Join() is like SQL join. We can combine columns from different DataFrames based on conditions. It supports all basic join types such as Inner, left outer, right outer, left anti, left semi, cross, self.

1. Leftsemi() join similar to inner join but get columns only from left dataframe for matching rows.
2. Leftanti() opposite to leftsemi,it gets non matching rows from left dataframe.
3. Selfjoin joins data with same dataframe.

**Example-1 – Inner, left, right, full**

from pyspark.sql import \*  
from pyspark.sql.functions import \*  
  
spark = SparkSession.builder.appName('join()').getOrCreate()  
  
data1 = [(1,'Harsha',2000,2),  
 (2,'Mokshit',4000,1),  
 (3,'Harika',3500,4)]  
schema1 = ['id','name','salary','dept']  
  
data2 = [(1,'IT'),(2,'HR'),(3,'Payroll')]  
schema2 = ['id','name']  
  
emp\_df = spark.createDataFrame(data1,schema1)  
dept\_df = spark.createDataFrame(data2,schema2)  
  
emp\_df.join(dept\_df, emp\_df.dept == dept\_df.id,'inner').show()  
emp\_df.join(dept\_df, emp\_df.dept == dept\_df.id,'left').show()  
emp\_df.join(dept\_df, emp\_df.dept == dept\_df.id,'right').show()  
emp\_df.join(dept\_df, emp\_df.dept == dept\_df.id,'full').show()  
emp\_df.join(dept\_df, emp\_df.dept==dept\_df.id,'leftsemi').show()  
emp\_df.join(dept\_df, emp\_df.dept==dept\_df.id,'leftanti').show()

**Output:**

+---+-------+------+----+---+----+

| id| name|salary|dept| id|name|

+---+-------+------+----+---+----+

| 2|Mokshit| 4000| 1| 1| IT|

| 1| Harsha| 2000| 2| 2| HR|

+---+-------+------+----+---+----+

+---+-------+------+----+----+----+

| id| name|salary|dept| id|name|

+---+-------+------+----+----+----+

| 2|Mokshit| 4000| 1| 1| IT|

| 1| Harsha| 2000| 2| 2| HR|

| 3| Harika| 3500| 4|null|null|

+---+-------+------+----+----+----+

+----+-------+------+----+---+-------+

| id| name|salary|dept| id| name|

+----+-------+------+----+---+-------+

| 2|Mokshit| 4000| 1| 1| IT|

| 1| Harsha| 2000| 2| 2| HR|

|null| null| null|null| 3|Payroll|

+----+-------+------+----+---+-------+

+----+-------+------+----+----+-------+

| id| name|salary|dept| id| name|

+----+-------+------+----+----+-------+

| 2|Mokshit| 4000| 1| 1| IT|

| 1| Harsha| 2000| 2| 2| HR|

|null| null| null|null| 3|Payroll|

| 3| Harika| 3500| 4|null| null|

+----+-------+------+----+----+-------+

+---+-------+------+----+

| id| name|salary|dept|

+---+-------+------+----+

| 2|Mokshit| 4000| 1|

| 1| Harsha| 2000| 2|

+---+-------+------+----+

+---+------+------+----+

| id| name|salary|dept|

+---+------+------+----+

| 3|Harika| 3500| 4|

+---+------+------+----+

**Self\_join() – Example**

from pyspark.sql import \*  
from pyspark.sql.functions import \*  
  
spark = SparkSession.builder.appName('selfjoin()').getOrCreate()  
  
data = [(1,'Harsha',0),(2,'Mokshit',1),(3,'Harika',2)]  
schema = ['id','name','managerid']  
  
df = spark.createDataFrame(data,schema)  
  
df1 = df.alias('e').join(df.alias('m'),col('e.managerid') == col('m.id'),'left')\  
 .select(col('e.name').alias('emp\_name'),  
 col('m.name').alias('mgr\_name'))  
  
df1.show()

**Output:**

+--------+--------+

|emp\_name|mgr\_name|

+--------+--------+

| Harsha| null|

| Mokshit| Harsha|

| Harika| Mokshit|

+--------+--------+