DA-9-copy

November 6, 2024

0.0.1 Creating Spark Session

`The following packages are need to be implemented to perform Spark Session`

[1]:
```
from pyspark import SparkContext
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *
from pyspark.sql.types import Row
from datetime import datetime
```

Initializing Spark Session After necessary imports we have to initialize the spark session by the following Command

[2]:
```
sc = SparkContext()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
24/11/06 07:07:22 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
```

[3]:
```
spark=SparkSession.builder.appName("Python Spark SQL basic example").
 ↪config("spark.some.config.option","some-value").getOrCreate()
```

0.0.2 Creation of spark RDD

`Create a Spark RDD using the parallelize function`

[4]:
```
srecord=sc.parallelize([
    Row(roll_no=1,name="john",passed=True,marks={'Math':89,'Physics':
 ↪87,'Chemistry':
 ↪96},sports=['chess','football'],DoB=datetime(2012,5,1,12,1,5)),
    Row(roll_no=2,name="Vignesh",passed=False,marks={'Math':95,'Physics':
 ↪66,'Chemistry':
 ↪77},sports=['carrom','tennis'],DoB=datetime(2012,5,12,14,2,5)),
    Row(roll_no=3,name="Sidharth",passed=True,marks={'Math':95,'Physics':
 ↪100,'Chemistry':
 ↪95},sports=['football','kabadi'],DoB=datetime(2012,5,14,12,2,5))
])
```

0.0.3 Creating a DataFrame

[5]:

```
srdf=srecord.toDF()
```

```
24/11/06 07:07:34 WARN GarbageCollectionMetrics: To enable non-built-in garbage
collector(s) List(G1 Concurrent GC), users should configure it(them) to
spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or
spark.eventLog.gcMetrics.oldGenerationGarbageCollectors
```

[6]:

```
srdf.show()
```

```
+-------+--------+------+--------------------+------------------+---------------
----+
|roll_no|    name|passed|               marks|            sports|
DoB|
+-------+--------+------+--------------------+------------------+---------------
----+
|      1|    john|  true|{Math -> 89, Chem…| [chess, football]|2012-05-01
12:01:05|
|      2| Vignesh| false|{Math -> 95, Chem…|  [carrom, tennis]|2012-05-12
14:02:05|
|      3|Sidharth|  true|{Math -> 95, Chem…|[football, kabadi]|2012-05-14
12:02:05|
+-------+--------+------+--------------------+------------------+---------------
----+
```

0.0.4 Create Temporary View

[7]:

```
srdf.createOrReplaceTempView('records')
```

[8]:

```
spark.sql("SELECT * FROM records").show()
```

```
+-------+--------+------+--------------------+------------------+---------------
----+
|roll_no|    name|passed|               marks|            sports|
DoB|
+-------+--------+------+--------------------+------------------+---------------
----+
|      1|    john|  true|{Math -> 89, Chem…| [chess, football]|2012-05-01
12:01:05|
|      2| Vignesh| false|{Math -> 95, Chem…|  [carrom, tennis]|2012-05-12
14:02:05|
|      3|Sidharth|  true|{Math -> 95, Chem…|[football, kabadi]|2012-05-14
12:02:05|
+-------+--------+------+--------------------+------------------+---------------
----+
[9]: re=spark.sql("SELECT * FROM records")
     type(re)
```

```
[9]: pyspark.sql.dataframe.DataFrame
```

0.0.5 Accessing Elements of List or Dictionary within DataFrame

```
[10]: spark.sql('SELECT roll_no,marks["Physics"],sports[1] FROM records').show()
```

```
+-------+-------------+---------+
|roll_no|marks[Physics]|sports[1]|
+-------+-------------+---------+
|      1|           87| football|
|      2|           66|   tennis|
|      3|          100|   kabadi|
+-------+-------------+---------+
```

0.0.6 Usage of Where Clause

```
[11]: spark.sql("SELECT * FROM records where passed= True").show()
```

```
+-------+--------+------+--------------------+-------------------+---------------
----+
|roll_no|    name|passed|               marks|             sports|
DoB|
+-------+--------+------+--------------------+-------------------+---------------
----+
|      1|    john|  true|{Math -> 89, Chem…|  [chess, football]|2012-05-01
12:01:05|
|      3|Sidharth|  true|{Math -> 95, Chem…|[football, kabadi]|2012-05-14
12:02:05|
+-------+--------+------+--------------------+-------------------+---------------
----+
```

```
[12]: spark.sql('SELECT * FROM records where marks["Chemistry"]<40').show()
```

```
+-------+----+------+-----+------+---+
|roll_no|name|passed|marks|sports|DoB|
+-------+----+------+-----+------+---+
+-------+----+------+-----+------+---+
```

0.0.7 Creating Global View

```
[13]: srdf.createGlobalTempView('globalrecord')
```

[14]:

```
spark.sql("SELECT * FROM global_temp.globalrecord").show()
```

```
+-------+--------+------+--------------------+------------------+---------------
----+
|roll_no|    name|passed|               marks|            sports|
DoB|
+-------+--------+------+--------------------+------------------+---------------
----+
|      1|    john|  true|{Math -> 89, Chem...| [chess, football]|2012-05-01
12:01:05|
|      2| Vignesh| false|{Math -> 95, Chem...|  [carrom, tennis]|2012-05-12
14:02:05|
|      3|Sidharth|  true|{Math -> 95, Chem...|[football, kabadi]|2012-05-14
12:02:05|
+-------+--------+------+--------------------+------------------+---------------
----+
```

0.0.8 Dropping Columns from DataFrame

[15]:

```
srdf.columns
```

[15]:

```
['roll_no', 'name', 'passed', 'marks', 'sports', 'DoB']
```

[16]:

```
srdf=srdf.drop('passed')
```

0.0.9 Few more Queries

This covers the usage of some additional functions like avg, sum, etc.

[17]:

```
spark.sql("SELECT round((marks.Physics+marks.Chemistry+marks.Math)/3)avg_marks␣
 ↪FROM records").show()
```

```
+---------+
|avg_marks|
+---------+
|     91.0|
|     79.0|
|     97.0|
+---------+
```

[18]:

```
srdf=spark.sql("SELECT *,round((marks.Physics+marks.Chemistry+marks.Math)/
 ↪3)avg_marks FROM records")
```

[19]:

```
srdf.show()
```

```
+-------+--------+------+--------------------+------------------+---------------
----+---------+
|roll_no|    name|passed|               marks|            sports|
DoB|avg_marks|
+-------+--------+------+--------------------+------------------+---------------
----+---------+
|      1|    john|  true|{Math -> 89, Chem…| [chess, football]|2012-05-01
12:01:05|     91.0|
|      2| Vignesh| false|{Math -> 95, Chem…|  [carrom, tennis]|2012-05-12
14:02:05|     79.0|
|      3|Sidharth|  true|{Math -> 95, Chem…|[football, kabadi]|2012-05-14
12:02:05|     97.0|
+-------+--------+------+--------------------+------------------+---------------
----+---------+
```