University Mohamed Boudiaf M'sila **M1 Artificial Intelligence**
MCS Faculty, CS Department **M'sila, 18/01/2024**

Final Exam
Big Data and Data Science

Exercice 1 (4 pts) : Answer these questions

Q1- Give a definition of the CAP theorem. (1 pt)
Q2- Cite the steps of the KDD process. (1 pt)
Q3- Explain by an example the k-fold Cross Validation. (2 pts)

Exercice 2 (4 pts) : Consider an input HDFS folder logsFolder containing the files log2017.txt and log2018.txt. log2017.txt contains the logs of year 2017 and its size is 2052MB while log2018.txt contains the logs of year 2018 and its size is 252MB. Suppose that you are using a Hadoop cluster that can potentially run up to 10 instances of the mapper in parallel. The HDFS block size is 256MB. How many mappers are instantiated by Hadoop when you execute the application by specifying the folder logsFolder as input?

Exercice 3 (4 pts): Consider an input HDFS folder inputFold containing the files log1.txt and log2.txt. The size of log1.txt is 1024MB and the size of log2.txt is 256MB. Suppose that you are using a Hadoop cluster that can potentially run up to 5 instances of the mapper in parallel. Find the proper HDFS block size if you want to "force" Hadoop to run 5 instances of the mapper in parallel when you execute the application by specifying inputFold as input folder?

Exercice 4 (8 pts): PoliCars is an international car sharing company. It has cars in over 1000 cities around the world. In each city, it has hundreds of cars. PoliCars computes a set of statistics to characterize and identify frequent failures of its cars. The analyses are performed by considering the following input data sets/files.

- *Cars.txt* : a text file containing the list of cars managed by PoliCars. One line for each car of PoliCars is stored in *Cars.txt*. The number of cars is more than 40000. Each line of *Cars.txt* has the following format : *CarID,Model,Company,City* where *CarID* is the car identifier, *Model* is its model, *Company* is the name of its carmaker company, and *City* is the city in which that car is used. For example, the following line *Car12,Panda,FCA,Paris* means that car Car12 is used in Paris. Car12 is a Panda (i.e., Panda is the model of Car12) and it was manufactured by FCA.
- *CarsFailures.txt* : a text file where historical failures of the cars are stored. A new line is inserted in *CarsFailures.txt* every time a new failure occurs. CarsFailures.txt contains the historical data about the failures of the last 30 years. Each line of *CarsFailures.txt* has the following format : *Date,Time,CarID,FailureType* where *CarID* is the identifier of the car that had a failure of type *FailureType* at time *Time* of the date *Date*. For example, the following line *2015/01/05,08:45,Car15,Engine* means that car Car15 had a failure of type Engine at 08:45 (hour=08, minute=45) of January 5, 2015.

The managers of PoliCars are interested in performing some analyses about the failures of their cars. Especially, cars with frequent and different types of failures in year 2018. The application considers only the failures of years 2018 and selects the CarIDs of the cars that had at least 5 failures during year 2018 and at least two different types of failures in year 2018. The identifiers (CarIDs) of the selected cars, and the associated number of failures in year 2018, are stored in the output HDFS folder. Each output line contains one pair (CarIDs,NumberOfFailuresYear2018), one line per selected car. Write the corresponding code of the mapper and reducer classes (methods).

Dr. Tahar Mehenni

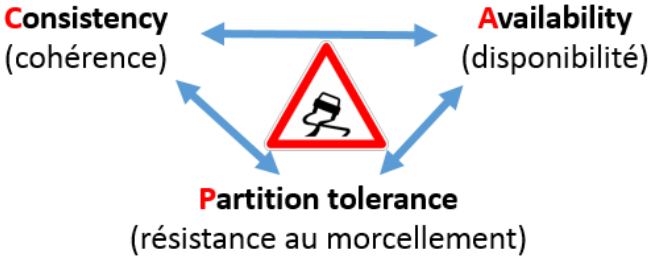
University Mohamed Boudiaf M'sila
MCS Faculty, CS Department

M1 Artificial Intelligence
M'sila, 18/01/2024

Solution of Final Exam
Big Data and Data Science

Exercice 1 (4 pts) : Answer these questions

Q1- Give a definition of the CAP theorem. (1 pt)

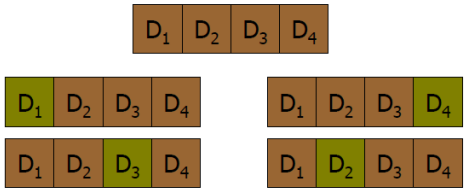


Q2- Cite the steps of the KDD process. (1 pt)

Raw data →Data wrangling →data selection →data preprocessing →data mining →data post-processing →data visualization →data interpretation

Q3- Explain by an example the k-fold Cross Validation. (2 pts)

Data randomly divided into k subsamples of equal sizes, one of which is used for predicting and the remaining k-1 for model testing. The process is only repeated k times. For example: For the dataset D divided in 4 subsamples D_1, D_2, D_3 and D_4, the process is performer 4 times.



Exercice 2 (4 pts) : 10 mappers are instantiated by Hadoop.

Exercice 3 (4 pts): Block size: 256 MB

Exercice 4 (8 pts):

```
class MapperBigData extends Mapper<LongWritable, // Input key type
                Text, // Input value type
                Text, // Output key type
                Text> {// Output value type

        protected void map(LongWritable key, // Input key type
                        Text value, // Input value type
                        Context context) throws IOException, InterruptedException {
                // Split record
                // Example: 2015/01/05,08:45,Car15,Engine
                String[] fields = value.toString().split(",");
                String date = fields[0];
                String carID = fields[2];
                String failureType = fields[3];
                // Select only failures of year 2018
                if (date.startsWith("2018")==true) {
                        // Emit (CarID,FailureType)
                        context.write(new Text(carID), new Text(failureType) );
                }
        }
}
```

```
class ReducerBigData extends Reducer<Text, // Input key type
                Text, // Input value type
                Text, // Output key type
                IntWritable> { // Output value type

        protected void reduce(Text key, // Input key type
                        Iterable<Text> values, // Input value type
                        Context context) throws IOException, InterruptedException {
                // Iterate over the set of values
                // Count the number of input values and
                // check if there are at least two different failure types
                int numFailures = 0;
                String previousFailureType = null;
                Boolean atLeastTwoFailureTypes = false;
                for (Text value : values) {
                        numFailures++;

                        // Check if the current failureType is different from the previous one
                        // If it is true there are at least two different failure types
                        if( previousFailureType!=null && previousFailureType.equals(value.toString())==false) {
                                atLeastTwoFailureTypes = true;
                        }

                        previousFailureType=value.toString();
                }

                // Emit the CarID and the number of failures only if
                // - number of failures >=5
                // - there are at least two failure types for this car
                if (numFailures>=5 && atLeastTwoFailureTypes==true)
                        context.write(new Text(key), new IntWritable(numFailures));
        }
}
```

Dr. T. Mehenni