

• A PySpark DataFrame is a distributed table-like structure that provides optimized processing for big data workloads. It is the most widely used abstraction in Spark for handling structured and semi-structured data efficiently.

- ★ 1. Creating a PySpark DataFrame
- 1.1 Create a SparkSession
 To use DataFrames, we first need a SparkSession:

from pyspark.sql import SparkSession

Initialize SparkSession
spark =
SparkSession.builder.appName("PySpark
Basics").getOrCreate()

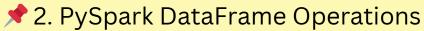
1.2 Creating DataFrames from Lists

We can create a DataFrame from Python lists:

data = [("Alice", 25), ("Bob", 30),
 ("Charlie", 35)]
 df = spark.createDataFrame(data,
 ["Name", "Age"])
 df.show()

◆ 1.3 Creating DataFrames from a CSV File

df = spark.read.csv("data.csv",
header=True, inferSchema=True)
df.show()



- df.printSchema()
- Select
- Filter
- withcolumn
- withColumnRename
- drop
- count
- groupBy
- Sorting & Ordering
- CreateOrReplaceTempView