# PySpark Beginner Guide

## 1. Introduction to PySpark

PySpark is the Python API for Apache Spark. It allows you to write Spark applications using Python.

## 2. Installation

If you are running locally (example in Jupyter Notebook or Colab):

!pip install pyspark

In **Databricks**, PySpark is already installed, so no need to install manually.

## 3. Starting SparkSession

from pyspark.sql import SparkSession  
  
# Create Spark session  
spark = SparkSession.builder.appName("PySparkBeginner").getOrCreate()

## 4. Creating DataFrame

data = [("Alice", 25), ("Bob", 30), ("Charlie", 28)]  
columns = ["Name", "Age"]  
  
df = spark.createDataFrame(data, columns)  
df.show()

## 5. Reading Data

# From CSV  
df\_csv = spark.read.csv("file.csv", header=True, inferSchema=True)  
  
# From JSON  
df\_json = spark.read.json("file.json")  
  
# From Parquet  
df\_parquet = spark.read.parquet("file.parquet")

## 6. Basic Operations

# Show schema  
df.printSchema()  
  
# Select column  
df.select("Name").show()  
  
# Filter rows  
df.filter(df["Age"] > 26).show()  
  
# Group by  
df.groupBy("Age").count().show()

## 7. Writing Data

df.write.csv("output\_csv", header=True)  
df.write.json("output\_json")  
df.write.parquet("output\_parquet")

## 8. RDD Basics (Optional)

rdd = spark.sparkContext.parallelize([1, 2, 3, 4, 5])  
print(rdd.collect())

## 9. Stop SparkSession

spark.stop()