Université M'sila **M1 IA/ Big Data at Science de données**
Département Informatique **Dr. Tahar Mehenni**

Série de TD N°3
HDFS

Exercice 1

Considérons un dossier HDFS nommé *patchesFolder* qui contient deux fichiers : *patches2018.txt* de taille 513MB et *patches2019.txt* de taille 515MB. Supposons que le cluster Hadoop utilisé peut supporter jusqu'à 5 instances mapper en parallèle. Le nombre d'instances reducer est 2 et la taille d'un bloc HDFS est 512MB. Quel est le nombre d'instances du mapper utilisées lorsqu'on exécute une application MapReduce sur les deux fichiers du dossier *patchesFolder* ?

Exercice 2

Considérons un dossier HDFS nommé *inputData* qui contient les fichiers suivants :

Nom du fichier	Taille	Contenu du fichier
Temperature1.txt	61 Octets	2016/01/01,00:00,0 2016/01/01,00:05,-1 2016/01/01,00:10,-1.2
Temperature2.txt	63 Octets	2016/01/01,00:15,-1.5 2016/01/01,00:20,0 2016/01/01,00:25,-0.5
Temperature3.txt	62 Octets	2016/01/01,00:30,-0.5 2016/01/01,00:35,1 2016/01/01,00:40,1.5

Supposons que le cluster Hadoop utilisé peut supporter jusqu'à 10 instances mapper en parallèle. Soit le programme MapReduce suivant exécuté sur les fichiers ci-dessus. Donner le résultat après exécution.

```
/* Mapper */
class MapperBigData extends Mapper<LongWritable, Text, Text, DoubleWritable> {
protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
String fields[] = value.toString().split(",");
String date = fields[0];
Double temperature = Double.parseDouble(fields[2]);
// Emit (date, temperature)
context.write(new Text(date), new DoubleWritable(temperature));
}
}
/* Reducer */
class ReducerBigData extends Reducer<Text, DoubleWritable, Text, DoubleWritable> {
@Override
protected void reduce(Text key, // Input key type
Iterable<DoubleWritable> values, // Input value type
Context context) throws IOException, InterruptedException {
double maxTemp = Double.MIN_VALUE;
// Iterate over the set of values and compute the maximum temperature
for (DoubleWritable temperature : values) {
if (temperature.get() > maxTemp) {
maxTemp = temperature.get();
}
}
// Emit (date, maximum temperature)
context.write(key, new DoubleWritable(maxTemp));
}
}
```

Exercice 3

On veut exécuter un programme MapReduce qui permet de sélectionner les lignes du fichier *logs.txt* qui contiennent les mots « ERROR » ou « WARNING ». Sachant que la taille du fichier *logs.txt* est 5000 MB, quelle est la taille du block HDFS qui sera choisi si on veut forcer Hadoop à exécuter 10 mappers en parallèle pour le programme MapReduce sur le fichier *logs.txt*.

a) Block size: 5000MB b) Block size: 2048MB c) Block size: 1024MB d) Block size: 512MB

Exercice 4

Considérons deux fichiers HDFS, *logs1.txt* et *logs2.txt* de tailles 1036MB et 500MB respectivement. Supposons que le facteur de réplication est 4 (c-à-d nombre de copies de chaque block) et que la taille du block HDFS est 512MB. Quel est le nombre total de blocks sont utilisés pour stocker les deux fichiers *logs1.txt* et *logs2.txt* (*Attention* : considérer aussi les copies).

a) 3 blocks b) 4 blocks c) 12 blocks d) 16 blocks