

(2^{1/2} Hours)

[Total Marks: 75]

- N. B.: (1) All questions are compulsory.
 (2) Make suitable assumptions wherever necessary and state the assumptions made.
 (3) Answers to the same question must be written together.
 (4) Numbers to the right indicate marks.
 (5) Draw neat labeled diagrams wherever necessary.
 (6) Use of Non-programmable calculators is allowed.

1. **Attempt any three of the following:** 15
 - a. Why Big Data? Explain characteristics of Big data.
 - b. Digital data can be broadly classified into structured, semi-structured, and unstructured data. Explain.
 - c. Write a short note on evolution of Big Data.
 - d. What are the main phases of the Data Analytics? Explain using suitable diagram.
 - e. Explain the ACID property in RDBMS.

2. **Attempt any three of the following:** 15
 - a. Explain the application of K-Mean in image processing, customer segmentation and medical analysis.
 - b. Write a short note on Association Rules.
 - c. Explain Linear Regression Model with Normally Distributed Errors.
 - d. Explain any three applications of logistic regression model applied to situations in Government and the Private sector.
 - e. What is clustering? State its advantages and disadvantages.

3. **Attempt any three of the following:** 15
 - a. Write a short note on Prediction Tree.
 - b. What is sentiment analysis? How it can be carried out? Explain it in detail.
 - c. What is Box-Jenkin Methodology? Explain.
 - d. Term frequency-inverse document Frequency (TFIDF) is widely used in information retrieval and text analysis. Explain
 - e. Explain Naïve Bayes Theorem using suitable example.

4. **Attempt any three of the following:** 15
 - a. What is data product? Explain.
 - b. Write a short note on the Hadoop Architecture.
 - c. Explain the Hadoop Big Data Operating System.
 - d. Write a short note on Hadoop Streaming.
 - e. Describe the concept of Spark Stack.

5. **Attempt any three of the following:** 15
 - a. MapReduce and Spark allow developers and data scientists the ability to easily conduct data parallel operations, where data is distributed to multiple processing nodes and computed upon simultaneously, then reduced to a final output. Explain distributed analysis and patterns using MapReduce and Spark.
 - b. Explain Design Patterns using suitable example.
 - c. Explain the concept of Data Mining and Data Warehousing.
 - d. What is Data Ingestion? Explain.
 - e. Explain with example Relations, Tuples and filtering in context of Pig.
