Nirma University

Institute of Technology

Semester End Examination (IR), December - 2016

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

IT7C4 Big Data Analytics

Roll / Exam No. + | Supervisor's Initial With Date

Time: 3 Hours | Total Marks: 100

SECTION-I

Q:1 Answer the following questions: (6 X 3) **[18]**

1 Why do you need scaling of data? Discuss any two platforms available for vertical scaling of big data.

2 Imagine a node n on rack r in data center d. Calculate the distance of following scenario
Distance(/d1/r1/n1 , /d2/r3/n4)
Distance((/d1/r1/n1 , /d2/r2/n3)

3 Deliberate the concept of CAP theorem.

4 Suppose there is a repository of ten thousand documents, andword w appears in 320 of them. In a particular document d, the maximumnumber of occurrences of a word is 15. Approximately what is the TF.IDFscore for w if that word appears (a) once (b) five times?

5 Evaluate the S-curve $1-(1-S^r)^b$for S = 0.1, 0.2, . . . , 0.9, forthe following values of r and b:

- r = 3 and b = 10.
- r = 6 and b = 20.
- r = 5 and b = 50.

6 Find the edit distances (using only insertions and deletions)between the following pairs of strings.
(a) abcdef and bdaefc.
(b) abccdabc and acbdcab.

Q:2 Answer the following questions: (4 X 4) **[16]**

1 Consider the following matrix with six rows.

Element	S1	S2	S3	S4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

A) Compute the minhash signature for each column if we use the

following three hash functions:
h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x) = 5x + 2 mod

2 Do as directed

I. Find the cosine distance for vectors X=[1,2,-1] and Y=[2,1,1]
II. Find the Manhattan distance between points (2,7) and (6,4).

3 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 a while on the different types of data generated. Mention each of them with proper logic.

4 Illustrate "Bonferroni's principle is used to show the statistical limit on data mining."

Q:3 Use MongoDB to develop a Library management system. Features of Library management systems are **[16]**

- Keep record of different categories like; Books, Journals, Newspapers, Magazines, etc.
- Classify the books subject wise.
- Easy way to enter new books.
- Keep record of complete information of a book like; Book name, Author name, Publisher's name, Date/ Year of publication, Cost of the book, Book purchasing date/ Bill no.
- Automatic fine calculation for late returns.
- Different criteria for searching a book.
- Different kind of reports like; total no. of books, no. of issued books, no. of journals, etc.
- Easy way to know how many books are issued to a particular student.
- Easy way to know the status of a book.
- Event calendar for librarian to remember their dates.
- Online access for registered user to see the status of their books.
- Completely cloud based Library Management System.

Create all the fields of your collections and your database according to above situation. For each feature there should be separate command/query to insert/Delete/Update/Search/sort the dataset.

OR

Q:3 Discuss the concept of sentiment analysis. Why analysis need the Hadoop framework in order to process the big data? Develop K-means clustering workflow using Hadoop map-reduce framework for sentiment analysis. Also throw some light on technical and algorithmic parameter in order to get good accuracy of same algorithm. **[16]**

SECTION –II

Q:4 Answer the following questions: (6 X 3) **[18]**

1 Explain the concept of Sharding in NoSQL.

2 Discuss the area under which pig is not applicable? Mention appropriate reason too.

3 Compare MapReduce with RDBMS.

4 What is data locality optimization? How it is achieved in HDFS?

5 Differentiate between MR1(MapReduce1) and MR2(MapReduce2).

6 Hadoop Distributed File System is not good fit for which application.

Q:5 Answer the following questions: (4 X 4) **[16]**

1 Discuss the detail scenario of reading file from HDFS with diagram.

2 What is serialization? How does it achieved in Hadoop framework?

3 Give the merits and demerits of Hive.

4 How exactly streaming (JAVA) and pipes (C++) execute the map and reduce task with detailed diagram.

Q:6 Answer the following questions: (8 X 2) **[16]**

1 Deliberate the detailed flowchart to show how Hadoop runs MapReduce job using YARN. **[8]**

2 If you are given the data of year and its temperature. How would you specify the combiner function with Mapper and Reducer to determine maximum global temperature of the year? Discuss the importance of combiner function. **[8]**

OR

2 Discuss at least three use cases where Pig is appropriately used. Also mention the specific reason of using Pig in specific areas. **[8]**