Distributed Databases

Database :- A database is an ordered collection of related data that is built for a specific purpose.

A database may be organized as a collection of multiple tables, where a table represents a real world element or entity. Each table has several different fields that represent the characteristic features of the entity.

For Example:- a company database may include tables for projects, employees, departments, products and financial records.

The fields in the employee table may be Name, company ID, Date of joining etc.

Database Management System :-

A DBMS is a collection of programs that Enables creation and maintenance of a database.

DBMS is available as a software package that facilitate defination, construction, Manipulation and sharing of data in a database.

"Defination" of a database includes description of the structure of a database.

"Construction" of a database includes actual storing of the data in any storage medium.

"Manipulation" refers to the retrieving information from the database, updating the database and generating reports.

"Sharing" of data facilitates data to be accessed by different users or programs.

Examples of DBMS Application Areas:

→ Automatic Teller Machine
→ Train Reservation System
→ Employee Management System
→ Student information system.

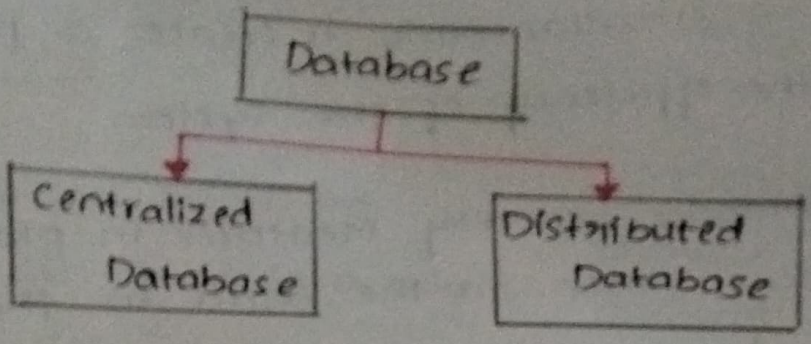
Examples of DBMS packages:

DBMS Packages Examples includes

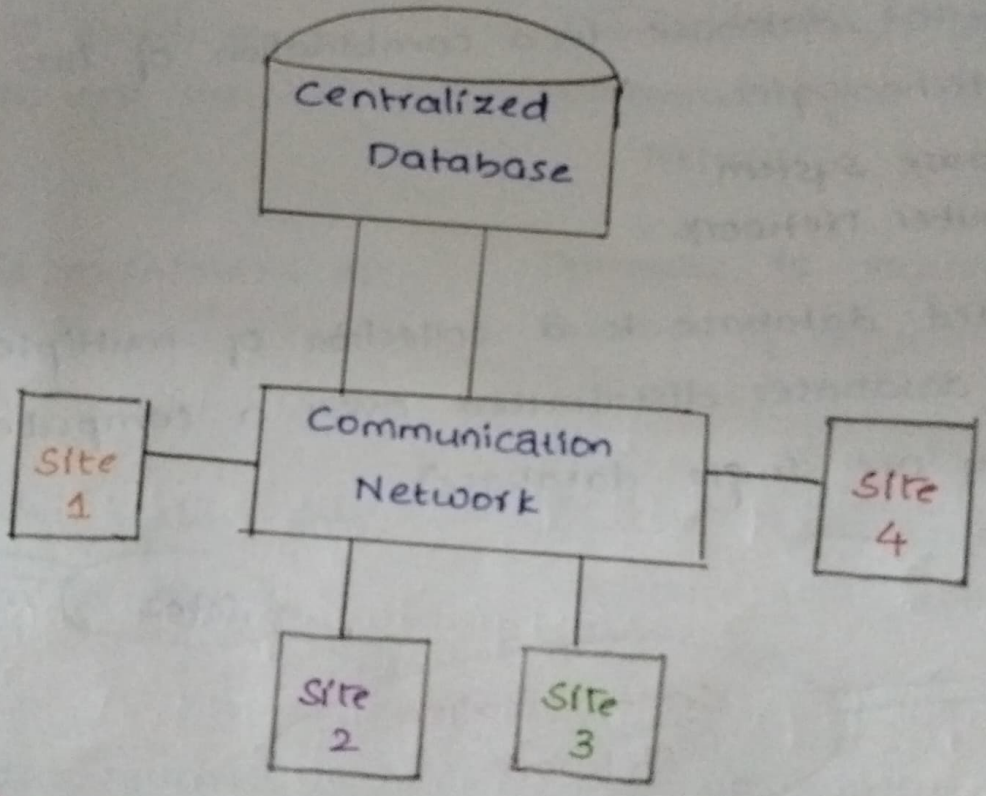
→ MYSQL
→ Oracle
→ SQL server dBASE
→ Foxpro postgreSQL
→ Microsoft Access
→ File Maker

RDBMS

Classification of Databases



1. Centralized Database:



Disadvantages of centralized Database:-

→ Since all the data is at one location, it takes more time to search and access it. If Network is slow, this process takes even more time

→ There is a lot of data access traffic for centralized database. This may create a bottleneck situation.

→ Since all the data is at the same location, if multiple users try to access it simultaneously it creates a problem. This may reduce the efficiency of the system.

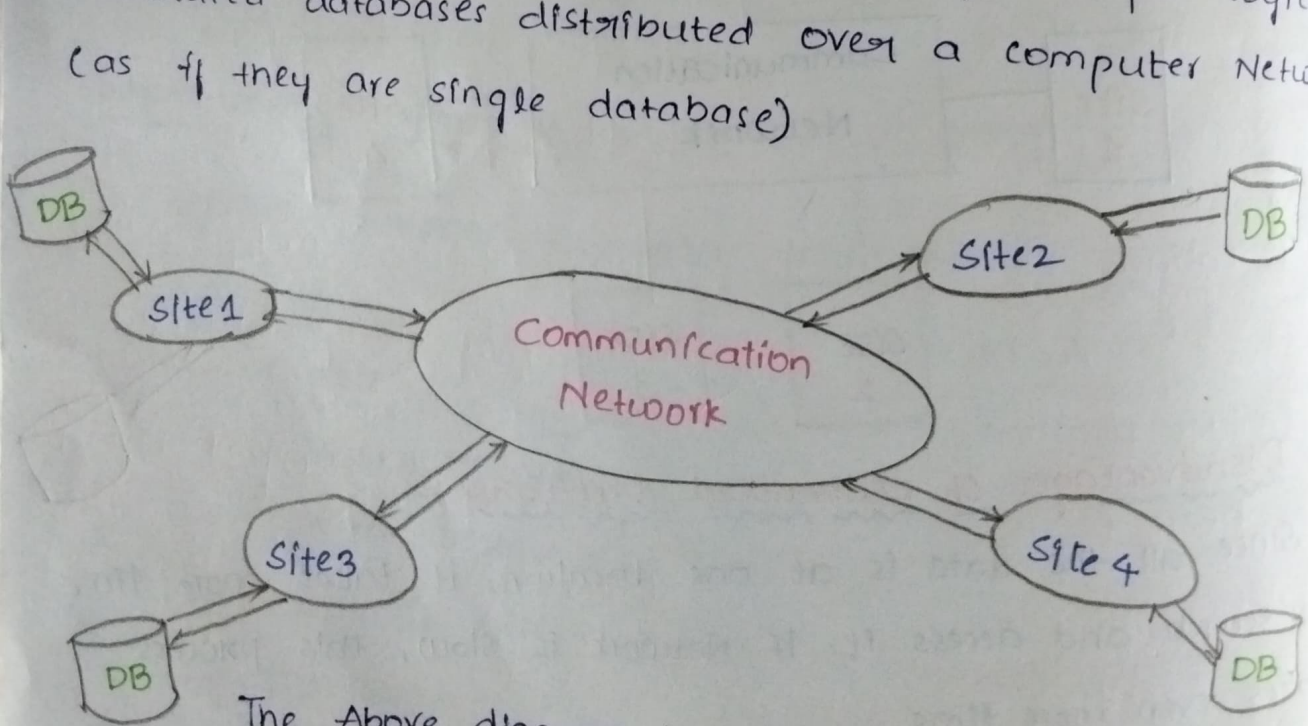
→ If there are no database recovery measures in place and a system failure occurs, then all the data in the database will be destroyed.

2. Distributed Database:-

A distributed database is a combination of two contrasting technologies.

(i) Database system
(ii) computer Network

A distributed database is a collection of multiple logically interelated databases distributed over a computer Network (as if they are single database)



The Above diagram is a typical example of distributed database. In which communication channel is used to communicate with the different locations and every system has its own database.

Distributed Database Vs Centralized Database.

Centralized DBMS	Distributed DBMS.
In centralized DBMS the database are stored in a only one site	In Distributed DBMS the database are stored on different site and help of network it can access it.
If the data is stored at single computer site, which can be used by multiple users.	Database and DBMS software distributed over many sites, connected by a computer Network.
Database is maintained at one site.	Database is maintained at a number of different sites.
If centralized system fails, Entire System is halted.	If one system fails, system continues work with other site.
It is a less reliable	It is a more reliable.

Advantages of Distributed Database:-

Modular Development:-

If the system needs to be expanded to new locations or new units, in centralized database systems, the action requires substantial efforts and disruption in the existing functioning.

However, in distributed databases, the work simply requires adding new computers and local data to the new site and finally connected them to the distributed system, with no interruption in current functions.

More Reliable:-

In case of database failures, the total system of centralized databases comes to a halt. However, in distributed systems, when a component fails, the functioning of the system continues may be at a reduced performance. Hence DDBMS is more reliable.

Better Response:-

If data is distributed in an efficient manner, the user requests can be met from local data itself thus providing faster response. on the other hand, in centralized systems, all queries have to pass through

the central computer for processing, which increases the response time.

Lower communication cost :-

In distributed database systems, if data is stored or located locally where it is mostly used. then the comm-unication costs for data manipulation can be minimized. This is not feasible in centralized systems.

Disadvantages of Distributed Databases.

Need for complex and Expensive software.

DDBMS demands complex and often expensive software to Provide data transparency and co-ordination across the several sites.

Processing overhead :-

Even simple operations may require a large number of communications and additional calculations to provide uniformity in data across the sites.

Data Integrity.

The Need for updating data in multiple sites Pose problems of data integrity.

Overheads for improper data distribution:.

Responsiveness of queries is largely dependent upon

Proper data distribution. often leads to very slow response to user requests.

Distributed DataBase:-

A distributed database is a collection of multiple interconnected databases, which are spread across physically 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 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 is a centralised 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 Sychronizes 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.

Factors Encouraging Distributed Database Management Systems:-

Distributed Nature of organisational units:-

Most organizations in the current times are subdivided in to multiple units that are physically distributed over the globe. Each unit requires it's own set of local data. thus, the overall database of the organization becomes distribut

Need for sharing of data:-

The Multiple organization units often need to communicate with each other and share their data and resources. This demands common databases or replicated databases that should be used in a synchronized manner.

Support for both OLTP and OLAP:-

Online Transaction processing and Online Analytical Processing work upon diversified systems which may have common data. Distributed database systems aid both these processing by Providing Synchronized data.

Database Recovery:-

One of the common techniques used in DDBMS is replication of data across different sites. Replication of data automatically helps in data recovery if database in any site is damaged. users can access data from other sites while the damaged site is being reconstructed. Thus database failure may become almost inconspicuous to users.

Support for Multiple Application Software:-

Most organizations use a variety of application software each with its specific database support. DDBMs Provides a uniform functionality for using the same data among different platforms.

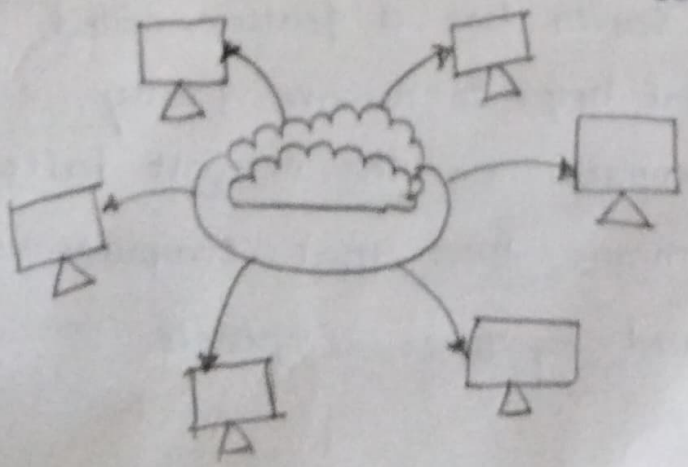
Distributed Data Processing (DDP)

Processing of data that is done online by different interconnected computers is known as distributed data processing.

We host our website on the online server nowadays cluster hosting is also available in which website data is stored in different clusters (remote computers).

When a visitor comes to the website then website pages are loaded from the server which is near to the user. Google also uses distributed data processing.

Google has database servers in all major countries. When computer user visits Google site from china then google website is loaded from china server.



Distributed Data Processing Diagram.

In Distributed processing, there Exists one main server which controls all other computers in the network.

Distributed Dataprocessing is done with high internet speed like querying the database.

Advantages:-

InExpensive:- Some companies buy a mainframe and Super computers to do large scale processing online but it cost those a hundred thousand dollars.

Buying mainframe and Supercomputers tend to centralized processing if that computer malfunction then all company data get into risk.

On the other hand, doing processing by connecting Personal computers from different locations can save money because they cost them a thousand bucks. data is also distributed so adding and removing (nodes) is easy.

Easy to replace remote computers:-

Microsoft windows server has a feature called Fail over clustering that helps to remove faulty computers. If any computer on the network fails or corrupted by some means then that computer is automatically replaced by other computers.

Optimized Processing:-

Managing data on online server solves slow processing on the personal computer, we can do extra tasks also. doing extra tasks consumes processor power. But the online computer is dedicated one type of processing and it is more likely to increase processing powers. Database server can only handle database queries and file server stores files. so data processing is optimized.

Easy to Expand:-

Suppose your company needs more data processing than expected then you can easily attach more computers to the distributed network.

Parallel Processing

Adding and removing computers from the network cannot disturb data flow. All data from different computers are processed in parallel processing means data is updated at the same time from all nodes.

Better Performance:-

The overall performance of the company gets better and data is filtered and processed more rapidly in the distributed environment.

Backup of data:-

Data can be backup from any computer connected to the network. so the user can backup data at a different time and work with that data locally and then upload the data to the server.

Local Data Synchronization :-

All the computers on the network can have local storage of important data. Suppose there are different office branches interconnected to each other. All branch computers are interlinks with main branch office. All office branch computers have a local copy of data. Office users edit and update data and then upload to the main server.

So, the data is synced and available to all computers. working locally with data is easy and fast and when the user thinks that his work is complete then at the end of the day he can sync that data with the main server.

Data Recovery :-

If some data like the database is a locs in any computer then it can be recovered by another interconnect-ed computer. i.e main database server.

Disadvantages:

Complexity :-

Computers attached in DDP are difficult to trouble-shoot, design and administrate.

Planning data synchronization is difficult:

Doing the correct synchronization of data is difficult to develop. Sometimes data is updated in wrong order, so administrators have to keep the focus on it before making a distributed network.

Data Security:

If the unauthorized computer is connected to distributed network then it can affect other computer performance and data can be a loss also.

Examples of DDP:

- Hosting a website on the online server.
- Online photo editing tools.
- Airline ticketing system.
- Processing user data by mobile computers
- Dropbox, Google drive, MSN drive, Google photos.
- Report Generation from Satelite.
- Weather forecast system.

Distributed DataBase System:-

A distributed database is basically a database that is not limited to one system, it is spread over different sites, i.e on multiple computers or over a network of computers.

A distributed database system is located on various sites that don't share physical components. This may be required when a particular database needs to be accessed by various users globally. It needs to be managed such that for the users it looks like one single database.

Types:

1. Homogeneous Database:

In a homogeneous database, all different sites store database identically. The operating system, database management system, and the data structure used.

All are the same at all sites. Hence they're easy to manage.

2. Heterogeneous Database:-

In a heterogeneous distributed database different sites can use different schema and software that can lead to problems in query processing transactions. also, a particular site might be completely unaware of the other sites.

Different computers may use a different database application. They may even use different data models for the database. Hence, translations are required for different sites to communicate.

Distributed Data Storage:

There are 2 ways in which data can be stored on different sites. These are

1. Replication:-

In this approach, the entire relationship is stored redundantly at 2 or more sites. If the entire database is available at all sites, it is a fully redundant database.

Hence, in replication, systems maintain copies of data

This is advantageous as it increases the availability of data at different sites. Also now query requests can be processed in parallel.

However, it has certain disadvantages as well. Data needs to be constantly updated. any change made at one site needs to be recorded at every site that relation is stored or else it may lead to inconsistency.

This is a lot of overhead. Also, concurrency control becomes way more complex as concurrent access now needs to be checked over a number of sites.

2. Fragmentation :-

In this approach, the relations are fragmented (i.e, they're divided into smaller parts) and each of the fragments is stored in different sites where they're required. It must be made sure that the fragments are such that they can be used to reconstruct the original relation (i.e, there isn't any loss of data).

Fragmentation is advantageous as it does not create copies of data, consistency is not a problem

Fragmentation of relations can be done in 2 ways.

Horizontal Fragmentation :- (Splitting by rows) :

The relation is fragmented into groups of tuples so that each tuple is assigned to at least one fragment.

Vertical Fragmentation:- Splitting by columns:-

The schema of the relation is divided in to smaller schemas. Each fragment must contain a common candidate key so as to ensure a loss less join.

In certain cases, an approach that is hybrid of fragmentation and replication is used.

Applications of Distributed Database:

- It is used in corporate management information system.
- It is used in multimedia applications.
- Used in military's control system, Hotel chains etc.
- It is also used in manufacturing control system.

Promises of DDBMS :-

A Distributed Database Management System (distributed DBMS) is then defined as the software system that permits the management of the distributed database makes the distribution transparent to the users.

The two important terms in this system are logically "interrelated and distributed" over a "computer Network."

What we are intrested in is an environment where data are distributed among a number of sites.

There are Four Fundamentals. which may also be viewed as Promises of DDBs Technology.

- Transparent Management of distributed and replicated data.
- Reliable access to data through distributed transactions.
- Improved Performance.
- Easier System Performance.

Transparent Management of Distributed and Replicated data.

A Transparent System is a system which "hides the implementation details from users.

Let us start our discussion with an example. consider an Engineering firm that has offices in Boston, waterloo, Paris, and San Francisco.

They run projects at each of these sites and would like to maintain a database of their employees. the projects and other related data.

Assuming that the database is relational we can store this information in 2 relations.

EMP (ENO, ENAME, TITLE) and

PROJ (PNO, PNAME, BUDGET)

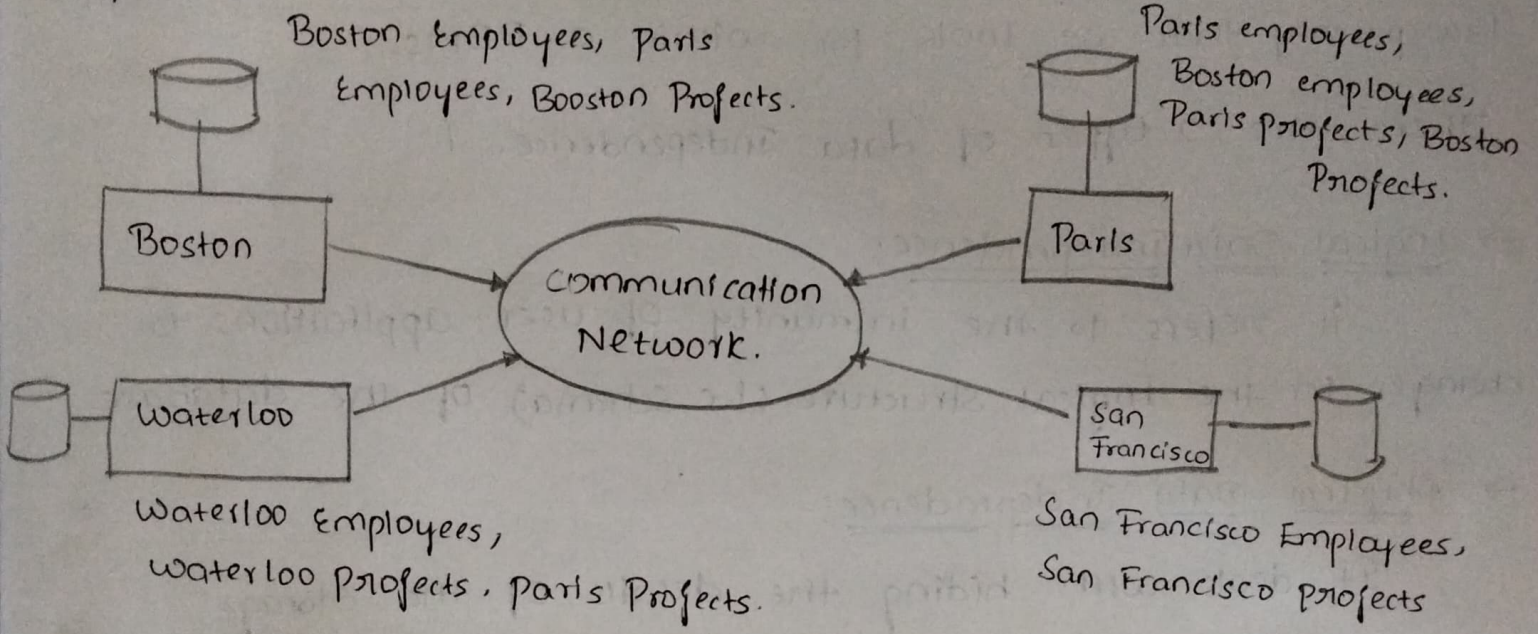
We also introduce a third relation to store salary information SAL (TITLE, AMT) and a fourth relation ASG which indicates which employees have been assigned to which projects for what duration with what responsibility

ASG (ENO, PNO, RESP, DUR)

If all this data were stored in a centralized DBMS, and we wanted to find out the names and employees who worked on a project for more than 12 months. we would specify this using the following SQL query

```
SELECT ENAME, AMT
FROM EMP, ASG, SAL
WHERE ASG DUR > 12
AND EMP.ENo = ASG.ENO
AND SAL. TITLE = EMP.TITLE
```

Furthermore, it may be preferable to duplicate some of this data at other sites for performance and reliability reasons. The result is a distributed database which is fragmented and replicated.



Fully Transparent access means that the users can still pose the query are as specified above, without paying any attention to the fragmentation, location or replication of data, and let the system worry about resolving these issues.

For a system to adequately deal with this type of query over a distributed fragmented and replicated database, it needs to deal with a number of different types of transparencies.

→ Data Independence
→ Network Transparency
→ Replication Transparency.
→ Fragmentation Transparency.

Data Independence :-

Data Independence is a fundamental form of Transparency that we look for within a DBMS.

There are 2 types of data Independence. !

→ **Logical Data Independence:**

it refers to the immunity of user applications to changes in the logical structure (i.e schema) of the database

→ **physical Data Independence:-**

It deals with hiding the details of the storage structure from user application

when a user application is written, it should not be concerned with the details of physical data organization.

Therefore, the user application should not need to be modified when data organization changes occur due to Performance considerations.

Network Transparency :-

User should be protected from the operational details of the network possibly even hiding the existence of the network. This type of transparency is referred to as network transparency or distribution transparency.

Replication Transparency :-

For performance, reliability and availability reasons. It is usually desirable to be able to distribute data in a replicated fashion across the machines on a network.

Furthermore, if one of the machines fails, a copy of data are still available on another machine on the network. In fact, the decision as to whether to replicate or not and how many copies of any database object to have depends to a considerable degree on user applications.

Fragmentation Transparency :-

There are two general types of fragmentation alternatives.

Horizontal Fragmentation :- a relation is partitioned in to set of sub relations, each of which have a subset of the tuples (rows) of the original relation.

Vertical Fragmentation :-

Where each sub relation is defined on a subset of the attributes (columns) of the original relation.

Distributed DBMs Problem Areas

→ Distributed Database Design:-

How to distribute the database

Replicated & Non replicated database distribution

A related problem in directory Management.

→ Query Processing:-

Convert user Transactions to data manipulation instructions.

Optimization problem

min [cost = data transmission + Local processing]

→ Directory Management:-

There are 3 Levels of directories

→ conceptual
→ Logical
→ physical.

Directories are consulted for most database operations.

There are many issues that concern wheather to distribute or centralize the directories.

→ Concurrency control.

Synchronization of concurrent accesses

consistency and isolation of transactions effects.

Deadlock Management

Deadlock Management:-

Similar to operating systems deadlock Management.
well known solutions.

→ Prevention
→ Avoidance
→ Detection/Recovery.

- Reliability:-

-How to make the system resilient to failures
Atomicity and durability.

Operating System support:-

O.S with Proper support for database operations.
Divide in 2 group blw general Purpose Processing requirements and database Processing requirements.

Heterogeneous databases:-

Sometimes called multi-databases.

Distributed databases are fully Autonomous usually databases already exist and distributed database system integrates them.

complimentary to distributed database systems.

Relationship blw issues