Université Libanaise
Faculté d'Information
Br I

الجامعة اللبنانية
كلية الاعلام
الفرع الاول

الامتحانات النهائية للفصل السادس
من العام الجامعي 2025-2026

المرحلة: الإجازة
السنة المنهجية: الثالثة
الاختصاص: علم البيانات - Data Science

المادة: Big Data
المدة: ساعتين
الأستاذ: د. حسين هزيمة

Documents are NOT authorized

Part 1: (20 pts) true-false questions:

1. Hadoop is an open-source framework for distributed storage and processing of large datasets. _____
2. HDFS stands for Hadoop Distributed File System. _____
3. The NameNode stores the actual data blocks in HDFS. _____
4. DataNodes are responsible for storing HDFS data blocks. _____
5. HDFS files are divided into blocks before storage. _____
6. HDFS typically replicates each block multiple times for fault tolerance. _____
7. Small files are always handled efficiently in HDFS. _____
8. MapReduce consists of Map and Reduce phases. _____
9. A Mapper usually processes data stored on its local node. _____
10. Shuffle and Sort occur after the Reduce phase. _____
11. Kafka follows a publish-subscribe messaging model. _____
12. Kafka messages are persisted on disk. _____
13. Kafka consumers publish messages to topics. _____
14. Kafka supports horizontal scalability through partitioning. _____
15. Kafka can integrate with Apache Spark. _____
16. Spark uses a master-slave architecture. _____
17. RDD stands for Resilient Distributed Dataset. _____
18. Spark RDD lineage helps recover lost partitions. _____
19. Spark SQL is used for structured data processing. _____
20. MLlib is Spark's machine learning library. _____

Part 2: (20 pts) MCQ questions

1. Which Hadoop component stores metadata about files?

A. DataNode
B. Reducer
C. NameNode
D. Executor

2. In HDFS, files are stored as:

A. Tables
B. Blocks
C. Topics
D. Streams

3. Which MapReduce phase generates intermediate key-value pairs?

A. Reduce
B. Shuffle
C. Map
D. Sort

4. Which Hadoop ecosystem project is commonly used for SQL-like queries?

A. Hive
B. Kafka
C. Docker
D. Jenkins

5. What is the default purpose of block replication in HDFS?

A. Compression
B. Encryption
C. Fault Tolerance
D. Visualization

6. Kafka mainly follows which architecture?

A. Peer-to-peer
B. Publish-subscribe
C. Ring topology
D. Mesh topology

7. Which Kafka component sends messages?

A. Consumer
B. Broker
C. Producer
D. Executor

8. Kafka stores messages primarily on:

A. GPU memory
B. Cache memory

C. Disk
D. ROM

9. Which service was traditionally used by Kafka for coordination?

A. Hadoop YARN
B. ZooKeeper
C. Docker
D. Hive

10. Kafka scalability is mainly improved through:

A. Compression
B. Replication
C. Partitioning
D. Encryption

11. Which Spark component coordinates application execution?

A. Executor
B. Worker
C. Driver
D. Reducer

12. Spark is generally faster than MapReduce because it relies heavily on:

A. Disk processing
B. In-memory computation
C. Tape storage
D. Manual caching

13. What does RDD stand for?

A. Resilient Distributed Dataset
B. Real Data Distribution
C. Replicated Data Driver
D. Reduced Distributed Dataset

14. Which Spark library supports machine learning?

A. Spark SQL
B. GraphX
C. MLlib
D. Kafka Connect

15. Which Spark module supports graph analytics?

A. GraphX
B. Hive
C. HDFS
D. Oozie

16. Spark Streaming is mainly used for:

A. Real-time analytics
B. File compression
C. Web hosting
D. Database backup

17. Which language was originally used extensively for Spark development?

A. C#
B. Scala
C. PHP
D. Ruby

18. Spark fault tolerance is achieved through:

A. Encryption
B. Replication only
C. Lineage
D. Compression

19. Which Spark component executes tasks?

A. Driver
B. Worker Executor
C. Producer
D. Broker

20. Which combination correctly matches technologies and purposes?

A. HDFS → Messaging, Kafka → Storage
B. Kafka → Messaging, Spark → Analytics
C. Spark → Storage, Hadoop → Messaging
D. MLlib → Storage, HDFS → Machine Learning

Part 3: (60 pts) Problem solving questions

Problem Solving Question 1 — Kafka (30 pts)

A smart hospital uses Apache Kafka to collect patient monitoring data in real time.

The hospital has:

- 5 patient monitoring devices acting as producers
- 1 Kafka broker
- 3 consumer applications:
  - Emergency Alert System
  - Patient Dashboard
  - Data Analytics Engine

Each device sends:

- 2,000 messages per second

Answer the following:

1. Calculate the total number of messages entering Kafka every second.
2. Explain the role of the Kafka broker in this architecture.
3. Identify which components are producers and which are consumers.
4. Explain why Kafka replication is important in this healthcare scenario.

Problem Solving Question 2 — Spark (30 pts)

A company uses Apache Spark to analyze website log files.

The dataset size is:

- 800 GB

The Spark cluster contains:

- 1 Driver node
- 4 Worker nodes
- Each worker has 2 Executors

Tasks:

1. Calculate the total number of executors in the cluster.
2. Explain how Spark distributes tasks among worker nodes.
3. Describe how RDD lineage helps Spark recover from failures.
4. Explain why Spark is faster than traditional disk-based MapReduce systems for iterative analytics.

Good Work