1. To get the random sampling subset from the large dataset.
2. Use fraction to indicate what percentage of data to return and seed value to make sure every time to get same random sample.

from pyspark.sql import \*  
from pyspark.sql.functions import \*  
  
*# Create a Spark session*spark = SparkSession.builder.appName("SalesAnalysis").getOrCreate()  
  
df = spark.range(start=1, end=101)  
  
df.sample(fraction=0.1).show()  
  
df.sample(fraction=0.1,seed=123).show()

**Output:**

+---+

| id|

+---+

| 10|

| 24|

| 25|

| 32|

| 37|

| 45|

| 55|

| 88|

+---+

+---+

| id|

+---+

| 36|

| 39|

| 42|

| 46|

| 72|

| 85|

| 88|

|100|

+---+