Core AWS Services

THE AWS GLOBAL INFRASTRUCTURE

Global infrastructure provided by AWS is made up of the latest and greatest networking and datacenter technologies and it spans almost the entire globe. To understand the global infrastructure provided by AWS we can break it down to three main components.

Regions

To provide efficient global infrastructure, the globe is divided into multiple geographic locations and each location is served independently and called Region. One region may contain one or more availability zones, each AZ has an independent power provider so that if one is not available, services in that region can be handled with other availability zones.

Availability Zones

Availability zones are the collection of data centers in one region. Data centers are separated as much as possible for high availability but connected with very high speed links. Each availability zone has its own power provider, or multiple power providers and has its own backup generators for high availability.

Edge Locations

These locations host cached content from your architecture for fast delivery to clients. The caching technology of AWS is called CloudFront. Edge locations further lowers the latency to serve media and other content to clients after regions and availability zones. Edge locations are also entry points into the AWS network when using CloudFront or S3 Transfer Acceleration for data ingestion.

Introduction to Virtual Private Cloud (VPC)

The VPC allows you to create virtual private networks and use the same concepts of traditional networking. With a VPC you have complete control of network configuration. You have the ability to isolate or expose the resources to the public internet or your to your private host systems inside corporations.

High Availability

Your VPC lives within a region and you can have multiple VPCs per account. As built into the regions and availability zones, VPCs are highly available.

Subnets

Just like private network structure VPCs are made up of subnets that you can use to provide segmentation at Layer 3 (Network Layer)

Route Tables

You can use route tables to route traffic entering and exiting your subnet. You get this familiar model without needing to about physical routers themselves.

Internet Gateway

Permits easy to configure access to the internet for your VPC.

NAT Gateway

Translate your privately addressed VPC resources to access the internet using public IP addresses.

NACLs

Network access control lists allow you to control access to your VPC subnets.

Introduction to Security Groups

Because security of your resources in the cloud is a prime concern for both you and Amazon, it is no big surprise that AWS provides you with built-in firewalls with your compute resources. These security groups help you easily control the accessibility of your EC2 resources. Perhaps you have a Web tier in your AWS architecture. You can configure the security group for this tier to permit HTTP and HTTPS traffic from customers using the Web tier, while at the same time you can permit your team of support engineers to access the Web tier using SSH and RDP. All other protocol attempts at accessing the Web tier are denied by the security group.

Introduction to Compute Services

Lambda

AWS Lambda is an exciting alternative to EC2 instances that you must operate and maintain. Lambda provides compute resources in a fully managed (by AWS) serverless compute cloud. You send compute requirements to Lambda in a variety of different manners (such as a call from a Web app), and Lambda takes care of the compute requirements for you. Subsecond metering is used for your cost calculations, so quite often it is a very

inexpensive way to provide the compute resources you require. It also supports many different programming languages to ease use.

Elastic Beanstalk

Elastic Beanstalk offers a very quick and simple method for getting your applications into the AWS Cloud. It is actually a Platform as a Service (PaaS) offering. The infrastructure and platform are quickly built for you in the cloud. This permits the quick deployment of your applications. Elastic Beanstalk also reduces the ongoing management complexity of your deployment. Importantly, you maintain control of the platform. For example, should you want to scale your applications more aggressively, you have complete control. Another great aspect to this service is that it supports a wide variety of languages and platforms, such as Go, Java SE, PHP, Python, and Node.js, just to name a few. Application upgrades are simple, as you just deploy them to Elastic Beanstalk as needed

EC2

Amazon Elastic Compute Cloud (EC2) is a web service that gives secure and resizable compute resources in the AWS Cloud. The EC2 service allows you to provision and configure capacity with minimal effort. It provides you with easy control of your computing resources. EC2 reduces the time required to obtain and boot new servers (EC2 instances) to just minutes. This efficiency allows you to scale capacity vertically (up and down, making your server resources bigger or smaller) and horizontally (out and in, adding more capacity in the form of more instances), as your computing requirements change. As you might recall from previous chapters, this property is known as elasticity.
The many benefits of EC2 in AWS include the following:

- EC2 allows for controlled expenditures as your business expands; you pay only for the resources you use as your business grows.
- EC2 provides you with the tools to build failure resilient applications that isolate themselves from common failure scenarios
- EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can commission one, hundreds, or even thousands of server instances simultaneously
- You have complete control of your EC2 instances. You have root access to each one, and you can interact with them as you would any traditional virtual machine.
- You can stop your EC2 instance while retaining the data on your boot partition and then subsequently restart the same instance using web service APIs. Instances can be stopped and started remotely using web service APIs
- You can choose among multiple instance types, operating systems, and software packages. Instance types inside AWS permit the choice of emphasis on CPU, RAM, and/or networking resources
- EC2 integrates with most AWS services, such as Simple Storage Service (S3), Relational Database Service (RDS), and Virtual Private Cloud (VPC). This tight integration allows you to use EC2 for a wide variety of compute scenarios.

- EC2 offers a reliable environment where replacement instances can be rapidly and predictably commissioned. The service runs within Amazon’s proven network infrastructure and data centers. AWS offers as much as 99.95 percent availability for each region.