ISSN: 2349 - 4891
International
Journal of Recent Research and Applied Studies
(Multidisciplinary Open Access Refereed e-Journal)

An Introductory Study on Distributed Database

M.Natarajan[1] & R.Geetha[2]

[1]*Asst. Prof. in Comp. Sci., Thanthai Hans Roever College, Perambalur, Tamilnadu, India.*
[2]*Research Scholar in Comp. Sci., Thanthai Hans Roever College, Perambalur, Tamilnadu, India.*

Received 18th August 2018, Accepted 10th September 2018

Abstract

A distributed database is a collection of multiple interconnected databases, which are spread physically across various locations that communicate via a computer network. This paper presents an overview of Distributed Database System along with their advantages and disadvantages. This paper also provides various aspects like replication, fragmentation and various problems that can be faced in distributed database systems. A centralized distributed database management system (DDBMS) integrates the data logically so it can be managed as if it were all stored in the same location.

Keywords: Homogeneous, Heterogeneous, Client-Server, Peer-to Peer, Multi DBMS.

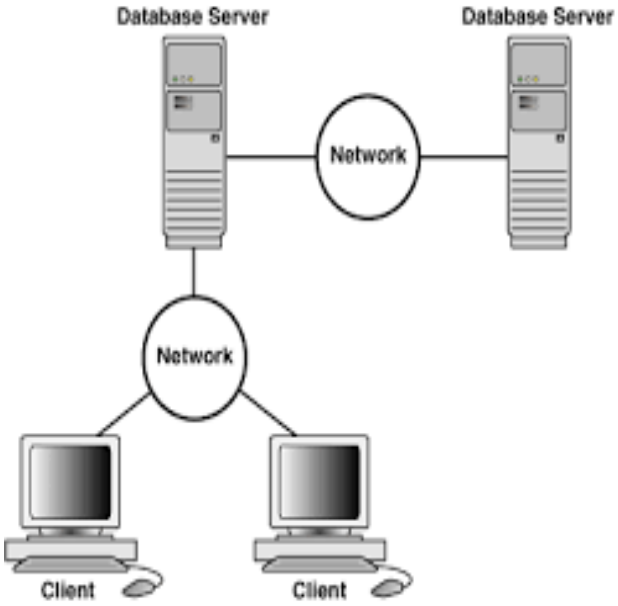
Introduction

The DDBMS synchronizes all the data periodically and ensures that updates and deletes performed on the data at one location will be automatically reflected in the data stored elsewhere. In a distributed database the data at each site is not necessarily an independent entity, but can be rather related to the data stored on the other sites. [1]A distributed database (DDB) is a collection of multiple, logically interrelated databases distributed over a computer network. A major motivation behind the development of database systems is the desire to integrate the operational data of an organization and to provide controlled access to the data. Although integration and controlled access may imply centralization, this is not the intention. In fact, the development of computer networks promotes a decentralized mode of work. The share ability of the data and the efficiency of data access should be improved by the development of a distributed database system that reflects this organizational structure, makes the data in all units accessible, and stores data proximate to the location where it is most frequently used.

Correspondence
M.Natarajan

Architecture

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. A Distributed System is the one in which hardware and software components at networked computers communicate and coordinate their activity only by passing messages.



A distributed database is a database in which storage devices are not all attached to a common processing unit such as the CPU. It may be stored in multiple computers, located in the same physical

location; or may be dispersed over a network of interconnected computers. A distributed database system consists of loosely-coupled sites that share no physical components [2].

Distributed Systems are required for Functional distribution, Inherent distribution in application domain, Economics, Better performance, and increased Reliability.

Types of Distributed Databases:

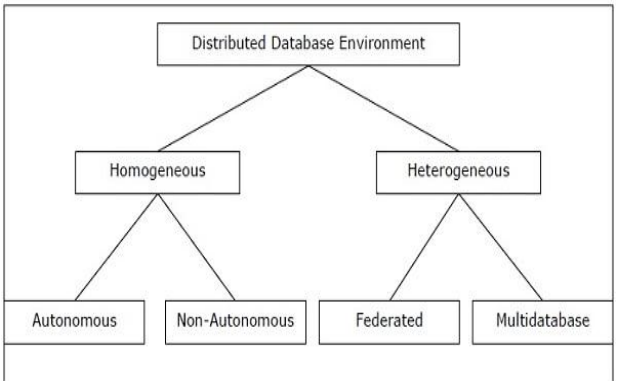
Distributed databases can be broadly classified into homogeneous and heterogeneous distributed database environments, each with further sub-divisions.

Homogeneous:

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

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:

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

Types of Heterogeneous Distributed database -

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

Architectural Model:

Client - Server Architecture- 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.

Peer- to-Peer Architecture- 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.

- Global Conceptual Schema
- Local Conceptual Schema
- Local Internal Schema
- External Schema

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

- Multi-database View Level
- Multi-database Conceptual Level
- Multi-database Internal Level
- Local database View Level
- Local database Conceptual Level
- Local database Internal Level

Concurrency problems in distributed databases.

1. Failure at local locations - When system recovers from failure the database is out dated compared to other locations. So it is necessary to update the database.

2. Failure at communication location-System should have a ability to manage temporary failure in a communicating network in distributed databases. In this case, partition occurs which can limit the communication between two locations.

3. Dealing with multiple copies of data-It is very important to maintain multiple copies of distributed data at different locations.

4. Distributed commit-While committing a transaction which is accessing databases stored on multiple locations, if failure occurs on some location during the commit process then this problem is called as distributed commit.

5. Distributed deadlock-Deadlock can occur at several locations due to recovery problem and concurrency problem (multiple locations are accessing same system in the communication network).

Concurrency Controls in distributed databases

There are three different ways of making distinguish copy of data by applying:

1) Lock based protocol: A lock is applied to avoid concurrency problem between two transaction in such a way that the lock is applied on one transaction and other transaction can access it only when the lock is released. The lock is applied on write or read operations. It is an important method to avoid deadlock.

2) Shared lock system (Read lock)-The transaction can activate shared lock on data to read its content. The lock is shared in such a way that any other transaction can activate the shared lock on the same data for reading purpose.

3) Exclusive lock-The transaction can activate exclusive lock on a data to read and write operation. In this system,

no other transaction can activate any kind of lock on that same data.

Database Security and Threats:

Data security is an imperative aspect of any database system. It is of particular importance in distributed systems because of large number of users, fragmented and replicated data, multiple sites and distributed control.

Threats in a Database.

- **Availability loss** − Availability loss refers to non-availability of database objects by legitimate users.
- **Integrity loss** − Integrity loss occurs when unacceptable operations are performed upon the database either accidentally or maliciously. This may happen while creating, inserting, updating or deleting data. It results in corrupted data leading to incorrect decisions.
- **Confidentiality loss** − Confidentiality loss occurs due to unauthorized or unintentional disclosure of confidential information. It may result in illegal actions, security threats and loss in public confidence.

Measures of Control

The measures of control can be broadly divided into the following categories −

- **Access Control** − Access control includes security mechanisms in a database management system to protect against unauthorized access. A user can gain access to the database after clearing the login process through only valid user accounts. Each user account is password protected.
- **Flow Control** − Distributed systems encompass a lot of data flow from one site to another and also within a site. Flow control prevents data from being transferred in such a way that it can be accessed by unauthorized agents. A flow policy lists out the channels through which information can flow. It also defines security classes for data as well as transactions.
- **Data Encryption** − Data encryption refers to coding data when sensitive data is to be communicated over public channels. Even if an unauthorized agent gains access of the data, he cannot understand it since it is in an incomprehensible format.

A distributed system needs additional security measures than centralized system, since there are many users, diversified data, multiple sites and distributed control.

In distributed communication systems, there are two types of intruders −

- **Passive eavesdroppers** − They monitor the messages and get hold of private information.
- **Active attackers** − They not only monitor the messages but also corrupt data by inserting new data or modifying existing data.

Security measures encompass security in communications, security in data and data auditing.

Communications Security

In a distributed database, a lot of data communication takes place owing to the diversified location of data, users and transactions. So, it demands secure communication between users and databases and between the different database environments.

Security in communication encompasses the following −

- Data should not be corrupt during transfer.
- The communication channel should be protected against both passive eavesdroppers and active attackers.
- In order to achieve the above stated requirements, well-defined security algorithms and protocols should be adopted.

  Two popular, consistent technologies for achieving end-to-end secure communications are −
- Secure Socket Layer Protocol or Transport Layer Security Protocol.
- Virtual Private Networks (VPN).

Data Security

In distributed systems, it is imperative to adopt measure to secure data apart from communications. The data security measures are −

- **Authentication and authorization** − These are the access control measures adopted to ensure that only authentic users can use the database. To provide authentication digital certificates are used. Besides, login is restricted through username/password combination.
- **Data encryption** − The two approaches for data encryption in distributed systems are −
  - Internal to distributed database approach: The user applications encrypt the data and then store the encrypted data in the database. For using the stored data, the applications fetch the encrypted data from the database and then decrypt it.
  - External to distributed database: The distributed database system has its own encryption capabilities. The user applications store data and retrieve them without realizing that the data is stored in an encrypted form in the database.
- **Validated input** − In this security measure, the user application checks for each input before it can be used for updating the database. An un-validated input can cause a wide range of exploits like buffer overrun, command injection, cross-site scripting and corruption in data.

Data Auditing

A database security system needs to detect and monitor security violations, in order to ascertain the security measures it should adopt. It is often very difficult to detect breach of security at the time of occurrences. One method to identify security violations is to examine audit logs. Audit logs contain information such as −

- Date, time and site of failed access attempts.
- Details of successful access attempts.
- Vital modifications in the database system.

- Access of huge amounts of data, particularly from databases in multiple sites.

All the above information gives an insight of the activities in the database. A periodical analysis of the log helps to identify any unnatural activity along with its site and time of occurrence. This log is ideally stored in a separate server so that it is inaccessible to attackers.

Conclusion

Distribution of data has its own advantages and disadvantages. This paper presents a complete review on distributed databases. It is clear from the study that distribution of data involves the problem of deadlock. The Distributed database is basically used for to store large amount of data on different site or server. There are Various Partitions technique uses in distributed database

References

1. ArtiKadav et. al. Data Mining Standard
2. en.wikipedia.org/wiki/Database
3. Abadi, D., Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., and Zdonik, Aurora: A new model and architecture for data stream management. Summer-2003
4. Akal, F., Bohm, K., and Schek, H.-J. Olap query evaluation in adatabase ¨cluster: A performance study on intra-query parallelism.August 2002.
5. Zhili Zhang and WiliamPerrizo. 2000. Distributed Query Processing Using Active Networks. *ACM 1-58113-239- 5/00/003.*