Create a DataFrame in PySpark and apply basic operations such as viewing data and selecting columns.

In [1]:

```
sc
```

Out[1]: **SparkContext**

Spark UI

Version	`v4.0.0`
Master	`local[*]`
AppName	`PySparkShell`

In [2]:

```
from pyspark.sql import SparkSession

# Step 1: Initialize Spark Session
spark = SparkSession.builder.appName("BasicDataFrameOps").getOrCreate()
```

In [3]:

```
# Step 2: Read CSV file into DataFrame
df = spark.read.csv("students.csv", header=True, inferSchema=True)
```

In [4]:

```
# === Basic Operations ===

# 1. View first 5 rows
print("=== First 5 rows ===")
df.show(5)
```

```
=== First 5 rows ===
+---+-------+---+------+----+-------+-------+
| id|   name|age|gender|math|science|english|
+---+-------+---+------+----+-------+-------+
|  1|  Alice| 20|     F|  66|     92|     44|
|  2|    Bob| 20|     M|  82|     52|     77|
|  3|Charlie| 22|     F|  43|     57|     76|
|  4|  David| 19|     M|  95|     69|     46|
|  5|    Eva| 19|     F|  62|     44|     96|
+---+-------+---+------+----+-------+-------+
only showing top 5 rows
```

In [5]:

```
# 2. Print schema (structure of DataFrame)
print("=== Schema ===")
df.printSchema()
```

```
=== Schema ===
root
 |-- id: integer (nullable = true)
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- gender: string (nullable = true)
 |-- math: integer (nullable = true)
 |-- science: integer (nullable = true)
 |-- english: integer (nullable = true)
```

In [6]:

```
# 3. Select specific columns: name and math
print("=== Select name and math columns ===")
```

```
df.select("name", "math").show(5)
```

```
=== Select name and math columns ===
+-------+----+
|   name|math|
+-------+----+
|  Alice|  66|
|    Bob|  82|
|Charlie|  43|
|  David|  95|
|    Eva|  62|
+-------+----+
only showing top 5 rows
```

In [7]:

```
# 4. Filter students with math >= 80
print("=== Students with math >= 80 ===")
df.filter(df.math >= 80).show(5)
```

```
=== Students with math >= 80 ===
+---+------+---+------+----+-------+-------+
| id|  name|age|gender|math|science|english|
+---+------+---+------+----+-------+-------+
|  2|   Bob| 20|     M|  82|     52|     77|
|  4| David| 19|     M|  95|     69|     46|
| 11| Kathy| 25|     M|  85|     71|     89|
| 12|   Leo| 24|     M|  97|     84|     83|
| 15|Olivia| 18|     M|  87|     90|     87|
+---+------+---+------+----+-------+-------+
only showing top 5 rows
```

In [8]:

```
# 5. Sort students by science marks (descending)
print("=== Sorted by science (desc) ===")
df.orderBy(df.science.desc()).show(5)
```

```
=== Sorted by science (desc) ===
+---+------+---+------+----+-------+-------+
| id|  name|age|gender|math|science|english|
+---+------+---+------+----+-------+-------+
| 27| Aaron| 25|     F|  81|     99|     44|
| 32| Fiona| 22|     F|  48|     96|     48|
| 33|George| 22|     M|  66|     95|     84|
| 29|  Carl| 22|     F|  53|     92|     52|
|  1| Alice| 20|     F|  66|     92|     44|
+---+------+---+------+----+-------+-------+
only showing top 5 rows
```

In [9]:

```
# 6. Count total rows
print("Total rows in dataset:", df.count())
```

```
Total rows in dataset: 50
```

In [10]:

```
# 7. Show column names
print("Columns:", df.columns)
```

```
Columns: ['id', 'name', 'age', 'gender', 'math', 'science', 'english']
```

In [11]:

```
# Stop Spark session
# spark.stop()
```