distinct( ) Transformation

This transformation is used to ensure there are no duplicates in the dataset

In [2]:
```
from pyspark.sql import SparkSession

spark = SparkSession \
        .builder \
        .master("local[*]") \
        .appName("Sample Transformation") \
        .getOrCreate()

#Two lists
list1 = [1, 2, 1, 4, 5, 3, 2, 5, 1, 10]

rdd = spark.sparkContext.parallelize(list1, 4)

#with distinct(), we find the output with not any duplicates.
distinct_rdd = rdd.distinct()

print(distinct_rdd.collect())
```

```
[4, 1, 5, 2, 10, 3]
```

In [6]:
```
#Applying the distinct() transformation on a dataset
str_rdd = spark.sparkContext.parallelize(['hi','John','how','are','you','doing','David how','how', 'coping'])


str_rdd = str_rdd.distinct()
print(str_rdd.collect())
```

```
['how', 'you', 'are', 'coping', 'David how', 'hi', 'John', 'doing']
```