CONCEPTS OF DISTRIBUTED AND PARALLEL DATABASE

Hiren H Darji[1], BinalS Shah[2], Manisha K Jaiswal[3]

[1] Assistant Professor, AIIS, Anand
Hirendarji7597@gmail.com

[2] Assistant Professor, AIIS, Anand
Binal.shah85@gmail.com

[3] Assistant Professor, AIIS, Anand
Mansha510@gmail.com

ABSTRACT

The purpose of this paper is to present anprologue to distributed databasesand Parallel Database. Also we compare Distributed and Parallel Database by using Different Characteristic. And also declare some research areas of Distributed Database that have to be solved.

Keywords: Distributed databases, Parallel Database,Information Retrieval.

1 Introduction to Distributed Database

In today's world of information systems, all course of people need access to companies' databases. In count to a company's own employees, these include the company's customers, likely customers, suppliers, and salespersons of all types. It is possible for a company to have all of its databases focused at one mainframe computer site with worldwide access to this site provided by communications networks, including the Internet. Although the management of such a centralized system and its databases can be controlled in a well-contained manner and this can be beneficial, it poses some problems as well. For example, if the single site goes down, then everyone is blocked from accessing the databases until the site comes back up again. Also the interactions costs from the many far PCs and terminals to the middle site can be expensive. One solution to such problems, and adiverse design to the centralized database concept, is known as distributed database. The idea is that in its place of having one, centralized database, we are going to spread the data out among the cities on the distributed network, each of which has its own computer and data storage facilities. [2]

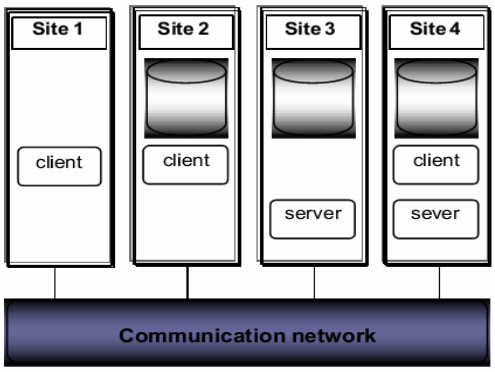
1.1 Definition

A distributed database (DDB) is a collection of multiple, logically interrelated databases distributed over a computer network. A distributed database management system (distributed DBMS) is the software system that authorities the management of the distributed database and makes the distribution transparent to the users [1]. The term distributed database system (DDBS) is usually used to refer to the mixture of DDB and the distributed DBMS. Distributed DBMSs are like to distributed file systems in that both assist access to distributed data.

1.2 Architecture of DDBs:

1.2.1 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.Thus, a client process could run on a personal computer and send queries to a server running on a mainframe.



Reward of client-server architecture:

1. Simple to implement because of the centralized server and division of functionality.

2. posh server machines are not underutilized with simple user interactions which are now pressed on to inexpensive client machines.

3. The users can have a known and friendly client side user interface rather than unfamiliar and unfriendly server interface. [3]

1.2.2 Collaborating Server

In CollaboratingServer system, we can have collection of database servers, each capable of running transactions beside local data, which cooperatively execute transactions spanning multiple servers.When a server receives a query that requires access to data at other servers, it generates appropriate sub queries to be executed by other servers and puts the results together to compute answers to the original query. [3]

1.2.3 Middleware

Middleware system is as special server, a layer of software that coordinates the execution of queries and transactions across one or more self-governing 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.We need just one database server that is capable of managing queries and transactions spanning multiple servers; the remaining servers only need to handle local queries and transactions.[3]

1.3 Types of Distributed Databases

1.3.1 Homogeneous Distributed Database is where the data stored across multiple sites is managed by same DBMS software at all the sites.

1.3.2 Heterogeneous Distributed Database is where multiple sites which may be autonomous are under the control of different DBMS software.[3]

1.4 Properties of Distributed Database

1.4.1 Distributed Data Independence: - The user should be able to access the database without having the need to know the location of the data. [3]

1.4.2 Distributed Transaction Atomicity: - The concept of atomicity should be distributed for the operation taking place at the distributed sites. [3]

2. Introduction to Parallel Databases

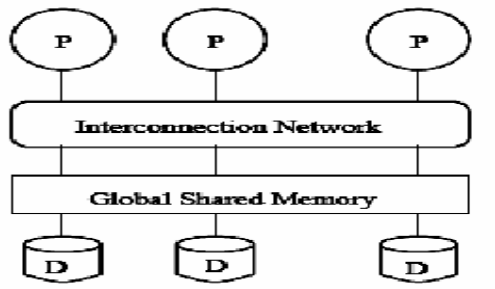
A parallel database system is one that seeks to get better performance through parallel implementation of various operations such as loading data, building indexes, and evaluating queries.

2.1 Architecture of Parallel Database

Parallel Database is implemented by using three types of Architecture and is:

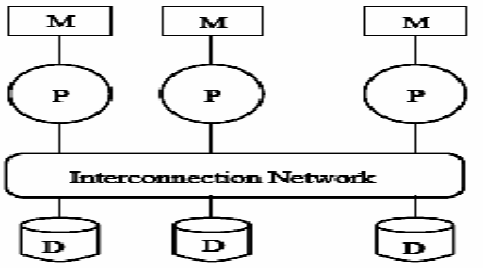
2.1.1 Shared-memory system

In this Architecture multiple CPUs are attached to an interconnection network and can access a common region of main memory. (See the following diagram)[4]



2.1.2 Shared-Disk system

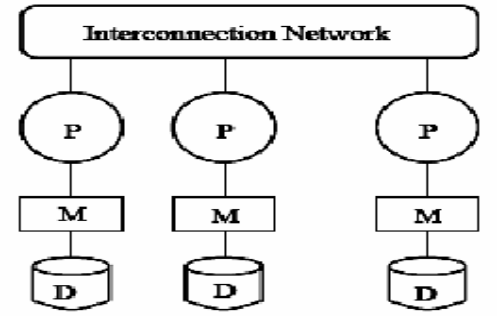
In this architecture each CPU has a private memory and direct access to all disks through an interconnection network. (See the following diagram)[4]



2.1.3 Shared-nothing system

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. (See the following diagram)[4]

Characteristic	Parallel DBMS	Distributed DBMS
Definition	It is a software system where multiple processors or machines are used to execute and run queries in parallel.	It is a software system that manages multiple logically interrelated databases distributed over a computer network.
Types	Shared Memory (Tightly coupled) Shared Disk (Loosely coupled) Shared nothing architecture Hierarchical architecture	HomogeneousHeterogeneousFederated DB system Multi database system
Geographical Location	The nodes are located at geographically same location.	The nodes are usually located at geographically different locations.
Execution Speed	Quicker	Slower
Overhead	Less	More
Node types	Compulsorily Homogeneous	Need not be homogeneous
Performance	Lower reliability & availability.	Higher reliability & availability.
Scope of Expansion	Difficult to expand	Easier to expand
Backup	Backup at one site only	Backup at multiple sites



3. Parallel vs. Distributed Databases [5]

4. Research Issues

4.1 Approximate Query Processing

It has been observed that in most typical data analysis and data mining applications, timeliness and interactivity are more important considerations than accuracy - thus data analysts are often willing to overlook small inaccuracies in the answer provided the answer can be obtained fast enough. This observation has been the primary driving force behind recent development of approximate query processing (AQP) techniques for aggregation queries in traditional databases and decision support systems. Many AQP techniques have been developed, the most popular ones based on random sampling, where a small random sample of the rows of the database is haggard, the query is executed on this small sample, and the results extrapolated to the whole database. In addition to simplicity of implementation, random sampling has the forcefulbenefit that in addition to an estimate of the aggregate, one can also provide self-belief intervals of the error with high probability. generally, two types of sampling-based approaches have been investigated: (a) Pre-computed samples - where a random sample is pre-computed by scanning the database, and the same sample is reused for several queries, and (b) Online samples - where the sample is drawn "on the fly" upon encountering a query.

4.2 Information Retrieval (Top-k ...)

IR is the art and science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases, whether relational standalone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data. There is anordinary confusion, however, between data retrieval, document retrieval, information retrieval, and text retrieval, and each of these have their own bodies of text, theory, praxis and technologies.

Automated information retrieval (IR) systems were originally used to manage information explosion in scientific text in the last few decades. Many universities and public libraries use IR systems to provide admission to books, journals, and other documents. IR systems are often related to object and query. Queries are official statements of information needs that are put to an IR system by the user. An object is an entity which keeps or stores information in a database. User queries are matched to documents stored in a database. A document is, therefore, a data object. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates.

4.3 P2P (Peer-to-Peer) Databases

A peer-to-peer (or P2P) computer network is a network that relies on the computing power and bandwidth of the participants in the network rather than fixed it in a relatively few servers. P2P networks are typically used for connecting nodes via largely ad hoc connections. Such networks are useful for many purposes. Sharing content files (see file sharing) containing audio, video, data or anything in digital format is very common, and real-time data, such as telephony traffic, is also passed using P2P technology. The term "P2P network" can also mean grid computing.

Some networks and channels, such as Napster, OpenNAP, or IRC @find, use a client-server structure for some tasks (e.g., searching) and a peer-to-peer structure for others. Networks such as Gnutella or Freenet use a peer-to-peer structure for all purposes, and are sometimes referred to as true peer-to-peer networks, although Gnutella is greatly facilitated by directory servers that inform peers of the network addresses of other peers.

Conclusion

Through this paper, we want to draw readers towards the helpful side of Distributed database and Parallel Database. We also described Architecture of Distributed and Parallel Database, Comparison of Distributed database and Parallel also in order to make readers totally aware about the topic being described here.

References:

1. Muhammad ShahidJamal,MohsinNazir, “Fundamental Research on Distributed Database “April 2012.

2. HarounRababaah, “DISTRIBUTED DATABASES FUNDAMENTALS AND RESEARCH “,Spring 2005.

3.Korth, Silberchatz, Sudarshan, “Database System Concepts” McGraw Hill.

4. Elmasri and Navathe, “Fundamentals of Database Systems”, Person Education.

5.http://student.ritzsoftec.com/content/parallel-vs-distributed-databases.

6. M. Tamer ¨Ozsu, Patrick Valduriez , “Distributed and Parallel Database Systems “

7. M. Tamer Özsu ,Patrick Valduriez , “DISTRIBUTED DATABASE SYSTEMS: WHERE ARE WE NOW? “