University of Abdelhamid Mehri-Constantine2
Faculty: NTIC
Department: IFA
Master's second year SDIA 2024/2025

Lab 1
Introduction to Spark RDD

In this first lab, we want to perform some basic operation using Spark RDD (Resilient Distributed Dataset) data type.

Consider the following data:

data1 = [1 , 15 , 17 , 20 , 5]

data2 = [("Ali", 25) , ("Brahim", 18) , ("Cherif", 30) , ("Djamel", 20) , ("Eliane",15)]

Such as, data1 contains IDs of the people represented in data2 which contains tuples of people name and ages.

1. Getting Started

We will be using pyspark within the google colab environment. Run the following line to install/check pyspak in google colab

```
!pip install pyspark py4j
```

2. Creating a spark Session

We need to create a spark session to create our first spark application. and a spark context object with which we can create RDDs.

```
from pyspark.sql  import SparkSession
from pyspark import SparkContext

spark = SparkSession.builder.appName("sparkFirstLab").getOrCreate()
sc = SparkContext.getOrCreate()
```

3. Creating RDDs

Create 2 RDDs `rdd1` and `rdd2` from the given data using the parallelize function of the sc object.

```
sc.parallelize(data1)
sc.parallelize(data1)
```

4. Performing operation in RDDs

For each of the RDD created, display the following information:

1. The number of elements in each RDD using `count()` operation
2. All the elements using `map()` function
3. The value of the first element using **take(number)** operation

4. The value of the last element in 2 different ways, the first one using `collect()` operation and the second one is index-based
5. All people over the age of 20 using **filter()** operation
6. The name and the ID of the youngest person. you can `use min(key)` function.

Now we want to create a third RDD rdd3 that contains the name, Age, and ID from the previous data.

7. Combine `rdd1` and `rdd2` in `rdd3` using `zip()` operation
8. Redo question 6 using rdd3