

## EXAMEN SESSION PRINCIPALE M.30H

<b>A.U:</b>	2021/2022	<b>Cycle:</b>	Engineering
<b>Module:</b>	Framework&Technologies BigData	<b>Level:</b>	3rd Level
<b>Time:</b>	09H.00 - 11H.00	<b>Field</b>	ILSI
<b>Date</b>	01/12/2021	<b>Duration:</b>	2h
<b>Documents:</b>	Not Authorized	<b>N° pages:</b>	3

<b>Exercise</b>	<b>1(8pts)</b>	<b>2(12pts)</b>
<b>C.L.Os Assesment</b>	<b>K2, K3, S2, S4</b>	<b>K2, S2, S4, V1, V3</b>

**Exercise1: (8pts)**

Choose the correct answer(s):

1 - What is Spark?

- a- Performs processing by batch or on the streaming.
- b- Performs batch processing only.
- c- Spark is a fast data processing engine dedicated to Big Data.
- d- Is a distributed file system.

2 - Machine learning:

- a- Machine learning or artificial learning.
- b- Allows a system to learn from data and not from explicit programming.
- c- Allows a system to learn from explicit programming.
- e- Non-artificial learning.

3- To obtain a modification of an RDD:

- a- A transformation must be applied to it, which will return a new RDD, the original will remain unchanged.
- b- An action must be applied to it, which will return a new RDD, the original will remain unchanged.
- c- A transformation must be applied to it, which will return a new RDD, the original will be changed.
- d- An action must be applied to it, which will return a new RDD, the original will be changed.

4- Big Data data are:

- a- Big data.
- b- Bulky data, Varied data, and having an increasing flow over time.

## Exercise2 (12pts)

A - Big data analysis is the use of advanced analytical techniques against very large and diverse big data sets that include structured, semi-structured and unstructured data, from different sources and in different sizes ranging from terabytes to zettabytes.

1- Give the 5 characteristics of Big Data.

2- Explain briefly each of the characteristics.

B - Spark is the new In-Memory brick of Hadoop distributions. Thanks to the richness of its libraries, Spark meets your Big Data needs or those requiring fast response times or to perform advanced calculations. The Spark solution interfaces with Yarn to benefit from the allocated resources.

Figure 1 shows a request from a Hadoop client to write a .txt file to HDFS.

Figure 2 shows a request from a Hadoop client to read a .txt file on HDFS.

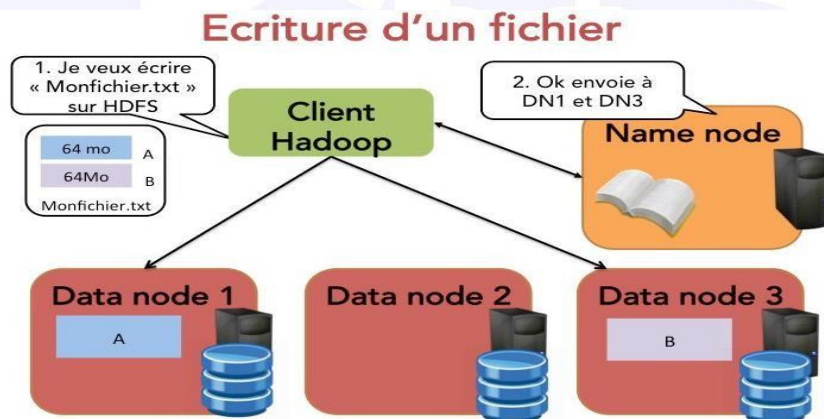


Figure1- Writing a file

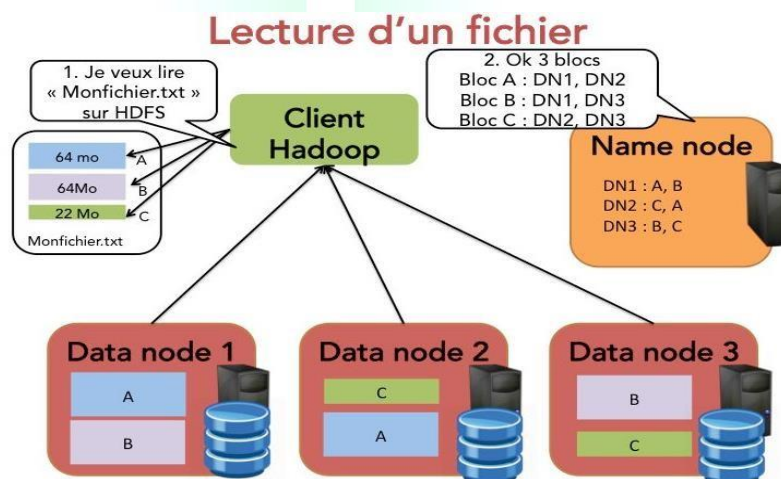


Figure2- Reading a file

1- Define Apache Spark.

2- Give the modules of Spark.

3- Define briefly the modules of Spark.

- 4- Specify the type of Hadoop architecture.
- 5- Give the steps to satisfy the need of the Hadoop customer (write a file).
- 6- Give the steps to satisfy the need of the Hadoop customer (read a file).
- 7- Give the role of the name node.
- 8- For the writing request of a file will we use the Data node2. Justify your answer.
- 9- For the reading request of a file will we use the Data node1. Justify your answer.

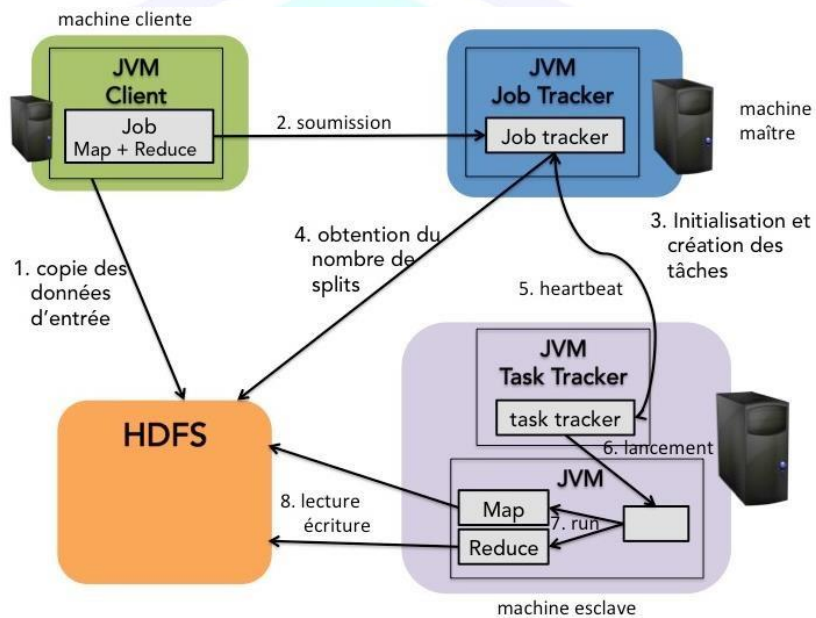


Figure3- The job submission and execution scheme in Hadoop MapReduce

- 10- Based on Figure3, give the job submission and execution components in Hadoop MapReduce.
- 11- Give the role of each component.
- 12- Compare the 2 environments: Apache Spark and Hadoop based on your knowledge.

Ecole Supérieure d'Ingénieurs  
Privée de Gafsa