316318

25226

3 Hours / 70 Marks Seat No. | | | | | | | | |

Instructions – (1) All Questions are *Compulsory.*

(2) Answer each next main Question on a new page.

(3) Illustrate your answers with neat sketches wherever necessary.

(4) Figures to the right indicate full marks.

(5) Assume suitable data, if necessary.

(6) Mobile Phone, Pager and any other Electronic Communication devices are not permissible in Examination Hall.

Marks

1. Attempt any FIVE of the following : **10**

a) Define term Big Data.

b) Discuss the characteristics of Big Data. (Any two)

c) State the use OR key-value pair in MapReduce.

d) Explain the use of data replication in HDFS.

e) Define schema-less model.

f) List any two applications of Pig.

g) Name any two features of Spark.

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

a) List and explain the classification of data.

b) Explain the challenges involved in Big Data management.

c) State the functions of DataNode.

d) Illustrate how to improves performance in Hadoop with an example.

P.T.O.

Marks

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

a) List the features of NoSQL databases.

b) Explain the trade-off between Consistency and Availability in CAP theorem.

c) Explain NoSQL to manage Big Data.

d) Define Apache Hive and list its features.

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

a) Describe how Hive queries are converted into MapReduce jobs.

b) Explain the concept of schema-on-read in Hive.

c) List and explain the characteristics of Spark streaming platform.

d) Explain why Spark is faster than Hadoop MapReduce.

e) Implement a Spark program flow using RDD transformations and actions.

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

a) Explain the importance of Big Data analytics in business intelligence.

b) Discuss the limitations of Hadoop MapReduce and their implications.

c) Demonstrate how YARN manages resources for multiple Hadoop jobs.

6. Attempt any TWO of the following : **12**

a) Demonstrate how MongoDB manages large datasets using collections and documents.

b) Apply partitioning and file formats in Hive to optimize query performance.

c) Apply MLlib algorithms to perform classification or clustering on a dataset.
