joins

February 9, 2026

[ ]:

```
from pyspark.sql import SparkSession

spark=SparkSession.builder.appName("Joins").getOrCreate()

iris_df1=spark.read.csv("iris/iris.csv",header=True,sep=",")
iris_df1.show(10)
```

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

[2]:

```
# left df + right df --> one df

from pyspark.sql import Row

species_df = spark.createDataFrame(
    [
        ("setosa", "small", "low"),
        ("versicolor", "medium", "medium"),
        ("virginica", "large", "high"),
    ],
    ["Species", "Size", "Risk"],
)

species_df.show()
```

```
+----------+------+------+
|   Species|  Size|  Risk|
+----------+------+------+
|    setosa| small|   low|
|versicolor|medium|medium|
| virginica| large|  high|
+----------+------+------+
```

[3]:

```
# Inner Join
# keeps rows only if match exists in both tables
iris_df1.join(species_df,on="Species",how="inner").show(10) # setosa rows␣
 ↪kept,species not found dropped
```

```
+-------+------------+-----------+------------+-----------+-----+----+
|Species|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width| Size|Risk|
+-------+------------+-----------+------------+-----------+-----+----+
| setosa|         5.0|        3.3|         1.4|        0.2|small| low|
| setosa|         5.3|        3.7|         1.5|        0.2|small| low|
| setosa|         4.6|        3.2|         1.4|        0.2|small| low|
| setosa|         5.1|        3.8|         1.6|        0.2|small| low|
| setosa|         4.8|        3.0|         1.4|        0.3|small| low|
| setosa|         5.1|        3.8|         1.9|        0.4|small| low|
| setosa|         5.0|        3.5|         1.6|        0.6|small| low|
| setosa|         4.4|        3.2|         1.3|        0.2|small| low|
| setosa|         4.5|        2.3|         1.3|        0.3|small| low|
| setosa|         5.0|        3.5|         1.3|        0.3|small| low|
+-------+------------+-----------+------------+-----------+-----+----+
only showing top 10 rows
```

[4]:

```
# left join
# keep all rows from left table,macthes from right if exists
iris_df1.join(species_df, on="Species", how="left").show(10) # if some species␣
 ↪left in speices_df size/risk becomes null
```

```
+-------+------------+-----------+------------+-----------+-----+----+
|Species|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width| Size|Risk|
+-------+------------+-----------+------------+-----------+-----+----+
| setosa|         5.1|        3.5|         1.4|        0.2|small| low|
| setosa|         4.9|        3.0|         1.4|        0.2|small| low|
| setosa|         4.7|        3.2|         1.3|        0.2|small| low|
| setosa|         4.6|        3.1|         1.5|        0.2|small| low|
| setosa|         5.0|        3.6|         1.4|        0.2|small| low|
| setosa|         5.4|        3.9|         1.7|        0.4|small| low|
| setosa|         4.6|        3.4|         1.4|        0.3|small| low|
| setosa|         5.0|        3.4|         1.5|        0.2|small| low|
| setosa|         4.4|        2.9|         1.4|        0.2|small| low|
| setosa|         4.9|        3.1|         1.5|        0.1|small| low|
+-------+------------+-----------+------------+-----------+-----+----+
only showing top 10 rows
```

[11]:

```
# right join
iris_df1.join(species_df,on="Species",how="right").show(10)
```

```
+-------+------------+-----------+------------+-----------+-----+----+
|Species|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width| Size|Risk|
+-------+------------+-----------+------------+-----------+-----+----+
| setosa|         5.0|        3.3|         1.4|        0.2|small| low|
| setosa|         5.3|        3.7|         1.5|        0.2|small| low|
| setosa|         4.6|        3.2|         1.4|        0.2|small| low|
| setosa|         5.1|        3.8|         1.6|        0.2|small| low|
| setosa|         4.8|        3.0|         1.4|        0.3|small| low|
| setosa|         5.1|        3.8|         1.9|        0.4|small| low|
| setosa|         5.0|        3.5|         1.6|        0.6|small| low|
| setosa|         4.4|        3.2|         1.3|        0.2|small| low|
| setosa|         4.5|        2.3|         1.3|        0.3|small| low|
| setosa|         5.0|        3.5|         1.3|        0.3|small| low|
+-------+------------+-----------+------------+-----------+-----+----+
only showing top 10 rows
```

[ ]:

```
# left semi join
# --> return rows from left df
# --> only checks existence in right df
# --> does not add columns from right
iris_df1.join(species_df, on="Species", how="left_semi").show(10)
# equivalent to Where species in subquery
```

```
+-------+------------+-----------+------------+-----------+
|Species|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width|
+-------+------------+-----------+------------+-----------+
| setosa|         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|
+-------+------------+-----------+------------+-----------+
only showing top 10 rows
```

[ ]:

```
# Left anti join
# returns rows from left df where no match in right df
iris_df1.join(species_df, on="Species", how="left_anti").show()
# equivalent to where species not in subquery
```

```
+-------+------------+-----------+------------+-----------+
|Species|Sepal_Length|Sepal_Width|Petal_Length|Petal_Width|
+--------+------------+-----------+------------+-----------+
+--------+------------+-----------+------------+-----------+
```

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

[7]:
```
# 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
```

[9]:

```python
# UNION

# two data frames with similar structures can be joined row-wise using the union function
iris1_df1 = spark.read.csv("iris/union/iris_union1.csv", sep=",", header=True)
iris1_df2 = spark.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|
+------------+-----------+------------+-----------+
```