SASTRA

School of Computing
First CIA Exam – Aug 2024
Course Code: **INT404R01**
Course Name: **BIG DATA ANALYTICS**
Duration: **90** minutes Max Marks: **50**

Answer all FIVE questions. **5 X 10 = 50 Marks**

1. Describe the characteristics of big data and explain the layered architecture of big data technology stack and the interaction among the key components with a neat diagram.

2. Given the task of calculating bigram frequencies using the MapReduce algorithm, how would you design the Mapper function to process each document and emit bigrams? After the Mapper function emits bigrams and their counts, what role does the Reducer function play in the MapReduce algorithm? Suppose you have the following sample documents:
   a. Document 1: "Machine learning is a field of artificial intelligence."
   b. Document 2: "Artificial intelligence and machine learning are interrelated."
   c. Document 3: "Deep learning is a subset of machine learning."

   What would be the output of the MapReduce job for these documents?

3. For the given set of m samples in S, where m = 10,
   S = {152, 173, 144, 182, 11, 22, 46, 85, 92, 210}
   with two hash functions, and a bit array of size 17 - bits
   i) $h(x) = (2x + 10) \% 17$ ii) $h(x) = 4x \% 17$

In the bit series, count the number of ones and find the error rate of false positives that the bloom filter produces.

4. Explain the following:
   i) Model of Data Stream management system 3 M
   ii) Execution of Relational algebra operations (Join and group) using map reduce algorithm. 7 M

5. The figure below shows the network connectivity between the systems. Here, the five separate systems are represented by nodes and are connected by a solid line. Construct the V matrix from the node value (i.e., A node value is 12) and the M matrix from the link value (i.e., A to B is 32 and B to A is 32). Perform a MapReduce based Matrix vector multiplication to determine the significance of each system and write an algorithm for the same.