DCC-Chapter 04

HDFS - Hadoop Distributed file System

For it to be fault-tolerant:

1> Block Replication.

→ Input split is repicated to multiple data node.

→ If in datnode the split is corrupted it informs the namenode then the next copy of the same input split is fetched and the mappers starts working on the copy.

→ Max. of 3 copies of each block.

→ 3 copies of same split.

→ one copy in one data node in one rackspace.

Second " " another " " in same as first rack space as it is easy to fetch

→ Third copy in another data node in other rack space.

→ Rack Space → is huge serve with multiple racks (nodes)

2> Replica placement.

3> Heartbeat and blockreport message

↓ ↓

Name node polling input splits in data node.

Namenode indicated about node corruption

Namenode - has metadata (location of datnode, structure of datanode, size of data node)

HDFS - configured by user.

HDFS Architecture.

→ 2 layers. MapReduce Engine over HDFS, workes.
Master → Job tracker (in MRE) ⇉ (tasktrackers).
→ Namenode. (in HDFS)

→ A cluster made up of racks. In cluster there will be a master at each layers.
→ Multiple nodes.

→ If something happens in task tracker message is send to Job tracker via block report message.
→ Job tracker can sense heartbeat of task tracker.
→ Each mapper output is given to reducer.
→ How many splits in data node = No. of mappers.

→ HDFS is storage manager.
→ Distributed and parallel programming paradigm

Dataflow.

GFS
Masters and clients   Searching done by multiple workers.
↓
where
data is kept
→ Next copy
→ Next split

(most, 4) - (most, 4)
(people, 6) - (most, 4)
(ignore, 6) - (most, 4)
(most, 4) - (most, 4) } 4
(poetry, 6) - (people, 6)
(people, 6)
(most, 4) -
(poetry, 6) - (ignore, 6) } 5
(ignores, 7) (poetry, 6)
(most, 4) - (poetry 16)
(people, 6) -
(ignores, 7)] 1

most, 4, 4
people, ignore, poetry, 6, (
ignores, 7, 1

DCC Chapter -05

Cloud Resource Management & Services.

RM → functionality
RM → performance
RM → cost

Master node takes care of scheduling jobs and has global state info such as no. of nodes, vms in each node, the memory allocated to each node.

policies → principles guiding decisions.
mechanisms → the means to implement policies

1) Admission Control
→ Given to the system that how many users has to be admitted on the cloud at a point in time.
Depending workload the no. changes.
Based on global state of info.

2) Capacity Control
→ what or how much resources has to be allocated, where it has to be allocated based on user requirement.

3) Load Balancing & Energy Optimization.
→ distribute loads equally on VM's.
→ CMS (central management system).
→ With energy optimization. No node is under/overloaded.

80, 60, 40, 20
100, 100, off, off
} 2 VM's are shutdown ∴ saving energy

→ least no. of servers to serve user.

4) Quality of Service (QoS) guarantees.

→ SLA (service level agreement) mutual agreement b/w user and CSP.

→ CMS should abide by SLA.

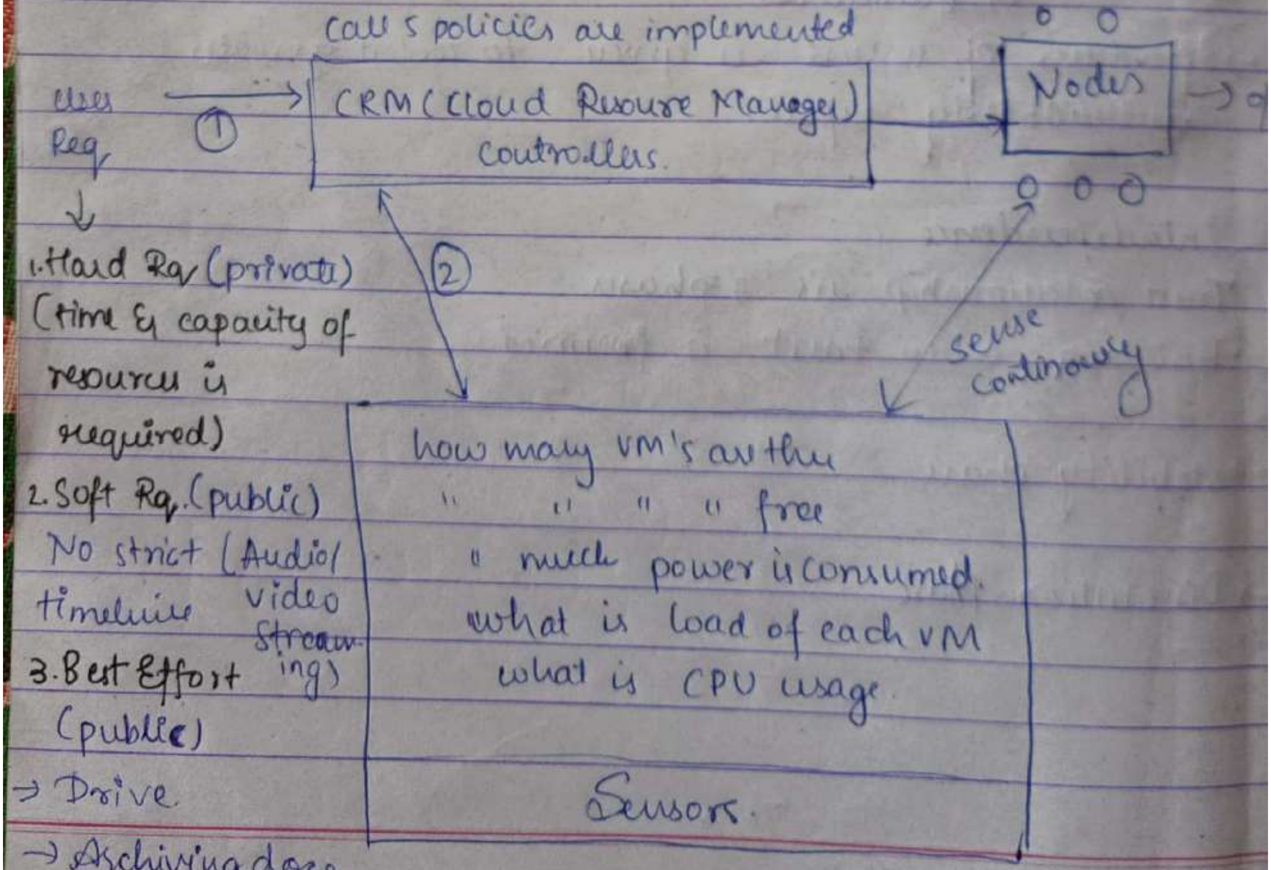
Mechanisms.

1. Control Theory → Closed loop (feedback from o/p) / Open loop (no " " ")
2. Utility Based → Platinum, Gold, Normal User type priority is given to the user.
3. Market Oriented.
4. ML Based

Control Theory

The feed back from o/p should go to controller so that desired o/p is obtained. Controller can be improved/ modified on the go.

Ex: Bread & Timer.



User Req
↓

1. Hard Rq (private) (time & capacity of resources is required)
2. Soft Rq. (public) No strict timeline (Audio/video Streaming)
3. Best Effort (public)

→ Drive

→ Archiving docs it takes time.

→ No critical timeline/capacity

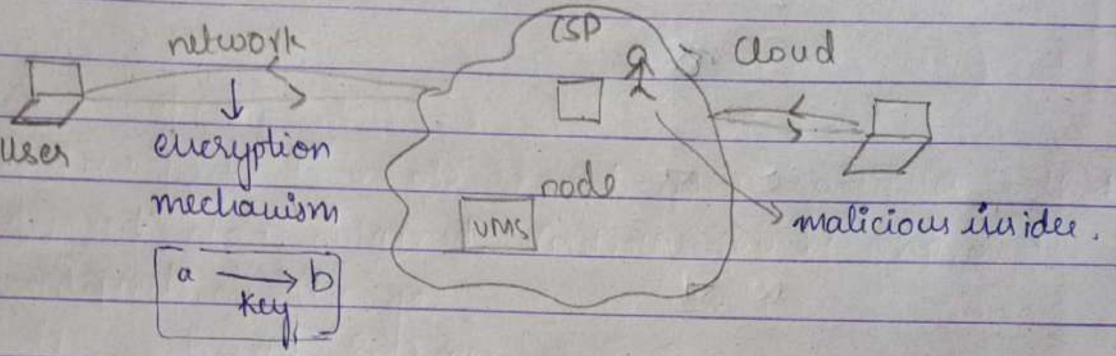
DCC - Chapter 06.

→ The security isn't wrt. single system it must be at all levels User level, Network level, CSP level.

→ In CSP there can be malicious insider also.

→ Who can block the service to legit user by flooding ping request (DoS attacks).

→ It is an attempt to not allow unauthorized user to acess or modify the data (either from network, data) Unauthorized user should not tamper the data while transitting b/w CSP and user and vice-versa



→ Integrity, Confidentiality, Availability & Privacy

authentication (login & pwd) & authorization (role/permissions)

→ user's data can't be exposed to another.

→ Proper isolation among processes running on VM's & b/w VM's.

→ Network layer
- ↳ Hashing (ensures data is not tampered)
- ↳ Encryption (ensures security while transiting)

→ User level
- ↳ firewalls (act as filters)
- ↳ authentication & authorizations
- ↳ IDS (Intrusion detection System) like Antiviruses

→ CSP level
- ↳ firewalls
- " ")

At Cloud level.

→ They outsource computation / data on third party.

→ So CSP should ensure the security the user's data on third party resource.

→ Multi Tenancy: Multiple process in multiple VM's.

→ VM vulnerability.

Threat Modelling

→ Model helps to analyse security problems & what are different strategies to mitigate the problem & evaluate the solution

→ After analysing different security threats we rank them

→ After ranking we give strategy

→ After that solution is implemented

→ Two type
  ↳ insiders
  ↳ outsiders.

→ If address of service is given to tool it generates vulnerability report.

Interdependence

Trust relationship in 3-phases

→ Building - when trust is formed

→ stability phase

→ Dissolution phase