Université Libanaise
Faculté d'Information
Br I

الجامعة اللبنانية
كلية الاعلام
الفرع الاول

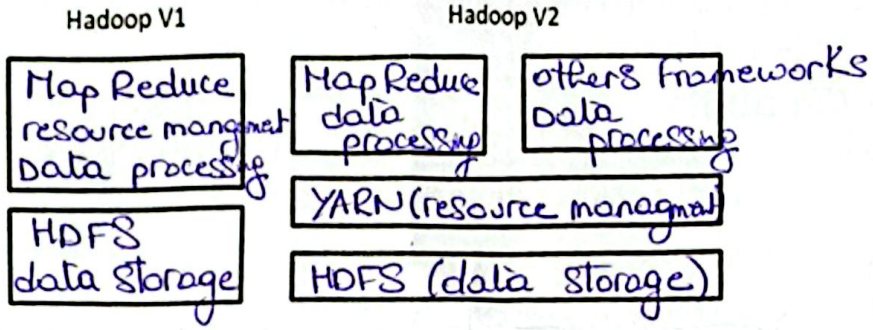
الامتحانات النهائية للفصل السادس - الدورة الثانية
من العام الجامعي 2024-2025

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

Documents are NOT authorized

Question 1: (20 pts) Hadoop

Compare between Hadoop V1 and V2 by filling the following diagram:



Question 2: (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	Spark concept don't require MapReduce
3. Hadoop hdfs can run on any programming language.	T	(F)	written in Java
4. Hadoop supports graph analytics tasks.	(T)	F	tools like Giraph
5. Spark supports only machine learning tasks.	T	(F)	SQL, ML (MLlib), streaming, Graph (GraphX)
6. Kafka publish-subscribe mechanism is similar to the RSS feeds mechanism.	(T)	F	
7. Hadoop MapReduce is execution is lazy evaluation.	T	(F)	Spark RDDs uses Lazy evaluation
8. Data variety means the generation of many different types of data.	(T)	F	
9. When installing Hadoop on Windows, 5 processes will be started simultaneously?	T	(F)	4 processes (DataNode, NameNode, NodeManager, Resource manager)
10. Hadoop can run normally without JVM?	T	(F)	Java based require JVM

Question 3: (25 pts) Hadoop, 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

   1.1 Are these blocks stored on namenode or 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: (25 pts) Hadoop

Suppose that you have a Hadoop multi-node cluster is configured on your machine. Write the necessary Hadoop commands to run the following word count job.

1. Format your namenode.
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 "word-count".
5. Upload a dataset named "word-dataset" from your pc to the created hdfs folder. The dataset size is 1 TBs.
6. Read the content of the dataset on your command line.
7. Run wordcount job.

Question 5: (10 pts) Big data architecture

1. Complete the following table:

MapReduce1	YARN
Jobtracker	Resource manager + application master + Timeline server
TaskTracker	Node manager
Slot	Container

2. Which one considered more scalable: YARN or MapReduce? Why?

Good Work