ADVANCED DATABASE TECHNIQUES

Some Viva Questions

1. what is parallel database system
A parallel database system is one that seeks to improve performance through parallel implementation of various operations such as loading data, building indexes, and evaluating queries

2. What is distributed database system
In a distributed database system, data is physically stored across several sites, and each site is typically managed by a DBMS that is capable of running independently of the other sites.

3. What are the architecture of parallel database

1. Shared - memory system, where multiple CPUs are attached to an interconnection network and can access a common region of main memory.

2. Shared - disk system, where each CPU has a private memory and direct access to all disks through an interconnection network.

3. Shared - nothing system, where each CPU has local main memory and disk space, but no two CPUs can access the same storage area; all communication between CPUs is through a network connection.

4. what is Data Partitioning
Here large datasets are partitioned horizontally across several disk, this enables us to exploit the I/O bandwidth of the disks by reading and writing them in parallel. This can be done in the following ways:

a. Round Robin Partitioning : If there are *n* processors, the ith tuple is assigned to processor *i mod n*

b. Hash Partitioning : A hash function is applied to (selected fields of) a tuple to determine its processor.

Hash partitioning has the additional virtue that it keeps data evenly distributed even if the data grows and shrinks over time.
c. Range Partitioning : Tuples are sorted (conceptually), and *n* ranges are chosen for the sort key values so that each range contains roughly the same number of tuples; tuples in range *i* are assigned to processor *i*.

Range partitioning can lead to data skew**;** that is, partitions with widely varying numbers of tuples across partitions or disks. Skew causes

Mission MCA

processors dealing with large partitions to become performance bottlenecks.

5. What is Sorting

Sorting could be done by redistributing all tuples in the relation using range partitioning.

Ex. Sorting a collection of employee tuples by salary whose values are in a certain range.

6. Distributed Data Independence: -
The user should be able to access the database without having the need to know the location of the data.

7. **Distributed Transaction Atomicity: -** The concept of atomicity should be distributed for the operation taking place at the distributed sites.

8. **Architecture of DDBs**

a)Client-Server:
A Client-Server system has one or more client processes and one or more server processes, and a client process can send a query to any one server process. Clients are responsible for user-interface issues, and servers manage data and execute transactions.

b) **Collaborating Server:**
In Collaborating Server **system,** we can have collection of database servers, each capable of running transactions against local data, which cooperatively execute transactions spanning multiple servers.

c) Middleware
Middleware system is as special server, a layer of software that coordinates the execution of queries and transactions across one or more independent database servers.

The Middleware architecture is designed to allow a single query to span multiple servers, without requiring all database servers to be capable of managing such multi site execution strategies. It is especially attractive when trying to integrate several legacy systems, whose basic capabilities cannot be extended.

10. Fragmentation

It is the process in which a relation is broken into smaller relations called fragments and possibly stored at different sites.

Mission MCA

It is of 2 types

1. **Horizontal Fragmentation** where the original relation is broken into a number of fragments, where each fragment is a subset of rows. The union of the horizontal fragments should reproduce the original relation. **2. Vertical Fragmentation** where the original relation is broken into a number of fragments, where each fragment consists of a subset of columns. The system often assigns a unique tuple id to each tuple in the original relation so that the fragments when joined again should from a lossless join. The collection of all vertical fragments should reproduce the original relation.

2 Replication:
 Replication occurs when we store more than one copy of a relation or its fragment at multiple sites.

3. Distributed catalog management :

Naming Object
Its related to the unique identification of each fragment that has been either partitioned or replicated.

 This can be done by using a global name server that can assign globally unique names.

11. Distributed Data Independence

It means that the user should be able to query the database without needing to specify the location of the fragments or replicas of a relation which has to be done by the DBMS

 Users can be enabled to access relations without considering how the relations are distributed as follows:
The *local name* of a relation in the system catalog is a combination of a *user name* and a user-defined *relation name*.

12. Distributed query processing

In a distributed system several factors complicates the query processing.

One of the factors is cost of transferring the data over network.

This data includes the intermediate files that are transferred to other sites for further processing or the final result files that may have to be transferred to the site where the query result is needed.

Mission MCA

13. semijoins and bloomjoins

Semi join and Bloom Join are methods of joining which are used in query processing in case of distributed database. In case of distributed databases the data has to be transferred between the databases for processing queries. These databases are usually located at different sites. To save on the cost of operation the queries are optimized so that minimum amount of data may need to be transferred. This is where these two methods come into picture.

semi join reduces the amount of data transferred between these sites. Only the join column is transferred between the sites.
In case of bloom join representation of join column is transferred between the remote sites instead of the join column like in semi join. This representation is created by using bloom filter with bit vector for executing membership queries. Bloom join is more efficient than semi join because the amount of data transferred is far less in case of bloom join

14. Two Phase Commit Protocol

When the user decides to commit the transaction and the commit command is sent to the coordinator for the transaction.
In a "normal execution" of any single distributed transaction, i.e., when no failure occurs, which is typically the most frequent situation, the protocol comprises two phases:
1. The commit-request phase (or voting phase), in which a coordinator process attempts to prepare all the transaction's participating processes (named participants, cohorts, or workers) to take the necessary steps for either committing or aborting the transaction and to vote, either "Yes": commit (if the transaction participant's local portion execution has ended properly), or "No": abort (if a problem has been detected with the local portion), and

2. **The commit phase**, in which, based on voting of the cohorts, the coordinator decides whether to commit (only if all have voted "Yes") or abort the transaction (otherwise), and notifies the result to all the cohorts. The cohorts then follow with the needed actions (commit or abort) with their local transactional resources (also called recoverable resources; e.g., database data) and their respective portions in the transaction's other output (if applicable).

15. What is data warehouse?

A data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis.

Mission M

16. What is Data marts
Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization. Data marts are often derived from subsets of data in a data warehouse.

17. What is the difference between Data Mart and Data warehousing?
Firstly, Data mart represents the *programs, data, software and hardware* of a **specific department**. For example, there is separate data mart for finance, production, marketing and sales department. None of the data mart resembles with any other data mart. However, it is possible to coordinate the data of various departments. Data mart of a specific department is completely focused on individual needs, requirements and desires. Data in data mart is highly indexed but is not suitable to support huge data as it is designed for a particular department.
Data warehousing is not limited to a department of office. It represents the database of a complete corporate organization. Subject areas of data warehousing includes all corporate subject areas of corporate data model.

18. What is Extraction
Extraction is the operation of extracting data from a source system for further use in a data warehouse environment. The source systems might be very complex and poorly documented, and thus determining which data needs to be extracted can be difficult. The data has to be extracted normally not only once, but several times in a periodic manner to supply all changed data to the data warehouse and keep it up-to-date. Moreover, the source system typically cannot be modified, nor can its performance or availability be adjusted, to accommodate the needs of the data warehouse extraction process.

19. What is Transformation
The transform stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the target database:

20. What is Loading
After the data is extracted from appropriate sources and transformed in the required format, it needs to be loaded in the database. Data loading can be performed in one of the following phases : **Initial load:** Data warehouse tables are populated for the very first time in single run. **Incremental loads:** The changes that are made to the database are applied periodically to the database. **Full Refresh:** Contents of one or more tables are erased completely and they are reloaded.

Mission MCA

21. What is Metadata

Metadata in a DW contains the answers to questions about the data in DW. The metadata component serves as a directory of the contents of your DW. All the answers are kept in metadata repository.

Metadata repository is a general purpose information directory that classifies and manages Business and Technical metadata. **Types of metadata:-**

- Operational metadata
- Extraction and transformation metadata
- End-user metadata

22. What is star schema

The star schema is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of one or more fact tables and the points of the star are the dimension tables.

Fact Tables: A fact table typically has two types of columns: those that contain numeric facts (often called measurements), and those that are foreign keys to dimension tables. A fact table contains either detail-level facts or facts that have been aggregated. **Dimension Tables:** A dimension is a structure, often composed of one or more hierarchies, that categorizes data. Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values.

23 What is snowflake schema

The snowflake schema is a variation of the star schema used in a data warehouse.

The snowflake schema (sometimes callled snowflake join schema) is a more complex schema than the star schema because the tables which describe the dimensions are normalized.

24. What is OLAP

OLAP means Online Analytical Processing which is used for processing of data so as to manipulate it for analysis. OLAP is used as a vehicle for carrying out analysis in data warehouse which is best for analysis. In due course we will be able to understand need of OLAP, major functions and features of OLAP, different models of OLAP so that one can understand which one to apply for their own computing environment and various tools of OLAP.

25. What is Relational OLAP

ROLAP is an abbreviation of Relational OLAP. It is an alternative to the MOLAP (Multidimensional OLAP) technology. It is one of the forms of OLAP server. While both ROLAP and MOLAP analytic tools are designed to allow analysis of data through the use of a multidimensional data model, there is a slight difference among them. ROLAP does not require the pre-computation and storage of information. Instead, ROLAP tools access the data in a relational database and

Mission MCA

generate SQL queries to calculate information at the appropriate level when an end user requests it. With ROLAP, it is possible to create additional database tables which summarize the data at any desired combination of dimensions.

26. ROLAP vs MOLAP

One of the most important elements of OLAP environments is their dependence upon multi-dimensional data values. Data warehouses represent subject-oriented records rather than transaction-oriented ones. As such, aggregated values can be viewed as existing within a logical *cube*, where the user is free to index the cube on one or more dimensional axes. This type of
conceptual representation of the data gives rise to what is known as data cube.

> **1 MOLAP:** MOLAP is the more traditional way of OLAP analysis. Here, data is stored in a multidimensional cube. The storage is not in the relational database, but in named formats.

27. Drill-Down and Roll-Up:

The two basic hierarchical operations when displaying data at multiple levels of aggregations are the ``drill-down'' and ``roll-up'' operations. Drill-down refers to the process of viewing data at a level of increased detail, while roll-up refers to the process of viewing data with decreasing detail.

Our system provides smooth and continuous level-of-detail control in all drilling operations. The control parameter is based on a measure of cluster sizes

Selective Drill-down and Roll-up

We also coupled our drilling operations with brushing. Our system permits selective drill-down and roll-up of the brushed and non-brushed region independently. This flexibility is important as it allows the viewing of a subset of elements in varying levels of detail in relation to elements outside the subset.

28 slice and dice

slice and dice is to arrange the the data in differnt ways. it can chage as rows to columns and columns to rows.

29 Bitmap indexes

The collection of bit vectors for one column is known as bitmap index for that column. For eg: Consider a table that describes employees
Employees(empid:integer, name:string, gender: boolean, rating: integer)

The rating value is an integer in the range 1 to 10, and only two values are recorded for gender. Columns with few possible values are called sparse. We

Mission MCA

can exploit sparsity to construct a new kind of index that greatly speeds up queries to these columns.

30. What is Data mining
Data mining is often defined as finding hidden information in a db. Through Data Mining we dont get a subset of data stored in db, instead, we get analysis of contents of db.

31. Association - looking for patterns where one event is connected to another event

32. Sequence or path analysis - looking for patterns where one event leads to another event

33. Classification - looking for new patterns

34. Clustering - finding and visually documenting groups of facts not previously known

35. Forecasting - discovering patterns in data that can lead to reasonable predictions about the future (This area of data mining is known as predictive analytics.)

36. DATA MINING ALGORITHMS

1 CLUSTERING
Here we partition the available data into groups such that records in one group are similar to each other and records in different groups are dissimilar to each other. Each group is called a cluster. Similarity between records is measured by distance function. The clustering algorithm□s output consists of summarized representation of each cluster.

1) Compute the distance between record r and each of the existing cluster centers. Let I be the cluster index such that the distance between r and Ci is the smallest.(Ci- Center of cluster).

2) Compute the value of new radius R´i of the ith cluster under the assumption that r is inserted into it. If R´i<= ε, then the ith cluster remains compact and we assign r to the ith cluster by updating its center and setting its radius to R´i. If R´i > ε, then the ith cluster is no longer be compact and if we insert r into it. Therefore we start with new cluster containing only the record r.

2 CLASSIFICATION:
Classification is a data mining (machine learning) technique used to predict group membership for data instances. Unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined.

Mission MCA

37. what is KDD PROCESS

The Knowledge discovery and data mining can roughly be described in 4 steps:

1) Data selection:

The target subset of data and attributes of interest are identified by examining the entire raw dataset.

2) Data Cleaning :

Noise and outliers are removed and some new fields are created by combining existing fields

3) Data Mining :

Apply data mining algorithms to extract interesting patterns

4) Evaluation:

The patterns are presented to the end user in an understandable form, for example, through visualization.

38. What is DECISION TREES

A decision tree is a graphical representation of a collection of classification rules. The tree directs the record from the root to a leaf. Each internal node of the tree is labeled with a predictor attribute called as splitting attribute because the data is split based on condition over this attribute.

39. What is NEURAL NETWORKS

Neural Network is information processing system graph the processing system and the various algorithms that access that graph. Neural Network is a structured graph with many nodes(processing elements) and arcs(interconnections) between them.

40. What is SEARCH ENGINES

There are two categories of search operators: Boolean and **non-Boolean** operators. An operator is a word or symbol that you type in that gives the search engine directions to help it know what to search for. Using these operators can narrow or widen your search, helping you find web sites that may be useful to you. You will have to check out the individual search engines to find out which operators work.

Boolean Searching: AND means the search results must have both terms – often it is typed in UPPER CASE, but not always AND decreases the number of web sites that will be found, narrowing in on specific topics.

Mission MCA

Example: pollution AND water will look for web sites about both pollution and water **OR** means the search results can have just one of the terms OR increases the number of web sites that will be found, broadening your search.

Non-Boolean Searching: + works like AND, making the term required in the search results the + should be placed directly in front of the search term without any spaces. **Example:** pollution +water will look for web sites about pollution that also mention water.

41.What is Access Control what are it type

In an organisation, database has most important data and many users. Many of the users are able to have access to the limited part of database. A DBMS offers two main approaches to access control: a) Discretionary access control and b) Mandatory access Control.

Discretionary access control (**DAC**) is a kind of access control defined by the Trusted Computer System Evaluation Criteria as "a means of restricting access to objects based on the identity of subjects and/or groups to which they belong.

mandatory access control (**MAC**) refers to a type of access control by which the operating system constrains the ability of a *subject* or *initiator* to access or generally perform some sort of operation on an *object* or *target*.

42. What is COVERT CHANNELS

Information can flow from a higher classification level to a lower classification level via indirect means, called Covert Channels, even if DBMS is forcing the mandatory access control. For example, if a transaction accessed data at more than one site in a distributed DBMS, then the actions at two sites must be coordinated properly.

43 what is Encryption:

The basic idea behind encryption is to apply an encryption algorithm to the data, using a user-specified encryption key. The result of the algorithm is the encrypted data. There is also a decryption algorithm which takes the encrypted data and decryption key as input and returns the original data. Both the algorithms i.e. Encryption and Decryption algorithm are known to the public but the keys either both or anyone is secret. In practice, the public sector uses database encryption to protect citizen privacy and national security.

And all the questions from previous university question papers.

All the Best!!!

Mission MCA