aws is a global cloud provider (has *regions* all over the world)

region - cluster of data centers — a region has multiple availablity zones (AZ)

- **Availablity Zone (AZ)** - one or more discrete data centers with redundant power, networking and connectivity

Identity Access Management [IAM]

aws security

- Users - a physical person
- Groups - containes users (functions, teams)
- Roles - internal usage within aws resources (given to machines)

policies - are JSON documents that defines what users, groups, roles can and cannot do

least privilege principle - give the users the minimal amount of permissions they need to perform their job IAM Federation

- big enterprises usually integrate their own repository of users with IAM
- can use company credentials to log in
- uses SAML standard [Security Assertion Markup Language] (e.g. Active Directory)

Note:

- 1 IAM user per PHYSICAL PERSON
- 1 IAM role per APPLICATION
- IAM credentials should NEVER BE SHARED
- Never write IAM credentials in code or commit them
- Never use ROOT account except for initial setup

Elastic Compute Cloud [EC2]

EC2

- renting VMs (EC2)
- storing data on virual drives (EBS)
- distributing load across machines (ELB)
- scaling the services using an auto-scaling group (ASG)

key pair - consists of a public key that AWS stores, and a private key file that you store. Together, they allow you to connect to your instance securely.

- think of public key as a lock and private key as a key that can unlock that lock. You can give out your public key to anyone and they can use it to lock their messages which you can only unlock with your private key
- *.perm file is the private key will have 0644 permissions by default. It is required for the private key to not be accessible by others. Changing the permissions to 0400 will fix it since it will only be readable by the owner
- `ssh -i <private_key>.perm ec2-user@<IP> # To connect to an instance`

Elastic IP - fixed public IPv4 IP you own as long you don't delete it (avoid using this)

EC2 User Data

- bootstrapping - launching commands when a machine start
- it is possible to bootstrap instances using an EC2 User data script
- runs only once at the instance first start
- used to automate tasks such as
  - installing updates, software, common files from the internet, basically anything

EC2 Instance Launch Types (remember the hotel analogy)

- On Demand: short workload, predictable pricing
  - pay for what you use, has the highest cost but no upfront payment, no long term commitment
  - recommended for short-term, un-interrupted workloads where you can't predict how the application will behave (elastic workloads)
- Reserved: (minimum 1 year)
  - reserved instances: long workloads
    - Up to 75% discount compared to on-demand
    - pay upfront for what you use with long term commitment
    - recommended for databases
  - convertible reserved instances: long workloads with flexible instances
    - up to 45% discount

        - Scheduled reserved instances: specific times (every thursday between 3PM-6PM)
    - Spot Instances short workloads, for cheap (cost-efficient), can lose instances (less reliable)
        - 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
    - Dedicated Instances: no other customers will share your hardware
        - instances running on hardware that's dedicated to you
        - may share hardware with other instances in same account
    - Dedicated Hosts: book an entire physical server, control instance placement (3 year period reservation)

Elastic Network Interfaces (ENI)

- logical component in a VPC that represents a virtual network card
- can be created independently and attach them on the fly (move them) on EC2 instances for failover
- bound to specific availbility zone

AMI (American Machine Image) *(are built for specific regions)*

- pre-installed packages
- faster boot time (no need for long EC2 user data at boot time)
- comes configured with monitoring/enterprise software
- security concerns
- installing your app ahead of time (for faster deploys when auto-scaling)
- using someone else's AMI that is optimized for running an app, DB

Security Groups

fundamental of network security, control how traffic is allowed into or out of EC2 machines, act as virtual firewall
security group:

- can be attached to multiple instances
- locked down to a region
- lives "outside" the EC2 - if traffic is blocked the EC2 instance won't see it
- By default, all inbound traffic is blocked and all outbound traffic is authorized
- security groups can reference other security groups, IPs, CIDR blocks but not DNS names

Acronyms

EC2 - Amazon Elastic Compute Cloud
EBS - Elastic Block Storage
VPC - Virtual Private Cloud **ENI** - Elastic Network Interfaces
AMI - Amazon Machine Image