dataFrameOperationsBasics

February 8, 2026

0.1 Basic Data Frame Operations

[15]:
```
from pyspark.sql import SparkSession

spark=SparkSession.builder.appName("Data Frame").getOrCreate()

sc=spark.sparkContext
sc.stop()
```

the entry point to programming spark with the data frame is spark session

with a sql context, applications can create data frames from an existing RDD, from a hive table or from data sources

[16]:
```
from pyspark import SparkContext
from pyspark.sql import SparkSession,SQLContext

# old way of creating spark context
sc1=SparkContext(master='local',appName='test1')

# creating spark session
spark1=SparkSession(sc1)

# creating SQL context
sqlcontext1=SQLContext(sc1)
```

```
c:\Users\emada\AppData\Local\Programs\Python\Python311\Lib\site-
packages\pyspark\sql\context.py:112: FutureWarning: Deprecated in 3.0.0. Use
SparkSession.builder.getOrCreate() instead.
  warnings.warn(
```

[ ]:
```
spark=SparkSession.builder.appName("test").getOrCreate()

# for rdd learning
sc=spark.sparkContext
rdd=sc.textFile("test.csv")

#  for dataframe purpose
df=spark.read.csv('test.csv',header=True)
```

[20]:
```
# Importing data using spark session
df1=spark1.read.csv(path='iris/iris.csv',sep=',',header=True)

# importing data using sql context
df2=sqlcontext1.read.csv(path='iris/iris.csv',sep=',',header=True)
# sqlcontext is old api, used for sql & dataframes(old style)

print(df1)
# print(df1.show())
print(df2)

iris1_df1=spark1.read.json('iris/iris.json')

iris1_df1.show()
```

```
DataFrame[Sepal_Length: string, Sepal_Width: string, Petal_Length: string,
Petal_Width: string, Species: string]
DataFrame[Sepal_Length: string, Sepal_Width: string, Petal_Length: string,
Petal_Width: string, Species: string]
+------------+-----------+------------+-----------+-------+
|Petal_Length|Petal_Width|Sepal_Length|Sepal_Width|Species|
+------------+-----------+------------+-----------+-------+
|         1.4|        0.2|         5.1|        3.5| setosa|
|         1.4|        0.2|         4.9|        3.0| setosa|
|         1.3|        0.2|         4.7|        3.2| setosa|
|         1.5|        0.2|         4.6|        3.1| setosa|
|         1.4|        0.2|         5.0|        3.6| setosa|
|         1.7|        0.4|         5.4|        3.9| setosa|
|         1.4|        0.3|         4.6|        3.4| setosa|
|         1.5|        0.2|         5.0|        3.4| setosa|
|         1.4|        0.2|         4.4|        2.9| setosa|
|         1.5|        0.1|         4.9|        3.1| setosa|
|         1.5|        0.2|         5.4|        3.7| setosa|
|         1.6|        0.2|         4.8|        3.4| setosa|
|         1.4|        0.1|         4.8|        3.0| setosa|
|         1.1|        0.1|         4.3|        3.0| setosa|
|         1.2|        0.2|         5.8|        4.0| setosa|
|         1.5|        0.4|         5.7|        4.4| setosa|
|         1.3|        0.4|         5.4|        3.9| setosa|
|         1.4|        0.3|         5.1|        3.5| setosa|
|         1.7|        0.3|         5.7|        3.8| setosa|
|         1.5|        0.3|         5.1|        3.8| setosa|
+------------+-----------+------------+-----------+-------+
only showing top 20 rows
```

[22]:
```
# Convert RDD to Data Frame
# using createDataFrame function
iris1=sc1.textFile('iris/iris_site.csv')
```

```
iris1_split=iris1.map(lambda line:line.split(","))

df1=spark1.createDataFrame(iris1_split)
df1.show(10)
```

```
+---+---+---+---+------+
| _1| _2| _3| _4|    _5|
+---+---+---+---+------+
|5.1|3.5|1.4|0.2|setosa|
|4.9|3.0|1.4|0.2|setosa|
|4.7|3.2|1.3|0.2|setosa|
|4.6|3.1|1.5|0.2|setosa|
|5.0|3.6|1.4|0.2|setosa|
|5.4|3.9|1.7|0.4|setosa|
|4.6|3.4|1.4|0.3|setosa|
|5.0|3.4|1.5|0.2|setosa|
|4.4|2.9|1.4|0.2|setosa|
|4.9|3.1|1.5|0.1|setosa|
+---+---+---+---+------+
only showing top 10 rows
```

[23]:
```
# convert dataframe to rdd
iris1_df1=spark1.read.csv('iris/iris.csv',sep=',',header=True)
iris1_df1.rdd.map(tuple).take(10)
```

[23]:
```
[('5.1', '3.5', '1.4', '0.2', 'setosa'),
 ('4.9', '3.0', '1.4', '0.2', 'setosa'),
 ('4.7', '3.2', '1.3', '0.2', 'setosa'),
 ('4.6', '3.1', '1.5', '0.2', 'setosa'),
 ('5.0', '3.6', '1.4', '0.2', 'setosa'),
 ('5.4', '3.9', '1.7', '0.4', 'setosa'),
 ('4.6', '3.4', '1.4', '0.3', 'setosa'),
 ('5.0', '3.4', '1.5', '0.2', 'setosa'),
 ('4.4', '2.9', '1.4', '0.2', 'setosa'),
 ('4.9', '3.1', '1.5', '0.1', 'setosa')]
```

[25]:
```
# Display contents of data frame in table format
iris_df1=spark1.read.csv('iris/iris.csv',sep=',',header=True)
iris1_df1.show(5) # shows only top 5 rows

iris1_df1.collect() # display content of dataframe as a list of rows

iris1_df1.head(10) # shows 1st 1 rows of data frame as a list of rows
```

```
+------------+-----------+------------+-----------+-------+
|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width|Species|
+------------+-----------+------------+-----------+-------+
|         5.1|        3.5|         1.4|        0.2| setosa|
```

```
|          4.9|          3.0|          1.4|          0.2| setosa|
|          4.7|          3.2|          1.3|          0.2| setosa|
|          4.6|          3.1|          1.5|          0.2| setosa|
|          5.0|          3.6|          1.4|          0.2| setosa|
+-------------+-------------+-------------+-------------+-------+
only showing top 5 rows
```

[25]:
```
[Row(Sepal_Length='5.1', Sepal_Width='3.5', Petal_Length='1.4',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='4.9', Sepal_Width='3.0', Petal_Length='1.4',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='4.7', Sepal_Width='3.2', Petal_Length='1.3',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='4.6', Sepal_Width='3.1', Petal_Length='1.5',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='5.0', Sepal_Width='3.6', Petal_Length='1.4',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='5.4', Sepal_Width='3.9', Petal_Length='1.7',
Petal_Width='0.4', Species='setosa'),
 Row(Sepal_Length='4.6', Sepal_Width='3.4', Petal_Length='1.4',
Petal_Width='0.3', Species='setosa'),
 Row(Sepal_Length='5.0', Sepal_Width='3.4', Petal_Length='1.5',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='4.4', Sepal_Width='2.9', Petal_Length='1.4',
Petal_Width='0.2', Species='setosa'),
 Row(Sepal_Length='4.9', Sepal_Width='3.1', Petal_Length='1.5',
Petal_Width='0.1', Species='setosa')]
```

[26]:
```
# Data Selection
# selecting any particular column
iris1_df1=spark1.read.csv('iris/iris.csv',sep=',',header=True)
iris1_df1.select("Sepal_Length","Species").show()
```

```
+------------+-------+
|Sepal_Length|Species|
+------------+-------+
|         5.1| setosa|
|         4.9| setosa|
|         4.7| setosa|
|         4.6| setosa|
|         5.0| setosa|
|         5.4| setosa|
|         4.6| setosa|
|         5.0| setosa|
|         4.4| setosa|
|         4.9| setosa|
|         5.4| setosa|
|         4.8| setosa|
```

```
|         4.8| setosa|
|         4.3| setosa|
|         5.8| setosa|
|         5.7| setosa|
|         5.4| setosa|
|         5.1| setosa|
|         5.7| setosa|
|         5.1| setosa|
+------------+-------+
only showing top 20 rows
```

[28]:
```
# Joins
iris1_df1 = spark1.read.csv(path="iris/merge/iris_merge1.csv", sep=",",␣
 ↪header=True)
iris1_df2 = spark1.read.csv(path="iris/merge/iris_merge2.csv", sep=",",␣
 ↪header=True)

iris1_df1.join(other=iris1_df2,on='ID',how='inner').show()
```

```
+---+------------+-----------+------------+-----------+-------+
| ID|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width|Species|
+---+------------+-----------+------------+-----------+-------+
|  1|         5.1|        3.5|         1.4|        0.2| setosa|
|  2|         4.9|          3|         1.4|        0.2| setosa|
|  3|         4.7|        3.2|         1.3|        0.2| setosa|
|  4|         4.6|        3.1|         1.5|        0.2| setosa|
|  5|           5|        3.6|         1.4|        0.2| setosa|
|  6|         5.4|        3.9|         1.7|        0.4| setosa|
|  7|         4.6|        3.4|         1.4|        0.3| setosa|
|  8|           5|        3.4|         1.5|        0.2| setosa|
|  9|         4.4|        2.9|         1.4|        0.2| setosa|
| 10|         4.9|        3.1|         1.5|        0.1| setosa|
| 11|         5.4|        3.7|         1.5|        0.2| setosa|
| 12|         4.8|        3.4|         1.6|        0.2| setosa|
| 13|         4.8|          3|         1.4|        0.1| setosa|
| 14|         4.3|          3|         1.1|        0.1| setosa|
| 15|         5.8|          4|         1.2|        0.2| setosa|
| 16|         5.7|        4.4|         1.5|        0.4| setosa|
| 17|         5.4|        3.9|         1.3|        0.4| setosa|
| 18|         5.1|        3.5|         1.4|        0.3| setosa|
| 19|         5.7|        3.8|         1.7|        0.3| setosa|
| 20|         5.1|        3.8|         1.5|        0.3| setosa|
+---+------------+-----------+------------+-----------+-------+
only showing top 20 rows
```

[29]:
```
# once two dataframes are joined, required columns from the two tables
# can be retrieved using select function along with the join function
```

```
iris1_df1.join(other=iris1_df2,on='ID',how='inner').select(iris1_df1.
  ↪Sepal_Length,iris1_df2.Petal_Length).show()
```

```
+------------+------------+
|Sepal_Length|Petal_Length|
+------------+------------+
|         5.1|         1.4|
|         4.9|         1.4|
|         4.7|         1.3|
|         4.6|         1.5|
|           5|         1.4|
|         5.4|         1.7|
|         4.6|         1.4|
|           5|         1.5|
|         4.4|         1.4|
|         4.9|         1.5|
|         5.4|         1.5|
|         4.8|         1.6|
|         4.8|         1.4|
|         4.3|         1.1|
|         5.8|         1.2|
|         5.7|         1.5|
|         5.4|         1.3|
|         5.1|         1.4|
|         5.7|         1.7|
|         5.1|         1.5|
+------------+------------+
only showing top 20 rows
```

[30]:
```
# joining two tables where the joining columns present in the two
# tables have a different name
iris1_df1.join(other=iris1_df2,on=(iris1_df1.ID==iris1_df2.ID),how='inner').
  ↪show()
```

```
+------------+-----------+---+---+------------+-----------+-------+
|Sepal_Length|Sepal_Width| ID| ID|Petal_Length|Petal_Width|Species|
+------------+-----------+---+---+------------+-----------+-------+
|         5.1|        3.5|  1|  1|         1.4|        0.2| setosa|
|         4.9|          3|  2|  2|         1.4|        0.2| setosa|
|         4.7|        3.2|  3|  3|         1.3|        0.2| setosa|
|         4.6|        3.1|  4|  4|         1.5|        0.2| setosa|
|           5|        3.6|  5|  5|         1.4|        0.2| setosa|
|         5.4|        3.9|  6|  6|         1.7|        0.4| setosa|
|         4.6|        3.4|  7|  7|         1.4|        0.3| setosa|
|           5|        3.4|  8|  8|         1.5|        0.2| setosa|
|         4.4|        2.9|  9|  9|         1.4|        0.2| setosa|
|         4.9|        3.1| 10| 10|         1.5|        0.1| setosa|
|         5.4|        3.7| 11| 11|         1.5|        0.2| setosa|
```

```
|         4.8|        3.4| 12| 12|         1.6|        0.2| setosa|
|         4.8|          3| 13| 13|         1.4|        0.1| setosa|
|         4.3|          3| 14| 14|         1.1|        0.1| setosa|
|         5.8|          4| 15| 15|         1.2|        0.2| setosa|
|         5.7|        4.4| 16| 16|         1.5|        0.4| setosa|
|         5.4|        3.9| 17| 17|         1.3|        0.4| setosa|
|         5.1|        3.5| 18| 18|         1.4|        0.3| setosa|
|         5.7|        3.8| 19| 19|         1.7|        0.3| setosa|
|         5.1|        3.8| 20| 20|         1.5|        0.3| setosa|
+------------+-----------+---+---+------------+-----------+-------+
only showing top 20 rows
```

[32]:
```
# UNION

# two data frames with similar structures can be joined row-wise using the␣
 ↪union function
iris1_df1=spark1.read.csv('iris/union/iris_union1.csv',sep=',',header=True)
iris1_df2=spark1.read.csv('iris/union/iris_union2.csv',sep=',',header=True)

iris1_df1.union(iris1_df2).show()
```

```
+------------+-----------+------------+-----------+
|Sepal.Length|Sepal.Width|Petal.Length|Petal.Width|
+------------+-----------+------------+-----------+
|           5|          3|           1|          0|
|         4.6|       NULL|           2|        0.1|
|         7.2|        3.1|         5.1|          1|
|           8|          4|           7|          2|
|          10|          6|           2|          0|
|         9.2|          0|           4|        0.2|
|        14.4|        6.2|        10.2|          2|
|          16|          8|          14|          4|
+------------+-----------+------------+-----------+
```