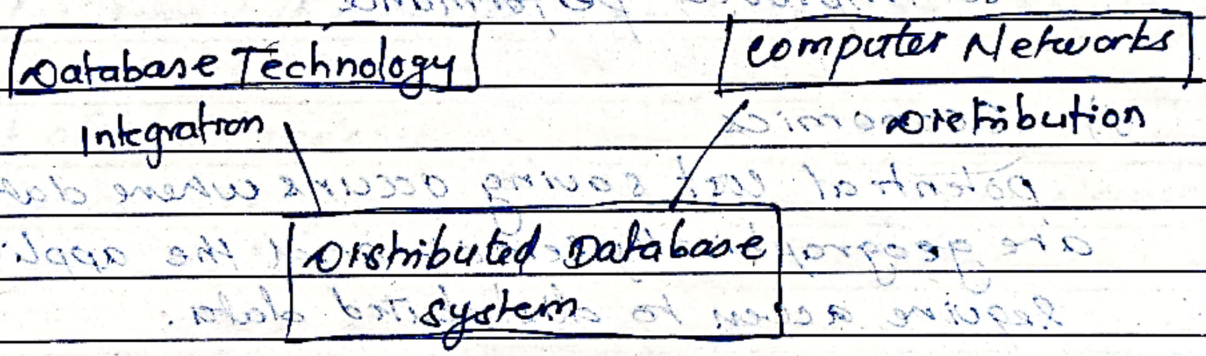
PHYSICS PRACTICAL SHEETS

Date: ..... CAMPUS
Class: Purusottam Adhikari Experiment No.: .....
Roll No: ..... Group: .....
Shift: ..... Sub: .....
Object of the Experiment (Block Letter) Set: .....

chapter 4

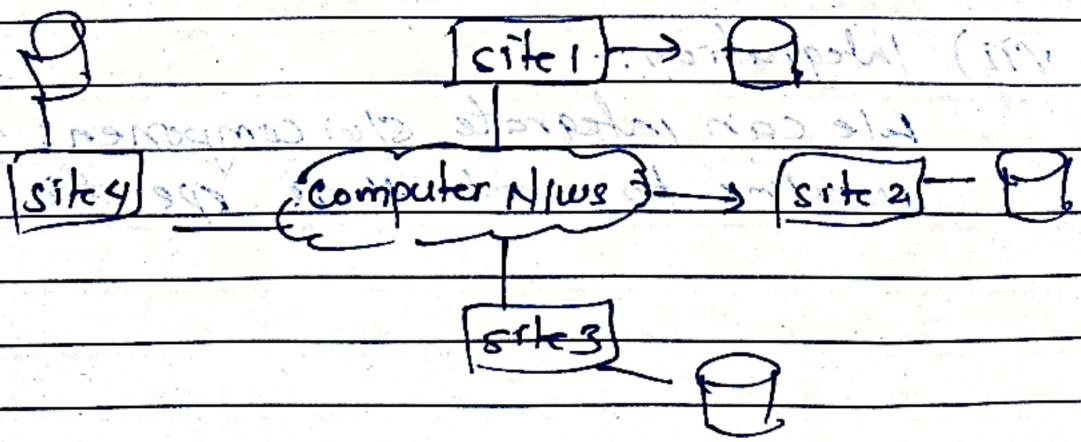
1. Discuss the distributed database concepts with advantages and disadvantages.

A distributed database (DDB) is a collection of multiple, logically interrelated databases distributed over a computer network. A distributed database management system (DDBMS) is the s/w that manages the DDB and provides an access mechanism that makes this distribution transparent to the users. The distributed database (DDB) and distributed database management systems (DDBMS) together is called Distributed Database System (DDBS)



Integration (Integration != Centralization)

fig: Distributed Database System



Advantages

i) Reflect organizational structure which are naturally distributed over several locations.

ii) Improved shareability and local autonomy:
Data can be placed at the site close to the user who normally use that data

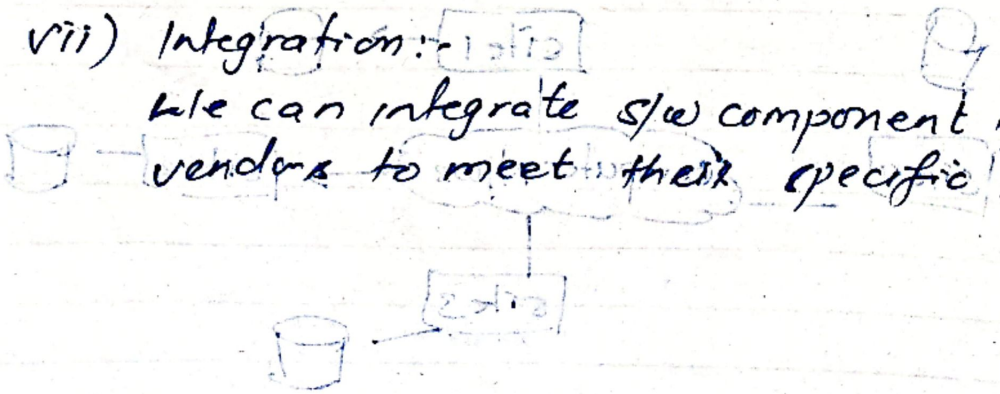
iii) Improved availability:
designed to continue if even if there failure of communication ~~lines~~ links.

iv) Improved reliability
Data exists more than one site so that each site handles only a part of entire database so improved performance

v) Economics
potential cost saving occurs where databases are geographically remote and the application require access to distributed data.

vi) Modular growth:
New site can be added to H/W without affecting operations of other site.

vii) Integration:-
We can integrate S/W component from different vendors to meet their specific requirements



Disadvantages

i) complexity: A distributed DBMS that hides the distributed nature from users and provides an acceptable level of performance, reliability which is more complex than centralized DBMS.

ii) cost :- Due to complex in nature procurement and maintenance costs are higher.

iii) Security:- In distributed DBMS not only does access to replicated data have to be controlled in multiple location but n/w itself has to be made secure

iv) Lack of standards:-
There no tools or methodologies to help users convert a centralized to distributed DBMS.

v) Lack of experience:-
We do not have same level of experience in industry as centralized.

vi) Database design are more complex than centralized database.

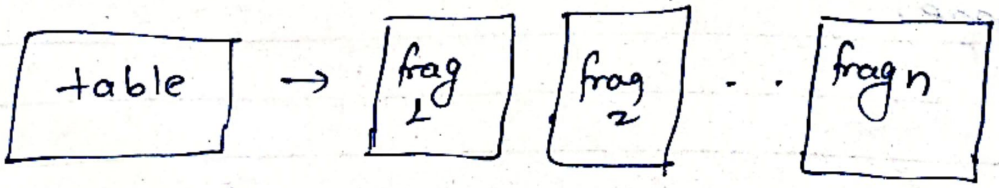
2. Discuss the data fragmentation, Replication and allocation techniques for distributed database design.

Data Fragmentation:
Fragmentation is the task of dividing a table into a set of smaller tables. Subset of tables are called fragments. These fragments may be stored at different locations Moreover, fragments increases parallelism and provides better disaster recovery.

Types
i) Vertical fragmentation
ii) Horizontal fragmentation
iii) Hybrid fragmentation

Fragmentation should be done in a way so that original table can be reconstructed from fragments.

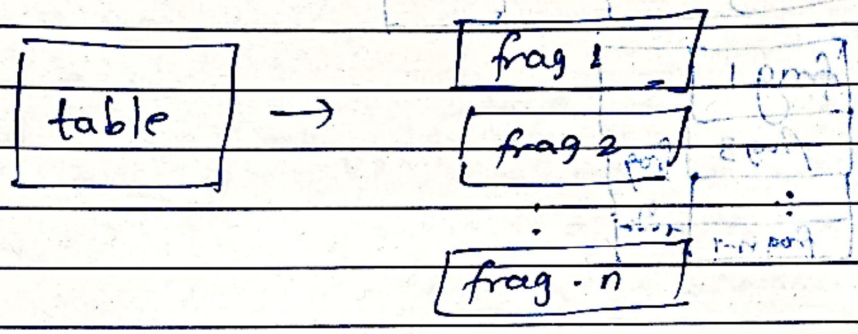
Vertical fragmentation:-
In vertical fragmentation, the fields or columns of table are grouped into fragments. In order to maintain reconstructiveness, each fragment should contain primary key field of table and it can be used to enforce privacy of data.



eg. CREATE TABLE std_address as SELECT stu_id, stu_Address FROM Student;

Horizontal fragmentation

- In horizontal fragmentation, the tuples of a table are grouped into fragments according to the values of one or more fields.
In order to maintain re-constructiveness, each fragment should contain all field(s) of table.



eg: CREATE TABLE Department AS
SELECT * FROM student where deptid = 1

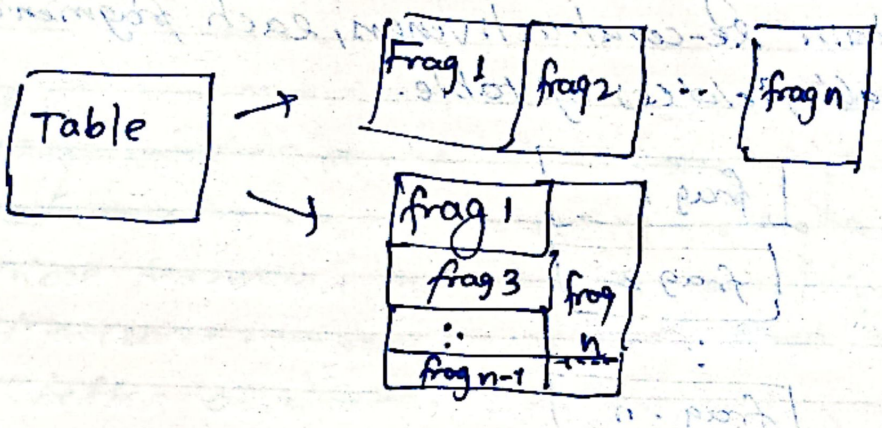
Hybrid fragmentation:

In hybrid fragmentation, a combination of horizontal and vertical fragmentation technique is used. This is most flexible fragmentation technique since it generates fragments with minimal extraneous information. Reconstruction of the original table is often an expensive task. Hybrid fragmentation can be done in two alternative ways:

i) At first, generate a set of horizontal fragments; then generate vertical fragments from one or more of the horizontal fragments

ii) At first, generate a set of vertical fragments, then generate horizontal fragments from one or more vertical fragments

eg: CREATE TABLE Hybrid AS
SELECT stuid, stu-name from student
WHERE stuid = 12;



Data Replication:

Data Replication is the process of generating and reproducing multiple copies of data at one or more sites. Replication is an important mechanism, because it enables organizations to provide users with access to current data where and when they need it. It is intended to increase the fault tolerance of a system such that if one database fails another can continue to serve queries or update requests.

Replication is sometimes described using the publishing industry metaphor of publishers, distributors and subscribers.

publisher:

A DBMS that makes data available to other locations through replication. The publishers can have one or more publications, each defining a logically related set of objects and data to replicate.

· Distributer:
A DBMS that stores replication data and metadata about the publication and in some cases acts as a queue for data moving from publishers to the subscriber. A DBMS can act as both publishers and distributor.

· Subscriber:
A subscriber can receive data from multiple publishers and publications. Depending on the type of replication chosen, the subscriber can also pass data changes back to publishers or republish data to other subscribers

Replication Purpose:
- It remove single point of failure by replicating data. (System availability)

Performance: Replication enables us to locate the data closer to their access point which will reduce response time.

Scalability: Replication allows for way to support growth geographically with acceptable response time

Application Requirements:
Replication may be dictated by application which may wish to maintain multiple copies of data.

Challenges

1. placement of replicas:
   major challenge in replication is where to put replicas.
   1. permanent replicas:
      consists of cluster of servers that may be geographically dispersed
   2. Server initiated replicas:
      placing replicas in hosting servers and server caches.
   3. Client initiated replicas include web browser cache.

2. propagation of updates among replicas:
   The net challenge is to how to propagate the updates in one replica among all the replicas efficiently and faster as possible.
   1. push based propagation: push updates to all other replicas
   2. pull based propagation: A replica requests another replica to send the newest data it has.

3. Lack of consistency:
   If a copy if modified, the copy becomes inconsistent from rest of copies. It takes sometime for all copies to be consistent.

Advantages:

i) Reliability because it contains available copy, incase of failure

ii) Due to multiple local copies it will reduce the N/w load.

iii) Due to available copies it has quicker response
iv) It is simpler in nature.

Disadvantage:
i) Increased storage requirements due to multiple copies
ii) Due to complex synchronization technique and protocal needs to update copies of data will increase cost and complexity of data updating
iii) Undesirable Application-Database coupling:
If complex update mechanism are not used, removing data inconsistency requires complex coordinate at application level.

Data Allocation:
each fragment or each copy of a fragment is stored at a particular site in the distributed system with 'optimal' distribution. This procen is called data distributon / alloca
The choice of sites and degree of replication depend on the performance and availability goals of the system and on the types and frequencies of transactions submitted at each site.
eg: If high availability is required: transactions can be submitted at any site, and most transactions are retriev only, a fully replicated database is a good choice.

There are four alternative strategies regarding placem of data:

i) Centralized:
consists of single database & DBMS stored at one site with user distributed ~~acce~~ across N/w.

ii) Fragmented:
partitions the database into disjoint fragments, with each fragment assigned to one site.

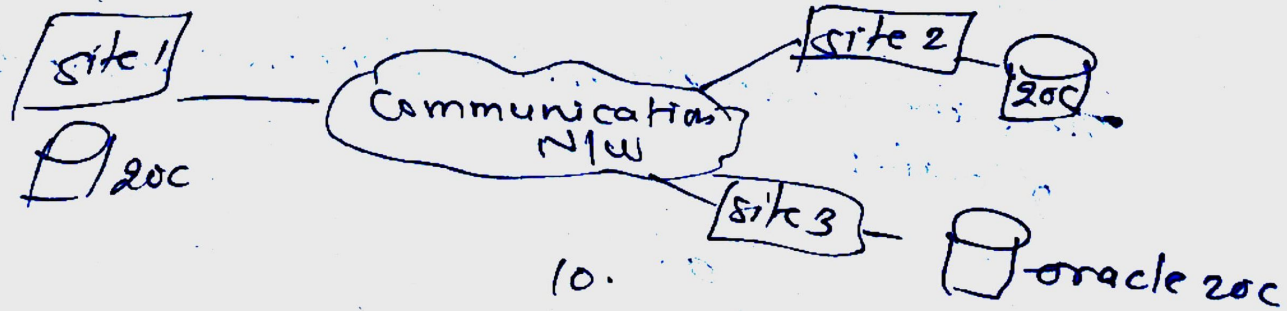
iii) complete Replication:
consists of maintaining a complete copy of database at each site

iv) Selective Replication
combination of fragmentation, replication and centralization. It has advantages of all other approach but none of disadvantages.

3. Explain the types of distributed database system.

i) Homogenous DDBS
In homogenous system all sites use the same DBMS product Homogenous system are much easier to design and manage. This approach provides incremental growth, making the additional of new site to DDBMS easy by allowing increase performance using parallel processing.

eg: Consider we have three sites using same oracle 20c to DBMS. If some changes are made in one department then it would update the other ~~de~~ site also



Types:

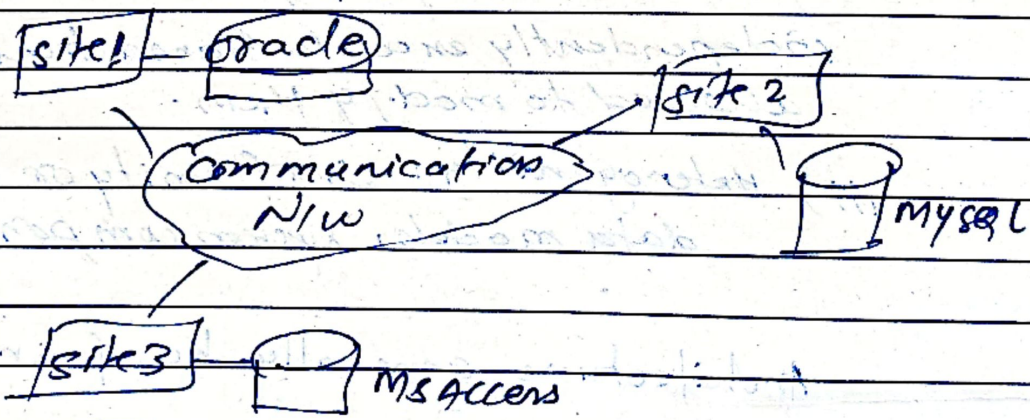
i) Autonomous: Each database is independent that functions on its own. They are integrated by a controlling application and use message to share data updates

ii) non-autonomous: Data is distributed across the homogenous node and a central DBMS coordinates data updates across sites.

2. Heterogenous DDBS

In a heterogenous system, sites may run different dbms products, which need not be based on same data model. Usually individual sites are implemented in their own database and integration is considered at last

eg:



If h/w is different but DBMS products are same, translation is straight forward and if DBMS are different translation is complicated

Types

i) Federated: independent in nature and integrates together as single database.

ii) Unfederated: central coordinating module is used to access database.

4. Explain the architecture of distributed database.

DDBS architecture are generally developed depending on three parameters:

i) Distribution: It states the physical distribution of data ~~access~~ across different sites.

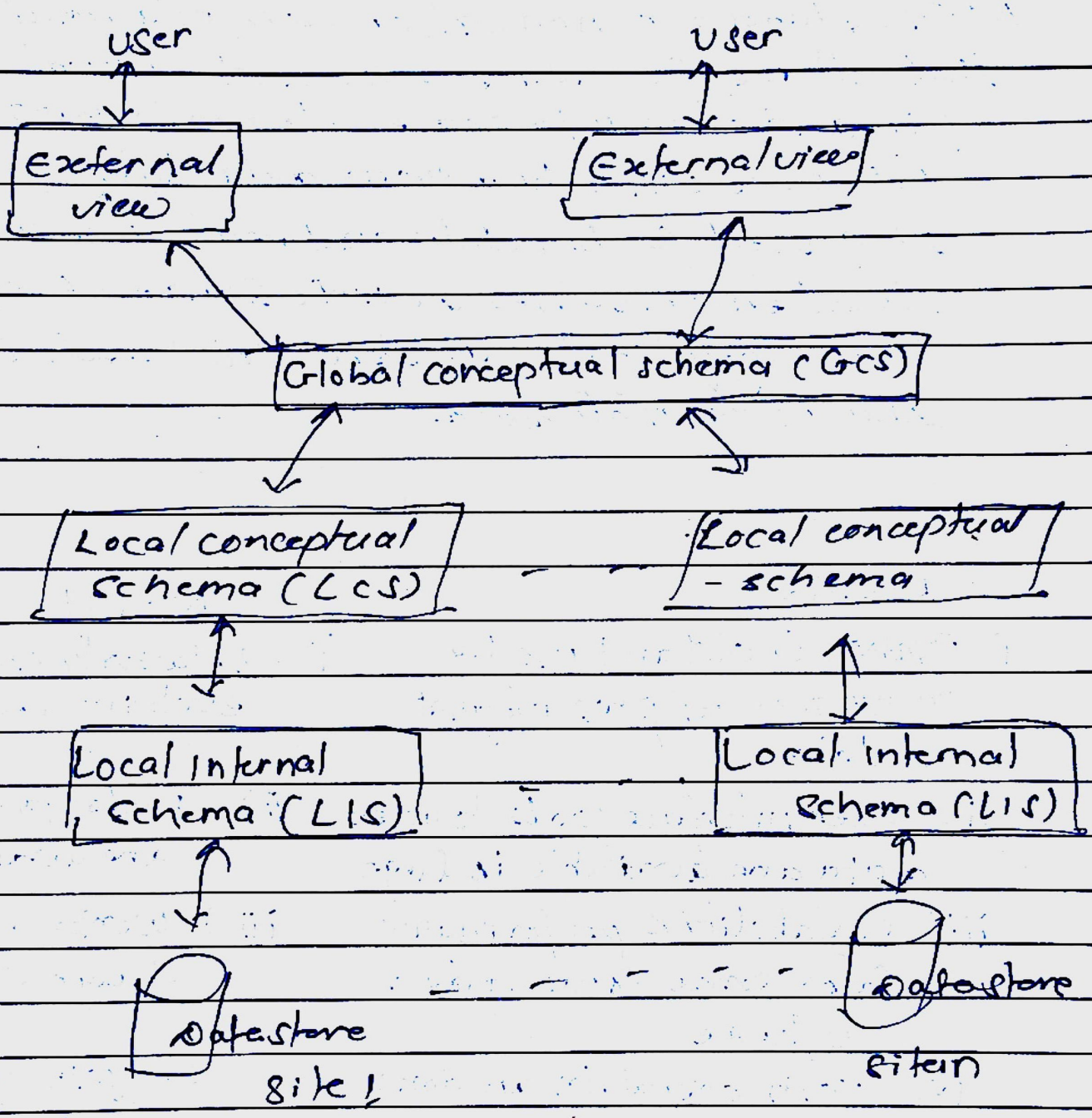
ii) Autonomy: It indicates the control (distribution) of database system. It is a function of a number of factors such as whether the component systems exchange information, whether they are independently execute transactions or whether one is allowed to modify them.

iii) Heterogeneity: Uniformity or dissimilarity of data models, system components and database.

Architecture generally has four level of schemas

- External view or schema (EV or ES): Depicts the user view of data.
- Global conceptual schema (GCS): Depicts the global logical view of data which provides NW transparacy
- Local conceptual schema (LCS): Depicts logical data organization at each site.

• Local Internal schema (LIS) : Depicts physical data organization at each site.



5. Describe nosql. Differentiate Nosql with RDBMS.

Nosql stand for non sql or non relational database that allows for data storage and retrieval. It avoids join and is easy to scale. Its main purpose is data store with humogous data storage needs. It is used for big data and real-time web apps. Nosql encompasses a wide range of database technology that can store structured, semi-structured and polymorphic data.

RDBMS	Nosql
i) RDBMS is old and use by many org. for proper format of data	i) Relatively new and evolving day by day
ii) User interface tools to access data are available is large	ii) tools to access data are available less
iii) Scalability & performance faces some issue if data is huge	iii) Works well in huge data
iv) Join operation are done	iv) Join operation can't be done
v) availability of data are mostly available	v) availability of data are highly available
vi) Documents can't be stored	vi) documents are stored
vii) eg Mysql, oracle etc.	vii) eg oracle Nosql, Apache HBase etc

6. Discuss different types of Nosql database.

i) Document Database:
objects are stored in document like JSON. Each document has a set of fields and values pair.
They can expand horizontally to accomodate enormous data volumes.

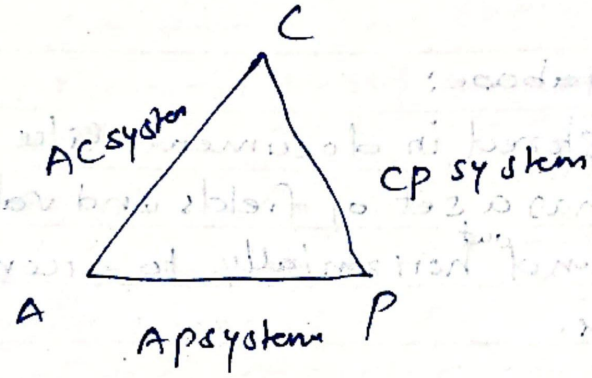
ii) key-value Database:
Simpler form of database that has key value for each item. It is usually straightforward because a value only be accessible by its key.

iii) Wide-column stores:
Similar to RDBMS but names and formats of columns can vary from row to row across table. It groups columns of related data together.

iv) Graph database:
Data is stored in nodes and edges in graph database. Edge hold information about relationship between nodes whereas nodes store information about people, locations and objects.

7. Describes CAP Theorem:
CAP theorem helps to understand the limitation of Nosql. CAP stands for consistency, Availability and partitioning. Nosql can't provide consistency and high availability together. CAP states that we can only achieve at most two out of three

guarantees for database - consistency, Availability and partition Tolerance.



consistency: every nodes of database has exactly same information at given time.

Availability: ability of database to be always be available no matter what happens

partitioning: having ~~abai~~ ability to support broken links within the cluster in database is distributed in.

CP database:
deliver consistency and partition tolerance at expense of availability. When a partition occurs between any two nodes, system has to be shutdown the non-consistent node until partition is resolved.

AP database: deliver availability and partition tolerance at expense of consistency.
When partition occurs, all nodes remains available.

CA database:- deliver consistency and availability in the absense of any N/W partition. often single nodes DB servers are categorized as CA system. Single node DB servers do not deals to with partitions tolerance.

8. Discuss Big data with its characteristics.

Big data is a collection of data that is huge in volume yet growing exponentially with time. It solves the issue of traditional databases. Big data can be structured or unstructured. Big data can be collected from publicly shared comment on social sites and website. Big data is most often stored in computer database and analyzed with software specifically designed to handle large, complex database/sets.

characteristics

i) Volume of Data:
It contains large volume of data which are in petabytes. Big data is vast 'volumes' of data generated daily from online and offline transactions

ii) Variety:
Big data can be structure, unstructured and semi-structured. Data may be in forms of images, text videos etc.

iii) Velocity:
velocity creates the speed by which data is created in real time. It contains online or offline data. It contains the linking of incoming data sets speeds rate of change and activity bursts.

iv) value: Business value of data collected

g. Write down pros and cons of big data.

Pros/Advantages

i) Voluminous collections:
large amount of market data can be generated using Big data analytics and various graphical and mathematical representation can be made for easy analysis

ii) Future insights:
With the prediction and analysis we can control business prospects.

iii) Big Data Analytics is cost effective so we can do market analysis.

iv) Research will take less time:
New s/w can easily analyze and interpret data sets, which helps make decisions and save lots of time.

v) Fraud Detection and prevention:
Big data is capable of stopping fraudulent transactions in banking, finance services.

Cons/disadvantages:

i) Unstructured data:
More variety of data can create difficulty in processing results and generating solutions.

ii) Security concerns:
For highly secured data or confidential information highly secured N/w are needed for its transfer and storage.

iii) Expensive :-
process of data generation and its analysis is costly without surety of favorable results.

iv) Need professionals or skilled manpower for analysis

v) H/w and storage :-
Requires high costly and hard to built server and h/w.

10. Describe Map reduce with its phases.
Map reduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. Map reduce is a Hadoop framework used for writting applications that can process vast amounts of data on large clusters. There are two primary task in Mapreduce : map and reduce. First map job is done and then reduce. In map job, the I/p dataset is split into chunks. and output of map job is used as input for reduce tasks. Map reduce framework enhances the scheduling and monitoring of tasks.

phases :-
(1) Mapping phase :-
There are two step in these phase : splitting and mapping A data set is split into two equal unit called chunks in splitting step which transform input into key-value pairs. and key-value pair are then used as input in mapping step. this is only a format that mapper can read. Mapping step contains a coding logic applied to these data blocks.

ii) shuffling phase
consists of two main steps: sorting & merging.
In sorting step, key-value pairs are sorted using keys and merging ensures that key value pairs are merged or combine. this phase removes the duplicates values and grouping of values.

iii) Reducer phase:
output of shuffling phase is used as input. It further reduces the intermediate values into smaller value
It provides summary of entire dataset

iv) Combiner phase:
optional phase that is used to optimizing Mapreduce process. It increased speed in shuffling phase by improving performance of jobs.

11. List out the benefits of map reduce
i) speed: process huge unstructed data in short time
ii) fault tolerance: handle failures
iii) cost-effective: process or store data in cost effective manner
iv) scalability: provides highly scalable framework. allows user to run application from many nodes.
v) Data availability: Replicas of data are sent to various nodes.
vi) parallel processing which reduce processing time.

12. Describe hadoop with its advantages.

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computer using simple programming models. It is designed to scale up from single servers to thousands of machine each offering local computation and storage. It is written in java and used from batch/offline processing.

Advantages:

1. It is highly scalable platform which will store and distribute very large data.

2. Hadoop offers a cost effective storage solution for businesses exploding data sets

3. Hadoop is flexible to access data and new data source i.e. structured or unstructured

4. Hadoop is fast which will be able to process terabytes of data in a minutes

5. Hadoop has its fault tolerance

6. Multiple language support

7. opensource

8. compatibility

OR


support me at
eSewa™
9810867824