Homework 1

Course Code: CISC3018

Course Title: Cloud Computing and Big Data Systems

Assignment Submission Instructions

Please submit your assignment in **PDF or Word format (*.pdf or *.docx).**
Name your file using the following format: **StudentID_Name_AssignmentNumber (e.g., 20250101_LiMing_HW1).**
You do not need to rewrite the questions. Simply provide your answers in order by question number, for example:

Part B
1. True
2. False

...

For short answer questions, write your responses clearly under each corresponding question number.

Submission platform: **UMMoodle**
Deadline: **April 19, 2025**
Late submissions will be penalized according to the course policy.

Part A: Multiple Choice (Please note that each question has only one correct answer.)

1. Which of the following is not a mode in which Hadoop can run?
A. Standalone (local) mode
B. Pseudo-distributed mode
C. Interconnected mode
D. Fully distributed mode

2. Who is the author of Hadoop?
A. Martin Fowler
B. Kent Beck
C. Doug cutting

3. What is the default Block Size in HDFS?
A. 32MB
B. 64MB
C. 128MB

4. Which of the following is usually the main performance bottleneck in a cluster?
A. CPU
B. Network
C. Disk I/O
D. Memory

5. Which of the following is responsible for data storage in HDFS?
A. NameNode
B. JobTracker
C. DataNode
D. SecondaryNameNode

6. A bank plans to have a cluster with 90 nodes. If 3 control nodes are planned, how many data nodes should be planned for the cluster?
A. 87
B. 85
C. 90
D. 86

7. In Huawei's FusionInsight HD cluster, which of the following services cannot Spark read data from?
A. YARN
B. HDFS
C. Hive
D. HBase

8. Which of the following statements about clients uploading files to the HDFS file system is correct?
A. The client's file data is passed through the NameNode to the DataNode.
B. The client divides the file into multiple Blocks and writes them sequentially to each DataNode based on the DataNode's address information.
C. The client writes the entire file sequentially to each DataNode based on the DataNode's address information, and then the DataNode divides the file into multiple Blocks.
D. The client uploads data to only one DataNode, and then the NameNode is responsible for Block replication.

9. IaaS layer services provide customers with over the network.
A. Virtual servers, storage, and network resources
B. Application server platforms or development environments as services
C. Software applications (such as CRM, ERP, OA, etc.) as services

10. SaaS layer services are__.
A. Virtual servers, storage, and network resources

B. Application server platforms or development environments as services

C. Software applications (such as CRM, ERP, OA, etc.) as services

11. ____ is the cutting-edge technology of big data.
A. Data processing
B. Data analysis
C. Data definition
D. Data structure

12. The strategic significance of big data technology lies not in mastering vast amounts of data but in the _____ of meaningful data.
A. Data information
B. Professional processing
C. Speed processing
D. Content processing

13. Cloud computing is the development and application of ____ technologies.
A. Parallel computing
B. Grid computing
C. Distributed computing
D. All of the above

14. The common scenario of "installing a Linux virtual machine on Windows using VMware" belongs to _____.
A. Storage virtualization
B. Memory virtualization
C. System virtualization
D. Network virtualization

15. Which of the following is not a main characteristic of virtualization?
A. High scalability
B. High availability
C. High security
D. Simple implementation technology

16. From the current research perspective, which of the following is not a characteristic of cloud computing?
A. Ultra-large scale
B. Virtualization
C. Privatization
D. High reliability

17. _____ provides the interface for cloud users to request services and is also the entry point for users to use the cloud. Users can register, log in, customize services,

configure, and manage users through a web browser. Opening application instances is like operating a local desktop system.

A. Cloud client

B. Service catalog

C. Management system and deployment tools

D. Monitoring end

18. Web QQ belongs to _____.

A. SaaS

B. IaaS

C. PaaS

D. VaaS

19. In the FusionInsight HD system, what is the smallest processing unit of HBase? Where is the routing information between User Region and RegionService stored?

A. Zookeeper

B. HDFS

C. Master

D. Meta table

20. Which of the following HDFS commands can be used to check the integrity of data blocks?

A. hdfs fsck /

B. hdfs fsck / -delete

C. hdfs dfsadmin -report

D. hdfs balancer -threshold 1

Part B: True/False Questions

1. MapReduce and HBase are not tightly coupled; HBase can run normally without MapReduce, and MapReduce can directly access HBase.
2. If the NameNode crashes unexpectedly, the SecondaryNameNode will take over to keep the cluster running.
3. Since Hadoop is developed in Java, MapReduce only supports programming in Java.
4. Since slave nodes store data, the larger the disk, the better.
5. Hadoop's default scheduler is FIFO, and it supports job submissions from multiple pools.
6. Because HDFS stores multiple replicas, the NameNode does not have a single point of failure.
7. A MapReduce input split is exactly one block.

8. Before installing the data platform software package, the /etc/hosts file on the local node must be configured with the hostname and business IP mapping of all nodes.
9. Cloud computing evolved from grid computing and is capable of providing resources on-demand.
10. Consumers of cloud computing need to manage or control the underlying infrastructure such as networks, operating systems, and storage.
11. Cloud computing is a computing model that provides on-demand, convenient, and ubiquitous network access to a shared pool of configurable computing resources.
12. A smart tourism platform can collect and summarize customer-related data, analyze tourist source regions and travel trajectory characteristics, and provide real-time visitor traffic statistics.
13. The Internet is a massive cloud.
14. Hadoop's default scheduling policy is FIFO.
15. Big data only refers to large data volume.
16. Big data consists only of structured data.
17. Data cleansing means removing duplicate records.
18. Each node in the cluster should be equipped with RAID to prevent single disk failure from affecting the entire node.
19. Because HDFS stores multiple replicas, the NameNode does not have a single point of failure.
20. Hadoop supports random write operations on data.

Part C: Short Answer Questions

1. Briefly explain big data technology.
2. Describe the characteristics of big data (5Vs).
3. Briefly explain the concept and characteristics of cloud computing.
4. Briefly describe the categories of cloud computing service models and define each.
5. Briefly describe the core components of Hadoop and their functions.