AWS Certified Solutions Architect - Associate

Week 1 – Content Review (Batch 2)

26 August 2021

DHAVAL SONI

infostretch

Agenda

- Week 1 - Review
- Topics we'll cover in next session
- Q&A

infostretch

CONTENT REVIEW

infostretch

What is Cloud Computing

Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing.

| Trade capital expense for variable expense | Increase speed and agility | Benefit from massive economies of scale | Stop spending money on running and maintaining data centers | Stop guessing capacity | Go global in minutes |

AWS Global Infrastructure

Global Infrastructure

#	Region & Number of Availability Zones		New Region (coming soon)

24 Geographic Regions

76 Availability Zones

205 Edge Locations

Region

AWS Availability Zone

Interconnected using high-speed private

AZ

AZ

Independent failure Zone

Region & Number of Availability Zones

US East
N. Virginia (6),
Ohio (3)

US West
N. California (3),
Oregon (5)

Asia Pacific
Mumbai (2),
Seoul (2),
Singapore (3),
Sydney (3),
Tokyo (4),
Osaka-Local (1)[1]

Canada
Central (2)

China
Beijing (2),
Ningxia (3)

Europe
Frankfurt (3),
Ireland (3),
London (3),
Paris (3),
Stockholm (3)

South America
São Paulo (3)

GovCloud (US)
US-East (3),
US-West (3)

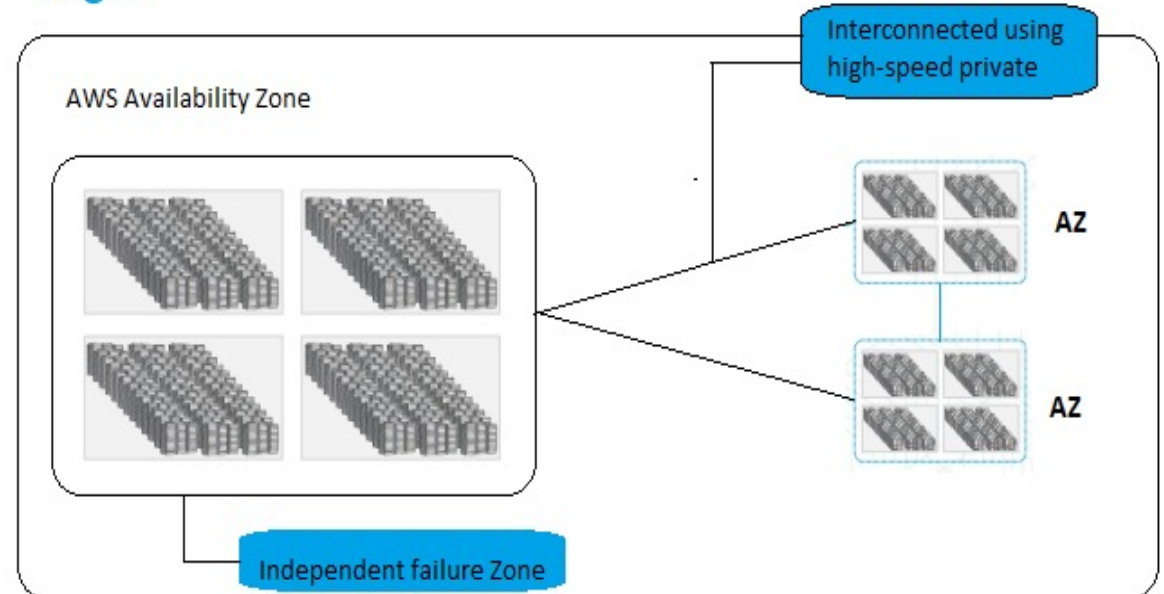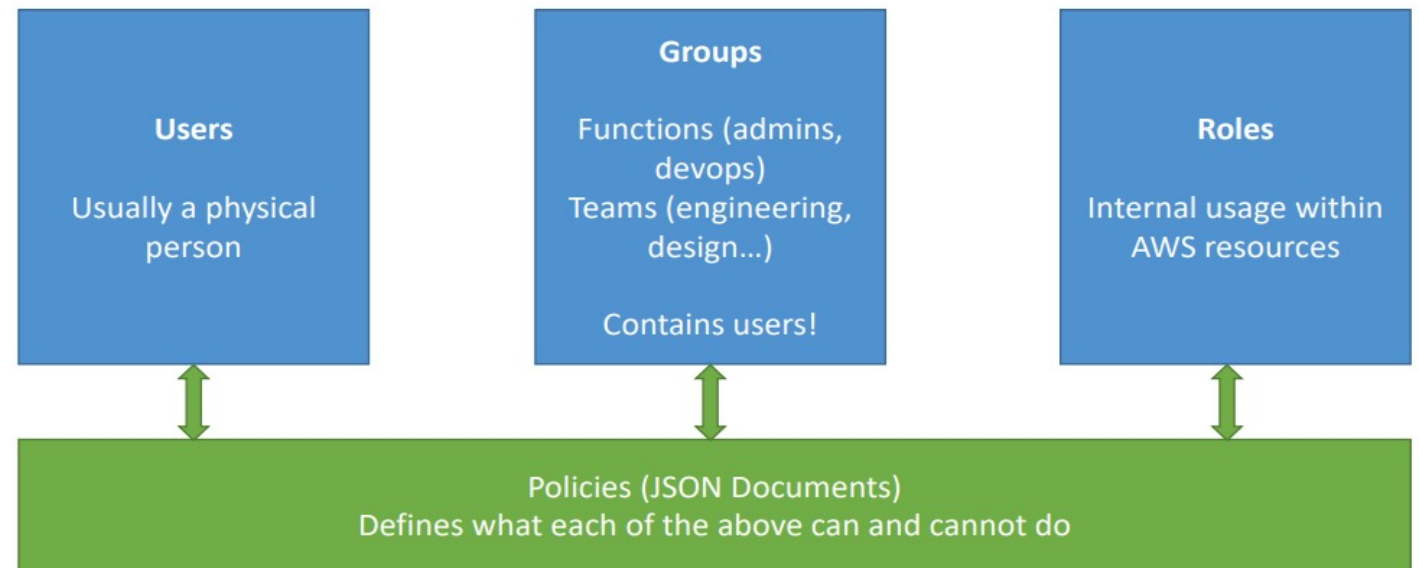**New Region (coming soon)**
Bahrain
Cape Town
Hong Kong SAR
Milan

Amazon IAM

Amazon Identity & Access Management

- IAM has a global view
- Permissions are governed by Policies (JSON)
- MFA (Multi Factor Authentication) can be setup
- IAM has predefined "managed policies"
- It's best to give users the minimal number of permissions they need to perform their job (least privilege principles
- Your Whole AWS Security is:
  - Users
  - Groups
  - Roles

AWS IAM

Users

Usually a physical person

Groups

Functions (admins, devops)
Teams (engineering, design...)

Contains users!

Roles

Internal usage within AWS resources

Policies (JSON Documents)
Defines what each of the above can and cannot do

Amazon EC2

Amazon Elastic Compute Cloud

- EC2 is one of most popular of AWS offering
- Knowing EC2 is fundamental to understand how the Cloud works
- Support numerous distributions of Linux or Microsoft Windows
- Complete control of your host operating system with root and administrator accounts
- Responsible for all installed applications
- It mainly consists in the capability of :
  - Renting virtual machines (EC2)
  - Storing data on virtual drives (EBS)
  - Distributing load across machines (ELB)
  - Scaling the services using an auto-scaling group (ASG)

EC2

EC2 User Data

- It is possible to bootstrap our instances using an EC2 User data script.
- bootstrapping means launching commands when a machine starts
- That script is only run once at the instance first start
- EC2 user data is used to automate boot tasks such as:
  - Installing updates
  - Installing software
  - Downloading common files from the internet
  - Anything you can think of
- The EC2 User Data Script runs with the root user

EC2 Instance Launch Types

- On Demand Instances: short workload, predictable pricing
- Reserved: (MINIMUM 1 year)
  - Reserved Instances: long workloads
  - Convertible Reserved Instances: long workloads with flexible instances
  - Scheduled Reserved Instances: example – every Thursday between 3 and 6 pm
- Spot Instances: short workloads, for cheap, can lose instances (less reliable)
- Dedicated Instances: no other customers will share your hardware
- Dedicated Hosts: book an entire physical server, control instance placement

EC2

EC2 Instance Types

- R: applications that needs a lot of RAM – in-memory caches
- C: applications that needs good CPU – compute / databases
- M: applications that are balanced (think "medium") – general / web app
- I: applications that need good local I/O (instance storage) – databases
- G: applications that need a GPU – video rendering / machine learning
- T2 / T3: burstable instances (up to a capacity)
- T2 / T3 - unlimited: unlimited burst

EC2

Placement Groups

- Placement strategy can be defined using placement groups
- When you create a placement group, you specify one of the following strategies for the group:
  - Cluster—clusters instances into a low-latency group in a single Availability Zone
  - Spread—spreads instances across underlying hardware (max 7 instances per group per AZ)
  - Partition—spreads instances across many different partitions (which rely on different sets of racks) within an AZ. Scales to 100s of EC2 instances per group (Hadoop, Cassandra, Kafka)

aws

Scalability & High Availability

- Scalability means that an application / system can handle greater loads by adapting.
- There are two kinds of scalability:
  - Vertical Scalability
  - Horizontal Scalability (= elasticity)
- Vertically scalability means increasing the size of the instance
- Horizontal Scalability means increasing the number of instances / systems for your application
- High Availability usually goes hand in hand with horizontal scaling
- High availability means running your application / system in at least 2 data centers (== Availability Zones)
- The goal of high availability is to survive a data center loss
- The high availability can be passive (for RDS Multi AZ for example)
- The high availability can be active (for horizontal scaling)

Amazon ELB

Amazon Elastic Load Balancer

- Spread load across multiple downstream instances
- Expose a single point of access (DNS) to your application
- Seamlessly handle failures of downstream instances
- Provide SSL termination (HTTPS) for your websites
- Enforce stickiness with cookies
- High availability across zones
- Separate public traffic from private traffic
- AWS has 3 kinds of managed Load Balancers:
  - Classic Load Balancer (v1 - old generation) – HTTP, HTTPS, TCP
  - Application Load Balancer (v2 - new generation) – HTTP, HTTPS, WebSocket
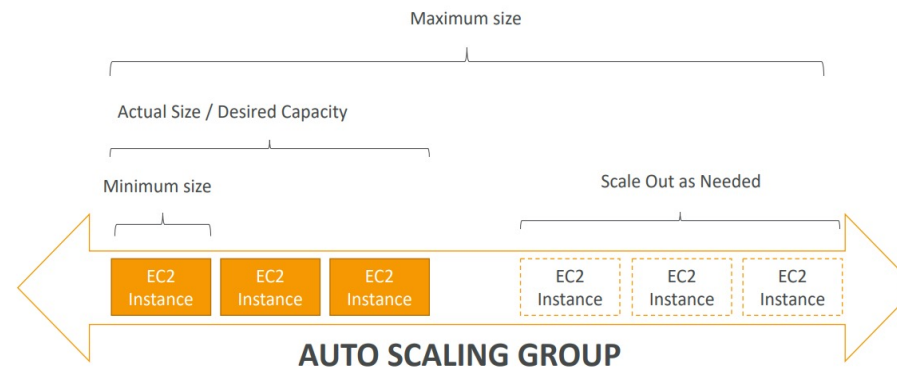 - Network Load Balancer (v2 - new generation) – TCP, TLS (secure TCP) & UDP

Amazon ELB

Amazon ASG

Amazon Auto Scaling Group

- In real-life, the load on your websites and application can change
- In the cloud, you can create and get rid of servers very quickly
- The goal of an Auto Scaling Group (ASG) is to:
  - Scale out (add EC2 instances) to match an increased load
  - Scale in (remove EC2 instances) to match a decreased load
  - Ensure we have a minimum and a maximum number of machines running
  - Automatically Register new instances to a load balancer

AWS ASGs

Maximum size

Actual Size / Desired Capacity

Minimum size

Scale Out as Needed

| EC2 Instance | EC2 Instance | EC2 Instance | EC2 Instance | EC2 Instance | EC2 Instance |

AUTO SCALING GROUP

Topics we'll cover in next session

- AWS Fundamentals: RDS, Aurora & ElastiCache
- Route53
- Amazon S3
- AWS Athena
- Amazon CloudFront
- Amazon Global Accelerator
- AWS Storage: Amazon FSx, Storage Gateway, Snow Family

Q & A

Thank you

CONNECT WITH US

EMAIL   INFO@INFOSTRETCH.COM

CALL   +1-408-727-1100

infostretch