Sharding - type of database partitioning that seprates large database into smaller & faster to manage parts.

Big table -

distributed, highly scalable & no SQL database system developed by goggle.

designed to handle massive ammount of structured data accross large no. of servers

Big table provides → ~~sorted map~~, multidimensional distributed and sorted map

Data is organized based on (row key, column key & timestamp).

It is optimized for handling large scale workloads & widely used for web indexing, data analysis & time series data storage.

key features →

① scalability → petabytes, (distributed over 1000+ machine) allows for horizontal scaling.

② High Performance

③ Automatic sharding* → BT automatically shards data accross servers for efficent storage & retrival.

④ Replication

⑤ Integrated with GCP -
allows seemless processing & analysis accross BigQuery, DataFlow, Dataproc.

Overall, BT is flexible & scalable DB for large-scale

col-based Data base - stores & retrives data by columns rather than rows.

Camlin Page
Date / /

HBase-

open source, distributed, column-oriented DB built on top of Apache hadoop ecosystem

modeled after BT & provides- scalability, fault tolerance & high performance.

It is used for handling large volume of structured & semi-structured data.

key features-

① column oriented - organizes data into column families for efficient storage & retrival of specific columns of data rather than entire rows.

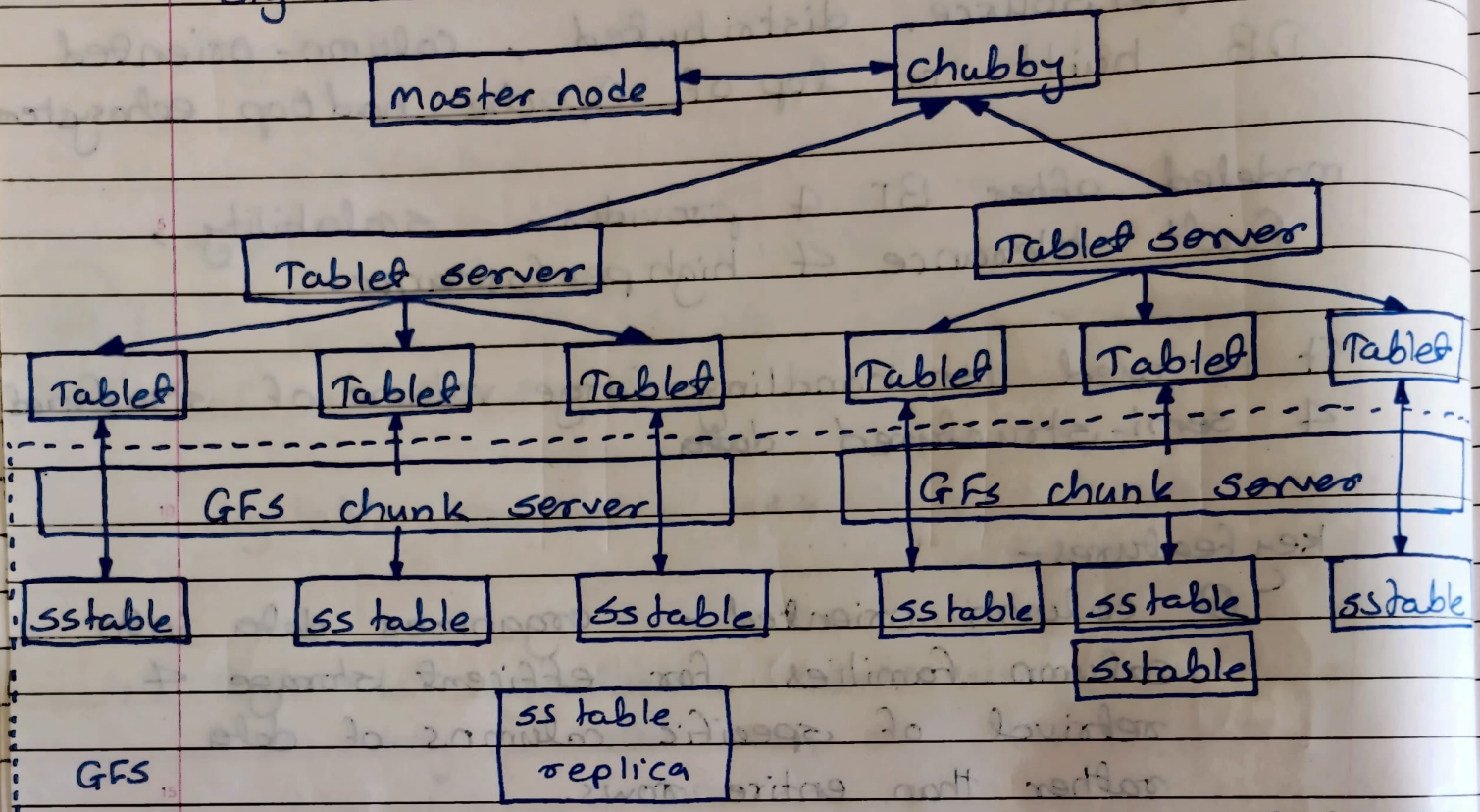
② Horizontal scaling- (can add more servers)

③ High Availability & fault-Tolerance- (Because of replication)

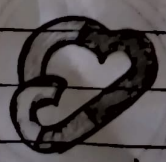
④ Strong Consistency - (for read/write ops)

⑤ Integration with Hadoop Ecbsystem - HDFs, mapreduce, Hive, for eff processing & analysis.

Big Table arch-



- SS table- file format store ordered immutable map from key to value. (binary search can be applie
- Master node- assign tablets to tablet servers, addition & expiration detection of tablet load balancing of tablet servers garbage collection in GFS.
- Tablet server- • manages set of tablets
  • handles read & write req to tablets.
  • splits the tablets that have grown too large.



HTF this arch works.

Working –

- Chubby is used to keep track of tablet servers.

① When a tablet server starts, it creates a uniquely named file in specific chubby directory & acquires a lock on it.

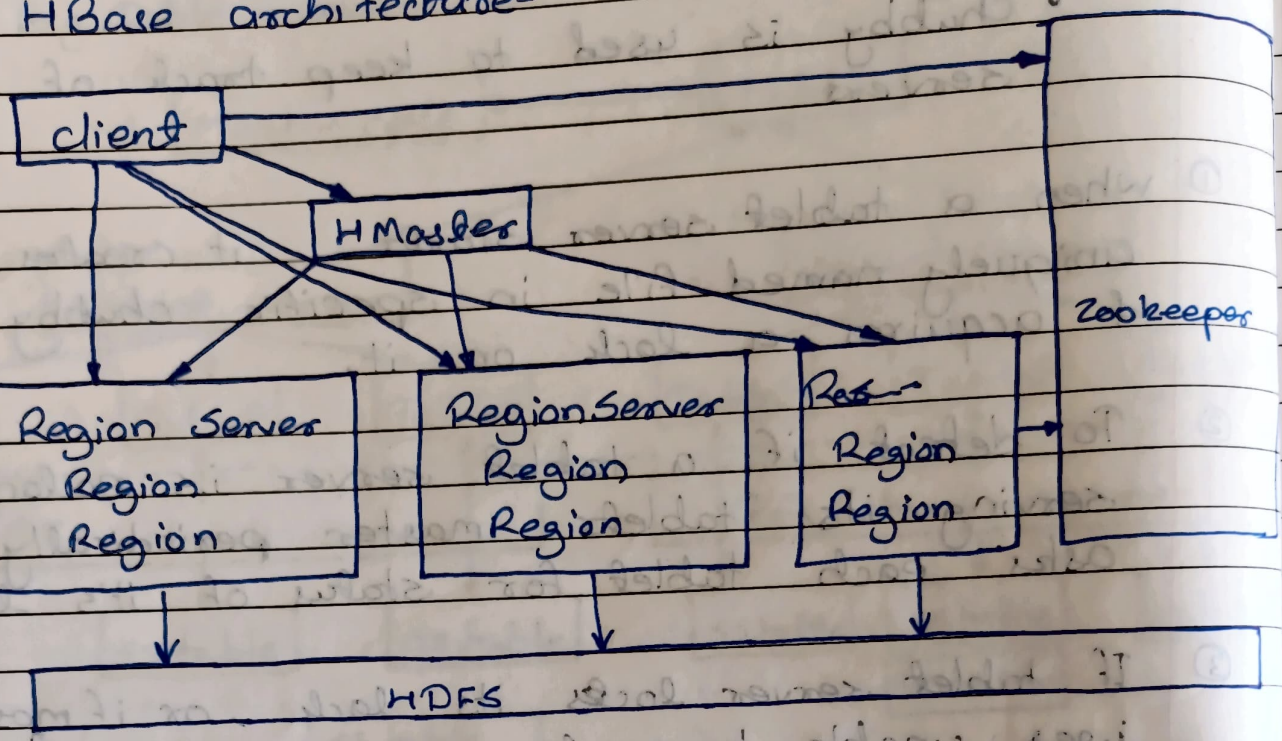
② To detect if a tablet server is no longer serving its tablets master periodically asks each tablet for status of it's locks.

③ If tablet server loses its lock, or if master was unable to reach server then it tries to acquire ~~lock externally.~~ exclusive lock on server.

④ If master is able to acquire lock ∴ Chubby is live & server is dead.
∴ assigns new server to those tablets.
& deletes the dead server.

HDFS	HBase
① File sys For storing large files	① database built on HDFS
② no fast individual record lookup	② it has ✓
③ only sequential access to data	③ Hash table for random acce
④ for high latency ops	④ for low latency ops.
⑤ data accessed through Mapreduce Jobs	⑤ provides access to single col from billion of rec
⑥ No random read & write	⑥ can perform

HBase architecture-



- Zookeeper- centralized monitoring server
  maintains config & info
  provides distributed synchronization

If client want to communicate with regions server, client has to ~~communicate~~ approach Zookeeper.

- Hmaster- for administrative operations.
  of the cluster.

- HRegions Servers -
  1. Hosting & managing regions
  2. Splitting regions automatically
  3. Handling read & write requests.
  4. Communicating with client directly.

- HRegions- maintain a store for each col family
  They are designed to accomodate semi-structured
  data that could vary in field size,
  data types & columns.

Dynamo - is key value ~~pair~~ structured storage system. It can act as database & distributed hash table.

It was created to help scalability issues that amazon faced during holiday seasons. & then it was used in AWS.

DynamoDB is built on the principles of Dynamo. but Dynamo is based on leaderless replication, DynamoDB is based on single-leader replication.

Principles -

- Incremental scalability - should be able to scale out one node at a time.
- Symmetry - every node should have same set of responsibilities
- Decentralization - design should favor peer-to-peer techniques.
- Heterogeneity - system should be able to exploit heterogeneity in the infrastructure.

work distribution should be proportional to capacity of the node.
∴ new nodes with higher capacity can be added

The index layer of S3 implements core features of Dynamo.

It also inspired many NoSQL databases implementations such as Apache Cassandra, Project Voldemort & Riak.

cool names ha!!

simple DB has been departed by AWS and users are encouraged to use DynamoDB as replacement (in June 30, 2019) just after our syllabus updated.

Google cloud Datastore	Amazon SimpleDB
provided by GCP	provided by AWS
NoSQL document database service	NoSQL Database service (now depricated)
Highly scalable & designed for large ammount of structured & semistructured data.	Designed for small scale apps with lower throughput requirements.
offers strong consistancy for read & write oprations.	Provides eventual consistancy.
Supports schema-less data model with entities and properties.	offers only schema-less data models.
enable complex query with filtering, sorting & projection.	supports bacic queries.
ACID-compliant transaction for data integrity	No built-in support for transactions.
Automatic distribution of data accross multiple servers for scalability.	Limited scalability compared to Datastore.
Integration with GCP	Integration with AWS.

geo-redundancy- distribution of file in servers accross multiple data centres that reside in diff geographical regions.

(Storage)

Cloud storage- data accessed over the internet

it is buit on virtualization techniques.

advantages- reduce cost, simplify IT mangent improve user experiance, remote collab. can be accesed via web service API, cloud storage gateway or web-based UI.

any device anywhere access, backups, easy file sharing.

Cloud storage Providers-

- Amazon S3- objeet storage service, scalable & durable storage for various type of data. offers features like encryption, versioning & lifecycle management.

common in all

- Google Cloud Storage- scalable, highly available objeet storage service by GCP. integration with Gcloud services.

- Microsoft Azure Blob Storage- optimized for storing large ammount of unstructured data integration with Azure services.

- Dropbox- offers personal & bussiness plans provides file synchronization, sharing & calabration accross multiple devices.

- IBM Cloud objeet Storage- for storing & managing large ammount of unstructured data.

features- geo-redundancy, integration with IBMCS.

Camlin Page
Date / /

Securing the Cloud.

General security advantages of cloud based solutions.

Cloud based solutions offer several security advantages compared to traditional on premises systems.

① Data Protection & Redundancy -
cloud provider implement robust data protection, data encryption, access control, regular data backups.

They also have redundant storage sys & geographically dispersed data centres.

② Scalable Secure Infrastructure -
teams of security experts to monitor, regularly update & patch their sys.

③ Access Control & Identity management -
multifactor authentication,
role based access control,
OAuth.
∴ reduce risk of unauthorized access.

④ Physical security measures -
surveillance system,
environmental safe guards.

⑤ Compliance & auditing -
cloud providers adhere to industry standards ISO 27001, SOC 2, HIPAA, & GDPR.

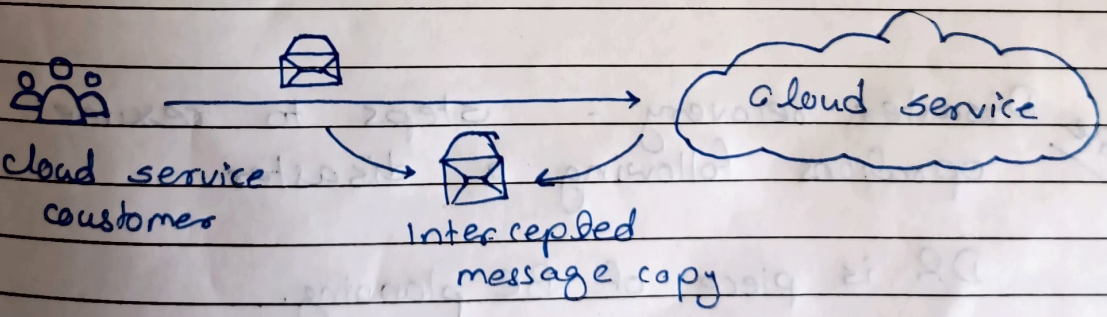
⑥ Automatic update & patching-
reduces the burden of organizations.
& helps to ensure systems are protected against known vulnerabilities.

⑦ Security Monitoring & incident Response-

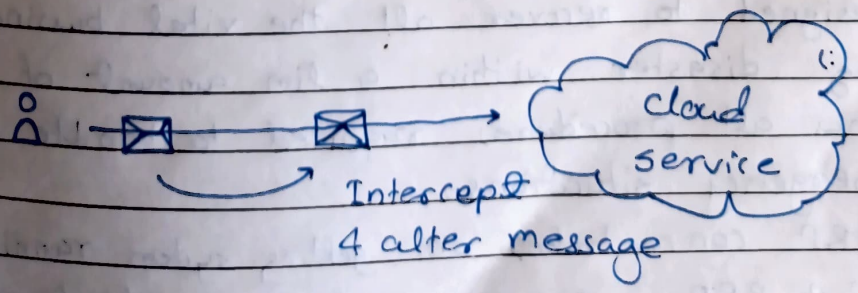
cloud providers employ advanced threat detection systems, perform security incident response and provide visibility into security events, allowing organizations to respond to potential threats more effectively.

- cloud security Threats-

① Traffic eavesdroping

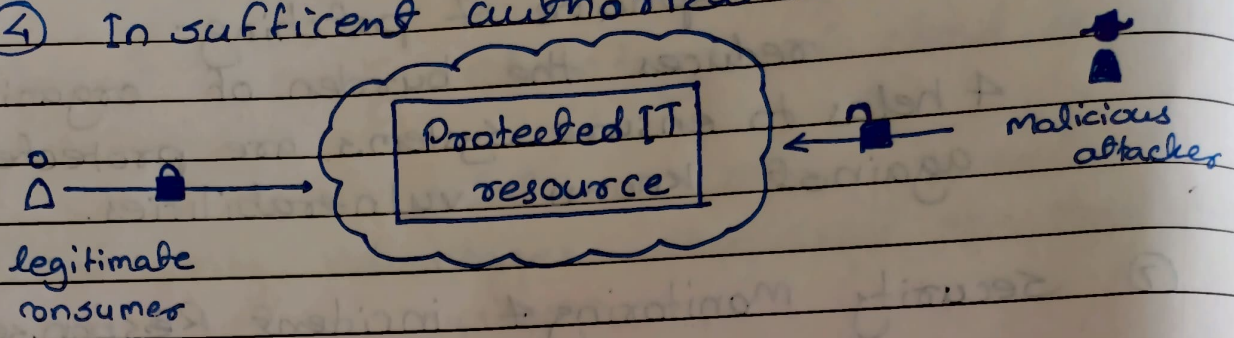


② Malicious intermediary



③ Denial of service- overloading resources so that resources can hardly be allocated to others.

④ Insufficient authorization.



⑤ Virtualization attack.

- Business Continuity & disaster Recovery.

Proactive

Business Continuity refers to process to ensure that critical function can work during & after a disaster.

Involves planing towards long term challanges.

Reactive

Disaster recovery - steps to resume operations following a disaster.

DR is piece of BC planning.

- Disaster recovery plan -
  - designed to recover all the vital business process during disaster within a lim amount of time.
  - It has all procedures required to handle the emergency situations.
  - DRP concentrates on getting system running asap
  - RTO & RPO recovery ~~point~~ time objective and recovery ~~time~~ Point objective are targets of DRP.

- most successful disaster recovery plan is the one which will never be implemented (risk avoidance)

Bussiness continuity plan -

- activites required to keep org running during period of displacement or interruption.
- BCP helps in continuing bussiness even after disaster.
- Bussiness has to stay active otherwise orgs will eriperience losses.
- legal issues can arise if critical services are not provided to clients.
- here effecient BCP is required to actively run and maintain bussiness activites.

Understanding the Threats -

- First step for perparing bussiness for dissaster.

Threats : -

1. Natural & localized disasters, earthquakes, floods, tornadoes & fires.

2. Failure of IT system - network, file servers, software applience.

3. Power outages, such as utility failures.

A BCP takes such things in account. &
It also takes in account tangible & non-tangible costs your bussiness may need.