01-fundamentals

Fundamentals

AWS Global infrastructure

- Regions, AZ, Data centers and Edge locations / Points of Presence
- Region = cluster of Data centers
  - Most AWS Services are region scoped
  - How to choose a Region : Compliance, Proximity (reduce latency), Available services and Pricing
- Edge Location : Endpoints for caching content (typically CloudFront CDN)
- Each region has many availability zones (usually 3, min is 2, max is 6)
  - Example for Region Sydney (ap-southeast-2) : ap-southeast-2a, ap-southeast-2b, ap-southeast-2c
  - Each availability zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity
  - They’re separate from each other, so that they’re isolated from disasters
  - They’re connected with high bandwidth, ultra-low latency networking

Ownership and responsibilities

- Shared Responsibility Model

Well architected framework White paper

- Six pillars
  - Operational Excellence
  - Perf. efficiency
  - Security
  - Cost opt.
  - Reliability
  - Sustainability

02-iam

IAM = Identity and Access Management, Global service

IAM is Global (Universal) Not bound to Region

Abstract

- Root account created by default, shouldn't be used or shared (Add MFA to it)
- Users and Groups
- Groups only contain users, not other groups
- Users don't have to belong to a group, and user can belong to multiple groups
- Permissions are managed by JSON documents = Policies
- Creating IAM User : Account ID / Alias / Signin URL
- Roles
- Can add tags to users (or any resources)

Policies

- Add policies to users / groups in the console
- Browse policies
- Create custom policy in the console builder (VisualEditor)
- Policies types
  - AWS Managed
  - Job Function (preconfigured permissions like PowerUser)
  - Customer Managed
- Default new User Permission : IAMUserChangePassword
- Deny always overrides Allows

Policy JSON example :

```
{
  "Version": "2012-10-17",
  "Id": "....", // Optional
  "Statement": [
    {
      "Sid" : "1",  // Optional
      "Effect": "Allow", // Allow or Deny
      "Action": "ec2:Describe*", // String or Array
      "Principal": {
        "AWS" : ["arn:aws:iam:545454455:root"] // account user role to
        which policy applies
```

```
      },
      "Resource": "*" // String or Array
      "Condition" : ... // Optional
    },
    {
      "Effect": "Allow",
      "Action": "elasticloadbalancing:Describe*",
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "cloudwatch:ListMetrics",
        "cloudwatch:GetMetricStatistics",
        "cloudwatch:Describe*"
      ],
      "Resource": "*"
    }
  ]
}
```

Password Policy

- Password length, strength
- Expiration and rotation
- Identical to name / email prevention
- Old password prevention

MFA - Multi Factor Authentication

- Virtual MFA devices : Google Auth, Authy
- U2F Key
- Hardware TOTP

AWS Access

- AWS Management console (Web)
- AWS CLI (access keys - generated from AWS console)
- AWS SDK (access keys)

AWS Cli

```
aws configure # this will prompt for Access Key ID, Secret and Region
name
# example (if permissions OK)
aws iam list-users
aws iam list-users --region .....

aws s3 ls
```

AWS CloudShell (not available for all regions)

- Web Cloud based terminal

IAM Roles for Services

- Role = like users but assigned to services to do operations on behalf of actual users (kind of machine account)
- Role: Services (example : EC2 Instance) + Permissions
- Trusted entity type - user or resource that can assume the Role
  - AWS Service
  - AWS Account (log as Role)
    - Can be related to this account or another one
    - Attach Role to user by “Trust relationships”
  - etc.

IAM Security Tools (Audit)

- IAM Credentials Report (account-level)
- IAM Access Advisor (user-level)

Identify providers

- Add new Identity Provider (SAML for AD or OIDC)

MISC

- ARN Amazon Resource Name ex. arn:aws:iam::339712790957:user/user-1

03-s3

S3 - Object storage

- S3 operates across global space (but bucket deployed regionally)
- Upload any file type
- Not used for OS or DB Storage
- Unlimited storage (volume + number of objects)
- Objects up to 5 TB
- Ensures High availability and High durability (data is spread across multiple DC)
  - 99.95-99.99 Availability
  - 11 nines durability (9 decimals)
- Standard S3
  - multi-devices / AZ redundancy (>= 3 AZ)
  - 4 nines Availability / 11 nines Durability
  - Frequent access
- Tiered Storage (different storage classes)
- Lifecycle management
- Versioning
- Security
  - Server side encryption
  - ACLs : can be granted to individual objects within bucket
  - Bucket policies (allowed/denied actions on buckets - all its objects)
- Strong read (get file, list, etc.) after write consistency
- Object can have prefixes (folder structure)
- By default, you have the ability to create a maximum of 100 buckets in every one of your AWS accounts. In case you require more buckets, you have the option to raise your account bucket quota to a maximum of 1,000 buckets by submitting a request to increase your quota.

S3 buckets

- Universal Namespace (globally unique)
  - All AWS Accounts share the S3 namespace
- URL format : https://[bucket-name].s3.[region].amazonaws.cm/[key-name]
- Successfull upload -> HTTP 200

Key-Value Store

- Key (string) / value (bytes)
- Version ID (multiple versions of same object)

- Metadata (content-type, last-modified, etc.)

Enable static web site hosting

- Specify index and error .html files

Security

- Block public access
  - Buckets are private by default
  - Have to allow public access on both bucket and objects
    - Add bucket policy that applies to all bucket objects
    - Bucket policy editor in the console
    - Ex.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "PublicReadGetObject",
      "Effect": "Allow",
      "Principal": "*",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": [
        "arn:aws:s3:::BUCKET_NAME/*"
      ]
    }
  ]
}
```

Versioning

- Activate “Enable bucket versioning”
- All writes, it remains even after object delete
- Use case : backup tool
- Once enabled, it cannot be disabled (only suspended)
- Integration with lifecycle rules
- Supports MFA (multi-factor auth) to delete object
- Version One ID = null, subsequent versions have version ID strings
- When deleting object type = “Delete marker”
  - All are versioned even deleted
  - In order to restore -> Delete the “Delete marker”

Storage classes

- S3 standard
  - 4 nines availability and 11 nines durability
  - Frequent access
  - Use cases : Web content, mobile, data analytics, gaming, etc.
- S3 Standard IA (infrequent access)
  - 3 nines availability and 11 nines durability
  - Rapid access
  - Pay per GB storage and per GB retrieval fee
  - Use cases : Long term storage like Backup, Disaster recovery
- S3 One Zone IA (infrequent access)
  - 99.5 availability and 11 nines durability
  - One AZ
  - cost 20% less
  - Use case : same as IA but for non critical Data
- S3 Intelligent tiering
  - 3 nines availability and 11 nines durability
  - auto move data to the most cost effective tier based on access frequency
- Glacier
  - 4 nines availability and 11 nines durability
  - Very infrequent access (archive)
  - Pay per access
  - Option 1 - Glacier Instant retrieval
  - Option 2 - Glacier Flexible retrieval (minutes or up to 12 hours retrieval time)
    - No cost retrieval
  - Option 3 - Glacial Deep Archive
    - Cheapest Storage class
    - 7-10 years retention
    - Standard retrieval time = 12 hours
    - Bulk retrieval = 48 hours

Lifecycle Management

- Automates object movement between storage classes, archive or delete
  - Keep period (after creation) before move
  - Lifecycle policies can’t work backwards. (from less freq. access storage class to more)
  - Use object filter or apply to all bucket
  - Use on current version or old versions

Locks

- S3 Object Lock
  - WORM - Write Once, Read Many
  - Prevent deletion and update for fixed amount of time
  - Retention period
    - protects an object version for a fixed amount of time

        - AWS use metadata to set timestamp attribute
        - Object is again changeable after retention period - unless legal hold is placed
    - Legal Hold
        - Placed on object to prevent update/deletion
        - Can be placed or removed by users having s3:PutObjectLegalHold permission
    - 2 Modes
        - Governance Mode
            - All users cannot delete or overwrite unless explicitly permitted
        - Compliance Mode
            - A protected object version cannot be overwritten or deleted at all (including by root)
            - This for the duration of the retention period
            - Retention mode can’t be changed
            - Retention period can’t be shortened
- Glacier Vault Lock
    - Apply WORM model to Glacier by vault lock policy
    - Once locked it can no longer be changed

Encryption

- Types
    - Encryption in Transit : SSL/TLS - HTTPS
    - Encryption at Rest (server side encryption)
        - SSE-S3 : fully managed by S3
        - SSE-KMS: Use AWS KMS to manage the keys
        - SSE-C : Customer side keys
    - Encryption at Rest (client side) - Encrypt before upload
- Can change default encryption on bucket creation
- SSE-S3 enabled by default
- Add x-amz-server-side-encryption header in PUT File request
    - x-amz-server-side-encryption: AES256 (SSE-S3) - even for recent S3 version it is active by default
    - x-amz-server-side-encryption: aws:kms (SSE-KMS)
- One can create a Bucket Policy that denies any S3 request without the x-amz-server-side-encryption header

Performances

- 3.5K request write & 5.5 K request read per second per prefix
- => The more the prefixes the better the performances
- Limitations
    - If SSE-KMS used keep in mind KMS limits
    - Upload/download count toward the KMS quota (5.5K, 10k or 30K request per second depending on region)
    - Cannot request KMS quota increase
- Multipart uploads (parallelize uploads)
    - Recommended for files over 100 MB
    - Required for over 5GB files

- Byte Range fetches (parallelize downloads)
  - If failure it’s only for specific byte range

Backup

- Replicate object from one bucket to another
- Replication requires versioning enabled
- Can apply replication to all object or to some objects using filters
- Delete markers are not replicated by default

04-ec2

EC2 - Elastic Compute Cloud

Introduction

- Compute services on EC2 :
  - Renting virtual machines (EC2)
  - Storing data on virtual drives (EBS)
  - Distributing load across machines (ELB)
  - Scaling the services using an auto-scaling group (ASG)
- Bootstrap EC2 instances using an EC2 User data script (only 1 time on first boot)
- Sizing - Instance types (ex. t2.micro)

Creating EC2 Instance

- Name and tags
- Select Base image (Amazon Machine Image) from Catalog (ex. Amazon Linux 2 AMI)
- ◦ Choose architecture
- Instance type : size and performances (t2.micro is Free tiers eligible)
  - https://aws.amazon.com/ec2/instance-types/
  - m5.2xlarge (m = class; 5 = generation; 2xlarge = size)
  - General purpose => t classes
  - Compute optimized => c classes
  - Memory Optimized (Cache stores, In mem db; databases; real time processing; etc.) => R classes, x1, etc.
  - Storage Optimized => (I, D, H classes)
- Create RSA (or other) Key pair for SSH access
- Security Group (Network rules - ex Allow HTTP and SSH traffic)
- Storage Config
  - Size
  - Delete on termination
  - Etc.
- User data : shell script launched on first machine boot
- States : Pending - Running
- Stopping / Starting (Public IP may change) / Terminating
- Note : stopped instances = no billing

Security Groups

- Kind of virtual firewalls
- Security groups only contain “Allow” rules

- Security groups rules can reference by IP or by security group
- 1 EC2 instance may have Many Security groups
- By default : All inbound is blocked and all outbound is authorized
- Regulates
  - Access to ports
  - Authorized IP Ranges
  - Inbound Network
  - Outbound Network
- SG can be attached to multiple instances
- Locked down to region/VPC combination
- Known ports :
  - 22 SSH and SFTP
  - 21 FTP
  - 3389 RDP (Remote Desktop Protocol / Windows)
- Note : Timeout errors are always due to SG
- SSH connection and PEM files
- EC2 Instance Connect (bash from Web)

Add role to instance

- Attach a Service Role to an instance
- No more need to aws configure to access services allowed by role (having the good policies)

Purshasing Options

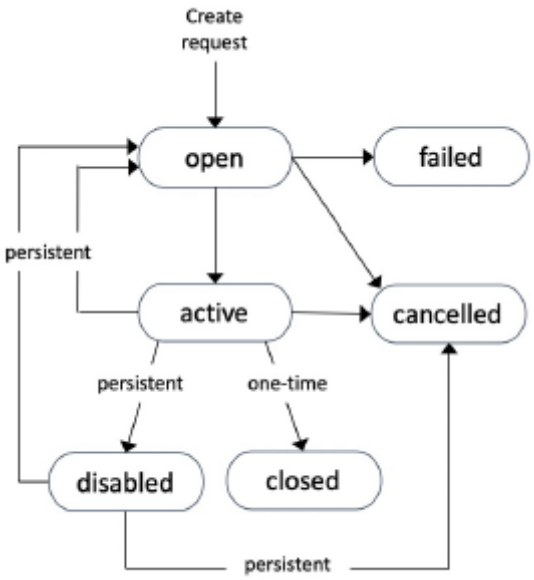
- AWS Pricing Calculator : calculator.aws
- On-Demand : short workload
  - predictable pricing (Pay by second for Linux or Windows / per hours for others)
  - Good for : flexibility, short term, testing the water
- Reserved (1 or 3 years) - up to 72% discount
  - Good for : predictable usage, specific capacity requirements,
  - Reserve a specific instance attributes (Instance Type, Region, Tenancy, OS)
  - Payment Options – No Upfront (+), Partial Upfront (++), All Upfront (+++)
  - Reserved Instance’s Scope – Regional or Zonal (reserve capacity in an AZ)
  - You can buy and sell in the Reserved Instance Marketplace
  - Convertible reserved instances - up to 54% discount
    - Can change the EC2 instance type, instance family, OS, scope and tenancy
  - Scheduled reserved instances : launch within a time window
- Savings Plans (1 or 3 years)
  - Get a discount based on long-term usage
  - Commit to a certain type of usage ($10/hour for 1 or 3 years)
  - Locked to a specific instance family & AWS region
  - Flexible across:
    - Instance Size (e.g., m5.xlarge, m5.2xlarge)
    - OS (e.g., Linux, Windows)

        - Tenancy (Host, Dedicated, Default)
- Spot Instances : short workload, cheap, less reliable (can lose instances)
    - Can get a discount of up to 90% compared to On-demand
    - Instances that you can “lose” at any point of time if your max price is less than the current spot price
    - Useful for workloads that are resilient to failure
        - Batch jobs
        - Data analysis
        - Image processing
        - Any distributed workloads
        - Workloads with a flexible start and end time
        - High performance computing
        - CI/CD and testing
        - Containerized Workloads
- Dedicated Hosts : book an entire physical server, control instance placement
    - compliance requirements and use your existing server bound software licenses
    - /! exam : Special licensing requirements = Dedicated Hosts
    - The most expensive option
    - Purchasing Options: On-demand or Reserved (1 or 3 years) with up to 70% off the on-demand price
- Dedicated Instances : no other customers will share your hardware
    - Instances run on hardware that’s dedicated to you
    - May share hardware with other instances in same account
    - No control over instance placement (can move hardware after Stop / Start)
- Capacity Reservations : reserve capacity in a specific AZ for any duration
    - No time commitment (create/cancel anytime), no billing discounts
    - Combine with Regional Reserved Instances and Savings Plans to benefit from billing discounts
    - You’re charged at On-Demand rate whether you run instances or not
    - Suitable for short-term, uninterrupted workloads that needs to be in a specific AZ

Spot instances

- Define max spot price and get the instance while current spot price < max
    - The hourly spot price varies based on offer and capacity
    - If the current spot price > your max price you can choose to stop or terminate your instance with a 2 minutes grace period.
- Spot Block :
    - “block” spot instance during a specified time frame (1 to 6 hours) without interruptions
    - In rare situations, the instance may be reclaimed
    - no longer supported since 31/12/22

- Spot Request :
  - Max price
  - Desired number of instances
  - Launch spec
  - Req Type : one-time | persistent
  - Valid from ; valid until
- Spot request states :



Spot request states

- You can only cancel Spot Instance requests that are open, active, or disabled.

- Cancelling a Spot Request does not terminate instances

- You must first cancel a Spot Request, and then terminate the associated Spot Instances

- Spot Fleets = set of Spot Instances + (optional) On-Demand Instances

  - The Spot Fleet will try to meet the target capacity with price constraints
  - Can have multiple launch pools, so that the fleet can choose
  - Define possible launch pools: instance type (m5.large), OS, Availability Zone
  - Spot Fleet stops launching instances when reaching capacity or max cost
  - Strategies to allocate Spot Instances:
    - lowestPrice (from the pool with lowest price)
    - diversified (availability) - get for different pools
    - capacityOptimized
    - priceCapacityOptimized: pools with highest capacity available, then select the pool with the lowest price
    - InstancePoolsToUseCount (in combination with lower price) : distribute across a nbr of specified pools

Public (default) IP vs Elastic IP

- All public IPs are charged by seconds
- EC2 default public IPs changes on stop/start
- Elastic IP : Fixed IP for EC2 instance
  - Up to 5 possible IPs
  - Try to avoid and use DNS

Placement Groups

- Control how EC2 instances are placed in infrastructure
- Only compute optimized, GPU, memory or storage optimized instances can be launched in a PG
- Stop existing instances to move them to a PG
- Strategies :
  - Cluster (Cluster instances into low latency group in a single AZ)
    - Same Rack (hardware) + same AZ => Great network vs failure risk
    - Use case : Big data; low latency apps
    - AWS recommends homogenous instances within cluster PG
  - Spread
    - Instances in difference hardware
    - Limit 7 instances by AZ by Placement Group
    - Scenario question: Individual instance (primary / secondary db for example)
  - Partition
    - Each partition = rack in AWS
    - Up to 7 partitions per AZ
    - Up to 100s of EC2 instances
    - Use case : Bigdata that are partition aware (like HDFS)
    - Scenario questions : Multiple instances

Elastic Network Interface (ENI)

- ENI = Virtual Network Card
  - Primary IPv4 and one or more secondary ones
  - One Elastic IPv4 per private IPv4
  - One Public IPv4
  - One or more SG
  - A MAC Address
  - …
- ENI are bound to AZ
- ENI can be created independently
- Can attach an ENI to an EC2 instance; detach it and move it to another instance
- Helpful for failover

Enhanced Networking (EN)

- Single root I/O Virtualization to provide high performance

- 10 Gbps - 100 Gbps
- Depending on instance type
  - Elastic Network Adapter (ENA) up to 100 Gbps
  - Intel 82599 Virtual Function (VF) Interface (10 Gbps used on older instances)
  - /! In the exam : Always choose ENA over VF

Elastic Fabric Adapter (EFA)

- Accelerates HPC (high performing computing) and machine learning apps
- Lower and more consistent latency
- Can use OS-BYPASS - lot faster with much lower latency
  - For HPC and machine learning apps
  - By passing OS and directly communicate with EFA Device
  - Only linux support

EC2 Hibernate

- Stop instance -> data on disk (EBS) kept
- Terminate instance -> data on disk (EBS volumes) are destroyed unless instructed not to
- Hibernate instance -> RAM is preserved
  - Faster boot time
  - RAM must be less than 150GB
  - Under the hood : use EBS Root volume (must be encrypted) to dump RAM
  - Use cases : Long running tasks ; saving ram sate; start time
  - Works on On demand, spot and reserved instances

EC2 Metadata

- curl to specific URL to get metadata from instance
  - TOKEN = curl
  - curl … $TOKEN …/latest/meta-data/public-ipv4 for example

vCenter in AWS

- use cases
  - Hybrid Cloud
  - Cloud migration
  - Disaster Recovery
  - Leverage AWS
- Runs on dedicated hardware
- Each host has 2 sockets with 18 cores per socket, 512 GiB RAM & 15.2 TB Raw SSD Storage
- Multiple VMware instances by host up to 100s
- Cluster can start with 2 hosts up to 16 hosts

AWS Outposts

- AWS Data center on-premises
- Large variety of AWS Service in private DC
- Sizes : 1U to 42U racks and multi rack deployments
- Use cases
  - Hybrid Cloud
  - Full managed (AWS can managed the infra for you)
  - Consistency with public AWS Services
- Outposts Rack
  - starting with single 42U rack and scale up to 96 racks
  - AWS compute, storage, database and other services
- Outposts Servers
  - Individual servers in 1U or 2U form factor
  - Use cases :
    - Small space requirements (retail stores, branch offices, healthcare provider locations, etc.)
    - Provide local compute and networking services
- Process
  - Order
  - Install (including, hardware, power, networking and connectivity)
  - Launch
  - Build

05-ebs-efs

Instance Storage (EBS, EFS)

EBS : Elastic Block Store

- Network drive volume
- Attached to one instance at a time (except multi-attach for some EBS)
- Allow data persistence after instance termination
- Highly available across AZ
- Locked to AZ / snapshot to move across AZ
  - Cannot attach to instances from another AZ
- Provisioning : Size and IOPS (IO per second)
- Billed for capacity
- Scalable
  - Capacity can be increased over time without downtime
  - Volume type can be changed on the fly
- Delete on Termination attribute = True / false
  - Default to true to root EBS volume
- Always in the same AZ of the instance using them
-

EBS Snapshot

- Can be copied across AZ/Regions and are shared only in their region
- Can create volume from snapshot
- EBS Snapshot archive
  - 75% cheaper
  - Take within 24 to 72 hours to restoring the archive
- Recycle Bin for EBS Snapshots
  - Rule for snapshot retention after deletion
  - Retention period : from 1 day to 1 year
- Fast Snapshot Restore (FSR)
  - Full initialization of snapshot with no latency on first use
  - Costly
- Snapshots are incremental
- Recommendation : stop instance before taking snap

EBS Volume Types

- 6 Types (only the first 4 can be used as boot)
  - gp2 / gp3 SSD - General purpose
    - 1GiB to 16TiB
    - Cost effective / Low latency
    - use case high performance / lower cost
  - io1 / io2 Block Express SSD (low latency / high throughput workload)
    - 4GiB to 16TiB
    - Sustained IOPS perfs
    - Support multi-attach
    - Use cases ! Big data, data warehouses, ETL, log processing => throughput (ex. 500MB/s per TB)
  - st1 HDD : Low cost HDD for frequently accessed, throughput intensive workloads
    - 125 GiB to 16TiB
    - Big Data, Log processing, Data warehouses
    - Max throughput 500 MiB/s / max IPOS 500
  - sc1 HDD : Lowest cost HDD for less frequently accessed workload
    - Cold HDD ##### IOPS vs Throughput

IOPS	Throughput
Nb of R/W per second	nb of bits R/W per s
Quick transactions, low latency, transactional	important for larger datasets, complex queries
Choose IO types	Choose Throughput optimized HDD

EBS multi-attach

- Only io1/io2 family
- Same EBS to multiple EC2 instances in the same AZ (up to 16 EC2)
- Each instance has full r/w permissions
- Must use cluster aware file system
- Use case : higher application availabilty (ex: Teradata); concurrent write operations

EBS Encryption

- Encrypted EBS Volume
  - At rest
  - in flight
  - snapshot
  - volumes created from snapshots
- Minimal impact on latency
- Keys from KMS (AES-256)

- To encrypt an existing one : snapshot then re-attach (launch instance from AMI if root volume)

EC2 Instance Store

- High performance hardware disk
- Better I/O
- Ephemeral (lost on stop)
- Use case : buffer, cache, temp, etc.
- Risk of data loss if hardware fail

EFS - Elastic File System

- Managed NFS (only for linux instances uses NFSv4 protocol)
- Support 1000s of concurrent connections and scale to petabytes
- Works with EC2 instances in multi AZ (basically it s a shared storage)
- No capacity plan (FS Scale) Pay per use (expensive = 3 x gp2)
- Highly available and scalable
- Use security groups to access
- Performance Mode
  - General Purpose (default) - latency sensitive use cases
  - Max IO - higher latency , throughput (big data, media processing)
- Throughput mode
  - Bursting : 1 TB = 50MiB/s + burst of up to 100MiB/s
  - Provisioned : Throughput set regardless of storage size
  - Elastic : Auto scales based on workloads
- Storage classes
  - Storage Tiers : Standard or Infrequent access (EFS-IA)
    - Must use lifecycle policy to enable EFS-IA
  - Availability and durability
    - Standard / Multi AZ
    - One Zone (good for Dev / cost saving)
- Use case : content management, web serving, data sharing, etc.

FSx

- FSx for Windows
  - Managed native Windows file system (built on windows server)
  - Supports AD users, ACLs, etc.
  - Support SMB protocaol
  - Encrypts keys via KMS
- FSx for Lustre
  - Lustre file system to process massive datasets
  - HPC: Millions of IOPS, GBs of throughput’s, sub milliseconds latencies
  - Can store data directly on S3

AMI - Amazon Machine Image

- Customization of EC2 instance (faster boot / prepackage software)
- Region specific : Built for specific region and can be copied across regions
  - So shared across same region AZ
- Types
  - AWS Provided Public AMI
  - Custom API
  - AWS Marketplace AMI
- Can build AMI from EC2 instance (start -> customize -> stop -> Build AMI)
  - This will create EBS snapshots
- AMIs are either backed by
  - EBS : the root device is an EBS volume created from an EBS snapshot
  - Instance Store : the root device is a EC2 Instance Store volume created from a template stored in S3

AWS Backup

- Consolidate backups across multiple AWS Services
  - EC2, EBS, EFS, FSx, Storage Gateway, DRS, DynamoDB
  - Can be used with Organizations
  - Benefits : central management, automation, improved compliance

06-databases

Databases

RDS - Relational DB Services

- Types : Postgres, MySQL, MariaDB, Oracle, SQL Server, IBM DB2 and Aurora (AWS Proprietary DB)
- RDS is Managed :
  - Auto prov. os patching
  - Continuous Backup and restore
  - Monitoring
  - Read replicas for perfs
  - Multi AZ setup for DR (Disaster Recovery)
  - Maintenance windows for upgrades
  - Scaling capabilities both vert and horiz
  - EBS Backed Storage (gp2 or io1)
  - Failover support
- Usage : OLTP - Online transaction processing
- Not suitable for OLAP : use redshift instead
- Cannot do SSH
- Linked to SG

RDS - Storage Auto scaling

- All DB engines
- Max storage Threshold to be set
- Auto scale up or down storage if
  - Free space is less than 10% of allocated
  - And Low storage lasts at least 5 minutes
  - And 6 hours have passed since last modification

Read replicas

- Only Select (for performance not DR)
- Up to 15 read replicas (15 aurora, 5 mysql)
  - Async replication between main RDS instance and read replicas
  - Reads are Eventually consistent (because replication is async)
  - Can be within AZ, cross AZ or cross Region
- Replicas can be promoted to their own DB (breaks replication)
- Sample Use case : reporting and analytics
- Network cost :
  - free (no charges) if Read Replicas in same Region
  - charged for cross region

RDS Multi AZ (disaster recovery)

- master (or primary) and standby instances (different AZ)
- Synchronous replication
- One DNS name - automatic app failover to standby promoted to master db
- Not used for scaling or performance - cannot use standby db when master is active
- /! Note : a Read Replica can be set as Multi AZ
- Zero downtime to go from single to multi AZ

RDS Custom

- Managed Oracle and SQL Server Database with OS and db customization
- Access to underlying database and OS (EC2 instances) using SSH or SSM Session Manager
- Need to De-activate automation mode

Amazon Aurora

- Cloud optimized -> 5x perf. improvements overs MySQL and 3x over Postgres on RDS
- Uses shared storage volume
- Note : Aurora snapshots can be shared with other AWS Accounts
- Posgtres and MySQL format are support (for drivers use)
- Auto grow in increments of 10GB up to 128 TB
- Master + up to 15 replicas (faster replication + possible autoscaling)
    - Support cross region replication
- Instantaneous failover - HA native
- Cost 20% more than RDS
- HA and read scaling
    - Min 2 copies across 3 AZ = 6 copies
    - 4 needed for writes , 3 for reads
    - Self healing with p2p replication
    - storage striped across 100s of volumes
    - Auto failover for master in less than 30 seconds (a reader become master)
- Writer endpoint (master) and reader endpoint (Connection Load Balancing over Readers)
- Backtrack turn on by default - restore data at any point of time
- Advanced
    - Replicas Auto scaling (ex. CPU usage)
    - 3 types of Replicas : Aurora (15), MySQL (5), Postgres (5)
        - Auto failover if only available with Aurora replicas
    - Custom Endpoints (ex. for Analytics with 2xlarge read replicas)
    - Aurora Serverless
        - Automated database Client instantiation and autoscaling based on actual usage

        - Good for infrequent, intermittent or unpredictable workloads
        - No capacity planning needed
        - Pay per second, can be more cost-effective
    - Global Aurora (different from Aurora cross region)
        - Decrease latency all over the world
        - 1 primary regison (r/w)
        - up to 5 secondary region (read only) and 16 replicas per secondary
        - Promoting a secondary region have RTO < 1 minute
        - Aurora Machine Learning #### RDS & Aurora Backups
- Daily full backup
- Transaction logs backed up every 5 minutes
- point in time recovery
- 1 to 35 days retention (set to 0 to disable - onyl RDS; not in Aurora)
- Manual DB Snapshot
- Trick : to save costs snapshot and terminate DB
- Can restore into new DB
- Backup and restore from S3
- Possible de backup and restore from on premise MySQL to Aurora using Percona XtraBackup
- Aurora Database Cloning (create cluster from existing one)
    - Copy on write protocol
    - Good for staging DB

RDS & Aurora Security

- At rest encryption (launch time)
- Read replicas cannot be encrypted if main instance is not
- For existing DB (snapshot and restore)
- In flight encryption
- Can use IAM Roles to connect to DB (except for Oracle)
- Security Group
- Audit logs can be enabled and sent to CloudWatch for longer retention

RDS Proxy

- Allow pool and sharing DB connection
    - Minimize connection opening
- Serverless / autoscaling and multi-AZ (highly available)
- Reduce RDS and Aurora failure time by up to 66%
- Enforce IAM auth
- RDS Proxy are never publicly accessible - must be accessed from VPC
- Util for Lambda functions (since they multiply quickly)

ElastiCache

- Redis or Memcached

- Support IAM Authentication for Redis only
- Patterns
  - Lazy loading (data can become stale)
  - Write through (update cache on DB update - no stale state)
  - Session store
- Redis
  - Multi AZ with auto failover
  - Read replicas
  - Data durability using AOF persistence
  - Backup/restore
  - Supports Sets and Sorted Sets (guarantee both uniqueness and element ordering)
  - Supports SSL in flight encryption
- Memcached
  - Multi-node (data partition = sharding)
  - No replication (no high availability)
  - Non persistent
  - No backup and restore
  - Multi-threaded
  - Supports SASL-based authentication

DynamoDB

- NoSQL Database (Eventually consistent)
- Pay per request pricing / balances cost an perf / no minimum capacity
- Create by creating Table
- Support both Document and Key/Value data model
- Stored on SSD
- Spread across 3 geographically distinct DC
- Strongly consistent reads : returns a result reflecting all successful writes
- DAX - DynamoDB Accelerator
  - In memory cache
  - 10x faster : from milli to micro seconds
  - Compatible with DynamoDB API (transparent)
- Security
  - KMS
  - Site to site VPN
  - Direct Connect (DX)
  - IAM Policies and roles
  - Fine grained access
  - CloudWatch and CloudTrail
  - VPC endpoints
- DynamoDB Transactions
  - ACID across 1 or more tables
  - Allowed up to 25 items or 4MB of data per transaction
  - 3 read options : eventual consistency, strong consistency and transactional
  - 2 options for writes : standard and transactional

- DynamoDB Backups
  - On demand
    - Zero impact
    - Consistent within second
    - Retained until deleted
    - Same region as source table
  - Point in TIme recovery (PITR)
    - Not enabled by default
    - Restore to any point in the last 35 days
    - Protects against accidental writes or deletes
    - Incremental backups
    - Latest restorable : 5 minutes in the past

DynamoDB Streams

- Time ordered sequence of item level changes in a table (steam records)
- Stored for 24 hours
- Stream records grouped by shards

DynamoDB Global tables

- Managed multi master, multi region replication
- Transparent and Native to DynamoDB
- Based on DynamoDB streams
- Replication latency = 1 second

Amazon DocumentDB (MongoDB compatible)

- Allows to Run mongodb on AWS as managed service

Amazon Keyspaces

- Allows to Run Casandra Workloads on AWS as managed service

Amazon Neptune

- Graph Database
- Use cases : Connections, Graph applications, detect fraud patterns, Security, etc.

Amazon Quantum Ledger Database

- NoSQL immutable, transparent having cryptographically verifiable tx logs

Amazon Timestream

- Time series Database

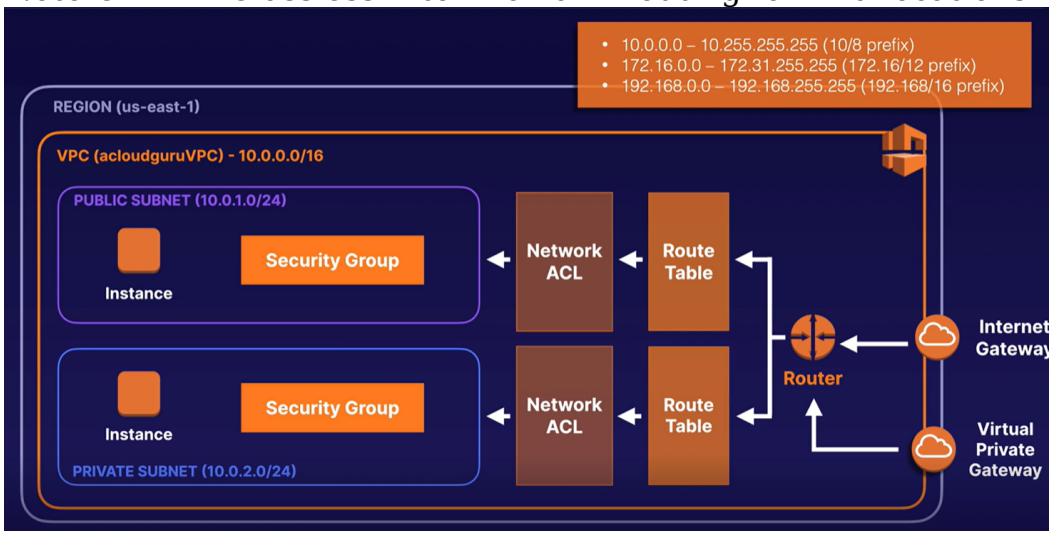
- Trillions of events by day
- Usecases : IoT, agriculture, analytics, DevOps apps

07-vpc

VPC

- Virtual Data Center in the Cloud / Fully customizable Network
- Logically isolated part of AWS Cloud where you can define your own network
- Complete control of virtual network including :
  - Your own IP address range
  - Subnets
  - Route tables
  - Network ACLs (NACL)
  - Network gateways
  - Better instances and network security
- Every region in every AWS account has default VPC
- Can create a Hardware VPN connection between on prem and VPC
  - AWS Cloud as extension to corporate DC
- When create a VPC (ex. 10.0.0.0/16 => wide range)
  - Internet Gateway
  - Router
  - Route table
  - Network ACL
  - Subnets (private / public) - (ex. 10.0.1.0/24, 10.0.2.0/24, etc.)
    - Route table and Network ACL configuration
    - SG, instances, etc.
  - /! by default it creates a Route table, a Network ACL and a Security Group
- When creating a subnet in a VPC
  - IP CIDR Block (ex. 10.0.1.0/24)
  - AZ
  - Auto assign public IP addresses is by default “NO”
  - Note: the first 4 and the last address are AWS reserved (5 in all)
    - ex. 10.0.1.0 => Network address
    - ex. 10.0.1.1 => AWS VPC Router
    - ex. 10.0.1.3 => AWS DNS Server
    - ex. 10.0.1.4 => Future use
    - ex. 10.0.1.255 => Network broadcast
- /! Every subnet is in 1 AZ and cannot spread across AZs

- /! IP addresses CIDR Block must be between /16 and /28 network mask
  - Note CIDR = Classless Inter-Domain Routing for IP allocations



- Default VPC vs Custom one
  - Default is user friendly / Custom takes time to configure
  - In default, all subnets have route out to internet
  - In default, all instances have private and public IPs
- VPC Tenancy = default or dedicated (dedicated infrastructure)

Internet Gateway

- Create a new Internet Gateway
  - Attach it to a VPC (only custom VPC)
  - /! Cannot attach an Internet Gateway to the Default VPC
- Create new Route table to associated with Internet Gateway
  - 0.0.0.0/0 destination
  - Attach to Internet Gateway as target
  - Associate explicitly the public subnet to the route table
    - This removes it from implicit association in the default Route table
  - Note: default Route has local target

NAT Gateway

- Issue : Getting out internet access from private subnet instances
- NAT => enables instances in private subnet to connect to internet (or other AWS services) while preventing IN internet access
- NAT Gateway if provisioned in the public subnet to gain private ones access to internet
- Caracteristics
  - Redundant inside an AZ
  - 5 Gbps and scales to 45 Gbps
  - No need to patch
  - Not associated with SG
  - Auto assigned public IP
- Need to allocate an Elastic IP

- In the private route table add a route to NAT
  - 0.0.0.0/0 destination (internet)
  - target = NAT Gateway created in the public subnet
  - => now private instances have access to internet

Network ACLs

- /! Note on SG : response of allowed inbound rules are allow to flow out regardless of outbound rules (This is called stateful behaviour of SG in contrast with NACLs which are Stateless)
- NACL = Optional layer of security controlling traffic in/out a subnet
- Rules can be similar to SG
- Default NACL for VPC allows are in/out traffic
- Custom NACL = by default deny all in/out traffic
- Each subnet must be associated with NACL / either custom or default one
  - A Subnet can be associated with only one NACL (but same NACL can serve multiple Subnets)
- /! NACLs can block IP addresses but SG cannot
- NACLs contain a numbered list of rules evaluated in order (starting from lowest 100 then 200 then 300 …)
  - Inbound or Outbound rule
  - Each rule can either allow or deny
  - They are stateless
  - No rules by default
- /! open ephemeral port 1024-65535 as outbound rule since web servers replies in custom ports
- Deny specific IP address
  - 23.24.25.26/32 as source
  - port 80
  - deny
  - rule number = 50 (< 100) so it s evaluated first

VPC Endpoints

- Connect VPC to AWS services without Internet Gateway or NAT
- Uses AWS Backbone network
- No public IP required
- Horizontally scaled, redundant and highly available
- No availability risk or network bandwidth constraints
- 2 types
  - Interface Endpoint
    - Elastic network interface with private IP
    - Large number of AWS Services
  - Gateway Endpoint
    - Similar to NAT Gateway
    - Virtual device
    - Supports S3 and DynamoDB

VPC peering

- Multi VPC
- Direct networking with private IP
- Instances behave as if it was same private network
- Can peer VPC with other AWS Accounts
- Peering are in star configuration
- No transitive peering (A -> B and B -> C does not give A -> C - must be explicit)
- Can peer between regions
- VPCs must have different CIDR address ranges
- Opens all services between VPC

PrivateLink

- More restrictive than VPC
- need Network load balancer on service VPC and ENI - Elastic Network Interface on consumer VPC
- Scales well (10s, 100s or 1000s of customers)

VPN CloudHub

- Connect multiple sites each with their VPN connection
- Hub and spoke model
- Low cost
- Operates over public internet with encrypted communication

Direct connect

- Connect On Premise DC with VPC
- Reduce Network cost, increase bandwidth throughput and give more consistent network than with internet connection
- 2 Types
  - Dedicated connection : Physical ethernet connection
  - Hosted connected : Physical ethernet connection that AWS Direct CConnect partner provision on behalf of client
- Uses DX Location + AWS Cage with routers configuration in each
- Fast, secure, reliable and take massive throughput compared to VPNs
- Can run VPN over a DX (Direct connection)

Transit Gateway

- Simplifies the network topology
- Allow to have transitive peering between 1000s of VPC and DX
- Hub and Spoke model
- Regional but can be multi-region
- Can be used across multiple AWS Accounts use RAM = Resource Access Manager
- Support IP Multicast (not supported by other AWS Services)

5G networks and AWS Wavelength

- AWS Wavelength Embeds AWS Compute and store services in 5G Networks
  - Mobile Edge computing infrastructure

08-route53

Route 53

- Route 53 = Managed and Authoritative DNS Services
- Authoritative = customer can update DNS records
- Route record
    - SOA : Start of Authority record
    - NS Records : Name server records : used by top level domain servers
    - Domain / Sub domain names
    - Record type : A, AAAA, etc.
        - CNAME hostname -> hostname (cannot be created for Top zone - like example.com)
            - Only non root domains (example.com est un root domain)
        - NS Name server for the Hosted Zone
        - Alias hostname -> IP or AWS Resource
            - A hostname -> IPv4 (or aws resource behind it)
            - AAAA hostname -> IPv6 (or aws resource behind it)
            - Works even for root domains (naked domains - in contrast with cnames)
            - Free of charge
            - Native health check
    - Value : IP
    - Routing Policy
    - TTL (mandatory for each record exception for Alias ones)
- Private vs Public Hosted Zone (billed monthly)
- Private Hosted Zone = route traffic within one of more VPCs
    - Internal names -> VPC resources
- Registering domain then create records
- Alias record targets :
    - ELB
    - CloudFront Distributions
    - API Gateway
    - Elastic Beanstalk env
    - S3 Websites (not buckets)
    - VPC Interface Points

    - Global Accelerator accelerator
    - Route 53 record in the same hosted zone

- /! Cannot set Alias to an EC2 DNS name

- /! record can be multi-valued (value chosen randomly from client)

- Domain Registrar = domain seller != DNS Service

- AWS is also a domain registrar #### Routing Policy

- Simple record -> single or multi IP values

- Weighted (same name and type but different record ID)

    - weight = number between 0 and 255
    - record id
    - can be associated with health check

- Latency based (same name and type but different record ID)

    - On ID by region (latency type)
    - Traffic between users and AWS Regions

- Failover (same name and type but different record ID)

    - If Health check unhealthy -> route to secondary

- Geolocation

    - Based on user location
    - Should have default

- Geoproximity (Using Traffic Flow only)

    - Traffic Flow -> AWS service to build complex routing flow (Traffic Policy)
    - Based on user / resources proximity
    - Geo region size = bias attribute (value from 1 to 99 et from -1 to -99)
    - AWS Resources will use Region
    - Non AWS Resources => have to specify Latitude and Longitude
    - Must activate Route 53 Traffic flow

- IP Based

    - Based on user IP (CIDRs)

- Multi Value (or Multi answer)

    - Routing to multiple resources (client side load balancing)
    - Can be associated with health checks
    - Up to 8 healthy records are returned

Health checks

- Only for public resources
- Types
  - Endpoint
    - 2xx or 3xx responses
    - can look to 5120 bytes of response
  - Calculated - combines Health checks
    - AND, OR, NOT
    - Monitor up to 256 child HC
    - Can specify how many HC need to pass to consider parent HC OK
    - Usage : maintenance without failing all Health checks
  - CloudWatch Alarms
- Health checks are integrated with CW
- About 15 global checkers
  - Threshold : 3 by default
  - if > 18% of checkers are OK then healthy
  - 30 sec (can be set to 10 sec at higher cost)

09-elb

ELB - Elastic Load Balancer

- Managed Load balancer
- Health checks : port and route
- 4 Kinds
  - CLB - Classic Load Balancer (deprecated - not in exam)
  - ALB - Application Load Balancer
    - HTTP(S), Websockets
  - NLP - Network Load Balancer (high performances)
    - TCP, TLS, UDP (Layer 4)
  - GLP - Gateway Load Balancer (no in exam according to Cloud guru)
    - IP Protocol (Layer 3)
- Some can be setup as Internal / External (internet-facing = requests from internet directly to targets)
- Load balancer SG (ex. 80 & 443) and App SG (ex. 80)
  - In App SG can add “allow traffic only from Load Balancer” rule by adding its SG rule
- Can route across AZ

ALB

- Operates on Layer 7
- HTTP(S), Websockets
  - on HTTPS at least one SSL/TLS server certificate is required
- ALB has listener (to ports) and each listener has rules
- Routing tables (rules) to Target Groups (or redirect URL or fixed response)
  - Based on URL Path
  - Based on Hostname
  - Based on Query/Headers
- LB Rules can have priorities (from 1 highest to 50000)
- ALB can route to multiple TG (target groups)
- Target groups
  - EC2 instances - can be managed by Auto Scaling Froup - HTTP
  - ECS tasks - managed by ECS - HTTP

    - Lambda functions (HTTP request translated to JSON Event)
    - IP Addresses (private IPs)

- Health checks are at the target group level

- Notes :

    - Fixed hostname for ALB
    - X-Forwared-For assigned
    - Error 504 Gateway timeout -> LB OK but service is down (servers, db, etc.)

- Note : In AWS, an Application Load Balancer (ALB) typically uses dynamic IP addressing. If you need your ALB to have a static IP address, you can achieve this by using a Network Load Balancer (NLB). You can associate the NLB with an Elastic IP address, which is a static, public IPv4 address, and then register your ALB as a target of the NLB. #### NLB

- Operates on Layer 4

- TCP, TLS, UDP (Layer 4) - Any port

- Handle millions of request / second

- Less latency

- NLB has one static IP per AZ and supports assigning Elastic IP

- Target Groups : EC2 Instances, IP Addresses (private), ALB

- Health check : TCP, HTTP and HTTPS

- Can decrypt traffic but have to install the certificate in the NLB

GLB

- Usage : Firewalls, Intrusion detection, Deep Packet Inspection, payload manipulation, etc.
- Another example : GENEVE Protocol on port 6081
- Traffic is routed to 3d party Target Group before accessing Application
- Combine 2 functions
    - Transparent Network Gateway
    - Load balancing

Sticky session - Session Affinity

- Available for CLB, ALB and NLB
- Uses Cookie
    - Custom Application based Cookies
        - Generated by the target
        - Can include any custom attributes required by the app

        - Cookie name must be specified individually for each target group
        - Reserved names : AWSALB, AWSALBAPP, AWSALBTG
    - Application cookie
        - AWSALBAPP named Cookie Generated by the LB
    - Duration based Cookies
        - Generated by the LB and expires based on LB specification
        - Named AWSALB
- Caution to imbalance

Cross zone balancing

- Distribute traffic across AZ based on instance number into each
- ALB - enabled by default (no extra charges)
    - Can be disabled at target group level
- NLB - disabled by default (extra charges if enabled)

SSL Certificates

- One can manage certificates using ACM - AWS Certifite Manager
- Or create and upload your own certificates
- HTTPS (LB) Listener requires :
    - Default certificate
    - Optional list of certs (to support multi domain)
    - Client can use SNI - Server Name Indication
    - Can support Legacy SSL
- SNI - newer protocol that requires the client to indicate the hostname in initial SSL handshake
    - This solves the problem of multiple SSL certs in one server
    - Only works for ALB and NLB, CloudFront

Connection Draining (Deregistration)

- Named Deregistration Delay for ALB and NLB
- Time to complete in flight requests while the instance is de-registering or unhealthy
- Stop sending new requests to EC2 instance
- Between 1 to 3600 seconds (default to 300) - can be set to 0

10-cloudwatch

Cloudwatch

- Managed Monitoring service in AWS
- System metrics (got out of the box - the more services are used, the more metrics there are)
- Application metrics (by installing CloudWatch agent - informations from inside EC2 instances)
- Alarms
- Default metrics (no need to install agent)
  - CPU Utilization
  - Network throughput
- Custom metrics (need agent)
  - EC2 Memory utilization
  - EBS Storage capacity
- Basic monitoring = 5 minutes (no extra fee)
- Detailed monitoring = 1 minutes (extra fee)

Create Alarm

- Metric and conditions
  - Select metric category (grouped by namespace)
  - Select instance/metric
  - Select statistics
  - Selector operator (Greater, Lower, etc.)
  - Datapoints to alarm (number of Err of OK 1/1, 5/10, etc.)
  - Missing data treatment (nothing, good, bad, etc.)
- Configure action :
  - Notification
  - Autoscaling action
  - EC2 action (stop, terminate, reboot, recover)
  - System manager action : create incident
- Etc.

CloudWatch Logs

- Managed tool for logs files from different sources to
  - Monitor
  - Store
  - Access
- CloudWatch Agent for custom logs (on AWS instances or on-premises servers)
- Terminology
  - Log Event : timestamp + data

    - Log stream : log events from same source
    - Log Group : Collection of log streams (ex. httpd logs across hosts)
- Features
    - Filter patterns : Look for specific terms in logs (ex. 400 errors)
        - Can send it to lambda or automation routine
    - Logs Insights : query using SQL-like interactive solutions
    - Alarms
- /! if logs processing is not needed, just send logs to S3

Amazon Managed Grafana

- Query, correlate, visualize operational metrics, logs and traces
- Workspaces : logical Grafana server (separation of data visualizations and querying)
- AWS Managed and Secure
- Pricing by active user in a Workspace
- Data sources :
    - CloudWatch
    - Prometheus
    - Amazon OpenSearch
    - Amazon Timestream
    - Etc.
- Use cases
    - Container Metrics (Prometheus, EKS, ECS, ON prem kubernetes cluster, etc.)
    - IoT and Edge device data (vast data plugins)
    - Troubleshooting (centralized dashboards)

Amazon Managed Prometheus

- AWS Managed, Serverless, Prometheus-compatible service
- Securely monitoring container metrics at scale
- Auto scale based on ingestion, storage and querying of metrics
- High Availability : replicates across 3 AZ in same Region
- Works with Kubernetes cluster running on Amazon EKS or on premises Kubernetes
- Uses PromQL : query language
- Data is stored in workspaces for 150 days then deleted

11-scaling

Scaling

- The 3 W’s of scaling
  - What ? EC2, DB, etc ?
  - Where ? how many AZ, what Load Balancers, scale DB or Web ?
  - When ? Most of the time Cloudwatch alarms

Launch template

- Launch template
  - Settings needed to configure EC2
  - Capable or leveraging all EC2 auto scaling features
  - Support versioning
  - More granularity
- Launch configuration (predecessor of Launch template)
  - Only for certain EC2 auto scaling features
  - Immutable
  - Limited configuration options
  - Dont include Networking information

Vertical scaling

- Turn off and resize instances (t2.micro -> other class)

ASG - Auto scaling group (Horizontal scaling)

- Scale out (add EC2) on increased load
- Scale in (remove EC2) on decreased load
- Only for EC2
- Ensure min / max instances
  - Min, max and desired capacities (desired = only initial)
- Note : Bake AMIs to reduce build times
- Auto register new instances to load balancer
- re-create EC2 instance in case previous is terminated (ex. if unhealthy)
- ASG are free
- EC2 instances can be registered behind a Load Balancer
  - Auto scaling group can be set to respect the LB health checks
- To create ASG
  - Networking
    - Select subnets in multi AZ for High availability
    - /! do not change the default network interface

    - ASG Launch template
        - AMI + Instance type
        - User Data
        - EBS Volumes
        - SG
        - IAM Roles
        - Network + subnet inf
        - LB informations
        - etc.
    - Can be attached to existing or new LB (with target group)
    - Scaling Policies
        - CloudWatch alarms (avg cpu, custom metric, etc.)
        - Dynamic scaling
            - Target Tracking (based on target state - ex. avg cpu = 50%)
            - Step Scaling (based on alarms)
                - Ex. Scale out add 5 instances when memory is between 60-80 % and 3 instances when between 80%-100%
                - Ex. scale in Terminate 5 instances when memory between 40%-10%
        - Scheduled scaling (ex. on Fridays increase to 10)
        - Predictive Scaling (forecasts)
    - Notification : when scaling activity happens
        - Using Amazon SNS Notifications
    - Purchase Option : On-Demand vs Spot %
- Example of scaling metrics :
    - AVG CPU
    - Request Count per target
    - Avg Network I/O
    - Any custom metric pushed to CloudWatch
- Warm up period - time before placing behind LB and do health check
- Cool down period (default to 300 seconds) after scaling activity
- Lifecycle Hooks
    - Perform custom actions on instances when lifecycle events occur

Scaling Relational DB

- Types of scaling
    - Vertical scaling
    - Scaling storage : only scale up, not down
    - Read replicas
    - Aurora serverless (unpredictable workloads)

Scaling non Relational DB

- Dynamo DB
    - Capacity model
        - Provisioned capacity (predictable workload)

        - On-Demand Capacity mode (sporadic workload)
        - Can switch between models (2 times within 24 hours period)
    - Read Capacity Unit RCU = Reads per second for in item up to 4KB in size
        - 1 RCU strongly consistent
        - 2 RCU eventually consistent
    - Write Capacity Unit WCU = Writes per second for an item up to 1KB in size

Disaster Recovery Strategies

- RPO : Recovery Point Objective
    - At what point in time
- RTO : Recovery Time Objective
    - How quickly
- Strategies
    - Backup and restore
    - Pilot Light : replicate DB (live replication) and deploy services (Web and others)
    - Warm Standby : replicate DB (live replication) and scale-up services (Web and others)
    - Active/Active : most expansive

Exam tips

/! Consider Switching DB as an answer in exam /! Spread out - select multiple AZ

12-decoupling-workflow

Decoupling Workflow

- Services
  - SQS - Simple Queue Service : Managed Message queue service
  - SNS - Simple Notification Service
    - A2A (Application 2 Application)
    - A2P (Application 2 Person)
  - API Gateway
    - Create, publish, maintain, monitor and secure APIs

SQS

- SQS - Simple Queue Service : Managed Message queue service
- Asynchronous
- SQS Settings
  - Delivery Delay : default to 0, can be set up to 15 minutes
  - Message size : can be up to 256KB of text in any format
  - Encryption :
    - in transit : by default yes
    - at rest : by default no
  - Retention :
    - Default is 4 days, can be between 1 min and 14 days
  - Long (connect and wait - more cpu time) vs Short polling (frequently)
    - Short polling is default
  - Queue Depth
    - Can be a trigger for autoscaling
  - Visibility timeout
    - Polled message is locked (invisible) for 30 seconds
    - Instance having polled the message send ACK to the queue to delete the message
    - If no ACK sent, the message is again visible

SQS FIFO

- Standard Queue
  - Best effort ordering
  - Can order on application level
  - Duplicate messages
  - unlimited tx/sec
- FIFO Queue
  - Guaranteed ordering
  - Must append “.fifo” to queue name

    - No message duplications (explicit message deduplication ID)
        - Can be based on message content
        - Message deduplication ID
    - Message Group ID
    - 300 tx/sec
    - Batching can achieve up to 3K tx/sec
    - if FIFO High throughput activated
        - 9K and 90K (with batching) tx/sec

Dead Letter queues - DLQ

- targets for messages that cannot be processed successfully
    - Technically DLQ are just other SQS queues
    - Same retention period as source
    - For FIFO SQS Queues, corresponding DLQs must be of type FIFO
- Work with both SQS and SNS
- Usefully for debugging and isolating unconsumed messages to troubleshoot
- Redrive capability : Move message back into the source queue
- Can configure alarms based on availability counts
- Troubleshoot consumer permissions

SNS

- Push based messaging service
- Proactively delivers messages to endpoints that are subscribed
    - Subscriber types : SQS, Lambda, email, HTTP(S), SMS, platform application endpoint, Kinesis Data Firehose
- On message can be sent to many receiver
- Message size : can be up to 256KB of text in any format
    - If SNS Extended Library installed -> Message can go to 2 GB in size (using S3 intermediate storage)
- Supports DLQs
- FIFO SNS only support FIFO SQS as a subscriber
- Encryption :
    - in transit : by default yes
    - at rest : by default no
- Access Policies
- SNS Fanout
    - Messages are replicated to multiple endpoint subscriptions
    - Fully decoupled parallel asynchronous processing
- SNS Architectures
    - SNS -> Multi SQS -> Processing
    - SNS -> Kinesis Data Firehose -> Massive amount of data to be ingested / transformed -> Services
- Message filtering
    - Filter subscriber based on content
- /! Custom retry policies for HTTP(s) endpoints only
- Use cases : real time alerts or push based messaging

API Gateway

- Managed AWS Service to Create, publish, maintain, monitor and secure APIs
- Integrates with Lambda, HTTP Endpoints, etc.
- Auto attach WAF
- Stop Abuse : DDoS protection
- API Options
  - REST API
  - HTTP API
  - Websocket API
- Endpoint types
  - Edge-Optimized : Default option
  - Regional : for APIs only reserved to particular region
  - Private : Only accessible via VPCs using VPC Endpoints
- Security
  - User authentication : IAM, AWS Cognito, custom authorizer (Lambda function)
  - Edge optimized endpoints require ACM certs in us-east-1
  - Regional endpoints require ACM certs in the same region
  - Can leverage AWS WAF for DDoS protection

AWS Batch

- Batched workload on EC2 or ECS/fargate
- Optimize the resource computation size accuracy based on number of submitted jobs
- Optimize the workload distribution
- Components
  - Job (shell scripts, executable or docker dimages)
  - Job definition : Blueprint of resources in the job
  - Job Queues : Jobs are submitted to queue an reside there until scheduled to run
  - Compute environment : Set of manager or unmanaged compute resources
- Fargate : recommended approach for most Workload
  - Fast start time < 30s
  - Require 16 vCPI or less and no GPUs
  - Require 120 GiB of memory or less
- EC2 :
  - If more controle is needed
  - Requires GPU, Elastic Fabric Adapter of custom AMIs
  - High levels of concurrency
  - Access to Linux parameters
- Note : Batch vs Lambda
  - Lambda have 15 minutes execution time limit
  - Lambda have limited disk space and EFS requires functions live within VPC
  - Lambda is fully serverless - no runtime
  - Batch uses Docker
- Note lambda can initiate Batch command

- Managed vs Unmanaged environment

AWS Amazon MQ

- Message broker server that allow easier migration of existing apps in the Cloud
- Supports both ActiveMQ and RabbitMQ
- Required private networking like VOC, DDirect Connect or VPN
  - In contrast SNS and SQS are publicly accessible
- Amazon MQ Brokers
  - ActiveMQ
    - Active/standby deployment
  - RabbitMQ
    - Logical grouping of three broker nodes across multiple AZs
    - Behind NLB

AWS Step functions

- Serverless orchestration service (integrates many AWS Services)
- Graphical console
- COpponents : state machines, tasks
  - State machine : Workflow with event-driven steps
  - Task; Specific states within a state machine representing a Single Unit of Work
  - States Every single step in the state machine
  - Execution : instance of state-machine
- Amazon State Language to describe states
- Two types
  - Standard
    - Have only one execution
    - Execution can run for up to 1 year
    - long-running workflows that have auditable history
    - Rates up to 2K executions/sec
    - Pricing based on state transition
  - Express
    - At least once
    - Execution can run for up to 5 minutes
    - High event rate workloads
    - Use case : IoT data streaming and ingestion
    - Pricing based on number of executions, durations and memory consumed
- Different states
  - Pass : pass any input to output
  - Task
  - Choice
  - Wait
  - Succeed
  - Fail
  - Parallel
  - Map : runs a set of steps based on elements of an input array

Amazon AppFlow

- Integrate data between SaaS (Salesforce, Slack, etc.) and AWS Services (S3, Redshift) via ingestion
- Bi-directional data transfers with limited combinations
- Concepts
  - Flow : Transfer data between sources and destinations
  - Data Mapping
  - Filters
  - Trigger
    - Run on demand
    - Run on event
    - Run on schedule

Exam Tips

- /! Never Tightly couple - always Loose Couple
- /! Every level of an application should be Loosely coupled
- Loosely Coupled = LB + Gateways + Async Messaging

13-big-data

Bigdata

- 3V’s
  - Volume : ranges from TB to PB of data
  - Variety
  - Velocity : speed of collection, storage, processing and analysis

Amazon Redshift

- Managed PB scaled Data warehouse (up to 16PB)
- Large relational DB : Standard SQL and BI Tools
- Based on PosgtreSQL - no OLTP Workloads neither and RDS replacement
- 10x other DWH
- Columnar
- Multi-AZ : only 2 AZ
- Snapshots are incremental and point in time : can be automated or manual
  - Always contained in S3 (you cannot managed the bucket)
- No conversion from Single AZ to multi-AZ and vice-versa
- /! Exam tip : always favor large batches to optimize Redshift perfs

Redshift spectrum

- Query and retrieve S3 Data without loading data into Redshift tables
  - Uses independent Redshift servers

EChanced VPC Routing

- All COPY and UNLOAD Traffic between cluster and data repositories are forced through a VPC
  - Enhances data security an controls
  - Use of VPC Features like Endpoints and Flow Logs

AWS EMR - Elastic MaPReduce

- AWS Managed ETL
  - Use cases : Web indexing, machine learning, large-scale genomics, etc.
- Uses open source tools such as Spark, Hive, HBase, Flink, Hudi and Presto

- 3 types of storage
  - HDFS : popular for caching during processing
  - EMR File System : extends Hadoop with the ability to access S3 as if part of HDFS
  - Local file system : EC2 instance store volumes
- Clusters and undoes
  - Cluster = group of EC2 instances (node)
  - Primary Node : manage the cluster, distribute data and tasks, tracks health status
  - Core Node (long running) : run tasks and stores data in HDFS
  - Task Node (optional) : run tasks with no storage - typically Spot instances
- Purchase options
  - On-Demand
  - Reserved
  - Spot
  - Type : Long-running or transient

AWS Kinesis

- Ingest, process and analyze real-time streaming data (huge data highway)
- 2 Versions
  - Kinesis Data Streams : real time
    - Consumer creation is managed by architect (app dev)
    - Model
      - Producers
      - Shards (n shards)
      - Consumers (EC2 instances pointing to output storage (S3, Redshift, Data Firehose, EMR, DynamoDB etc.))
  - Kinesis Data Firehose : Near real time
    - Sends only to S3, Redshift, Elasticsearch or Splunk
    - Plug and play
    - Model
      - Input -> Data Firehose -> output storage (S3, Redshift, etc.)
- Kinesis Data Analytics and SQL
  - Analyse data using standard SQL
  - No servers, transparent
  - Pay as u use
- Kinesis vs SQS
  - In contrast with SQS - Kinesis is real time
  - SQS is Simpler

AWS Athena & Glue

- Athena : Serverless interactive Query service to analyse S3 using SQL
  - Directly query S3 Data without loading into DB
- Glue : Serverless data integration service
  - ETL without EC2 instances

    - Replace EMR
- Athena and Glue
    - S3 -> Glue Crawlers (build structure for data) -> Glue Data Catalog -> Athena -> Quick Sight
    - Alternatively : S3 -> Glue Crawlers (build structure for data) -> Glue Data Catalog Redshift Spectrum

AWS Quick Sight

- Serverless, Fully managed BI Data viz service
- Create dashboards
- Share dashboards with users and groups
- Use cases
    - Data viz, ad-hoc analytics, business insights
- Integrates with : RDS, Aurora, Athena, S3, etc.
- SPICE : Robust in memory engine - advanced calculations
- Enterprise offering : Column-Level Security (CLS)
- Pricing : Per session, per user
- Create Users and Groups (EEnterprise version) for Quick sights - not same as IAM ones
- Create Dashboards and share them

AWS Data pipeline

- Managed ETL used for automated movement of data
- Data Driven Workflows
- Define parameters for data transformations
- Highly Available, Fault tolerant
- Handles Failures and integrates with SNS
- Works with AWS Storage services
- Works with AWS Compute services (EC2, EMR)
- Components
    - Pipeline Definition
    - Managed Compute
    - Task Runners
    - Data nodes
    - Activities : pipeline components to define work to perform
- Flexibility to schedule tasks
- Use cases :
    - Processing data in EMR using Hadoop streaming
    - Importing or exporting DynamoDB data
    - Copying CSV files between S3 buckets
    - Exporting RDS data to S3
    - Copying data to Redshift

AWS MSK - Managed Streaming for Apache Kafka

- Managed streaming service for Apache Kafka
- Provides Control plane operations : creates, updates and delete clusters

- Leverage Kafka data plane operations for producing, consuming streaming data
- Open source version of Apache Kafka - support of existing apps, tools and plugins
- Component
  - Broker nodes
  - ZooKeeper nodes
  - Cluster Operations
  - Producers, Consumers and Topics
- Residency
  - Automatic Recovery - same state (broker IP) after successful recovery
  - Failure Detection
  - Reduce data: Reuse storage from older brokers
  - Time required
- MSK Connect : stream data to and from Apache Kafka clusters
- Security and logging
  - Integrates with SSE
  - Encrypt data at rest by default
  - TLS 1.2 for encryption in transit
  - Broker logs to CloudWatch, S3 and Kinesis Data Firehose
  - Metrics are sent to CloudWatch
  - All MSK API calls are logged to AWS CloudTrail

AWS OpenSearch

- Managed services for search and analytics engines
- Successor of Amazon Elasticsearch Service
- Used for managed analytics and viz service
- Features
  - Quick analysis : ingest search and analyse data
  - Scalable
  - Security : IAM, VPC SG, encryption at rest and in transit and field level security
  - Stability : Multi-AZ Capable (master and snapshots)
  - Flexible : Can use SQL for BI apps
  - Integrates with CloudWatch, CloudTrail, S3 and Kinesis
- /! OpenSearch service is widely used for log analytics

14-serverless

Serverless

- Run code
- Event based
- Pay as you go : provisioned resources and length of runtime

Lambda

- Free tier 1M request, 400K GBs of compute per month
  - After that, pay per request
- Integrates with numerous AWS services (S3, DynamoDB, EventBridge, SQS/SNS, Kinesis)
- Built-in Monitoring and Logging (CloudWatch)
- Up to 10, 240 MB Memory - CPUs scale with memory
- 900 seconds (15minutes) execution length (short time executions)
- Support large variety of languages : Python, Golan, Java, Node.js, etc.
- Run inside or outside (default) a VPC
  - Ex. if we need to access private RDS instance we need to run inside VPC
- Configuration
  - Runtime
  - Permission (attach Role)
  - Networking (VPC/subnet/SG/etc.)
  - Resources : Memory
  - Trigger : Event that runs Lambda
- Quotas
  - 1K concurrent execution
  - 512 MB - 10 GB dist storage /tmp
  - EFS if needed
  - 4 KB for all env variables
  - 128 MB - 10 GB memory allocation
  - Compressed deployment package (.zip) siez <= 50 MB
  - Uncompressed deployment package must be <= 250 MB
  - Request and response payload up to 6 MB
  - Streamed responses up to 20 MB
- Use case 1
  - Put object in S3
  - Launch trigger
  - Execute Lambda
  - Send output processed to S3 bucket or store to DynamoDB
- Use case 2
  - EventBridge Rule (Rate or CRON)

    - Launches Lambda function
    - The lambda function shuts down Dev Instances

AWS Serverless Application Repository

- Repository : Find, deploy, publish serverless applications
- Ability to privately or publicly share applications
- Manifest file : SAM Template
- Deeply integrated with AWS Lambda service
- 2 Options
    - Publish
        - Private by default
        - Explicit share
    - Deploy
        - Find and deploy published applications
        - Public applications don’t need AWS Account to browse
        - Be careful of trusting all applications

AWS ECS, EKS

- Running containers, Orchestration
- ECS - AWS Elastic Container service
    - Manage up to 1000s of containers
    - Place containers and keep them online
    - Appropriately registered with chosen load balancers
    - IAM : individual roles by containers
- Kubernetes : large scale orchestration
- EKS - AWS Elastic Kubernetes Service
    - Managed Kubernetes
    - Best used when you’re not all in on AWS
- ECS and EKS : Good for long running applications

AWS EKS Distro

- Kubernetes distribution based on Amazon EKS
- Fully managed on prem / in cloud / anywhere (self managed)

AWS EKS Anywhere and ECS Anywhere

- EKS Anywhere
    - On prems kubernetes cluster based on EKS Distro
    - Full lifecycle management of multiple k8s clusters and operates indep of AWS
    - Curated packages (extended core functionalities) Enterprise subscription
- ECS Anywhere
    - ECS on prems
    - Inbound traffic : no ELB support
    - Requirements : SSM Agent, ECS agent, Docker installed
    - Register external instances as SSM Managed instances

    - Installation script within ECS console (with SSM activation keys and commands)
    - Execute scripts on on prem VMSs or bare metal
    - Deploy containers using external launch type

Fargate

- Serverless compute engine for Docker Containers
- Requires the use of ECS or EKS
- Supports Windows and Linux apps
- ECS Launch types :
    - EC2
        - EC2 pricing model
        - User is responsible for underlying OS
        - Multi containers can share same host
        - Capable of mounting EFS
    - Fargate
        - AWS Managed the infra entirely (including OS and Host)
        - Pay based on resources and time run
        - Short-running tasks
        - Isolated Env by container
        - Can use IAM roles for tasks / services
        - Capable of mounting EFS
- Fargate vs Lambda
    - Chose Fargate for more consistent Workloads
    - To allow Docker use and better developers control
    - Chose Lambda fr unpredictable or inconsistent Workloads

EventBridge (formally CloudWatch Event)

- Serverless Event Bus
- Pass events from a source to and endpoint
    - Connects serverless application together
- Concepts
    - Events : Recorded change in AWS Env, SaaS partner or configured app/service including Scheduled or Realtime events
    - Rules : Criteria to match incoming events and send them to appropriate targets (patterns or schedules)
        - Event Pattern : Ex. EC2 Terminated
        - Scheduled : Rate Based or Cron Based
    - Event Bus : A router between events and targets
        - Every account have default bus
        - Can create custom bus
- Ex. Use case
    - User terminates instance
    - Event matches rule
        - Trigger lambda -> retarts instance
        - Publish SNS Message

ECR - Elastic Container Registry

- AWS Managed Container Image Registry
  - Each registry is regional
- Private container image repositories with resource based permissions via IAM
- Support Open Container Initiative OCI images, Docker images and OCI Artifacts
- Component
  - Registry : provided to each Account (one or more) for image storage
  - Authorization token
  - Repository : contain images
  - Repository Policy
  - Image
- Note: AWS ECR Public if for public image repositories
- Lifecycle Policies
  - RUses for cleaning up
  - can test before run
- Image scanning : identify software vulnerabilities
  - Scan on push
  - Generate reports
- Sharing :
  - Cross region
  - Cross account
  - Configured Per repo per region
- Cache Rules :
  - Caching public repos privately
- Tag mutability
  - Prevents image tags from being overwinter
- Integration : ECS, EKS, on Prems,

AWS Aurora Serverless

- Aurora provisioned vs Aurora Serverless
- Aurora Serverless
  - On Demand and Auto scaling (based on app demands)
  - Automation and adjusting capacity
  - Charged only for resources consumed per-second (budget friendly)
- Concepts
  - ACUs : Aurora Capacity Unit
  - Set min and max ACUs for scaling requirements - can be zero
  - AWS Managed Warm pools : Allocated quickly
  - Each ACU : 2 GiB mem with matching CPU and Networking capability
  - Same data resiliency as Aurora provisioned : 6 copies of data across 3 AZ
  - Multi-AZ for high availability
- Use cases :
  - Variable Workloads

    - Multi-tenant Apps
    - New Apps : unsure what db instance needed
    - Dev and Test
    - Mixed Use Apps : different traffic spikes
    - Capacity Planning : Easy to swap from provision to serverless and vice-versa

X-Ray - application insights

- Collects application data for insights about requests / responses
- View calls to downstream AWS resources
- Use traces
    - Tracing headers, trace data, X-Ray Daemon
- Concepts
    - Segments : resource names, request details, other data
    - Subsegments : more granular info
    - Service graph : graphical representation of interacting services
    - Traces : Trace ID
    - Tracing Header : named X-Amzn-Trace-Id
- X-Ray Daemon : Listen to port 2000, collects raw segment data and sends it to X-Ray API
    - When daemon is running it works along with AWS X-Ray SDKs
- Integrations : EC2, ECS Tasks, Lambda, SNS, SQS, API Gateway, Elastic Beanstalk

AWS AppSync

- GraphQL Interface
- Data from multiple sources

15-security

Security

DDos

- Layer 4 DDos attack : SYN flood attack
  - Amplification Attack : NTP, SSDP, DNS, CharGEN, SNMP attacks, etc.
    - Using Spoofed IP Address
- Layer 7 Attacks : Floods of GET and POST attacks

AWS CloudTrail

- Record AWS Management Console actions and API calls
- Increase visibility into user (user, source IP, time) and resource activity
- Logged Data
  - Metadata
  - Identify of caller
  - Time
  - Source IP address
  - Request
  - Response elements
- Can enable log file integrity validation within the CloudTrail console for trails.
- Use cases
  - After the fact incident investigation
  - Near real-time intrusion detection (coupled with Lambda)
  - Compliance

AWS Shield

- Free DDOS Protection
- Applies to all AWS Accounts on ELB, CloudFront and Route 53
- Protects against SYN/UDP floods, reflection attacks and other layer 3 and 4 attacks
- AWS Shield advanced
  - Protects against larger and more sophisticated attacks
  - Always on, flow based monitoring of network traffic - near real time notification of DDos
  - 24/7 access to DDoS Response Team (DRT)
  - Protects AWS bill against higher fees during spikes of DDoS Attach
  - Cost 3K$ per month

AWS WAF

- Web application Firewall (layer 7)
- Applies to CloudFront or ALB
- Configure conditions (IP, request params, content, etc.)
- HTTP 403 if blocking
- 3 behaviours : Allow all except, Block all except, count requests matching property
- Conditions
  - IP
  - Country
  - Values in Request header
  - Presence of SQL
  - Presence of script
  - Strings and regex in requests

AWS Firewall manager

- Security management service in a single pane of glass
- Centrally set up and manage firewall (and WAF) rules across multiple AWS accounts and applications
- Need AWS Organizations
- Force compliance

GuardDuty

- Threat detection service with AI
  - 7-14 days to set a baseline
  - 1 month free then charged based on quantity of activity
- Uses machine learning to continuously monitor for malicious behavior
  - Unusual API calls
  - Calls from known malicious IP
  - Attempts to disable CloudTrail logging
  - Unauthorized deployments
  - Compromised instances
  - Reconnaissance by would-be attackers
  - Port scanning, failed log in
- Alerts : GuardDury console and CloudWatch Events
- Receives feeds from third parties : Proofpoint, CrowdStrike, AWS Security, etc.
- Monitors CloudTrail logs, VPC Flow logs and DNS logs
- Centralize threat detection across multiple AWS accounts
- Automated response using CloudWatch and Lambda

AWS Macie

- PII - Personally Identifiable Information (HIPAA and GDPR)
  - Plus PHI and financial data

- Monitor S3 buckets (ML and Pattern matching) to discover sensitive data
  - Alerts about unencrypted buckets
  - Alerts about public buckets
  - Alerts about shared buckets (outside of organization)
- Alerts :
  - Macie Console
  - Integrates with EventBridge
  - Integrates with AWS Security Hub and other AWS Services

AWS Inspector

- Inspects Network and EC2 instances
- Assessment findings
  - Network assessment: Network config analysis - Check for ports reachable from outside VPC
  - Host assessment : Vulnerable Software, host hardening (CIS Benchmarks), secu best practices
- Need Agents on EC2 instances
- Run once or weekly

AWS KMS

- Key management service - create and control (lifecycle) encryption keys
- Integrates with other AWS Services
- CMK : Customer Master Key
  - Logical representation of a master key
  - Includes metadata : Key ID, creation date, description and key state
  - Contains Key material used to encrypt / decrypt data
- Start with creating CMK - you control its lifecycle and who can use and manage it
- HSM : Hardware security module : physical computing device that safeguards and manages digital keys and performs encryption / decryption functions
- CloudHSM : rent HSM from AWS
- 3 ways to generate CMK
  - AWS Created - generated with HSMs managed by AWS
  - Import key material from customer Key management infra
  - Import key material on AWS CloudHSM cluster as custom key store feature
- Key rotation
  - Can auto rotate CMK each year
  - Auto Rotation not supported for
    - Imported keys
    - Asymmetric keys
    - CloudHSM custom generated keys
- Policies : who has access to what
  - IAM policies vs Resource based policies
  - Key policies : resource based policies attached to CMK

    - 3 Ways
        - Key policy
        - IAM Policies in combination with Key policy
        - Grants in combination with key policy
- KMS vs CloudHSM
    - Shared tenancy of underlying hardware in KMS
    - Dedicated HSM to the account in CloudHSM
    - No automatic key rotation in CloudHSM

AWS Secret Manager

- Stores, encrypts and rotates database credentials and other secrets
- Encryption in transit and at rest - using KSM
- Fine grained access control using IAM policies
- Highly scalable
- Secrets are retrievable by API call
- /! if rotation enabled, Secrets Manager immediately rotates once to test the configuration
    - All applications should not hard code secrets but retrieve them from secret manager
    - Do not enable rotation if applications use embedded credentials

Parameter Store

- Secure, hierarchical storage for configuration data management and secrets management
- Data such as : passwords, database strings, license codes, etc.
- Plain text or encrypted data
- free service
- Limitation
    - number of parameters 10K max
    - no rotation

Temporarily share S3 Objects

- All object in S3 are private by default
- Only the owner has access
- Owner can share presigned URL using their security credentials to grant time limited permission to download
    - Security credentials
    - Bucket name and object key
    - HTTP method
    - expiration date time
- Any one who received the presigned URL can access
- Presigned Cookies
    - Access to multiple restricted files
    - Users having Cookies can access to the entire contents

Advanced IAM Policy Document

- ARNs : Amazon resource names syntax
  - arn:partition:service:region:account_id:[resource; resource_type/resource; resource_type/resource/qualifier; etc.]
  - partition can be : aws, aws-cn, etc.
  - service : iam, s3, ec2, etc.
  - Region is omitted for global services : arn:aws:iam:::: 123456789:user/ryan
  - Account is omitted in S3 : arn:aws:s3:::bucket_name/image.png
  - Wild cards : arn:aws:ec2:123456:instances/*
- IAM Policies
  - JSON Document that defines permissions
  - Identity Policy (to users and groups)
  - Resource Policy (to resources)
  - Policies have no effects until attached
  - List of statements
  - Each statement matches an AWS API Request
  - Statement
    - SID
    - Effect : Allow |Deny
    - Action : generally resource:Action (ex. dynamodb:Query, dynamodb:BatchGet*)
      - Uses wildcards
    - Resource : what the action is against (ex. arn:aws:dynamodb:::table/MyTable)
      - ARN using wildcards
  - Permission Boundaries
    - Delegate admin to other users
    - Prevent privilege escalation or unnecessarily broad permissions
    - Use cases :
      - Dev. creating roles for Lambda functions
      - Admins creating ad hoc users
  - Not explicitly allowed = implicitly denied
  - AWS joins all applicable policies
  - AWS Managed vs Customer managed policies

AWS Certificate manager

- Create, manage and deploy public and private SSL certificates for user with other services
- Integrates with ELB, CloudFront, API Gateway, etc.
- No paying for SSL
- Automate renewals and Deployment

AWS Audit manager

- Continuously audit AWS usage
- Compliance with industry standards and regulations

- Produces specific reports to auditors for PCI compliance, GDPR, etc.

AWS Artifacts

- Artifact : single source to get compliance related informations
  - AWS Security and compliance reports
  - Select online agreements
- Huge number of reports available
- Download compliance document

Amazon Cognito

- Authentication, authorization and user management for Web and mobile apps (Identity broker)
- Single service without custom code
- Access server-side resources (token with permissions)
- Access AWS AppSync (remember : GraphQL / Serverless) Resources
- Third party sign-in
- Sign-up / Sign-in
- Guest user access
- Sync user data across multiple devises
- Recommended for mobile applications
- Two components
  - User pools
    - Directory of users providing sign-up / sign-in
  - Identity pools
    - Allow to give user access to other AWS Services
    - Exchange tokens and get AWS credentials to access AWS Services with them
  - User and Identity pools can be used separately or together
- STS - AWS Security Token Service to validate tokens

Amazon Detective

- Analyse, investigate and identify the root cause of potential security issues or suspicious activities
- Uses ML, Statistical analysis and graph theory
- Sources : VPC Flow Logs, CLoudTrail logs, EKS audi logs, GuardDuty findings, etc.

AWS Network firewall

- Managed service to deploy a physical firewall protection across VPCs
- Has a managed infrastructure (physical firewalls by AWS)
- Complete control over network traffic
- Use cases :
  - Filtering traffic before it reaches internet Gateway

    - Intrusion prevention system
    - hardware requirements

AWS Security Hub

- Single place to view all security alters from services like
    - GuardDuty
    - Inspector
    - Macie
    - Firewall manager
- Works across multiple accounts

16-automation

Automation

CloudFormation

- Infrastructure as a code : YAML or JSON Templates
  - Immutable
  - Versioning
  - Leverage stack tags
  - Consistency
- Resources and relationships
- Process
  - Write declarative template
  - Deploy the code as Stack or Stack Set
    - CloudFoundation do the AWS API calls to create and configure resources
    - Admin account can create and manage stack across multiple accounts and regions with single operation
      - This is stack sets
    - Stacks are regional resources
    - Can deploy portable stacks into other regions
    - Change preview (change sets) on existing stack
    - Caution to replacement attribute (replace resource if true - lose data)
- Rollback on error
- Prevent hard coded values to prevent CloudFormation failures
- Template sections
  - AWSTemplateFormatVersion (Optional)
  - Parameters (Optional)
  - Mappings (Optional)
    - Parameter based on criteria
    - Ex. RegionMap => based on region mappings
  - Resources (Required)
  - Outputs (Optional)
    - To S3 or others
  - Transform (Optional)
    - Macros and custom processing to transform the template before transforming
- Can use designer for visual representations
- Note : The immutable pattern specifies a deployment of application code by starting an entirely new set of servers with a new configuration or version of application code. When we go immutable, we don’t want to ever upgrade in place. Once the cloud resource exists, it can’t be modified.

Beanstalk

- PaaS : Platform as a service
- Build infrastructure from app code
- Developer, from code based instantiation
- Free to use
- Components :
  - Application : collection of Beanstalk components (env, versions, configs, etc.)
  - App version
  - Environment
    - Collection of AWS resources
    - Tiers
      - Web Server Tier : Web Env with ELP
      - Worker Tier : No client Env with SQS Queue
    - Can have multiple environments
      - Supports : Go, Java, Tomcat, .NBet, NodeJS, PHP, Python, Ruby, Packer Builder, Docker, multi container Docker, preconf Docker
- Deployment modes
  - Single instance (Elastic IP)
  - High available with Load balancer (ALB, ASG, RDS Redundancy)
- Need IAM rights
- Uses Cloud Formation

Systems Manager

- View, patch, manage and configure EC2 instances and On prems resources
- Need IAM permissions
- SSM Agent
  - Software component to install on instances (EC2 or on prems)
- Capabilities
  - Automation : use predefined or custom playbooks (documents) to manage resources
  - Run Command : Remotely execute commands without SSH or DP
  - Patch Manager : Automate patching (OS and application patching)
  - Parameter Store
  - Maintenance Windows : Define a schedule for performing actions on instances
  - Session Manager : Securely connect without needing SSH access
- Logging : SM logs all usage to CLoudWatch and CloudTrail
- SSM Agent
  - Amazon software that runs on your compute
  - Preinstalled on majority of AMIs
  - Make sure to have the IAM permissions
  - Possible de install Agent on on prems resources

- Parameter Store
  - Free feature to store config data and secret values in hierarchical manner with parameter policies (expiration dates)
  - Types : String, StringList, SecureString (with KMS)
  - ex. /dev/squid_conf=……
  - can be references this way : {{ssm:/dev/squid_conf}} in commands
- Connect with SSM in Console (user : ssm-user)
- Run commands

17-caching

Caching

- AWS Cache
  - Externally (o user) vs Internally (to local services)
  - When ever possible favor answer with caches

CloudFront

- CloudFront is CDN distributing data, videos, applications and APIs using AWS Edge locations
- Used for speed
- Settings
  - Security
    - Default to HTTPS
    - Can use custom SSL Certificate
  - Global distribution
    - Cannot pick specific countries, just general areas (use WAF if wanna blocking connections/countries)
  - Endpoint support
    - Can be used to front AWS Endpoints along with non AWS application
  - Expiring Content
    - TTL
    - Can force the expiration

ElastiCache

- Managed version of Memcached and Redis
  - Generally Sits in front of DB (excels in front of RDSs)
- Memcached
  - Simple database caching solution
  - Not a database
  - No failover or multi AZ
  - No backups
- Redis
  - Caching solution
  - Functions as a standalone database
  - Failover and multi-AZ
  - Supports backups

DAX - DynamoDB Accelerator

- In Memory Cache : reduces DynamoDB response times to microseconds
- Highly available and lives inside the VPC
- Custom Node size and count for the cluster, TTL for the data and maintenance windows for updates

Global Accelerator (GA)

- IP Caching -> think GA
- Networking services that sends user traffic through AWS Global network infrastructure via accelerators
- Increase performance and deal with IP caching by leveraging Anycast IP
- Meant for TCP/UDP traffic
- Concepts
  - Accelerators : Directs user traffic to optimal AWS Endpoints
  - Listeners : Processes inbound connections based on ports and protocols
  - Endpoint : Resources that GA directs traffic to
- Function
  - 2 Static Anycast IP addresses for accelerators
  - Dual stack receives 4 static IP addresses (2 v4 and 2 v6)
  - Static IPs single fixed entry for ALL client traffic
  - Traffic routed based on location, health checks and wights
  - Traffic routed to specified EC2 instances and ports in VPC

18-governance

Governance

AWS Organizations

- Free governance tool to manage multiple AWS Accounts
- Controls accounts from single location
- Account types
  - Management account - Payer account
  - Member account - all others
- Features
  - Consolidated billing
  - Isolate costs by account
  - Usage Discounts : Aggregate usage discounts
  - Shared Savings : share reserved instances and savings plans across the org
- Concepts
  - Multi-account : improved security, cost management
  - Tag enforcement
  - Organizational Unit (OU) : Logical grouping of multiple accounts
  - Service Control Policies (SCPs) : JSON Policies (IAM like) applied to OUs aor accounts to restrict actions
  - Management account : Not affected by SCPs
  - Account Best Practices :
    - Centralized logging account for organizational CloudTrail logs
    - Cross-account roles for accessing member accounts

AWS RAM

- Resource Access Manager
  - Facilitate sharing resources with other accounts (even out of organization)
  - Free to use
- Shared resources
  - Transit gateways
  - VPC subnets
  - License Manager
  - Route 53 resolvers (Rules and Endpoints)
  - Dedicated hosts
  - Etc.
- Owners and participants
  - Owner create and manage VPC resources that get shared

    - Owner cannot delete or modify resources deployed by participants
    - Participant is able to provision services into the shared VPC subnets
    - Participant Cannot modify or delete the shared resource

Cross account Role access

- Cross account role access gives the ability to set up temporary access that can be managed
- No need for long-term access keys or IAM users
- Role can be revoked as needed
- Temporary (credentials expires)
- Steps
    - Create role with desired privileges
    - Update Trust Policy to allow role assumption from other account identified by its ARN ID (Principal = AWS account and action = sts:AssumeRole)
    - Provide ARN of the Role to the external account

AWS Config

- Inventory Management and control tool
- Pay per item
- Historical record of configuration history of infrastructure over time
- Can create rules to make sure resources conform to requirements
    - Flag when something is going wrong
- Can receive alerts via SNS
- Configured by region
- Aggregated results across regions and accounts
- Rules
    - AWS predefined managed config rules
    - Custom config rules
    - Rules evaluated on a schedule or by a trigger
    - AWS Config is monitoring and assessment but NOT PREVENTIVE
- Remediation
    - Auto remediation via SSM Automation Documents
    - AWS Managed or Custom automation documents
    - Custom automation documents can leverage Lambda function or custom logic
    - Can retry on failure
- Alerts and Events
    - SNS Topics integration
    - Alert configuration changes or compliance state notification
    - EventBridge integration : send events to SQS or Lambda

Directory Service

- AWS Directory Service - managed Active Directory

- Types
  - Managed Microsoft AD : Entire AD suite
  - AD Connector : Creates a tunnel between AWS and on-prems AD
  - Simple AD : Standalone Linux Samba AD

AWS Cost explorer

- Visualize and analyse cloud costs
- Generate custom reports based on variety of factors including resource tags
- Break down cost by Monthly, Hourly, etc.
- Built in forecasting up to 12 months
- Features
  - Time
  - Filter (tags, categories, accountIDs, etc.)
  - Service

AWS Budgets

- Service to plan and set expectations around cloud costs
- Track ongoing spend, create alerts when close de exceeding allotted spend
- Types of budgets
  - Cost budgets : How much to spend on a service
  - Usage budgets : How much to use on one or many services
  - RI Utilization Budgets : Utilization threshold (RIs are used or under-utilized)
  - RI Coverage Budgets : Coverage threshold (how much instance usage is covered by a reservation)
  - Savings Plans Utilization : Utilization threshold
  - Savings Plans Coverage Budgets : Coverage threshold

AWS Cost and usage reports (CUR)

- Most comprehensive set of cost and usage data
- Publish billing reports to S3
- Break costs down by hours, day and month, service and resource or by tags
- Update reports in S3 once a day
- Easily integrate with Athena, Redshift or QuickSight
- Use within AWS Org, OU groups or individual accounts

AWS Compute optimizer and Savings Plans

- Compute optimizer
  - Disabled by default
  - Analyzes configuration and utilization metrics of AWS resources
  - Reports current usage

    - Produce recommendations
    - Visualize history and projected utilization metrics
    - Used to make informed decisions based on graphs, metrics and recommendations
    - Resources
        - EC2, ASG, EBS, Lambda
    - Supported Accounts : Standalone, Org Member account, Org Management account
- Savings Plans
    - Flexible pricing models for up to 72% savings
    - Lower prices for EC2 instances, Lambda, AWS Fargate and SageMaker
    - Regardless of instance family, size, OS, tenancy or Regions
    - Long-term commitments : 1 or 3 years
    - Pricing plan : All Upfront, Partial Upfront, No Upfront
    - Savings Plans Types
        - Compute Savings
        - EC2 Instance Savings
        - SageMaker Savings
    - Use recommendations section of AWS billing console

AWS Trusted Advisor

- Best practices auditing tool
- Recommendations for 5 categories
    - Cost optimization
    - Performance
    - Security
    - Fault Tolerance
    - Service Limits
- Uses industry and customer established best practices
- Works at an account level
- Basic or Developer support plans
- Business, Enterprise, On-Ramp support (full Advisory checks)
    - ■ integration with EventBridge

AWS Console tower

- Orchestration service that automates account creation, management and security controls
- Extends AWS Organizations to prevent governance drift
- Leverage different guardrails
- Central Admin Compliance policies
- Features
    - Landing zone : multi account environment
    - Guardrails :
        - High level rules
        - Preventative and detective
    - Account factory : Account template - pre-approved configs
    - CloudFormation StackSet : for repeated resources for governance (like config rules)

    - Shared accounts : 3 accounts used by Control Tower created during landing zone creation

AWS License manager

- Centrally manage licenses across AWS accounts and on prems env.
- Set Usage Limits
- Reduce Overages and prevent license abuse
- Versatile

AWS Personal Health Dashboard

- Visibility of resource performances and availability of AWS services or accounts
- View how health events affects services, resources and accounts
- Naming : Shifting to call the service AWS Health
- Timely, automations and alerts (near real time)
- EventBridge integrations
- Concepts
    - Health event : notification
    - Account specific event
    - Public event
    - AWS Health dashboard
    - Event type code
    - Event type category
    - Event status : open, closed or upcoming
    - Affected entity
- Examples : EC2 system reboot maintenance scheduled, EC2 Operational issue, Billing suspension notice, etc.

AWS Service Catalog and AWS Proton

- AWS Service Catalog
    - Create and Manage Catalog of Approved IT Services
    - Multipurpose : AMIs, servers, software, DBs, preconfigured components
    - Centralized service using Organizations
    - End user friendly : easy deploy of pre-approved catalog items
    - CloudFormation templates based
- AWS Proton
    - Create and manage infrastructure and deployment tooling for users as well as serverless and container based apps
    - Automate Infra as Code provisioning and deployments
    - Define standardized infra for serverless and container based apps
    - Use templates to define and manage app stacks that contain ALL components
    - Auto provision resources, configure CI/CD and deploy code
    - Supports AWS CloudFormation and Terraform IaC providers

AWS Well architected tool

- Well-architected Framework - Six pillars
  - Operational Excellence
  - Reliability
  - Security
  - Performance Efficiency
  - Cost optimization
  - Sustainability
- Well-architected Tool
- Measure cloud architecture
- Assistance with documentation
- Guides
- Measure workloads againts years of AWS Best pratices
- Specific audiences : tech teams, CTOS, archi and ops teams

19-migration

Migration

Snow family

- Load data in truck and ship it physically to AWS
- Both ways : On Prem <-> AWS
- Types
  - Snowcone
    - 8TB, 4GB RAM, 2vCPUs
    - IoT sensor integration
    - Perfect for edge computing
  - Snowball Edge
    - 48 to 81 TB storage
    - Varying amount of CPU and RAM
    - Perfect for off the grid and migration
  - Snowmobile
    - 100 PB of storage
    - Designed for exabyte scale data

Storage Gateway (Hybrid solution)

- Hybrid Cloud storage
  - Merge on prem resources with the cloud
  - One time migration or long term pairing
- File Gateway
  - Caching local files
  - NFS or SMB mount
  - Extend on prem storage
- Volume Gateway (backup drives)
  - ISCSI mount
  - Cached or stored mode
  - Create EBS snapshots
- Tape Gateway
  - Replace physical tapes

AWS DataSync (One time migration)

- Agent based solution for on-prem to AWS storage migration (S3, EFS, FSx)
- Move data between NFS and SMB shares and AWS storage solutions

AWS Transfer Family

- Move files in and out S3 or EFS using SFTP, FTPS or FTP

AWS Migration Hub

- Single place to track the progress of application migration to AWS
- Integrates with Server Migration Service (SMS) and DB Migration Service (DMS)
- SMS
  - Schedule migration
  - Uploads
  - Convert
  - Create AMI
- DMS
  - AWS Schema conversion Tool
  - Upload to Aurora

AWS Application Discovery Service and Application Migration Service

- Application Discovery Service
  - Plan migrations to AWS via usage and config data from on prems servers
  - Integrates with AWS Migration Hub
  - Discover servers and group by application
  - Track each application migration
  - Discovery types
    - Agent less
      - Agentless collector
      - OVA within VMware vCenter
    - Agent Based
      - AWS Application Discovery Agent to install on each VM and Physical server
      - Windows and linux
- Application Migration Service (AWS MGN)
- Automated Lift and shift service for expediting migration of apps to AWS
- Flexible
- Avoid cutover windows or disruptions
- Features : RTO (minutes - OS boot time) and RPO (sub second range)

Zoom to AWS Database migration Service

- Migrates relational, non relational and data warehouses
- On prems <-> AWS
- One time or On going (continuously replicate changes)
- Endpoints : Source and target data stores
- Same engine migration vs different engine migrations

- Must have ONE endpoint at AWS
- SCT - Schema Conversion Tool
  - Supports both OLAP and OLTP and data warehouses
  - Any supported Amazon datastore type
  - Can even use converted schemas on DB running on EC2 or data stored in S3
- 3 migration types
  - Full load
  - Full load and CDC - Change Data Capture (only one that guarantee transactional integrity)
  - CDC only
- Can migrate data stores with AWS Snowball
  - Still use SCT
  - CDC compatible

20-web-n-mobile

Web and mobile

AWS Amplify

- Tools for front end web and mobile developers to quickly build full stack apps on AWS
- 2 Services
  - Amplify hosting
    - SPA : React, Angular and Vue
    - Gatsby and Hugo
    - Separate prod and staging envs
    - SSR - Server Side Rendering
  - Amplify Studio
    - Easy auth and authorizations
    - Simplify developments
    - Ready to use components

AWS Device Farm

- Application testing services
- Android, iOS, Web
- Uses actual hosted devices
- 2 testing methods
  - Automated : upload script
  - Remote access : Swipe, gesture and interact with devices in real time via browser

AWS Pinpoint

- Engage with customers via message channels (email, SMS, push messages, etc.)
- Features
  - Projects : Information’s, segments, campaigns, journeys
  - Channels
  - Segments : categorising users
  - Campaigns : Initiatives for specific audience
  - Journeys : Customized multi step engagement
  - Message template
  - Machine learning
- Usage :
  - Marketing
  - Transactions (order, billing, etc.)
  - Bulk

21-machine-learning

Machine learning

- /! Rekognition and Sagemaker are most used

Amazon Comprehend

- Comprehend : NLP - Natural language processing
  - Use cases : Call center analytics, Index and search product reviews, process financial documents
- Kendra : Intelligent search service
  - Use cases : Accelerate R&D, improve customer interactions
- Textract : Auto extract text, handwriting or scanned
  - Use cases : Financial services, Health case, etc.

Amazon Forecast

- Time series data forecasting service using ML algorithms
- Use cases : IoT, Analytics, DevOps, etc.

Amazon Fraud Detector

- Account fraud detection
- Suspicious online payments, account fraud, prevent trail and loyalty
- Improve account takeover detection

Amazon Text and speech

- Transcribe : Speech to text
- Lex : build conversational interfaces using NLM - Natural Language models
  - ■ Voice assistants
- Polly : Text to Lifelike speech

Amazon Rekognition

- AWS Computer vision product for photos and videos
- Use cases
  - Content moderation
  - Face detection and analysis
  - Celebrity recognition
  - Streaming Video Events Detection

Amazon Sagemaker

- Manage, package and deploy ML Models
- Components
  - Ground Truth : Labeling jobs for training datasets using active learning and human labeling
  - Notebook : Managed Jupyter Notebook environment
  - Training : train and tune models
  - Inference : Package and deploy ML Models at scale
- 2 Deployment types
  - Offline Usage (async or batch)
  - Online Usage (real time) - if need immediate response
- Stages
  - Create a Model
  - Create and Endpoint configuration
  - Create an Endpoint
- Model training
  - Load data in container registry
  - Etc.
- SageMaker Neo
  - Customize ML models for specific Hardware : ARM, Intel, NVidia
  - Use TensofFlow, ONNX, etc. software
- Elastic Inference
  - EI speeds up throughput and decrease latency of real time inferences using only CPU instances
  - Much more cost effective thant full GPU instance
- High availability
  - Recommendation : Deploy ML in different AZ

Amazon Translate

- Auto language translation
- Uses deep learning and neural network
- Scalable, Cost effective, easy to integrate with application

22-media

Media

- AWS Elastic transcoder
  - Media converter
- AWS Kinesis Video streams
  - Streaming media content from large number of devices to AWS
  - Run analytics, machine learning, playback, and other processing
  - Elastically scales to millions of devices
  - Encryption and indexes
  - Use cases : Smart home, smart city, industrial automation

55-glossary

Glossary

Security

- AWS CloudTrail : Record AWS Management Console actions and API calls
- AWS Shield : Free DDOS Protection
- AWS Shield advanced : Payed DDOS Protection
- AWS WAF : Applies to CloudFront or ALB
- AWS Firewall manager : Security management service in a single pane of glass
- AWS Network firewall : Managed service to deploy a physical firewall protection across VPCs
- AWS GuardDuty : Threat detection service with AI - 7-14 days to set a baseline
- AWS Inspector : Inspects Network and EC2 instances
- AWS Detective : Use ML to analyse, investigate and identify the root cause of potential security issues or suspicious activities
- AWS Security Hub : Single place to view all security reports from services
- AWS KMS : Key management service - create and control (lifecycle) encryption keys
- AWS Certificate manager : Create, manage and deploy public and private SSL certificates for user with other services
- AWS Secret Manager : Stores, encrypts and rotates database credentials and other secrets
- Parameter Store : Free, secure, hierarchical storage for configuration data management and secrets management (10 K max)
- AWS Macie : PII - Personally Identifiable Information (HIPAA and GDPR)
- AWS Audit manager : Continuously audit AWS usage, produces specific reports to auditors for PCI compliance, GDPR, etc.
- AWS Artifacts : Single source to get compliance related informations (reports to download)
- Temporarily share S3 Objects / CloudFront content : presigned URL or Cookies
- AWS Cognito : Authentication, authorization and user management for Web and mobile apps (Identity broker)

Governance

- AWS Organizations : Free governance tool to manage multiple AWS Accounts
  - Management vs member accounts
- AWS RAM - Resource Access Manager : Facilitate sharing resources with other accounts (even out of organization)
- Cross account Role access : Cross account role access gives the ability to set up temporary access that can be managed
- AWS Config : Inventory Management and control tool
- AWS Directory Service : Managed AD - Types : Managed Microsoft AD; AD Connector and Simple AD
- AWS Cost explorer : Visualize, analyse and generate reports for cloud costs and costs breakdown
- AWS Cost and usage reports (CUR) : Most comprehensive set of cost and usage data
  - Publish billing reports to S3
- AWS Budgets : Service to plan and set expectations around cloud costs (create alerts)
- AWS Compute optimizer : Analyzes configuration and utilization metrics of AWS resources then produce recommendations
- AWS Trusted Advisor : Best practices auditing tool (cost, perf, secu, fault tolerance and service limits)
- AWS Console tower : Orchestration service that automates account creation, management and security controls
  - Guardrail rules
- AWS License manager : Centrally manage licenses across AWS accounts and on prems env.
- AWS Personal Health Dashboard : Visibility of resource performances and availability of AWS services or accounts
  - EventBridge integrations, automations and alerts (near real time)
- AWS Service Catalog : Create and Manage Catalog of Approved IT Services
- AWS Proton : Create and manage infrastructure and deployment tooling for users as well as serverless and container based apps
- Savings Plans Types
  - Compute Savings
  - EC2 Instance Savings
  - SageMaker Savings

Migration

- Snowcone (8TB) < Snowball Edge (48 to 81 TB) < Snowmobile (100 PB)
- AWS Storage Gateway : Hybrid Cloud storage (File - NFS or SMB, Volume (EBS Snapshot - think backup) or Tape gateway)
- AWS DataSync (One time migration) : Agent based solution for on-prem to AWS storage migration (S3, EFS, FSx)
- AWS Transfer Family : Move files in and out S3 or EFS using SFTP, FTPS or FTP

- AWS Application Discovery : Plan migrations to AWS via usage and config data from on prems servers (Agent or Agent-less)
- AWS Application Migration Service: Automated Lift and shift service for expediting migration of apps to AWS
- AWS Database migration Service : Migrates relational, non relational and data warehouses
  - On prems <-> AWS
  - One time or On going (continuously replicate changes)
  - SCT - Schema Conversion Tool
  - 3 migration types : Full load; CDC - Change Data Capture; Full load and CDC (only one that guarantee transactional integrity)
- AWS Migration Hub : Single place to track the progress of application migration to AWS (AMS and DMS)

Machine learning

- AWS Comprehend : NLP - Natural language processing
- AWS Kendra : Intelligent search service
- AWS Textract : Auto extract text, handwriting or scanned
- AWS Translate : Auto language translation
- AWS Transcribe : Speech to text
- AWS Lex : build conversational interfaces using NLM - Natural Language models
- AWS Polly : Text to Lifelike speech
- AWS Forecast : Time series data forecasting service using ML algorithms
- AWS Fraud Detector : Suspicious online payments, account fraud, prevent trail and loyalty
- AWS Rekognition : Computer vision product for photos and videos
- AWS Sagemaker : Manage, package and deploy ML Models

Automation

- CloudFormation
- Beanstalk : PaaS
- Systems Manager : View, patch, manage and configure EC2 instances and On prems resources

Caching

- CloudFront : CDN

- Global Accelerator (GA) : Networking services that sends user traffic through AWS Global network infrastructure via accelerators
  - IP Caching -> think GA
- DAX : In Memory Cache - reduces DynamoDB response times to microseconds
- Elasticache : Managed version of Memcached and Redis

Web and mobile

- AWS Amplify (hosting and studio) : Web and mobile Dev
- AWS Device Farm : devices for tests
- AWS Pinpoint : Customer message channels (email, SMS, push messages, etc.)

Media

- AWS Elastic transcoder : Media converter
- AWS Kinesis Video streams : Streaming media content from large number of devices to AWS

Decoupling

- AWS Batch : Batched workload on EC2 or ECS/fargate
- AWS Amazon MQ : Supports both ActiveMQ and RabbitMQ
- AWS Step functions : Serverless orchestration service (integrates many AWS Services)
- Amazon AppFlow : Bi-directional data integration between SaaS (Salesforce, Slack, etc.) and AWS Services (S3, Redshift)

Big data

- Redshift : Managed PB scaled Data warehouse (up to 16PB)
- Redshift spectrum : Query and retrieve S3 Data without loading data into Redshift tables
- EMR - Elastic MaPReduce : AWS Managed ETL
- AWS Kinesis : Ingest, process and analyze real-time streaming data (huge data highway)
  - Kinesis Data Streams : Real time
  - Kinesis Data Firehose : Near real time
  - Kinesis Data Analytics : Analyse data using standard SQL
- AWS Athena : Serverless interactive Query service to analyse S3 using SQL
- AWS Glue : Serverless data integration service
- AWS Quick Sight : Serverless, Fully managed BI Data viz service
- AWS Data pipeline : Managed ETL used for automated movement of data
- AWS MSK - Managed Streaming for Apache Kafka
- AWS OpenSearch : Managed services for search and analytics engines

Serverless

- Lambda
- AWS Serverless Application Repository
- AWS ECS, EKS
- AWS EKS Distro
- AWS EKS Anywhere and ECS Anywhere : On prems
- Fargate : Serverless compute engine for Docker Containers
  - ECS Fargate vs EC2
- EventBridge (formally CloudWatch Event) : Serverless Event Bus
- ECR - Elastic Container Registry
- X-Ray - application insights : request tracing and analysis
- AWS AppSync : GraphQL Interface

99-exam

Exam

Pass mark

- 65 questions in 130 minutes
- Minimum 720/1000

Response types

- Scenario based
- Multi choice
- Multi responses (2 correct over five)

Budget setup

- Account can activate Billing and payment views to Admin users
- Bill view - Breakdown, forecast, charges by services etc.
- Budget setup : Creating budget alerts when threshold exceeded