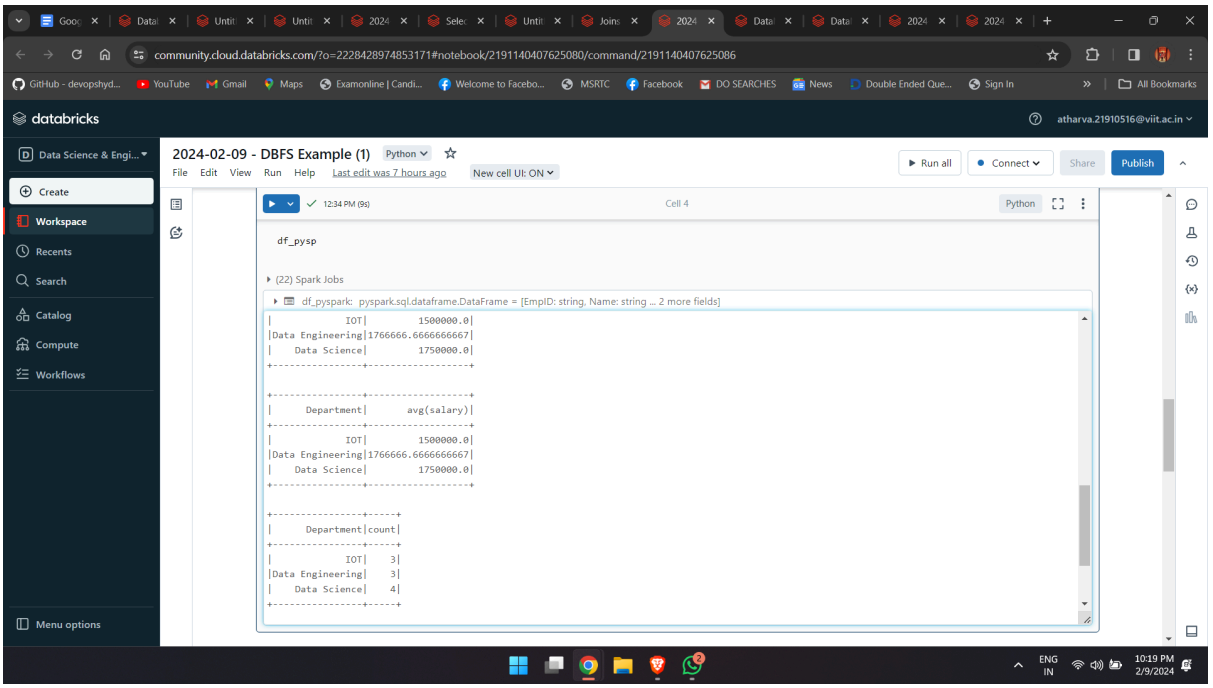

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246710/command/1349101789246711
databricks
atharva.21910516@viit.ac.in
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
SelectingFile
Python
File Edit View Run Help
Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
02:51 PM (2s)
Cell 1
Python
from pyspark.sql import SparkSession
spark =SparkSession.builder.appName('pyspark - example join').getOrCreate()
data = [(('Ram'),'1991-04-01','M',3000),
(('Mike'),'2000-05-19','M',4000),
(('Rohini'),'1978-09-05','M',4000),
(('Maria'),'1967-12-01','F',4000),
(('Jenis'),'1980-02-17','F',1200)]
columns = ['Name','DOB','Gender','salary']
df = spark.createDataFrame(data=data,schema=columns)
df.show()
(3) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Name: string, DOB: string ... 2 more fields]
+------+----------+------+------+
| Name| DOB|Gender|salary|
+------+----------+------+------+
Ram	1991-04-01	M	3000
Mike	2000-05-19	M	4000
Rohini	1978-09-05	M	4000
Maria	1967-12-01	F	4000
Jenis	1980-02-17	F	1200
+------+----------+------+------+
ENG
IN
10:17 PM
2/9/2024

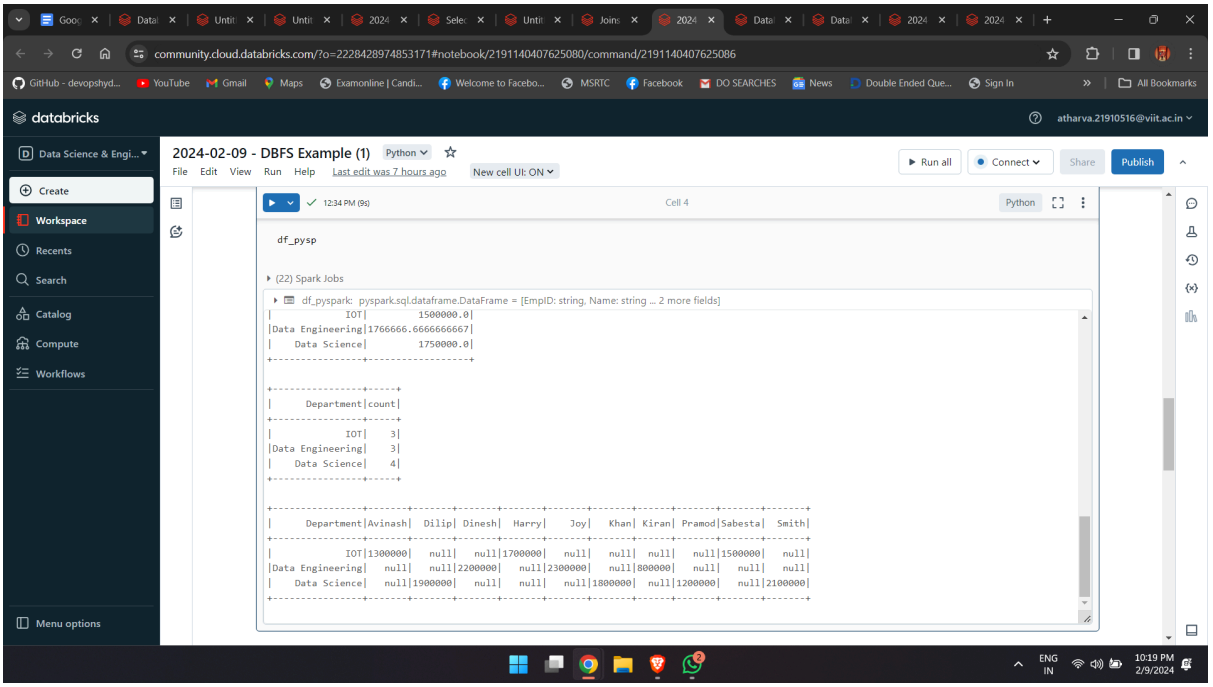
GroupBy with different Aggregate Functions :



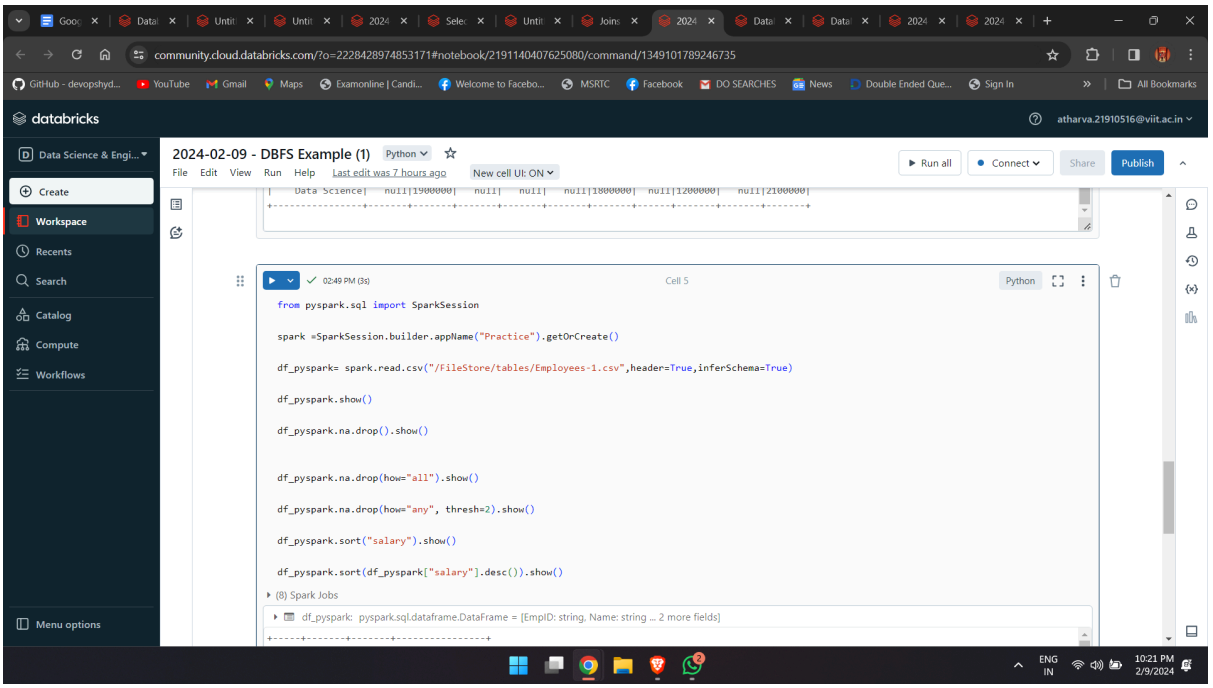


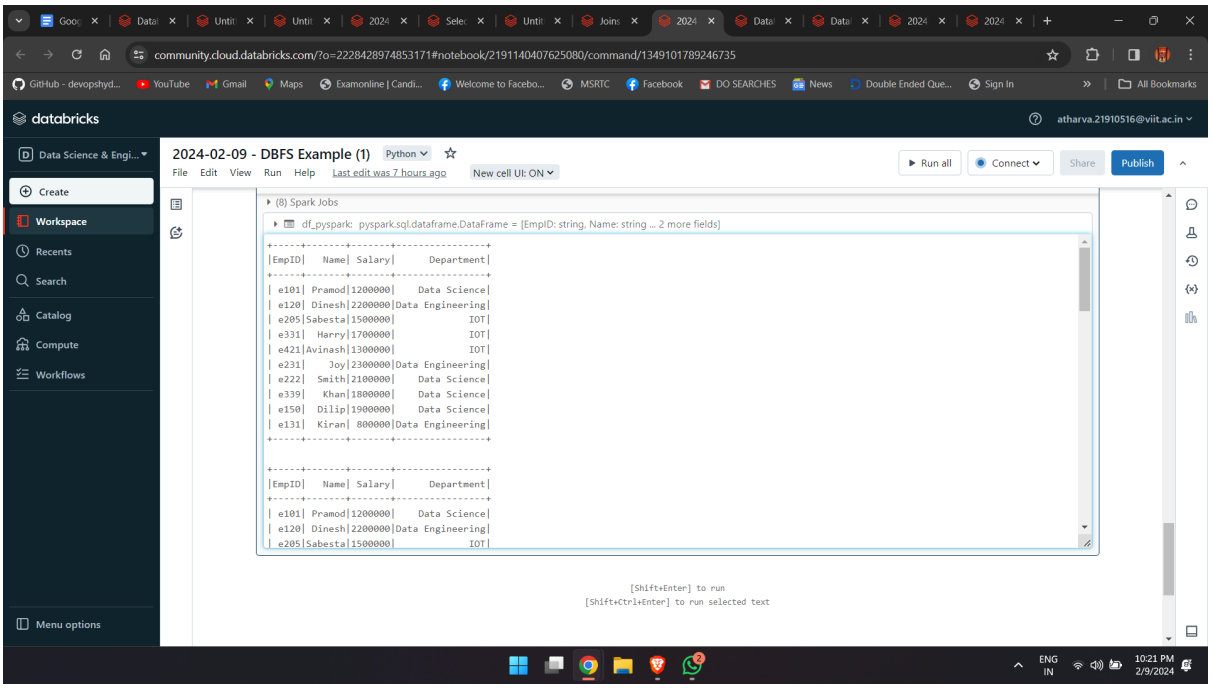

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/2191140407625086
databricks
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
12:34 PM (9s)
Cell 4
Python
df_pysp
(22) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
+---------------+-----------+
| Department|min(salary)|
+---------------+-----------+
| IOT| 1300000|
|Data Engineering| 800000|
| Data Science| 1200000|
+---------------+-----------+
+---------------+-----------+
| Department|max(salary)|
+---------------+-----------+
| IOT| 1700000|
|Data Engineering| 2300000|
| Data Science| 2100000|
+---------------+-----------+
+---------------+------------------+
| Department| avg(salary)|
+---------------+------------------+
ENG IN
10:18 PM 2/9/2024

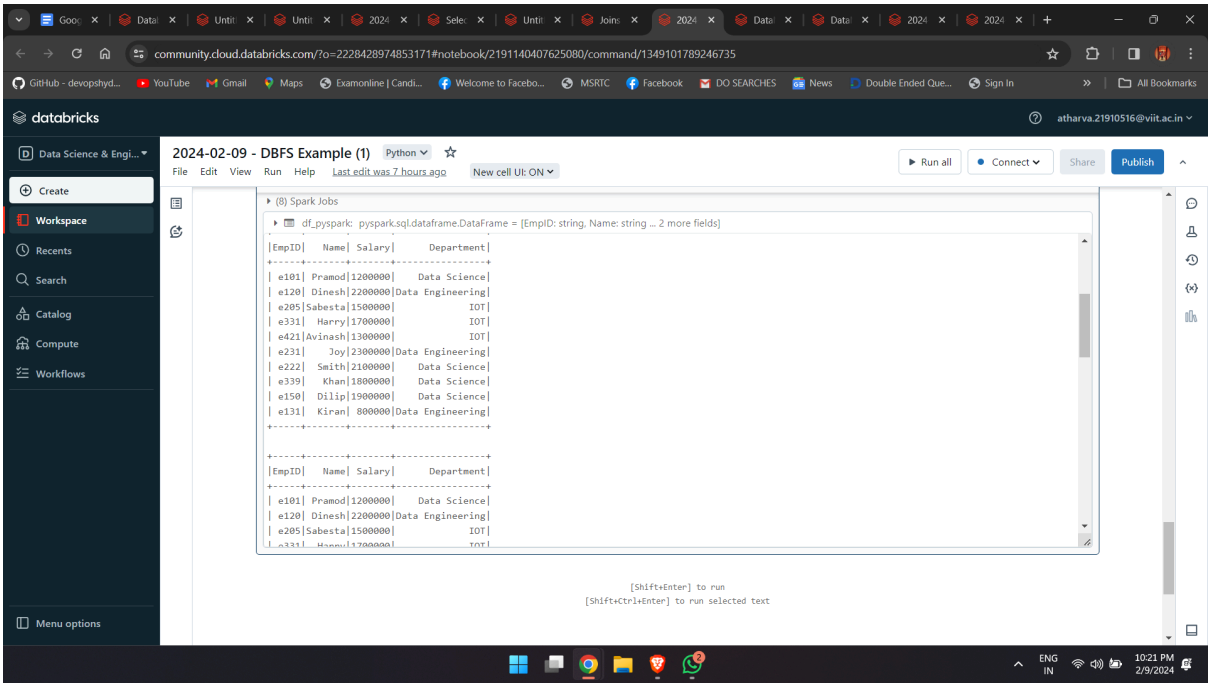

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/2191140407625086
databricks
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
12:34 PM (9s)
Cell 4
Python
df_pysp
(22) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
IOT	1500000.0
Data Engineering	1766666.6666666667
Data Science	1750000.0
+---------------+------------------+
+---------------+------------------+
| Department| avg(salary)|
+---------------+------------------+
IOT	1500000.0
Data Engineering	1766666.6666666667
Data Science	1750000.0
+---------------+------------------+
+---------------+-----+
| Department|count|
+---------------+-----+
| IOT| 3|
|Data Engineering| 3|
| Data Science| 4|
+---------------+-----+
ENG IN
10:19 PM 2/9/2024

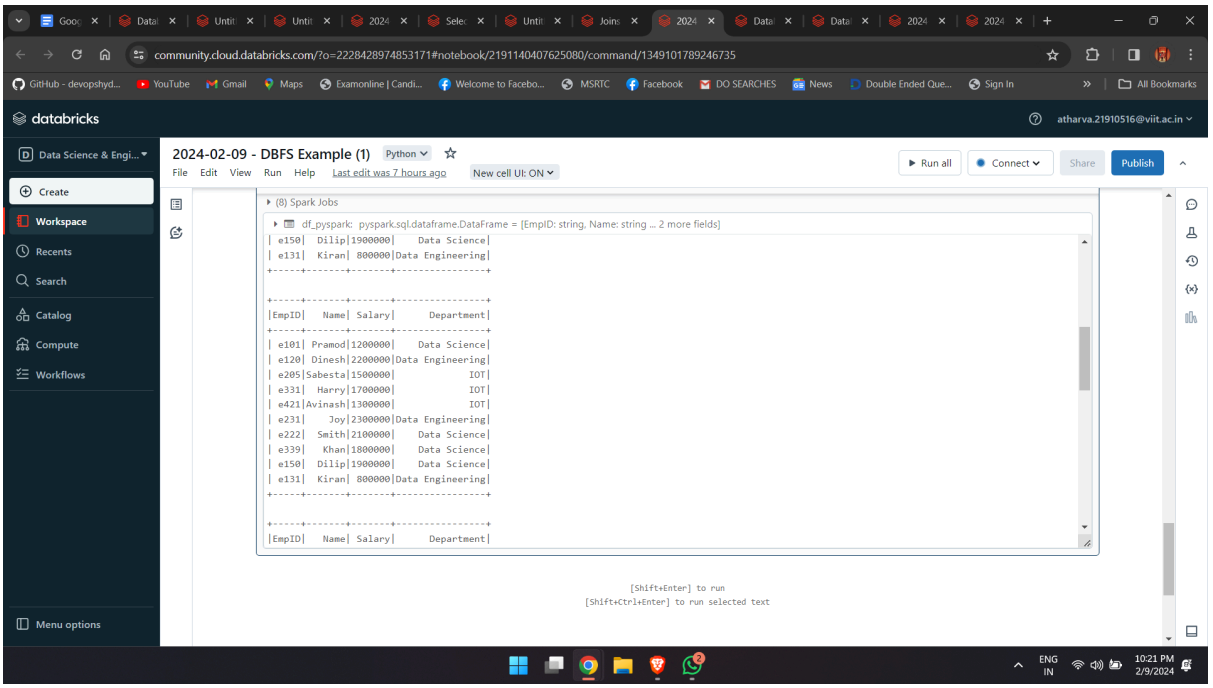

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/2191140407625086
databricks
atharva.21910516@viit.ac.in
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help
Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
12:34 PM (9s)
Cell 4
Python
df_pysp
(22) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
IOT	1500000.0
Data Engineering	1766666.6666666667
Data Science	1750000.0
+---------------+------------------+
+---------------+-----+
| Department|count|
+---------------+-----+
| IOT| 3|
|Data Engineering| 3|
| Data Science| 4|
+---------------+-----+
+---------------+-------+-------+-------+-------+-------+-------+------+-------+-------+-------+
| Department|Avinash| Dilip| Dinesh| Harry| Joy| Khan| Kiran| Pramod|Sabesta| Smith|
+---------------+-------+-------+-------+-------+-------+-------+------+-------+-------+-------+
| IOT|1300000| null| null|1700000| null| null| null| null|1500000| null|
|Data Engineering| null| null|2200000| null|2300000| null|800000| null| null| null|
| Data Science| null|1900000| null| null| null|1800000| null|1200000| null|2100000|
+---------------+-------+-------+-------+-------+-------+-------+------+-------+-------+-------+
ENG
IN
10:19 PM
2/9/2024

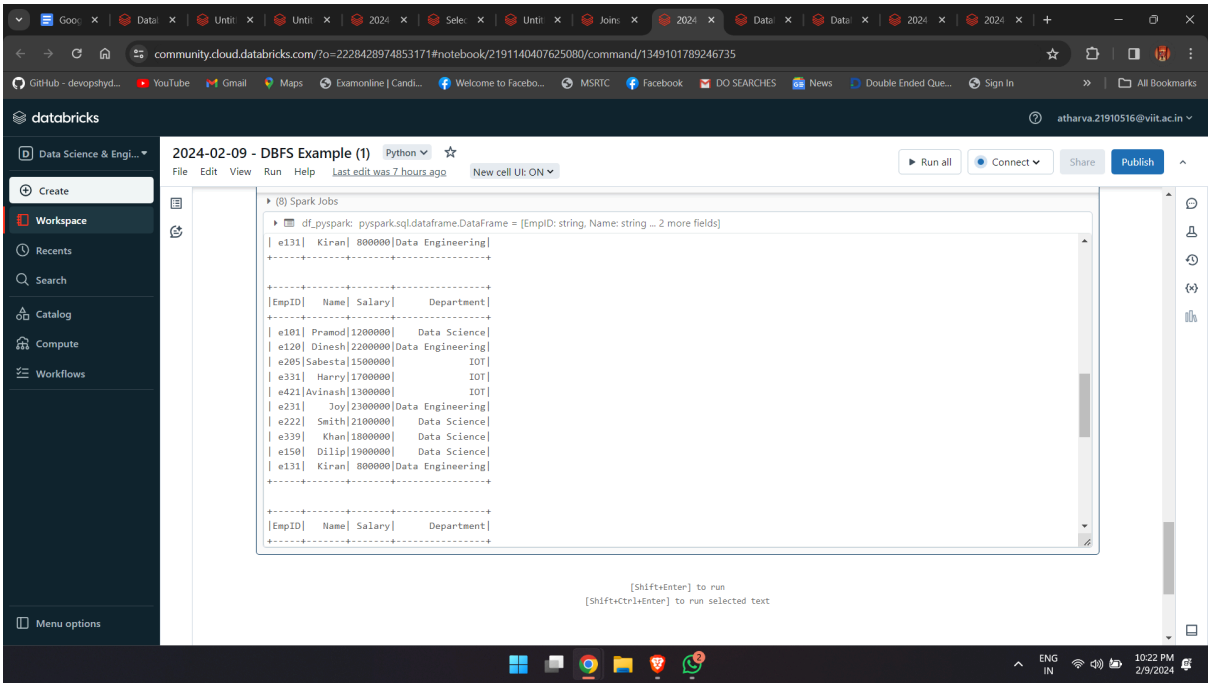
Sorting data :

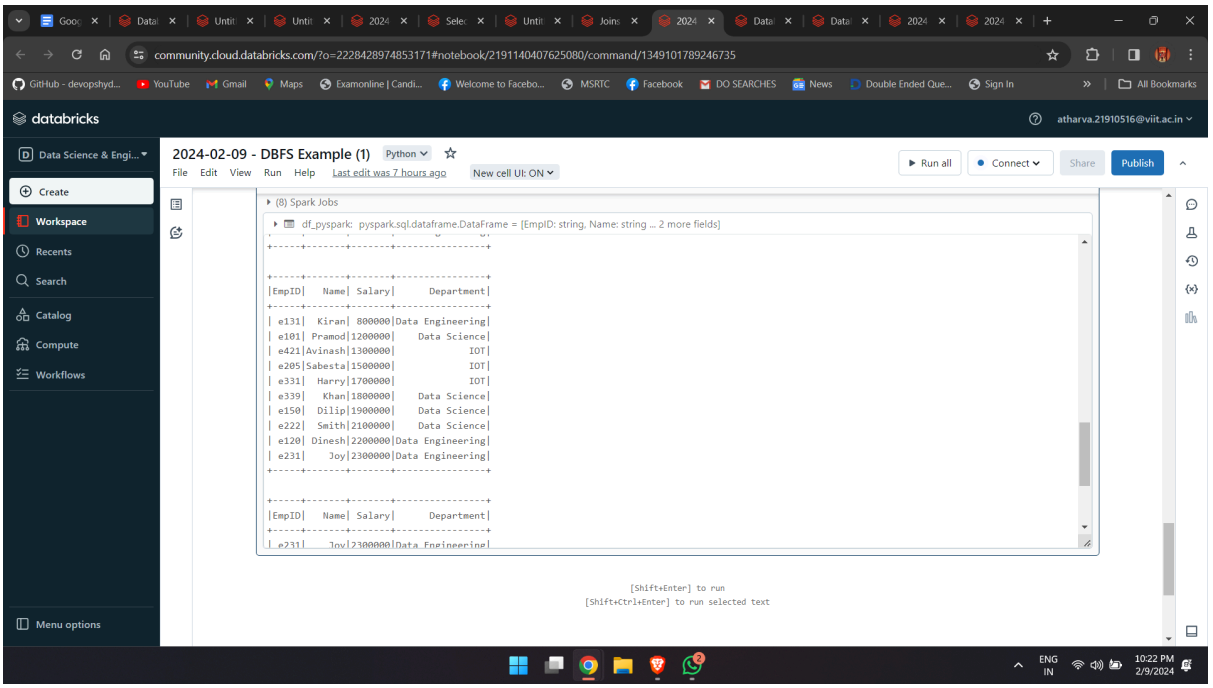


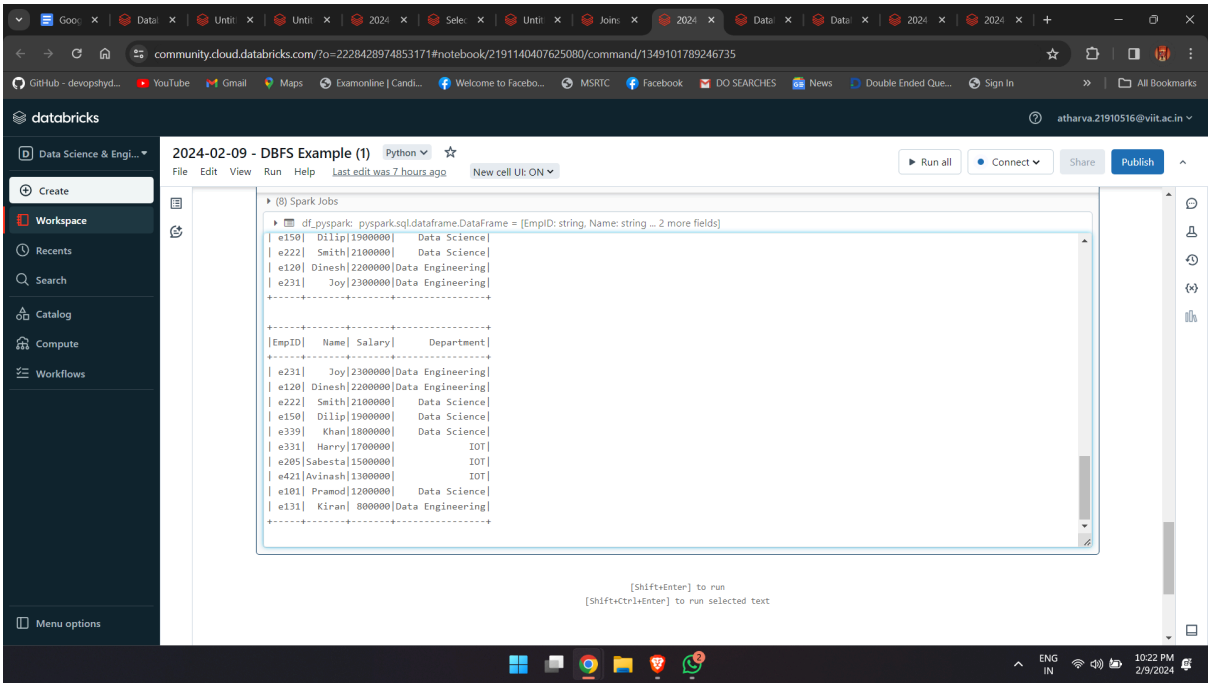



community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/1349101789246735
databricks
atharva.21910516@viit.ac.in
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
(8) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
|EmpID| Name| Salary| Department|
e101	Pramod	1200000	Data Science
e120	Dinesh	2200000	Data Engineering
e205	Sabesta	1500000	IOT
e331	Harry	1700000	IOT
e421	Avinash	1300000	IOT
e231	Joy	2300000	Data Engineering
e222	Smith	2100000	Data Science
e339	Khan	1800000	Data Science
e150	Dilip	1900000	Data Science
e131	Kiran	800000	Data Engineering
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
ENG IN
10:21 PM 2/9/2024

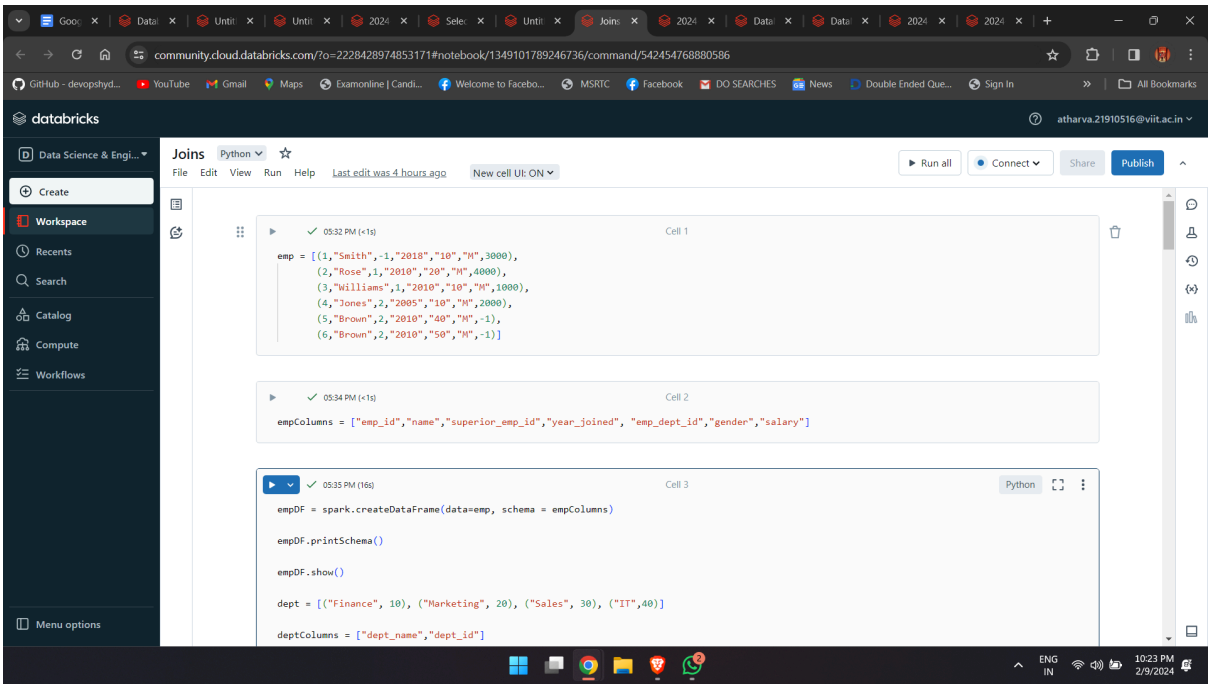

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/1349101789246735
databricks
atharva.21910516@viit.ac.in
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
(8) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
| e150| Dilip|1900000| Data Science|
| e131| Kiran| 800000|Data Engineering|
|EmpID| Name| Salary| Department|
e101	Pramod	1200000	Data Science
e120	Dinesh	2200000	Data Engineering
e205	Sabesta	1500000	IOT
e331	Harry	1700000	IOT
e421	Avinash	1300000	IOT
e231	Joy	2300000	Data Engineering
e222	Smith	2100000	Data Science
e339	Khan	1800000	Data Science
e150	Dilip	1900000	Data Science
e131	Kiran	800000	Data Engineering
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
ENG IN
10:21 PM 2/9/2024

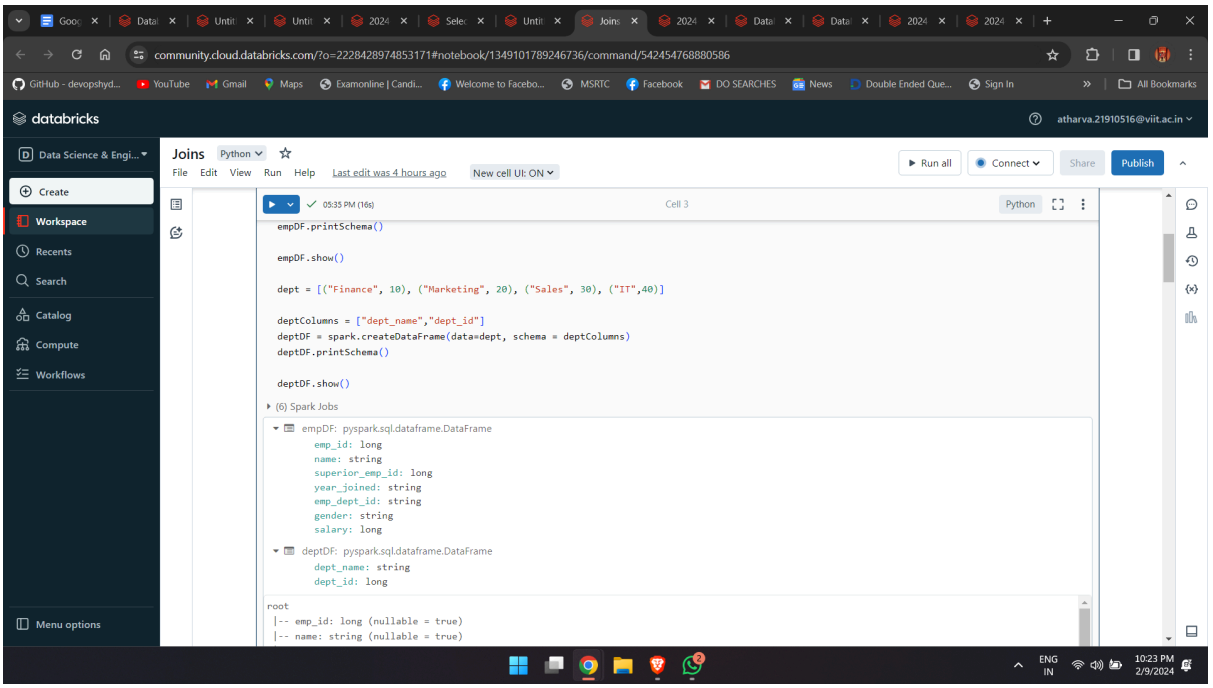

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/1349101789246735
databricks
atharva.21910516@viit.ac.in
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
(8) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
| e131| Kiran| 800000|Data Engineering|
|EmpID| Name| Salary| Department|
e101	Pramod	1200000	Data Science
e120	Dinesh	2200000	Data Engineering
e205	Sabesta	1500000	IOT
e331	Harry	1700000	IOT
e421	Avinash	1300000	IOT
e231	Joy	2300000	Data Engineering
e222	Smith	2100000	Data Science
e339	Khan	1800000	Data Science
e150	Dilip	1900000	Data Science
e131	Kiran	800000	Data Engineering
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
ENG
IN
10:22 PM
2/9/2024

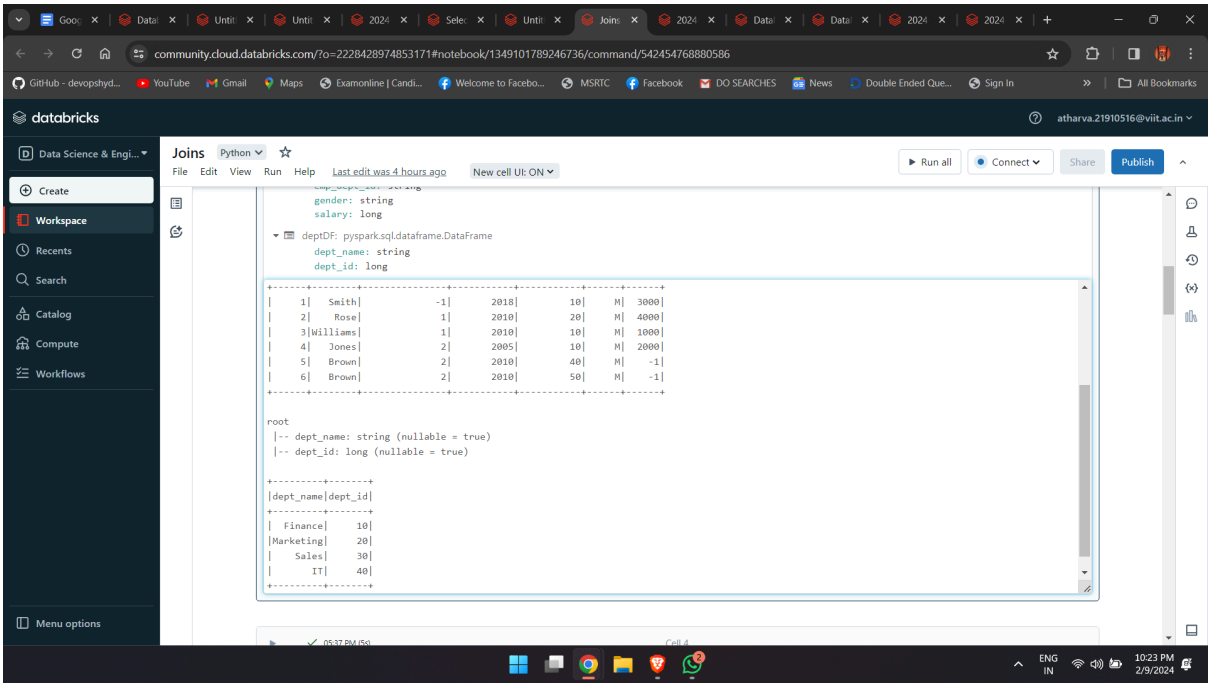

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/1349101789246735
databricks
atharva.21910516@viit.ac.in
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
(8) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
|EmpID| Name| Salary| Department|
e131	Kiran	800000	Data Engineering
e101	Pramod	1200000	Data Science
e421	Avinash	1300000	IOT
e205	Sabesta	1500000	IOT
e331	Harry	1700000	IOT
e339	Khan	1800000	Data Science
e150	Dilip	1900000	Data Science
e222	Smith	2100000	Data Science
e120	Dinesh	2200000	Data Engineering
e231	Joy	2300000	Data Engineering
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
ENG
IN
10:22 PM
2/9/2024

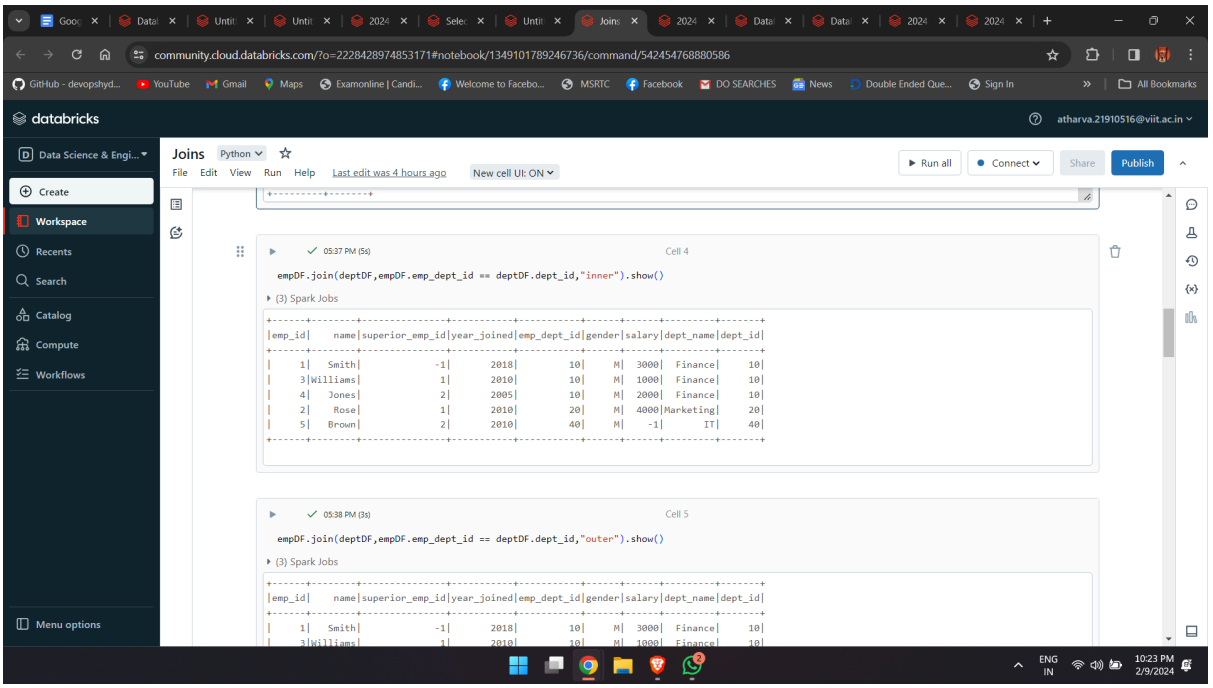

community.cloud.databricks.com/?o=2228428974853171#notebook/2191140407625080/command/1349101789246735
databricks
atharva.21910516@viit.ac.in
2024-02-09 - DBFS Example (1)
Python
File Edit View Run Help
Last edit was 7 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
(8) Spark Jobs
df_pyspark: pyspark.sql.dataframe.DataFrame = [EmpID: string, Name: string ... 2 more fields]
e150	Dilip	1900000	Data Science
e222	Smith	2100000	Data Science
e120	Dinesh	2200000	Data Engineering
e231	Joy	2300000	Data Engineering
+-----+-------+-------+----------------+
+-----+-------+-------+----------------+
|EmpID| Name| Salary| Department|
+-----+-------+-------+----------------+
e231	Joy	2300000	Data Engineering
e120	Dinesh	2200000	Data Engineering
e222	Smith	2100000	Data Science
e150	Dilip	1900000	Data Science
e339	Khan	1800000	Data Science
e331	Harry	1700000	IOT
e205	Sabesta	1500000	IOT
e421	Avinash	1300000	IOT
e101	Pramod	1200000	Data Science
e131	Kiran	800000	Data Engineering
+-----+-------+-------+----------------+
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
ENG
IN
10:22 PM
2/9/2024

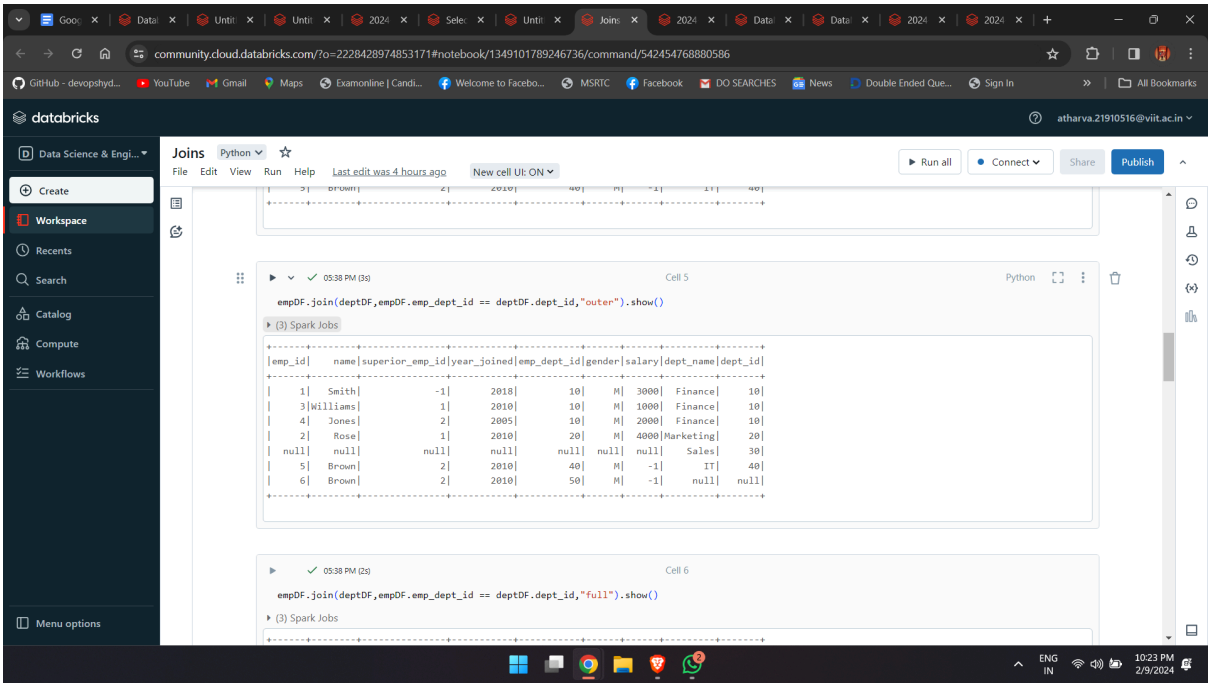
Different types of JOINS :

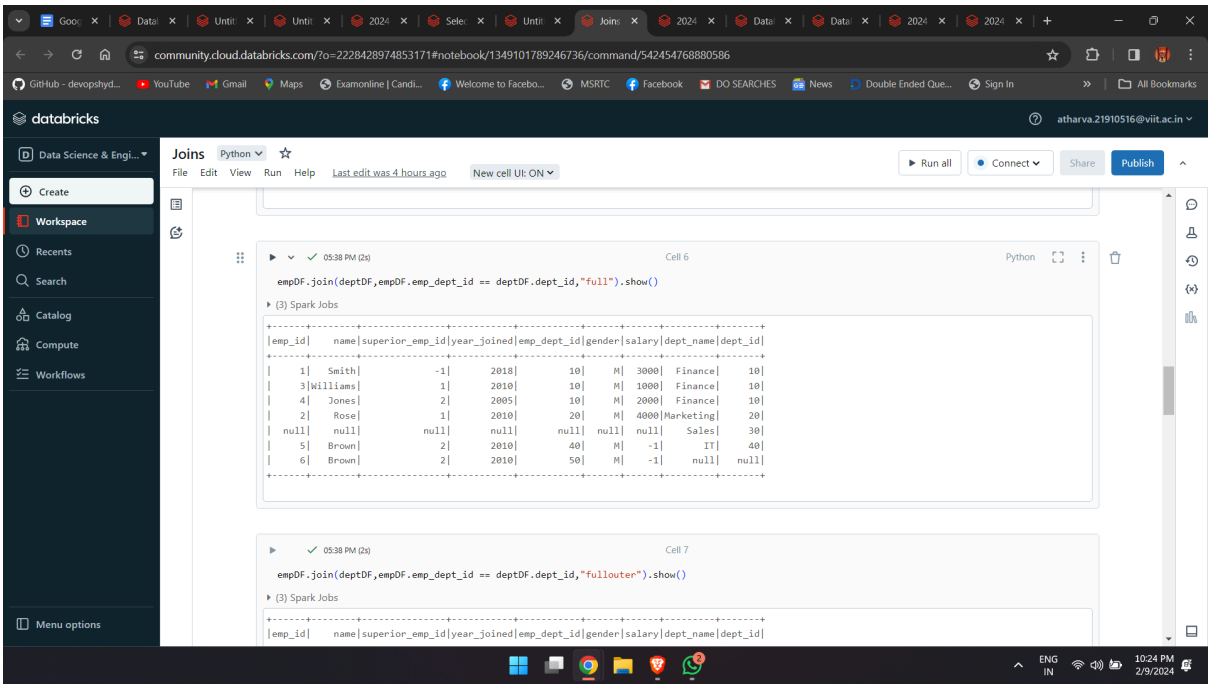


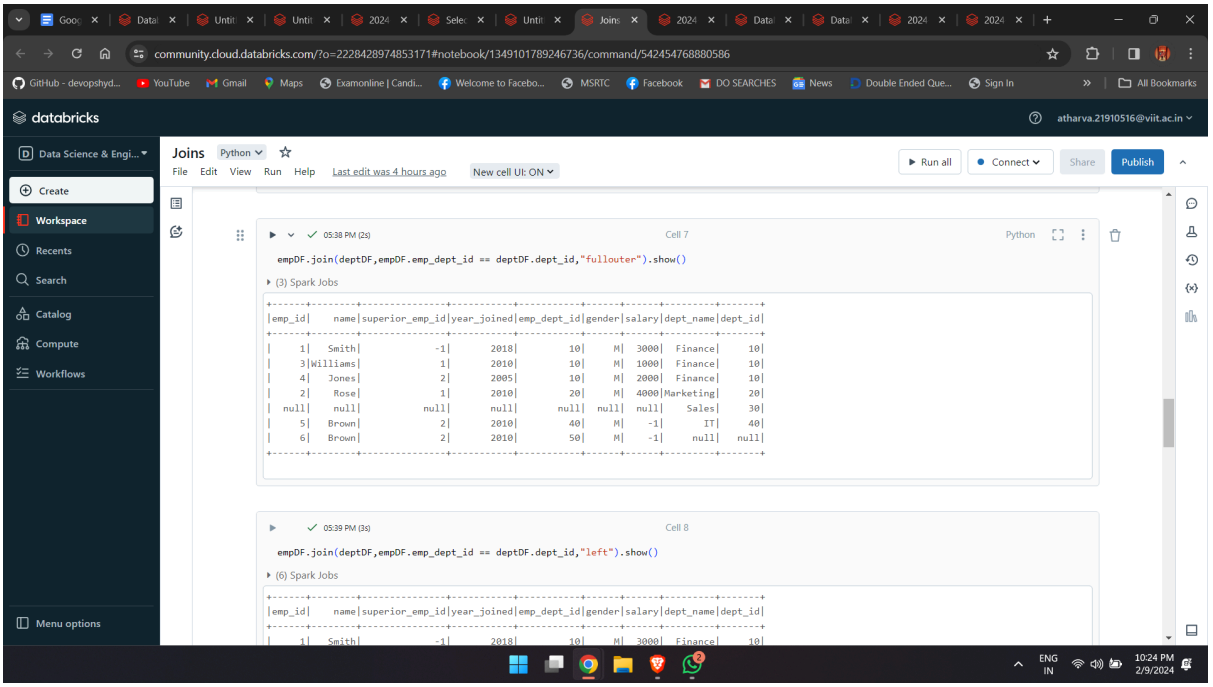


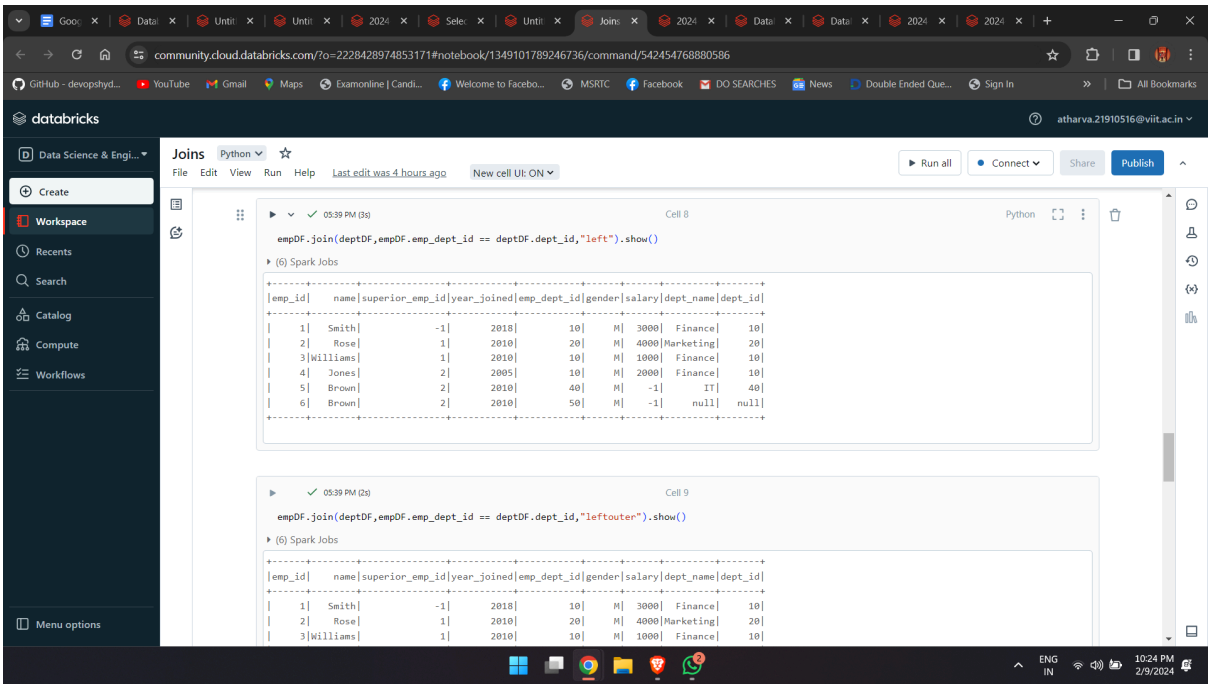

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
Joins
Python
File Edit View Run Help Last edit was 4 hours ago New cell UI: ON
Run all Connect Share Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
gender: string
salary: long
deptDF: pyspark.sql.dataframe.DataFrame
dept_name: string
dept_id: long
+------+--------+--------------+----------+----------+------+------+
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	M	2000
5	Brown	2	2010	40	M	-1
6	Brown	2	2010	50	M	-1
+------+--------+--------------+----------+----------+------+------+
root
|-- dept_name: string (nullable = true)
|-- dept_id: long (nullable = true)
+---------+-------+
|dept_name|dept_id|
+---------+-------+
| Finance| 10|
|Marketing| 20|
| Sales| 30|
| IT| 40|
+---------+-------+
05:37 PM (5s)
Cell 4
ENG IN
10:23 PM 2/9/2024

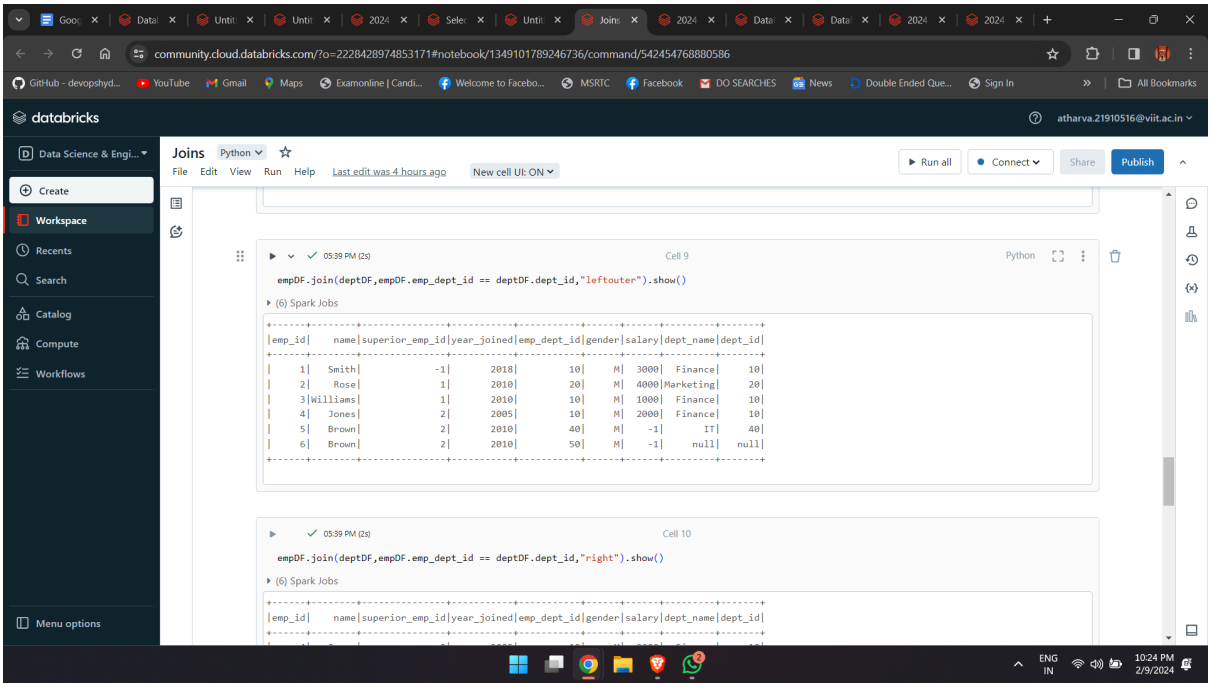

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
Joins
Python
File Edit View Run Help Last edit was 4 hours ago New cell UI: ON
Run all Connect Share Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:37 PM (5s)
Cell 4
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	M	2000	Finance	10
2	Rose	1	2010	20	M	4000	Marketing	20
5	Brown	2	2010	40	M	-1	IT	40
+------+--------+--------------+----------+----------+------+------+---------+-------+
05:38 PM (3s)
Cell 5
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|
ENG IN
10:23 PM 2/9/2024

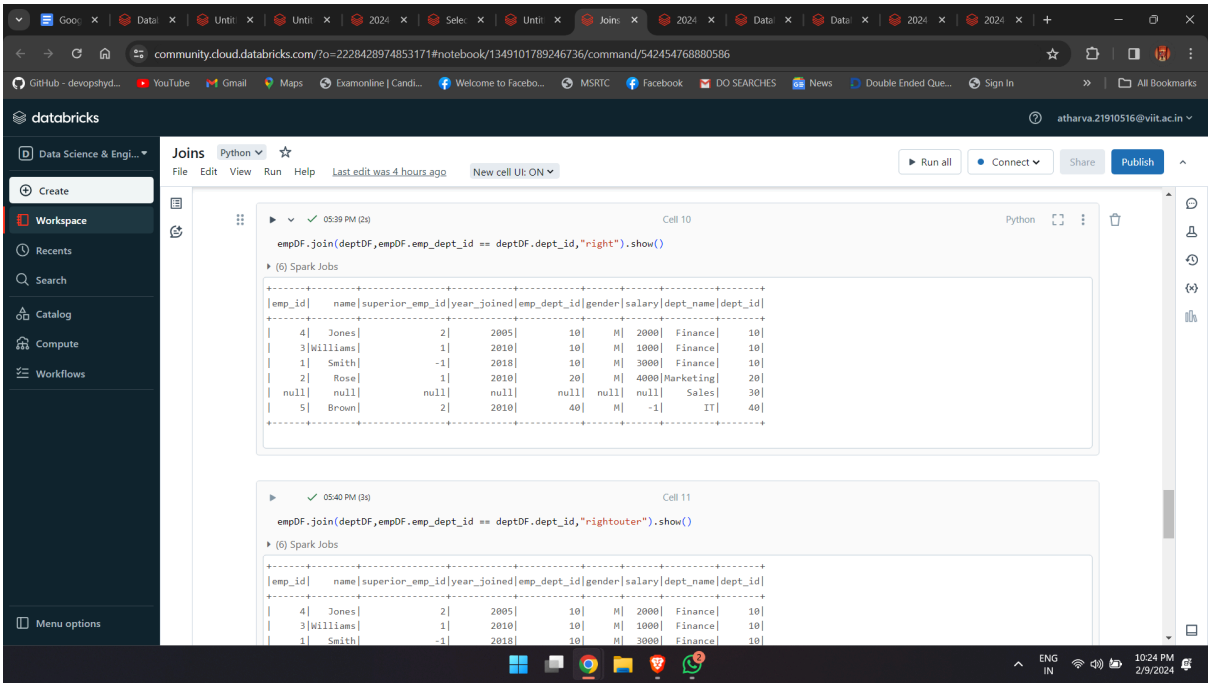

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help Last edit was 4 hours ago New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:38 PM (3s)
Cell 5
Python
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	M	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	M	-1	IT	40
6	Brown	2	2010	50	M	-1	null	null
05:38 PM (2s)
Cell 6
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"full").show()
(3) Spark Jobs
ENG IN
10:23 PM 2/9/2024

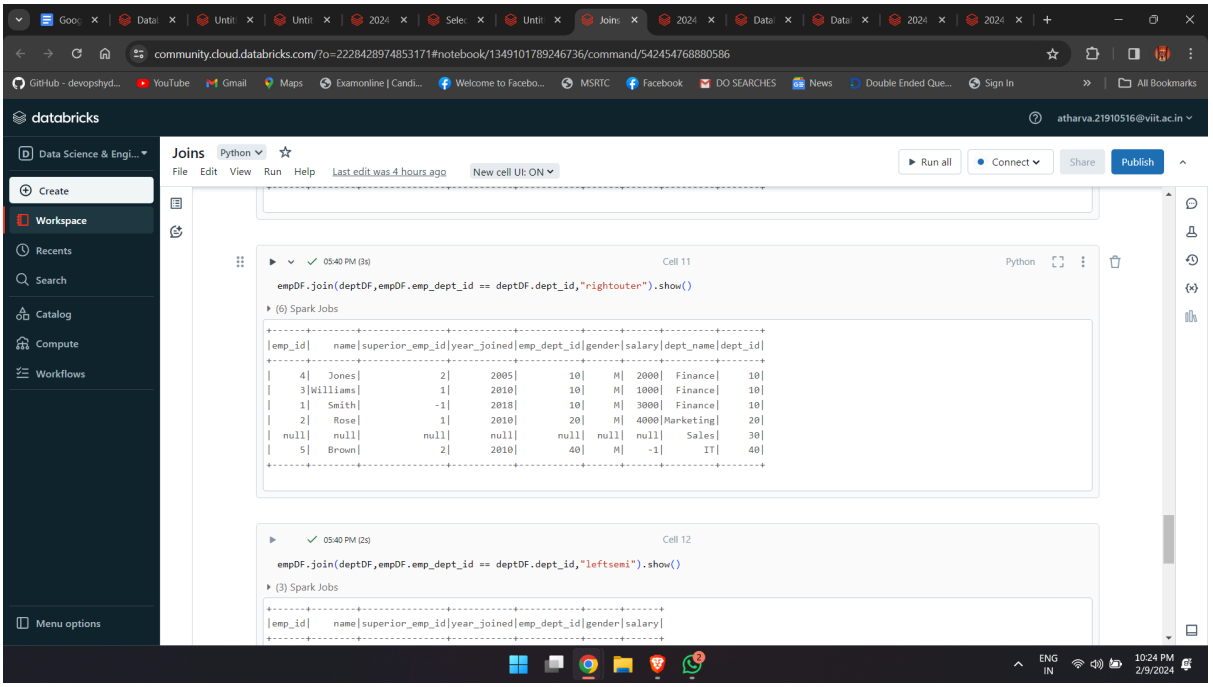

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help Last edit was 4 hours ago New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:38 PM (2s)
Cell 6
Python
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"full").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	M	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	M	-1	IT	40
6	Brown	2	2010	50	M	-1	null	null
05:38 PM (2s)
Cell 7
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"fullouter").show()
(3) Spark Jobs
|emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
ENG IN
10:24 PM 2/9/2024

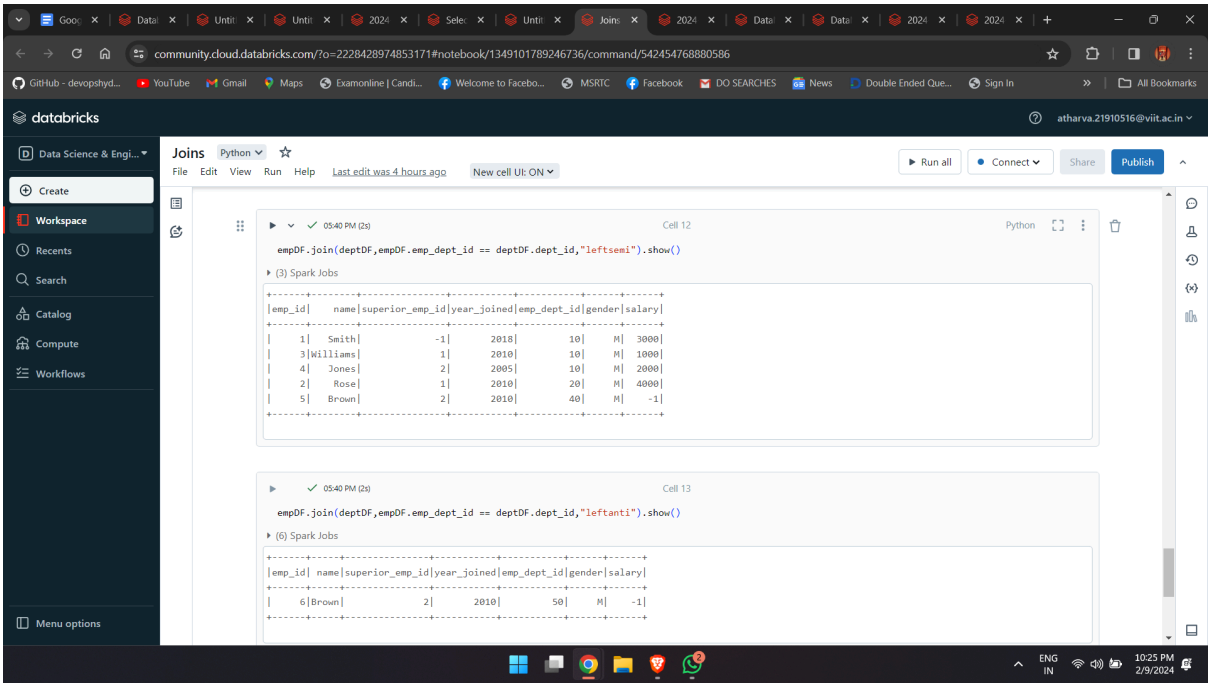

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help Last edit was 4 hours ago New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:38 PM (2s)
Cell 7
Python
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"fullouter").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	M	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	M	-1	IT	40
6	Brown	2	2010	50	M	-1	null	null
+------+--------+---------------+----------+-------------+------+------+---------+-------+
05:39 PM (3s)
Cell 8
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|
ENG
IN
10:24 PM
2/9/2024

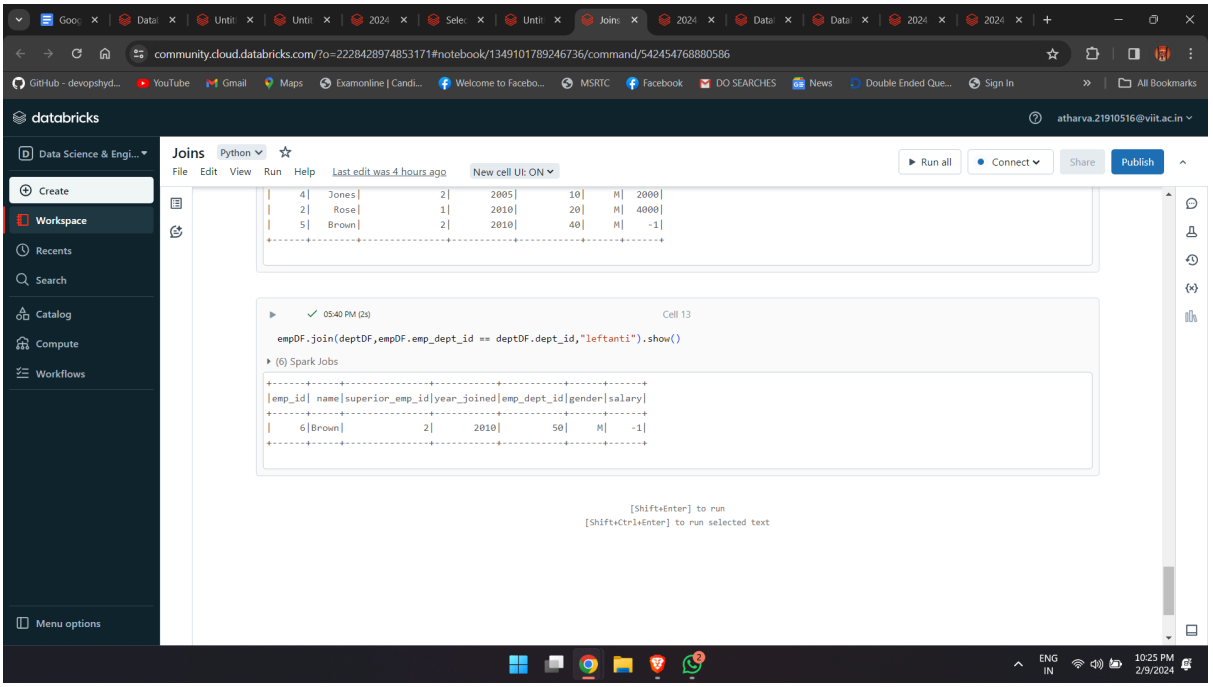

community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help Last edit was 4 hours ago New cell UI: ON
Run all
Connect
Share
Publish
Data Science & Engi...
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:39 PM (3s)
Cell 8
Python
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	M	2000	Finance	10
5	Brown	2	2010	40	M	-1	IT	40
6	Brown	2	2010	50	M	-1	null	null
+------+--------+---------------+----------+-------------+------+------+---------+-------+
05:39 PM (2s)
Cell 9
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"leftouter").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
ENG
IN
10:24 PM
2/9/2024


community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help Last edit was 4 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:39 PM (2s)
Cell 9
Python
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"leftouter").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	M	2000	Finance	10
5	Brown	2	2010	40	M	-1	IT	40
6	Brown	2	2010	50	M	-1	null	null
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
05:39 PM (2s)
Cell 10
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"right").show()
(6) Spark Jobs
ENG IN
10:24 PM 2/9/2024


community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help Last edit was 4 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:39 PM (2s)
Cell 10
Python
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	M	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	M	-1	IT	40
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
05:40 PM (3s)
Cell 11
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"rightouter").show()
(6) Spark Jobs
ENG IN
10:24 PM 2/9/2024


community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help
Last edit was 4 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:40 PM (3s)
Cell 11
Python
empDF.join(deptDF,empDF.emp_dept_id == deptDF.dept_id,"rightouter").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	M	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	M	-1	IT	40
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
05:40 PM (2s)
Cell 12
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|
+------+--------+---------------+-----------+-----------+------+------+
ENG
IN
10:24 PM
2/9/2024


community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help
Last edit was 4 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
05:40 PM (2s)
Cell 12
Python
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	M	2000
2	Rose	1	2010	20	M	4000
5	Brown	2	2010	40	M	-1
+------+--------+---------------+-----------+-----------+------+------+
05:40 PM (2s)
Cell 13
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| M| -1|
+------+----+---------------+-----------+-----------+------+------+
ENG
IN
10:25 PM
2/9/2024


community.cloud.databricks.com/?o=2228428974853171#notebook/1349101789246736/command/542454768880586
databricks
atharva.21910516@viit.ac.in
Joins
Python
File Edit View Run Help
Last edit was 4 hours ago
New cell UI: ON
Run all
Connect
Share
Publish
Create
Workspace
Recents
Search
Catalog
Compute
Workflows
Menu options
4	Jones	2	2005	10	M	2000
2	Rose	1	2010	20	M	4000
5	Brown	2	2010	40	M	-1
05:40 PM (2s)
Cell 13
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| M| -1|
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
ENG
IN
10:25 PM
2/9/2024