(2 Hours.) **[Total Marks: 60]**

N. B.: (1) **All** questions are **compulsory.**
(2) Numbers to the **right** indicate **marks**.
(3) Make **suitable assumptions** wherever necessary and **state the assumptions** made.
(4) Answers to the **same question** must be **written together.**
(5) **Mixing** of Sub-Questions is not allowed.
(6) Draw **neat labelled diagrams** wherever **necessary**.

1. Attempt any two of the following: **12**

a. What is Big Data? State its characteristics.
b. Discuss the challenges of Big data.
c. Describe the current analytical architecture for data scientists.
d. What are different phases of the Data Analytics Lifecycle? Explain in detail.

2. Attempt any two of the following: **12**

a. What is clustering? State its advantages & disadvantages.
b. Explain the concept of K-means clustering algorithm.
c. Write a brief note on Apriori algorithm.
d. What is Logistic regression? Explain in detail.

3. Attempt any two of the following: **12**

a. Write a short note on Decision trees.
b. Explain a probabilistic classification method based on Naive Bayes' theorem.
c. Explain Time Series Analysis in detail.
d. What is tokenization? Explain how it is used in text analysis.

4. Attempt any two of the following: **12**

a. Describe the concept of MapReduce.
b. State the components of Spark Stack.
c. Write a short note on Hadoop Distributed File System.
d. Explain in detail the Hadoop Architecture.

5. Attempt any two of the following: **12**

a. What is the functionality of the filter mapper? Explain in detail with example.
b. Write a short note on Data Ingestion.
c. Explain with example Relations, tuples and Filtering in context of Pig.
d. Write the entire procedure with appropriate commands for importing data from MySQL to HDFS.

****