SPRING MID-SEMESTER EXAMINATION-2023

School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University
Big Data
[CS-3032]

Time: 1 1/2 Hours **Full Mark: 20**

Answer any four Questions including Q.No.1 which is Compulsory.
The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered in one place only.

01. Answer all the questions [1 x 5]

a) Differentiate between data analysis and data modeling?

b) Describe any two vital characteristics of semi-structured data?

c) Bring out any two key differences between batch processing and real-time processing in big data analytics?

d) According to the CAP theorem, what is the maximum number of attributes that a distributed system can satisfy simultaneously?

e) What is the purpose of sample size calculation?

02. a) Explain what are the benefits of using machine-readable data, and how does it improve data processing? [2.5Marks]
b) Illustrate any five real-world applications of big data which will help society? [2.5Marks]

03. Draw the schematic diagram of the big data architecture and explain each component in brief. What are the emerging trends and technologies in big data architecture design? [5Marks]

04. a) Demonstrate the design of the Bloom filter Suppose we have a set of five numbers: {3, 7, 11, 15, 19}. We want to encode this set into a Bloom filter with an array of 20 bits and three hash functions as follows: [3 Marks]

$h1(x) = x \% 20$
$h2(x) = (2x + 3) \% 20$
$h3(x) = (5x + 7) \% 20$

b) Can a Bloom filter provide false positives or false negatives? Explain How can you minimize these errors? [2Marks]

05. a) Explain the best practices for deploying Big Data Analytics solutions in a production environment? [2.5Marks]

b) Enumerate the common challenges faced during data validation, representation, cleaning, and aggregation, and how can they be addressed. [2.5Marks]

*** Best of Luck ***