Assignment No:- 2

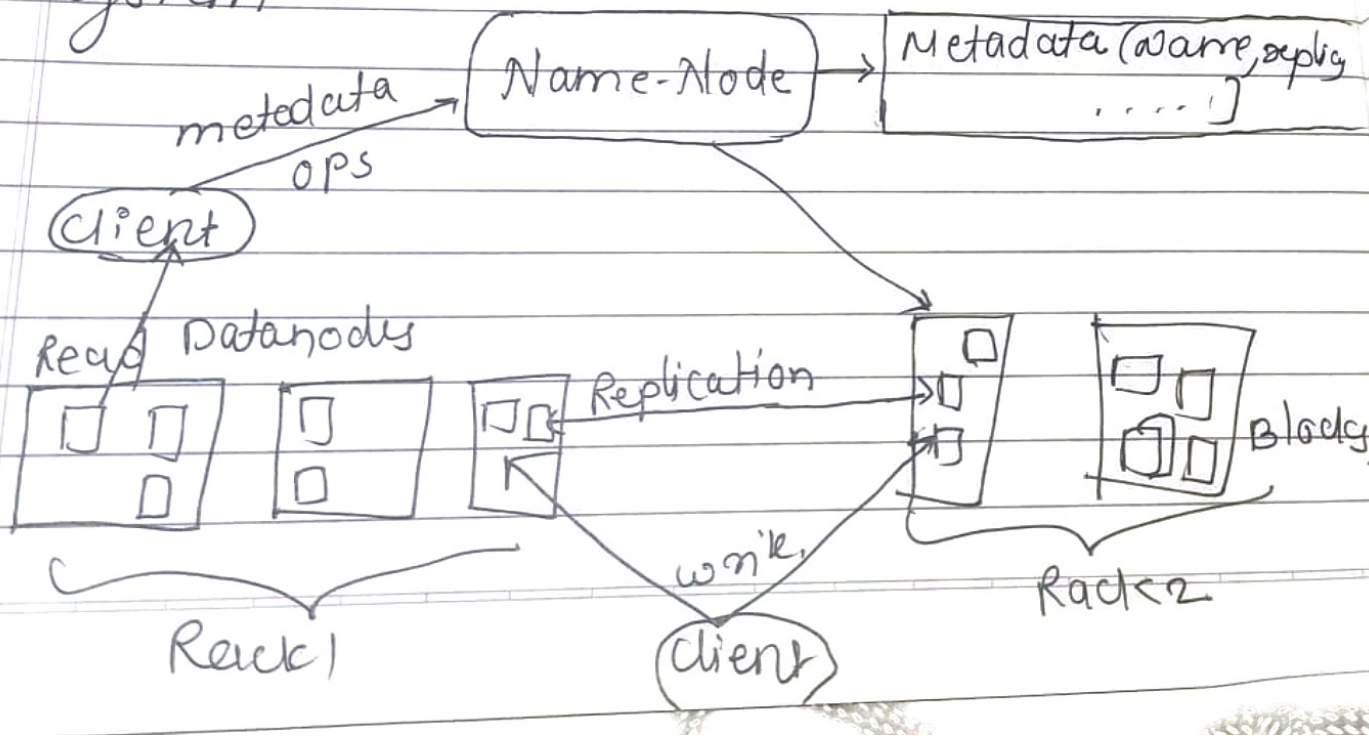
Q1] Write a note on cloude file system with Architecture

→. Cloud storage is the abstraction, pooling & sharing of store relources through the internet. file storing is the dominant technology used on NAS system & is responsible for organizing data & representing it to users. It is hierachical structure allow us to navigation data from top to bottom easily, but increase processing time.

- Cloud File storage is a storage service that is delivered over thet internet billed on a pay-per-use basis & has an architechure based on common file level protocols such Server Msg Block (SMB), common Internet File system (CIFS) & Network File system (NFS)
- Types of cloud file system.

i] Ghost File system:- Amazon Web services (AWS) use Ghost cloud file system (GFS). GFS run over Amazon's EC2, SimpleDB web service & S3

ii] Hadoop File system:- is a distributed file system inspired by GFS that organizes files & store their data on a distributed computing system



iii] Kosmos file system (KFS) :- is a open source project written in C++ by search startup kosmix. Three components are :- i] One or more chunk server that store the data on their own hard disks, b) A metaserver that keeps an eye on the chunks server & c) An appln that quickly gets rids of a single large file!

Q2] Discuss Hbase Data Model?

→ • It is a database that is open source platform & it is the implementation of storage architecture by Google's Big Table. HBase databases column-oriented thus it makes it unique from other data bases. One of the unique qualities of Hbse is it doesn't care about data types because we can store different data types of data for the same column in different rows.

• It contains different sets of tables that maintain the data in key-value format. Hbase is best suitable for sparse data sets which are very common in the case of big data. It can be used to managed structure & semi-structured data & it has many built-in features such as :- scalability, versioning, compression, garbage coll

• Two types of data storage medium :-

i] Row-oriented :- data is stored & retrived one row at a time. This cloud lead to serval problems. Suppose we want only some part of the data from the row but according to this approach you have to retrieve the complete row even if you don't need it.

- Column-Oriented :- data is stored & retrieved based on the columns. Thus the problem which ~~were~~ were facing in the case of the row-oriented approach has been solved because in the column-oriented approach we can fill out the data which is required to us from the whole set of data with the help of corresponding columns. OA the read & write oper^n are slower than other but it can be efficient while performing operations on the entire database hence it permits very high compression rate

column families

Row Key	personal data		professional data	
emp id	name	city	designation	salary
1	A	AAA	manager	₹150,000
2	B	BBB	sr. engg	₹70,000
3	C	CCC	Jr. engg	₹50,000

Q3) Draw & Explain cloud file system GFS?

→ • A file system in cloud is a hierarchical storage system that provides shared access file data. Users can create, delete, modify read & write file & can organize them logically in directory trees for intuitive access

- cloud file storage is most appropriate for unstructure data or semi-structured data, such as documents, spreadsheets, presentation & other file-base data
- CFS is a method for storing data in the cloud that provides servers & appln access the data through shared file system. This compatibility

makes cloud file storage ideal for workloads that rely on shared file systems & provides simple integration without code changes.

- Ghost File System:- It given high redundant elastic mountable, cost effective & standard based file system. A fully featured scalable & stable cloud file system is provided by GCFS.
- GFS run over Amazon's S3, EC2 & SimpleDB web services.
- When using GFS, user can have complete control of the data & can be accessed as a standard n/w disk drive. It is highly secure. It can be mounted on server, client or access file via web page.

Q4) Explain Working of cloud-Data storage?

→ Data store is a connection to a store of data, whether the data is stored in a database or in one or more files. The data store may be used as the source of data for a process.

- Datastore is a repository for storing, managing & distributing data sets on an enterprise level.
- Distributed Data store:- Computer n/w where information is stored on more than one node by means of data replication it is termed as DDS. It is used to refer to distributed db where user store infr on no. of nodes or a computer n/w in which store information on no. of peer n/w nodes.

- DDS is one in which file, scripts & images are stored in more than on server or volumes rather than a single server as in traditional system
- Eg:- Google's Big Table, Amazon's Dynamo
- These types of data store are non relational databases that searches data quickly over a large multiple nodes.
- Data store Types

i] Big table:- is a distributed storage system that is used for managing & storing structure data at Google Bigtable is designed to reliably scale to petabytes of data & thousands of Machine Bigtable has multiple goals like applicability, high availability, scalability, high performance. Bigtable is build on Google file system for storing the data for scheduling large scale data processing. It stored data in form of rows, columns & timestamp that means it maps with arbitrary string value like row key & column key as well as timestamp

ii] Dynamo:- is a property key value structured storage system. It can act as database & also distributed hash table. Dynamo dynamically partitions a set of keys over a set of storage nodes. It is most powerful relational database available in WWW. Dynamo does not support replication. It is used to manage the state of service that have very high reliability requirements & need tight control over the trade-off b/w availability consistency, cost effective & performance

5] How does cloud file System are different from Normal file system.

	Cloud file storage	Traditional File storage
i]	stored on remote server in the cloud	• Typically stored on local hard drives
ii]	Accessible from anywhere with internt	• limited to local n/w or device
iii]	Highly scalable, can easily scale up or down based on demand.	• Limited scalability, capacity tied to hardware.
iv]	Build-in redundancy through data replication & backups.	• Relies on manual backup or RAID for redundancy
v]	pay-as-you-go mode based on storage usage	• Upfront costs for hardware & maintaunance.
vi]	Managed & maintained by cloud service provider	• Requires manual management & updates.
vii]	Build-in collaboration features such as shared folder & access controls	• Collaboration may require complex setups.
viii]	Often includes robust security measures provided by the cloud provider.	• security features depend on local configurations.

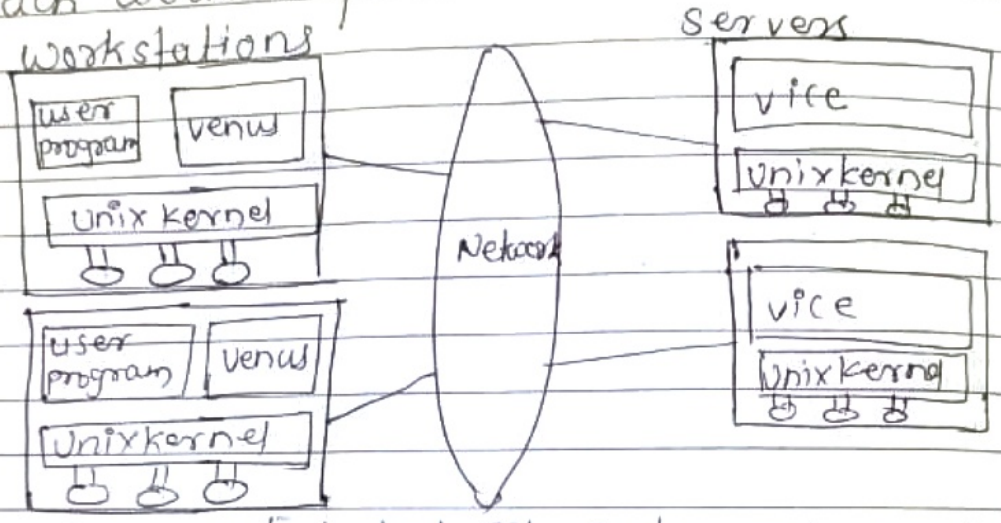
Q6) Explain AFS Architecture?

→ • Andrew File System Architecture.

i] Vice :- AFS provides a homogeneous, location-transparent file namespace to all client workstation by utilizing a group of trustworthy servers known as vice.

ii] Venus :- caches files from vice & returns updated versions of those file to those file to the servers from which they originated.

- File system architecture was largely inspired by the need for scalability. To increase the no. of clients a server can service. Venus performs as much work as possible rather than Vice.



- AFS is a distributed file system. It uses the client server model, where all the files are stored on file server machines.

97) Difference b/w SAN & NAS.

SAN	NAS
i] storage area n/w	• Network Attached storage
ii] In SAN data is identified by disk block	• In NAS data is identified by file name as well as byte offset.
• In SAN file system is managed by servers	• File system is managed by Head Unit
• It is more costly & it is more complex	• It is less expensive. • It is less complex than SAN
• Protocol used in SAN are SCSI, SATA	• Protocol :- File servers, CIFS (common Internet FS)
• Suitable for that environment which has high speed traffic.	• Not suitable for that environment which has high speed traffic

- Has lower latency
- It support virtualizat^n

- has higher latency
- Do not support virtualizat^n

Q8) Elaborates HDFS Architecture in details.

→ • Hadoop File System:- Hadoop is an open source software framework that supports data-intensive distributed appl^n, licensed under the Apache v2 licence. It provide software framework for distributed processing of large datasets in real-time appl^n.

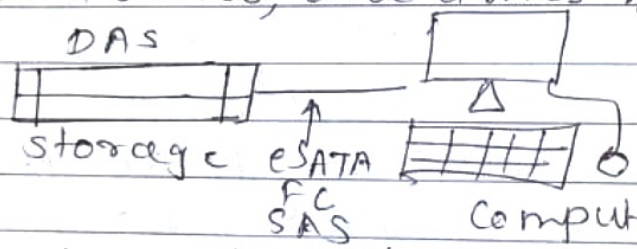
- Hadop Distributed file System is a block-structured file system where each file is divided into blocks of pre-determined size. These blocks are stored across a cluster of one or several machines.
- Blocks are the nothing but the smallest continuous location on your hard drive where data is stored. Similarly, HDFS store each file as blocks which are scattered throughout the Apache Hadoop cluster.
- It is the primary storage system used by Hadoop applicatiah. HDFS operates as a distributed file system designed to run on commodity h/w.

Q9] Explain the Data storage (EDS, DAS, SAN, NAS)

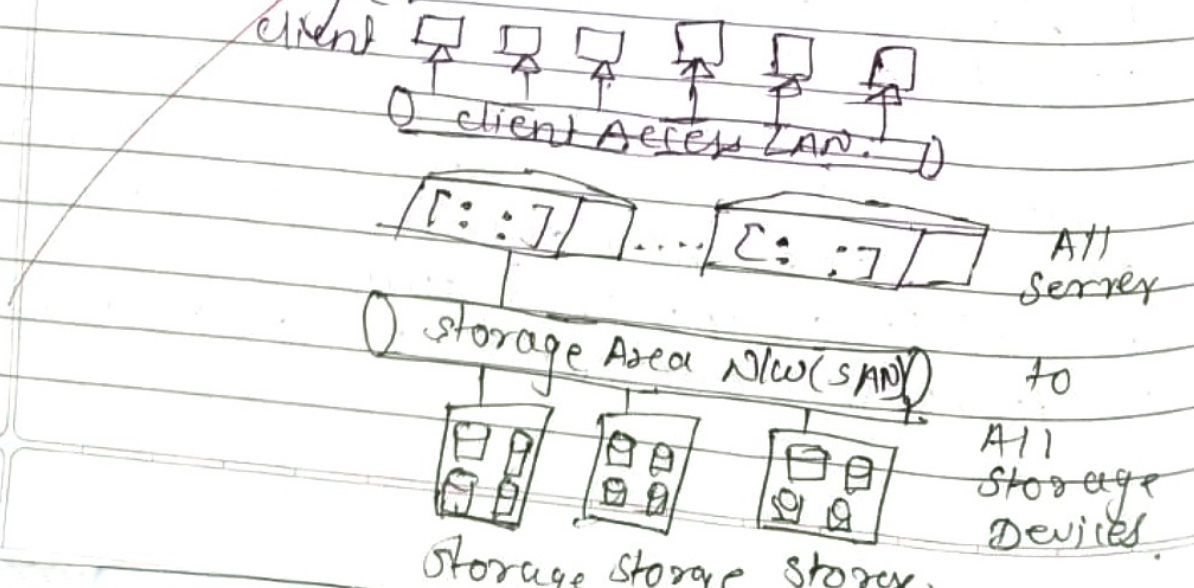
→ i] EDS (Enterprise Data storage) :- Enterprise storage is a centralized repository for business-critical information that provides data sharing, data management & data protection across multiple computer system.

- Enterprise storage may include a SAN, NAS, DAS.

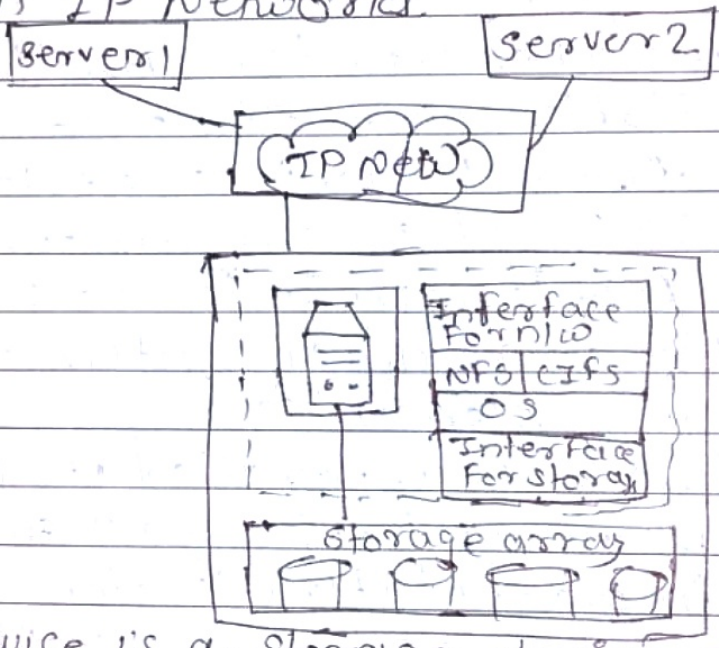
- Direct Attached storage is hard drisk drives or solid-state drives connected directly inside or outside to a single computer or server that cannot be accessed by other computer or server. DAS is not networked through Ethernet or FC switches
- Storage device may includes one or more drives built into a server or with an appropriate host bus adapter, may be configured as RAID array. Eg:- hard drives, disc drives & storage on external



- Storge Area Network :- is a dedicate high-performance n/w or subnetwork dedicated to storage that is independent of an organization's common user Network.
- It interconnects pools of disk or solid-state storage & shares it to multiple servers so each one can access data as if it was directly attached. A SAN is a dedicated n/w that provides access to consolidated, block level data storage.



- Network-attached storage :- is a file level computer data storage server connected to a computer n/w providing data access to a heterogeneous group of clients. It is specialization for serving file either by it's hardware, s/w or configuration.
- NAS devices are storage arrays or gateways that support file-based storage protocols such as NFS & CIFS & are typically connected via an IP Networks.



- NAS device is a storage device conneded to a n/w that allow storage & retrieval of data from central location for authorised n/w user & varied clients.

Q10) Explain the Data Intensive Technologies for cloud computing.

→ • DI system encompass terabytes to petabytes of data. Such systems require massive storage & intensive computational power in order to execute complex queries & generate timely results.

- Data Intensive computing is "a class of parallel computing applications which use a data parallel approach to processing large volumes of data"

1) Processing approach :- current data-intensive computing platforms use a "divide & conquer" parallel processing approach combining multiple processors & disks in large computing clusters connected using high-speed comm^n n/w.
   - It allows the data to be partitioned among the available computing resources & process independently to achieve performance & scalability based on the amount of data.

2) System Architecture :- For data-intensive computing an array of system architectures have been implemented.

a) MapReduce :- is a parallel programming model proposed by Google & available open source implementation known as Hadoop. It aims at supporting distributed computation on large datasets by using a large No. of computers with scalability & fault tolerance guarantees.
   - Map & reduces are two primitives in functional programming languages, such as Lisp, Haskell etc. A map function processes a fragment of a key-value pairs list to generate a list of intermediate key-value pairs.
   - Reduce function merges all intermediates values associated with a same key & produces a list of key-value pairs as output.

b) High Performance computing cluster (HPCC) also known as DAS (Data Analytics Supp computer) is an open sources, data-inten- sive computing system platform developed by LexisNexis Risk Soln. DAS is platform designed to refine, link & fuse large amou of data from disparate sources for complex analysis & queries.

- Architecturally, DAS is a HPCC based on comm dity server hw, which can be scaled up to thousands of processors to handle any amount of data & runs on the Linux O.S.
- Custom system software & middleware parts were created & layered to provide th execution environment & distributed file system support that is essential data-intensive computing on the base of Linux O.S.

Q/