In [2]:

```
sc
```

Out[2]:

SparkContext

Spark UI

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

In [3]:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, avg, max, min, round, count
 # Step 1: Initialize Spark Session
spark = SparkSession.builder.appName("StudentsAnalytics").getOrCreate()
```

In [4]:

```
df = spark.read.csv("students.csv", header=True, inferSchema=True)
```

In [5]:

```
 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 [6]:

```
 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 [7]:

```
print("Total rows:", df.count())
```

```
Total rows: 50
```

In [8]:

```
 print("=== Summary Statistics ===")
 df.describe().show()
```

```
=== Summary Statistics ===
+-------+------------------+-----+------------------+------+------------------+-----
-------------+-----------------+
|summary|                id| name|               age|gender|              math|
science|           english|
+-------+------------------+-----+------------------+------+------------------+-----
-------------+-----------------+
|  count|                50|   50|                50|    50|                50|
50|               50|
|   mean|              25.5| NULL|              21.5|  NULL|             68.94|
70.16|            69.36|
| stddev|14.577379737113251| NULL|2.2337851101588404|  NULL|17.609610085034216|14.63
6214521186957|18.74507826560544|
|    min|                 1|Aaron|                18|     F|                40|
44|               42|
|    max|                50| Zoey|                25|     M|               100|
99|              100|
+-------+------------------+-----+------------------+------+------------------+-----
-------------+-----------------+
```

In [9]:

```python
print("=== Students with math >= 80 ===")
df.filter(col("math") >= 80).select("id", "name", "math").show(10)
```

```
=== Students with math >= 80 ===
+---+------+----+
| id|  name|math|
+---+------+----+
|  2|   Bob|  82|
|  4| David|  95|
| 11| Kathy|  85|
| 12|   Leo|  97|
| 15|Olivia|  87|
| 20|  Tina| 100|
| 21|   Uma|  89|
| 22|Victor|  96|
| 25|  Yara| 100|
| 27| Aaron|  81|
+---+------+----+
only showing top 10 rows
```

In [10]:

```python
print("=== Average marks per subject ===")
df.select(
round(avg("math"),2).alias("avg_math"),
round(avg("science"),2).alias("avg_science"),
round(avg("english"),2).alias("avg_english")
).show()
```

```
=== Average marks per subject ===
+--------+-----------+-----------+
|avg_math|avg_science|avg_english|
+--------+-----------+-----------+
|   68.94|      70.16|      69.36|
+--------+-----------+-----------+
```

In [11]:

```
df_with_avg = df.withColumn("average", round((col("math")+col("science")+col("engli
print("=== Dataset with 'average' column ===")
df_with_avg.show(5)
```

```
=== Dataset with 'average' column ===
+---+-------+---+------+----+-------+-------+-------+
| id|   name|age|gender|math|science|english|average|
+---+-------+---+------+----+-------+-------+-------+
|  1|  Alice| 20|     F|  66|     92|     44|  67.33|
|  2|    Bob| 20|     M|  82|     52|     77|  70.33|
|  3|Charlie| 22|     F|  43|     57|     76|  58.67|
|  4|  David| 19|     M|  95|     69|     46|   70.0|
|  5|    Eva| 19|     F|  62|     44|     96|  67.33|
+---+-------+---+------+----+-------+-------+-------+
only showing top 5 rows
```

In [12]:

```
print("=== Topper ===")
df_with_avg.orderBy(col("average").desc()).limit(1).show()
```

```
=== Topper ===
+---+----+---+------+----+-------+-------+-------+
| id|name|age|gender|math|science|english|average|
+---+----+---+------+----+-------+-------+-------+
| 12| Leo| 24|     M|  97|     84|     83|   88.0|
+---+----+---+------+----+-------+-------+-------+
```

In [13]:

```
print("=== Average marks by gender ===")
df_with_avg.groupBy("gender").agg(
    round(avg("math"),2).alias("avg_math"),
    round(avg("science"),2).alias("avg_science"),
    round(avg("english"),2).alias("avg_english"),
    round(avg("average"),2).alias("overall_avg")
).show()
```

```
=== Average marks by gender ===
+------+--------+-----------+-----------+-----------+
|gender|avg_math|avg_science|avg_english|overall_avg|
+------+--------+-----------+-----------+-----------+
|     F|   63.86|      68.55|      70.55|      67.66|
|     M|   75.95|      72.38|      67.71|      72.02|
+------+--------+-----------+-----------+-----------+
```

In [15]:

```
print("=== Min & Max of each subject ===")
df.select(
min("math").alias("min_math"), max("math").alias("max_math"),
min("science").alias("min_science"), max("science").alias("max_science"),
min("english").alias("min_english"), max("english").alias("max_english")
).show()
```

```
=== Min & Max of each subject ===
+--------+--------+-----------+-----------+-----------+-----------+
|min_math|max_math|min_science|max_science|min_english|max_english|
+--------+--------+-----------+-----------+-----------+-----------+
|      40|     100|         44|         99|         42|        100|
+--------+--------+-----------+-----------+-----------+-----------+
```

In [ ]: