Register No.: ……………………………… Name: …………………………………………………………….

SAINTGITS COLLEGE OF ENGINEERING (AUTONOMOUS)

(AFFILIATED TO APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY, THIRUVANANTHAPURAM)

SECOND SEMESTER M.TECH DEGREE EXAMINATION (Regular), JULY 2022

COMPUTER SCIENCE AND SYSTEMS ENGINEERING

(2021 Scheme)

Course Code: **21SE206-D**

Course Name: **Big Data Management and Analytics**

Max. Marks: **60** **Duration: 3 Hours**

PART A

(Answer all questions. Each question carries 3 marks)

1. How did Internet contribute to the concept of Big data?
2. What are the major goals of a distributed file system?
3. What is meant by Fair scheduling?
4. Why is it important not to enforce strict consistency in distributed storage?
5. What is meant by entropy in data analytics?
6. What is the importance of avoiding class imbalance in supervised learning?
7. Describe the plot used to analyze the probability distribution of the data?
8. What are the major types of attributes in a dataset?

PART B

(Answer one full question from each module, each question carries 6 marks)

MODULE I

9. How does prescriptive analytics differ from predictive analytics? (6)

OR

10. Explain the various steps involved in analytics methodology. (6)

MODULE II

11. Explain the GFS architecture with diagram. (6)

OR

12. Explain the garbage collection mechanism in GFS. (6)

MODULE III

13. Use map reduce methodology to detect the total number of words in the pages of a digital document. Write the pseudo code for the mapper and reducer class. (6)

OR

14. Consider the case of counting respective car brands from a database of total car purchase in two years. Illustrate how shuffle and sort mechanisms can be implemented prior to giving input to reducer class (6)

MODULE IV

15. Explain the concept of column family document model with example. (6)

OR

16. Explain how particular document parts can be retrieved from data stored using document model. (6)

MODULE V

17. Explain the concept of Decision tree with an example. (6)

OR

18. With a pseudo code explain the process of determining centroids in kmeans clustering algorithm. (6)

MODULE VI

19. You have been given a student dataset with five features including name,gender,address,marks,pass/fail. Create a set of sample data and demonstrate how stem charts can be useful. (6)

OR

20. Explain the data analytics life cycle. (6)

****