CSPE733

USN	1	M	S							

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(AUTONOMOUS INSTITUTE, AFFILIATED TO VTU)

BANGALORE – 560 054

SEMESTER END EXAMINATIONS – MAY / JUNE 2014

Course & Branch	:	**B.E.- Computer Science and Engineering**	**Semester**	:	**VIII**
Subject	:	**Big Data and Data Science**	**Max. Marks**	:	**100**
Subject Code	:	**CSPE733**	**Duration**	:	**3 Hrs**

Instructions to the Candidates:

- Answer one full question from each unit.

UNIT – I

1. a) Explain in detail the different challenges of analyzing data at scale. (08)
   b) With a simple schematic, explain different stages in a generic data processing pipeline. (06)
   c) Explain the following with examples (06)
      a. Overfitting
      b. Bias in large amounts of Data
      c. Indexing

2 a) Explain the following Data Store terminologies (08)
      a. tuple, record, object, document
      b. Key-Value store, document store, record store and relational store
   b) Explain the challenges in big data analytics. (08)
   c) Explain the following data transformation terminologies (06)
      a. Machine Translation
      b. Autocoding
      c. Term Extraction

UNIT – II

3. a) Explain in detail Different Component parts of Data Science. (10)
   b) Explain a). Vector Space Model and Cosine Similarity b) Feature selection (06)
   c) Explain Different classes of analytic techniques. (04)

4. a) Explain some of the classes of implementation constraints found in a data intensive problem with suitable examples. (08)
   b) Explain briefly feature selection, data varacity and curse of dimensionality. (06)
   c) Explain the fractal analytic model. (06)

UNIT – III

5. a) Explain in detail, K-Means clustering Algorithm and bring out the drawbacks/limitations or problems with K-Means clustering. (12)
   b) Explain Data Reduction in context to parameter describing data. (04)
   c) Explain the analytical tasks : Recommending and Modeling. (04)

6. a) Explain Naive Bayes Classification mechanism with relevant mathematical assumptions and a suitable example and explain the advantages and disadvantages of Naïve Bayes Classification. (12)
   b) Explain different corrections involved in data normalization. (04)
   c) Explain the significance and use of measures of central tendency. (04)

UNIT – IV

7. a) List and explain the commonly used features in text processing. (08)
   b) Explain with a schematic the System architecture for generic text mining system. (12)

8. a) Explain Frequent set mining in a textual context. (08)
   b) Explain Text categorization performed using Supervised Bayesian classifier. (12)

UNIT – V

9. a) Explain with a schematic the workflow of a mapreduce program execution in a distributed framework. (12)
   b) Explain how fault tolerance is achieved in a distributed set-up explained in Google's Mapreduce implementation. (08)

10. a) Explain how Hadoop solves various execution challenges at scale with suitable diagrams. (10)
    b) Explain the different components of HDFS (a) NameNode (b) DataNode. (04)
    c) What are the problems, which HDFS solves when compared to other distributed file systems? Also explain the assumption made in order to achieve the kind of performance HDFS offers. (06)

****