Set up your PySpark environment:

- PySpark is the Python API for Spark, which allows you to interface with Spark using Python. To set up PySpark, you need to have Apache Spark installed on your system along with the necessary Python libraries.
- You can install PySpark using pip: pip install pyspark.

Read data into PySpark DataFrames:

- PySpark DataFrames are distributed collections of data, similar to pandas DataFrames but distributed across a cluster. You can create PySpark DataFrames from various data sources such as CSV files, JSON files, Hive tables, etc.
- Use spark.read to read data into PySpark DataFrames.

Perform SparkSQL joins:

- SparkSQL is a module in Spark that provides support for querying structured and semi-structured data using SQL syntax. You can perform joins, filters, aggregations, etc., using SparkSQL.
- After reading the data into PySpark DataFrames, you can create temporary views or register DataFrames as tables to use them in SparkSQL queries. Then, you can write SQL queries to perform joins.

Convert the resulting PySpark DataFrame into a Pandas DataFrame:

- While PySpark is great for distributed processing, sometimes you may want to work with the results locally, especially if the dataset is small enough to fit into memory. In such cases, you can convert the PySpark DataFrame into a Pandas DataFrame.
- Use the toPandas() method to convert a PySpark DataFrame into a Pandas DataFrame.

Apply functions to the Pandas DataFrame:

- Once you have the data in a Pandas DataFrame, you can apply various transformations, calculations, or custom functions to the data using pandas' rich set of functions.
- Use the apply() method along with a custom function to apply transformations to columns or rows of the Pandas DataFrame.
- Now, let's put these steps into a sample example:

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

In [2]:
```
spark =SparkSession.builder.appName("example spark-sql").getOrCreate()
```

In [3]:
```
spark.sql("CREATE DATABASE IF NOT EXISTS customer_db COMMENT 'This is customer database'WITH DBPROPERTIES (ID=1, Name='John')");
```

In [6]:
```
spark.sql("DESCRIBE DATABASE EXTENDED customer_db").show();
```

```
+--------------+--------------------+
|     info_name|          info_value|
+--------------+--------------------+
|  Catalog Name|       spark_catalog|
|Namespace Name|         customer_db|
|       Comment|This is customer ...|
|      Location|file:/C:/Users/Ab...|
|         Owner|             Abhishek|
|    Properties|((ID,1), (Name,Jo...|
+--------------+--------------------+
```

In [5]:
```
spark.sql("SHOW DATABASES").show()
```

```
+-----------+
|  namespace|
+-----------+
|customer_db|
|    default|
+-----------+
```

In [10]:
```

```

```
  Cell In[10], line 3
    val sampleDF = Seq(
        ^
SyntaxError: invalid syntax
```

In [16]:
```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
Cell In[16], line 8
      1 from pyspark.sql import SparkSession
      3 spark = SparkSession \
      4     .builder \
      5     .appName("Python Spark SQL basic example") \
      6     .config("spark.some.config.option", "some-value") \
      7     .getOrCreate()
----> 8 import spark.implicits._

ModuleNotFoundError: No module named 'spark'
```

In [18]:
```
df = spark.read.json("people.json")
```

In [19]:
```
df.show()
```

```
+----+-------+
| age|   name|
+----+-------+
|NULL|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+
```

In [20]:
```
df.printSchema()
```

```
root
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)
```

In [21]:
```
df.select("name").show()
```

```
+-------+
|   name|
+-------+
|Michael|
|   Andy|
| Justin|
+-------+
```

In [23]:
```
df.select(df['name'], df['age'] + 1).show()
```

```
+-------+---------+
|   name|(age + 1)|
```

```
+-------+---------+
|   name|(age + 1)|
+-------+---------+
|Michael|     NULL|
|   Andy|       31|
| Justin|       20|
+-------+---------+
```

In [24]:
```
df.filter(df['age'] > 21).show()
```

```
+---+----+
|age|name|
+---+----+
| 30|Andy|
+---+----+
```

In [26]:
```
df.groupBy("age").count().show()
```

```
+----+-----+
| age|count|
+----+-----+
|  19|    1|
|NULL|    1|
|  30|    1|
+----+-----+
```

In [27]:
```
df.createOrReplaceTempView("people")
```

In [28]:
```
sqlDF = spark.sql("SELECT * FROM people")
sqlDF.show()
```

```
+----+-------+
| age|   name|
+----+-------+
|NULL|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+
```

In [29]:
```
df.createGlobalTempView("people")
```

In [30]:
```
spark.sql("SELECT * FROM global_temp.people").show()
```

```
+----+-------+
| age|   name|
+----+-------+
|NULL|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+
```

In [32]:
```
spark.newSession().sql("SELECT * FROM global_temp.people").show()
```

```
+----+-------+
| age|   name|
+----+-------+
|NULL|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+
```

In [ ]:
```

```

databricks

abhishek7621cse@gmail.com

Pyspark function

Python

File Edit View Run Help Last edit was 4 minutes ago New cell UI: OFF

Run all | My Cluster | Share | Publish

Cmd 1

```python
from pyspark.sql import SparkSession
import pandas as pd

# Initialize SparkSession
spark = SparkSession.builder \
    .appName("PySpark - SparkSQL Joins") \
    .getOrCreate()

```

(8) Spark Jobs

```
+-------+---+-----+-------+
|   name|key| name|address|
+-------+---+-----+-------+
|  Alice|  1|Alice|  addr1|
|    Bob|  2|  Bob|  addr2|
|Charlie|  3| null|   null|
+-------+---+-----+-------+


<command-3611120654214249>:33: FutureWarning: reindexing with a non-unique Index is deprecated and will raise in a future version.
  pandas_df['name_length'] = pandas_df['name'].apply(len)
ValueError: cannot reindex on an axis with duplicate labels
```

Command took 13.29 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:11:43 PM on My Cluster

Cmd 2

```python

# Sample data for DataFrame 1
data1 = [("Alice", 1), ("Bob", 2), ("Charlie", 3)]
schema1 = ["name", "key"]
df1 = spark.createDataFrame(data1, schema1)
df1.createOrReplaceTempView("table1")

```

df1: pyspark.sql.dataframe.DataFrame = [name: string, key: long]

Command took 0.29 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:12:47 PM on My Cluster

Cmd 3

```python

# Sample data for DataFrame 2
data2 = [("Alice", "addr1"), ("Bob", "addr2"), ("David", "addr3")]
schema2 = ["name", "address"]
df2 = spark.createDataFrame(data2, schema2)
df2.createOrReplaceTempView("table2")

```

df2: pyspark.sql.dataframe.DataFrame = [name: string, address: string]

Command took 0.17 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:12:59 PM on My Cluster

Cmd 4

```python

# Perform SQL join
result = spark.sql("""
    SELECT *
    FROM table1
    LEFT JOIN table2 ON table1.name = table2.name
""")
result.show()

```

(4) Spark Jobs

result: pyspark.sql.dataframe.DataFrame = [name: string, key: long ... 2 more fields]

```
+-------+---+-----+-------+
|   name|key| name|address|
+-------+---+-----+-------+
|  Alice|  1|Alice|  addr1|
|    Bob|  2|  Bob|  addr2|
|Charlie|  3| null|   null|
+-------+---+-----+-------+
```

Command took 1.65 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:13:09 PM on My Cluster

Cmd 5

```python

# Convert PySpark DataFrame to Pandas DataFrame
pandas_df = result.toPandas()

```

(4) Spark Jobs

Command took 1.38 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:13:17 PM on My Cluster

Cmd 6

```python
# Apply functions to Pandas DataFrame
pandas_df['name_length'] = pandas_df['name'].apply(len)
```

```
<command-3611120654214254>:2: FutureWarning: reindexing with a non-unique Index is deprecated and will raise in a future version.
  pandas_df['name_length'] = pandas_df['name'].apply(len)
ValueError: cannot reindex on an axis with duplicate labels
```

Command took 0.28 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:15:20 PM on My Cluster

Cmd 7

```
1
2  # Display Pandas DataFrame
3  print(pandas_df.head())
```

```
      name  key   name address
0    Alice    1  Alice   addr1
1      Bob    2    Bob   addr2
2  Charlie    3   None    None
```

Command took 0.08 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 12:13:27 PM on My Cluster

Cmd 8

```
1
```

[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text

databricks

abhishek7621cse@gmail.com

Pyspark3 Python

File Edit View Run Help Last edit was 5 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

Cmd 1

```
1
```

Cmd 2

Python

```
1
```

Cmd 3

Python

```
1  # Joins
2
```

Command took 1.47 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 4

```
1  from pyspark.sql import SparkSession
2  spark =SparkSession.builder.appName('pyspark - example join').getOrCreate()
3  emp = [(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)]
4  empColumns = ["emp_id","name","superior_emp_id","year_joined", "emp_dept_id","gender","salary"]
5
6  empDF = spark.createDataFrame(data=emp, schema = empColumns)
7  empDF.printSchema()
8  empDF.show()
9
10 dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
11 deptColumns = ["dept_name","dept_id"]
12 deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
13 deptDF.printSchema()
14 deptDF.show()
```

(6) Spark Jobs

empDF: pyspark.sql.dataframe.DataFrame = [emp_id: long, name: string ... 5 more fields]

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|     F|  2000|
|     5|   Brown|             2|       2010|         40|      |    -1|
|     6|   Brown|             2|       2010|         50|      |    -1|
+------+--------+--------------+-----------+-----------+------+------+

root
 |-- dept_name: string (nullable = true)
 |-- dept_id: long (nullable = true)

+---------+-------+
|dept_name|dept_id|
+---------+-------+
|  Finance|     10|
|Marketing|     20|
|    Sales|     30|
|       IT|     40|
+---------+-------+
```

Command took 15.36 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 5

```
1  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|     F|  2000|  Finance|     10|
|     2|    Rose|             1|       2010|         20|     M|  4000|Marketing|     20|
|     5|   Brown|             2|       2010|         40|      |    -1|       IT|     40|
+------+--------+--------------+-----------+-----------+------+------+---------+-------+
```

Command took 6.38 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 6

```
1  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"outer").show()
2  #Or
3  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"full").show()
4  #Or
5  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"fullouter").show()
```

(9) Spark Jobs

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

+------+--------+--------------+-----------+-----------+------+------+---------+-------+
```

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

Command took 5.58 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 7

```
1  # Left Join
2  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"left").show()
3  #Or
4  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftouter").show()
```

▸ (12) 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|     F|  2000|  Finance|     10|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     null|   null|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+


|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|     F|  2000|  Finance|     10|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     null|   null|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

Command took 5.46 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 8

```
1  # Right Join
2  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"right").show()
3  #Or
4  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"rightouter").show()
5
```

▸ (12) Spark Jobs

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


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

Command took 4.91 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 9

```
1  # Left semi Join
2  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftsemi").show()
3
```

▸ (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|     F|  2000|
|     2|    Rose|              1|       2010|         20|     M|  4000|
|     5|   Brown|              2|       2010|         40|      |    -1|
+------+--------+---------------+-----------+-----------+------+------+
```

Command took 1.90 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 10

```
1  # Leftanti Join
2  empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"leftanti").show()
3
```

▸ (6) Spark Jobs

```
+------+-----+---------------+-----------+-----------+------+------+
|emp_id| name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+-----+---------------+-----------+-----------+------+------+
|     6|Brown|              2|       2010|         50|      |    -1|
+------+-----+---------------+-----------+-----------+------+------+
```

Command took 1.70 seconds -- by abhishek7621cse@gmail.com at 2/12/2024, 11:38:21 AM on My Cluster

Cmd 11

```
1
```

[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text