Time: 3 Hours   Max. Marks: 75

Note: This question paper contains two parts A and B.

Part A is compulsory which carries 25 marks. Answer all questions in Part A. Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks and may have a, b, c as sub questions.

PART- A

(25 Marks)

1.a) Is it desirable to build a system that helps users perform appropriate analysis? Explain. [2]

 b) Whether use of a centralized server for the MDW database is the best Practice and Caveat for Management Data Warehouse? Justify.   [3]

 c) What is Apache Spark? Explain key features of Spark.   [2]

 d) Define HDFS. Discuss the HDFS Architecture and HDFS Commands in brief.   [3]

 e) Explain some of the applications of Big data.   [2]

 f) What is impedance mismatch? What are the major difficulties faced by big data application developers?   [3]

 g) What do you mean by hypothesis? Explain.   [2]

 h) Discuss the criteria for evaluating case study.   [3]

 i) What do you mean by data visualization? Explain.   [2]

 j) Write names of some data visualization tools. Also, discuss properties of different tools.
                                                                                   [3]

PART-B

(50 Marks)

2. What role should data quality and governance play in any organization? Also how are data quality and governance related? Explain.   [10]

OR

3. What software requirements does user end data analysis impose upon a data management project? Explain with a suitable example.   [10]

4. Write Map Reduce code for counting occurrences of specific words in the input text file(s). Also write the commands to compile and run the code.   [10]

OR

5. What is Hbase? Discuss in detail the data model and Implementation aspect of Hbase.
                                                                                   [10]

6. What are the benefits of Big Data? Discuss challenges under Big Data. How Big Data Analytics can be useful in the development of smart cities.   [10]

OR

7. What is RDD? Explain about transformations and actions in the context of RDDs. State and explain RDD operations in brief.   [10]

8. Describe the steps involved in sampling design. Discuss the criteria for selecting a sampling procedure. [10]

OR

9. Differentiate type i error and type ii error. How is a hypothesis tested? Explain with an example. [10]

10. Explain the principle of linear interpolation along a line segment P_0P_1, when P_0 and P_1 are data points. Using diagrams show how this principle can be extended to bi-linear interpolation in a square grid cell with vertices $P_0...P_3$, and to tri-linear interpolation in a cubic grid cell with vertices $P_0...P_7$. [10]

OR

11. A general model of the visualization process is a pipeline with four stages: data generation, pre-processing (filtering), mapping, and rendering. The user can interact with the visualization process at each of these stages. Indicate for each of the following input actions at which stage it will influence the process:
a) Choosing a color scale
b) Selecting a part from a data set to be visualized.
c) Changing measurement parameters
d) Selecting an iso-value
e) Specifying a viewpoint and lighting conditions. [10]

--ooOoo--