join Transformation

When called on datasets of type (K, V) and (K, W), returns a dataset of (K, (V, W)) pairs with all pairs of elements for each key. Outer joins are supported through leftOuterJoin, rightOuterJoin, and fullOuterJoin. Below, these are explored.

syntax of join for DataFrame - join(dataset,joinExprs,joinType)

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

spark = SparkSession \
        .builder \
        .master("local[2]") \
        .appName("Join Transformation") \
        .enableHiveSupport() \
        .getOrCreate()

#Key-value pairs from dataset 1
kvPair = [(1,"v"), (1,"i"), (2, "s"), (3,"h"), (4,"a") ]

kvPairRDD = spark.sparkContext.parallelize(kvPair)
print(kvPairRDD.collect())

#Key-value pairs from dataset 2
otherKvPair = [(1,"y"), (1,"l"), (2, "d"), (3,"m"), (8,"z2") ]

otherKvPairRDD = spark.sparkContext.parallelize(otherKvPair)
print(otherKvPairRDD.collect())


#join() transformation to join both datasets to return (K, (V,W))
joined = kvPairRDD.join(otherKvPairRDD)
print("joined.count(): ", joined.count())
print("joined.collect(): ", joined.collect())
```

```
22/10/14 00:33:10 WARN Utils: Your hostname, Vaishalis-MacBook-Pro.local resolves to a loopback address: 127.0.0.1; using 192.168.0.105 instead (on interface en0)
22/10/14 00:33:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/10/14 00:33:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

[(1, 'v'), (1, 'i'), (2, 's'), (3, 'h'), (4, 'a')]
[(1, 'y'), (1, 'l'), (2, 'd'), (3, 'm'), (8, 'z2')]
```

```
joined.count():  6
joined.collect():  [(1, ('v', 'y')), (1, ('v', 'l')), (1, ('i', 'y')), (1, ('i', 'l')), (2, ('s', 'd')), (3, ('h', 'm'))]
```

In [2]:
```
#Let us look at another example to understand various joinType

data1 = [(1,"John","-1,"2019","10","M"), \
      (2,"Alex",1,"2015","20","M"), \
      (3,"Williams",1,"2016","10","M"), \
      (4,"Arjun",2,"2010","10","F"), \
      (5,"Brown",2,"2012","40",""), \
      (6,"Brown",2,"2012","50","") \
   ]

data1_columns = ["employee_id", "name", "superior_emp_id", "year_joined", "emp_dept_id","gender"]

df1 = spark.createDataFrame(data= data1, schema = data1_columns)
df1.printSchema()
df1.show(truncate=False)
```

```
root
 |-- employee_id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- superior_emp_id: long (nullable = true)
 |-- year_joined: string (nullable = true)
 |-- emp_dept_id: string (nullable = true)
 |-- gender: string (nullable = true)

+-----------+--------+---------------+-----------+-----------+------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|
+-----------+--------+---------------+-----------+-----------+------+
|1          |John    |-1             |2019       |10         |M     |
|2          |Alex    |1              |2015       |20         |M     |
|3          |Williams|1              |2016       |10         |M     |
|4          |Arjun   |2              |2010       |10         |F     |
|5          |Brown   |2              |2012       |40         |      |
|6          |Brown   |2              |2012       |50         |      |
+-----------+--------+---------------+-----------+-----------+------+
```

In [3]:
```
data2 = [("Finance",10), \
         ("Marketing",20), \
         ("Sales",30), \
         ("IT",40)]

data2_columns = ["dept_name","dept_id"]

df2 = spark.createDataFrame(data = data2, schema = data2_columns)
df2.printSchema()
df2.show(truncate=False)
```

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

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

In [4]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"inner") \
     .show(truncate=False)
```

```
[Stage 10:=============================================================(2 + 0) / 2]
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [5]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"cross") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [6]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"outer") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|null       |null    |null           |null       |null       |null  |Sales    |30     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
|6          |Brown   |2              |2012       |50         |      |null     |null   |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [7]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"fullouter") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|null       |null    |null           |null       |null       |null  |Sales    |30     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
|6          |Brown   |2              |2012       |50         |      |null     |null   |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [8]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"left") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|6          |Brown   |2              |2012       |50         |      |null     |null   |
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [9]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"leftouter") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|6          |Brown   |2              |2012       |50         |      |null     |null   |
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [10]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"right") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|null       |null    |null           |null       |null       |null  |Sales    |30     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [11]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"rightouter") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|dept_name|dept_id|
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
|4          |Arjun   |2              |2010       |10         |F     |Finance  |10     |
|3          |Williams|1              |2016       |10         |M     |Finance  |10     |
|1          |John    |-1             |2019       |10         |M     |Finance  |10     |
|2          |Alex    |1              |2015       |20         |M     |Marketing|20     |
|null       |null    |null           |null       |null       |null  |Sales    |30     |
|5          |Brown   |2              |2012       |40         |      |IT       |40     |
+-----------+--------+---------------+-----------+-----------+------+---------+-------+
```

In [12]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"leftsemi") \
     .show(truncate=False)
```

```
+-----------+--------+---------------+-----------+-----------+------+
|employee_id|name    |superior_emp_id|year_joined|emp_dept_id|gender|
+-----------+--------+---------------+-----------+-----------+------+
|1          |John    |-1             |2019       |10         |M     |
|3          |Williams|1              |2016       |10         |M     |
|4          |Arjun   |2              |2010       |10         |F     |
|2          |Alex    |1              |2015       |20         |M     |
|5          |Brown   |2              |2012       |40         |      |
+-----------+--------+---------------+-----------+-----------+------+
```

In [13]:
```
df1.join(df2, df1.emp_dept_id == df2.dept_id,"leftanti") \
     .show(truncate=False)
```

```
+-----------+-----+---------------+-----------+-----------+------+
|employee_id|name |superior_emp_id|year_joined|emp_dept_id|gender|
+-----------+-----+---------------+-----------+-----------+------+
|6          |Brown|2              |2012       |50         |      |
+-----------+-----+---------------+-----------+-----------+------+
```