| **Examination** | **: \_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Seat No** | **: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| --- | --- | --- | --- |
| **Date** | **: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Day** | **: \_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Time** | **: \_\_\_\_\_\_\_\_ to \_\_\_\_\_\_\_\_\_** | **Max. Marks** | **: 60** |

| **INSTRUCTIONS:** | | | | |
| --- | --- | --- | --- | --- |
| 1. | Answer each section in a separate answer book. | | | |
| 2. | Figures to the right indicate maximum marks for that question. | | | |
| 3. | The symbols used carry their usual meanings. | | | |
| 4. | Assume suitable data, if required & mention them clearly. | | | |
| 5. | Draw neat sketches wherever necessary. | | | |
| **SECTION – I** | | | | |
| **Q.1** | **Do as directed.** | | **[10]** | |
|  | (a) | Fill in the blanks:  (i) Data science involves multiple disciplines: \_\_\_\_ , \_\_\_\_\_ and \_\_\_\_.  (ii) The full form of data format CSV is \_\_\_\_\_\_\_\_\_\_ and the full form of data  format JSON is \_\_\_\_\_\_. | [2] | |
|  | (b) | Match the following:   | Column A | Column B | | --- | --- | | HDFS master component | ResourceManager | | HDFS worker component | DataNode | | YARN master component | NodeManager | | YARN worker component | NameNode | | [2] | |
|  | (c) | List key advantages of Hadoop. | [2] | |
|  | (d) | Mention the learning by CAP theorem with respect to Distributed Systems. | [2] | |
|  | (e) | Compare and contrast SQL, NoSQL and NewSQL. (any two topics) | [2] | |
| **Q.2** | **Attempt the following questions.** | | **[10]** | |
|  | (a) | Explain any four CLI operations to support usage of Hadoop Distributed File System (HDFS). Show how we can do something similar via GUI. | [5] | |
|  | (b) | Describe various ways to categorize types of data with examples. | [5] | |
| **Q.3** | **Attempt the following questions:** | | **[10]** | |
|  | (a) | Relate mongoDB terminology for handling various information i.e. table, row, column, etc.  Describe briefly mentioned terms with respect to MongoDB: Sharding, Upsert, \_id. | [5] | |
|  | (b) | Provide steps to setup hadoop cluster. | [5] | |
|  | | | | |
| **SECTION – II** | | | | |
| **Q.4** | **Do as directed.** | | **[10]** | |
|  | (a) | Fill in the blanks:  (I) \_\_\_\_\_\_\_\_ algorithm helps perform market basket analysis with association rules.  (II) An example of operation in Data Warehouse for Data mining is \_\_\_\_\_\_\_\_\_. | [2] | |
|  | (b) | State true or false.  (I) Presence of outliers does not let analysis come to a good decision.  (II) Decision trees are pruned to address the problem of overfitting/anomalies in the  training due to noise or outliers. | [2] | |
|  | (c) | What is the purpose of the elbow method in clustering? | [2] | |
|  | (d) | List categories of NoSQL databases. | [2] | |
|  | (e) | Analyze box plot in relevance to five number textual/numeric summary. | [2] | |
| **Q.5** | **Attempt the following questions.** | | **[10]** | |
|  | (a) | What is Big Data Analytics? Describe Big Data characteristics using famous Vs. | [5] | |
|  | (b) | Knowing various types of attributes, support how to handle peculiarities of data in the  process of knowledge discovery from data (KDD). | [5] | |
| **Q.6** | **Attempt the following questions.** | | **[10]** | |
|  | (a) | Explain working of map-reduce using word count diagram or program example. | [5] | |
|  | (b) | Draw Hadoop EcoSystem components diagram with appropriate labels. | [5] | |