

You
Need to
Know

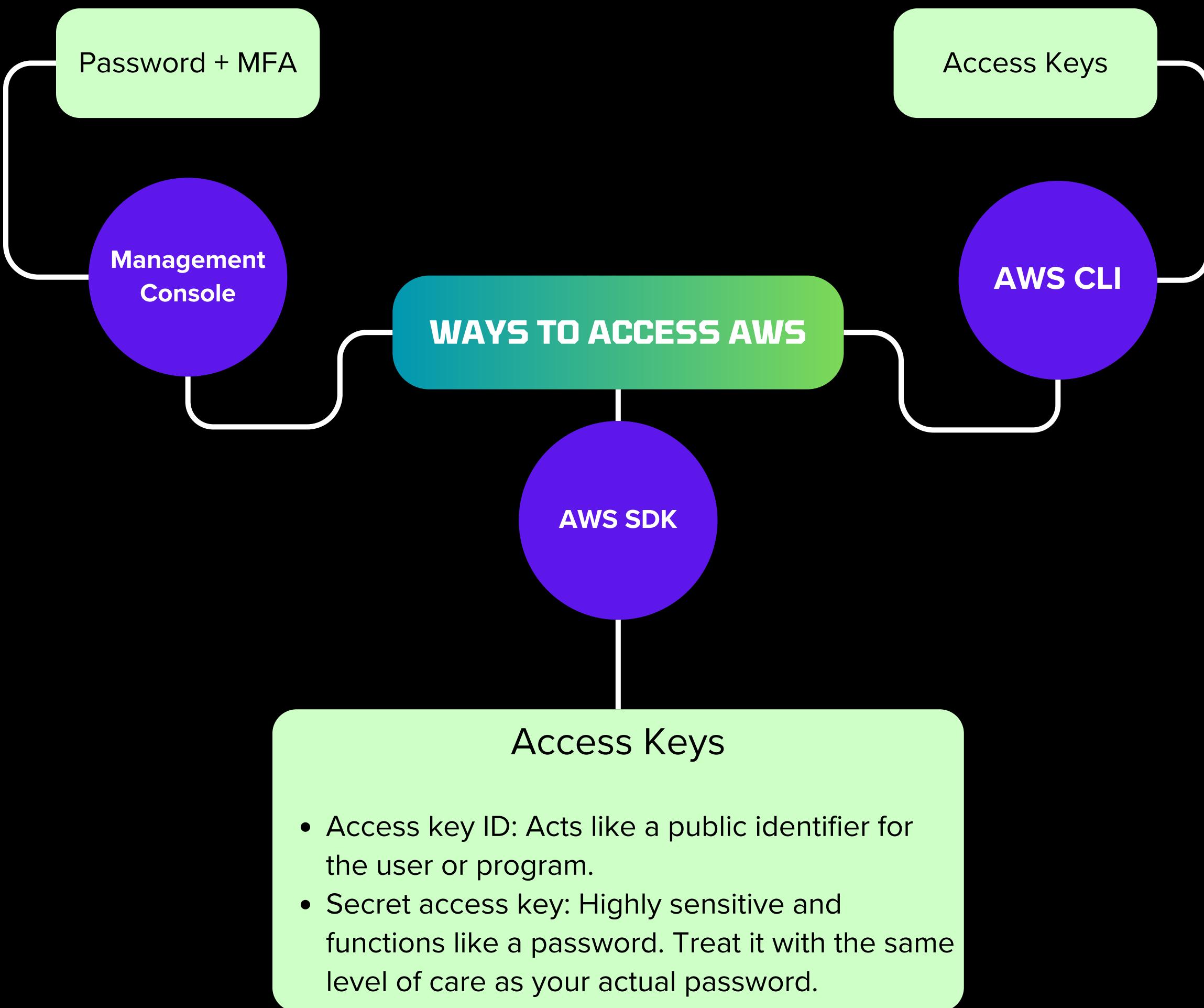


Your Ultimate SAA Exam Guide





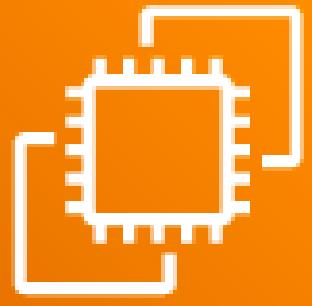
Ways to Access AWS



Rajan Kafle



REPOST



EC2 INSTANCE TYPES



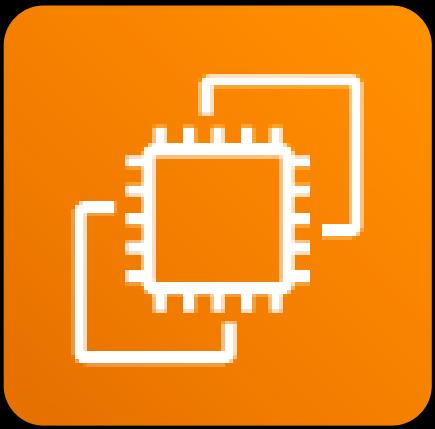
- **General purpose (t)**: Balanced compute, memory, and networking for web servers, development, and small databases.
- **Compute optimized (c)**: High processing power for scientific computing, video encoding, and high-performance web servers.
- **Memory optimized (m)**: Large amounts of RAM for databases, in-memory analytics, and enterprise applications.
- **Storage optimized (u)**: High-performance storage for data warehousing, log processing, and content repositories.
- **Accelerated computing (a)**: Specialized hardware (GPUs, FPGAs) for machine learning, video processing, and scientific simulations.



Rajan Kafle



REPOST



EC2 INSTANCE TYPES

Different AWS EC2 Instance Types

General Purpose

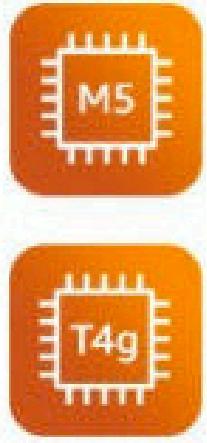
Compute-Optimized

Memory-Optimized

Storage-Optimized

Accelerated Computing

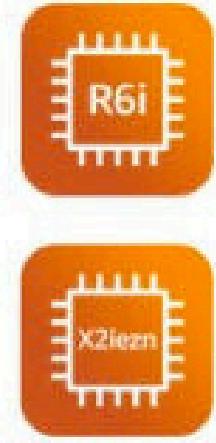
M and T families



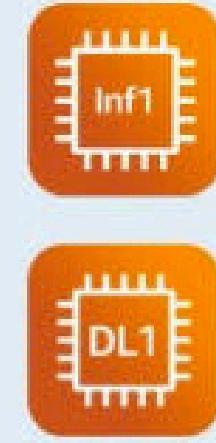
C-type family



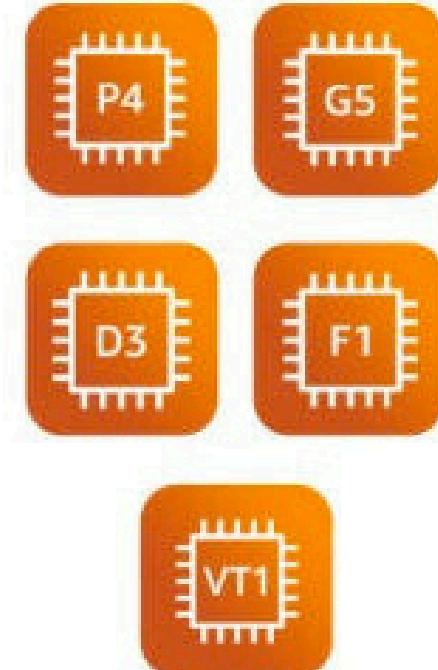
R and X families



I, D, and H families



P, G, D, F, and V families



Naming Convention: AWS has the following naming convention

m5.4xlarge

m: instance class

5: generation (AWS improves them over time)

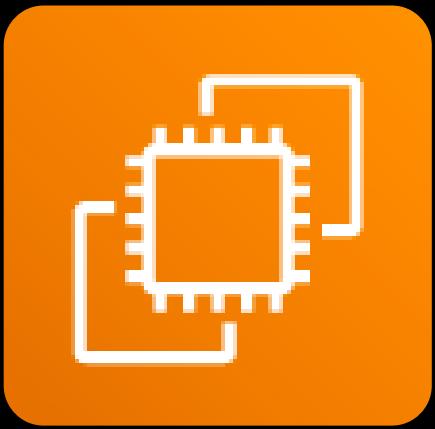
4xlarge: size within the instance class



Rajan Kafle



REPOST



Security Groups (like Firewall)

Inbound and Outbound Traffic Control:

Security groups define rules that specify which traffic is allowed to enter (inbound) and leave (outbound) your resources.

Allow/Deny Mechanism

Security groups operate on a “default deny” basis. All traffic is blocked unless explicitly permitted by a rule.

Regulate access to ports + authorized IP ranges

You can define security group rules to regulate access to specific ports for authorized IP ranges.

Can be attached to multiple instances

Security groups can be attached to one or more instances within a VPC.

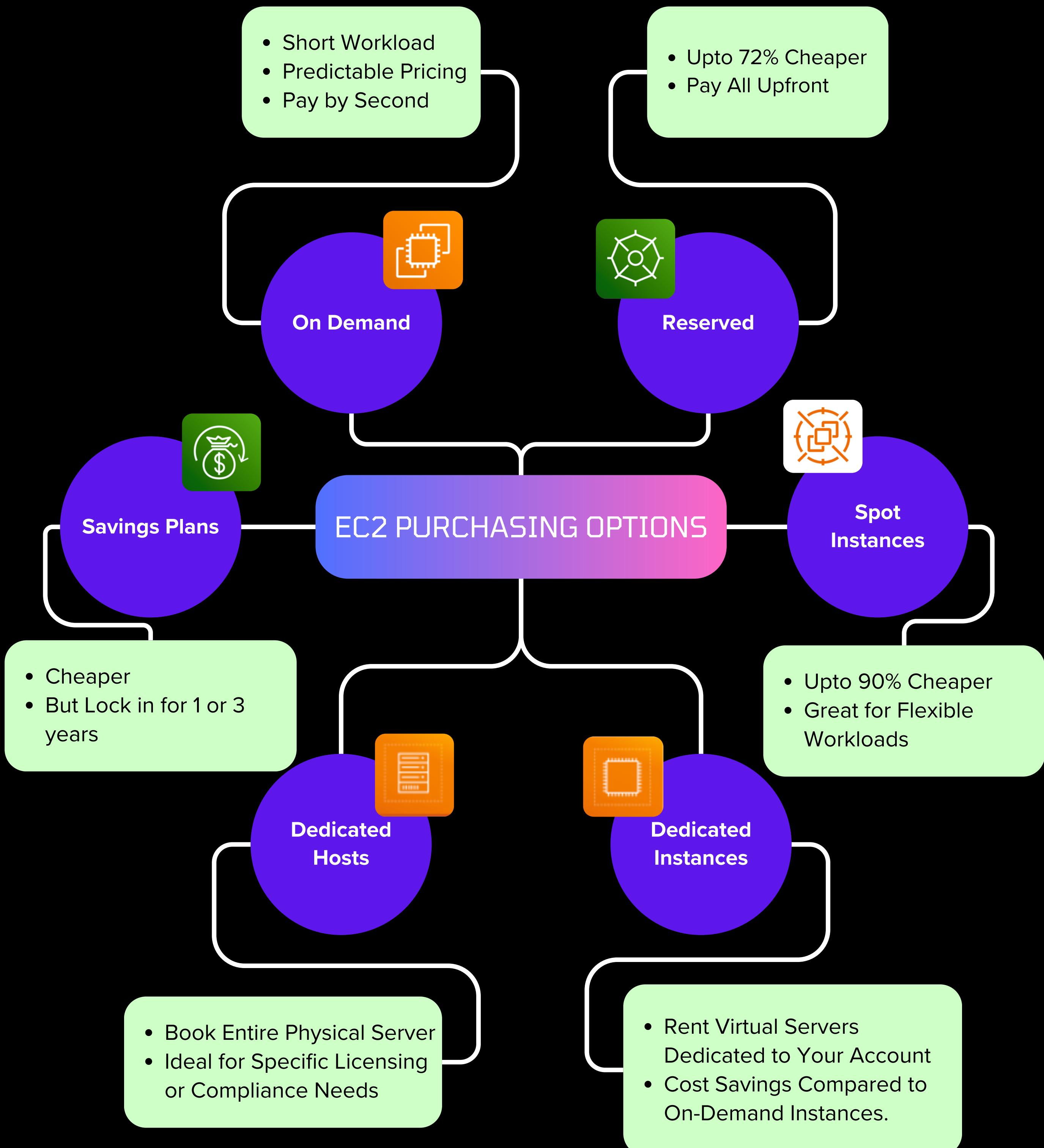


Rajan Kafle



REPOST

EC2 instances purchasing option





Elastic IPs (5 per account)

Each AWS account has a default **limit of 5 Elastic IPs per region**. This means you can have a maximum of 5 static public IP addresses associated with your resources in a specific region.

Public IP addresses assigned to EC2 instances are dynamic.

When you stop and start an instance, its public IP address typically changes. This can disrupt your application's functionality if it relies on a specific IP for access.

The Solution for Fixed Public IPs:

Elastic IPs are static public IP addresses that you can allocate to your EC2 instances or network interfaces in a VPC. They remain constant even when you stop and start your instances.

Note:

You Can only Attach Elastic IP to **One Instance**. Elastic IPs are designed to provide a static public IP address for a single resource at a time.



Rajan Kafle



REPOST



Elastic network interface (ENI)

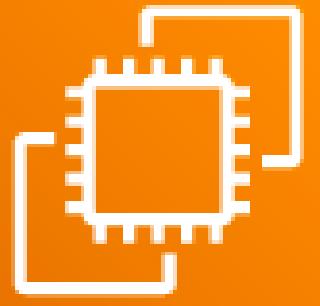
- **Virtual Network Interface:** Acts like a network card for your EC2 instances.
- **Private IP Addressing:** 1 primary + optional secondary private IPs for internal communication.
- **Elastic IP Attachment:** Optional for assigning a public IP to a private IP.
- **One or More Security Groups + One MAC Address.**
- **Attachable to EC2 Instances:** Provides network connectivity.
- **Bound to a Specific Availability Zone (AZ):** Limited mobility across AZs.



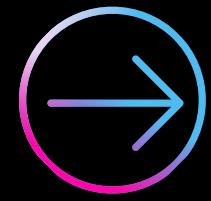
Rajan Kafle



REPOST



Hibernation in EC2



Hibernation: Sleep Mode for Your EC2

- **Saves Instance State:** Saves RAM contents to EBS storage when hibernating.
- **Stops Instance:** Powers down the instance like stopping normally.
- **Faster Startup:** Resumes from saved state for quicker restarts.
- **Cost-Effective:** Saves on compute costs while maintaining instance state.
- **Requires Enabled AMI:** Only works with AMIs configured for hibernation.



Rajan Kafle



REPOST



Elastic Load Balancer

Managed Load Balancer – AWS guarantees it will be working + takes care + upgrade. Health Checks for eg. ec2 instance (status 200).



Rajan Kafle

ELASTIC LOAD BALANCER



Security Practice

- ELB acts as a "doorman" for your EC2 instances.
- Security groups ensure only the ELB can "knock" on your instances' door.
- No direct access to EC2s, keeping them secure.

Routing

- Different Routes can send to different Target Group
- Target Group: EC2 Instances, Lambda Functions, IP, etc

Sticky Sessions

- Redirect to Same EC2 Instance / Target Group Only
- Use Cookie with Expiration Date → Ensure User doesn't Lose Session Data

Cross Zone Load Balancing

- Distribute evenly across all registered instance in AZ
- Not Supported on Classic Load Balancers (CLB)

SSL/TLS ELB

- Load Balancer uses X.509 certificate.
- Can manage certificates using ACM (upload own certs + integrate).
- HTTPS listener: specify default listener.

Connection Draining (CLB)

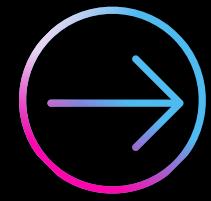
- Time to complete in-flight request while instance is deregistering or unhealthy.
- Between 1 and 3600 seconds (default=300s).
- Set to LOW value if requests are short.

Server Name Indication (SNI)

- Solves problem of loading multiple SSL certs onto one web server (multiple domains)
- Newer protocol, requires client to indicate hostname of target server in initial SSL handshake
- Server then finds correct certificate, or return default one
- Doesn't work on classic LB



Auto Scaling Group



Auto Scaling Intro

- Can Scale In and Scale Out as traffic increase/decrease.
- Scale In: Less Traffic, Remove Instances.
- Scale Out: More Traffic, Add Instances.

Dynamic Scaling Policies

- Target tracking scaling → want average CPU to stay around 40%
- Simple / Step Scaling → when cloudwatch alarm triggered, add X units
- Scheduled Actions → increase min capacity from X to Y time on monday

Predictive Scaling

- Generate Forecast using ML → scale accordingly
- Eg. Cpu Utilization, Request count per target, average network in/out

Scaling cooldowns

- After Scalling Activity, Cooldown period 300 seconds.
- ASG Doesn't Launch / Terminate instances → allows metrics to stabilize



Rajan Kafle



REPOST



Virtual Private Cloud (VPC)

Understanding subnet mask

- 192.168.0.0/32 → 1 address → 192.168.0.0
- 192.168.0.0/31 → 2 address → 192.168.0.0, 192.168.0.1
- 192.168.0.0/30 → 4 address → 192.168.0.0 to 192.168.0.3
- 192.168.0.0/29 → 8 address → 192.168.0.0 to 192.168.0.7
- 192.168.0.0/28 → 16 address → 192.168.0.0 to 192.168.0.15
- 192.168.0.0/16 → $2^{16} = 65536$ address → 192.168.0.0 to 192.168.255.255
- 192.168.0.0/0 → ALL IPs → whole internet

Private IP range

- 10.0.0.0 to 10.255.255.255 (10.0.0.0/8) → in big networks
- 172.16.0.0 to 172.31.255.255 (172.16.0.0/12) → AWS default VPC in this range
- 192.168.0.0 to 192.168.255.255 (192.168.0.0/16) → home networks
- Everything else are public

Default VPC on AWS

- All new AWS accounts have a default VPC
- New ec2 instances launch into defualt VPC if no subnet is specified
- Defualt VPC has internet connectivity, all ec2 instances inside it also have ipv4
- We get public n private ipv4 DNS names.

ROUTE TABLE

- Destination to target
- Eg. destination = 0.0.0.0/0
- Eg. target = igw-{rest of id} OR local → all device within network can talk to one another.



Rajan Kafle



REPOST



Virtual Private Cloud (VPC)



Subnets In VPC

- AWS reserves 5 IP addresses
 - first 4 and last 1 in each subnet
- 5 IP addresses not available for use, cannot be assigned to EC2 instances

CIDR block 10.0.0.0/24, reserved IP addresses are

- 10.0.0.0 → network address
- 10.0.0.1 → reserved by AWS for VPC router
- 10.0.0.2 → reserved by AWS for mapping to Amazon-provided DNS
- 10.0.0.3 → reserved by AWS for future use
- 10.0.0.255 → network broadcast address. AWS does not support broadcast, thus reserved

Internet gateway

- Allows resources e.g. EC2 in VPC to connect to public internet

Bastion Hosts

- Can use Bastion Host to SSH into private EC2 instances
- Bastion == Public Subnet which is connected to all other private subnets
- Bastion Host security group must allow inbound from internet port 22
- Security Group of EC2 instances must allow security group of bastion host.

NAT Gateway (Network Address Translation)

- AWS managed NAT with higher bandwidth, availability etc
- Pay per hour for usage/bandwidth
- Requires Internet Gateway (private subnet → NATGW → IGW)
- Created in specific AZ, uses elastic IP

Network Access Control List

- Checks if inbound traffic is allowed, if allowed, send inside subnet
- Checks if outbound traffic is allowed, if not allowed, will block
- Like firewall, controls traffic from and to SUBNETS
- ONE NACL per subnet, new subnets assigned default NACL

VPC Peering

- Privately connect 2 VPCs using AWS private network
- Makes them behave as if they are in same network
- Must not have overlapping CIDRs
- VPC peering connection NOT transitive + must update route tables.

VPC Endpoints (AWS PrivateLink)

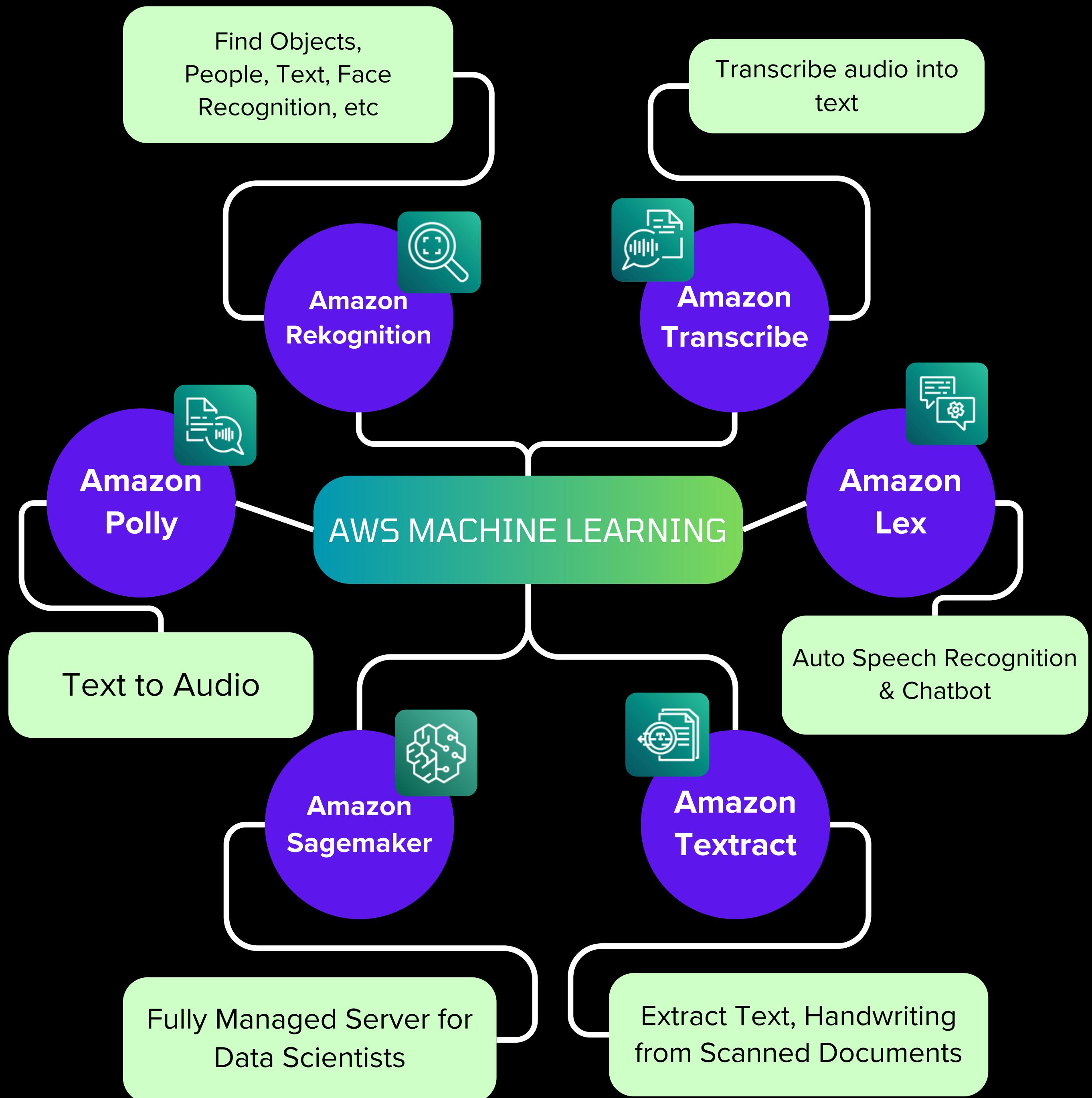
- VPC endpoints allow connect to AWS services using private network instead of public

TYPES OF ENDPOINTS

- Interface endpoints (powered by privatelink)
- Gateway endpoints



Machine Learning in AWS





AWS S3

S3 Intro

- Allow people to store objects (files) in buckets (directories)
- Buckets defined at regional level
- Objects (file) → key is FULL PATH → no folders in s3 buckets , just names with slashes
- S3 buckets PRIVATE by default, can access using SIGNED URL

S3 Security

User-based Access → specific user from IAM + resource-based

Bucket Policy

- JSON based policies
 - Effect: Allow/Deny
 - Principal: who can access
 - Resource: which resource this applies.

S3 Versioning

- By default Disabled, enabled at bucket level (all obj in bucket)
- Keeps all previous versions of your objects when you upload new ones.

Naming Convention

- No uppercase, no underscore, not IP, must start with lowercase letter or number
- Must NOT start with prefix xn — , must NOT end with suffix -s3alias

Public S3 + static website hosting

- Make bucket public
- Enable static website hosting

S3 Replication → lower latency

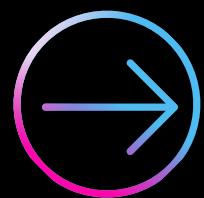
- Cross region replication (CRR)
- Same region replication (SRR)
- bucket can be in different AWS accounts, must give proper IAM permissions
- ONLY NEW OBJECTS ARE REPLICATED
- no chaining of replication

S3 durability & availability

- 99.999... % (119s) across multiple AZs
- 99.99% availability → not available 53 minutes a year.



S3 Storage Classes



- Frequently Accessed Data
- Low latency and high throughput for fast retrieval.

- Less frequently accessed. Rapid access when needed
- Standard Infrequent Access → 99.9% availability → disaster recovery, backups
- One-Zone Infrequent Access → 99.5% availability → secondary backups

Standard

Infrequent Access

Glacier

Glacier Instant Retrieval

S3 STORAGE CLASSES

- Very large archiving / backup
- Very low price for storage + object retrieval cost

Glacier Deep Archive

Glacier Flexible Retrieval

- Long term storage (cheapest)
- Standard (12 hours), bulk (48 hours)
- Minimum storage 180 days

- Expedited retrieval (1–5 minutes)
- Standard retrieval (3–5 hours)
- Bulk (5–12 hours) (free)



AWS S3



S3 Intelligent Tiering

- Auto switch storage classes for you, but charges you spare amount of monitoring & auto-tiering fee.
- Moves objects across access tiers + no retrieval charge

S3 Requestor Pays Bucket

- Requester who downloads object pays networking cost
- Helpful for large datasets
- Requester must be authenticated AWS user

S3 Performance

- Multi-part upload (recommended for large files > 100mb, must for > 5gb)
- Divide into parts → parallel uploads → converge into S3

S3 Transfer Acceleration (make upload faster)

- Increase transfer speed, transferring file to AWS edge location.
- Forward data to S3 bucket + compatible with multi-part upload



S3 Lifecycle rules

- Transition Actions → Configure obj to transition after X days
- Expiration Actions → configure obj to delete after X days

S3 Event Notifications

- S3:ObjectCreated, S3:ObjectRemoved, S3:ObjectRestore
- Amazon EventBridge for event handling

S3 Byte Range Fetches

- Retrieve on partial data
- Parallelize GETs by requesting specific byte ranges
- Better resilience in case of failures

S3 select, S3 Glacier Select

- Retrieve less data using SQL n server-side filtering
- Less Data == Less Cost

S3 batch operations

- Perform Bulk Operations on existing S3 objects.



AWS CloudFront



Cloudfront (CDN)

- Content Delivery Network
 - cache content at the edge
- 216 edge locations globally
- DDOS Protection
- Integration with AWS security stuff

CloudFront origins

- **S3 Bucket:** Stores your content in an Amazon S3 storage bucket.
- **Application Load Balancer:** Distributes traffic across multiple instances in your application.
- **S3 Website:** An S3 bucket configured to serve static content directly.
- **Any HTTP Backend:** CloudFront can fetch content from any server that speaks the HTTP language.

CloudFront Simple Terms

- **Edge Locations:** Think of these as mini data centers strategically placed around the world. Each location holds a cache of your content.
- **Cache Invalidation:** Allows you to remove specific content from the cache using the Create Invalidations API. This ensures users always get the latest version.

CloudFront Simple Terms

- **CloudFront's Cache Behaviors** are like traffic directors for your content. They allow you to define specific rules for different parts of your website.
- **Time-to-Live (TTL):** Set how long objects stay in the cache before checking for updates with your origin server.



Rajan Kafle



REPOST



AWS Global Accelerator



Global Accelerator lets your application go as fast as possible through AWS network to reduce latency.

Unicast IP

- One Server holds one IP address

Anycast IP

- All servers hold same IP address
- Client routed to the nearest



AWS global accelerator works using anycast IP

- 2 Anycast IP created for your app
- Anycast IP send traffic directly to edge locations
- Edge locations send traffic to your app
- Performs Health Check + DDOS protection due to AWS shield.
- EXPENSIVE



Rajan Kafle



REPOST



AWS Serverless



Developers don't have to manage servers anymore → they just deploy functions.

THERE ARE SERVERS. YOU JUST DON'T MANAGE / PROVISION / SEE THEM



Rajan Kafle



REPOST



AWS Lambda

AWS Lambda Intro

- No Servers to manage, runs On-Demand, Scaling is Automated, Limited by Time
- Pay Per Request & Compute Time → free tier of 1mil requests + 400k gb of compute time

LIMITS

- Memory Allocation 128mb — 10gb
- Max Execution Time == 15 minutes
- Disk Capacity is 512mb to 10gb
- Deployment Size (compressed zip file): 50mb
- Size of Uncrompressed Deployment: 250mb
- Size of Environment Variables: 4kb

Lambda Quick Notes

- Lambda@Edge: More powerful (Node.js, Python), network access, regional edge locations.
- **Lambda VPC Configuration:**
 - Default: Outside VPC, no access to VPC resources.
 - In VPC: Define VPC ID, subnets, security groups, Lambda creates ENI.
- **RDS Proxy:**
 - Handles many connections to RDS (like an ALB for RDS).



Rajan Kafle



REPOST



Amazon DynamoDB



Amazon DynamoDB Intro

- Fully Managed, NoSQL,
- Multiple AZs
- Auto Scaling
- Low Cost

Read / Write Capacity Modes

- Provisioned Mode (default) → specify num read/write per second
- On-Demand Mode → auto-scale num read/write per second (more \$\$)

DynamoDB Accelerator (DAX)

- Same as Read Replica as RDS
- In-memory cache for dynamoDB + 5min for TTL
- Solves read congestion, reduce latency

Dynamodb Streams

- Log stuff when request made to dynamodb
- 24 hours retention (Kinesis 1 year)
- Limited number of customers (kinesis high number)

Dynamodb Global Tables

- Makes DynamoDB table accessible with low latency in multiple regions.
- active-active Replication

DynamoDB Time To Live (TTL)

- Auto delete items after expiry timestamp (reduce stored data)

DynamoDB backup for disaster recovery

- Continuous backup using Point-In-Time Recovery (PITR)
- Optionally enabled for last 35 days
- Recovery process creates a new table

On-demand backups (managed using AWS Backup)

- Full backups for long-term retention until explicitly deleted
- Recovery process also creates a new table



API Gateway

API gateway Intro

- Lambda + API Gateway → no infra to manage
- Supports WebSocket, handles security, swagger/openapi imports to quickly define api
- Cache API responses, generate SDK n api specifications

ENDPOINT TYPES

- Edge-Optimized (default) → for global clients, requests routed through cloudfront
- Regional → for clients within same region
- Private → can only be accessed from VPC using ENI

SECURITY

- User Auth
 - IAM roles
 - Cognito (external users)
 - Custom authorizer
- Custom Domain Name HTTPS
 - Use AWS certificate manager (ACM)



Rajan Kafle



REPOST



AWS ECS & Docker



Docker Flow

- Dockerfile build → docker image
- Upload docker image to docker repository eg. DockerHub or Amazon ECR
- ECS / EKS pull image from repository

AWS Tools

- Amazon Elastic Container Service (Amazon ECS) → container platform
- Amazon Elastic Kubernetes Service (EKS) → managed kubernetes (opensource)
- AWS fargate → serverless container platform (works with ECS, EKS)
- ECR → stores container images

ECS Infrastructure options

- AWS fargate (serverless)
- EC2 instances (manual configuration)
- External instances using ECS anywhere

Fargate launch type

- Launch Docker containers on AWS, no need to manage servers
- Just create task definitions
- To scale, increase number of ec2 instances

Roles in ECS

- ECS task role
 - Need to access other AWS services eg RDS
- EC2 instance profile
 - Role for permission when pulling image from ECR

Create task definition

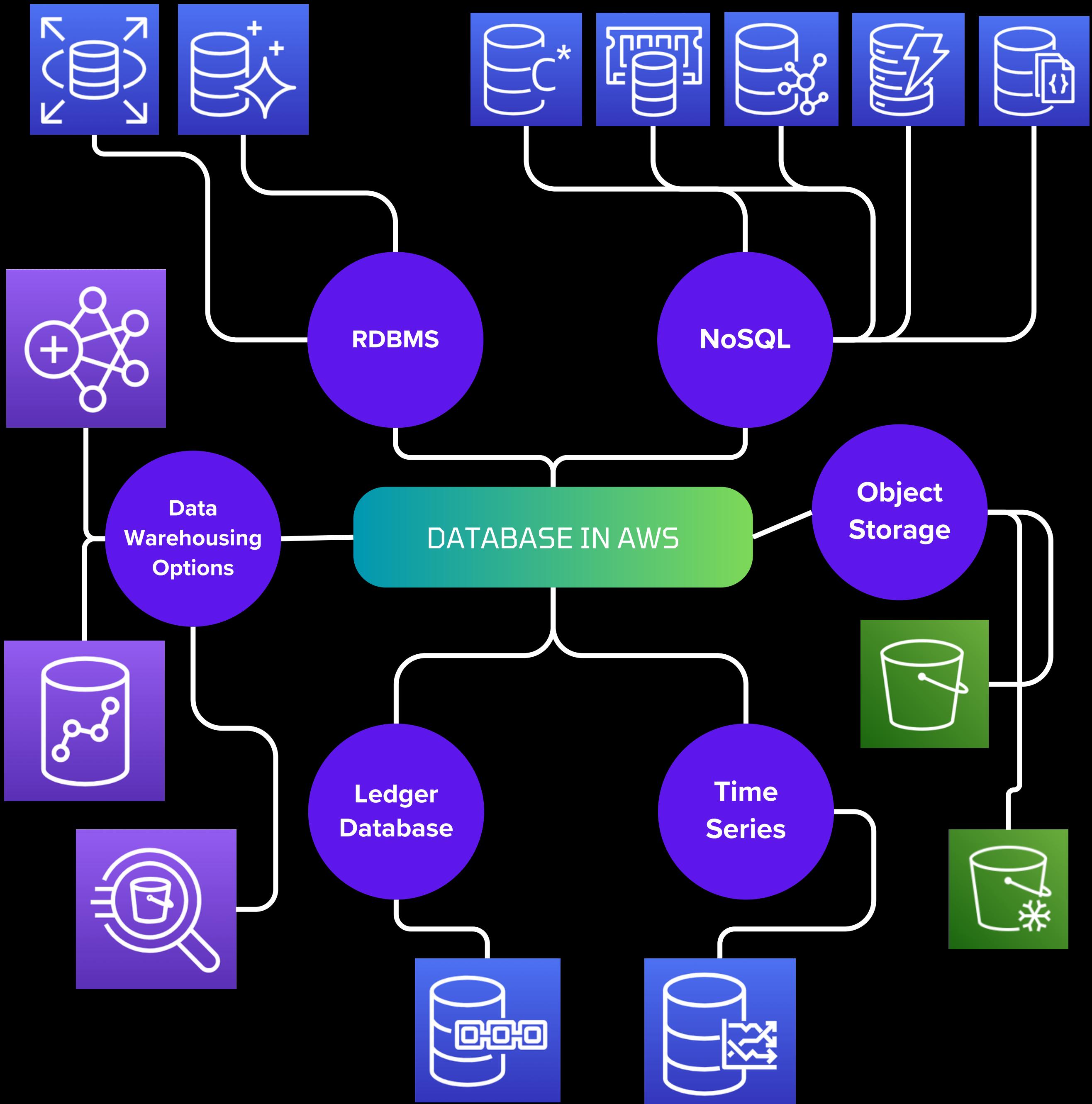
- ImageURI → copy from ECR
- Define ports, protocol etc, env variables, healthchecks
- Task role → if task accessing some other AWS service

ECS Scaling

- ECS capacity providers
- Cloudwatch metric
 - create alarm → auto scales ECS



Database in AWS



Database Recovery in AWS



- How often backup data → what data loss can we sustain
- Eg. Can sustain loss of 1 day worth of data, so back up every 1 day

- How much downtime can app sustain

Recovery Point Objective

Recovery Time Objective

Backup Strategies

DISASTER RECOVERY IN AWS

Backup & Restore

- Backup and Restore
- Pilot Light
- Warm Standby
- Hot Site / Multi-Site approach

Warm standby

Multi-site / Hot Site

- Cheapest, but Highest RTO & RPO
- Server Die, Restore & Backup → Delays
- Schedule Regular Snapshots

- Full System UP & Running, but at minimum size
- Master DB → Data Replication
- Upon Disaster, scale to production load

- Least RTO n RPO
- Keep exact copy of production app running
- Very Expensive \$\$\$

Database Recovery Tips



- EBS snapshots, RDS automated backups, snapshots
- Regular pushes to S3, etc

- Use route 53 to migrate DNS over to failover
- RDS multi-AZ, elasticache, EFS, S3
- Site-to-Site VPN as recovery from Direct Connect

Backup

High Availability

Replication

Automation

DISASTER RECOVERY TIPS

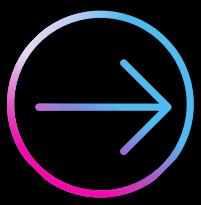
Chaos

- RDS replication (cross region), AWS Aurora + global databases
- DB Replication from on-premises to RDS
- Async Replication between master & Standby DB.
- Storage Gateway

- CloudFormation / Elastic Beanstalk to re-create whole new environment
- Recover / Reboot EC2 instances with CloudWatch if alarms fail
- AWS Lambda functions for customized automations

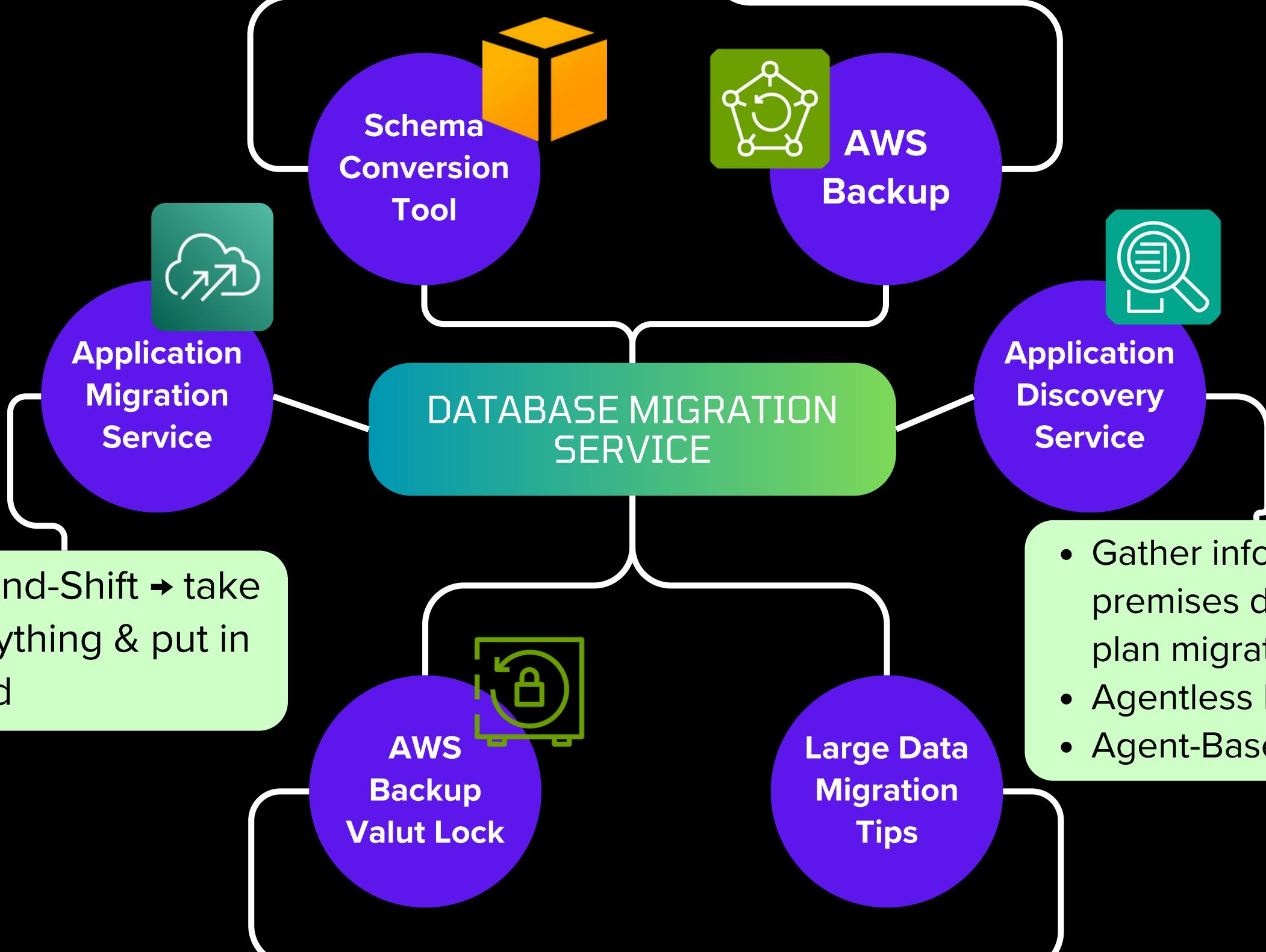
- Purposely cause disaster Eg. DDOS
 - Just to check how system reacts to this

Database Migration Service



- Convert Schema from one engine to another
- Eg. Oracle SQL to Aurora etc

- Fully Managed service, centrally manage & automtae backups across everything
- No Custom Scripts
- Supports Cross-Region & Cross-Account Backups



- Lift-and-Shift → take everything & put in cloud

- Gather info about on-premises data center → plan migration projects
- Agentless Discovery
- Agent-Based Discovery

- Enforce a WORM — write once ready many state for all backups
- Protect your data from being deleted by malicious ppl
- Even Root User cannot delete data when in place.

- Over the Internet / Site-to-Site VPN → immediate to setup, but take long time
- Over Direct Connect 1gbps → long setup time (1month), but less time transferring
- Over Snowball → 2–3 snowballs in parallel → 1 week to finish



AWS CloudWatch



Cloudwatch Metrics

- Provides Metrics for every service in AWS Eg. CPUUtilization, Networking..
- Metrics Belongs to Namespaces → Easier to Organize Metrics
- Dimension is Attribute of Metric eg instance_id, environment (up to 10 per metric)
- Can create Cloudwatch Dashboards containing metrics
- Can create Custom Metrics eg. RAM

Metric Streams

- Continuously Streams Cloudwatch Metric, send to Destination
- Can send to Kinesis Data Firehose or 3rd party service provider.

Cloudwatch Logs Agent

- Older Version of Agent
- Can only send to Cloudwatch Logs

Cloudwatch Unified Agent

- Newer, Collect Additional System Level Metrics
- Collect Logs, send to Cloudwatch Logs
- Centralized Configuration using SSM parameter store

Cloudwatch Container Insights

- Collect, Aggregate, Summarize Metrics and Logs from Containers
- Containerized Version of Cloudwatch Agent to Discover Containers in EKS.

Cloudwatch Lambda Insights

- Monitor, Troubleshoot Serverless Applications
- Collect, Aggregate, Summarize System Level Metrics



Cloudwatch Logs

- Log Groups: Arbitrary Name
- Log Stream: instances within apps
- Log Expiration Policies eg. delete logs after 7 days
- Send Logs to lots of AWS services.

Cloudwatch Alarms

- Trigger Notification for Metric
- States → OK, INSUFFICIENT_DATA, ALARM
- Period → Num of Seconds to Evaluate Metric
- Note: can use AWS- CLI to Trigger Alarm

Cloudwatch alarm targets

- Stop, Terminate, Reboot, Recover EC2 Instance
- Trigger Auto-Scaling Action
- Notify SNS

Composite Alarms

- Alarm Containing Alarms
- AND condition, OR condition etc.

Cloudwatch Contributor Insights

- Analyze Log Data, Create time-series that Display Contributor Data
- Find Bad Hosts, Identify Heaviest Network Users

Cloudwatch Application Insights

- Auto Dashboard Shows Potential Problems with Monitored Apps
- Powered by Sagemaker

CloudTrail & Config

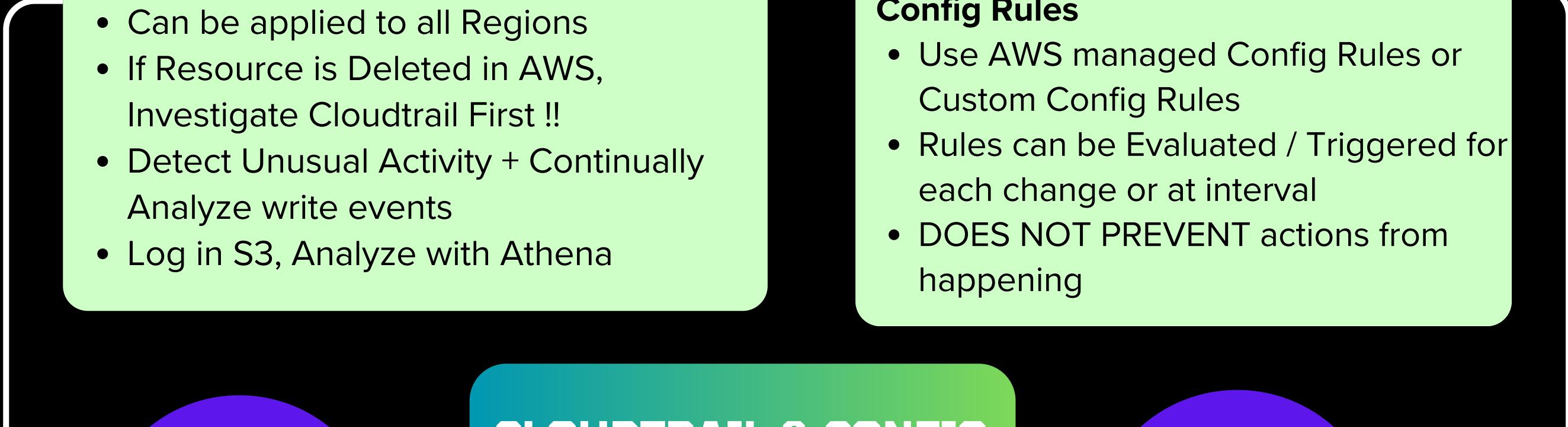


- Provides Governance, Compliance & Audit for AWS Account (Enabled by Default)
- Get HISTORY or Events / API Calls made by AWS Account
 - Console, SDK, CLI, AWS services
- Can be applied to all Regions
- If Resource is Deleted in AWS, Investigate Cloudtrail First !!
- Detect Unusual Activity + Continually Analyze write events
- Log in S3, Analyze with Athena

- Helps with Auditing & Recording Compliance of AWS Resources.
- Records Configurations and Changes over time.
- Receive Alerts for any Changes + Per-Region Service.

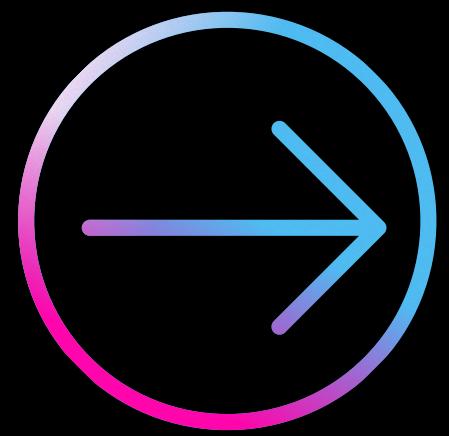
Config Rules

- Use AWS managed Config Rules or Custom Config Rules
- Rules can be Evaluated / Triggered for each change or at interval
- DOES NOT PREVENT actions from happening



Cloudwatch vs Cloudtrail vs Config

- Cloudwatch → Monitors Performance, Events, Alert, Logs
- Cloudtrail → Records Actions made by Account
- Config → Records Config Changes, Evaluate against Compliance Rules



DM Me for Full Guide

**or Subscribe to my
Free Newsletter**

rajankfl.substack.com



Rajan Kafle



REPOST



REPOST

FOLLOW FOR MORE GUIDES!

