📝 RDD Exercises – Set 2

Setup (once)

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("RDD-Exercises-Set2").getOrCreate()
sc = spark.sparkContext
```

---

1. Numbers Practice

- Create an RDD with numbers 1–15.
- Find all numbers divisible by 3.
- Create a new RDD with each number doubled.
- Count how many numbers are greater than 10.

---

2. String Processing

- Create an RDD: ["apple", "banana", "grape", "banana", "apple", "mango"].
- Find distinct fruits.
- Count how many times each fruit appears.
- Find the longest word in the list.

---

3. Sentence Split

- RDD: ["spark makes big data easy", "rdd is the core of spark", "python with spark"].
- Split into words using flatMap.
- Convert words to lowercase and remove duplicates.
- Count total number of unique words.

---

4. Pair RDD Operations

- Create an RDD of (student, marks) like:
  [("Rahul", 85), ("Priya", 92), ("Aman", 78), ("Rahul", 90), ("Priya", 88)].
- Find total marks for each student.
- Find average marks for each student.
- Find the student with the highest marks overall.

---

5. Reduce & Aggregate

- RDD: [5, 10, 15, 20, 25].
- Find the sum using reduce.
- Find the product of all numbers using reduce.
- Compute average manually using reduce (total sum ÷ count).

---

6. Word Length Analysis

- RDD: ["data", "engineering", "spark", "rdd", "pyspark", "analytics"].
- Map each word to (word, length).
- Find the longest word.
- Find average word length.

---

7. Joins

- Students RDD: [(1, "Rahul"), (2, "Priya"), (3, "Aman")]
- Courses RDD: [(1, "Python"), (2, "Spark"), (4, "Databases")]
- Do:
    - Inner join (students with valid courses).
    - Left outer join.
    - Right outer join.

---

8. Mini Real-World

- Orders RDD: [(1, 200), (2, 500), (3, 300), (1, 150), (2, 250)] → (customer_id, order_amount).
- Find total spend per customer.
- Find customer with maximum spend.
- Find total revenue from all customers.