University of Eloued
Faculty of Exact Sciences
Department of Computer science

For Second-year Master: IA & Data Science

University season: 2025/2026

Correction Make-up exam in the Module:Big Data Analytics

EX 01 : (06 p) : Answer these questions briefly

1- Distinguish between **batch processing** and **stream processing** with one example for each?

Batch = periodic, large-scale processing with higher latency.
Example: Calculating monthly payroll for all employees.

Stream = continuous, real-time processing with low latency.
Example: Real-time fraud detection in credit card transactions.

2- Compare **RDDs**, **DataFrames**, and **Datasets**. When would you prefer each one?

Feature / API	RDD	DataFrame	Dataset
Schema	❌ No	✅ Yes	✅ Yes
Query Optimization (Catalyst)	❌	✅	✅
Type Safety	❌	❌	✅
Ease of Use	Low	High	High (typed)
Best Use Case	Custom low-level ops	Structured data processing	Structured + type safety

When to prefer each:

RDD: When you need full control, custom logic, or are processing unstructured data where schemas don't help.

DataFrame: For most structured data pipelines, SQL-like queries, and performance-focused analytics.

Dataset: When you need *type safety for compile-time checking while still getting optimized execution.

3- Explain in detail the concept of **shuffling** within the context of **MapReduce**. Why is it considered a costly step in terms of performance?

he shuffle step is the process that:Redistributes intermediate data ,Prepares this data before the Reduce phase begins.shuffle ensures the reduce function receives all values for a given key in one place. Without shuffle, reducers wouldn't know which data belongs together.

Why Shuffling Is Considered Costly: Large Data Transfers Over the Network, Disk I/O Overhead,Sorting and Grouping Overhead, Load Imbalance (Data Skew).

EX 02 : (04 p) Mark √ for the correct answer

1) What is the definition of an RDD in Apache Spark? b)A Resilient Distributed Dataset

2) Which component for the reduction of intermediate data before the final reduce operation? b)Combiner

3)Which technology is primarily used for real-time processing of massive datasets? a)Apache Spark

EX 03 : (04 p) :

A text file, clients.txt, is provided on HDFS. Write a MapReduce program (Java or pseudocode) that:

1) public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {
String line = value.toString();
String[] tokens = line.split("\\s+"); // découpe sur les espaces
for (String token : tokens) {
if (!token.isEmpty()) {
word.set(token);
context.write(word, one); // émettre (mot, 1) } } }
public void reduce(Text key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException {
int sum = 0;
for (IntWritable val : values) {
sum += val.get(); // additionne chaque 1 }
context.write(key, new IntWritable(sum)); // (mot, fréquence) }

2) public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {
String line = value.toString();
String[] tokens = line.split("\\s+");
for (String token : tokens) {
if (!token.isEmpty()) {
word.set(token);
context.write(word, one); // Emit (word,1) } } }
public void reduce(Text key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException {
int sum = 0;
for (IntWritable val : values) {
sum += val.get(); // Summing all occurrences }
context.write(key, new IntWritable(sum)); }

EX 04 : (06 p) :

Instructions: Assume that sc is your SparkContext

1)

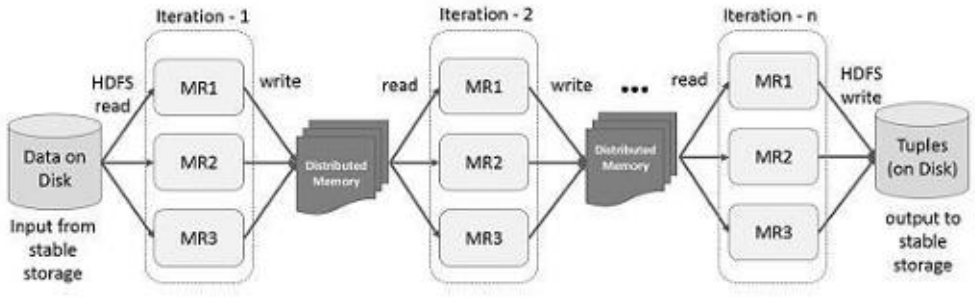

Figure: Iterative operations on Spark RDD

2)

```
rddA = sc.parallelize([10,20,30,40,50])
print(rddA.count())
```

3)

```
doubleRDD = rddA.map(lambda x: x*2)
print(doubleRDD.collect())
```