University of Djilali Bounaama - Khemis Miliana
Faculty of Sciences of Matter and Computer Science

Second year of Master's degree	**Full Name**: ……..
Exam in Big Data and Cloud Computing	**Grade:**
Duration: 1h30min	…….
University Year: 2024-2025	

Exercise 1 (15pt): Choose the correct answer. Each question has only one correct answer. Correct answer: +1, Incorrect answer: 0.

1- Why is structuring Big Data important? ☐ It increases the amount of raw data stored without any processing. ☐ It eliminates the need for data governance and security measures. ☐ It improves data management, enhances analysis, and ensures scalability.	5- What is the purpose of data cleansing in Big Data structuring? ☐ To encrypt data for better security. ☐ To remove duplicates, errors, and irrelevant information. ☐ To divide data into smaller chunks for parallel processing.
2- What is governance in Big Data? ☐ Policies, procedures, and standards ensuring data is managed, protected, and used effectively and ethically. ☐ Techniques for compressing and storing large datasets without losing quality. ☐ A system for automating all data analysis tasks without human intervention.	6- What is a single-core processor? ☐ A CPU with one processing unit that executes tasks sequentially. ☐ A single-core processor can perform high-level data analysis and machine learning tasks. ☐ Many tasks can be performed by this type of processor.
3- What is the primary advantage of a data lake? ☐ It stores large volumes of raw, unprocessed data in its native format. ☐ A data lake ensures all data is stored in a structured and organized format. ☐ A data lake always completes the unprocessed data imported from the data warehouse.	7- What is metadata in the context of a data warehouse? ☐ Raw data stored in its original format without processing. ☐ Encrypted data used for securing sensitive information. ☐ Information about the data, such as schema, relationships, and lineage.
4- What is the role of the ETL process in Big Data? ☐ To analyze and visualize data directly from its raw format. ☐ To encrypt and store data without any processing. ☐ To extract, transform, and load data into a staging area for consistency and integration.	8- What is data wrangling? ☐ Data wrangling is the process of encrypting data to enhance security. ☐ Data wrangling is the process of cleaning, organizing, and preparing data so it can be analyzed properly. ☐ Data wrangling is the process of storing raw data without any modifications.

9- How does isolation benefit transactions in ACID theorem?

- ☐ It allows transactions to be executed simultaneously, even if they conflict.
- ☐ It is the process of retrieving information after an operating system failure.
- ☐ It ensures that transactions do not interfere with each other.

10- Which of the following data is structured data?

- ☐ Tables with rows and columns.
- ☐ Social media posts and videos.
- ☐ JSON or XML files.

11- What is an example of data variety in Big Data?

- ☐ Storing all data in a single, uniform format for easier processing.
- ☐ Combining structured data from databases with unstructured data like social media posts.
- ☐ Using only numerical data for analysis and ignoring text or images.

12- Why is velocity critical in Big Data?

- ☐ It guarantees that all data is stored in a single location for better security.
- ☐ It focuses on reducing the size of data to save storage space.
- ☐ It ensures that fast-moving data is processed in real time for timely insights.

13- What is the main advantage of replication?

- ☐ It provides fault tolerance and data availability.
- ☐ It reduces the overall storage requirements by storing data in a single location.
- ☐ It increases network traffic and reduces data access speed.

14- Why is unstructured data harder to process?

- ☐ Because it lacks a predefined structure.
- ☐ Because it is always encrypted and requires decryption before use.
- ☐ Because it is stored in a relational database.

15- What is the architecture of HDFS?

- ☐ A peer-to-peer architecture where all nodes have equal responsibilities.
- ☐ A client-server architecture with a single central server handling all data storage.
- ☐ A master/slave architecture with a NameNode and DataNodes.

16- What type of data is stored in relational databases?

- ☐ Structured data.
- ☐ Unstructured data like videos and images.
- ☐ Semi-structured data such as JSON files.

17- What is a cluster in Big Data storage?

- ☐ A cluster is a collection of servers (nodes) connected via a network to work as a single unit.
- ☐ A cluster is a single high-performance computer used for processing data.
- ☐ A cluster is a database used exclusively for storing structured data.

18- What resources does each node in a cluster have?

- ☐ Each node has its own memory, processor, and hard drive.
- ☐ Each node shares all resources equally with other nodes.
- ☐ Each node has only a processor and relies on the network for memory and storage.

19- How does a cluster complete tasks efficiently?

- ☐ By using a single powerful server to handle all tasks.
- ☐ By dividing tasks into manageable chunks and distributing them among nodes.
- ☐ By storing all data in one location to reduce processing time.

20- What does Hadoop Common provide?

- ☐ A user interface for managing and monitoring Hadoop clusters.
- ☐ Tools and libraries for other Hadoop components.
- ☐ A distributed file system for storing large datasets across multiple machines.

21- What is Hadoop YARN?

- ☐ It stores and processes large datasets across distributed clusters.
- ☐ It manages resources and coordinates the execution of applications across the system.
- ☐ It provides a user interface for monitoring and managing Hadoop jobs.

22- How are the write and read operations applied in the master-slave model?

- ☐ Writing and reading are both done on the slave side to reduce master load.
- ☐ Both write and read operations are handled by the master, with no distinction between them.
- ☐ Writing is done at the master level, while the read operation is performed on the slave.

23- What is the purpose of the Analysis Layer in a data warehouse?

- ☐ The analysis layer is used for storing backup copies of data.
- ☐ The analysis layer is where the data is collected from external sources.
- ☐ The analysis layer is where users interact with the data warehouse to extract insights.

24- How does horizontal scaling benefit Big Data systems?

- ☐ By adding more resources to increase capacity and performance.
- ☐ By upgrading the existing hardware to improve processing power.
- ☐ By reducing the number of nodes in the system to simplify management.

25- Can a system achieve all three aspects of the CAP theorem simultaneously?

- ☐ No, it can only achieve two out of the three at a time.
- ☐ Yes, it can achieve consistency, availability, and partition tolerance simultaneously.
- ☐ No, but in a business intelligence system, all aspects can be applied.

26- What are Data Marts in the context of data warehousing?

- ☐ Data Marts are large, centralized repositories of all company data.
- ☐ Data Marts are the same as Data Lakes.
- ☐ Data Marts are smaller subsets of the data warehouse focused on specific business domains.

27- What does atomicity ensure in ACID theorem?

- ☐ That the transaction can be partially completed, with some operations applied and others not.
- ☐ That all operations in a transaction are applied in sequence, but without any guarantee of success.
- ☐ That all operations in a transaction are completed successfully, or none are applied.

28- What does the NameNode do?

- ☐ It determines the mapping of blocks to DataNodes.
- ☐ It stores and manages the actual data in the Hadoop cluster.
- ☐ It processes the data stored in the HDFS and runs MapReduce jobs.

29- What is the role of DataNodes in HDFS?

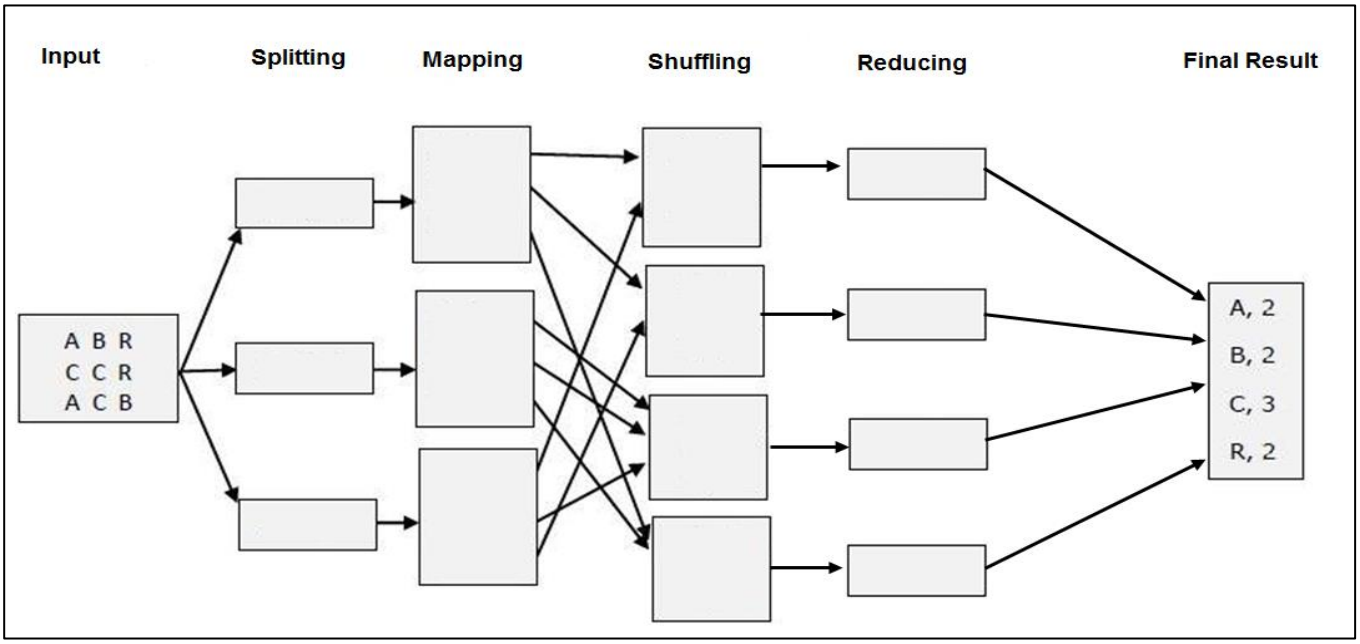
- ☐ DataNodes manage the HDFS namespace and handle file system metadata.
- ☐ DataNodes coordinate communication between clients and the NameNode.
- ☐ They handle storage and serve read/write requests.

30- How is the relationship between the operating system and the file system?

- ☐ File systems are used by operating systems to store and retrieve data for programs.
- ☐ The operating system uses the file system only for temporary data storage.
- ☐ The file system controls the operating system's memory allocation.

Exercise 2 (2 pt):

- The image above illustrates the stages of the **MapReduce** process to calculate the **frequency of each letter** in the original data (input). Each stage has a specific function in processing data.
- Your task is to **complete the missing information.**



Exercise 3 (3 pt):

Parallel processing is a method of computation where multiple tasks are executed simultaneously across multiple processors or cores. So:

Give me the **advantages** of parallel Data processing:

- ……..
  ……..
- ……..
  ……..
- ……..
  ……..

Give me the **disadvantages** Of Parallel Data Processing:

- ……..
  ……..
- ……..
  ……..
- ……..
  ……..