Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous College Affiliated to University of Mumbai)

SYNOPTIC Set-I

April -May 2018

Max. Marks	**: 100**	**Duration: 180 Minutes**
Class	**: BE Computer**	**Semester: VIII**
Course Code	**: CPE 8035**	**Branch : Computer Engineering**
Name of the Course:	**Big Data Analytics**	

Instructions:

(3) All Questions are Compulsory

(4) Draw neat diagrams

Q1 A) Explain 4 ways by which big data problems are handled by NoSQL

A client needs a database design for his blog. Website has following requirements.

a)Every post can have one or more tag.

b)Every post has unique title, description and url

c)Every post has name of its publisher and total no. of likes.

d)On each post, there can be 0 or more comments.

Design the normalized schema(tables and columns) for RDBMS and the MongoDB schema.

Synoptic 1A)

4 M - 4ways

3 M - RDBMS schema

3 M - MongoDB schema

4 ways

1. Moving queries to the data, not data to the queries

Use commodity processors that each hold a subset of the data on their local shared-nothing drives. When a client wants to send a general query to all nodes that hold data, it's more efficient to send the query to each node than it is to transfer large datasets to a central processor. Keeping all the data within each data node in the form of logical documents and moves query itself and the final result over a network. This keeps your big data queries fast.

2. Using hash rings to evenly distribute data on a cluster:

One of the most challenging problems with distributed databases is figuring out a consistent way of assigning a document to a processing node. Using a hash ring technique to evenly distribute big data loads over many servers with a randomly generated 40-character key is a good way to evenly distribute a network load.

Hash rings take the leading bits of a document's hash value and use this to determine which node the document should be assigned. This allows any node in a cluster to know what node the data lives on and how to adapt to new assignment methods as your data grows.

3. Using replication to scale reads:

Databases use replication to make backup copies of data in real time. Using replication allows you to horizontally scale read requests.

If a client does a write and then an immediate read from that same node, there's no problem. The problem occurs if a read occurs from a replica node before the update happens. This is an example of an inconsistent read. The best way to avoid this type of problem is to only allow reads to the same write node after a write has been done. This logic can be added to a session or state management system at the application layer. Almost all distributed databases relax database consistency rules when a large number of nodes permit writes. If your application needs fast read/write consistency, you must deal with it at the application layer.

4. Letting the database distribute queries evenly to data nodes:

In order to get high performance from queries that span multiple nodes, it's important to separate the concerns of query evaluation from query execution. NoSQL systems move the query to a data node, but don't move data to a query node. In this example, all incoming queries arrive at query analyzer nodes. These nodes then forward the queries to each data node. If they have matches, the documents are returned to the query node. The query won't return until all data nodes (or a response from a replica) have responded to the original query request. If the data node is down, a query can be redirected to a replica of the data node.

Database schema – RDBMS

Comment(comment_id,post_id,by_user,date_time,likes,messages)
post(id,title,description,like,url,post_by)
tag_list(id,post_id,tag)

Database schema – MongoDB

```
{ id:POSTID
title : TITLE_OF_POST,
description : POST_DESCRIPTION
by: POST_BY,
usl: URL_OF_POST,
tags:[TAG1,TAG2,TAG3],
likes: TOTAL_LIKES,
comments: [
        {
user: COMMENT_BY
message: TEXT
dateCreated : DATE_TIME
like: LIKES
        }
        {
user: COMMENT_BY
message: TEXT
dateCreated : DATE_TIME
like: LIKES
        }
}
```

Q1 B) How does clickstream analytics using Hadoop help sales and marketing team to prepare their campaigning strategies?

Synoptic 1B)
2 M - Clickstream data types
3M – How interactive and customized experience is given using Hadoop

Clickstream data is generated when customers interact with website. This data include pages they load, time spent by them on each page, links they clicked, frequency of visit, from which page do they exit etc. Website content, its organization, navigation and transaction completion can be improved, This helps to understand customer better and improve website content, its organization, navigation and transaction completion.This ensures that customers can easily find what they want. For analysis and prediction, more than 3 years data is collected to find correlation. Hadoop can load structured and unstructured data efficiently. This helps in performing repetitive queries. With big data analytics, interactive and cutomized experience is given to the customers.

Q2) How does Pagerank differ from HITS algorithm? Give 2 differences.
Calculate the authority and hub scores for the given matrix that represent a graph of four vertices (n1 to n4) , using HITS algorithm with k = 2.
Identify the best authority and hub nodes.

$$\begin{vmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

Synoptic 2)

2 M – 2 difference
6 M – calculation 2 iterations
2 M – Best hub and authority

OR

Q2) What are the categories the spammer divides the web pages into. Explain? Mathematically prove the improvement in pagerank of a target page using spam farm.

Synoptic 2)
3 marks - name 3 categories (Inaccessable pages, Accessable pages, Own pages)
3 marks - explain 3 categories
4 marks – Mathematical proof

Q3 A) Write a Map Reduce program to find out Designation Wise Customer count like How many Doctors are there in the file or How many Lawyers are in the file.

Consider following scenario for taking input data :

Let there be a file : custs

where col1 - cust_id
col2 - first_name
col3 - last_name
col4 - age
col5 - designation

Synoptic :

A] Mapper code Total 5 marks

[3M] for correct logic of mapper + [2M] for correct defination of functions with correct data types

B] Reducer Code Total 5 Marks

[2M] for correct logic of reducer + [2M] for correct defination of functions with correct data types and [1] clearly indicating output

OR

Q3 A) Write a Map Reduce program to find the total number of words from the file as given below :

Let the input data is:

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow.
final output will be = total words, 17

Synoptic :

A] Mapper code Total 5 marks = 3M for correct logic of mapper + 2M for correct definition of functions with correct data types

B] Reducer Code Total 5 Marks = 2 M for correct logic of reducer + 2 M for correct definition of functions with correct data types and 1M clearly indicating output

Q3 B) How to compute RDBMS Selection Operation by Map Reduce. Clearly explain Map and Reduce functions and Give its pseudo code.

Synoptic :

Correct explaination of each function 2 M each , correct pseudocode 6 M = 2+2+6

Q4) Consider the following two texts, Find the similarity between 1 and 2.

1. Julie loves me more than Linda loves me
2. Jane likes me more than Julie loves me

Synoptic : Now we count the number of times each of these words appears in each text:

me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

[2M]

the two vectors are, again:

a: [2, 0, 1, 1, 0, 2, 1, 1]

b: [2, 1, 1, 0, 1, 1, 1, 1] [2M]

The cosine of the angle between them is about 0.822. [1M]

Q5A) What is Data Stream Management System? Explain with block diagram.

CO-4 [10M]

Synoptic :

1. Explanation about – how DSMS supports on-line analysis of stream data i.e real time, continuous sequence of items and handle continuous queries = 2 Marks
2. Comparison between DBMS and DSMS = 2 Marks
3. Generic DSMS Architecture Diagram 2 Marks
4. Description of 4 components – Input Monitor, Storage Area, Query Repository, Query Processor = 4 Marks

Q5B) Explain with diagram how the PCY algorithm helps to perform frequent itemset mining for large datasets CO-4 [10M]

Synoptic :

1. To overcome drawback of A-priori algorithm = 1 Mark

2. Diagram of Memory Structure with two passes of PCY = 1 Mark
3. PCY Algorithm – Pass-I = 2 Marks
4. PCY Algorithm – Between passes = 2 Marks
5. PCY Algorithm – Pass-II = 2 Marks
6. Example = 2 Marks

OR

Q5B) Explain the concept of CURE Clustering Algorithm with example.

CO-4 [10M]

Synoptic :

1. Overview of CURE Algorithm = 2 Marks
2. CURE Initialization Step = 2 Marks
3. CURE Completion Phase = 2 Marks
4. Working of CURE algorithm with example = 4 Marks

Q6A) Define Count Distinct Problem with constraints.

Suppose data stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Let the hash function being used is h(x) = 3x +1 mod 5; Show how Flajolet-Martin algorithm will estimate the number of distinct elements in this stream. **CO-4** **[10M]**

Synoptic :

1. Deinition of Count Distinct Problem with constraints = 2 Marks
2. Calculation of Hash Function for complete stream = 4 Marks
3. Count trailing zeroes r(a) for each hash function bit = 2 Marks
4. Record R i.e. maximum r(a) = 2 = 1 Mark
5. Estimate 2^R = 4 = 1 Mark

OR

Q6A) Define following terms with respect to Association Rule Mining -

A) Support of Item B) Confidence of Rule

Consider a small database with four items I = {Bread, Butter, Eggs, Milk} and four transactions as

shown in following table. Suppose, that the minimum support and minimum confidence of an association rule are 40% and 60% respectively. Find out several potential association rules.

CO-4 [10M]

Transaction ID	Items
T1	{Bread, Butter, Eggs}
T2	{Butter, Eggs, Milk}
T3	{Butter}
T4	{Bread, Butter}

Synoptic :

1. Definition of Support and Confidence = 2 Marks
2. Determination of itemset with support values = 4 Marks
3. Determination of Association rule with confidence value = 4 Marks

Q6B) What is a community in "Social Network Graph"? What are the standard Clustering Techniques for Graph Clustering? Explain one algorithm for finding communities in a Social Graph.

CO-5 **[10M]**

Synoptic :

1. Concept of Community in Social Network Graph = 2 Marks
2. Clustering Techniques – 1) Binary Distance Measure 2) Betweenness Measure = 2 Marks
3. Grivan-Newman Algorithm = 6 Marks

Q7) By computing Pearson Coefficient users determine the rating of User-ID-3 for Item-ID-1 and

Item-ID-6 using following table. **CO-5** **[10M]**

Item ID →	1	2	3	4	5	6
User ID↓						
1	7	6	7	4	5	6
2	6	7	?	4	3	4
3	?	3	3	1	1	?
4	1	2	2	3	3	4
5	1	?	1	2	3	3

Synoptic :

1. Determination Mean Rating of all users = 2.5 Marks
2. Determination Pearson Coefficient for every user = 2.5 Marks
3. Computation of rating of User-3 for Item 1 = 2.5 Marks
4. Computation of rating of User-3 for Item 6 = 2.5 Marks