Working on PySpark in Jupyter notebook

-Open an Anaconda Navigator and click **launch** and in top right corner **click new** in that choose **python.**

-Lets takes a students values as example.

Create CSV file:

-create a csv file named as Student in **notepad**.

ROLL_NO,CGP,ATTEDENCE_PERCENTAGE

,,,

1,9.012,89%

2,7.894,85%

3,8.543,90%

-save the file as student.csv in all types of file.

Create values in MySql :

-create a connection with password.

-create a **schema** and create **database** and **tables**.

-refer the video for your reference :

https://youtu.be/wALCw0F8e9M?si=xol-1BA7rkBVHbHw

https://youtu.be/UzodkZUt5JY?si=9wW1T1KLnx057NkP

- Store the vales of student with ROLL_NO as key

```
+-------+-----+----------+---+
|ROLL_NO| NAME|DEPARTMENT|AGR|
+-------+-----+----------+---+
|      1|Alice|       CSE| 19|
|      2|  Bob|        IT| 19|
|      3| Chen|      AIDS| 19|
+-------+-----+----------+---+
```

PySpark Code:

-lets create a csv file and mysql and **joined** them and **write to another csv file and mysql.**

Cell 1 : !pip install pyspark

Cell 2 : from pyspark.sql import SparkSession

Replace this **path** with your actual JDBC driver location

jdbc_driver_path = "C:\\jar\\mysql-connector-j-9.2.0 (1)\\mysql-connector-j-9.2.0"

Create Spark session

spark = SparkSession.builder \

    .appName("MySQL + CSV in Jupyter") \

    .config("spark.jars", jdbc_driver_path) \

    .getOrCreate()

Cell 3 : # install mysql-connector

!pip install mysql-connector-python

 pip install sqlalchemy mysql-connector-python

Cell 4 : from pyspark.sql import SparkSession

from pyspark.sql import SparkSession

Initialize Spark Session

spark = SparkSession.builder \

    .appName("MySQL to Spark") \

    .config("spark.jars", "mysql-connector-java-8.0.33.jar") \

    .getOrCreate()

```
# Define JDBC properties
mysql_url = "jdbc:mysql://localhost:3306/company"
mysql_properties = {
    "user": "root",
    "password": "My@3066Sql",  # Note: don't URL-encode special characters here
    "driver": "com.mysql.cj.jdbc.Driver"
}

# Table name
table_name = "student"

# Read MySQL table into PySpark DataFrame
df_mysql = spark.read.jdbc(url=mysql_url, table=table_name, properties=mysql_properties)

# Show top records
df_mysql.show()
```

output:

```
+-------+-----+----------+---+
|ROLL_NO| NAME|DEPARTMENT|AGR|
+-------+-----+----------+---+
|      1|Alice|       CSE| 19|
|      2|  Bob|        IT| 19|
|      3| Chen|      AIDS| 19|
+-------+-----+----------+---+
```

Cell 4 : # Load CSV file (should be in same dir or give full path)

df_csv = spark.read.csv("D:\\csv file\\Student.csv", header=True, inferSchema=True)

Show CSV data

print("📄 Data from CSV:")

df_csv.show()

output :

```
📄 Data from CSV:
+-------+-----+--------------------+
|ROLL_NO|  CGP|ATTEDENCE_PERCENTAGE|
+-------+-----+--------------------+
|   NULL| NULL|                NULL|
|      1|9.012|                 89%|
|      2|7.894|                 85%|
|      3|8.543|                 90%|
+-------+-----+--------------------+
```

Cell 5 : #join the csv file and mysql

df_joined = df_csv.join(df_mysql, on="ROLL_NO", how="full")

Show result

df_joined.show()

output :

```
+-------+-----+--------------------+-----+----------+----+
|ROLL_NO|  CGP|ATTEDENCE_PERCENTAGE| NAME|DEPARTMENT| AGR|
+-------+-----+--------------------+-----+----------+----+
|   NULL| NULL|                NULL| NULL|      NULL|NULL|
|      1|9.012|                 89%|Alice|       CSE|  19|
|      2|7.894|                 85%|  Bob|        IT|  19|
|      3|8.543|                 90%| Chen|      AIDS|  19|
+-------+-----+--------------------+-----+----------+----+
```

Cell 6 : # write the joined table to mysql

from pyspark.sql import SparkSession

```
spark = SparkSession.builder \
    .appName("SaveToMySQL") \
    .config("spark.jars", "C:/path/to/mysql-connector-j-8.0.32.jar") \
    .getOrCreate()


jdbc_url = "jdbc:mysql://localhost:3306/students"
table_name = "student_report"
connection_properties = {
    "user": "root",
    "password": "My@3066Sql",
    "driver": "com.mysql.cj.jdbc.Driver"
}


df_joined.write \
    .jdbc(url=jdbc_url, table=table_name, mode="overwrite",
properties=connection_properties)
df_joined.show()
```

output :

```
+-------+-----+--------------------+-----+----------+----+
|ROLL_NO|  CGP|ATTEDENCE_PERCENTAGE| NAME|DEPARTMENT| AGR|
+-------+-----+--------------------+-----+----------+----+
|   NULL| NULL|                NULL| NULL|      NULL|NULL|
|      1|9.012|                 89%|Alice|       CSE|  19|
|      2|7.894|                 85%|  Bob|        IT|  19|
|      3|8.543|                 90%| Chen|      AIDS|  19|
+-------+-----+--------------------+-----+----------+----+
```

Cell 7 : # converting to python because with pyspark we cannot store values in csv file

df_pandas = df_joined.toPandas()

Cell 8 :#writhing to csv file

import os

Ensure the directory exists

os.makedirs("D:/csv_file", exist_ok=True)

Now save the DataFrame

df_pandas.to_csv("D:/csv_file/student_report_single.csv", index=False, header=True)

print(df_pandas) # Prints the whole DataFrame

output :

	ROLL_NO	CGP	ATTEDENCE_PERCENTAGE	NAME	DEPARTMENT	AGR
0	NaN	NaN	None	None	None	NaN
1	1.0	9.012	89%	Alice	CSE	19.0
2	2.0	7.894	85%	Bob	IT	19.0
3	3.0	8.543	90%	Chen	AIDS	19.0

-The joined values written to student_report csv file and mysql database.

-You can check in mysql workbench and csv file that given in d drive.

- we have done **collect** data from two structure data and **transform** them into into single source data and **load** to two different structured data.

How to run this code in cmd :

-In notebook click **file** and **choose Save and Export notebook as** and click **executable py.**

-The file is downloaded and remove emoj or install pip commands from text file.

-Open your cmd and type commands as :

Before executing the script

-pip install pyspark

-pip install mysql-connector-python

-chcp 65001 (This switches the console to UTF-8 encoding (65001) which supports emojis.)

-pip install pandas

-Now execute your script like

spark-submit --jars "C:\path\to\mysql-connector-j-9.2.0.jar" C:\path\to\your_script.py

or

spark-submit --jars "C:\\path\\to\\mysql-connector-j-9.2.0.jar" C:\\path\\to\\your_script.py

Output of cmd :

MySql :

```
25/04/16 18:38:21 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
25/04/16 18:38:22 INFO CodeGenerator: Code generated in 19.314 ms
25/04/16 18:38:22 INFO JDBCRDD: closed connection
25/04/16 18:38:22 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1633 bytes result sent to driver
25/04/16 18:38:22 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 204 ms on PREAMKUMAR (executor driver)
(1/1)
25/04/16 18:38:22 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
25/04/16 18:38:22 INFO DAGScheduler: ResultStage 0 (showString at NativeMethodAccessorImpl.java:0) finished in 0.423 s
25/04/16 18:38:22 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/16 18:38:22 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished
25/04/16 18:38:22 INFO DAGScheduler: Job 0 finished: showString at NativeMethodAccessorImpl.java:0, took 0.465922 s
25/04/16 18:38:22 INFO BlockManagerInfo: Removed broadcast_0_piece0 on PREAMKUMAR:52903 in memory (size: 6.2 KiB, free:
434.4 MiB)
25/04/16 18:38:22 INFO CodeGenerator: Code generated in 18.9394 ms
+-------+-----+----------+---+
|ROLL_NO| NAME|DEPARTMENT|AGR|
+-------+-----+----------+---+
|      1|Alice|       CSE| 19|
|      2|  Bob|        IT| 19|
|      3| Chen|      AIDS| 19|
+-------+-----+----------+---+

25/04/16 18:38:22 INFO InMemoryFileIndex: It took 14 ms to list leaf files for 1 paths.
25/04/16 18:38:23 INFO InMemoryFileIndex: It took 1 ms to list leaf files for 1 paths.
25/04/16 18:38:23 INFO FileSourceStrategy: Pushed Filters:
25/04/16 18:38:23 INFO FileSourceStrategy: Post-Scan Filters: (length(trim(value#25, None)) > 0)
25/04/16 18:38:23 INFO CodeGenerator: Code generated in 13.3401 ms
25/04/16 18:38:23 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 200.0 KiB, free 434.2 M
iB)
```

CSV file:

```
25/04/16 18:38:24 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1614 bytes result sent to driver
25/04/16 18:38:24 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 250 ms on PREAMKUMAR (executor driver)
(1/1)
25/04/16 18:38:24 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
25/04/16 18:38:24 INFO DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:0) finished in 0.266 s
25/04/16 18:38:24 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/16 18:38:24 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
25/04/16 18:38:24 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.274148 s
25/04/16 18:38:24 INFO CodeGenerator: Code generated in 9.6786 ms
+-------+-----+--------------------+
|ROLL_NO|  CGP|ATTEDENCE_PERCENTAGE|
+-------+-----+--------------------+
|   NULL| NULL|                NULL|
|      1|9.012|                 89%|
|      2|7.894|                 85%|
|      3|8.543|                 90%|
+-------+-----+--------------------+

25/04/16 18:38:24 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
25/04/16 18:38:24 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
25/04/16 18:38:24 INFO DAGScheduler: Got job 4 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
25/04/16 18:38:24 INFO DAGScheduler: Final stage: ResultStage 4 (showString at NativeMethodAccessorImpl.java:0)
25/04/16 18:38:24 INFO DAGScheduler: Parents of final stage: List()
25/04/16 18:38:24 INFO DAGScheduler: Missing parents: List()
25/04/16 18:38:24 INFO DAGScheduler: Submitting ResultStage 4 (MapPartitionsRDD[19] at showString at NativeMethodAccesso
rImpl.java:0), which has no missing parents
25/04/16 18:38:24 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 12.3 KiB, free 434.1 Mi
B)
25/04/16 18:38:24 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 6.2 KiB, free 434
```

Joined data :

```
25/04/16 18:38:25 INFO Executor: Finished task 0.0 in stage 9.0 (TID 7). 6361 bytes result sent to driver
25/04/16 18:38:25 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 7) in 282 ms on PREAMKUMAR (executor driver)
(1/1)
25/04/16 18:38:25 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool
25/04/16 18:38:25 INFO DAGScheduler: ResultStage 9 (showString at NativeMethodAccessorImpl.java:0) finished in 0.297 s
25/04/16 18:38:25 INFO DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/16 18:38:25 INFO TaskSchedulerImpl: Killing all running tasks in stage 9: Stage finished
25/04/16 18:38:25 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.328576 s
25/04/16 18:38:25 INFO CodeGenerator: Code generated in 14.5873 ms
+-------+-----+--------------------+-----+----------+----+
|ROLL_NO|  CGP|ATTEDENCE_PERCENTAGE| NAME|DEPARTMENT| AGR|
+-------+-----+--------------------+-----+----------+----+
|   NULL| NULL|                NULL| NULL|      NULL|NULL|
|      1|9.012|                 89%|Alice|       CSE|  19|
|      2|7.894|                 85%|  Bob|        IT|  19|
|      3|8.543|                 90%| Chen|      AIDS|  19|
+-------+-----+--------------------+-----+----------+----+

25/04/16 18:38:25 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
25/04/16 18:38:25 INFO FileSourceStrategy: Pushed Filters:
25/04/16 18:38:25 INFO FileSourceStrategy: Post-Scan Filters:
25/04/16 18:38:25 INFO MemoryStore: Block broadcast_12 stored as values in memory (estimated size 199.9 KiB, free 433.6
MiB)
25/04/16 18:38:25 INFO MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 34.4 KiB, free 4
33.6 MiB)
25/04/16 18:38:25 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on PREAMKUMAR:52903 (size: 34.4 KiB, free:
434.2 MiB)
25/04/16 18:38:25 INFO SparkContext: Created broadcast 12 from jdbc at NativeMethodAccessorImpl.java:0
25/04/16 18:38:25 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is conside
red as scanning 4194304 bytes.
```

Mysql joined values :

```
25/04/16 18:38:25 INFO Executor: Finished task 0.0 in stage 9.0 (TID 7). 6361 bytes result sent to driver
25/04/16 18:38:25 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 7) in 282 ms on PREAMKUMAR (executor driver)
(1/1)
25/04/16 18:38:25 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool
25/04/16 18:38:25 INFO DAGScheduler: ResultStage 9 (showString at NativeMethodAccessorImpl.java:0) finished in 0.297 s
25/04/16 18:38:25 INFO DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie tasks for this job
25/04/16 18:38:25 INFO TaskSchedulerImpl: Killing all running tasks in stage 9: Stage finished
25/04/16 18:38:25 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.328576 s
25/04/16 18:38:25 INFO CodeGenerator: Code generated in 14.5873 ms
+-------+-----+--------------------+-----+----------+----+
|ROLL_NO|  CGP|ATTEDENCE_PERCENTAGE| NAME|DEPARTMENT| AGR|
+-------+-----+--------------------+-----+----------+----+
|   NULL| NULL|                NULL| NULL|      NULL|NULL|
|      1|9.012|                 89%|Alice|       CSE|  19|
|      2|7.894|                 85%|  Bob|        IT|  19|
|      3|8.543|                 90%| Chen|      AIDS|  19|
+-------+-----+--------------------+-----+----------+----+

25/04/16 18:38:25 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
25/04/16 18:38:25 INFO FileSourceStrategy: Pushed Filters:
25/04/16 18:38:25 INFO FileSourceStrategy: Post-Scan Filters:
25/04/16 18:38:25 INFO MemoryStore: Block broadcast_12 stored as values in memory (estimated size 199.9 KiB, free 433.6
MiB)
25/04/16 18:38:25 INFO MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 34.4 KiB, free 4
33.6 MiB)
25/04/16 18:38:25 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on PREAMKUMAR:52903 (size: 34.4 KiB, free:
434.2 MiB)
25/04/16 18:38:25 INFO SparkContext: Created broadcast 12 from jdbc at NativeMethodAccessorImpl.java:0
25/04/16 18:38:25 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is conside
red as scanning 4194304 bytes.
```

CSV Joined values :

```
25/04/16 18:38:33 INFO BlockManagerInfo: Removed broadcast_21_piece0 on PREAMKUMAR:52903 in memory (size: 7.4 KiB, free:
 434.3 MiB)
25/04/16 18:38:33 INFO BlockManagerInfo: Removed broadcast_16_piece0 on PREAMKUMAR:52903 in memory (size: 34.4 KiB, free
: 434.3 MiB)
25/04/16 18:38:33 INFO BlockManagerInfo: Removed broadcast_22_piece0 on PREAMKUMAR:52903 in memory (size: 7.5 KiB, free:
 434.3 MiB)
25/04/16 18:38:33 INFO BlockManagerInfo: Removed broadcast_17_piece0 on PREAMKUMAR:52903 in memory (size: 7.4 KiB, free:
 434.3 MiB)
   ROLL_NO    CGP ATTEDENCE_PERCENTAGE   NAME DEPARTMENT   AGR
0      NaN    NaN                 None   None       None   NaN
1      1.0  9.012                  89%  Alice        CSE  19.0
2      2.0  7.894                  85%    Bob         IT  19.0
3      3.0  8.543                  90%   Chen       AIDS  19.0
25/04/16 18:38:34 INFO SparkContext: Invoking stop() from shutdown hook
25/04/16 18:38:34 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/04/16 18:38:34 INFO SparkUI: Stopped Spark web UI at http://PREAMKUMAR:4040
25/04/16 18:38:34 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/04/16 18:38:34 INFO MemoryStore: MemoryStore cleared
25/04/16 18:38:34 INFO BlockManager: BlockManager stopped
25/04/16 18:38:34 INFO BlockManagerMaster: BlockManagerMaster stopped
25/04/16 18:38:34 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/04/16 18:38:34 INFO SparkContext: Successfully stopped SparkContext
25/04/16 18:38:34 INFO ShutdownHookManager: Shutdown hook called
25/04/16 18:38:34 INFO ShutdownHookManager: Deleting directory C:\Users\pream\AppData\Local\Temp\spark-a3254a4d-864d-433
4-9427-e738b2115d0a\pyspark-52dcbca8-28a6-4d50-8cee-a682cd0417aa
25/04/16 18:38:34 INFO ShutdownHookManager: Deleting directory C:\Users\pream\AppData\Local\Temp\spark-e8288056-599b-48b
b-a5a6-533ba458e57a
25/04/16 18:38:34 INFO ShutdownHookManager: Deleting directory C:\Users\pream\AppData\Local\Temp\spark-a3254a4d-864d-433
4-9427-e738b2115d0a
PS C:\Users\pream>
```