Global Knowledge®

Expert Reference Series of White Papers

Amazon Web Services: An Overview

Amazon Web Services: An Overview

Rich Morrow, Global Knowledge Instructor, Cloud and Big Data Analyst

Introduction

Getting a clear understanding of what Amazon Web Services (AWS) is and how it can help your business can be a daunting task. The depth and breadth of AWS is significant, comprising more than 48 services in dozens of data centers located at 11 Regions throughout the globe. They offer computing, storage, networking, deployment, management, and a host of supporting services like queues and email. Learning the details of just one of those services can be a multi-day adventure, and just as soon as you've got it down, AWS will introduce a new feature or competing service that requires you to re-evaluate your designs.

It's likely that AWS has more than a few products to help your company work faster, smarter, and more cost-effectively, but many people can be overwhelmed and wonder, "Where do we start?" In this white paper, we hope to provide a good understanding of AWS, how it works, and how your company can get started.

Some History

Like many successful dot-com era startups, Amazon found itself with an enviable problem at the turn of the century: the scale of their business had grown beyond the capacities of any available pre-packaged software solutions. They had to rethink their entire infrastructure from the ground up, and they had to design all their systems to deal with a new set of requirements from their users. Amazon set out to ensure this new infrastructure would provide:

- **High availability:** Via geographical fault tolerance, redundancy, and horizontal linear scale.
- **Auto-scaling:** The ability to dynamically respond to spikes in demand and increase or decrease capacity.
- **Integrated backup and disaster recovery:** Operations like recurring database backups and snapshot restores become as simple as checkboxes and buttons on web forms.
- **"Infinite" scale:** Users should not have to consider ever outgrowing a service. Simple Storage Service (S3) allows users to store an unlimited number of objects and Elastic Cloud Compute (EC2) lets users spin up thousands of virtual servers.
- **Ease of use:** Modular, API-driven services that can be used either independently or with one another.

When Amazon finished building much of their required infrastructure in the early 2000s, they realized they had another enviable problem—extra capacity that they rarely and sporadically used. In 2006, they opened AWS for limited public beta, and although the offerings were only a fraction of what is currently available, the product line became wildly popular. By 2007, they had attracted over 300,000 users, and in 2008, the "beta" moniker was dropped. Subsequent years have seen rapid releases of new services, feature additions to existing services, and prominent customers including Pinterest, NASA, Nasdaq, and Netflix. In addition to the previous requirements, Amazon tacked on the following new features to serve external customers:

- **Flexible pricing:** Nearly every service is provided via consumption-based pricing with no upfront fees. Pay for only what you use, only when you use it.
- **Heightened security:** Stateful firewalls known as security groups provides heightened security. Consolidated user account management is enabled via the Identity and Access Management (IAM) product, and integrated encryption capabilities come with some services like S3. Top-of-the-line hardware based key management systems (HSM) are available for compliance and regulation heavy scenarios.
- **Proactive price cuts:** Because AWS operates with significant scale, they can purchase hardware, software, power, bandwidth, and nearly everything else at much, much lower prices than competitors. Instead of pocketing those savings as profits, AWS has traditionally passed those savings on to customers, reducing prices more than 40 times in the last seven years.

For these reasons and many more, AWS has become an attractive way for businesses of all sizes to deploy and serve the computing and storage needs of their customers. For some companies like Netflix, trusting AWS with all of their computing needs allows them to operate with significantly reduced headcount and a great deal more agility.

Global Architecture

AWS calls their individual data centers Availability Zones or AZs. Each data center has multiple redundant power and bandwidth providers, and AZs are organized into Regions which are collections of physically close data centers that are interconnected via high-bandwidth, low-latency fiber.

Amazon currently has 11 Regions (although one is strictly for U.S. government use), and each region has between one and five AZs. For production use, AWS recommends only using regions with two or more AZs, but they have launched some Regions, like Beijing, in beta mode with only a single AZ.

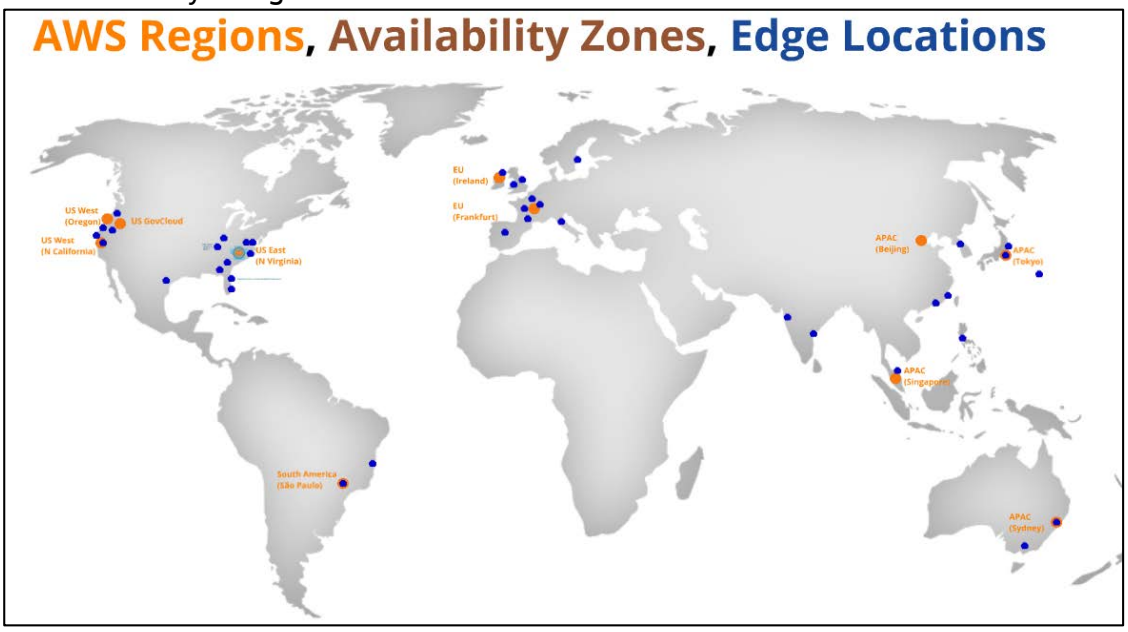

Figure 1: AWS Regions, Availability Zones, Edge Locations

Most customers will decide to deploy in one Region (usually the one closest to their users), and use multiple AZs within that region for high-availability. Many customers leverage another

Region across the country or across the globe to house backups and act as disaster recovery. Some services, like the storage product S3 (discussed later), deploy to a specific region, using multiple servers at multiple AZs to allow stored objects to be both incredibly redundant and fault tolerant.

Other services, like the virtual server product EC2, deploy to only one AZ (you can only launch a virtual server into a single data center). However, by using AutoScaling groups, one could launch a few servers in two or more AZs and then balance across multiple AZs within the region to achieve redundancy and fault tolerance.

These multi-AZ deployments are the preferred way to serve web and mobile traffic within AWS.

Figure 2: An AWS Region with 5 AZs

In addition to regions and AZs, AWS also provides dozens of lightweight caching servers known as Edge Locations which are sometimes referred to as Points of Presence or (PoPs) locations. Currently used only by the content delivery network (CDN) product called CloudFront and the DNS product Route 53—these Edge Locations can cache content and DNS records closer to individual users, thereby reducing latency and decreasing page load times. AWS currently has 53 Edge Locations around the globe, and they're adding about a dozen per year.

AWS is also aggressively adding AZs and Regions, typically introducing one new region per year. End customers get immediate access to all of these new locations as they are added.

The Top Services

Although AWS currently offers nearly 50 services, there are a select few that provide the majority of the utility and benefits. If your company starts using AWS, it's likely that it will be for one or more of the following services:

- Simple Storage Service (S3)
- Elastic Compute Cloud (EC2)
- Elastic Block Store (EBS)
- Relational Database Service (RDS)
- Elastic Load Balancer (ELB)
- AutoScaling
- CloudWatch

Next, we'll explain each of these services and what they do.

Amazon S3

S3 is where many AWS customers begin. It's a simple Write Once, Read Many (WORM) object store that lets users securely, durably, and cost-effectively store assets in a region. "Write once" means that objects cannot be changed after it is written, and "read many" means that multiple copies of the object are made and many users can concurrently retrieve the file quickly by doing parallel reads across many disk heads.

When a user puts an object into S3, multiple copies of that object are made on multiple servers across multiple AZs. This means that the file is highly durable, and data loss is nearly impossible. S3 delivers a target of 11 nines of durability, meaning that over the course of a year, a customer *may* lose one file out of every 100 ***billion*** that they store in the service. This high level of durability would be nearly impossible to achieve on your own.

S3 is especially well-suited for storing and serving static web assets like HTML, CSS, JavaScript, and image files. When a user loads a web page in their browser, the browser makes a remote call every time it hits a CSS, JavaScript, and image tag. If S3 is used to serve those assets, it can typically deliver the assets much faster than your company website, which now runs much faster since it's only providing a much smaller amount of dynamic content. To speed the delivery of content, AWS offers the CloudFront content delivery network (CDN) service. Due to the bandwidth pricing being lower than standard S3 bandwidth, customers using CloudFront to deliver S3 backed data both better serve their customers with reduced latency, and better serve themselves with a lower bill from AWS.

Amazon EC2 and Amazon EBS

EC2 is likely to be at the heart of your AWS deployment, and it's estimated that most customers spend north of 70 percent of their AWS bill on EC2 instances alone. At its essence, EC2 is simply a virtual server that contains a hardware footprint known as instance type and a software footprint known as an Amazon Machine Image (AMI). There are currently 53 instance types available, with varying amounts of CPU, RAM, and local storage, some even offering fast SSD disks.

You'll typically start out with an Amazon-provided vanilla basic that is OS only (both Linux and Windows supported), and then install and configure your application into that AMI. Amazon gives you the ability to burn a custom AMI (also known as Golden Image) of your software,

and then for instance, you can use that image to deploy other servers of that type – picture a farm of web servers, all configured exactly the same (because they come from the same AMI), all sitting behind a load balancer.

Amazon also offers a Marketplace which contains thousands of vendor-provided AMIs that you can launch. The Marketplace can be a great way to evaluate a particular piece of software before deciding to purchase.

Many companies use EC2 with AutoScaling (discussed later) to allow their web apps to automatically scale up to and serve variable peak demand, while automatically scaling down to very small numbers of servers, which saves money, during low-usage times like nights and weekends.

Although most instance types come with some local disks, those disks are volatile and tied directly to the instance's lifetime. If, like most AWS customers, you need a persistent disk that exists independent of an instance lifetime, you'll soon be using Elastic Block Store (EBS).

EBS functions nearly the same as a physical 3.5-inch disk that you would install directly into a server. You first provision the size of the volume, and then attach one or more volumes to a single EC2 instance, format, and use them for EC2-accessible storage. EBS is very fast, and a user could even stripe multiple volumes together for even faster disk reads and writes. EBS is a great place to store database data, but if you're looking for a database, AWS has an even better option: the Relational Database Service.

Amazon RDS

AWS' Relational Database Service (RDS) is a fully managed, administration-free install of MySQL, Oracle, SQLServer, or PostgreSQL. Like many of AWS' managed services, the user gives up some control (like fine-tuning performance) in exchange for having an easy-to-use, scalable database.

With RDS, a user provisions a server type similar to the EC2 instance types, and receives a client interface URL against which they can interact with their database. The exposed interface can then be used to create, insert into, and query database tables.

Like EC2, the user can resize their instance up or down, and can also attach security groups which are stateful, easy-to-configure firewalls that limit access to the resource.

With RDS, setting up weekly or daily database backups is as simple as using a checkbox in the web GUI. For MySQL particularly, RDS allows quick, easy setup of slaves and read replicas, allowing your database layer to be fault tolerant and scale up with read volume.

AutoScaling and ELB/CloudWatch

With the popularity of smartphones and tablets, more users are increasingly creating greater demands on websites and the back-end servers. There's also arguably more variability in the minute-to-minute load on a website—with bursts of traffic from email drops and press releases becoming trickles of traffic later on that evening when users are in bed.

Non-cloud websites would've dealt with that traffic by over-provisioning and under-utilizing large farms of servers configured for the expected peak demand, plus maybe 10 percent. The problem with these architectures was not only the waste associated with powering and cooling say, for example, 100 servers at night when you only need two, but also with an even worse scenario of not keeping ahead of demand (having 100 servers when you really need 200).

AWS' AutoScaling solves this problem by dynamically growing and shrinking the number of servers you need based on real-time demand. AutoScaling typically involves three services: AutoScaling groups, CloudWatch alarms, and ELBs (collectively referred to by AWS as the "triangle services").

You simply pre-configure your AutoScaling group to absolute maximums, minimums, EC2 instance types, and AMI, and then set up several CloudWatch alarms to monitor metrics like load balancing requests, and dynamically add and remove instances from the group and from the load-balancing rotation.

As your ELB sees request volume increase and decrease, it fires the pre-configured CloudWatch alarms, which then immediately add capacity to, or take capacity from, your web tier. No more underutilization, and most importantly, no more serving 404 "page not found" messages to users because you don't have enough servers to handle their requests.

Interfaces

To interact with all of these various services AWS provides three interfaces, each with various benefits:

- AWS Web GUI
- API (and the command line tools that wrap them).
- Language and platform specific software development kits (SDKs).

The Web GUI is probably how most users begin using AWS, and it's the first thing you see when you create an account online. It exposes most of the features of the popular services that AWS provides, and has great point-and-click support for EC2, ELB, EBS, RDS, CloudWatch and S3, and even AutoScaling.

Although the Web GUI works perfectly well for most beginner users, in order to access the full power of AWS, one might eventually need to use the REST APIs. Only the REST APIs offer 100 percent access to all the services and features that AWS provides.

Almost all the services have command line interface (CLI) tools that wrap the REST APIs, making them even simpler to use. For example, by using the EC2 command line tools, one could create two servers simply by making the following call:

```
ec2-run-instances ami-1a2b3c4d -n 2 -k my-key-pair –availability-zone us-east-1a
```

In 2014, AWS moved the CLI tools to a "unified" download where customers can simply download all of the available CLI tools in one single action (previously, users had to download each independently). The download is available from AWS at http://aws.amazon.com/cli.

The last method of access to AWS is the language and platform-specific SDKs. Supported languages and platforms vary from service to service, but commonly, PHP, Java, Python, and most other popular web languages like Ruby and Node.js are supported, in addition to Android and iOS platforms. Using the SDKs, one could write custom web or mobile apps that interact directly with AWS using the language your developers use daily.

Learn More

To learn more about how you can improve productivity, enhance efficiency, and sharpen your competitive edge, Global Knowledge suggests the following courses:

- AWS Business Essentials
- AWS Technical Essentials
- Architecting on AWS
- Developing on AWS
- Systems Operations on AWS

Visit **www.globalknowledge.com** or call **1-800-COURSES (1-800-268-7737)** to speak with a Global Knowledge training advisor.

About the Author

Rich is a full-stack generalist with particular depth in and love for cloud and big data technologies. When he's not flying around the country for Global Knowledge providing training on AWS and Hadoop to Fortune 500 companies, he's probably writing or speaking about cloud, big data, mobile development, or IoT topics for VentureBeat or Gigaom.

Additional Reading

Hopefully, we've been able to clarify some of the inner workings of AWS and share its value with you. If you're inspired to learn more about the services, the links below are helpful resources.

AWS Regions, AZs and Edge Locations Prezi
AWS Documentation
EC2 Documentation
AutoScaling Documentation
ELB Documentation
S3 Documentation
RDS Documentation
CloudWatch Documentation
AWS Free Tier
Getting Started with AWS Guide
Global Knowledge AWS Training