PYSPARK ASSIGNMENT

Case Study: Telecom Call Records Analysis (RDD Only)

Objective

This assignment is designed to help you understand **RDDs (Resilient Distributed Datasets)** in PySpark by working with **raw data** and applying **functional transformations** such as `map`, `filter`, `reduceByKey`, and `reduce`.

You must **NOT use DataFrames or Spark SQL**.
All processing must be done using **RDDs only**.

Business Context

You are working as a data engineer for a telecom company.

Every day, the company receives **Call Detail Records (CDRs)** as CSV files.
Each record represents one phone call made by a customer.

Your task is to analyze this raw data using **RDD-based processing**.

Dataset

File name

`call_records.csv`

Schema

`call_id,caller,receiver,city,call_type,duration_seconds,cost`

Sample data to be used (students must create this file)

```
call_id,caller,receiver,city,call_type,duration_seconds,cost
C001,Amit,Rahul,Hyderabad,Local,180,2.5
C002,Neha,Arjun,Bangalore,STD,320,6.0
C003,Rahul,Pooja,Delhi,Local,60,1.0
C004,Pooja,Neha,Mumbai,ISD,900,25.0
C005,Arjun,Amit,Chennai,STD,400,7.5
C006,Sneha,Karan,Hyderabad,Local,240,3.0
C007,Karan,Sneha,Delhi,Local,120,2.0
C008,Riya,Vikas,Bangalore,STD,360,6.5
C009,Vikas,Riya,Mumbai,ISD,1100,30.0
C010,Anjali,Sanjay,Chennai,Local,90,1.5
C011,Farhan,Ayesha,Delhi,STD,420,7.0
C012,Ayesha,Farhan,Hyderabad,ISD,950,28.0
C013,Suresh,Divya,Bangalore,Local,150,2.0
C014,Divya,Suresh,Mumbai,STD,380,6.8
C015,Nikhil,Priya,Delhi,Local,200,2.8
C016,Priya,Nikhil,Chennai,STD,410,7.2
C017,Rohit,Kavya,Hyderabad,Local,170,2.3
C018,Kavya,Rohit,Bangalore,Local,140,2.1
C019,Manish,Tina,Mumbai,ISD,1000,27.0
C020,Tina,Manish,Delhi,STD,350,6.2
```

Constraints

- Use **RDDs only**
- Do not convert to DataFrame
- Do not use Spark SQL
- Do not use Pandas
- Handle header manually
- Explicitly convert data types where required

Tasks / Exercises

Task 1

Read the CSV file using `sparkContext.textFile` and display the first 5 records.

Task 2

Remove the header row and create a clean RDD containing only data rows.

Task 3

Split each row into individual fields using a delimiter.

Task 4

Calculate the **total call cost per city**.

Task 5

Identify the **city with the highest total call cost**.

Task 6

Calculate the **total call duration per call type** (Local, STD, ISD).

Task 7

Count the **number of calls per city**.

Task 8

Calculate the **average call cost per city** using RDD transformations.

Task 9

Filter and list all **high-value calls** where call cost is greater than 20.

Task 10

Count the number of **ISD calls per city**.

Task 11

Identify the **longest call** based on call duration.

Task 12

Calculate the **total revenue generated by each caller**.

Task 13

Detect **suspicious calls** based on the following rule:

- duration greater than 900 seconds
- cost greater than 25
