Apache Spark Tutorial for Beginners – Command Guide

# 1. Prerequisites

• Apache Spark (Databricks handles this for you)  
• Python (PySpark in notebooks)  
• Sample CSV or JSON file uploaded to `/FileStore/`

# 2. Start Spark Session

from pyspark.sql import SparkSession  
  
spark = SparkSession.builder \  
 .appName("Beginner Spark App") \  
 .getOrCreate()  
  
\*Note: In Databricks, `spark` session is automatically available.\*

# 3. Read Data

# CSV  
df = spark.read.option("header", True).csv("/FileStore/sample.csv")  
  
# JSON  
df = spark.read.json("/FileStore/sample.json")  
  
# Parquet  
df = spark.read.parquet("/FileStore/sample.parquet")

# 4. Show Data

df.show()  
df.show(5)

# 5. Print Schema

df.printSchema()

# 6. Select Columns

df.select("name", "age").show()

# 7. Filter Rows

df.filter(df.age > 25).show()  
df.where("age > 25").show()

# 8. Add New Column

from pyspark.sql.functions import col  
df.withColumn("age\_plus\_10", col("age") + 10).show()

# 9. Rename Column

df.withColumnRenamed("age", "AgeYears").show()

# 10. Drop Column

df.drop("age").show()

# 11. GroupBy and Aggregation

df.groupBy("city").count().show()  
df.groupBy("city").agg({"salary": "avg"}).show()

# 12. Sort Data

df.sort("age").show()  
df.orderBy(df.salary.desc()).show()

# 13. Join DataFrames

df1 = spark.read.csv("/FileStore/emp.csv", header=True, inferSchema=True)  
df2 = spark.read.csv("/FileStore/dept.csv", header=True, inferSchema=True)  
  
df1.join(df2, df1.dept\_id == df2.dept\_id, "inner").show()

# 14. Write Data

df.write.mode("overwrite").csv("/FileStore/output\_data/")

# 15. Use SQL Queries

df.createOrReplaceTempView("people")  
result = spark.sql("SELECT name, age FROM people WHERE age > 30")  
result.show()

# 16. Count, Distinct, Describe

df.count()  
df.select("city").distinct().show()  
df.describe().show()

# 17. Common PySpark Functions

from pyspark.sql.functions import \*  
  
df.select(upper(col("name"))).show()  
df.select(length(col("name"))).show()  
df.withColumn("updated\_age", col("age") + 5).show()

# 18. Sample Dataset to Practice

id,name,age,city,salary  
1,Alice,28,Delhi,50000  
2,Bob,32,Mumbai,62000  
3,Charlie,24,Bangalore,45000  
  
df = spark.read.option("header", True).csv("/FileStore/people.csv")