Université Libanaise
Faculté d'Information
Br I

الجامعة اللبنانية
كلية الاعلام
الفرع الاول

الامتحانات النهائية للفصل السادس
من العام الجامعي 2023-2024

المادة: Big Data	المرحلة: الإجازة
المدة: ساعة ونصف	السنة المنهجية: الثالثة
الأستاذ: د. حسين هزيمة	الاختصاص: علم البيانات - Data Science

Documents are NOT authorized

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

Answer the following questions by true or false, and explain the false statements.

1. RDD must be created in MapReduce architecture.	(T)	F	
2. Cluster manager can be replaced with YARN in standalone Spark installation.	T	(F)	
3. Hadoop hdfs can run on any programming language.	T	(F)	
4. Hadoop supports graph analytics tasks.	(T)	F	
5. Spark supports only machine learning tasks.	(F)	(F)	
6. Kafka publish-subscribe mechanism is similar to the email massaging mechanism.	(T)	F	
7. Hadoop MapReduce is execution is lazy evaluation.	(T)	(F)	
8. Data velocity means the generation of many different types of data.	T	(F)	
9. When installing Hadoop on Windows, four processes will run; if one of them is failed can we execute a MapReduce job?	T	(F)	
10. Hadoop can run normally without JVM?	T	(F)	

Question 2: (20 pts) definitions

Describe briefly the following software:

HBase	Kafka	Zookeeper
Hbase is an Apache Hadoop software created for processing data blocks stored in files. HBase provides parallel processing for data streams, which makes it an efficient software for big data generation.	Kafka is a Hadoop processing software used to process large amount of data. It allows for the partitioning of data in files, where partitions are distributed on different blocks and in different datanodes.	It's the software in which ~~Spark~~ Kafka is built on top of it.

Question 3: (30 pts) Kafka and Spark

Assume we have the following data stored in a file named "data.txt".

```
Lorem Ipsum is simply dummy text of the printing and
typesetting industry. Lorem Ipsum has been the
industry's standard dummy text ever since the 1500s,
when an unknown printer took a galley of type and
scrambled it to make a type specimen book. It has
survived not only five centuries, but also the leap into
electronic typesetting, remaining essentially unchanged.
```

The data is extracted from 5 different resources.

1. Assume the data above is stored on hdfs, given the following settings, how many blocks are needed to store the above file on hdfs?
   a. Data size: 745 MB
   b. Block size: 65 MB

   65 × 12 = 780; 650 + 130 = 780; 650 + 65 = 715

   1.1 Are these blocks stored on namenode or datanodes? datanodes
2. How many topics do you think we must create for this data to be published in a Kafka broker? Justify your answer.
3. If this data is streamed in real-time and stored on hdfs, can we run a Spark graph analytics task on it? Justify your answer?
4. Write a Scala code to create an RDD for the input data in the above file using three different methods. Name each method and describe it.
5. Select and use an RDD from part 3). How many transformations and actions we can run? Write two of them.

Question 4: (30 pts) Hadoop

Assume that we have a Hadoop single-node cluster is configured on your machine. Write the necessary Hadoop commands to run the following MapReduce job.

1. Format your namenode. hadoop name -format
2. What is the result of this command `start-dfs.cmd` ?
3. What is the result of the following command `start-yarn.cmd` ?
4. Create a directory on hdfs named "bigdata".
5. Upload a dataset named "big-dataset" from your pc to the created hdfs folder. The dataset size is 1.5 TBs.
6. Read the content of the dataset on your command line.
7. Run wordcount built-in MapReduce job.
8. How many blocks created for your dataset on hdfs? What is the size of each block?

Good Work

1 TB = 1024 GB
1 GB = 1024 MB

Big Data (2023, 2024)

Question 1

1) F: RDDs are Spark-specific abstraction (independent of MapReduce)
2) T
3) F: HDFS requires Java/JVM (but we can use APIs to interact with other languages)
4) T: (Hadoop ecosystem include tools like Giraph for graph processing)
5) F: Spark support ML, streaming, SQL, graph processing
6) T
7) F: Lazy evaluation applies to Spark RDDs not Hadoop MapReduce
8) F: velocity → speed, variety → different types
9) F: if any process fail Hadoop won't run
10) F: built in java → require JVM

Question 2

- HBase: distributed, column-oriented NoSQL db. that runs on top of HDFS. Provides real time read/write access to large datasets with low latency

- Kafka: distributed streaming platform used for building real-time data pipeline. Works on publish-subscribe model and handle high-throughput low-latency data streams

- ZooKeeper: A centralized coordination service for distributed systems. Manages configuration, naming, synchronization and group services.
  Critical for Kafka, HBase distributed systems

Question 3

1) Block size : 65 MB
=> nb. of blocks: 745/65 = 11.46 = 12 block

1.1) Blocks are stored on DataNodes
(NameNode only store metadata)

2) 5 topics one for each resource
↳ maintain data isolation by source
↳ enable independent processing per source
↳ Aligns with Kafka's partitioning strategy for parallel processing

3) Yes, but with preprocessing
Spark GraphX can process data in HDFS. but:
. the text must be converted to graph structure (vertices/edges)
. Real time streaming requires Spark Streaming to ingest Kafka data
. Graph algo. ~~can~~ can then be applied to the processed graph

4) same as in 2022, 2023

5). unlimited Transformations (e.g. map, filter, flatMap, distincts...)
. Actions Trigger computation (e.g. count, collect, reduce, saveAsTextFile...)

```
// Transformation 1: Tokenize words
val words = rdd1.FlatMap(line => line.split(" "))
```

```
// Transformation 2 : Filter long words
val longWords = words.filter(word => word.length>8)

//Action 1: Count total words
val wordCount = words.count()
//Action 2: Save filtered words
longWords.saveAsTextFile("hdfs:///output/long_words")
```