

THE KAVERY ENGINEERING COLLEGE*(An Autonomous Institution)***MCA DEGREE END SEMESTER THEORY EXAMINATIONS, APRIL-MAY 2025***Third Semester**Master of Computer Applications***DS4015 - Big Data Analytics***Regulations - 2021***Duration : 3 Hrs****Maximum : 100 Marks****PART – A (ANSWER ALL QUESTIONS) (Marks 10*2=20)**

Q.No	Questions	BLT	CO	Marks
1.	Distinguish analysis Vs reporting in big data.	L4	1	2
2.	Define prediction error.	L2	1	2
3.	Analyze the convergence behavior of simulated annealing for solving an optimization problem.	L4	2	2
4.	How can you evaluate the effectiveness of a visualization technique?	EV	2	2
5.	Define the need for filtering data streams.	L2	3	2
6.	List few examples for real time analytics platform application.	L1	3	2
7.	Compare NoSQL databases over traditional relational databases?	L4	4	2
8.	Analyze the performance implications of different sharding strategies.	L4	4	2
9.	Write R code to create a vector of numbers from 1 to 10.	L3	5	2
10.	Given a CSV file containing sales data with columns for "Product," "Region," and "Sales," write an R code snippet to read this file into an R data frame, and then calculate the total sales for each region.	L3	5	2

PART – B (ANSWER ALL QUESTIONS) (Marks 5*16= 80)

Q.No	Questions	BLT	CO	Marks
11.	a i Explain the challenges of conventional system and need for big data analytics.	L2	1	8
	ii Explain the various statistical concepts applied for data analysis.	L2	1	8
	Or			
	b i Explain Intelligent data analysis and terms related to IDA.	L2	1	8
	ii Explain the various analytic process tools and methods used for data.	L2	1	8
12.	a i Describe in detail about genetic programming with examples.	L2	2	6
	ii Analyze the travelling salesperson problem as a case study and find the goal to find the shortest possible route that visits each city exactly once and returns to the starting city. Use a genetic algorithm to solve this problem by representing each possible route as a chromosome. The fitness of a chromosome is determined by its total distance. Genetic operators like crossover and mutation can be applied to explore the solution space and find the optimal solution.	L4	2	10

Or

	b	i	Analyze the impact of data visualization aid in decision-making and problem-solving in a business context. Give appropriate example to justify.	L4	2	10
		ii	Summarize the various data types and give appropriate example.	L2	2	6
13.	a	i	Explain stream data model and architecture with neat diagram.	L2	3	10
		ii	Summarize the issues in stream processing.	L2	3	6
			Or			
	b	i	Apply counting distinct elements algorithm for a stream.	L3	3	6
		ii	You are a data scientist working for an e-commerce company. The company wants to analyze customer reviews to understand customer sentiment and identify areas for improvement. Perform the following: • Data Preprocessing • Sentiment Analysis Model • Aspect-Based Sentiment Analysis • Visualization • Interpretation	L4	3	10
14.	a	i	Given a large dataset of Electronic Health Records (EHRs) to improve patient care and medical research. Design the solution for the following operations • Data Ingestion and Preparation • Data Analysis and Insights • Ethical Considerations	L3	4	8
		ii	Model a scalable architecture to handle large volumes of data for the above case study on HER.	L3	4	8
			Or			
	b	i	Take up a case study of working for a social media company. Your task is to develop strategies to protect user privacy and prevent inference attacks on sensitive information. Analyze common types of inference attacks, such as demographic inference, relationship inference, and interest inference. Explore data anonymization and pseudonymization techniques to mask sensitive information.	L4	4	10
		ii	How can you ensure data privacy and security in big data systems?	L2	4	6
15.	a	i	Explain the concept of R Data Structures in detail with examples.	L3	5	10

Or

15. b A dataset containing information about students' grades in different subjects. The dataset includes columns for student ID, name, and grades in Mathematics, Science, and English. L3 5 16

Student_ID	Name	Maths	Science	English
1	Alice	85	92	88
2	Bob	78	75	82
3	Charl	95	90	98
4	David	65	72	70
5	Emily	80	88	85
6	Frank	72	78	75
7	Grac	90	95	92
8	Henry	68	70	65
9	Ivy	83	86	81
10	Jack	75	80	77

Design the solution for the following:

- Import the dataset into R using appropriate functions
- Calculate the average, median, and standard deviation of grades for each subject
- Identify the top-performing students in each subject
- Determine the correlation between different subjects
- Create a function to calculate the grade point average (GPA) for each student
- Create a function to visualize the distribution of grades using histograms