# Nirma University

## Institute of Technology

Semester End Examination (IR), December - 2022

B. Tech. in Computer Science and Engineering, Semester-VII

2CS702 Big Data Analytics

Roll / Exam No. | Supervisor's Initial with Date

Time: 3 Hours | Max Marks: 100

Instructions:
1. Attempt all questions
2. Figures to right indicate full marks
3. Assume necessary data.
4. Use section-wise separate answer book.
5. Draw neat sketches wherever necessary.

**SECTION – I**

**Q-1. Do as directed.** **[16]**

(A) CO2 BL2 Explain the terms, (8)
1) HDFS block
2) Input split
3) RDD
4) Record reader

(B) CO2 BL3 Explain the sort and shuffle phase in the MapReduce programming model. Discuss the problems that may be occurred. How to optimize the function of the sort and shuffle process? (8)

**Q-2. Do as directed.** **[18]**

(A) CO2 BL3 How CAP theorem is supported in NoSQL databases? List out the pros, cons and database names of NoSQL databases. (6)

(B) CO2 BL4 Explain the types of operations performed on RDD in Spark architecture. (6)

**OR**

(B) CO2BL4 Describe any three types of input format in map-reduce programming with its appropriate application. (6)

(C) CO2 BL2 Explain the anatomy of file write operation in HDFS architecture. (6)

**OR**

(C) CO2 BL2 Describe the major limitations of HDFS architecture. (6)

**Q-3. Do as directed.** **[16]**

(A) CO1BL3 Briefly discuss scalable platforms with their features and drawbacks. (8)

(B) CO3BL3 What are the major drawbacks of Hadoop v1.0? Describe the tool used to overcome the listed drawbacks. (8)

## SECTION – II

**Q-4. Do as directed.** **[18]**

(A) CO3BL5 Consider the following dataset, (8)
***Employee (Emp_id, Emp_name, Qualification, Experience).***
Write a MapReduce pseudo code to filter employee names who are Ph.D. with more than 10 years of experience.

(B) CO3 BL4 Define Fault Tolerance. At what extent fault tolerance is supported in horizontal scaling platforms and vertical scaling platforms? Also justify your answer. (5)

(C) CO1BL4 Explain the types of data analytics with suitable case studies. (5)

**Q-5. Do as directed.** **[16]**

(A) CO1BL3 Discuss any one application in detail which present the limitation of MapReduce programming. How Apache Spark overcome this limitation? Explain the features and advantages of using Spark. (6)

(B) CO2BL2 Differentiate between MapReduce, Hive, and PIG architectures. (5)

**OR**

(B) CO2BL2 Differentiate between MapReduce, Cassandra, and RDBMS. (5)

(C) CO2BL5 What is the default size of the HDFS block? How to configure HDFS block size? If the input file size is of 1TB and the block size is 0.5 GB, then find the number of mapper tasks. Assume MapReduce input split size is equal to HDFS block size. (5)

**OR**

(C) CO2BL5 What is the difference between MapReduce Input Split and HDFS block? If the number of mappers tasks = 100, then how many MapReduce Input Splits are there? (5)

**Q-6. Do as directed.** **[16]**

(A) CO4BL 4 Consider a blog database in MongoDB. Create collection posts in the blog database. Assume that each document can be described by title, description, category, likes, tags, and date attributes. Assume data as per the requirement. Write a query for the followings,

1) Print all posts posted in the month of November 2022.
2) Update the category of the posts with 'Recent' which were posted one day before the current date.
3) Delete all posts which have no tags.
4) Sort all the posts based on likes.
5) Add one post under the category News.
6) Filter all posts which have likes more than 15.
7) Display all posts of the collection posts.
8) Filter all posts which contain pandemic words in the description.