

In [1]: `sc`

Out[1]: **SparkContext**

[Spark UI](#)

Version	v4.0.0
Master	local[*]
AppName	PySparkShell

In [2]: `from pyspark.sql import SparkSession`
`# Step 1: Initialize Spark Session`
`spark = SparkSession.builder.appName("SamplingExample").getOrCreate()`

In [3]: `df = spark.read.csv("students.csv", header=True, inferSchema=True)`
`print("=== First 5 rows of dataset ===")`
`df.show(5)`

=== First 5 rows of dataset ===

id	name	age	gender	math	science	english
1	Alice	20	F	66	92	44
2	Bob	20	M	82	52	77
3	Charlie	22	F	43	57	76
4	David	19	M	95	69	46
5	Eva	19	F	62	44	96

only showing top 5 rows

In [4]: `print("=== Schema of dataset ===")`
`df.printSchema()`

=== Schema of dataset ===

```
root
 |-- id: integer (nullable = true)
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- gender: string (nullable = true)
 |-- math: integer (nullable = true)
 |-- science: integer (nullable = true)
 |-- english: integer (nullable = true)
```

In [5]: `print("=== Sample (30% without replacement) ===")`
`df.sample(withReplacement=False, fraction=0.3, seed=42).show(10)`

```

=== Sample (30% without replacement) ===
+---+-----+---+-----+---+-----+-----+
| id|  name|age|gender|math|science|english|
+---+-----+---+-----+---+-----+-----+
| 4| David|19|  M| 95|  69|  46|
| 8| Henry|21|  F| 53|  82|  60|
|17|Quincy|18|  M| 65|  79|  54|
|19|  Sam|18|  F| 76|  70|  65|
|27| Aaron|25|  F| 81|  99|  44|
|28| Bella|19|  F| 54|  76|  76|
|32| Fiona|22|  F| 48|  96|  48|
|37|  Kyle|21|  M| 57|  86|  92|
|39|  Matt|25|  M| 64|  71| 100|
|41| Oscar|20|  M| 87|  72|  81|
+---+-----+---+-----+---+-----+-----+
only showing top 10 rows

```

```

In [6]: print("=== Sample (20% with replacement) ===")
        df.sample(withReplacement=True, fraction=0.2, seed=42).show(10)

```

```

=== Sample (20% with replacement) ===
+---+-----+---+-----+---+-----+-----+
| id|  name|age|gender|math|science|english|
+---+-----+---+-----+---+-----+-----+
| 6| Frank|22|  F| 70|  78|  94|
| 7| Grace|24|  F| 67|  66|  93|
|14|Nathan|23|  F| 71|  66|  60|
|17|Quincy|18|  M| 65|  79|  54|
|21|  Uma|19|  F| 89|  70|  76|
|22|Victor|22|  M| 96|  75|  56|
|31| Ethan|24|  M| 53|  57|  45|
|32| Fiona|22|  F| 48|  96|  48|
|35|  Ian|21|  F| 72|  75|  70|
|38| Laura|23|  M| 84|  73|  56|
+---+-----+---+-----+---+-----+-----+
only showing top 10 rows

```

```

In [7]: print("=== takeSample: 5 rows (without replacement) ===")
        sampled_rows = df.rdd.takeSample(False, 5, seed=42)
        for row in sampled_rows:
            print(row)

```

```

=== takeSample: 5 rows (without replacement) ===
Row(id=35, name='Ian', age=21, gender='F', math=72, science=75, english=70)
Row(id=26, name='Zoey', age=18, gender='M', math=42, science=48, english=42)
Row(id=17, name='Quincy', age=18, gender='M', math=65, science=79, english=54)
Row(id=43, name='Quinn', age=18, gender='F', math=56, science=60, english=87)
Row(id=38, name='Laura', age=23, gender='M', math=84, science=73, english=56)

```

```

In [8]: print("=== takeSample: 5 rows (with replacement) ===")
        sampled_rows_wr = df.rdd.takeSample(True, 5, seed=42)
        for row in sampled_rows_wr:
            print(row)

```

```
=== takeSample: 5 rows (with replacement) ===  
Row(id=47, name='Umar', age=21, gender='F', math=75, science=80, english=59)  
Row(id=17, name='Quincy', age=18, gender='M', math=65, science=79, english=54)  
Row(id=10, name='Jack', age=19, gender='F', math=44, science=59, english=60)  
Row(id=38, name='Laura', age=23, gender='M', math=84, science=73, english=56)  
Row(id=23, name='Wendy', age=24, gender='M', math=57, science=83, english=81)
```

```
In [9]: print("Total rows in dataset:", df.count())
```

Total rows in dataset: 50

```
In [ ]:
```