S. No.	Date	Title	Page No.	Teacher's Sign / Remarks
		AWS		
		1) Cloud Introduction		
		2) EC2		
		3 AMI		
		4 Load balancer		
		5 Cloudwatch		
		6 SNS		
		7 Auto scaling Groups		
		8 EBS		
		9 S3		
		10 Cloudfront		
		11 IAM		
		12 SES		
		13 RDS		
		awscli		
		14 cloudformation (awscli)		
		15 Route53		
		16. (λ) lambda function		
		17 Elastic Beanstalk		

Scanned with ACE Scanner

(*) what is cloud?

→ Without buying the hardware, if a customer can use H/w resources of vendor, on demand of customer, can access globaly through internet is called as cloud.

eg:- google drive

→ The cloud is all about "pay for what you use"

→ In the case of google drive we are not buying the hardware of googledrive We are just utilizing the H/w of google drive by just creating an account, upto 15GB is free, beyond that you have to pay;

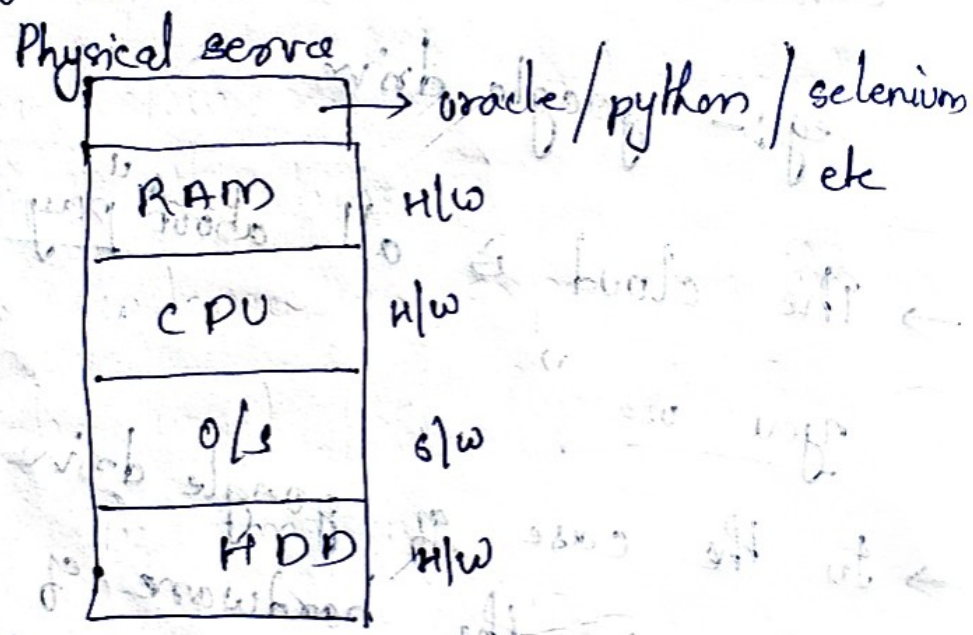
→ Google drive is a storage service of vendor (google)

Scanned with ACE Scanner

Server:-

→ It is serving something (an applicati to you is called server. It will take a request from you and process it and give response to you.

H/w peripherals



→ If it is oracle, it is called database server.

→ If it is python, it is called development server

→ If it is selenium, it is called testing server.

Scanned with ACE Scanner

→ The purpose of the s/w you use on it will define the server of that purpose.

Physical Server in real time

→ It needs atleast 64 GB of RAM, it is not free, let's say 50K rupees
→ CPU → atleast 2-3 lakhs
→ o/s → software licencing → 2-3 lakhs
→ HDD → 50K
→ oracle → 3 lakhs for enterprise addition of database
→ We are spending lakhs of rupees to setup one server. for our business.
→ This total process is called H/W procurement (buying / getting)
→ It will be taking more than 2-3 weeks of time to setup a realtime server.

Scanned with ACE Scanner

→ To install it we need different teams in our organisation,

Let's say,

We setup a server for flipcart which can proces 100 usrs per second.

→ On certain day of flipcart sale, 175 usrs are hitting the server, the server becomes slow or it won't respond it is called downtime,

→ If the customer is experiencing the down time, he won't use flipcart, this is loss to the company.

→ He has to maintain the no. of servers depending the workload,

→ let's say on the flipcart sale day he needs 100 servers and remaining days he needs only 10 servers, In this case 90 servers are sitting are idle

Scanned with ACE Scanner

→ These all factors makes the loss to the company. because to buy and setup 100 servers, is time taking and lots of money he has to spend to run a business.

→ To avoid this problem, the service provider is providing the same infrastructure on rent i.e pay for what you use.

→ In a physical server, we can create no. of servers for our use and delete them after completing our work, this creation of servers in a physical server is called virtual Server, and it is given by EC2 service in AWS

Advantages :-

→ lots of money saving

→ mainly time saving

Scanned with ACE Scanner

→ We can easily setup a virtual Server within 2 to 3 min of time. and we can also delete.

Advantages:-

1) cost benefits
   → No upfront cost
   → pay for what you use
   → purchasing options
2) Automation options
   → Aws cli
   → Aws cloudformation
3) Global availability
4) Low total cost of IT

Cloud Providers:-

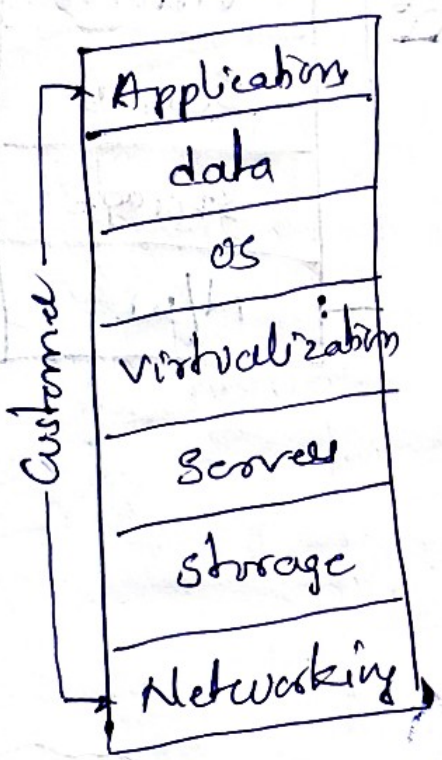
→ amazon webservices
→ Microsoft Azure
→ Google Cloud Platform
→ IBM Bluemix

Scanned with ACE Scanner

Service Models:

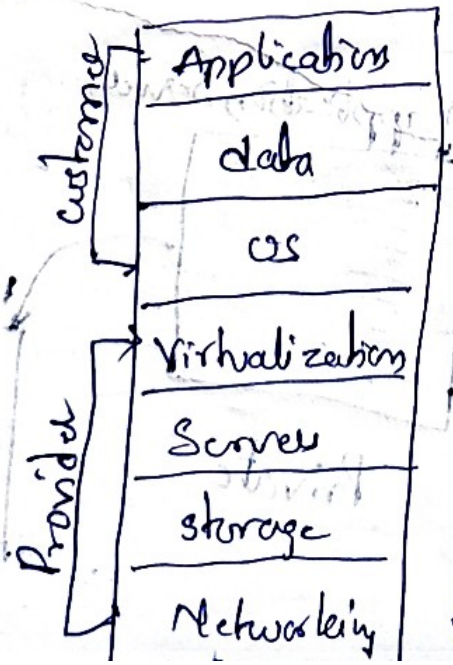
→ IaaS (Infrastructure as a Service)
→ Paas (Platform as a Service)
→ Saas (Software as a Service)

Private Cloud



→ customer has to buy all these H/w and s/w

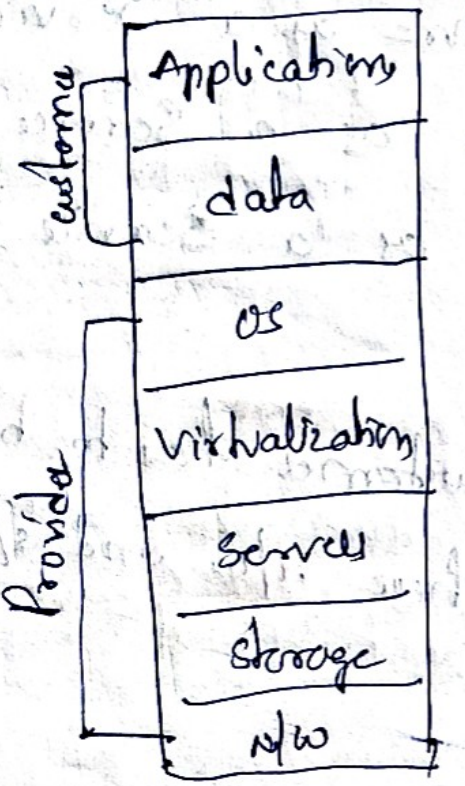
Infrastructure as cloud



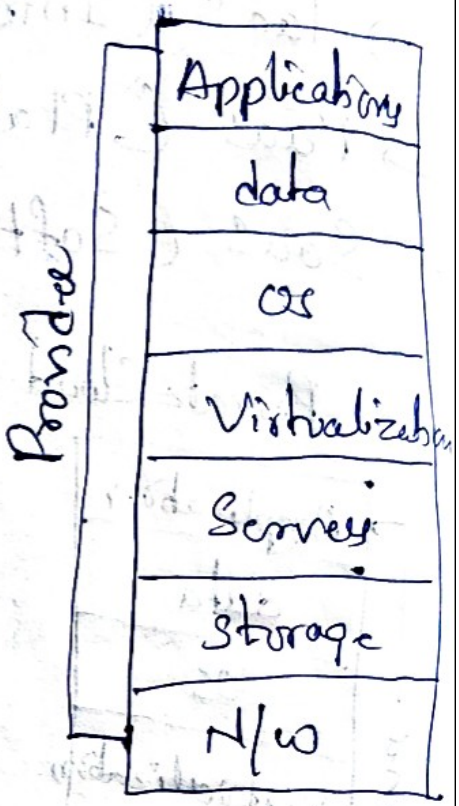
→ In this case from n/w to virtualization, this H/w is provided by service provider, we are cutting our cost for this

Scanned with ACE Scanner

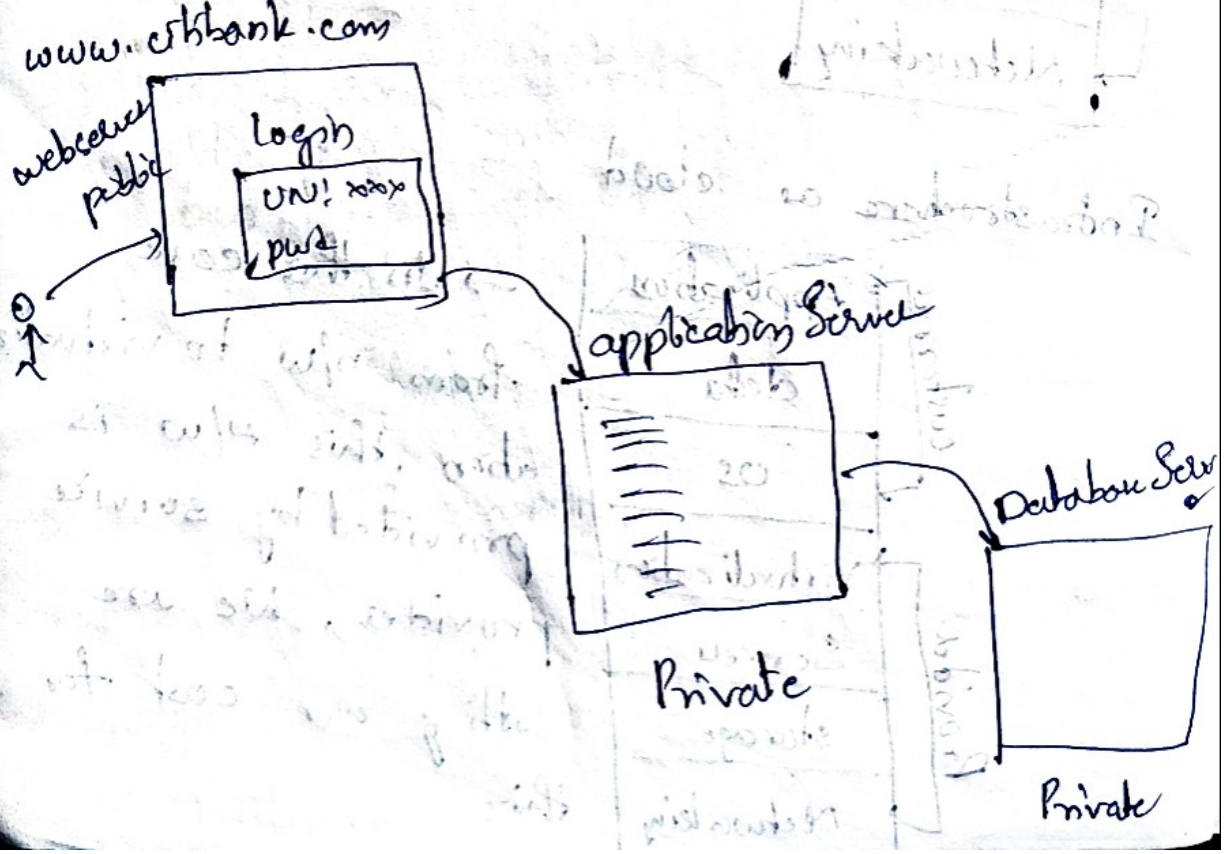
Platform as a Service



Software as a Service



3 tier presentation



Scanned with ACE Scanner

→ We can't expose application server (it has code for executing the applicati
→ Database servers are also not exposing to the customers, these two are private servers

⭐ webservers are more in number for any application, where as application server and database servers are less in number

→ If you are combining both public and private servers is called hybrid cloud platform.

→ Databases and application servers are created on premises are also on public cloud platforms.

Public Cloud :-

→ Infrastructure managed by cloud vendor
→ The customer has no visibility and

Scanned with ACE Scanner

control over where the computing infrastructure is hosted
→ The computing infrastructure is shared across organisations
→ Inexpensive
→ Secured

Private Cloud

→ Infrastructure is dedicated to a particular organisation
→ Most expensive and more secured
→ customer has to manage the infrastructure and s/w.
→ Build your own cloud using open source or commercial softwares like
  (i) apache cloud stack
  (ii) Eucalyptus
  (iii) openstack

Scanned with ACE Scanner

Hybrid cloud:-

→ Organisations may host critical applications on private clouds and applications with relatively less security concerns on the public cloud

→ The usage of both private and public clouds together is called hybrid cloud

→ Use databases and application servers on private and webservers on public cloud, this will for normal usage use private cloud and high/peak load use public clouds.

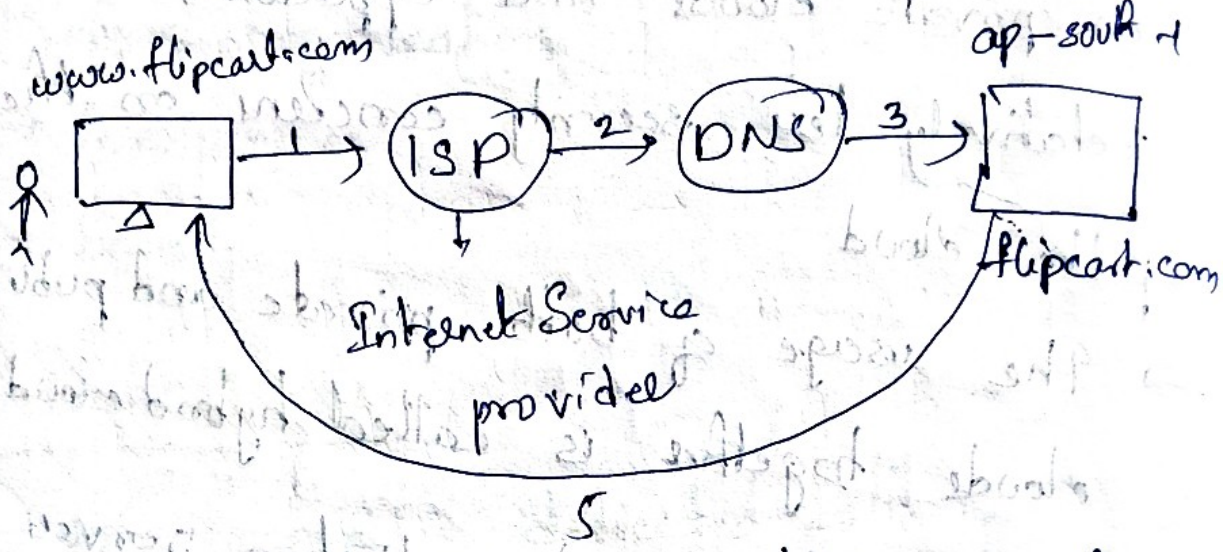
Aws Global Infrastructure

Region:-

It is a geographical location where amazon is maintaining their data centers

→ Now it has 22 regions

latency :- The time taken to process a request and give response is called latency.



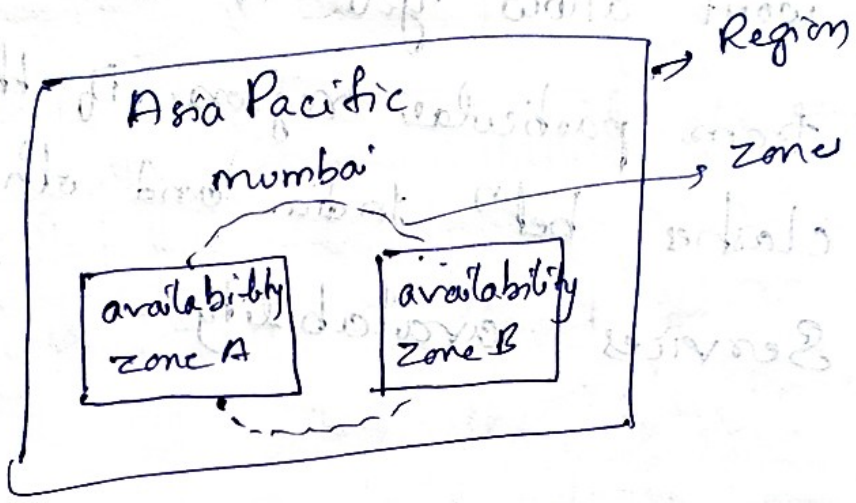
→ If the server is near to you, the latency is less.

→ By creating more servers to reduce the latency we can just enable cache servers (in this servers frequently used things are stored) the cause is calling this cache servers as edge locations to enable we have a service called cloudfront, they are calling this cache servers as "edge locations"

Availability zones :-

The region in which servers are kept in a location is called zone (or) the house where servers are kept is called zone.

eg :-



→ If a client is in India, the server region to be selected near by him

Advantages :-

→ Global reach to customers in minutes
→ High availability across regions
→ Disaster recovery sites

Scanned with ACE Scanner

The region to use / on what basis region is choosed.

① Proximity :- (Closest location to customer)

② Legal Compliances → this legal team won't allow you to use the servers from particular region if there is any clashes betn India and other countries

③ Services availability

(EC2) Elastic Compute Cloud

→ It is service provided by Aws, to create virtual machines on Aws cloud platform.

→ Based on the client requirement create the instances

→ EC2 is giving us pre installed o/s images

→ Marketplace :- It is like playstore of Aws, We can select the o/s we want and install it on the instance

EC2 Instances = Virtual Servers

① Resizable compute capacity in 200+ instance types.

② Reduces the time required to obtain and boot new server to minutes or seconds

Scanned with ACE Scanner

③ scale capacity as your computing requirements change

④ pay only for capacity that you actually use

⑤ Choose linux or windows

⑥ Deploy across Regions and Availability zones for reliability

⑦ Flexible networking (NAT/classic, VPC, Elastic IPs)

⑧ Support for virtual n/w interfaces that can be attached to EC2 instance in your VPC

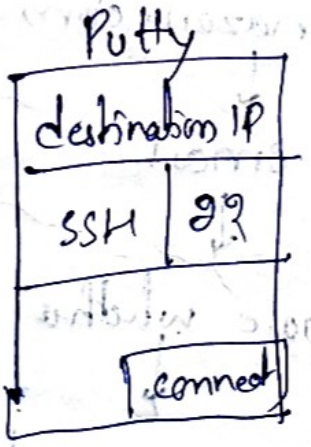
→ Based on the clients requirement, the AWS has created instance Family, there are 400+ instance families are there

→ Similar kind of work that a instance can do, we group them as a family

eg:- general purpose family
cpu optimised family
graphic card "
- memory based " " etc

→ The default port no. of SSH is 22
the linux understands SSH protocol

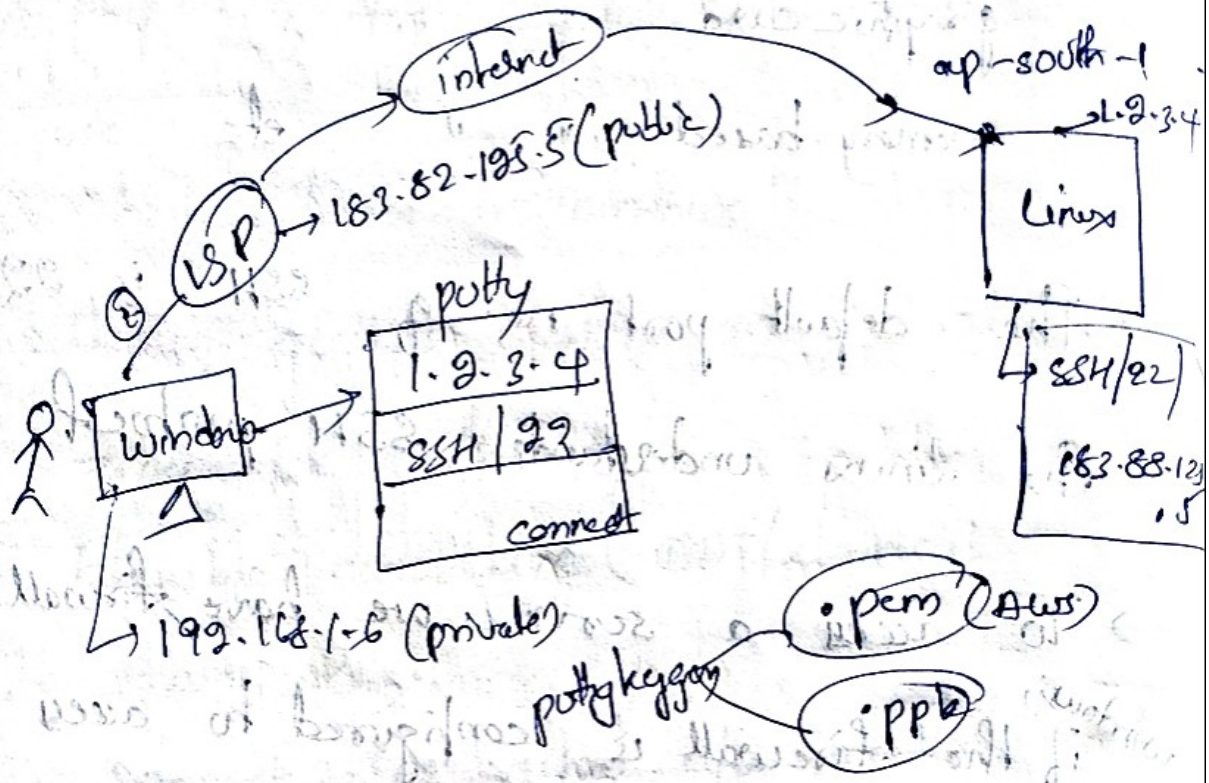
→ To access a server we have firewall in windows, if the firewall is configured to access then only we can access. where as in linux it is called as Security group.



→ AWS will give the public key, the extension is .pem (public encryption module)

Scanned with ACE Scanner

→ To convat .pem key into .ppk file
we need to have puttykeygen



→ To create AWS account, these are 5 steps

aws.amazon.com
↓
email
↓
choose whether ① Personal ② Business
↓
payment
↓
personal information
↓
select plans
↓
complete sign up

Scanned with ACE Scanner

→ To know the ip address of your laptop `ip config` in putty only

To connect to linux server from windows laptop

→ We need to have putty and puttygen to be installed

→ with the help of puttygen convert the public key pair .pem file into .ppk

→ In putty configuration

- give public ip of your server
- SSH
- 22 port no
- then select expand SSH which is present in left hand side
  - → In that just select Auth
  - → then provide private key file
  - → open/connect.

Scanned with ACE Scanner

Public ip :- It is used to access the server from outside the AWS n/w.

Private ip :- It is used to access the server from within the n/w.

→ when you start a server, the AWS community will assign a public ip address to it. At the moment you stop the instance it releases the ip address, because assigning a public ip to a stopped instance is no use.

→ whenever you stop and start the instance we are getting new ip ~~for~~ the instances.

→ To avoid that problem we have a service called Elastic ip address,

→ It won't change whenever you stop or start.

Scanned with ACE Scanner

→ These elastic ip addresses are chargeable

Keypair :-

→ the scope of a keypair is within the region.

→ You can launch 'n' no. of instances with one keypair in the same region only.

→ The same keypair will not work in another region

To create a elastic ip and add it to the server :

→ Network and Security

→ select Elastic IP's

→ click on Allocate Elastic IP address (this option is present on right corner)

→ Click on Amazon's pool of IPv4 addresses

→ Then click on Allocate

Then an elastic ip is created.

Scanned with ACE Scanner

Now attach that ip address to the server.

→ select the elastic ip address
→ Now go to Actions
→ select Associate Elastic IP address
→ click on instance and select the instance
~~then~~ ..
→ Associate.

For deleting elastic ip address

→ select the elastic ip address
→ go to Actions
→ select Release elastic ip address

→ In organisation, to change anything we have to raise a request ie CR (change Request)

Interview

Sir on the daily basis we are if an issue is happening we will be raising a CR to aws team, then they will solve it.

Scanned with ACE Scanner

→ yum is a installer tool in Linux

→ It will download, install, verify the package. (automate)

Note :-

In production we use only "rpm" to download and install any package.

→ yum installs whatever dependencies are in the package also. where as in rpm downloads whatever packages we select.

→ rpm is also installer tool (manual)

How to host/configure/install a website on Linux server:

→ install httpd (to configure a webserver)

→ by default httpd will be in inactive state

→ To check the status of httpd

| service httpd status/start/stop/restart |

Scanned with ACE Scanner

service service_name status

→ Now start the httpd

service httpd start

→ All the webpages are created in a location called /var/www/html

→ Create a index.html file for webpage

→ After changing anything, we have to restart the service

service httpd restart

→ chkconfig httpd on → whenever anybody reboot the httpd will go in dead

→ To see the website, go to security

group and add a rule as

http | 80 | anywhere

→ From browser give public_ip : 80

Scanned with ACE Scanner

How to automate webserver configuration

→ To automate we use shell script in linux.

→ shell script :- sequence of linux commands with some logic is called shell script

Rules or conditions to write shell script

① file must end with .sh

eg:- webserver.sh

② The first line of a script must and should start with

```
#!/bin/bash
```

→ shebang line

we are telling our o/s, that I want to write many linux commands, the commands library is present in /bin/bash only. then o/s will picks them and use it

Scanned with ACE Scanner

vim webserver.sh

#!/bin/bash → no spaces in the firstline

yum install -y httpd git

service httpd start

chkconfig httpd on

cd /var/www/html

git clone pasturl .

service httpd restart

→ write above script in notepad and save with extension as .sh

→ while launching the server attach that file to it, thats all, then take public ip and port no. and access it from browser

→ while launching server, you find advanced details section

Scanned with ACE Scanner

→ go to user data section → select As file and launch the server.

→ At the time of booting the server we are installing the script.

→ It is called bootstraping

strapping = attaching

Scenario:

① We deployed a website on prod server and it is running successfully.

② Now client came and asking I need to include few more modules in the same website, what will you do?

→ Never add those two modules directly on prod servers because there may be compatability issues, these may cause the on going business to go down.

Scanned with ACE Scanner

→ Then go for testing in low production servers as QA/deployment servers.

→ Now development team is asking prod manager, that create a copy of prod server and give it to me, we will add two more modules and test it, if it is going well, we will let you know.

→ Then prod team will create a Image out of that server i.e AMI.

→ Now the development team can create server from that AMI.

AMI → s/w library + softwares installed on it.

→ Direct an image can be copied but direct an server can't be copied

→ If we have AMI, we can create a server.

Scanned with ACE Scanner

AMI creation :-

→ How to copy an instance? → AMI

→ How to backup of instance ? → AMI

→ How can you do backup or restore prod server ? → AMI.

→ How to migrate an instance ? → AMI

Scenarios :-

* If prod servers are in same aws account and same region, → there no need of copying → he can create AMI out of it and use it

* Both are in same aws account and both are in different regions ?

→ We need to copy the AMI.

* Imagine that prod servers are maintain of seperate Aws account, non prod servers are maintaining seperate Aws account ?

Scanned with ACE Scanner

lets say TCS is maintaining AWS account and Bajaj insurance has gave project to implement and give it me and he is also maintaining AWS account.

→ then TCS will create a AMI and it will be copied into Bajaj companies account.

Creation of AMI

→ Let say you have prod server.

→ select that server and go to Actions

→ select image and template → create image

→ give some name

☆☆☆☆ No reboot ☑ enable

when you enable it, then it wont reboot production servers, reboot will be done by authorization or planned downtime

Scanned with ACE Scanner

→ create.

You can find the AMI in Images section

Copy a AMI into other region

→ select the AMI
→ go to Actions → select copy AMI.
→ select the destination region and specify your region
→ It will take 30 to 40 mins depending on the configurations etc of that AMI

copy AMI from one AWS ac to another AWS ac

→ select the AMI
→ go to Actions → select Edit AMI permissions
→ select private
→ In Shared accounts → select Add account ID and the AWS account ID of other person

Scanned with ACE Scanner

Elastic Load Balancer :-

Scenario:-

→ In production environment, we have two servers, If the load is increasing on them either of the server may crash or it will go down.

→ In that case we have to create one more server and do all necessary configurations, this may take time and bussiness will go down.

→ To overcome such problems, elastic load balancer is used, It will distribute load to the servers equally.

★ → If we have more than one server, we must have the loadbalancer

→ Now create two servers and one load balancer

Scanned with ACE Scanner

parameters to be configured in Load Balancer

ping

1. protocal : http
2. port : 80
3. path : /index.html
4. Interval : 30 sec
   → Load Balancer will ping the servers in every 30 sec
5. Response timeout : 5 sec
   → It will ~~receive~~ revive the respon for every 5 sec
6. unhealthy threshold : 2
   → If the first server or second server is not responding for first 5 sec it will show the server as in service and next time after 5 sec it again not responding, then load balancer shows that server is in out of service

→ then the requests of users are not sent

Scanned with ACE Scanner

to unhealthy server and all are forwarded to next server.

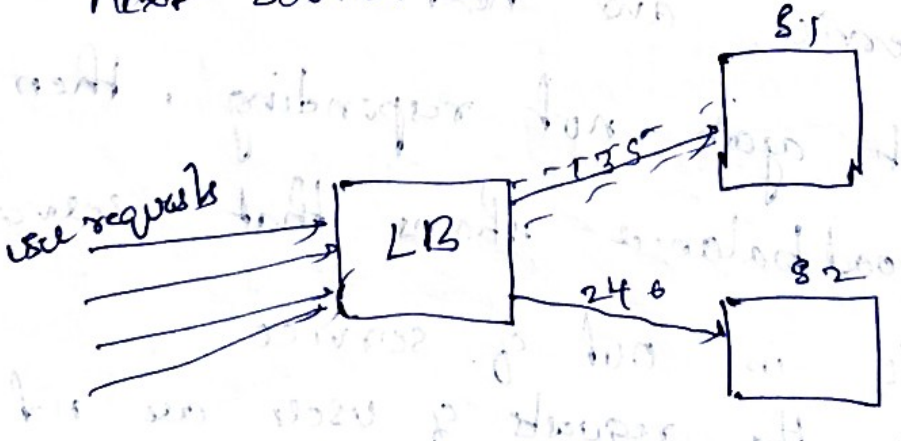
⑦ healthy threshold : 2

→ Load balancer keeps on pinging it when the server comes into working state, it will show as in-service after 2 thresholds.

⑧ connection draining ☑

enable it.

→ It will send less no. of request to unhealthy server in just 5 sec, then it shows as out of service and all the requests are redirected to next server.



Scanned with ACE Scanner

⑨ Cross zone load balancing ☑

→ one server may be in 1a zone
→ Second server may be in 1b zone.
→ But both must be in one region only and loadbalancer too.
→ ~~We can add~~ ~~no. of~~ ~~servers~~
→ We can add n no. of servers to a loadbalancer.
→ If both servers are in different zone enable it.

Interview:
⊗ On what basis this distribution of load happens ?

Ans - Round Robin

→ even distribution happens.

load balancer configuration

→ Create two servers and one load balancer
→ Now for our understanding create two

Scanned with ACE Scanner

different websites and deploy one in webserver1 and another in webserver two

```
#!/bin/bash
yum install httpd -y
service httpd start
chkconfig httpd on
cd /var/www/html
echo "welcome to webserver1" > index.html
service httpd restart
```

→ and do bootstrapping ~~for~~ one server

```
#!/bin/bash
yum install httpd -y
service httpd start
chkconfig httpd on
cd /var/www/html
echo "welcome to webserver2" > index.html
service httpd restart
```

Scanned with ACE Scanner

→ do bootstraping in server 2.

→ In lefthand side search for load balancer in AWS. and create load balancer and configure as required

Scenario:

→ one user is accessing a sbi bank's server 1 and doing transcations, he came to 1 page and waiting for otp and enter, meanwhile the server has gone down,

→ In that case the request will be redirected to next server, but the next server don't know where the stopped user is seeing for otp, then user starts from starting of the server page, we have to fill all the details and go upto otp page.

→ To avoid this AWS has given "stickeness", enable it then next server

starts from the same page where the user has stopped to do.

Stickiness! — ☑

→ enable it in loadbalanced,

→ when you enable, the AWS is maintaining buffer area, in that all the data is stored,

→ when one server is going down the next server will look into buffer area and starts from the same page where the other servers has stopped.

→ go to instances and scroll down, you will find the stickiness → enable it depending on the application.

→ You can have more no. of servers in the same loadbalancer.

(*) We have configured on load balance and running sucessfully, at any time a serve may down or go in out of service, then how cloud engineer know that the serve has gone down.?

→ He can knows when login into this account, and in the loadbalancer section he can see instances status;

(*) Is this possible to know the status 24/7 manually ?

→ It is not possible,

→ Then we depend on monitoring tools called (i) Nagios

(ii) bmcpatrol

(iii) splunk

(iv) prometheus

etc

→ These tools are not free,

→ If there is any error, it generates "Alarms"

Scanned with ACE Scanner

Cloud watch :-

It is a service given by AWS. to monitor the resources in AWS.

→ It is a managed service

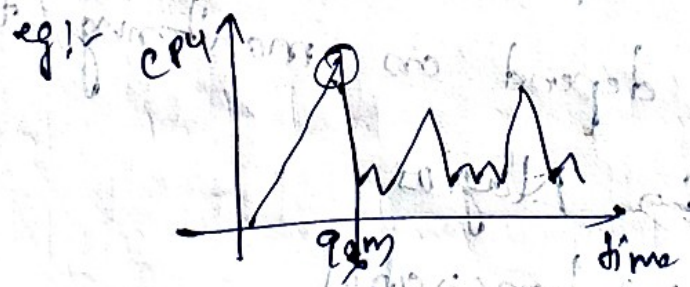
(*) How to monitor a linux server?

→ based on a metric

① Collection of information of a server

→ all the information is stored in log file

② Use visualization tool



take 9am time before 5mins and after 5min to 9am, and analyse the problems and troubleshoot it.

③ Root Cause Action

→ Understand the problems and take action

Cloud watch configuration

→ Write condition for this, as

whenever [ ] of cpuutilization ≥ 70% for-

min
max
avg
sum

- [1] interval of [5 min]

→ It is collecting info and storing in datapoint

(40) (50) (60) (35) (80)

→ min means → it takes 40 and compare with threshold 70%, if the condition satisfied it shows the status as

[OK]

→ If not satisfied, the status as

[IN-ALARM]

→ If the server it self crashes, it shows as

[INSUFFICIENT]

Scanned with ACE Scanner

Note:-

→ In real time we use avg or min

→ better set avg

(*) If alarm is generated, how do you know ?

→ For this I must get a notification as email or sms

Simple Notification Service (SNS)

Scenario:-

In e-commerce portal, let's say electronic goods,

→ event: Instock

→ Notifyme → whenever product is in stock

- email id
- mobile no.
- confirm subscription

→ All these details will be given,

→ whenever the concerned product is in stock

Scanned with ACE Scanner

it will send a email or a sms to the customer.

→ In the same way, in AWS, "event" is called as topic.

① Topic : HYDWEBSERVERHIGLOAD

→ whenever it gives a notification like that, cloud engineer must immediately login to server and rectify the problem

→ If I have 100 members, can we send emails individualy?

→ Create a group and assign them, group is called DL (distribution List)

→ There is a option called Alarm Notificat

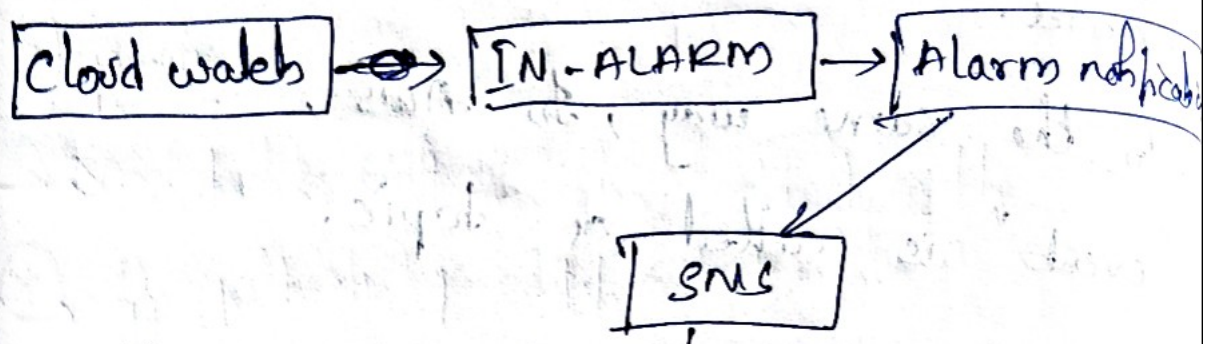
on

Alarm Notification [ ] D

topic will be present here

Scanned with ACE Scanner

→ If any webserver has HIGH LOAD



→ Assign the token for yourself and start working.

Create a webserver

↓

SNS

↓

cloudwatch → use SNS topic

↓

alarm notifications

step-1 :-

Create a webserver !

step-2 :-

Create SNS topic

→ go to SNS
→ click on create topic
→ give topic name
→ display name (with what name the msge is to be displayed)
→ In real time enable encryption
→ In Access policy → choose basic ✓
If you are programer choose Advanced ✓
→ Define who can publish msgs to the topic
✓ everyone
→ Define who can subscribe to the topic
✓ everyone
→ Delivery retry policy :-
How many time you can send msgs, whenever you send the msge their won't be network in that area.
→ enter no of times you want to send
→ Now create subscription
→ Protocol → select email option

→ endpoint → give email address

Step-5 :-

→ Create a alarm in cloud watch

→ go to cloud watch

→ select the metric

(or)

→ select the EC2 on which you are setting alarm

→ Actions → Monitor and troubleshoot → Manage cloud Watch alarms

→ Alarm notification → select the SNS topic you created

Alarm action :-

Be careful for enabling this,

Note :- Don't enable it without knowing the problem to do.

→ Without analysis don't enable it.

Select the parameter

Scanned with ACE Scanner

select 5minutes: → Interview

→ 5minutes is free upto 10 alarms, after that they are chargable

To check the CPU utilization increase

→ go to EC2 server and install ~~stress~~ stress package.

→ To install stress package, we need to have ~~amazon-linux-extras install epel~~

sudo yum install stress

epel → extra packages for enterprise linux

★ → top command will act like task manager in linux

→ Now execute the command

stress -c 100 --timeout 1800 &

100 → no. of times

1800 → seconds

& → to run in background

→ Now see the top comand and see CPU utilization

Scanned with ACE Scanner

Note:- Don't execute stress command in real time

→ Never execute apt-get update command on production servers, because update take more time, until the update is executed server won't be available for the business

→ rm -rf command too

→ To stop the process, we use kill command, if the 100 processes are running each process has id, we can't give 100 ids to do so

```
killall stress
```

```
killall process_name
```

Scanned with ACE Scanner

Scenario:-

→ whenever a application is runing on Let's say 3 servers, then one server has gone down, to create a server and replace it, is a time consuming process in the mean while my business will get down.

→ The creation of server must be automatically happen

Requirements:-

→ whenever the load increase, the server must be created automatically

→ whenever the load decreases, the servers must be decreased automatically

→ whenever the server goes down loadbalancer will sence it

→ For this I need to have

(i) Create AMI

Scanned with ACE Scanner

(ii) Launch configurations

→ server (AMI), config, type

→ EBS, tag, Security group, key

(iii) Load balancer :-

→ It must be empty at starting point.

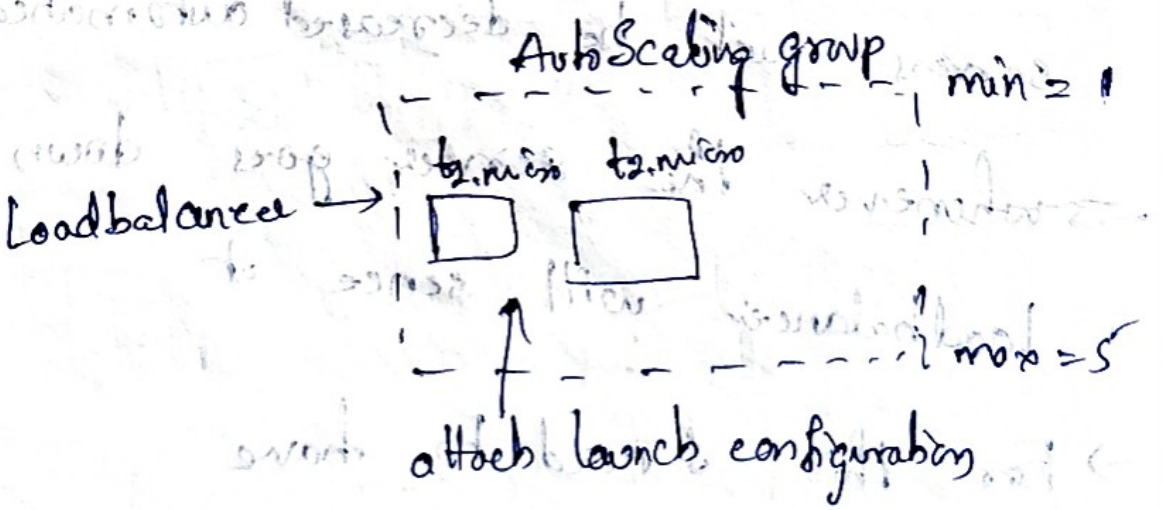
→ The addition of servers into Load balancer must be automatic

(iv) We need auto scaling group

→ whenever load increases & decreases the servers will come up automatically is called : Auto scaling :

(v) scaling polices

Create autoscaling group



Scanned with ACE Scanner

→ write scaling polices, i.e. write conditions for increase and decrease the servers

⊛ Policies:-

(i) increase group size i.e adding more no. of servers

(ii) CPU utilization ≥ 75% for 1 interval of 5min

- Take Action [Add] [8] instances

(iii) Add alarms, SNS

(iv) decrease group size, write condition whenever CPU utilization increases in my case it adds servers

→ when the addition of servers reaches the max limit you specified,

→ The autoscaling group is trying add more servers but addition of servers reached max limit; then it shows ERROR i.e failed to launch

↓
Interview

Scanned with ACE Scanner

→ decrease group size

whenever avg of CPU ≤ 30 % for 1 of 5 min.

Takes Action [remove] [2] Instances or your choice

→ whenever it reaches min limit while removing, it throws an ERROR

ie failed to terminate

Interview

(*) How to upgrade autoscaling group servers?

→ launch configuration

→ Now change values min=0 max=0 then old servers will be removed, do this in planned downtime only.

→ Then attach new launch configuration to the autoscaling group and now change values min = 1 max = 5

Scanned with ACE Scanner

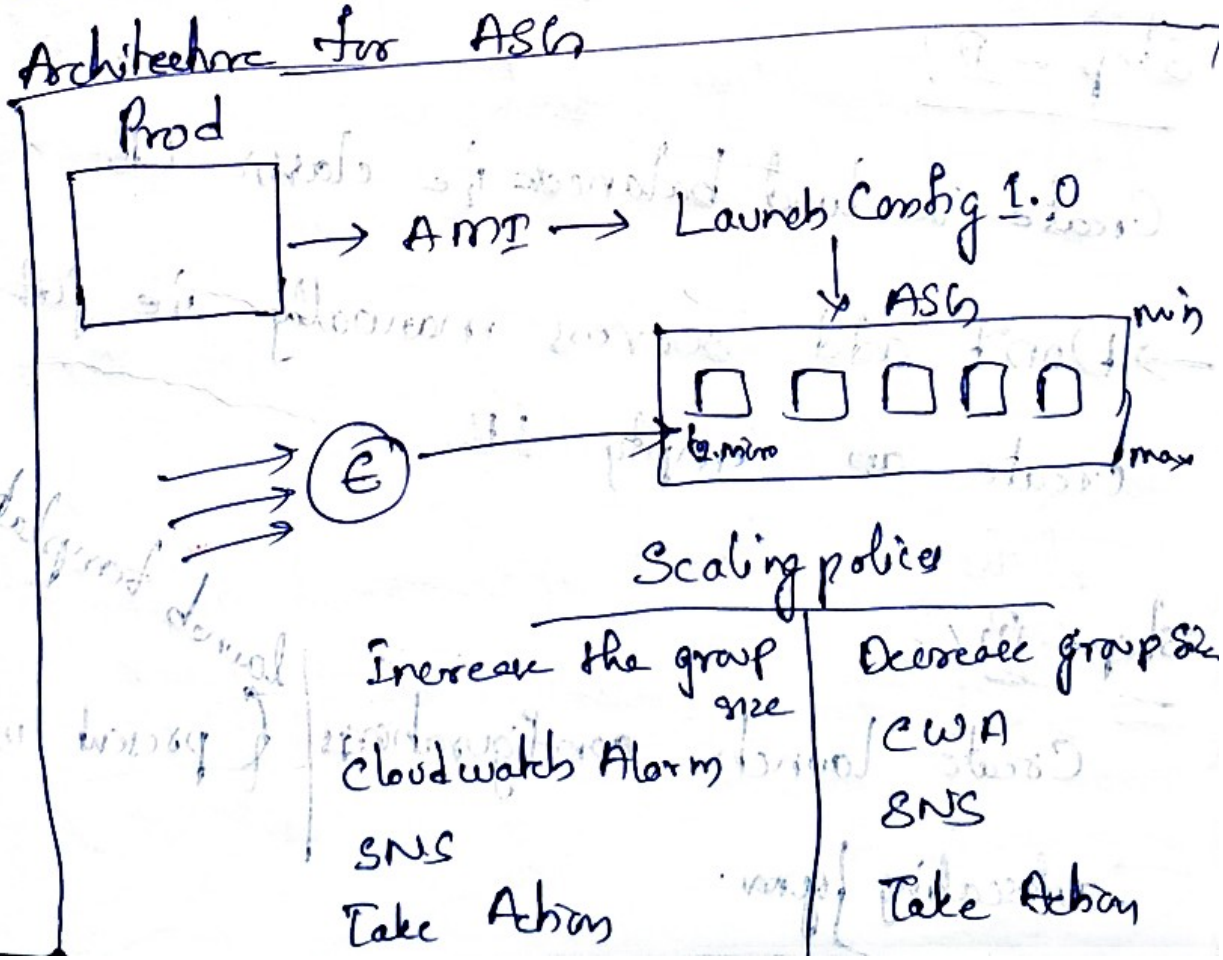
Scenario

(*) You have set it for CPU utilization but the server goes down due to another reason i.e a server crashed what is the autoscaling group do?

→ The autoscaling group has self healing mechanism i.e the server will come up, If a server is going unhealthy the autoscaling group will sence it and replace it with healthy server

→ It is called horizental Scaling

Architecture for ASG



Scanned with ACE Scanner

Target Tracking :- It is a service given by AWS to write policies automatically

→ just add metric and threshold

→ no. of servers to add and remove, adding alarm etc there all be taken care by target tracking.

step - I :-

Select PROD servers and create AMI out of it

step - II :-

Create a load balancer i.e classic LB

→ Don't add servers manually i.e just create an empty LB

step - III :-

Create launch configurations / launch template { present in autoscaling } grp.

Scanned with ACE Scanner

→ Click on creat launch config

→ Select AMI for which you have to create

→ choose instance type

Monitoring :-

☐ Enable EC2 instance detailed monitoring within cloud watch

EBS-optimized instance

☐ Launch as EBS-optimized instance

→ These all are chargeable

Interview

* Already you have webserver AMI, How can we add additional s/w's on top of same AMI ?

→ while creating launch config, in advanced details → write a script or attach the s/w file in user data

→ like this you can install more s/w's on top of AMI.

→ select all the requirements you want and create launch config

Step-IV:-

Create ASG

→ select Launch configured → Actions → create Auto Scaling group

→ Name

Refer (*) You are creating min 10 servers in ASG, should all the servers in one zone?

→ No, we should not create in zone, because to have high availability.

→ Select atleast two or more zones

→ Attach to a existing load balancer

Scanned with ACE Scanner

→ select classic LB

Health checks

☑ ELB

Additional settings

☑

Group size

Desired capacity :- It is the value between min and max,

→ We can't go ~~beyond~~ beyond the desired capacity,

min — desired capacity — max

eg:- desired capacity [2] / or give value between them as your wish

min [1]

max [5]

Scaling policies:-

○ Target tracking scaling policy ○ None

Scanned with ACE Scanner

→ If you select none, you have to write policies

→ If you select target, just select metric and threshold, other policies are written by Aws

Warm up instance [ ] seconds

→ The time taken by the service to get ready to work

→ give the value depending upon how big is your application

Instance scale in protection

whatever instances added, you don't want to remove them, then enable it

Scale in :- decreasing/removing

scale out :- adding/increasing

Add notifications :-

Scanned with ACE Scanner

select the events

Add notifications .one for each event

Tags :-

→ If is important, give some tag name

key	value
Name	ASG-Servey

Dyramic target policy :-

→ write your own policies

→ Name

→ select step scaling

→ Take the Action

Add ☐ servers

→ Depending on client use target, or dynamic policies

Scanned with ACE Scanner

~~Snapshots~~ :- Storage :-

* How many types of hard drives ?

Tape drives, cold HDD, throughput optimised HDD

↓ 12 mb/sec ↓ 123 mb/sec

SSD's → 3000 i/p, o/p's/sec

→ In ssp's → different types of SSD's.

Scenario :-

Two teams in the organisation require 5GB each to store their data,

→ To allocate this type of volumes to the teams is done by EBS

EBS (Elastic Block Store) :

→ It gives types of storage devices, select the type and use it

→ In `/dev/xvda` → it has all the s/w's required to the server, therefore don't store your applications data in it

xvda → extendable volume virtual device a

→ It is the first volume attached to any server

Note:-

The instance and the volume should in the same region and same availability zone.

→ It is a disadvantage

→ first volume name is /dev/xvda

→ second volume name is /dev/xvdf

(*) How can we do partition of 10GB and allocate to each team ?

→ fdisk is used to do partition

fdisk pathname of volume/HDD

eg:- fdisk /dev/xvdf

: n → new partitions

: 1 → it will ask for partition number, it starts from 1

: +5G → size you want to enter

Scanned with ACE Scanner

→ For next 5GB

:n ↵

:2 ↵

:t56

:w ↵ → to come out

10GB

xvdf1	xvdf2

② Format:-

To make any file system to know the OS

OS FS xfs → extendable file system

mkfs -t xfs /dev/xvdf1

mkfs -t xfs /dev/xvdf2

→ Now OS knows what kind of file to store in this

③. mount points:-

→ These are nothing but directories

mkdir /shift1 /shift2

application name/give any name

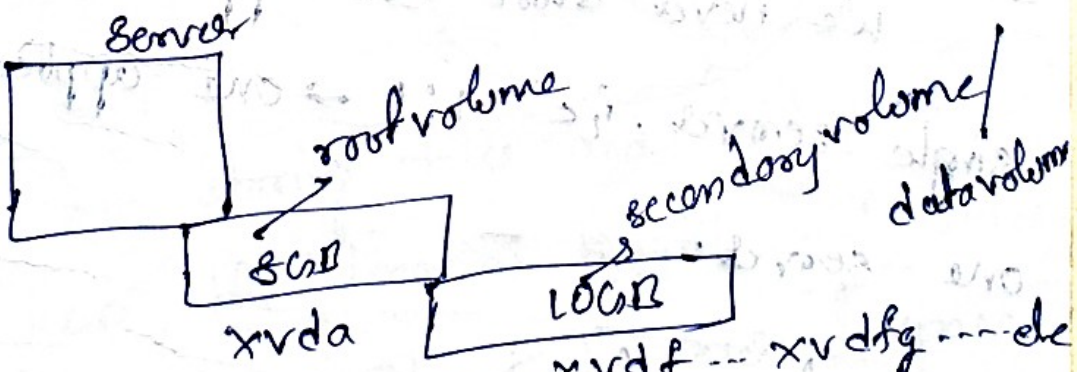
→ Now give shiftA directory to one team and shiftB directory to another team.

④ mounting :- attaching these folders to teams

```
mount /dev/xvdf1  /shiftA
mount /dev/xvdf2  /shiftB
```

definition:-
→ We have service called EBS, with which we are creating volumes and attach the volume to the server

→ while launching the server we create volumes



→ We can attach 12 volumes maximum for a server

Scanned with ACE Scanner

→ The max size of root volume is 2TB

→ The max size of secondary volume is 16TB

step-02 :-

→ To see the disk information

```
fdisk -l
```

→ Select the server and go to EBS service

→ first create a volume and attach to the server

Note :-

We never store two application data in single server. i.e 1:1 → one application one server.

→ Create volume.

→ After that select the volume → Actions

→ attach volume → select the server.

Scanned with ACE Scanner

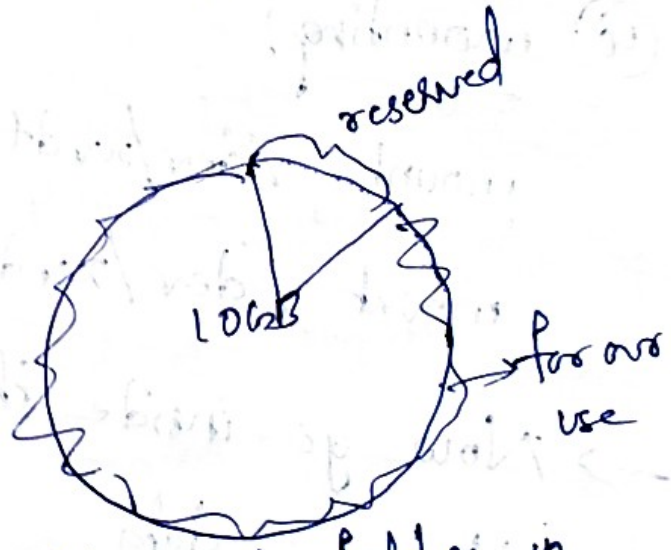
→ Now go inside the server and partition

① Partition

fdisk pathname of HDD

```
fdisk /dev/xvdf ↵
```

```
: n ↵
partition type p : ↵
First sector : ↵
: +5G ↵
```



→ We are storing file and folder in the HDD, the location of file and folder path/address is stored in reserved space

∴ For 16 GB pendrive, we get 14.83 gb free and remaining gb is reserved.

Scanned with ACE Scanner

:w ↵ for saving all these partitions

② Formating :-

mkfs -t xfs /dev/xvdf1

mkfs -t xfs /dev/xvdf2

lsblk
df -hT

③ mounting points :-

mkdir /shifta /shiftb

④ mounting :-

mount /dev/xvdf1 /shifta

mount /dev/xvdf2 /shiftb

→ Now go inside the mount point and add your data

cd shifta ↵

→ touch file{1..100} → It will create 100 files

→ go to another mount point shiftb and create files

Scanned with ACE Scanner

lsblk → To see the files added to volume.

→ umount is used to unmount the files

→ To keep the mount point permanently → go to vim /etc/fstab and write as follows

/dev/xvdf1 /filename xfs defaults 0 0

Snapshot:-

Taking backup of a volume is called snapshot.

→ We can read only the data but we can't write

→ To read write on the data, create a volume and attach it

→ Create a volume out of snapshot and attach it to the server

Note:-

After attaching the volume from snapshot, don't create partition, If you do so data will be lost.

→ partition ✗ → do only mount

→ formatting ✗

Scanned with ACE Scanner

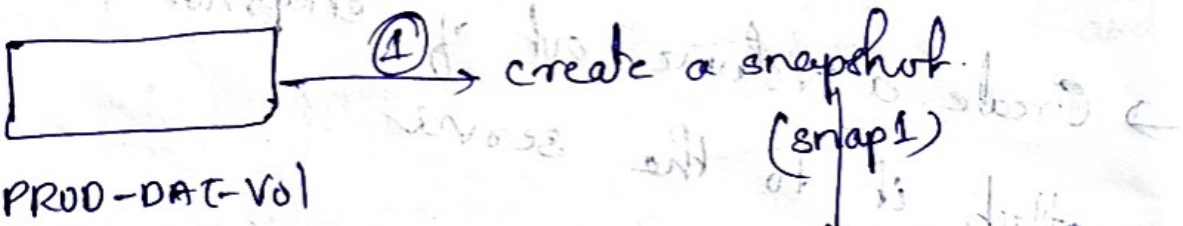
→ when you mount only you can see the data.

Interv:-

① How to copy data → snapshot
② How to backup and restore data → snapsht
③ How to migrate the data → snapshot

Scenario:

① Same Aws ac, same region
② same Aws ac, different region
③ Different accounts



Scanned with ACE Scanner

→ Now you can copy into another region

→ To copy snapshot into other AWS ac

click Actions → modify permissions → give Ac ID

→ We are using EBS volumes for dynamic data

→ To store huge static data, We have a service called S3

S3 :- (Simple Storage Service)

→ Huge static data can be stored in it.

eg :- google drive

→ It is called as "bucket"

→ It can store unlimited data

→ file is called as "object"

Scanned with ACE Scanner

→ Each file can be of 0-5TB, like this we can upload unlimited files.

eg:- netflix, gaana.com etc

(*) How many buckets can be created in Aws ac by default?

→ 100

(*) How can we create more than 100 buckets?

→ Yes

→ By raising a support ticket.. feature called increase limit

→ 1:1 bucket is required i.e 1 application 1 bucket.

Scenario:

→ wherever you upload a movie in netflix, many users are accessing it,

→ There shouldn't be latency.

Scanned with ACE Scanner

→ It must be accessed to everybody without latency.

→ For that we have "standard Storage class"

storage class :-

→ standard is a default storage class

→ they are using SSD's only (Solid state drive)

→ they will maintain 9 copies of a file by default.

→ It is managed by AWS

Scenario:

→ For the first 30 days everybody is accessing the data/movie

→ After 30 days very less no. of users are accessing it, then why to maintain 9 copies and pay for them,

→ So I want to reduce copies and cost cutting also.

→ For that we have a feature called Reduced Redundancy Storage (RRS)

Scanned with ACE Scanner

Reduced Redundancy (Repdition) storage: (180 days)

Redundancy means duplicate

→ RRS means less no. of duplicate copies

→ By default it is maintaing 2 copies

→ Therefore cost is reduced

→ Then write one policy i.e if my data age is more than 30 days change the class

Scenario:

→ If my endusers are accessing it very rarely, why to maintain 2 copies and pay for it.

→ For that we have a class called glacier

Glacier:- (slowest movement/slowest storage) (365) days

→ It maintains only 1 copy

→ It is very slow to access the data.

Scanned with ACE Scanner

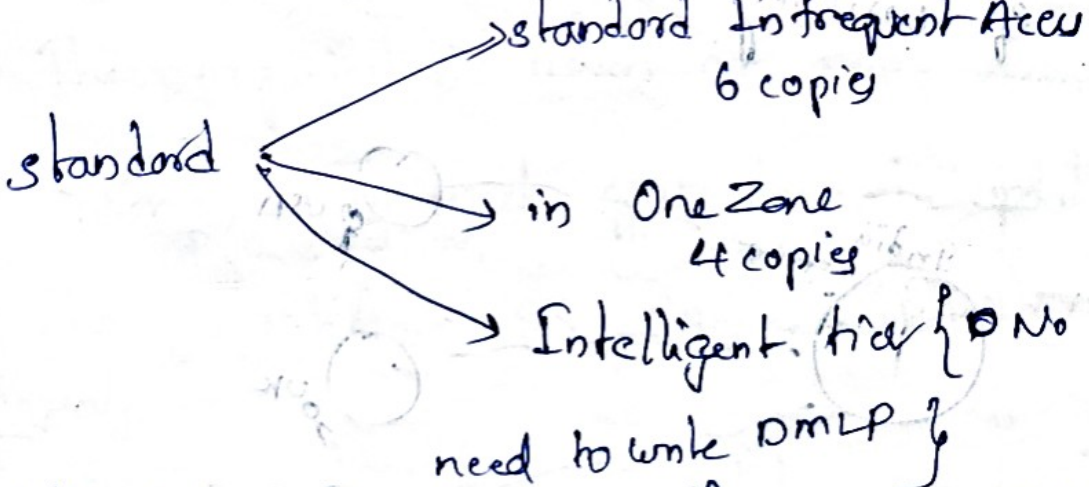
→ Cost will be reduced, archive your data
→ Tapes are used in this case.
→ It takes (3-4) hrs to access the data
→ Raise a request, After 365 days delete it or not upto you

DLMP → Data Life cycle Management Policy

90 days in (standard) — 180 days in (RRS) — 365 da in (Glacier) → delete it

→ Write DLMP → ~~force~~ on the wish of client

To overcome the DLMP, AWS given sub classes in standard only



Intelligent tier → It sees the no. of users

Scanned with ACE Scanner

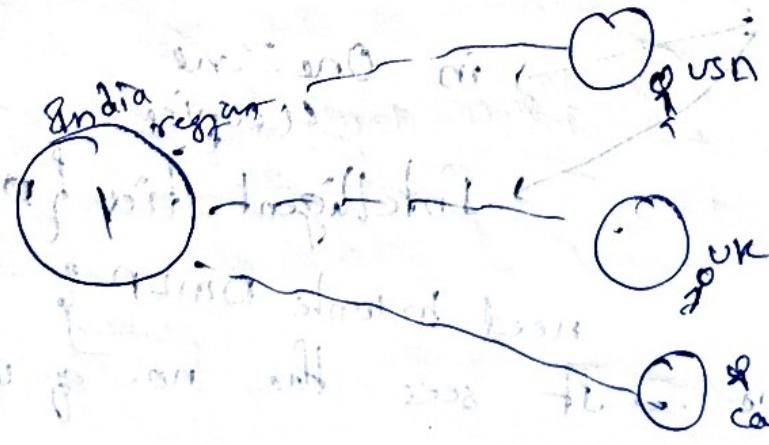
accessing it, if more no. of users are accessing it, it changes to standard IO, if no body accessing it changes to glacier

→ Here we no need of writting DMLP, depending on the users load, it changes automatically.

→ store 180 days in intelligent tier

→ 1 day store in standard, and next 180 days store in Intelligent tier, then IT will take care, billing is reduced.

→ Intelligent Tier, use it in organisation to reduced the cost, therefore no need of writting DMLP again and again.



Scanned with ACE Scanner

→ The objects are called Static object Store

Security:-

→ In S3 bucket, there is encryption ☑

→ Two types of encryption:

① S3 service is giving encryption algorithm

→ S3 is additionally charging you.

② Customer can manage encryption algorithm i.e he can buy encryption algorithm from third party and store, with that only he can encrypt his data and store

→ Encryption certificate can be uploaded

→ KMS / customer managed keys

→ The organisation which are using second option, every week you have to update the algorithm SSL (Secure Sockets Layer)

③ Property:-

→ We can also host a static website on S3, without going with EC2.

Scanned with ACE Scanner

→ enable the static website ⊙ by defau
it will be disabled

→ It will give you url, not the IP address

① index page → index.html

(or)

homepage.html

② Error page

error.html

→ save

③ All the data of your website upload in the S3

→ With this we are saving lot of money

II. property:-

→ I want to get notification, if anybody doing something in S3 bucket

→ For that we have Event notification

Scanned with ACE Scanner

11. Property! - Versioning

Scenario

→ The older and later versions of your data are maintained by S3

→ enable versioning while creating S3 bucket

→ Once you enable, there is no option to disable, if you disable it, the older objects will be lost

→ enabling is chargable.

→ We can suspend the versioning,

→ whenever you click on suspend button, the older versions will retain and next object onwards there is no versioning.

→ S3 service is Global, i.e we can access it from anywhere.

⇒ S3 bucket creation:

→ click on create bucket

→ Bucket name

Scanned with ACE Scanner

Intv:-

→ while creating new bucket, can you copy existing bucket settings?

→ Yes, click on copy existing bucket settings and click on the bucket you want

→ ACL (Access Control lists)

→ Block all public access

→ Bucket Versioning
  ⊙ enable

→ Default encryption
  ⊙ Disable
  ⊙ enable

→ In real time enable encryption

Intv:-

→ If anybody doing deleting the objects in s3, How can you block it?

→ Yes, by object lock

→ create bucket.

Scanned with ACE Scanner

Permissions! –

→ Others to access the bucket give the permissions

→ unblock public access permissions and write a policy in json format

→ In real time scenarios don't enable public access

→ give the ARN of your bucket and don't change anything

```
{
  "version": "2012-10-17",
  "statement": [{
    "sid": "PublicReadGetObject",
        "Effect": "Allow",
    "Principal": "*",
       "Action": ["s3:GetObject"],
      "Resource": ["arn:aws:s3:::nitbucketbyd task/folder1/*"
      ]
    }]
}
```

Scanned with ACE Scanner

→ On which folder, file you want to give access then specify the folder, filename in the code

→ Now you can open the file through url given in the bucket.

→ To suspend the bucket versioning → go to properties.

* Create event Notification:

→ select whatever option you want, this will send a notification

→ create SNS topic

Transfer acceleration:-

→ To transfer data faster, enable it.

* Static website hosting:-

→ By default, it is disabled,

Scanned with ACE Scanner

→ enable it and host static website

→ Enter two files

(i) index.html

(ii) error.html

→ If gives an url, it throws an error which means our project folders are not added, 404 Not found error

→ Then upload your project files in the S3 bucket,

→ Then, access it with url, and you can see the static website

Note:- If and only if static website choose S3 to host it

Management:- (Lifecycle rules)

→ create lifecycle rule

→ name

Scanned with ACE Scanner

Lifecycle rule actions

→ select all actions

Transition current versions of objects between storage class

choose storage class transition days after object creation

Intelligent-Tiering | 30

Transition non current versions

→ These are older versions of objects

Glacier | 180 | 5

↓

no. of newer versions to retain

Expire current versions of objects

→ I want to delete the current versions after 2 years

days

730

Permanently delete non current versions

day

365

5 → no. of copies to retain

Scanned with ACE Scanner

Delete expired object delete markers or incomplete multipart uploads

→ let's say a movie file has a size of 1GB

→ while you are downloading, there is a interruption, but you downloaded 70% i.e 700 mb,

→ With that 700mb you can't play the movie you need full file to play that movie.

→ With 70% of file no use, and you have to pay for it, but it is not in use.

→ Next time you can resume the same activity,

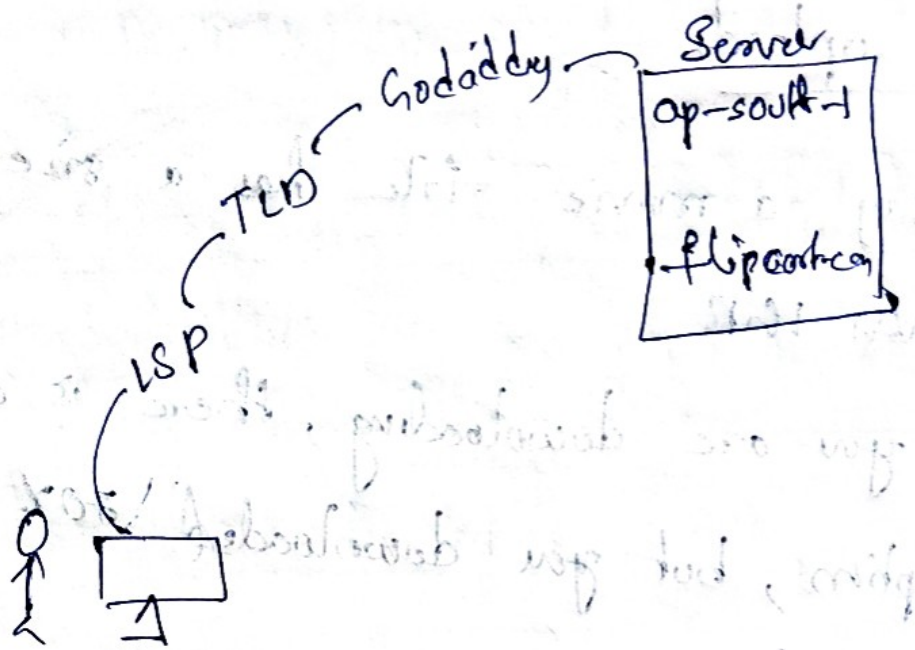
→ But these types of incomplete parts are retained upto 7 days

☑ Delete incomplete multipart uploads

days

7

Global Infrastructure



→ If the same person and server is in same region, latency will be less 5ms.

→ If the user is in USA, accessing the server in India, then that request has to travel long distance latency will be more.

→ To overcome this problem if we create more instances in all the regions, cost will be more.

→ To overcome that we have to enable cache servers (frequently accessed static data is stored in it)

Scanned with ACE Scanner

→ For this type of problem, In AWS. we have a service called CloudFront

CloudFront :-

It is a service, to create cache servers in AWS,

→ Just enable edge locations

→ This can be done through distribution i.e we are distributing our static data/content over the network

CloudFront → edgelocation → distribution

↓

① origin location

② origin location { loadbalancer (or) S3 bucket }

select it

③ price class

(i) USA, europe, canada

(ii) usa, europe, asia, africa, middle east

(iii) All edge locations

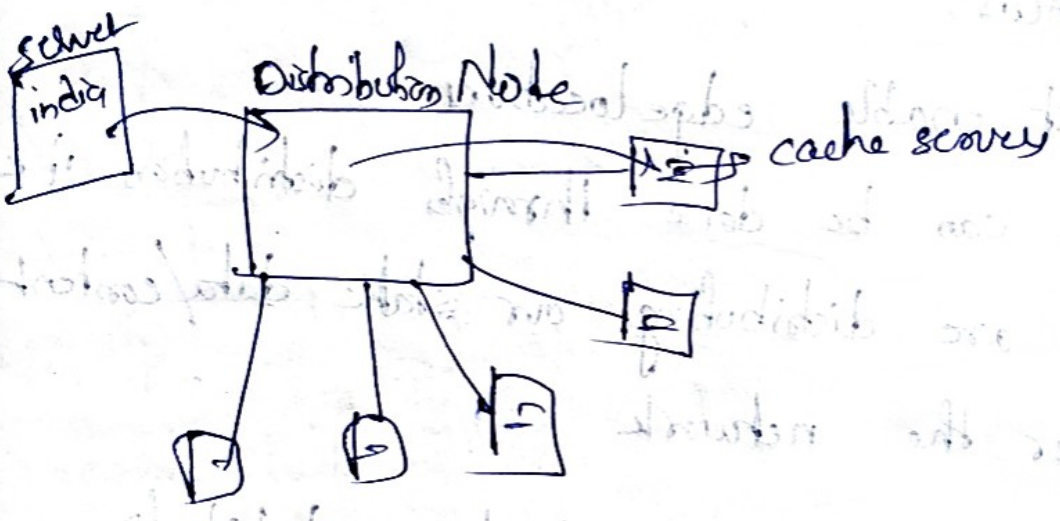
→ More money will be spent, depends on the option selected.

Scanned with ACE Scanner

→ Here we are saving considerable amount of money for client

→ AWS is maintaining 'n' no. of edge locations in all the regions.

→ Then are called distribution Nodes



→ All the edgelocations/cache servers are connected to distribution node,

→ The distribution node will forward all the request to nearest cacheserver of the user.

→ If the cacheserver gone down, the distribution node will forward the same request to the next cache server

→ DNode will pull the information from the

server and push that information into the subsequent cache servers, it will maintain same data,

→ To update the server we have a parameter

TTL → Time To Live

→ We are setting the time to live state data inside cache servers

$TTL = 7 \times 86400$ — for 1 week

↓

7 days

→ Because of this, the existing content in the nodes will be expired after 7 days and new content will be pulled again

→ Every company will be having their own update policy. (for 30 days, 7 days, 3 months etc)

→ To pull the content from the server to the node, Latency will be there, because it has to go through internet, to overcome that AWS is using its own network

Scanned with ACE Scanner

→ For the distribution it has secured n/w. and latency will be low.

Intro

(*) How do you know that secured n/w is reducing the latency

→ least conjested algorithm

~~comparison~~

Advantages

1) latency solved
2) cost reduced
3) security
4) least conjested algorithm

CloudFront:- To load the websites faster without creating virtual machines across the globe, we use cloudfront, We solved the problems like

(i) Latency
(ii) cost cutting
(iii) security

Scanned with ACE Scanner

Interview:

(*) what is cloudfront?

(*) what is cache server?

(*) what are edge locations?

(*) what is content Delivery Network?

Ans! - Cloudfront

Scenario! -

→ Create a S3 bucket and host a website in it.

→ Then create a cloudfront and create a distribution and use the S3 as origin in the cloudfront

→ Add new identity

(i) O/No → customer has to write policy

(ii) O/Yes → cloudfront has to write policy

→ Now create Cloudfront i.e CDN (Content Delivery Network)

→ We are enabling the cache servers to reduce the latency of my website

Creat OAI (Origin Acess Identity)

✓ Yes use OAI

→ Create OAI

→ Bocket policy

✓ Yes

*settings:-

Price class

✓ use all edge locations (best performance)

Intro:-

(x) My bussines is in India, Can I select USA as edgelocation?

→ No.

→ We have only 3 options, from that only you have to select

Default root object - optional

index.html

→ Cloud front gives SSL certificate, gives secure access

Intro:

(*) All of a sudden, if one of the caches server is going down, who will be taking care, who will be knowing that)

→ DN (Distribution Node)

→ The popular cache service is "AKAMAI"

→ In AWS we have cloudfront

→ Youtube data is handling by "AKAMAI"

→ We can use "AKAMAI" service, management has to enable it, the AKAMAI people come and install it.

(*) Instead of S3 bucket as origin, use loadbalancer as origin.

IAM

Identity Access Management

→ It is nothing but creating credentials to the employee.

→ Creating users and giving permissions on AWS account.

→ eg:- Create a dev environment, and create group of developers and give them access to in the dev environment

→ We can manage them by giving concerned group permissions to them

In windows

→ We have Active Directory to create users and manage them.

→ AD → User management

In Linux

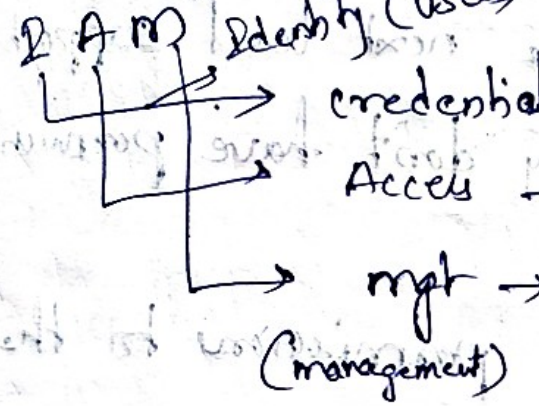
→ LightWeight Directory Access Protocal (LDAP)

Scanned with ACE Scanner

→ Employe ID is called as LDAP only

In AWS :-

We have "IAM", is a service given by AWS. It is "User Management" tool



(i) Creating a user in IAM

User Access Types :-

1) AWS mgt console Access (GUI)

Username : user1

pwd :

2) AWS programmatic Access (CLI)

(ii) Access :-

→ Provide permissions to the User.

→ Permissions are given on Service level only

Scanned with ACE Scanner

eg:- L1 support team.

→ They will just check the health of servers.

→ They have only read only permissions.

→ If any server is not running, they will inform to the next level support team, because they don't have permissions to modify.

→ Like wise we give permissions to the concerned team.

eg:- On EC2 service, we give read only permissions.

★ ROT (Rule of thumb) → Always provide least privileged Access

→ give only the permissions, the user is working on.

Scanned with ACE Scanner

Permission / policy :-

(i) AWS managed policy :-

→ It is default policy, created by AWS.
→ No-modifications
→ These are listed in the polices.
→ attach to any user
→ On every service we have two access they are (i) read only access
(ii) Full access

(ii) Customer managed Policy :

→ modification is possible
→ Listed.
→ attach to any user

(iii) In Line policy :

→ modification is possible
→ not listed
→ only for a particular user only

Scanned with ACE Scanner

Management :-

→ If I am having 100 users in a organisation, they are working on same operation

eg:- L1 support team has 100 users,

→ Create a policy and attach to the individual users is time taking.

→ Instead of that,

(i) create a group

(ii) Create a policy and attach to that group

(iii) Now add users to the group

→ then the permissions of the group will be applied to all the users in the group.

→ One user can be a part of multiple groups.

Scanned with ACE Scanner

Scenario:-

Let's say a website has login page, signup button will be there, We have fill all our details and upload photo etc.

→ The client says that store all the details in a database (RDS) and photos in S3 bucket.

→ For this there must be communication between RDS and S3.

→ website from EC2 service is uploading photos in S3 and details in RDS, i.e service to service communication is happening

→ We can't attach policy to the service.

★ policies can be attached to the User/group
Roles are attached to the service

Role:- It is nothing but a policy; giving this service credentials to other service to login into it.

Scanned with ACE Scanner

policies are for Users
Roles are for Service

eg:-

→ If a person is wearing a kaki uniform what is the role he has to control law and order

→ The uniform defines this role.

→ So the same way I am creating a role for EC2 instance, the EC2 instances

→ The role is S3 full access and RDS full access

→ If the EC2 instance has the role on it, then the can access S3 and RDS

MFA:- Multi Factor Access Authentication

→ This is an additional layer of password on top of credentials.

Scanned with ACE Scanner

* Users are global
* Access management Service is free

Creating a User:

→ click on add User

eg: Username : sudarshan@ibm.com

depends upon your company naming convention/ policy

Select AWS access type:-

- ⊙ Programmable access
- ⊙ AWS management Console Access

Console password :-

→ click on custom password

Require password reset :-

→ Policies :-

→ EC2readonly Access ☑

→ Every service has readonlyaccess & Fullaccess

Scanned with ACE Scanner

→ Create user
→ send the details to the user, login credentials to him via email.
→ Create alias to AccountID, because the ID should not be known to user..

Customer managed Policy:-
→ Give the user just start, stop, restart of instance on top of EC2 readonly acc
→ Then write the policy in json

To create a policy
→ go to policies in the left hand side of Aws ac page
→ click on policies
→ There will be visual editor and json
→ click on visual editor and follow the steps

Scanned with ACE Scanner

Inline Policy:-

→ To create it, go inside the user and click on Add inline policy

MFA to User:-

→ go inside user
→ click on security credentials
→ select Virtual MFA device
→ select one of the list of compatible applications
→ scan QR code, then assign MFA

Roles:-

→ giving permissions to services.

eg:- Role for EC2

→ select AWS service
→ Permission policies
  (i) S3 full access
  (ii) RDS full access
→ give Role name → create a role

Scanned with ACE Scanner

→ Then attach it any particular EC2 instance.

User groups :-

→ click on user groups

→ Give group name

eg :- L1 support

→ create a new policy to that group or add existing policy to that group

→ create group

H/W :-

① Create 3 buckets $S3_1$ $S3_2$ $S3_3$

② Create a user

③ Create a policy that, the User has acces only on $S3_1$ and $S3_3$, but not on $S3_2$

Scanned with ACE Scanner

Scenario!

How can we send bulk emails/marketing of IT products?

→ We need to have EC2 instance
→ os
→ SMTP (Simple mail transfer protocol)
→ Cost increases

★ Email marketing is used in IT

→ For this we have SES in AWS

SES

Simple Email Service

→ It is a service given by AWS, to send bulk emails for marketing

Advantages!

→ We are not creating EC2 instance
→ not installing SMTP
→ Not using third party subscription plan
→ We are not using third party s/w

Scanned with ACE Scanner

→ Cost is reduced for all these.
→ But SES is not free, but it can reduce the considerable cost
→ It is disabled by default
→ when you enable it, it can give you free tier i.e only 200 emails/24h i.e 1 email/sec

whenever you enable, it shows sandbox environment, it means you can do whatever you want in this, If it is getting succes and running good, It is like POC (Proof of Concept)

→ when you give POC to the client, then it satisfied to the client,
→ Then that will be deployed in ~~testing~~ prod environment

→ We can increase the limit
→ upgrade to production Acess

Scanned with ACE Scanner

→ increase the sending quota i.e for
eg:- 50,000 emails/day
→ It is cheaper than any other thirdparty service
→ If you are upgrading to production acces, they are giving 62,000 emails/month is free.

Suppose you are sending 70,000 emails/month
How they will charge you ?
→ For 62,000 is free
→ next 12,000 emails; they will charge like for every 1000 mails → 0.10$ only

Sandbox :-
It has some conditions, to use this sender email and receive email has to be verified. (must and should)
→ It is not possible, So organisations are upgrading to production accey

Creation of SES

→ Search for SES in AWS account
→ Amazon Simple Email Service
→ click on it
→ create identity
→ For organisation select ⊙ Domain
→ For single email select ⊙ Email address
→ Now we are going with email address
→ provide your email i.e the person who is sending emails
→ click on create identity
→ Now send email to the employer / customer
→ select the sender email and choose a Send email that
→ In that select ⊙ formatted
→ In Scenario → select [custom

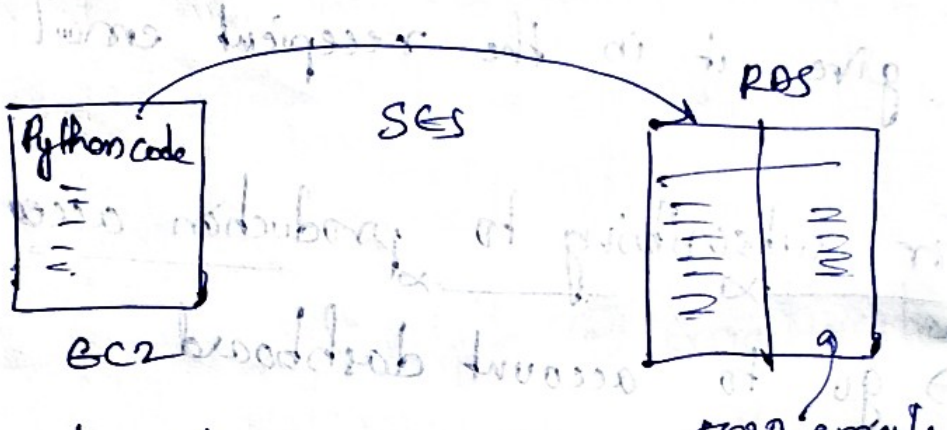
→ The provide recepient email address
→ send test email.
→ To send a email to group of users create a DL (Distribution list) and give it in the recepient email address

For subscribing to production access
→ go to account dashboard
→ Request production access
→ Then they will talk to you, and they will upgrade your request.

Scenario:
→ For the end users email information is not available, as distribution list.
→ whenever users are registrying in the website, they give their details, all the details are stored in a separate database in the form of tables

→ Now you can write a python program to automatically access the database from SES.

→ For this we need role



① We need Role to access the services

② Programmatic access we need for users

→ Programmatic access gives Access key

→ Using access key in my program I can access the emails in the database and send bulk emails

For above scenario we have

- SMTP settings

→ Create SMTP credentials

→ It is taking me to IAM user

→ Create

→ It will create SMTP credentials

Scanned with ACE Scanner

→ Copy those credentials and give to developers

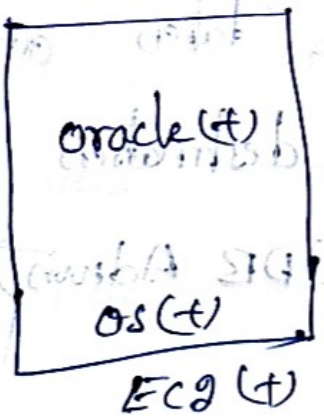
→ They will use this credentials in their programs, and they will send bulk emails

DataBase

→ It is a s/w, with this raw data is stored in structured (tables, rows, colums) format. by indexing.

→ We can retrive the data with the help of indexing very quickly.

Challenges of maintaining a DB in our own Virtual machine



(i) oracle DB admin

(ii) OS admin

→ Cost increase to maintain a DB

There are not free,

① If the DB, of EC2 machine is going down we can't access the data

SPOF → Single point of failure

→ To avoid this, we are maintaing another EC2 it is called as standby database server

② Backup :-

→ Every organisation maintain a backup policy

minute backup
hourly "
daily "
monthly "
yearly "

③ Restore DB

→ The backup is taken and it will be stored in a destination server.

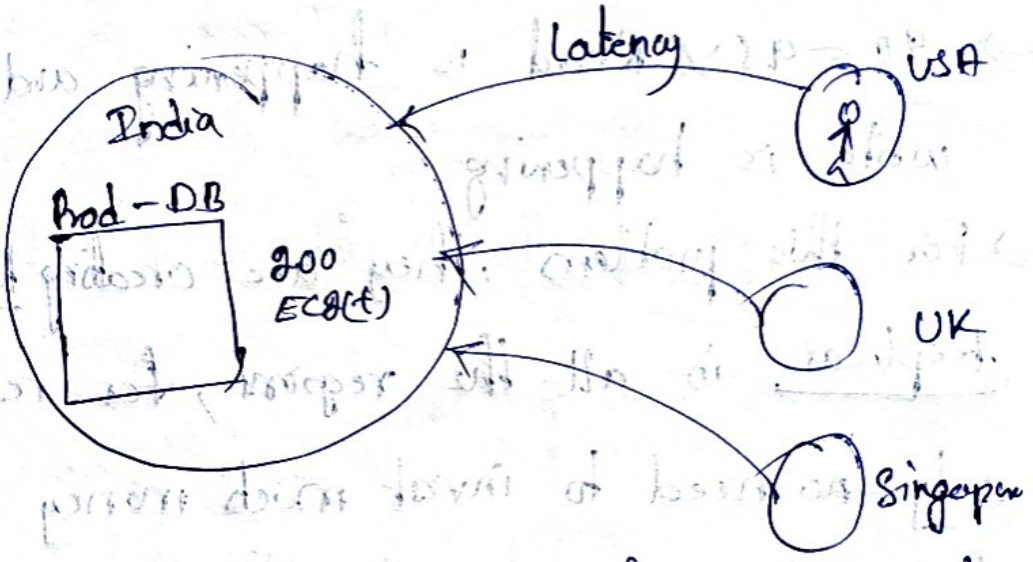
DBA (DB Admin)
↓
.sh
↓
ops (team) → start doing backup

Scanned with ACE Scanner

→ These backups will take 5-6-8 hrs, while doing it fails at 99%, again we have to start doing.

→ We have to start fresh backup, for this we have to clean backup servers.

(4) Scenario:-



→ To access the financial DB which is in India from other regions, latency will increase ∴ performance will be low, transactions will go down

→ To decrease the latency we have to create same DB servers in all the regions and again cost increase

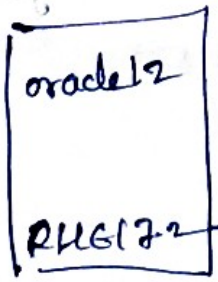
eg:- ATM

→ insert card → It will check with DB and give

Scanned with ACE Scanner

the response if it matches → Reading DB
→ PIN → Read
→ statement → Read
→ mini statement → Read
→ Balance enquiry → Read
→ withdraw → write
→ 90 - 95% Read is happening and 5% write is happening
→ For this problem, they are creating Read Replicas in all the regions, for reading only no need to invest much money and they will maintain 1 or 2 Primary DB servers in each region, by this cost is reduced and there will be no latency.
→ latency ↓ → performance ↑

(5) O/S administration



→ upgrade → O/S patching
→ security patching

Scanned with ACE Scanner

(6) oracle 12 → upgrade.
→ DB patching
→ DB security patching.

(7) All of a sudden oracle server is not running again it is a problem, Compatibility issues

→ As a customer, we are maintaing the DB, there all are the challenges he has to face

→ For all these problems to maintain we have a service, i.e RDS

RDS:

Relational DataBase Service

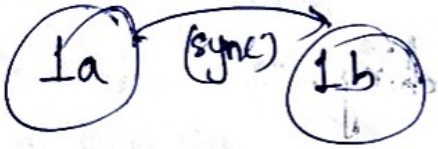
→ It takes care of all the challenges

→ select the readily available databases in the RDS,

→ It will automatically takes care of compatability issues

Scanned with ACE Scanner

→ check ☑ multi Available Zone Deployment



when you click this a DB is created in (1a) zone and another standby DB is created in (1b) zone and data is always synchronous with (1a) zone DB

→ Create Read Replica → select destination (UK)

→ To take DB backup

☑ automated backup → open window and specify time for backup

→ Restore is happen automatically.

→ It is a managed DB service

→ RDS is a service given by Aws to create database, to store raw data in structured format.

Scanned with ACE Scanner

RDS

Creation :-

→ click on create database

→ click on standard create

because it has, for options than easy create

→ Engine option

→ select any db engine

eg:- oracle, mysql

Templates :-

⊙ Production { for production purpose }

⊙ Dev/Test { for develop/testing purpose }

⊙ Free tier { for dev free applications and testing purpose }

→ settings

DB instance identifier

give name for your db

→ Master username

dbadmin / anyname

Scanned with ACE Scanner

DB instance class

⊙ Burstable classes (include t classes)

include previous generation classes

→ It give t class for lower cost and some are free also, thats why we are choosing this one

select t2.micro

storage :-

give some storage based on your requirements

Note:

(*) Storage Autoscaling

while we are creating we created with 10GB, but some transactions are happening we require more gb than what is there with us.

→ Then it automatically increases the storage space,

Scanned with ACE Scanner

→ It is not a free service

Availability & durability

→ To create a standby db instance is created in another zone

→ It is not free service

Public access

→ Don't enable it

mysql
mariadb
Amazon Aurora } port no is 3306.

oracle → 1521

→ Create a new security group and attach to it.

Database authentication

- Password and IAM db authentication
  → for this create a user, and give full access of RDS
- Password authentication
  → master password will be for this

Scanned with ACE Scanner

Backup retention period → 1 day

→ for everyday 1 backup will happens automatically, the older backup will be removed and new latest backup will be kept or retained.

Backup window

at what time you want to take backups set the time here, depends on your organi-sation.

To give the time in UTC, there will be online converters from ITC to UTC then specify the time

Intuit
→ The maintainace window and backup window should not overlap

Deletion protection

☒ enable it always.

Scanned with ACE Scanner

→ To access database from your laptop you have to install mysql workbench

→ for different different db's there are different client s/w's

for oracle → sql developer, toad

→ copy the url of dB and paste in the mysql workbench at the place of host ip, give port no, and password click on test connect

Scanned with ACE Scanner

Scenario:-

→ The IT requirements or the IT bussiness starts from the client requirement

Client → 4 env → Dev, QA, stage and Prod

→ We have 4 environments

→ 30 servers for Dev in 3 different zones in mumbai

Dev (20)

10 × mumbai × 1a × ubunt × key × Sg
10 × " × 1b × " × " × "
10 × mum × 1c × ubunt × key × Sg

QA (60)

20 × mum × 1a × ubuntu × key × Sg
20 × " 1b " " "
20 × " 1c " " "

PreProd (150)

50 × USA × 1a × ubuntu × key × Sg
50 × " × 1b × " " "
50 × " × 1c × " " "

Prod (300)

100 × USA × 1a × ubuntu × key × Sg
100 × canada × 1b × " " "
100 × africa × 1b × " " "

→ Total we require 540 servers

Scanned with ACE Scanner

To create a instance, we have to go through 7 steps

AMI → Inst. type → config → Gbs → tag → Sg → key pair

→ To create 540 instance, we have to go for 12 times all these steps

→ After doing all these things, you have done a mistake i.e Dev servers in wrong region, then client won't agree, and client will say please remove them,

→ The moment you launch Dev servers, 30 servers will be in the ownership of Dev manager, You have to get approval from him, billing will be there, getting the approval will take 1 week also,

→ Then manager have to write what is the impact of errors, conduct meeting, get approval,

→ Then you can remove wrong instances after the approval only.

Scanned with ACE Scanner

→ These all are human errors, manually.

→ To automate this activity Aws given us "awscli"

awscli

→ The awscli is a programatic access

→ It uses scripting

→ It has a user

→ For user → We go with IAM

→ with IAM → Create User → with programatic access → attach a policy (i.e EC2 full acc

→ In every organisation

→ The moment you click on create user it give two access key

AK : ****

SAK : xyz2tc

→ Now open a cli config server/ bash shell it will be present in every organisation.

Scanned with ACE Scanner

→ After loging into that server, first check whether there is awscli

```
aws --version
```

```
yum install awscli -y
```

→ After installation, then configure cli user

```
aws configure
```

→ AK: paste your access keys here
→ SAK: "
→ Default Region: ap-south-1
→ In which region you have to create servers, specify here
→ output format: text/json/table

→ It asks all the four question, answer it

→ Now to create 30 instances in mumbai

```
aws ec2 run-instances ami-id ami-0b52314425
instance-type t2.micro vpc-id <> a-zone ap-south-1a
count 10 --region ap-south-1a tag <>
sg-id <> ,key-name <>
```

Scanned with ACE Scanner

→ In this way, we can create by a single command.

(*) should you give all the parameters in an order step by step?

→ No

Advantages

→ execution time is less

→ Reusability

Drawbacks! –

→ If you are doing error, to rectify it automatic "roll-back" is not possible in aws cli

→ We have to delete it manually after getting approvals

→ To overcome all these errors, the aws given us a service called cloud Formation

cloudformation

It will help us to provision the Infrastruct

- re
→ We need to write 12 steps to create 540 instances.
→ It is letting us to write only 1 script
→ To create all 540 instances only one script.

Scenario:
→ suppose, the 1000 servers are running in singapore are prone to disaster, then you have to replace them very soon.
→ In this scenario's we go with these scripts and setup all the 1000 servers just in mater of minutes

→ To create a same environment, replica environment very quickly, you can create script in cloudformation

→ IAAC, Infrastructure as a Code
→ IAC is a service given by AWS ie

→ Script languages are allowed in cloudformation, they are

(i) json

(ii) yaml

→ The cloudformt calling the script as template (ie predefined) blue print.

Syntax

filename.json

{

cloudformation version → It is a mandatory line

{

Dev

{

Resources → mandatory section

{

Type : : aws : : ec2 : : instance

{

properties

{

ami-id

instance type

}

}

}

}

}

execution methods:-

① upload the code into S3 bucket
→ then open cloudformation service
→ then open first execution method ie S3
→ copy object url, copy into cloudform
-ation and then click [Next], then
code will be executed

② As a local file upload directly into
cloudformation.

* what is the significance of S3 bucket
incase of cloudformation?
→ S3 will be working as compiler
→ There will be inbuilt compiler in S3

→ S3 will execute the code,
→ If the code executed upto Dev, QA
and pre-prod in pre-prod the code has
got error and stopped executing
further.

Scanned with ACE Scanner

→ Then what about the created servers, 90 server

Dev → 30, QA → 60, → total 90 created out of 540

There are two options for this problem

1) Rollback on failure

→ It roll's back everything that is created, removes everything created

Advantages

→ No need of taking approvals from manager

→ No need of listening the songs of clients etc

Scenario:

I don't want to delete the 90 servers which are succesfully created

preserve succesfully created resources

Scanned with ACE Scanner

→ We can create any service ie EC2, S3, RDS etc for this we require IAM role is required, and attach it to cloudformation

Timeout valve:

→ If there is 2000 lines of code, to execute the code if can set time, under this time only it has to execute

awscli creation

→ We need a server

→ Now create a IAM user

→ for this we go with programmatic access

→ copy AKS, SAK

→ Now enter into a server, then

→ check whether awscli package is present or not

aws --version (or) yum install awscli -y

→ Now give command as

```
aws configure
```

AKS : paste here

SAK : "

- Default region : give region name

Default o/p format : text

→ lets check, can we access S3 bucket?

```
aws s3 ls
```

It shows permission ~~denied~~ denied because we have only ec2 permissions for this user.

→ We can access EC2 only

→ To see the list of instance

```
aws ec2 describe-instances --region region-name
```

→ To see the o/p of above command in different formats

```
aws ec2 describe-instances --region region-name --output /table/json
```

Scanned with ACE Scanner

→ Create a ec2 instance

```
aws ec2 run-instances --image-id ( )
--count 1 --instance type t2.micro --key-name
( ) --security-group-ids <  >
--subnet-id <  > --region <  >
```

json script for ec2 formation

filename.json

```
{
"AWSTemplateFormatVersion": "2010-09-09",
"Description": "Demo ec2 instance creation",
    {
    "Resources":
            {
            "DEVELOPMENT":
                    {
                    "Type": "AWS::EC2::Instance",
                    "Properties":
                            {
                            "AvailabilityZone": " ",
                            "ImageId": " ",
                            "InstanceType": "t2.micro",
                            "KeyName": "ut",
```

Scanned with ACE Scanner

"SecurityGroups": [" "]

}

}

}

}

→ Now go to cloudformation

→ create cloudformation stack

→ ◉ Template is ready

→ You have written the template

◉ Use a sample template

→ Aws provides it

→ Select any one

◉ Amazon S3 url ◉ upload a template file

→ Next

Permissions! —

IamRoleName ▽ [▽]

→ Without creating a role will it execute or not?

→ Yes, every cloudformation template will be

Scanned with ACE Scanner

using a default role

stack failur options

- Rollback all the stack resources
- Preserve succesfully provisioned resources

→ Notifications

→ Timout policy

→ Protection termination

→ Next

DNS

→ It gives a unique name for your ip address,

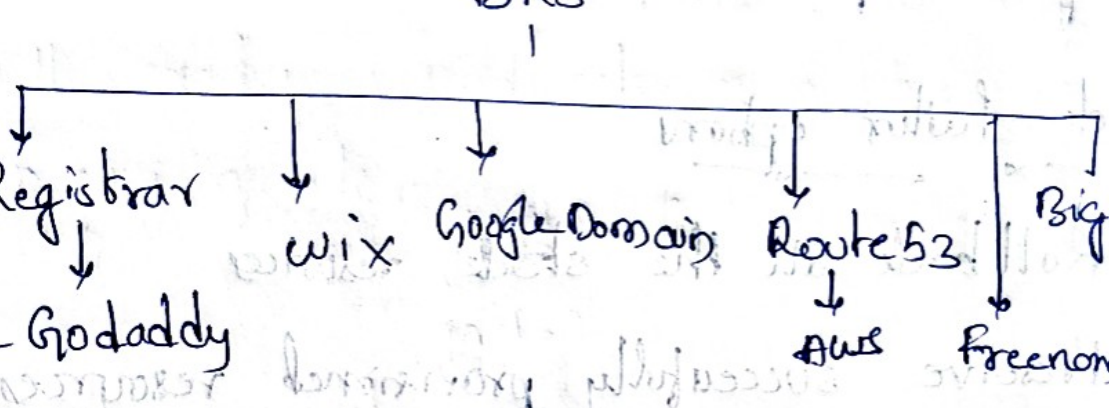
→ We can't remember ip's, and load balancer's url,

→ It is Domain Naming System

→ We need to provide a meaningfull name unique name that can be used across the globe is called DNS.

→ To maintain all the names, there must be a central body, it is DNS

Scanned with ACE Scanner



→ The distributor or dealer of domain names inside DNS is called Registrar.

→ There are many registrar's, Very famous one is eg:- Godaddy

→ In Registrar, You will register your domain name or you will buy your domain name

→ We will buy a domain name / registering a domain name, we are going to attach to our load balancer.

eg:- whenever a user types the domain name and hit enter, then

Domain Name → Registrar —will forward→ Actual Server

Scanned with ACE Scanner

Scenario-1

eg:- Infosys has its domain name www.infosys.com.

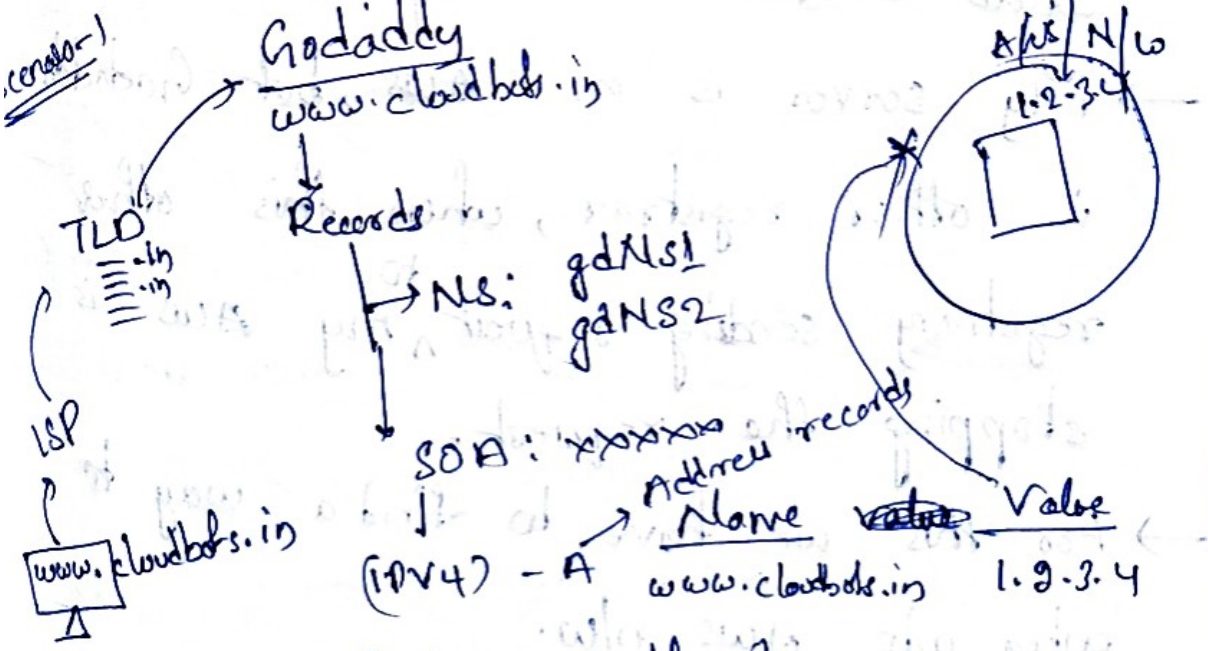
If they want to migrate their business to Aws How to manage their domain name?

Scenario-2

→ If there is a startup company, without buying a domain name outside, they can buy it from Aws itself

→ By it from Route53

→ Every domain name has expiry, Every year you have to renew, If you are not renewing your domain name, they will release it to others



SOA → Start of Authority

→ We registered our website name in Godaddy
→ It creates record sets, of number two
naming records
eg: gdNS1 (Name server records
gdNS2 "
and it creates SOA: xxxxxx xxxx
→ whenever a user types www.cloudbots.in
this request reaches to ISP
→ Then ISP looks into a TLD in which
all the domains ends with .in,
→ Then TLD sends the request to Goddady,
then if the name is registered in it,
it will check and sends to server
where website is running.
→ My server is on Aws, but Godaddy
is other registrar, when this other
registrary sending request to my Aws is
stopping the request,
→ For this we have to find a way to
enter into Aws n/w.

Scanned with ACE Scanner

Route 53

It is a service given by AWS to configure or to register a domain name or a existing domain name

→ 53 is the port no: of DNS in linux os

→ In this I am going to created a hosted zone with the same exact domain name in Route53.

→ This Route53 thinks that it is fresh name and gives ~~two~~ four NS records

→ If 3 NS records are going down, we can find answer with 4 NS record, it gives high availability

→ It is faster, than others because of 4 NS records

It has SOA : xxxx.com also

→ Now still request is going to Godaddy only, how can make that request, directly

Scanned with ACE Scanner

comes to Route 53

→ For that we replace Godaddy NS with AWS NS

aws NS1
aws NS2
aws NS3
aws NS4

→ Now create address records in Route 53

A | Name | value
www.cloudbob.in | 1.2.3.4

SOA → where you are buying the domain name, your request should start from them.

Routing policy :-

→ You are having single server or many servers for one website, then how can you make entry point for many servers as one entry point

① Simple routing :-

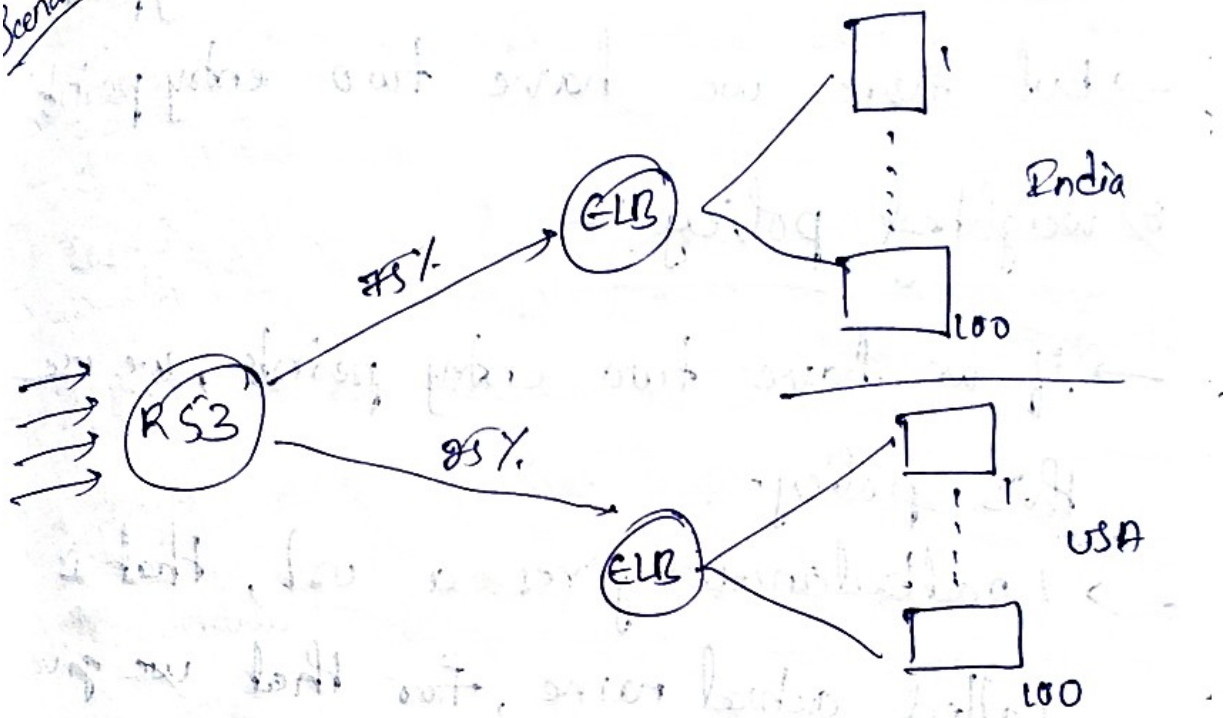
→ If you have one entry point, you can

Scanned with ACE Scanner

use this policy

eg:- single server, loadbalancer, S3 bucket url

Scenario:-



→ For my business 100 servers are running in India with 1 LB and 100 servers in USA with 1 LB.

→ Loadbalancer has a limit, it can work in one region only, i.e we can't include USA region servers in India ELB.

→ Then R53 has to forward request to one of the loadbalancer;

Scanned with ACE Scanner

Here we have two entry points, for this we can't go with simple routing policy.

→ For simple routing, we are resolving name to IP address for one entry point

→ But here we have two entry points

8) weighted policy :-

→ If we have two entry points, we use this policy.

→ Loadbalancer gives a url, that is called actual name, for that we give another name i.e alias name

→ Now I create one alias record for loadbalancer in India and one for another alias record for USA load balancer

→ let's say I have more users in India but less in USA,

→ For this I create 75% alias record for

Scanned with ACE Scanner

India and 25% alias record for USA
Loadbalancer

→ eg:- If R53 is receiving 100 requests
route 53 sends 75 requests to India and
25 requests to USA

③ Geo location routing policy:-

→ Based on the location, the request should
reach ELB of that region only.
eg:- Indian request should reach India LB
USA " " USA "

→ For every country there is a separate
IP range.
→ This information is having with Route53,
then the moment you type the
url, it reaches to R53, then it
understands from which region and
sends that request to that region
LB only

Scanned with ACE Scanner

④ Latency based policy:-

→ If my user is in India and typing my website

→ To reach India LB, the latency is 7ms and to reach USA it has latency of 5ms then R53 send request to USA LB only eventhough he is in India, because he choosed this policy.

→ The R53 calculates the latency based on heartbeat

⑤ Failover routing policy:-

→ If India LB is more efficient, then I choose it has Primary LB, then all the requests go to Primary only

→ In case Primary LB fails then only the requests are send to secondary LB in USA

Scanned with ACE Scanner

→ If your organisation is maintaining IPv4 we have to create Adress record of

'A' → IPV4

for IPv6 → "AAAA" record

Name —resolve→ IP

IP —resolve→ Name

↓

PTR → pointer Record

→ For an organisation while buying a domain name, you have to mail service record

eg:- sudarshan@bcs.com

→ If you already have mail service we can configure in AWS with 'MX' record 'mail server exchange record

→ for mx also → Name, Value will be there

Note:- whenever you are creating Record sets and Domain name it will take (24-48) hrs, to reflect into all TLD's and to access it from browser

Attaching a domain name to IP

→ Create instance and it has one IP
→ Launch a website in it
→ Now attach that IP to domain name
→ Now lets say we bought domain name now I have attach that domain name from Godaddy to R53
→ My domain name is wwww.cloudbots.in
→ For existing domain name in a registrar, to configure that name into Aws we create hosted zone
→ search for Route 53
→ select create hosted zone

Domain Name: cloudbots.in

comment: my org

create

→ Now it will give you two record sets

NS → ns-1449.awsdns-53.org.
ns-1005.awsdns-61.net.
ns-206.awsdns-25.com.
ns-1592.awsdns-07.co.uk.

Scanned with ACE Scanner

SOA ns-1449.awsdns-53.org. awsdns-hostmast
er.amazon

→ Now copy NS names and paste it
Godaddy NS.

Note!- copy till the name only leave the dot, copy it carefully.

Scenario-1 → don't copy the trailing dot

→ Now create RecordSets in R53

→ click on create RecordSet

Name: www.cloudbots.in

Type: A - IPv4 address

Value: paste IP of your server (pub-ip)

Create.

Note!- copy LB url, S3 url in the type section if and only if you have only one LB and one S3 only.

Scenario!-2

→ If my customer is not typing www only typing cloudbots.in, then we will loss the bussines

Scanned with ACE Scanner

→ For this create an alias RecordSet

Name: cloudbots.in

Type: A - IPV4-address

Alias ◉ Yes ○ No

Alias Target: www.cloudbots.in. ▼

↓ select it from drop down

Scenario-3

How to create a subdomain name.

Name: dev.cloudbots.in

Type: CNAME - Canonical name

value: paste DNS name given by Aws for server

eg: ec2-3-110-169-95 ap-south-1 .compute.amazonaws.com

→ Create

Scenario-4

Two entrypoints, one is India and other in USA

Scanned with ACE Scanner

→ Create RecordSet for Loadbalancer of India

Name: www.cloudbook.in

Type: A - IPV-4 address

Value: paste url of LB of India

Routing policy: [weighted ▼]

weight: 75

SetID: IndiaLb

→ Create

→ Create Record set for Loadbalancer of USA

Name: www.cloudbook.in

Type: A - IPV-4 address

value: paste url of LB of USA

Routing policy: [weighted ▼]

weight: 25

SetID: USALb

→ Create

Note:- Create Records based on Geolocation, latency, failover for the same website.

Scanned with ACE Scanner

(λ) LAMDA functions

Scenario:

→ In an organisation, we have 2000 servers and 8000 have moved to cloud, they are telling that we have to check the server whether it is open to internet (24×7×365 day) if any remediate to company n/w or company IP range?

→ We can check whether the server is open to internet through security groups,

→ open the security groups and check in inbound rules, if the IP address is (0.0.0.0/0) → it means open to internet. If not it must have company IP address eg 10.1.2.4/8

→ We can't check 8000 security groups manually,

→ Let's say we checked 1000 security groups manually, and while we are reading in next 1000 sg's mean while anyone may opened the server to internet in above 1000 servers.

→ How to restrict someone opening the server to internet? manually it is not possible to do/check sg always once?

→ For this we do automation, to do so we need to write program, for this we ~~have~~ have to buy

(i) server
(ii) IDE
(iii) OS, licencing
(iv) Java/python etc
(v) We need to have programmatic access
(vi) to automate aws account using python we need to import a library called 'boto3'

→ To write a code in a server, it takes
2% of RAM, time, HDD, cpu,
→ 95-98% of the time these all resources
are sitting idle, for this we have to
pay to AWS

→ To overcome all these problems, they
invented a service called Lambda,
which is serverless programing platform
It is a platform as a service (Paas)

→ We will assign a user to Lambda
function, he opens it and selects a
program for automation; they are giving
a readily available IDE

→ Lambda function wants to access Sg,
S3, EC2 etc, create role and
attach to Lamda function

→ whenever it is executing the code only
utilizes the resources we can cut

Scanned with ACE Scanner

cost.

→ I can control executing with timeout

* because when the lamda is executing it is charged (charged at runtime)

(*) How/when to execute the lamda function

(1) scheduled based execution

→ it is like cron jobs, i.e every ~~next~~ hourly

→ At 8.00 am, I am scheduling, it will run next time at 9:00am, next 10:00 am ...., In between can anyone open my server to interact, then how can we trace it, there is a chance of opening the server.

Ans-2:

→ Because you should it for hourly, next 9:00am only it will execute and trace if any server is open and ~~x~~ remediate to secure n/w.

Scanned with ACE Scanner

Case-II

→ If I have scheduled it for * * * * *
i.e for every minute, every hour. It is going to charge because we are executing for every execution it charges

(2) Event based execution

eg:- creating a service it is a event, starting a server, doing anything in aws is called event.

→ By default events/log files are not captured in Aws, for that they have a service it is called CloudTrail

Cloudtrail:-

→ It is not a free service,

→ The moment you enable it, it creates a S3 bucket and storing all the 23 regions log files into it

→ If anybody is changing the Security

group, what is the event for it ?

→ Event name : ~~Security group add~~

Security Group Ingress

→ I have to use Lambda function on security Group Ingress

(*) Now how can Lambda service is going to search in the S3 bucket for SecurityGroup Ingress

→ For this, after writing the logs in S3, cloud trail is going to stream (directly insert logs) into a service called Cloudwatch logs

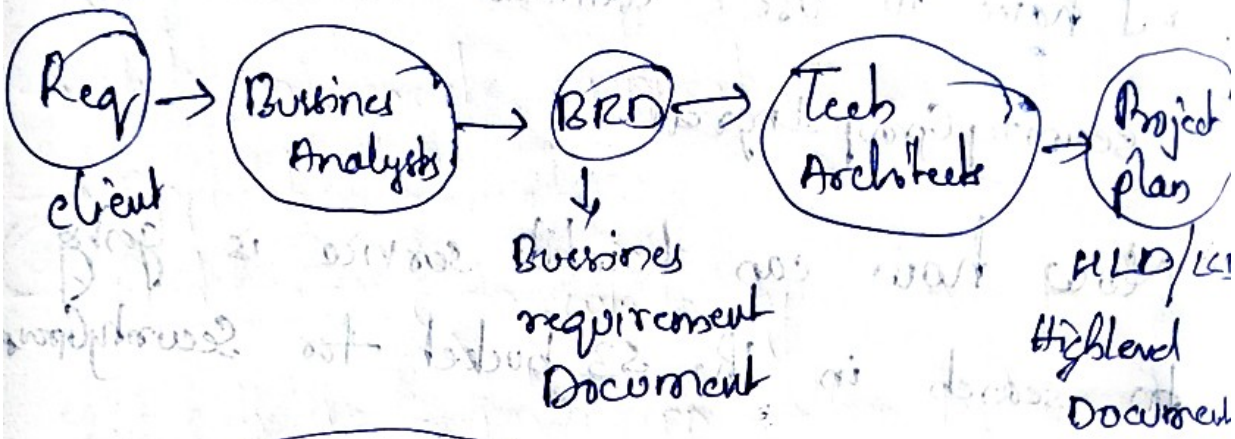
→ Now after writing logs into cloudwatch logs, then I will trigger Lambda function and from dropdown [Security Group Ingress] and save

→ whenever any event is done, my lambda function executes, whenever the lambda function executes the error is identified and solved immediately

Scanned with ACE Scanner

Elastic Beanstalk

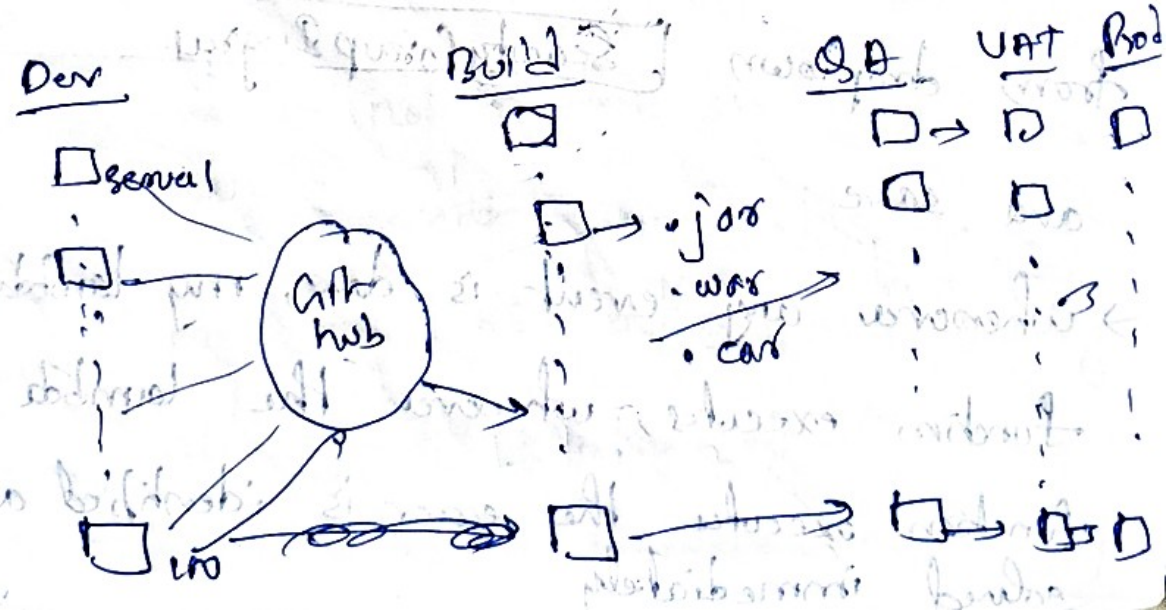
→ For a business in IT, we have a following cycle.



→ work distribution

→ Then cloud engineer has to create Dev, Prod build, QA/Deployment, UAT/preprod environments

UAT → User Acceptance Test environment



Scanned with ACE Scanner

Roles of cloud engineer

1. Create environment
2. Deployment
3. Management

→ To solve above challenges, there is a service to quickly setup entire environment is Elastic Beanstalk

→ Note:- This is only for web application

→ Upload your build file in that service it will deploy in all environments

→ You can deploy, undeploy, redeploy very quickly.

→ It will not charge for this service but it charges for the resources utilized in this service.

→ You simply upload the code and beanstalk automatically handles the deployment, from

capacity provisioning, loadbalancing and automatic scaling to web application, health monitoring, with ongoing fully managed patch and security updates

→ click on create application

→ Application name

→ Platform (on which you have to deploy)

eg: If it is java app → then deploy on tomcat

→ Application code:

- sample application
- ✓ upload your own code

→ source code origin

Version label [simplejavaapplication/any name]

- local file
- Public S3 url

→ If your file is in your machine then select localfile

→ If it is in S3 bucket, select S3 url

Scanned with ACE Scanner

→ create application

For redeploying another version of war file just select

upload and deploy

Then it will remove older version and deploy later version

Configure more options

→ You can select instance types
→ No. of instances
→ Deployment policy
  ○ Rolling update
  ○ All at once {it is dangerous, don't go with it}

Security :-

→ Add the keypair, it by default won't create any keypair
→ Create RDS data base and attach to it
→ The cloudformation is advanced compared to elastic beanstalk.

Scanned with ACE Scanner

Delete

→ With the help of cloudformation we can create no. of resources, no. of environments and deploy any kind of application

Delete

→ But with elastic beanstalk only 1 environment only for webapplications.

VPC

Virtual Private Cloud

→ It is nothing but a n/w.

→ AWS gives default n/w to any account.

→ We can create our own n/w.

→ The privat IP range is given created by the n/w.

→ We can create our own IP range i.e pvt IP through VPC

→ We are isolating the n/w from AWS given n/w i.e we are isolating our resource from AWS n/w, and keeping them separately in our own n/w.

Scanned with ACE Scanner

→ We have n/w classes A, B, C, D, E

→ D, E are used by govt sectors (space & research)

→ For commercial we can use only A, B, C

→ AWS is using class B range n/w

→ With the help of mask bits we can create no. of IP's, i.e that many no. of servers we can create

eg:- AWS IP range 172.31.0.0/16 → mask bit
↓
address bits

→ It creates 65534 ip address

→ depending upon the requirement in your organisation create IP addresses

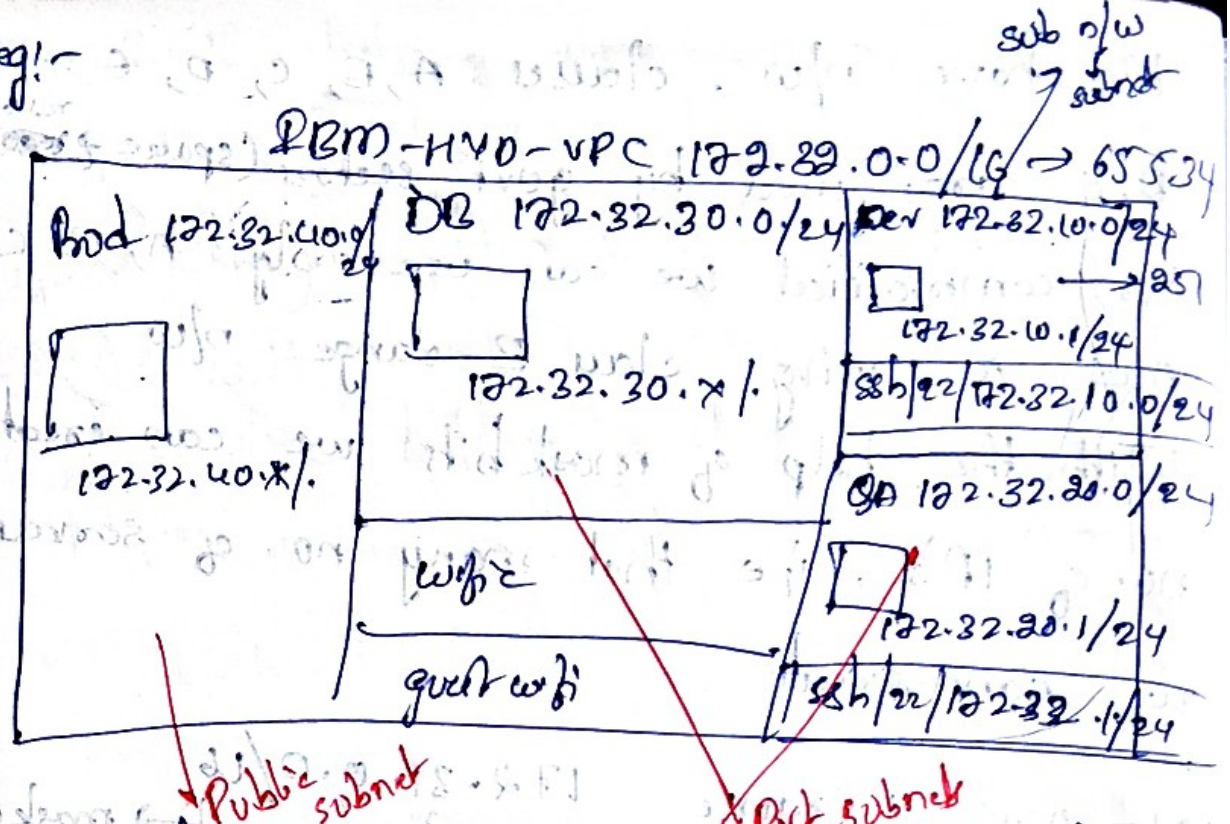
→ 172.31.0.0/16 → 65534

172.31.0.1 172.31.1.1 ... 65534
0.2 1.2
0.3 1.3
⋮ ⋮
0.255 1.255

eg:- We are creating a VPC for IBM

IBM-HYD-VPC

eg:-



→ We have created a our own n/w through VPC

→ I am creating a subnetwork in my main n/w, it is called as subnet. because I don't want all the Dev servers accessed by Testing team.

→ Like this I can create no. of n/w's in a main n/w and restrict the access by others and outside of this n/w too

→ eg:- 172.32.10.0/24 → 254 IP address from 3 IP address are used by internal purposes and only 251 can be utilized

Case 1:-
To communicate with the Dev servers we need the IP range of Dev servers only

→ To restrict the access from one subnet to another subnet, I am creating a security group

ie Sg | ssh/22/172.32.10.0/24 |

The Dev team members coming through the above Sg only can access the Dev Servers

Case-2
→ In QA environment, I create a subnet

g. IP range 172.32.20.0/24

→ All the team members can communicate in QA servers those who have Sg of

| ssh/22/172.32.20.0/24 |

Case-3
→ The IP range of 172.32.20.0 can't access

Scanned with ACE Scanner

the IP range of 172.32.10.0

→ because the Sg is restricting it, it acts as gate to each subnetwork.

Note!- The starting point of your project is networking .i.e create n/w and seperate subnetwork for each environment

Case-4:- Prod/live environment.

→ Their servers are accessed by the end users

→ But my prod server is in my own n/w.

→ The end users will be outside the n/w, to access it we need public IP

→ We have to expose prod servers to public for this we have to enable public IP for prod servers

Scanned with ACE Scanner

Internet Gate Way ! (IGW)

→ It acts a gatekeeper to every n/w.
→ It has permissions that whom to allow/ whom to restrict to enter into your VPC.
→ Create one IGW and attach it your VPC
→ It allows you to connect to your VPC only.

Note: We never expose Dev, testing, DB servers to public, for that we never enable public ip for these servers

→ We are not enabling these servers to public ip address, it means these are accessed by our employees only, then we can say it as private subnetwork. and Prod environment we can call it as Public subnetwork.

→ By default all the sub nets are pvt only, when we are exposing to public, it will become public subnet

Scanned with ACE Scanner

Scenario-1

(*) From outside the VPC, We have to login into Dev server and install httpd.

→ Now a employee who is doing work from home, login to Dev server and install httpd

→ For this he will enter into Dev server through IGW and loged in.

→ then he is installing httpd for that he again has to connect to internet and to download the httpd package and install it.

→ For VPC there is only one way to enter into it, and there is no out way for go out and connect to internet i.e there is a path to enter into it but there is no path to go out

→ without the out path he can't connect to outside n/w and download httpd.

→ For VPC there is only inlet and no outlet.

Scanned with ACE Scanner

→ From inside the n/w, to connect outside n/w we have routing tables.

Routing table:- (Path)

→ It creates a path from pvt VPC to internet, it is like a way/outgate to VPC

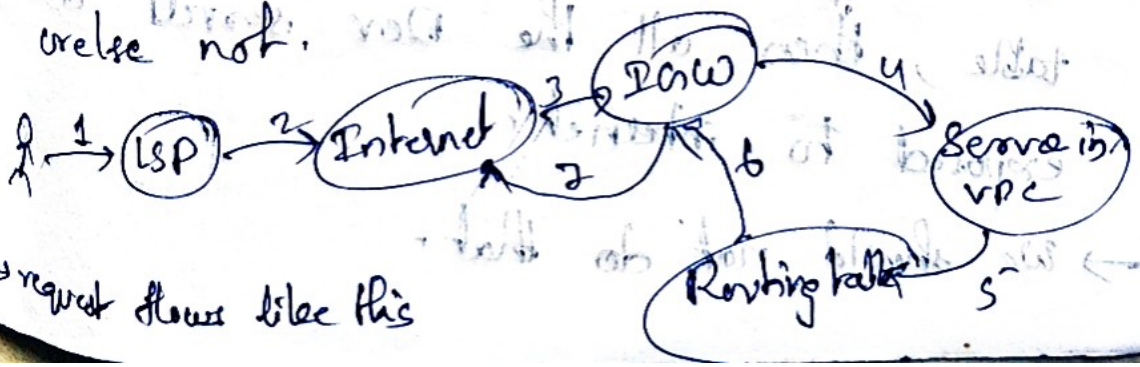
→ To connect a pvt subnet to outside there must be outgate/path

eg:- Route table

0.0.0.0/0	IGW

→ it means you can connect anywhere and anyone in the internet through IGW only.

→ All the servers in the subnet, contacting routing table, if the servers has access in that table they will be connected outside orelse not.



→ request flows like this

Scanned with ACE Scanner

→ Firewall is a security group at instance level and route table is security group at subnet level

Note:- IGW allows outsider to connect the resources in your n/w, and Route table allows inside servers to connect to Internet

Scenario:
To connect prod member to Dev team for that create a Security group in the Dev subnet as ssh/22/172.32.40.1/24
then only prod team servers can communicate with dev team servers

→ From Dev to connect internet create a routing table in the Dev subnet.

→ If we give IGW id in routing table, then all the Dev servers are exposed to internet.

→ We should not do that.

Scanned with ACE Scanner