Nirma University

Institute of Technology

Semester End Examination (IR), December - 2017

B. Tech. in Computer Engineering / Information Technology, Semester-VII

IT7C4 Big Data Analytics

Roll / Exam No [ ] Supervisor's Initials with Date [ ]

Time: 3 Hours Max Marks: 100

Instructions:
1. Attempt all the questions.
2. Figures to right indicate full marks.
3. Draw neat sketches wherever necessary.

Q.1 Do as directed **[18]**

A. What can be the real life applications of clustering in the Big data like Banking Data, Mega Store transaction data? What inferences can be made using clustering algorithms? [5]

B. Explain the Architecture of HDFS. Explain diagrammatically the read and write request sequence from HDFS. [5]

C. What are the challenges that Prevent Business from capitalizing on Big Data? [5]

D. Define Big data with all its characteristics. [3]

Q.2 Do as directed **[16]**

A. What are the fault tolerance mechanism to handle the failure instances in Hadoop? Write each scenario assuming one component failing and also the fault tolerance mechanism to handle the same in Hadoop. [6]

B. Compare Mongo DB to Cassandra. What are the difference in terminology, structure and the way data is organized. Identify one application for each which suits the data model of Mongo DB or Cassandra. [6]

C. What are the various forms of data how they can be classified based on the source, format, and Model of Data, [4]

OR

C. Compare No SQL to HDFS and discuss the characteristic of each. [4]

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

A. We have data in Mongo DB which is the result database of the University. The Field are like [8]
Name: Aadi , Roll Number: 1234 , Programme Name: BCE ,Semester: VII , Course Name: BDA , Marks: 87 , Date: (Day :24, Month:12, Year:2017) .
There will be multiple rows for multiple courses for the same students. Date signifies the date of the result
Write NoSQL Query in Mongo DB to find the following
1. Percentage of the candidate in each semester.
2. Highest Marks obtained in the given Year by any student in any course.

B. Below are the final exam scores of twenty introductory statistics [8]

students. 57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94 Create a box plot of the distribution of these scores. Label all the values required.

OR

B. For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. For comparison sketch dot plots of the distributions. [8]

(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20

b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30

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

A. Definitions:
M is a matrix with element m in row i and column j.
N is a matrix with element n in row j and column k.
P is a matrix = MN with element p in row i and column k, where P=M*N. The Matrix being sparse in nature is saved as relational Representation as below.
M with tuples (i, j, m)
N with tuples (j, k, n)

(i) Write the Map-reduce pseudo code for doing matrix multiplication. Clearly specify the key-Value pair. [6]

(ii) Can this be done without Map Reduce? Explain your answer [2]

OR

A. N dimensional numerical values are written in a row in the Text file.

(i) Write Map- Reduce pseudo code for implementing K Means clustering, clearly specify the key-Value pair. [6]

(ii) What are the challenges in doing K Means clustering using Map Reduce? [2]

B. Match the following [4]

Column-1(NoSQL type)	Column – 2(Example)
Key-Value Store	Neo4J
Graph database	CouchDB
Document oriented database	Cassandra
Column oriented database	Dynamo

C. Answer the following questions: [6]

(a) Which command should we use in order to change file permission to read write execute to a file on HDFS?

(b) If I want to see only first 3 documents of my collection. What would be the query in MongoDB?

(c) I want to change default replication factor. Which configuration file will help me to do so?

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

A. You are at city shopping mall. You see few people are browsing the items. Some of them are looking for discounts. Some of them are filling feedback form. Few people are at billing counter. You may consider other things and events happening in this scenario. Think for while on the different types of data generated. Categorize each data source into appropriat category, by considering the Variety and velocity of each source. [4]

B. Differentiate between MR1(MapReduce1) and MR2(MapReduce2). [4]

C. Explain the concept of Sharding and replication in NoSQL. Also specify the steps to achieve the same in MongoDB. (Syntax of the commands are not mandatory but action has to be clear) [8]

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

A Discuss at-least two real time applications of HIVE in detail with features and support of HIVE. [4]

OR

Discuss at least three use cases where Pig is appropriately used. Also mention the specific reason of using Pig in those areas. [4]

B Considering scenario of Movie review, design a graph database defining its nodes and relationships. The design should facilitate the prevalent types of searches required. Ex. actor , genre, production house, etc. [8]

C Compare the Hadoop map-reduce and Apache Spark in at least 12 parameters [4]