Université Mohamed Boudiaf M'sila **M1 Intelligence Artificielle**
Faculté MI, Dépt Informatique **M'sila, Le 17/01/2023**

Examen Final

Big Data et Science de Données

Exercice 1 (5 pts) : Choisir la bonne réponse *(1pt/réponse correcte).*

1- _____ NameNode est utilisé lorsque le NameNode primaire ne fonctionne plus.
   a) Rack
   b) Data
   c) Secondary
   d) Aucune des réponses précédentes
2- _____ est responsable de la consolidation des résultats produits par chacune des fonctions / tâches Map ().
   a) Reduce
   b) Map
   c) Reducer
   d) Toutes les réponses précédentes
3- Le nombre de Maps est généralement déterminé par la taille totale des :
   a) Sorties
   b) Entrées
   c) Tâches
   d) Aucune des réponses précédentes
4- Lesquelles des phases suivantes se produisent simultanément ?
   a) Shuffle & Map
   b) Reduce & Sort
   c) Shuffle & Sort
   d) Toutes les réponses précédentes
5- L'entrée du _____ est la sortie triée des Mappers.
   a) Reducer
   b) Mapper
   c) Shuffle
   d) Toutes les réponses précédentes

Exercice 2 (3 pts)

Soit un fichier HDFS log.txt de taille 1024MB. On suppose que Hadoop cluster peut exécuter jusqu'à 2048 mappers en parallèle. On veut exécuter une application MapReduce de comptage de mots sur le fichier log.txt. Si on dispose d'un système Hadoop qui définit automatiquement le nombre de mappers égal à 2, quelle est la taille d'un bloc du système HDFS ? *(1.5 pts)*. Expliquer et argumenter votre réponse *(1.5 pts).*

   a) Taille du bloc: entre 1024MB et 2048MB
   b) Taille du bloc: entre 512MB et 1023MB
   c) Taille du bloc: entre 256MB et 511MB
   d) Taille du bloc: entre 128MB et 255MB

Exercice 3 (4 pts)

On considère le répertoire HDFS « InputFolder » qui contient les deux fichiers suivants :

Nom du fichier	Taille	Contenu du fichier
HumidityA.txt	16 octets	51.45 9.55 8.15
HumidityB.txt	18 octets	40.53 12.98 52.99

On utilise un cluster Hadoop qui peut exécuter jusqu'à 4 mappers en parallèle. La taille du bloc HDFS est 2048MB. On suppose que le programme Mapper suivant est exécuté sur le répertoire « InputFolder ».

```
/* Mapper */
import …;
class MapperBigData extends Mapper<LongWritable, Text, DoubleWritable, NullWritable> {
        Double top1;
        protected void setup(Context context) {
                top1 = null;
        }
        protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
                Double val = new Double(value.toString());
                if (top1 == null || val.doubleValue() > top1) {
                        top1 = val;
                }
        }
        protected void cleanup(Context context) throws IOException, InterruptedException {
                // emit the content of top1
                context.write(new DoubleWritable(top1), NullWritable.get());
        }
}
```

Quelle est la sortie (output) du programme après exécution *(2 pts).* Expliquer votre réponse *(2 pts).*

a) Un seul fichier qui contient 52.99
b) Deux fichiers: l'un contient 52.99, l'autre est vide
c) Deux fichiers, l'un contient 51.45, l'autre contient 40.53
d) Deux fichiers, l'un contient 51.45, l'autre 52.99

Exercice 4 (8 points)

On veut analyser les ventes d'une société commerciale. Pour cela on dispose d'un fichier CSV qui contient les ventes par jour, où chaque ligne a le format suivant :

date\t *daily_income*\n

Ecrire les fonctions Map et Reduce pour déterminer les ventes mensuelles (c-à-d par mois) pour chaque année.

Exemple :

- Input file

2015-11-01	1000
2015-11-02	1305
2015-12-01	500
2015-12-02	750
2016-01-01	345
2016-01-02	1145
2016-02-03	200
2016-02-04	500

- Output

(2015-11,2305)
(2015-12, 1250)
(2016-01, 1490)
(2016-02, 700)

BONNE REUSSITE
Dr. Tahar Mehenni

Correction de l'Examen Final

Big Data et Science de Données

Exercice 1 (5 points)

1-c), 2-a), 3-b), 4-c), 5-a)

Exercice 2 (3 points)

Réponse : b) **(1.5 pts)**

Explication : nombre de mappers=2 . Taille du fichier =1024MB. Le bloc sera donc compris entre 512MB et 1023MB. Le fichier sera stocké sur deux blocs. **(1.5 pts)**

Exercice 3 (04 points)

Réponse : d). **(2 pts)**

La sortie contient deux fichiers, l'un est relatif au traitement du fichier en entrée HumidityA.txt, qui contient la valeur top (ou maximale) de l'humidité parmi les valeurs du fichier, elle est égale à 51.49. L'autre fichier est relatif au traitement du fichier en entrée HumidityB.txt, qui contient la valeur Top de l'humidité parmi les valeurs du fichier, elle est égale à 52.99 **(2 pts).**

Exercice 4 (8 points)

```
/**
 * Exercise 4 - Mapper
 */
class MapperBigData extends
           Mapper<Text, // Input key type
                      Text, // Input value type
                      Text, // Output key type
                      DoubleWritable> {// Output value type

      protected void map(Text key, // Input key type       (1 pt)
                 Text value, // Input value type
                 Context context) throws IOException, InterruptedException {

           String[] date = key.toString().split("-");  (0.5 pt)

           String month = new String(date[0] + "-" + date[1]); (0.5 pt)

           // emit the pair (month, value)    (2 pts)
           context.write(new Text(month), new
DoubleWritable(Double.parseDouble(value.toString())));
      }
}

/**
 * Exercise 4 - Reducer
 */
class ReducerBigData extends
           Reducer<Text, // Input key type
                      DoubleWritable, // Input value type
                      Text, // Output key type
```

```
                  DoubleWritable> { // Output value type

      @Override
      protected void reduce(Text key, // Input key type    (1 pt)
                  Iterable<DoubleWritable> values, // Input value type
                  Context context) throws IOException, InterruptedException {

            double totalIncome = 0;     (0.5 pt)

            // Iterate over the set of values and sum them
            for (DoubleWritable value : values) {             (1.5 pts)
                  totalIncome = totalIncome + value.get();
            }
            context.write(new Text(key), new DoubleWritable(totalIncome)); (1 pt)
      }
}
```

Rédigé par T. Mehenni