

Custom accumulator

- The `zero()` function is to initialize the accumulator
- The `addInPlace()` is the actual counter

In [2]:

```
from pyspark.accumulators import AccumulatorParam

class VectorAccumulatorParam(AccumulatorParam):

    def zero(self, value):
        return [0.0] * len(value)

    def addInPlace(self, v1, v2):
        for i in range(len(v1)):
            v1[i] += v2[i]

        return v1
```

In [3]:

```
vector_accum = sc.accumulator([10.0, 20.0, 30.0], VectorAccumulatorParam())

vector_accum.value
```

Out[3]:

```
[10.0, 20.0, 30.0]
```

In [4]:

```
vector_accum += [1, 2, 3]

vector_accum.value
```

Out[4]:

```
[11.0, 22.0, 33.0]
```

In [1]:

```
sc
```

Out[1]:

SparkContext

[Spark UI](#)

Version

```
v2.3.0
```

Master

```
local[*]
```

AppName

```
PySparkShell
```

Setting up the Data in Pyspark

In [5]:

```
valuesA = [('John', 100000), ('James', 150000), ('Emily', 65000), ('Nina', 200000)]

tableA = spark.createDataFrame(valuesA, ['name', 'salary'])
```

In [6]:

```
tableA.show()
```

```
+-----+-----+
| name|salary|
+-----+-----+
| John|100000|
|James|150000|
|Emily| 65000|
| Nina|200000|
+-----+-----+
```

In [7]:

```
valuesB = [('James', 2), ('Emily', 3), ('Darth Vader', 5), ('Princess Leia', 6),]
```

```
tableB = spark.createDataFrame(valuesB, ['name', 'employee_id'])
```

In [8]:

```
tableB.show()
```

```
+-----+-----+
|      name|employee_id|
+-----+-----+
|      James|          2|
|      Emily|          3|
|Darth Vader|          5|
|Princess Leia|        6|
+-----+-----+
```

Inner join

In [9]:

```
inner_join = tableA.join(tableB, tableA.name == tableB.name)
inner_join.show()
```

```
+-----+-----+-----+-----+
| name|salary| name|employee_id|
+-----+-----+-----+-----+
|James|150000|James|          2|
|Emily| 65000|Emily|          3|
+-----+-----+-----+-----+
```

Left outer join

In [10]:

```
left_join = tableA.join(tableB, tableA.name == tableB.name, how='left')
left_join.show()
```

```
+-----+-----+-----+-----+
| name|salary| name|employee_id|
+-----+-----+-----+-----+
|James|150000|James|          2|
| John|100000| null|        null|
|Emily| 65000|Emily|          3|
| Nina|200000| null|        null|
+-----+-----+-----+-----+
```

Right outer join

In [11]:

```
right_join = tableA.join(tableB, tableA.name == tableB.name, how='right')
right_join.show()
```

```
+-----+-----+-----+-----+
| name|salary|      name|employee_id|
+-----+-----+-----+-----+
|James|150000|      James|         2|
| null|  null|Princess Leia|         6|
|Emily| 65000|      Emily|         3|
| null|  null|  Darth Vader|         5|
+-----+-----+-----+-----+
```

Full outer join

In [12]:

```
full_outer_join = tableA.join(tableB, tableA.name == tableB.name, how='full')
full_outer_join.show()
```

```
+-----+-----+-----+-----+
| name|salary|      name|employee_id|
+-----+-----+-----+-----+
|James|150000|      James|         2|
| John|100000|      null|        null|
| null|  null|Princess Leia|         6|
|Emily| 65000|      Emily|         3|
| Nina|200000|      null|        null|
| null|  null|  Darth Vader|         5|
+-----+-----+-----+-----+
```

In []:

In []:

In []:

In []:

In []:

In []: