In [1]:

```
# import findspark
# findspark.init()
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
from pyspark.sql.types import *
```

In [2]:

```
spark = SparkSession\
    .builder\
    .appName("chapter-15-cluster")\
    .getOrCreate()
```

In [4]:

```
import os
SPARK_BOOK_DATA_PATH = os.environ['SPARK_BOOK_DATA_PATH']
```

In [5]:

```
spark
```

Out[5]:

SparkSession - hive

SparkContext

[Spark UI (http://192.168.1.2:4044)](http://192.168.1.2:4044)

Version

`v2.4.3`

Master

`local[*]`

AppName

`PySparkShell`

In [6]:

```
df1 = spark.range(2, 10000000, 2)
df2 = spark.range(2, 10000000, 4)
step1 = df1.repartition(5)
step12 = df2.repartition(6)
step2 = step1.selectExpr("id * 5 as id")
step3 = step2.join(step12, ["id"])
step4 = step3.selectExpr("sum(id)")

step4.collect() # 2500000000000
```

Out[6]:

```
[Row(sum(id)=2500000000000)]
```

In [7]:
```
step4.explain()
```

```
== Physical Plan ==
*(7) HashAggregate(keys=[], functions=[sum(id#6L)])
+- Exchange SinglePartition
   +- *(6) HashAggregate(keys=[], functions=[partial_sum(id#6L)])
      +- *(6) Project [id#6L]
         +- *(6) SortMergeJoin [id#6L], [id#2L], Inner
            :- *(3) Sort [id#6L ASC NULLS FIRST], false, 0
            :  +- Exchange hashpartitioning(id#6L, 200)
            :     +- *(2) Project [(id#0L * 5) AS id#6L]
            :        +- Exchange RoundRobinPartitioning(5)
            :           +- *(1) Range (2, 10000000, step=2, splits=4)
            +- *(5) Sort [id#2L ASC NULLS FIRST], false, 0
               +- Exchange hashpartitioning(id#2L, 200)
                  +- Exchange RoundRobinPartitioning(6)
                     +- *(4) Range (2, 10000000, step=4, splits=4)
```

In [8]:
```
print(spark.range(1000).where("id > 500").selectExpr("sum(id)").collect
```

```
[Row(sum(id)=374250)]
```

In [9]:
```
print(spark.range(11).where("id > 0").selectExpr("sum(id)").collect())
```

```
[Row(sum(id)=55)]
```

In [10]:
```
df = spark.range(11)
```

In [11]:
```
df.show()
```

```
+---+
| id|
+---+
|  0|
|  1|
|  2|
|  3|
|  4|
|  5|
|  6|
|  7|
|  8|
|  9|
| 10|
+---+
```

Spark UI

In [14]:
```
file_path = SPARK_BOOK_DATA_PATH + "/data/retail-data/all/online-retail
```

In [15]:

```python
spark.read\
  .option("header", "true")\
  .csv(file_path)\
  .repartition(2)\
  .selectExpr("instr(Description, 'GLASS') >= 1 as is_glass")\
  .groupBy("is_glass")\
  .count()\
  .collect()
```

Out[15]:

```
[Row(is_glass=None, count=1454),
 Row(is_glass=True, count=12861),
 Row(is_glass=False, count=527594)]
```

In [ ]: