University Mohamed Boudiaf M'sila **M1 Artificial Intelligence**
Faculty MI, Computer Science Department **M'sila, 19/12/2022**

Assignment

Big Data & Data Science

Exercice 1 (2 pts)

Consider the HDFS file data.txt. The size of data.txt is 10000MB. Suppose that you are using a Hadoop cluster that can potentially run up to 20 instances of the mapper class in parallel and suppose to execute a MapReduce based program that selects the rows of data.txt containing the words "BIG" and "DATA".
Which of the following values is a proper HDFS block size if you want to "force" Hadoop to run exactly 5 instances of the mapper class in parallel when you execute the application by specifying data.txt as input file?

a) Block size: 2048MB
b) Block size: 1024MB
c) Block size: 512MB
d) Block size: 256MB

Exercice 2 (2 pts)

Consider the input HDFS folder myFolder that contains the following two files:

- Temperatures2019.txt: size=1048MB
- Temperatures2020.txt: size=1000MB

Suppose that you are using a Hadoop cluster that can potentially run up to 10 instances of the mapper class in parallel. Suppose to execute a MapReduce application for Hadoop that computes some statistics about temperatures. This application is based on one single MapReduce job. The input of this Hadoop application is the HDFS folder myFolder. The HDFS block size is 1024MB. The number of instances of the reducer class is set to 5. How many mappers are instantiated by Hadoop (i.e., how many instances of the mapper class) when you execute this MapReduce application by specifying the folder myFolder as input?

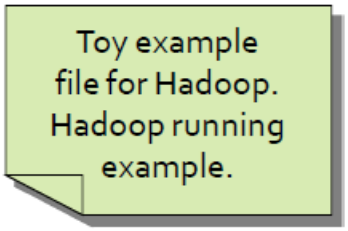
a) 10
b) 5
c) 3
d) 2

Exercice 3 (4 points)

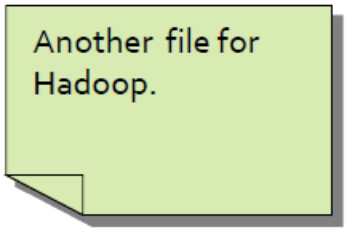
Consider the following Java program of Mapper and Reducer (see page 2).

1- Describe the output of this program.
2- Find the output of the program having the following input files:

File1.txt



File2.txt



```
/**
 * Exercise 3 - Mapper
 */
class MapperBigData extends Mapper<
                LongWritable, // Input key type
                Text,         // Input value type
                Text,         // Output key type
                IntWritable> {// Output value type

   protected void map(
         LongWritable key,   // Input key type
         Text value,         // Input value type
         Context context) throws IOException, InterruptedException {

         // Split each sentence in words. Use whitespace(s) as delimiter (=a space, a tab, a line break, or a form feed)
                        // The split method returns an array of strings
         String[] words = value.toString().split("\\s+");

         // Iterate over the set of words
         for(String word : words) {
            // Transform word case
            String cleanedWord = word.toLowerCase();

            // emit the pair (word, 1)
            context.write(new Text(cleanedWord), new IntWritable(1));
         }
   }
}


/**
 * Exercise 3 - Reducer
 */
class ReducerBigData extends Reducer<
            Text,          // Input key type
            IntWritable,    // Input value type
            Text,          // Output key type
            IntWritable> {  // Output value type

   @Override
   protected void reduce(
      Text key, // Input key type
      Iterable<IntWritable> values, // Input value type
      Context context) throws IOException, InterruptedException {

      int occurrances = 0;

      // Iterate over the set of values and sum them
      for (IntWritable value : values) {
         occurrances = occurrances + value.get();
      }
      context.write(key, new IntWritable(occurrances));
   }
}
```

Solution of Assignment

Big Data & Data Science

Exercice 1 (2 pts)

Answer: a)

Exercice 2 (2 pts)

Answer: c)

Exercice 3 (4 points)

1- Describe the output of this program.
The program counts the number of occurances of each word in a file or a collection of files.

2- Find the output of the program having the following input files

- Output pairs (another, 1)
(example, 2)
(file, 2)
(for, 2)
(hadoop, 3)
(running, 1)
(toy, 1)