Second Semester M.Tech. Degree Examination, June/July 2016
Managing Big Data

Time: 3 hrs.   Max. Marks:100

Note: *Answer any FIVE full questions.*

1 a. Discuss the variety dimension of Big Data. What is structured, unstructured and semi-structured data? Explain with examples. **(10 Marks)**
   b. How big data analytics can help in detecting fraud? Explain with industry example. **(10 Marks)**

2 a. Explain the three business models for big data. **(06 Marks)**
   b. What is impedance mismatch? Why is it a major source of frustration for application developers? **(06 Marks)**
   c. What are different ways of constructing version stamps? What are their pros and cons? **(08 Marks)**

3 a. Compare and contrast sharding and replication. Explain the advantages and limitations of sharding and replication. **(10 Marks)**
   b. Explain the trade off involved in consistency and availability with respect to the CAP theorem. **(10 Marks)**

4 a. Write brief note on materialized views. **(04 Marks)**
   b. A Hadoop cluster is configured to use block size of 128 MB and replication factor of 3. A user has a file "mylife" of size 520 MB and uses the following command:
      "hadoop fs-put myfile /user/hadoop".
      i) What is the size of the last block of the file when stored on HDFS?
      ii) How many blocks will be stored on HDFS for the file? **(04 Marks)**
   c. Discuss the factors considered for design of HDFS. Which are the areas where HDFS is not a good fit today? **(12 Marks)**

5 a. What is the mechanism used by HDFS to ensure data integrity? How data nodes and clients verify the data for integrity? **(06 Marks)**
   b. What is serialization and why is it needed? Explain the process of serialization and de-serialization in Hadoop. **(08 Marks)**
   c. Explain the following in brief,
      i) Hadoop streaming   ii) Reduce function. **(06 Marks)**

6 a. Explain mapreduce workflows. **(10 Marks)**
   b. Given a text file containing multiple words, we need to count how many times each word appears in the file using mapreduce. Explain shuffle and sort in detail for this word count problem using a neat diagram. **(10 Marks)**

7 a. Explain with a neat diagram, how hadoop runs a map reduce job using YARN. **(10 Marks)**
   b. Provide overview of HBASE data model. **(10 Marks)**

8 Write a brief notes on:
   a. PIG data model, i.e., data types.
   b. Hive QL.
   c. CASSANDRA
   d. AVRO. **(20 Marks)**

* * * * *