Time: 3 Hours **Marks: 80**

Note: 1. Question 1 is compulsory

2. Answer any three out of remaining questions.
3. Assume suitable data wherever required and justify the same.

Q1 a) Explain time series mining with an appropriate example. [5]

b) What is sharding? Explain the advantages of sharding. [5]

c) What is data leakage with respect to big data? [5]

d) What are the basic differences between relational database and HDFS? [5]

Q2 a) Use PCA to transform 2D data space to 1D data space for the given matrix A. [10]

$$A = \begin{matrix} 0 & 1 \\ -2 & -3 \end{matrix}$$

b) Explain singular value decomposition (SVD) with an example. [10]

Q3 a) Explain Gaussian (normal) distribution with respect to pdf and cdf and its use in statistics. [10]

b) Explain Independent Component Analysis (ICA) in descriptive modeling. [10]

Q4 a) What type of problem were you looking to solve with text mining? How did you know how to text mine? What could be the challenges when text mining? [10]

b) What is Recommendation System (RS)? Explain types of RS and various issues and challenges in RS. [10]

Q5 a) Draw and describe the information visualization process. [10]

b) How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regressions? [10]

Q6 a) Describe the working of the Map-Reduce with an example. [10]

b) What is No SQL? Compare No SQL with SQL. [10]
