Sample Transformation

Sample() Transformation is get random sample records from the dataset. This can be helpful when there is a larger dataset and wanted to test a subset of the data.

Syntax of the Sample() Transformation:

sample(WithReplacement, fraction, seed=None)

fraction - Fraction of rows to generate, range[0.0,1.0]. Note that it doesn't guarantee to provide the exact number the exact number of the fraction of records.

seed - Seed for sampling (default a random seed). Used to reproduce the same random sampling.

withReplacement - Sample with replacement or not (default False).

In [7]:
```
from pyspark.sql import SparkSession

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

#DataFrame created with 100 records and getting a 5% sample of records
rdd = spark.range(100)
print(rdd.sample(0.05).show())
```

```
+---+
| id|
+---+
|  0|
| 49|
| 60|
| 85|
+---+

None
```

In [2]:
```
#Get 5% sample of records and seed is used to reproduce the same random sample whenever there is a need to
#regenerate the same results as tested at a different phase.

print(rdd.sample(0.05, 25).show())
```

```
+---+
| id|
+---+
| 10|
| 49|
| 88|
| 93|
+---+

None
```

In [5]:
```
#Different Results with only fraction usage
print(rdd.sample(0.05).show())

#the same results as above for the same seed.
print(rdd.sample(0.05,25).show())
```

```
+---+
| id|
+---+
| 13|
| 70|
| 75|
| 76|
+---+

None
+---+
| id|
+---+
| 10|
| 49|
| 88|
| 93|
+---+

None
```

In [6]:
```
#To get random sample with repeated values, withReplacement has to be True.

print(rdd.sample(True,0.5,25).show())

print(rdd.sample(False,0.5,25).show())
```

```
+---+
| id|
+---+
|  3|
|  3|
|  3|
|  4|
|  6|
|  8|
| 13|
| 19|
| 21|
| 21|
| 26|
| 30|
| 31|
| 31|
| 31|
| 35|
| 37|
| 38|
| 45|
| 45|
+---+
only showing top 20 rows

None
+---+
| id|
+---+
|  0|
|  1|
|  2|
|  4|
|  7|
|  9|
| 10|
| 12|
| 14|
| 16|
| 17|
| 19|
| 22|
| 23|
| 25|
| 27|
| 29|
| 31|
| 35|
| 37|
+---+
only showing top 20 rows

None
```