# PySpark DataFrame Exercise

## **1. Create a Sample DataFrame**

**Question:** Create a PySpark DataFrame with columns id, name, age, and date.

**Solution:**

from pyspark.sql import SparkSession

# Initialize Spark Session

spark = SparkSession.builder.appName("PySparkExercises").getOrCreate()

# Sample Data

data = [

(1, "Alice", 25, "2024-02-01"),

(2, "Bob", 30, "2023-07-15"),

(3, "Charlie", 35, "2022-10-30"),

(4, "David", 40, "2021-12-22"),

]

# Define schema

columns = ["id", "name", "age", "date"]

# Create DataFrame

df = spark.createDataFrame(data, columns)

df.show()

# Department dataframe

data2 = [(1, "HR"), (2, "IT"), (3, "Finance")]

columns2 = ["id", "department"]

df2 = spark.createDataFrame(data2, columns2)

df2.show()

## **2. Select Specific Columns**

**Question:** Select only name and age columns from the DataFrame.

## **3. Select All Columns**

**Question:** Select all columns from the DataFrame.

## **4. Select with Alias**

**Question:** Rename the name column as full\_name while selecting it.

## **5. Access Columns Using df.column**

**Question:** Select name and age using dot notation.

## **6. Show DataFrame**

**Question:** Display the DataFrame.

## **7. withColumn()**

## **a) Rename id to user\_id and age to user\_age.**

**Question:** Rename all columns using toDF.

## b) Convert the age column to string type using WithColumn

## **8. Filter Data Using where**

**Question:** Retrieve rows where age is greater than 30.

## **9. Filter Data Using filter**

**Question:** Retrieve rows where age is less than 35.

## **10. Use where in Condition**

**Question:** Retrieve rows where the name is either "Alice" or "Bob".

## **11. Use or and and Conditions in Filter**

**Question:** Retrieve rows where age is greater than 25 **and** less than 40.  
**Also, retrieve rows where age is less than 30 or name is "Charlie".**

## **12. Date-Related Operations**

**Question:** Convert date column to date format and add a current\_date column.

## **13. Count Rows**

**Question:** Count the number of rows in the DataFrame.

## **14. Inner Join**

**Question:** create another DataFrame containing id and department.

**Perform an inner join with df & df2 based on id.**

data2 = [(1, "HR"), (2, "IT"), (3, "Finance")]

columns2 = ["id", "department"]

df2 = spark.createDataFrame(data2, columns2)

## **15. Get Distinct Values**

**Question:** Retrieve distinct age values from the DataFrame.

## **16. Group By & Find Max and Min**

**Question:** Find the maximum and minimum age from the DataFrame.

## **17. Aggregate Functions (agg)**

**Question:** Find the average and total sum of age.

## **18. Order By Ascending**

**Question:** Sort the DataFrame by age in ascending order.

## **19. Order By Descending**

**Question:** Sort the DataFrame by age in descending order.

## **20. Chain Multiple Operations**

**Question:** Filter for ages greater than 25, select name and age, and order the result by age in descending order.