DHARMSINH DESAI UNIVERSITY, NADIAD

FACULTY OF TECHNOLOGY

M. TECH. SEMESTER II.

COMPUTER ENGINEERING

SUBJECT: BIG DATA ANALYTICS (MT219)

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

| **INSTRUCTIONS:** | | | | |
| --- | --- | --- | --- | --- |
| 1. | Figures to the right indicate maximum marks for that question. | | | |
| 2. | The symbols used carry their usual meanings. | | | |
| 3. | Assume suitable data, if required & mention them clearly. | | | |
| 4. | Draw neat sketches wherever necessary. | | | |
| **SECTION – I** | | | | |
| **Q.1** | | Answer the following: | | **[10]** |
| (a) | Fill in the blanks:  (I) Basic clustering methods include \_\_\_\_ , \_\_\_\_\_ , \_\_\_\_\_\_ and \_\_\_\_\_. | [2] |
|  | (II) An observation that lies an abnormal distance from other values in a random  sample from a population is known as \_\_\_\_\_\_\_\_\_, which can be easily  identified visually by \_\_\_\_\_\_\_\_ plot. |  |
| (b) | State true or false:  (I) “hadoop fs -put” command copies files from local/native/host file system to  destination hadoop distributed file system/guest specified as args. | [2] |
|  | (II) “hadoop fs -rm” command can delete files from location within hadoop  distributed file system specified as args. |  |
|  | | (c) | Match the following to give it a correct meaning:   | **Column A** | **Column B** | | --- | --- | | 1. Pig | i. Support complex data types like tuple, bag and map. | | 2. RDBMS | ii. A blend of big data and data from legacy systems to yield  insights and offerings with speed and impact. | | 3. Fuzzy approach | iii. Overcomes the disadvantage of rule-based systems that  they involve sharp cut-offs for continuous attributes. | | 4. Analytics 3.0 | iv. Not suitable for storing and processing large files, images  and videos. | | [2] |
|  | | (d) | Explain the meaning/use of following keyword/term:  (I) Horizontal Scalability  (II) Elbow method | [2] |
|  | | (e) | What is the drawback of the K-Means clustering algorithm? Explain in brief. | [2] |
| **Q-2** | | Answer the following: | | **[10]** |
| (a) | Explain basic algorithm/example for inducing a decision tree from training tuples. | [5] |
| (b) | Describe various types of data. Describe characteristics of Big Data using V’s. | [5] |
|  | |  |  |  |
| **Q-3** | | Answer the following: | | **[10]** |
| (a) | Provide mongoDB commands to achieve below requirements:   1. Display all databases, start using a database named ‘ecommerce’, create a collection named ‘survey’, display all collections created thus far in the chosen database. 2. Create documents in ‘survey’ having field and value as mentioned & verify: doc#1 -> ‘name’:’James’ ‘age’:27 ‘education’:’BSc’ ‘scholarship’:’yes’   doc#2 -> ‘name’:’Harry’ ‘age’:25 ‘salary’: 25000  doc#3 -> ‘name’:’William’ ‘age’:25 ‘salary’: 15000  doc#4 -> ‘name’:’Michel’ ‘age’:20 ‘gender’:’female’ hobbies:’hockey’, ‘painting’, ‘singing’   1. Display all the documents having salary of ‘age’ <= 25 2. Display sorted all the documents in ascending order of age followed by descending order of salary nicely. 3. Display average salary as AvgSalary group by gender. | [5] |
| (b) | Describe and compare SQL-NoSQL-NewSQL. | [5] |

| **SECTION – II** | | | | |
| --- | --- | --- | --- | --- |
| **Q.4** | | Answer the following: | | **[10]** |
| (a) | Fill in the blanks:  (I) Data science involves multiple disciplines: \_\_\_\_ , \_\_\_\_\_ and \_\_\_\_. | [2] |
|  | (II) Data consistency model BASE means \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. |  |
| (b) | State true or false:  (I) According to Gartner, “Dark Data is the information assets organizations  collect, process and store during regular business activities, but generally fail to  use for other purposes.” | [2] |
|  | (II) Isolated data sets in Data Silos can lead to perfect decision-making. |  |
|  | | (c) | Match the following to give it a correct meaning:   | **Column A** | **Column B** | | --- | --- | | 1. Attribute Oriented Induction | i. Support for windowing functions. | | 2. Data Discretization | ii. Find frequent itemsets and perform association  rule mining | | 3. Stream Analytics | iii. Generalization based on the examination of  the number of each attribute’s distinct values  and more. | | 4. Apriori Algorithm | iv. Transforms numeric data by mapping values to  interval or concept labels. | | [2] |
|  | | (d) | Explain the meaning/use of following keyword/term:  (I) Cognitive Analytics  (II) Agglomerative hierarchical clustering | [2] |
|  | | (e) | Describe Brewer's CAP theorem for a distributed computer system. | [2] |
| **Q-5** | | Answer the following: | | **[10]** |
| (a) | Explain typical OLAP operations (Data Warehouse and Data Mining) with suitable example/s. | [5] |
| (b) | Describe steps involved in the knowledge discovery process (KDD-Knowledge Discovery in Databases). | [5] |
|  | |  |  |  |
| **Q-6** | | Answer the following: | | **[10]** |
| (a) | Describe Hadoop Map-reduce with Word Count example/program. | [5] |
| (b) | Draw Hadoop cluster high-level architecture describing storage and processing master-slave components. Provide high-level steps to setup hadoop cluster. | [5] |