Distributed Data Processing and Analysis (BIG DATA) (BİL401/501) - Erdoğan Doğdu
Spring 2014 – Midterm - TOBB University of Economics and Technology
Feb 20, 2014 – 90 min **Student Number**: **Name**:

Question	***1-9***	***10***	***11***	***12***	***13***	***Total***
Value	30	30	20	5	15	*100*
Points						

(Circle the right answer(s) for the following questions)

1. What is Hadoop used for? *(choose one)*
    a. Online transaction processing (bank EFT transfers, etc.)
    b. Intensive calculations on a small data set
    c. Fast access to data
    d. Sequential processing of a large data set
    e. None of the above
2. What is HDFS? *(choose one)*
    a. Hadoop Distributed File System
    b. Hadoop Data File System
    c. Hadoop Data Folders System
    d. Hash Data Files System
    e. None of the above
3. Which one of the below statements is NOT TRUE for HDFS? *(choose one)*
    a. Data access is via Map/Reduce
    b. 3 replicas are kept for each piece of data by default
    c. Can create, delete, copy, ~~update~~ data
    d. Designed for streaming reads, not for random access
    e. Data blocks are kept on separate data nodes
4. What is the default HDFS block size and BigInsight's block size? *(choose one)*
    a. 32 MB / 64 MB
    b. 64 MB / 64 MB
    c. 64 MB / 128 MB
    d. 64 MB / 252 MB
    e. 128 MB / 252 MB
5. Which statements below are correct about HDFS architecture? *(mark all that apply)*
    a. **HDFS has a master/slave architecture**.
    b. **NameNode manages the file system**.
    c. **DataNode manages storage attached to the nodes**.
    d. Consists of a DataNode and many NameNodes.
    e. **DataNode periodically reports status to NameNode**.

6. Which one of the following statements is NOT TRUE about parallel computing and distributed computing? *(choose one)*
    a. Parallel computing is about multiple CPUs processing over shared data.
    b. Distributed computing is about a cluster of multiple separate computers working on pieces of local data that is separated from a large data set.
    c. Distributed computing uses network messaging heavily.
    d. Parallel computing takes advantage of CPU parallelism using multi-threading.
    e. None of the above
7. Which one of the following computing paradigm(s) is MapReduce mainly using? *(choose one)*
    a. Parallel computing
    b. Distributed computing
    c. Grid computing
    d. Quantum computing
    e. None of the above
8. Which statements below are correct about the MapReduce engine? *(mark all that apply)*
    a. **Has a master/slave architecture**.
    b. TaskTracker controls job execution on multiple JobTrackers.
    c. **JobTracker accepts MapReduce jobs from clients, pushes map and reduce tasks to TaskTrackers**.
    d. **TaskTracker runs map and reduce tasks**.
    e. **Job Tracker monitors tasks and TaskTrackers**.
9. The input to a mapper takes the form <k1,v1>. What form does the mapper's output take? *(choose one)*
    a. <list(k2), v2>
    b. list(<k2,v2>)
    c. <k2, list(v2)>
    d. <k1,v1>
    e. None of the above

10. (30p) The classical MapReduce example "Word Count" is given below in pseudo code:

```
class Mapper
   method Map(docid id, doc d)
      for all term t in doc d do
         Emit(term t, count 1)

class Reducer
   method Reduce(term t, counts [c1, c2,...])
      sum = 0
      for all count c in [c1, c2,...] do
          sum = sum + c
      Emit(term t, count sum)
```

Using the same syntax, write MapReduce solutions for the following problems:

a. Scan all documents and find the keywords "President Gül", "President Obama", "President ...." and list the names of all presidents {"Gül","Obama","...."}. No count is needed, just the unique names. Hint: Find the word "President" first and the next word will be sent to the final list.

```
Ans:

class Mapper
   method Map(docid id, doc d)
      for all term t in doc d do
         if t.equals("President")
           prev=true
         else
           if prev
             Emit(term t, count 1)
             prev=false

class Reducer
   method Reduce(term t, counts [c1, c2,...])
      Emit(term t, count 0)
```

b. Scan all documents and find the list of <word_length, count> pairs, that is the number of words for each word length (number of words of length 1, 2, 3, etc.).

Ans:

```
class Mapper
   method Map(docid id, doc d)
      for all term t in doc d do
         Emit(term t.length(), count 1)

class Reducer
   method Reduce(term t, counts [c1, c2,...])
      sum = 0
      for all count c in [c1, c2,...] do
          sum = sum + c
      Emit(term t, count sum)
```

11. (20p) Complete the following sentences:

    a. _____ invented MapReduce. (**Ans:** Google)
    b. _____ is a widely used open source implementation of MapReduce that is developed and maintained by _____ with significant contributions from the company _____. (**Ans:** Hadoop, Apache, Yahoo).
    c. The main algorithmic strategy of MapReduce is _____ and _____. (**Ans:** divide, conquer)
    d. The name Hadoop comes from _____. (**Ans:** the name of the developer's son's toy elephant)
    e. Hadoop is written in _____ language. (**Ans:** Java)
    f. Pig is developed by _____. (**Ans:** Yahoo)
    g. Hive is developed by _____. (**Ans:** Facebook)
    h. Jaql is developed by _____. (**Ans:** IBM)
    i. The language of Pig is _____. (**Ans:** Latin)
    j. The language of Hive is _____. (**Ans:** HiveQL)
    k. Three main steps of a Pig program are _____, _____, _____. (**Ans:** Extract, Transform, Load)
    l. Hive language is similar to _____. (**Ans:** SQL)
    m. Pig, Hive, and Jaql translate programs written in high-level languages to _____ jobs. (**Ans:** MapReduce)

12. (5p) Twitter is one those Web sites that collect Big Data. Members send about 60 million tweets a day on Twitter. How many bytes of data is added to the big data on Twitter in a day and in a year approximately? Give your numbers in at most 3 digits (999) and the right unit of bytes, e.g. 128MB.

    Ans: 60M tweets/day x 140 char (max) x 2 bytes = 16.8GB a day (max) → 6.14TB a year (max)

13. (15p) Give three examples of big data (the sources of them), data type and your estimation of how big they are?

Source	*Data type*	*Size*
Mobese recordings from 1000 cameras in Ankara	Video	**1 camera**, 30fps, H.264 encoding, 640x480pixel, 24hours, 31days = **645GB** (http://www.ezwatch.com/dvr-storage-calculator) (x **1000 cameras**) = **645 TB/month**
Sağlık-Net data coming from 1483 hospitals	XML	1483 hospitals x 1MB/day (avg) x 365 days = **541GB/year**
Images of earth from Turkish RASAT Earth Observation Satellite	Image	3 million square km/year 900 sq.km/picture ~30MB/picture 3M/900 pictures x 30MB= 100GB/year http://www.tubitak.gov.tr/tr/haber/yerli-gozlem-uydusu-rasat-uzayda-ikinci-yilini-tamamladi http://uzay.tubitak.gov.tr/tr/haber/rasattan-alinan-goruntuler-iki-boyutlu-haritaya-donusecek