**Module 3**

**Overview of Distributed Database System**

**Motivation:**

To extend the usability of database systems and to share the workload among different systems in a network.

**Syllabus:**

|  |  |  |  |
| --- | --- | --- | --- |
| Lecture no | Content | Duration  (Hr) | Self-Study (Hrs) |
| 1 | Features of Distributed Databases | 1 | 1 |
| 2 | Design Issues of Distributed Databases | 1 | 1 |
| 3 | Types of Distributed Databases | 1 | 2 |
| 4 | Distributed Database Architectures: Client-Server | 1 | 2 |
| 5 | Distributed Database Architectures: Peer-Peer | 1 | 2 |
| 6 | Distributed Database Architectures: Multi - DBMS Architecture | 1 | 2 |

**Learning Objectives:**

Learners shall be able to:

1. To understand the characteristics of distributed databases

2. To classify the distributed databases based on the features

3. To understand different distributed database architectures

**Theoretical Background:**

A distributed database is a database in which data is stored across different physical locations. It may be stored in multiple computers located in the same physical location (e.g. a data centre); or maybe dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely coupled sites that share no physical components.

System administrators can distribute collections of data (e.g. in a database) across multiple physical locations. A distributed database can reside on organized network servers or decentralized on the Internet, on corporate intranets or extranets, or on other organization networks. Because distributed databases store data across multiple computers, distributed databases may improve performance at end-user worksites by allowing transactions to be processed on many machines, instead of being limited to one.

Two processes ensure that the distributed databases remain up-to-date and current: replication and duplication.

1. Replication involves using specialized software that looks for changes in the distributive database. Once the changes have been identified, the replication process makes all the databases look the same. The replication process can be complex and time-consuming, depending on the size and number of the distributed databases. This process can also require much time and computer resources.
2. Duplication, on the other hand, has less complexity. It identifies one database as a master and then duplicates that database. The duplication process is normally done at a set time after hours. This is to ensure that each distributed location has the same data. In the duplication process, users may change only the master database. This ensures that local data will not be overwritten.

Both replication and duplication can keep the data current in all distributive locations.

Besides distributed database replication and fragmentation, there are many other distributed database design technologies. For example, local autonomy, synchronous, and asynchronous distributed database technologies. The implementation of these technologies can and do depend on the needs of the business and the sensitivity/confidentiality of the data stored in the database and the price the business is willing to spend on ensuring data security, consistency and integrity.

**Key Definitions:**

**Distributed Database:** A distributed database is a collection of multiple interconnected databases, which are spread physically across various locations that communicate via a computer network.

**Heterogeneous Database:** A heterogeneous database system is an automated (or semi-automated) system for the integration of heterogeneous, disparate database management systems to present a user with a single, unified query interface.

**Homogeneous Database:** In a homogeneous distributed database, all the sites use identical DBMS and operating systems.

**Multi-Database System:** A multi-database system (MDBS) is a facility that allows users access to data located in multiple autonomous database management systems (DBMSs). In such a system, global transactions are executed under the control of the MDBS. Independently, local transactions are executed under the control of the local DBMSs. Each local DBMS integrated by the MDBS may employ a different transaction management scheme.

**Course Content:**

**Lecture 1**

**Features of Distributed Databases**

A distributed database is a collection of multiple interconnected databases, which are spread physically across various locations that communicate via a computer network.

Features

* Databases in the collection are logically interrelated with each other. Often they represent a single logical database.
* Data is physically stored across multiple sites. Data in each site can be managed by a DBMS independent of the other sites.
* The processors in the sites are connected via a network. They do not have any multiprocessor configuration.
* A distributed database is not a loosely connected file system.
* A distributed database incorporates transaction processing, but it is not synonymous with a transaction processing system.

Distributed Database Management System

A distributed database management system (DDBMS) is a centralized software system that manages a distributed database in a manner as if it were all stored in a single location.

Features

* It is used to create, retrieve, update and delete distributed databases.
* It synchronizes the database periodically and provides access mechanisms by the virtue of which the distribution becomes transparent to the users.
* It ensures that the data modified at any site is universally updated.
* It is used in application areas where large volumes of data are processed and accessed by numerous users simultaneously.
* It is designed for heterogeneous database platforms.
* It maintains confidentiality and data integrity of the databases.

*Let’s check the take away from this lecture*

1) A distributed database has which of the following advantages over a centralized database?

a) Software cost

b) Software complexity

c) Slow Response

**d) Modular growth**

2) A distributed database is which of the following?

**a) A single logical database that is spread to multiple locations and is interconnected by a network**

b) A loose collection of file that is spread to multiple locations and is interconnected by a network

c) A single logical database that is limited to one location.

d) A loose collection of file that is limited to one location.

3) A distributed database can use which of the following strategies?

a) Totally centralized at one location and accessed by many sites

b) Partially or totally replicated across sites

c) Partitioned into segments at different sites

**d) All of the above**

Exercise

Q.1 Define distributed database.

Q.2 State the advantages of distributed databases over centralized databases.

Q.3 List the features of distributed databases and brief them.

**Learning from this lecture**: Learners will be able to understand the need of distributed databases and their features.

**Lecture 2**

**Design Issues of Distributed Databases**

The following are the design issues related to distributed databases

1. Distributed Database Design

• One of the main questions that is being addressed is how database and the applications that run against it should be placed across the sites.

• There are two basic alternatives to placing data: partitioned (or no-replicated) and replicated.

• In the partitioned scheme the database is divided into a number of disjoint partitions each of which is placed at different site. Replicated designs can be either fully replicated (also called fully duplicated) where entire database is stored at each site, or partially replicated (or partially duplicated) where each partition of the database is stored at more than one site, but not at all the sites.

• The two fundamental design issues are fragmentation, the separation of the database into partitions called fragments, and distribution, the optimum distribution of fragments. The research in this area mostly involve mathematical programming in order to minimize the combined cost of storing the database, processing transactions against it, and message communication among site.

1. Distributed Directory Management

• A directory contains information (such as descriptions and locations) about data items in the database. Problems related to directory management are similar in nature to the database placement problem discussed in the preceding section.

• A directory may be global to the entire DDBS or local to each site; it can be centralized at one site or distributed over several sites; there can be a single copy or multiple copies.

1. Distributed Query Processing

• Query processing deals with designing algorithms that analyze queries and convert them into a series of data manipulation operations. The problem is how to decide on a strategy for executing each query over the network in the most cost-effective way, however cost is defined.

• The factors to be considered are the distribution of data, communication cost, and lack of sufficient locally-available information. The objective is to optimize where the inherent parallelism is used to improve the performance of executing the transaction, subject to the abovementioned constraints.

1. Distributed Concurrency Control

• Concurrency control involves the synchronization of access to the distributed database, such that the integrity of the database is maintained. It is, without any doubt, one of the most extensively studied problems in the DDBS field.

• The concurrency control problem in a distributed context is somewhat different that in a centralized framework. One not only has to worry about the integrity of a single database, but also about the consistency of multiple copies of the database. The condition that requires all values of multiple copies of every data item to converge to the same value is called mutual consistency.

• Let us only mention that the two general classes are pessimistic, synchronizing the execution of the user request before the execution starts, and optimistic, executing requests and then checking if the execution has compromised the consistency of the database.

• Two fundamental primitives that can be used with both approaches are locking, which is based on the mutual exclusion of access to data items, and time-stamping, where transactions executions are ordered based on timestamps.

• There are variations of these schemes as well as hybrid algorithms that attempt to combine the two basic mechanisms.

1. Distributed Deadlock Management

• The deadlock problem in DDBSs is similar in nature to that encountered in operating systems.

• The competition among users for access to a set of resources (data, in this case) can result in a deadlock if the synchronization mechanism is based on locking. The well-known alternatives of prevention, avoidance, and detection/recovery also apply to DDBSs.

1. Reliability of Distributed DBMS

• It is important that mechanisms be provided to ensure the consistency of the database as well as to detect failures and recover from them. The implication for DDBSs is that when a failure occurs and various sites become either inoperable or inaccessible, the databases at the operational sites remain consistent and up to date.

• Furthermore, when the computer system or network recovers from the failure, the DDBSs should be able to recover and bring the databases at the failed sites up-to date. This may be especially difficult in the case of network partitioning, where the sites are divided into two or more groups with no communication among them.

1. Replication

• If the distributed database is (partially or fully) replicated, it is necessary to implement protocols that ensure the consistency of the replicas, i.e. copies of the same data item have the same value.

• These protocols can be eager in that they force the updates to be applied to all the replicas before the transactions completes, or they may be lazy so that the transactions updates one copy (called the master) from which updates are propagated to the others after the transaction completes.

*Let’s check the take away from this lecture*

1) Which of the following is a disadvantage of replication?

a) Reduced network traffic

b) If the database fails at one site, a copy can be located at another site.

**c) Each site must have the same storage capacity.**

d) Each transaction may proceed without coordination across the network.

2) A distributed database can use which of the following strategies?

a) Totally centralized at one location and accessed by many sites

b) Partially or totally replicated across sites

c) Partitioned into segments at different sites

**d) All of the above**

3) Which of the following is not one of the stages in the evolution of distributed DBMS?

**a) Unit of work**

b) Remote unit of work

c) Distributed unit of Work

d) Distributed request

Exercise

Q.1 Define Replication.

Q.2 Write short notes on distributed concurrency control.

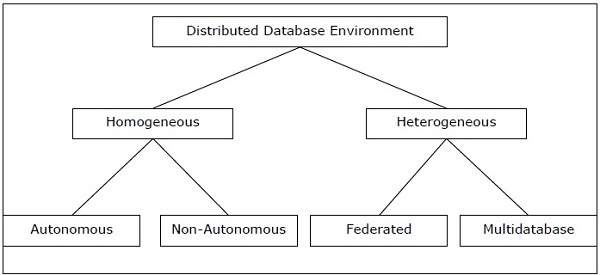
Q.3 Explain the design issues of distributed databases.

**Learning from this lecture**: Learners will be able to understand the various design issues of distributed databases.

**Lecture 3**

**Types of Distributed Databases**

Distributed databases can be broadly classified into homogeneous and heterogeneous distributed database environments, each with further sub-divisions, as shown in the following illustration.



Homogeneous Distributed Databases

In a homogeneous distributed database, all the sites use identical DBMS and operating systems. Its properties are −

* The sites use very similar software.
* The sites use identical DBMS or DBMS from the same vendor.
* Each site is aware of all other sites and cooperates with other sites to process user requests.
* The database is accessed through a single interface as if it is a single database.

Types of Homogeneous Distributed Database

There are two types of homogeneous distributed database −

* Autonomous − Each database is independent that functions on its own. They are integrated by a controlling application and use message passing to share data updates.
* Non-autonomous − Data is distributed across the homogeneous nodes and a central or master DBMS co-ordinates data updates across the sites.

Heterogeneous Distributed Databases

In a heterogeneous distributed database, different sites have different operating systems, DBMS products and data models. Its properties are −

* Different sites use dissimilar schemas and software.
* The system may be composed of a variety of DBMSs like relational, network, hierarchical or object oriented.
* Query processing is complex due to dissimilar schemas.
* Transaction processing is complex due to dissimilar software.
* A site may not be aware of other sites and so there is limited co-operation in processing user requests.

Types of Heterogeneous Distributed Databases

* Federated − Heterogeneous database systems are independent in nature and integrated together so that they function as a single database system.
* Un-federated − Database systems employ a central coordinating module through which the databases are accessed.

*Let’s check the take away from this lecture*

1) An autonomous homogenous environment is which of the following?

**a) The same DBMS is at each node and each DBMS works independently.**

b) The same DBMS is at each node and a central DBMS coordinates database access.

c) A different DBMS is at each node and each DBMS works independently.

d) A different DBMS is at each node and a central DBMS coordinates database access.

2) A heterogeneous distributed database is which of the following?

a) The same DBMS is used at each location and data are not distributed across all nodes.

b) The same DBMS is used at each location and data are distributed across all nodes.

c) A different DBMS is used at each location and data are not distributed across all nodes.

**d) A different DBMS is used at each location and data are distributed across all nodes.**

3) A homogeneous distributed database is which of the following?

a) The same DBMS is used at each location and data are not distributed across all nodes.

**b) The same DBMS is used at each location and data are distributed across all nodes.**

c) A different DBMS is used at each location and data are not distributed across all nodes.

d) A different DBMS is used at each location and data are distributed across all nodes.

Exercise

Q.1 What is homogeneous database?

Q.2 What is heterogeneous database?

Q.3 Discuss the types of homogeneous and heterogeneous databases.

**Learning from this lecture**: Learners will be able to understand the concept behind the types of distributed databases.

**Lecture 4**

**Distributed Database Architectures: Client-Server**

DDBMS architectures are generally developed depending on three parameters −

* **Distribution** − It states the physical distribution of data across the different sites.
* **Autonomy** − It indicates the distribution of control of the database system and the degree to which each constituent DBMS can operate independently.
* **Heterogeneity** − It refers to the uniformity or dissimilarity of the data models, system components and databases.

Architectural Models

Some of the common architectural models are −

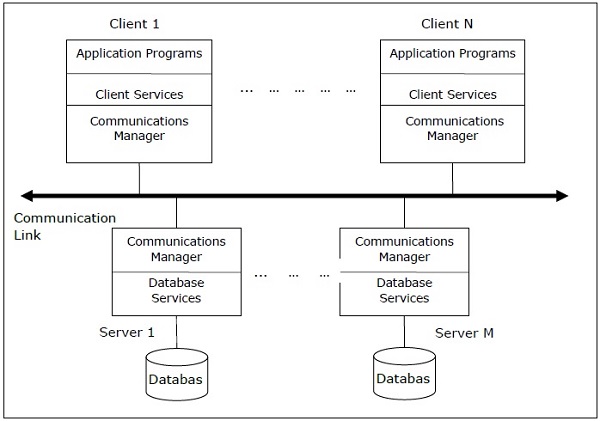
* Client - Server Architecture for DDBMS
* Peer - to - Peer Architecture for DDBMS
* Multi - DBMS Architecture

**Client - Server Architecture for DDBMS**

This is a two-level architecture where the functionality is divided into servers and clients. The server functions primarily encompass data management, query processing, optimization and transaction management. Client functions include mainly user interface. However, they have some functions like consistency checking and transaction management.

The two different client - server architectures are −

* Single Server Multiple Client
* Multiple Server Multiple Client (shown in the following diagram)



*Let’s check the take away from this lecture*

1) Depending on the situation each node in the Distributed Database system can act as, \_\_\_\_\_\_\_\_\_ .

a) A client

b) A server

**c) Both A & B**

d) None of the above

2) A(n) \_\_\_\_\_\_\_\_ is a database stored on multiple computers in multiple locations that are NOT connected by a data communications link.

a) distributed database

**b) decentralized database**

c) unlinked database

d) data repository

3) Which of the following are business conditions that encourage the use of distributed databases?

a) Companies with less than 10 employees

b) Lack of Data sharing needs

**c) Data communication reliability**

d) Companies that only store data on spreadsheets

Exercise

Q.1 Define Autonomy.

Q.2 Define Distribution.

Q.3 Draw and explain the client-server architecture.

**Learning from this lecture**: Learners will be able to understand the client-server distributed database architecture.

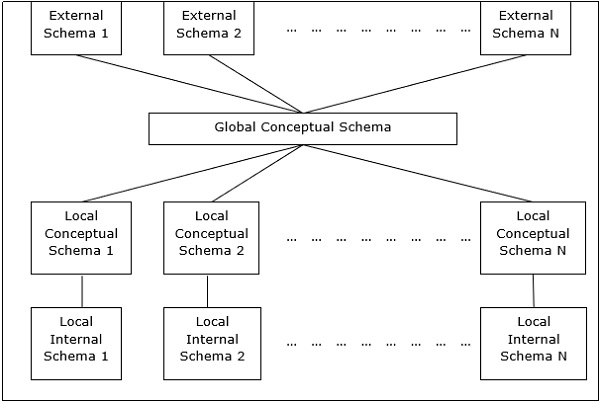
**Lecture 5**

**Distributed Database Architectures: Peer-Peer**

In these systems, each peer acts both as a client and a server for imparting database services. The peers share their resource with other peers and co-ordinate their activities.

This architecture generally has four levels of schemas −

* **Global Conceptual Schema** − Depicts the global logical view of data.
* **Local Conceptual Schema** − Depicts logical data organization at each site.
* **Local Internal Schema** − Depicts physical data organization at each site.
* **External Schema** − Depicts user view of data.



*Let’s check the take away from this lecture*

1) Which of the following environments uses the same DBMS at each node with a central or master DBMS coordinating database access across nodes?

a) Centralized; maximum

b) Centralized; minimum

**c) Homogeneous; non autonomous**

d) Federated; non autonomous

2) A transaction manager is which of the following?

a) Maintains a log of transactions

b) Maintains before and after database images

c) Maintains appropriate concurrency control

**d) All of the above.**

3) Location transparency allows for which of the following?

a) Users to treat the data as if it is at one location

b) Programmers to treat the data as if it is at one location

c) Managers to treat the data as if it is at one location

**d) All of the above**

Exercise

Q.1 Define external schema and local internal schema.

Q.2 Describe local conceptual schema.

Q.3 Explain peer-peer distributed database architecture with the help of diagram.

**Learning from this lecture**: Learners will be able to understand the peer-peer distributed database architecture.

**Lecture 6**

**Distributed Database Architectures: Multi - DBMS Architecture**

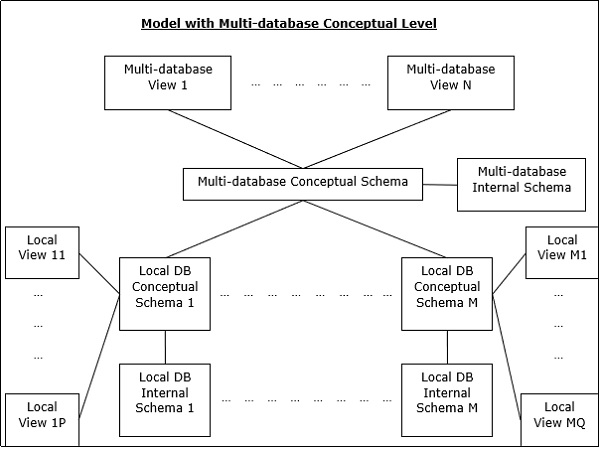
This is an integrated database system formed by a collection of two or more autonomous database systems.

Multi-DBMS can be expressed through six levels of schemas −

* **Multi-database View Level** − Depicts multiple user views comprising of subsets of the integrated distributed database.
* **Multi-database Conceptual Level** − Depicts integrated multi-database that comprises of global logical multi-database structure definitions.
* **Multi-database Internal Level** − Depicts the data distribution across different sites and multi-database to local data mapping.
* **Local database View Level** − Depicts public view of local data.
* **Local database Conceptual Level** − Depicts local data organization at each site.
* **Local database Internal Level** − Depicts physical data organization at each site.

There are two design alternatives for multi-DBMS −

* Model with multi-database conceptual level.
* Model without multi-database conceptual level.



*Let’s check the take away from this lecture*

1) A distributed transaction can be ............. if queries are issued at one or more nodes.

a) fully read-only

**b) partially read-only**

c) fully read-write

d) partially read-write

2 Which transaction contains statements that access more than one node?

a) A Remote Transaction

**b) A Distributed transaction**

c) Both A & B

d) None of the above

3) Which of the following parallel database architectures is/are mainly used by distributed database system?

a) Shared Memory

b) Shared Disk

**c) Shared Nothing**

d) Hierarchical

Exercise

Q.1 What is multi-database system?

Q.2 Define multi-database conceptual schema.

Q.3 Explain Multi-DBMS architecture with diagram.

**Learning from this lecture**: Learners will be able to understand the Multi-DBMS architecture.

**Conclusion**

The study of Distributed database system helps to know about its features and design goals. Also one can learn the types of distributed databases and understand the various architectures used.

**Short Answer Questions:**

1. What is distributed database?

**Ans:**

A distributed database is a collection of multiple interconnected databases, which are spread physically across various locations that communicate via a computer network.

2. What is homogeneous database?

**Ans:**

In a homogeneous distributed database, all the sites use identical DBMS and operating systems.

3. What is heterogeneous database?

**Ans:**

A heterogeneous database system is an automated (or semi-automated) system for the integration of heterogeneous, disparate database management systems to present a user with a single, unified query interface.

4. What is multi-database system?

**Ans.**

A multi-database system (MDBS) is a facility that allows users access to data located in multiple autonomous database management systems (DBMSs). In such a system, global transactions are executed under the control of the MDBS. Independently, local transactions are executed under the control of the local DBMSs. Each local DBMS integrated by the MDBS may employ a different transaction management scheme.

5. Define Autonomy.

**Ans.**

− It indicates the distribution of control of the database system and the degree to which each constituent DBMS can operate independently

6. Define Distribution.

**Ans.**

− It states the physical distribution of data across the different sites.

7. What is replication.

**Ans.**

Maintaining copy of data in another site is termed as replication.If the distributed database is (partially or fully) replicated, it is necessary to implement protocols that ensure the consistency of the replicas, i.e. copies of the same data item have the same value.

8. List the features of distributed database.

**Ans.**

* Databases in the collection are logically interrelated with each other. Often they represent a single logical database.
* Data is physically stored across multiple sites. Data in each site can be managed by a DBMS independent of the other sites.
* The processors in the sites are connected via a network. They do not have any multiprocessor configuration.
* A distributed database is not a loosely connected file system.
* A distributed database incorporates transaction processing, but it is not synonymous with a transaction processing system.

**Long Answer Questions:**

1. Explain the design issues of distributed database.

**Ans:** Refer the contents under lecture 2

2. Describe the types of distributed databases.

**Ans:** Refer the contents under lecture 3

3. Draw the diagram of client-server distributed database architecture and explain

**Ans:** Refer the contents under lecture 4

4. Explain peer-peer distributed database architecture with the help of diagram.

**Ans:** Referthe contents under lecture 5

5. Explain multi-DBMS architecture with diagram.

**Ans:** Refer the contents under lecture 6

**Set of Questions for FA/IA/ESE**

1. List the features of distributed database system. (5M)

2. State the difference between distributed database and centralized database. (5M)

3. Explain the design issues of distributed database. (10M)

4. Describe the types of distributed databases. (5M)

5. Draw the diagram of client-server distributed database architecture and explain. (10M)

6. Explain peer-peer distributed database architecture with the help of diagram. (10M)

7. Explain multi-DBMS architecture with diagram. (10 M)

**References:**

1. “Distributed Database Systems”, Chhanda Ray, PEARSON Education, First Edition 2009.

2. “Principles of Distributed Database Systems”, M. Tamer Ozsu, Patrick Valduriez, Springer, Third Edition 2011.

**Self-assessment**

Q.1) What is Distributed database? What is the need for distributed databases.

Q.2) Describe the difference between centralized database and distributed databse.

Q.3) Explain the types of distributed databases.

Q.4) Explain the different distributed database architectures.

**Self-evaluation**

|  |  |
| --- | --- |
| Name of Student |  |
| Class |  |
| Roll No. |  |
| Subject |  |
| Module No. |  |
| S.No |  | Tick  Your choice |
|  | Do you understand the need for distributed database? | * Yes * No |
|  | Do you understand the design issues of distributed databases? | * Yes * No |
|  | Do you understand the types of distributed databases? | * Yes * No |
|  | Do you understand the different distributed database architectures? | * Yes * No |
|  | Do you understand module 3 ? | * Yes, Completely. * Partialy. * No, Not at all. |