Overview of Amazon Web Services
March 2013

Jinesh Varia/Sajee Mathew

Table of Contents

Introduction

Managing the unique and groundbreaking changes in both technology and business over the past decade has created an ongoing IT infrastructure challenge for many senior technology executives. Indeed, over the past ten years, the typical business application architecture has evolved from a desktop-centric installation, then to client/server solutions, and now to loosely coupled web services and service-oriented architectures (SOA). Each evolutionary step has built on the previous one while adding new challenges, dimensions, and opportunities for IT departments and their business partners.

Recently, virtualization has become a widely accepted way to reduce operating costs and increase the reliability of enterprise IT. In addition, grid computing makes a completely new class of analytics, data crunching, and business intelligence tasks possible that were previously cost and time prohibitive. Along with these technology changes, the speed of innovation and unprecedented acceleration in the introduction of new products has fundamentally changed the way markets work. Along with the wide acceptance of software as a service (SaaS) offerings, these changes have paved the way for the latest IT infrastructure challenge: cloud computing.

What is "Cloud Computing"?

Cloud computing has become one of the most discussed IT paradigms of recent years. It builds on many of the advances in the IT industry over the past decade and presents significant opportunities for organizations to shorten time to market and reduce costs. With cloud computing, organizations can consume shared computing and storage resources rather than building, operating, and improving infrastructure on their own. The speed of change in markets creates significant pressure on the enterprise IT infrastructure to adapt and deliver. Cloud computing provides fresh solutions to address these changes. As defined by Gartner[1], "Cloud computing is a style of computing where scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies."

Cloud computing enables organizations to obtain a flexible, secure, and cost-effective IT infrastructure, in much the same way that national electric grids enable homes and organizations to plug into a centrally managed, efficient, and cost-effective energy source. When freed from creating their own electricity, organizations were able to focus on the core competencies of their business and the needs of their customers. Likewise, cloud computing liberates organizations from devoting precious people and budget to activities that don't directly contribute to the bottom line while still obtaining IT infrastructure capabilities.

These capabilities include compute power, storage, databases, messaging, and other building block services that run business applications. When coupled with a utility-style pricing and business model, cloud computing promises to deliver an enterprise-grade IT infrastructure in a reliable, timely, and cost-effective manner.

---

[1] Gartner IT Glossary, http://www.gartner.com/it-glossary/cloud-computing.

To understand the impact and promise of cloud computing, one may first analyze the significance of and lessons learned from business outsourcing. Focusing on a core competency and then shifting the peripheral business tasks to other organizations is a proven business strategy. Today, organizations outsource business functions such as logistics, human resources (HR), payroll, and facilities. Many organizations have taken advantage of IT outsourcing as a way to move some capabilities out of their internal organization altogether.

Superficially, at least, cloud computing resembles the trend of business outsourcing because it provides the benefits of leveraging the expertise of others and being cost efficient. However, cloud computing also provides flexibility, scalability, elasticity, and reliability. These additional benefits are why enterprise organizations see cloud computing as a powerful next step in their IT infrastructure evolution.

Amazon and Cloud Computing

Amazon has a long history of using a decentralized IT infrastructure. This arrangement enabled our development teams to access compute and storage resources on demand, and it has increased overall productivity and agility. By 2005, Amazon had spent over a decade and millions of dollars building and managing the large-scale, reliable, and efficient IT infrastructure that powered one of the world's largest online retail platforms. Amazon launched [Amazon Web Services](#) (AWS) so that other organizations could benefit from Amazon's experience and investment in running a large-scale distributed, transactional IT infrastructure. AWS has been operating since 2006, and today serves hundreds of thousands of customers worldwide. Today Amazon.com runs a global web platform serving millions of customers and managing billions of dollars' worth of commerce every year.

Using AWS, you can requisition compute power, storage, and other services in minutes and have the flexibility to choose the development platform or programming model that makes the most sense for the problems they're trying to solve. You pay only for what you use, with no up-front expenses or long-term commitments, making AWS a cost-effective way to deliver applications.

Here are some of examples of how organizations, from research firms to large enterprises, use AWS today:

- A large enterprise quickly and economically deploys new internal applications, such as HR solutions, payroll applications, inventory management solutions, and online training to its distributed workforce.

- An e-commerce website accommodates sudden demand for a "hot" product caused by viral buzz from Facebook and Twitter without having to upgrade its infrastructure.

- A pharmaceutical research firm executes large-scale simulations using computing power provided by AWS.

- Media companies serve unlimited video, music, and other media to their worldwide customer base.

The Differences that Distinguish AWS

AWS is readily distinguished from other vendors in the traditional IT computing landscape because it is:

- **Flexible**. AWS enables organizations to use the programming models, operating systems, databases, and architectures with which they are already familiar. In addition, this flexibility helps organizations mix and match architectures in order to serve their diverse business needs.

- **Cost-effective**. With AWS, organizations pay only for what they use, without up-front or long-term commitments.

- **Scalable and elastic**. Organizations can quickly add and subtract AWS resources to their applications in order to meet customer demand and manage costs.

- **Secure**. In order to provide end-to-end security and end-to-end privacy, AWS builds services in accordance with security best practices, provides the appropriate security features in those services, and documents how to use those features.

- **Experienced**. When using AWS, organizations can leverage Amazon's more than fifteen years of experience delivering large-scale, global infrastructure in a reliable, secure fashion.

Flexible

The first key difference between AWS and other IT models is *flexibility*. Using traditional models to deliver IT solutions often requires large investments in new architectures, programming languages, and operating systems. Although these investments are valuable, the time that it takes to adapt to new technologies can also slow down your business and prevent you from quickly responding to changing markets and opportunities. When the opportunity to innovate arises, you want to be able to move quickly and not always have to support legacy infrastructure and applications or deal with protracted procurement processes.

In contrast, the flexibility of AWS allows you to keep the programming models, languages, and operating systems that you are already using or choose others that are better suited for their project. You don't have to learn new skills. Flexibility means that migrating legacy applications to the cloud is easy and cost-effective. Instead of re-writing applications, you can easily move them to the AWS cloud and tap into advanced computing capabilities.

Building applications on AWS is very much like building applications using existing hardware resources. Since AWS provides a flexible, virtual IT infrastructure, you can use the services together as a platform or separately for specific needs. AWS run almost anything—from full web applications to batch processing to offsite data back-ups.

In addition, you can move existing SOA-based solutions to the cloud by migrating discrete components of legacy applications. Typically, these components benefit from high availability and scalability, or they are self-contained applications with few internal dependencies. Larger organizations typically run in a hybrid mode where pieces of the application run in their data center and other portions run in the cloud. Once these organizations gain experience with the cloud, they begin transitioning more of their projects to the cloud, and they begin to appreciate many of the benefits outlined in this document. Ultimately, many organizations see the unique advantages of the cloud and AWS and make it a permanent part of their IT mix.

Finally, AWS provides you flexibility when provisioning new services. Instead of the weeks and months it takes to plan, budget, procure, set up, deploy, operate, and hire for a new project, you can simply sign up for AWS and immediately begin deployment on the cloud the equivalent of 1, 10, 100, or 1,000 servers. Whether you want to prototype an application or host a production solution, AWS makes it simple for you to get started and be productive. Many customers find the flexibility of AWS to be a great asset in improving time to market and overall organizational productivity.

Cost-Effective

Cost is one of the most complex elements of delivering contemporary IT solutions. It seems that for every advance that will save money, there is often a commensurate investment needed to realize that savings. For example, developing and deploying an e-commerce application can be a low-cost effort, but a successful deployment can increase the need for hardware and bandwidth. Furthermore, owning and operating your own infrastructure can incur considerable costs, including power, cooling, real estate, and staff.

In contrast, the cloud provides an on-demand IT infrastructure that lets you consume only the amount of resources that you actually need. You are not limited to a set amount of storage, bandwidth, or computing resources. It is often difficult to predict requirements for these resources. As a result, you might provision too few resources, which has an impact on customer satisfaction, or you might provide too many resources and miss an opportunity to maximize return on investment (ROI) through full utilization.

The cloud provides the flexibility to strike the right balance. AWS requires no up-front investment, long-term commitment, or minimum spend. You can get started through a completely self-service experience online, scale up and down as needed, and terminate your relationship with AWS at any time. You can access new resources almost instantly. The ability to respond quickly to changes, no matter how large or small, means that you can take on new opportunities and meet business challenges that could drive revenue and reduce costs. If you want to consult with AWS for deeper technical discussions, our sales and solutions architecture teams are available.

For more information on how you can save money with AWS, consult the AWS Economics Center at aws.amazon.com/economics.

Scalable and Elastic

In the traditional IT organization, scalability and elasticity were often equated with investment and infrastructure. In the cloud, scalability and elasticity provide opportunity for savings and improved ROI. AWS uses the term elastic to describe the ability to scale computing resources up and down easily, with minimal friction. Elasticity helps you avoid provisioning resources up front for projects with variable consumption rates or short lifetimes. Instead of acquiring hardware, setting it up, and maintaining it in order to allocate resources to your applications, you use AWS to allocate resources using simple API calls.

Imagine what would happen to a traditional IT shop if traffic to an application doubled or tripled in a short period. For example, during benefits open enrollment periods, many corporate users generate significant traffic to internal applications. You need to be confident that your existing infrastructure can handle a spike in traffic, and that the spike will not interfere with normal business operations. Elastic Load Balancing and Auto Scaling can automatically scale your AWS cloud-based resources up to meet unexpected demand, and then scale those resources down as demand decreases.

The AWS cloud is also a useful resource for implementing short-term jobs, mission-critical jobs, and jobs repeated at regular intervals. For example, when a pharmaceutical company needs to run drug simulations (a short-term job), it can use AWS to spin up resources in the cloud, and then shut them down when it no longer needs additional resources. When an enterprise has to quickly deal with the effects of natural disaster on its data center (a mission-critical job), it can use AWS to tap into new storage and computing resources to accommodate demand. Furthermore, AWS can preserve computing resources and reduce costs for regularly repeated tasks, such as month-end payroll or invoice processing.

For more information on cloud architectures, consult the AWS Architecture Center at aws.amazon.com/architecture.

Secure

AWS delivers a scalable cloud-computing platform that provides customers with end-to-end security and end-to-end privacy. AWS builds security into its services in accordance with security best practices, and documents how to use the security features. It is important that you leverage AWS security features and best practices to design an appropriately secure application environment.

Ensuring the confidentiality, integrity, and availability of your data is of the utmost importance to AWS, as is maintaining your trust and confidence. AWS takes the following approaches to secure the cloud infrastructure:

- **Certifications and accreditations.** AWS has in the past successfully completed multiple SAS70 Type II audits, and now publishes a Service Organization Controls 1 (SOC 1) report, published under both the SSAE 16 and the ISAE 3402 professional standards. In addition, AWS has achieved ISO 27001 certification, and has been successfully validated as a Level 1 service provider under the Payment Card Industry (PCI) Data Security Standard (DSS). In the realm of public sector certifications, AWS has received authorization from the U.S. General Services Administration to operate at the FISMA Moderate level, and is also the platform for applications with Authorities to Operate (ATOs) under the Defense Information Assurance Certification and Accreditation Program

(DIACAP). We will continue to obtain the appropriate security certifications and conduct audits to demonstrate the security of our infrastructure and services. AWS will continue to obtain the appropriate security certifications and accreditations to demonstrate the security of our infrastructure and services.

- **Physical security**. Amazon has many years of experience designing, constructing, and operating large-scale data centers. The AWS infrastructure is located in Amazon-controlled data centers throughout the world. Knowledge of the location of the data centers is limited to those within Amazon who have a legitimate business reasons for this information. The data centers are physically secured in a variety of ways to prevent unauthorized access.

- **Secure services**. Each service in the AWS cloud is architected to be secure. The services contain a number of capabilities that restrict unauthorized access or usage without sacrificing the flexibility that customers demand.

- **Data privacy**. You can encrypt personal and business data in the AWS cloud, and publish backup and redundancy procedures for services so that your customers can protect their data and keep their applications running.

For more information on security policies and procedures for AWS, consult the AWS Security Center at aws.amazon.com/security.

Experienced

AWS provides a low-friction path to cloud computing by design. Nevertheless, as with any IT project, the move to the AWS cloud should be done thoughtfully. You should hold your cloud-computing partner to the same high standards that you would expect of any hardware or software vendor. The trust that you place in your cloud-computing vendor will be critical as your organization grows and your customers continue to expect the best experience.

The AWS cloud provides levels of scale, security, reliability, and privacy that are often cost-prohibitive for many organizations to meet or exceed. AWS has built an infrastructure based on lessons learned from over sixteen years' experience managing the multi-billion dollar Amazon.com business. AWS customers benefit as Amazon continues to hone its infrastructure management skills and capabilities. Today Amazon.com runs a global web platform serving millions of customers and managing billions of dollars' worth of commerce every year. AWS has been operating since 2006, and today serves hundreds of thousands of customers worldwide.

Moreover, AWS has a demonstrated track record of listening to its customers and delivering highly innovative new features at a rapid pace. These new releases have the same high standards of security and reliability that are demonstrated in all existing AWS infrastructure services.

In addition to new services, AWS constantly hones its operational expertise to ensure ongoing dependability, and we continue to incorporate both industry best practices and proprietary advances into the cloud infrastructure. Choosing AWS as a cloud-computing provider allows you to take advantage of these ongoing investments.

Amazon Web Services Cloud Platform

AWS is a comprehensive cloud services platform that offers compute power, storage, content delivery, and other functionality that organizations can use to deploy applications and services cost-effectively—with flexibility, scalability, and reliability. AWS self-service means that you can proactively address your internal plans and react to external demands when you choose.

Cross-Service

Support
Phone & email fast-response 24X7 Support

Marketplace
Bull and Sell Software and Apps

Management Console
UI to manage AWS services

SDKs, IDE kits and CLIs
Develop , integrate and manage services

Deployment & Management

CloudFormation
Templated AWS Resource Creation

CloudWatch
Resource and Application Monitoring

Data Pipeline
Orchestration for Data-Driven Workflows

Elastic Beanstalk
AWS Application Container

IAM
Secure AWS Access Control

OpsWorks
DevOps Application Management Service

CloudHSM
Hardware-based key storage for compliance

Database

DynamoDB
Predictable and Scalable NoSQL Data Store

ElastiCache
In-Memory Cache

RDS
Managed Relational Database

Redshift
Managed Petabyte-Scale Data Warehouse

Storage and Content Delivery

S3
Scalable Storage in the Cloud

EBS
Networked Attached Block Device

CloudFront
Global Content Delivery Network

Glacier
Archive Storage in the Cloud

Storage Gateway
Integrates On-Premises IT with Cloud Storage

Import Export
Ship Large Datasets

Compute & Networking

EC2
Virtual Servers in the Cloud

VPC
Virtual Secure Network

ELB
Load balancing Service

Auto Scaling
Automatically scale up and down

Elastic MapReduce
Managed Hadoop Framework

Direct Connect
Dedicated Network Connection to AWS

Route 53
Scalable Domain Name System

App Services

CloudSearch
Managed Search Service

Elastic Transcoder
Easy-to-use Scalable Media Transcoding

SES
Email Sending Service

SNS
Push Notification Service

SQS
Message Queue Service

SWF
Workflow Service for Coordinating App Components

**AWS Global Physical Infrastructure
(Geographical Regions, Availability Zones, Edge Locations)**

Compute

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers and system administrators.

Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment.

Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers and system administrators the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Amazon Elastic MapReduce (Amazon EMR)

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It uses a hosted Hadoop framework running on the web-scale infrastructure of Amazon EC2 and Amazon Simple Storage Service (Amazon S3).

Using Amazon EMR, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics research. Amazon EMR lets you focus on crunching or analyzing your data without having to worry about time-consuming set-up, management, or tuning of Hadoop clusters or the compute capacity upon which they sit.

Auto Scaling

Auto Scaling allows you to scale your Amazon EC2 capacity up or down automatically according to conditions you define. With Auto Scaling, you can ensure that the number of Amazon EC2 instances you're using increases seamlessly during demand spikes to maintain performance, and decreases automatically during demand lulls to minimize costs. Auto Scaling is particularly well suited for applications that experience hourly, daily, or weekly variability in usage. Auto Scaling is enabled by Amazon CloudWatch and available at no additional charge beyond Amazon CloudWatch fees.

Elastic Load Balancing

Elastic Load Balancing automatically distributes incoming application traffic across multiple Amazon EC2 instances. It enables you to achieve even greater fault tolerance in your applications, seamlessly providing the amount of load balancing capacity needed in response to incoming application traffic. Elastic Load Balancing detects unhealthy instances and automatically reroutes traffic to healthy instances until the unhealthy instances have been restored. Customers can enable Elastic Load Balancing within a single Availability Zone or across multiple zones for even more consistent application performance.

Networking

Amazon Virtual Private Cloud (Amazon VPC)

Amazon Virtual Private Cloud (Amazon VPC) lets you provision a private, isolated section of the AWS cloud where you can launch AWS resources in a virtual network that you define. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own data center. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.

You can easily customize the network configuration for your Amazon VPC. For example, you can create a public-facing subnet for your webservers that has access to the Internet, and place your backend systems such as databases or application servers in a private-facing subnet with no Internet access. You can leverage multiple layers of security (including security groups and network access control lists) to help control access to Amazon EC2 instances in each subnet.

Additionally, you can create a hardware virtual private network (VPN) connection between your corporate data center and your VPC and leverage the AWS cloud as an extension of your corporate data center.

Amazon Route 53

Amazon Route 53 is a highly available and scalable Domain Name System (DNS) web service. It is designed to give developers and businesses an extremely reliable and cost-effective way to route end users to Internet applications by translating human readable names, such as www.example.com, into the numeric IP addresses, such as 192.0.2.1, that computers use to connect to each other. Route 53 effectively connects user requests to infrastructure running in AWS, such as an EC2 instance, an elastic load balancer, or an Amazon S3 bucket. Route 53 can also be used to route users to infrastructure outside of AWS.

Amazon Route 53 is designed to be fast, easy to use, and cost effective. It answers DNS queries with low latency by using a global network of DNS servers. Queries for your domain are automatically routed to the nearest DNS server, and thus are answered with the best possible performance. With Route 53, you can create and manage your public DNS records with the AWS Management Console or with an easy-to-use API. It's also integrated with other Amazon Web Services. For instance, by using the AWS Identity and Access Management (IAM) service with Route 53, you can control who in your organization can make changes to your DNS records. Like other Amazon Web Services, there are no long-term contracts or minimum usage requirements for using Route 53—you pay only for managing domains through the service and the number of queries that the service answers.

AWS Direct Connect

AWS Direct Connect makes it easy to establish a dedicated network connection from your premises to AWS. Using AWS Direct Connect, you can establish private connectivity between AWS and your data center, office, or co-location environment, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

AWS Direct Connect lets you establish a dedicated network connection between your network and one of the AWS Direct Connect locations. Using industry standard 802.1Q virtual LANS (VLANs), this dedicated connection can be partitioned into multiple logical connections. This allows you to use the same connection to access public resources such as objects stored in Amazon S3 using public IP address space, and private resources such as Amazon EC2 instances running within an Amazon VPC using private IP space, while maintaining network separation between the public and private environments. Logical connections can be reconfigured at any time to meet your changing needs.

Storage

Amazon Simple Storage Service (Amazon S3)

Amazon S3 is storage for the Internet. It is designed to make web-scale computing easier for developers.

Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. The container for objects stored in Amazon S3 is called an Amazon S3 bucket. Amazon S3 gives any developer access to the same highly scalable, reliable, secure, fast, inexpensive infrastructure that Amazon uses to run its own global network of websites. The service aims to maximize benefits of scale and to pass those benefits on to developers.

Amazon Glacier

Amazon Glacier is an extremely low-cost storage service that provides secure and durable storage for data archiving and backup. In order to keep costs low, Amazon Glacier is optimized for data that is infrequently accessed and for which retrieval times of several hours are suitable. With Amazon Glacier, customers can reliably store large or small amounts of data for as little as $0.01 per gigabyte per month, a significant savings compared to on-premises solutions.

Companies typically over-pay for data archiving. First, they're forced to make an expensive upfront payment for their archiving solution (which does not include the ongoing cost for operational expenses such as power, facilities, staffing, and maintenance). Second, since companies have to guess what their capacity requirements will be, they understandably over-provision to make sure they have enough capacity for data redundancy and unexpected growth. This set of circumstances results in under-utilized capacity and wasted money. With Amazon Glacier, you pay only for what you use. Amazon Glacier changes the game for data archiving and backup because you pay nothing up front, pay a very low price for storage, and can scale your usage up or down as needed, while AWS handles all of the operational heavy lifting required to do data retention well. It only takes a few clicks in the AWS Management Console to set up Amazon Glacier, and then you can upload any amount of data you choose.

Amazon Elastic Block Storage (EBS)

Amazon Elastic Block Store (EBS) provides block level storage volumes for use with Amazon EC2 instances. Amazon EBS volumes are network-attached, and persist independently from the life of an instance. Amazon EBS provides highly available, highly reliable, predictable storage volumes that can be attached to a running Amazon EC2 instance and exposed as a device within the instance. Amazon EBS is particularly suited for applications that require a database, file system, or access to raw block level storage.

AWS Storage Gateway

AWS Storage Gateway is a service connecting an on-premises software appliance with cloud-based storage to provide seamless and secure integration between an organization's on-premises IT environment and AWS's storage infrastructure. The service enables you to securely upload data to the AWS cloud for cost-effective backup and rapid disaster recovery. AWS Storage Gateway supports industry-standard storage protocols that work with your existing applications. It provides low-latency performance by maintaining data on your on-premises storage hardware while

asynchronously uploading this data to AWS, where it is encrypted and securely stored in Amazon Simple Storage Service (Amazon S3).

Using AWS Storage Gateway, you can back up point-in-time snapshots of your on-premises application data to Amazon S3 for future recovery. In the event you need replacement capacity for disaster recovery purposes, or if you want to leverage Amazon EC2's on-demand compute capacity for additional capacity during peak periods, for new projects, or as a more cost-effective way to run your normal workloads, you can use AWS Storage Gateway to mirror your on-premises data to Amazon EC2 instances.

AWS Import/Export

AWS Import/Export accelerates moving large amounts of data into and out of AWS using portable storage devices for transport. AWS transfers your data directly onto and off of storage devices using Amazon's high-speed internal network and bypassing the Internet. For significant data sets, AWS Import/Export is often faster than Internet transfer and more cost effective than upgrading your connectivity.

AWS Import/Export supports importing and exporting data into and out of Amazon S3 buckets in the US Standard, US West (Oregon), US West (Northern California), EU (Ireland), and Asia Pacific (Singapore) regions. The service also supports importing data into Amazon Elastic Block Store (Amazon EBS) snapshots in the US East (N. Virginia), US West (Oregon), and US West (Northern California) regions.

Content Delivery

Amazon CloudFront

Amazon CloudFront is a web service for content delivery. It integrates with other Amazon Web Services to give developers and businesses an easy way to distribute content to end users with low latency, high data transfer speeds, and no commitments.

Amazon CloudFront can be used to deliver your entire website, including dynamic, static and streaming content using a global network of edge locations. Requests for objects are automatically routed to the nearest edge location, so content is delivered with the best possible performance. Amazon CloudFront is optimized to work with other Amazon Web Services, like Amazon S3 and Amazon EC2. Amazon CloudFront also works seamlessly with any origin server, which stores the original, definitive versions of your files. Like other Amazon Web Services, there are no contracts or monthly commitments for using Amazon CloudFront—you pay only for as much or as little content as you actually deliver through the service.

Database

Amazon Relational Database Service (Amazon RDS)

Amazon Relational Database Service (Amazon RDS) is a web service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while managing time-consuming database administration tasks, freeing you up to focus on your applications and business.

Amazon RDS gives you access to the capabilities of a familiar MySQL, Oracle or SQL Server database. This means that the code, applications, and tools you already use today with your existing databases can be used with Amazon RDS. Amazon RDS automatically patches the database software and backs up your database, storing the backups for a retention period that you define and enabling point-in-time recovery. You benefit from the flexibility of being able to scale the compute resources or storage capacity associated with your relational database instance by using a single API call. In addition, Amazon RDS makes it easy to use replication to enhance availability and reliability for production databases and to scale out beyond the capacity of a single database deployment for read-heavy database workloads.

Amazon DynamoDB

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. With a few clicks in the AWS Management Console, customers can launch a new Amazon DynamoDB database table, scale up or down their request capacity for the table without downtime or performance degradation, and gain visibility into resource utilization and performance metrics. Amazon DynamoDB enables customers to offload the administrative burdens of operating and scaling distributed databases to AWS, so customers don't have to worry about hardware provisioning, setup and configuration, replication, software patching, or cluster scaling.

Amazon DynamoDB is designed to address the core problems of database management, performance, scalability, and reliability. Developers can create a database table that can store and retrieve any amount of data, and serve any level of request traffic. DynamoDB automatically spreads the data and traffic for the table over a sufficient number of servers to handle the request capacity specified by the customer and the amount of data stored, while maintaining consistent, fast performance. All data items are stored on solid state drives (SSDs) and are automatically replicated across multiple Availability Zones in a Region to provide built-in high availability and data durability.

Amazon DynamoDB enables customers to offload the administrative burden of operating and scaling a highly available, distributed database cluster while only paying a low variable price for the resources they consume.

Amazon ElastiCache

Amazon ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud. The service improves the performance of web applications by allowing you to retrieve information from a fast, managed, in-memory caching system, instead of relying entirely on slower disk-based databases. Amazon ElastiCache is protocol-compliant with Memcached (a widely adopted memory object caching system) so code, applications, and popular tools that you use today with existing Memcached environments will work seamlessly with the service.

Amazon ElastiCache simplifies and offloads the management, monitoring, and operation of in-memory cache environments, enabling you to focus on the differentiating parts of your applications. Using Amazon ElastiCache, you can add an in-memory cache to your application architecture in a matter of minutes. With a few clicks of the AWS Management Console, you can launch a Cache Cluster consisting of a collection of Cache Nodes, each node running Memcached software. You can then scale the amount of memory associated with your Cache Cluster in minutes by adding or deleting Cache Nodes to meet the demands of your changing workload. In addition, Amazon ElastiCache automatically detects and replaces failed Cache Nodes, providing a resilient system that mitigates the risk of overloaded databases, which slow website and application load times. Through integration with Amazon CloudWatch, Amazon ElastiCache provides enhanced visibility into key performance metrics associated with your Cache Nodes.

Amazon Redshift

Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse service that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools. It is optimized for datasets ranging from a few hundred gigabytes to a petabyte or more and costs less than $1,000 per terabyte per year, a tenth the cost of most traditional data warehousing solutions.

Deployment & Management

AWS Identity and Access Management (IAM)

You can use AWS Identity and Access Management (IAM) to securely control access to AWS services and resources for your users. IAM enables you to create and manage users in AWS, and it also enables you to grant access to AWS resources for users managed outside of AWS in your corporate directory. IAM offers greater security, flexibility, and control when using AWS.

IAM enables identity federation between your corporate directory and AWS services. This enables you to use your existing corporate identities to grant secure and direct access to AWS resources, such as S3 buckets, without creating a new AWS identity for those users.

Amazon CloudWatch

Amazon CloudWatch provides monitoring for AWS cloud resources and the applications customers run on AWS. Developers and system administrators can use it to collect and track metrics, gain insight, and react immediately to keep their applications and businesses running smoothly. Amazon CloudWatch monitors AWS resources such as Amazon EC2 and Amazon RDS DB Instances, and can also monitor custom metrics generated by a customer's applications and services. With Amazon CloudWatch, you gain system-wide visibility into resource utilization, application performance, and operational health.

Amazon CloudWatch provides a reliable, scalable, and flexible monitoring solution that you can start using within minutes. You no longer need to set up, manage, or scale your own monitoring systems and infrastructure. Using Amazon CloudWatch, you can easily monitor as much or as little metric data as you need. Amazon CloudWatch lets you

programmatically retrieve your monitoring data, view graphs, and set alarms to help you troubleshoot, spot trends, and take automated action based on the state of your cloud environment.

AWS Elastic Beanstalk

AWS Elastic Beanstalk is an easy way for you to quickly deploy and manage applications in the AWS cloud. You simply upload your application, and Elastic Beanstalk automatically handles the deployment details of capacity provisioning, load balancing, auto scaling, and application health monitoring. At the same time, with Elastic Beanstalk, you retain full control over the AWS resources powering your application and can access the underlying resources at any time. Elastic Beanstalk leverages AWS services such as Amazon EC2, Amazon S3, Amazon Simple Notification Service (Amazon SNS), Elastic Load Balancing, and Auto Scaling to deliver the same highly reliable, scalable, and cost-effective infrastructure that hundreds of thousands of businesses depend on today. Elastic Beanstalk is easy to get started with and impossible to outgrow.

Most existing application containers or platform-as-a-service solutions, while reducing the amount of programming required, significantly diminish developers' flexibility and control. Developers are forced to live with all the decisions pre-determined by the vendor and have little to no opportunity to take back control over various parts of their application's infrastructure. However, with Elastic Beanstalk, you retain full control over the AWS resources powering your application. If you decide you want to take over some (or all) of the elements of their infrastructure, you can do so seamlessly by using Elastic Beanstalk's management capabilities.

To ensure easy portability of your application, Elastic Beanstalk is built using familiar software stacks such as the Apache HTTP Server for PHP and Python, Passenger for Ruby, Apache Tomcat for Java, and Microsoft IIS for .NET web applications. There is no additional charge for Elastic Beanstalk. You pay only for the AWS resources you need to store and run your applications.

AWS CloudFormation

AWS CloudFormation gives developers and systems administrators an easy way to create and manage a collection of related AWS resources, provisioning and updating them in an orderly and predictable fashion.

You can use AWS CloudFormation's sample templates or create your own templates to describe the AWS resources, and any associated dependencies or runtime parameters, required to run your application. You don't need to figure out the order in which AWS services need to be provisioned or the subtleties of how to make those dependencies work. AWS CloudFormation takes care of this for you. Once deployed, you can modify and update the AWS resources in a controlled and predictable way. This allows you to version control your AWS infrastructure in the same way as you version control your software.

You can deploy and update a template and its associated collection of resources (called a stack) using the AWS Management Console, AWS CloudFormation command line tools, or CloudFormation API. AWS CloudFormation is available at no additional charge, and you pay only for the AWS resources needed to run your applications.

AWS Data Pipeline

AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services as well as on-premise data sources at specified intervals. With AWS Data Pipeline, you can regularly access your data where it's stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon Elastic MapReduce (EMR). AWS Data Pipeline helps you easily create complex data processing workloads that are fault tolerant, repeatable, and highly available. You don't have to worry about ensuring resource availability, managing inter-task dependencies, retrying transient failures or timeouts in individual tasks, or creating a failure notification system. AWS Data Pipeline also allows you to move and process data that was previously locked up in on-premise data silos.

AWS OpsWorks

AWS OpsWorks is a DevOps solution for managing applications of any scale or complexity on the AWS cloud. AWS OpsWorks features an integrated experience for managing the complete application lifecycle, including resource provisioning, configuration management, application deployment, software updates, monitoring, and access control.

AWS OpsWorks lets you model and visualize your application with layers that define how to configure a set of resources that are managed together. You can also define the software configuration for each layer, including installation scripts and initialization tasks. When an instance is added to a layer, all the configuration steps are applied for you. AWS OpsWorks promotes conventions but is flexible enough to let you customize any aspect of your environment. Since AWS OpsWorks uses Chef recipes, you can leverage hundreds of community-built configurations such as PostgreSQL, Nginx, and Solr.

AWS OpsWorks uses automation to simplify operations. You specify how to deploy, scale, and maintain your applications and AWS OpsWorks performs the tasks for you. AWS OpsWorks can scale your applications using automatic load-based or time-based scaling, and maintain the health of your applications by detecting failed instances and replacing them. With AWS OpsWorks, you can deploy your applications to 1,000s of Amazon EC2 instances with the same effort as a single instance.

Application Services

Amazon Simple Queue Service (Amazon SQS)

Amazon Simple Queue Service (Amazon SQS) offers a reliable, highly scalable, hosted queue for storing messages as they travel between computers. By using Amazon SQS, developers and administrators can simply move data between distributed components of their applications that perform different tasks, without losing messages or requiring each component to be always available. Amazon SQS makes it easy to build an automated workflow, working in close conjunction with Amazon EC2 and the other AWS infrastructure web services.

Amazon SQS works by exposing Amazon's web-scale messaging infrastructure as a web service. Any computer on the Internet can add or read messages without any installed software or special firewall configurations. Components of

applications using Amazon SQS can run independently, and do not need to be on the same network, developed with the same technologies, or running at the same time.

Amazon Simple Notification Service (Amazon SNS)

Amazon Simple Notification Service (Amazon SNS) is a web service that makes it easy to set up, operate, and send notifications from the cloud. It provides developers and system administrators with a highly scalable, flexible, and cost-effective capability to publish messages from an application and immediately deliver them to subscribers or other applications. It is designed to make web-scale computing easier.

Amazon SNS provides a simple web services interface and browser-based console that can be used to create topics you want to notify applications (or people) about; subscribe clients to these topics; publish messages; and have these messages delivered over the clients' protocol of choice (such as HTTP, email, or SMS). Amazon SNS delivers notifications to clients using a "push" mechanism that eliminates the need to periodically check or "poll" for new information and updates. Amazon SNS can be leveraged to build highly reliable, event-driven workflows and messaging applications without the need for complex middleware and application management. The potential uses for Amazon SNS include monitoring applications, workflow systems, time-sensitive information updates, mobile applications, and many others.

Amazon Simple Workflow Service (Amazon SWF)

Amazon Simple Workflow Service (Amazon SWF) is a workflow service for building scalable, resilient applications. Whether automating business processes for finance or insurance applications, building sophisticated data analytics applications, or managing cloud infrastructure services, Amazon SWF reliably coordinates all of the processing steps within an application.

With traditional development approaches, it is both time-consuming and costly to build and track processing steps that run at different times and have different durations, while ensuring that the steps are performed reliably and without duplication. When you distribute the implementation of applications across multiple systems, coordinating the processing steps across those systems presents an added challenge. Using Amazon SWF, developers can structure the various processing steps in an application as "tasks" that drive work in distributed applications. Amazon SWF then coordinates these tasks in a reliable and scalable manner. Amazon SWF manages task execution dependencies, scheduling, and concurrency based on a developer's application logic. The service stores tasks, reliably dispatches them to application components, tracks their progress, and keeps their latest state.

Amazon SWF is a fully managed service, without hardware or software to administer, scale, tune, patch, or upgrade. Amazon SWF provides simple API calls that can be executed from code written in any language and run on your EC2 instances, or on any of your machines located anywhere in the world that can access the Internet. On-premises machines can simply open an Internet connection to request tasks from the Amazon SWF API, requiring no changes to firewall rules. Detailed reporting on the current status and execution history of your workflows is available through the AWS Management Console.

Amazon Simple Email Service (Amazon SES)

Amazon Simple Email Service (Amazon SES) is a highly scalable and cost-effective bulk and transactional email sending service for organizations and developers. Amazon SES eliminates the complexity and expense of building an in-house email solution or licensing, installing, and operating a third-party email service. The service integrates with other AWS services, making it easy to send emails from applications that are hosted on services such as Amazon EC2. With Amazon SES there is no long-term commitment, minimum spend, or negotiation required. Organizations can utilize a free usage tier and after that enjoy low fees for the number of emails sent plus data transfer fees.

Building large-scale email solutions to send marketing and transactional messages is often a complex and costly challenge for organizations. To optimize the percentage of emails that are successfully delivered, organizations must deal with email server management and network configuration, plus they must meet rigorous Internet service provider (ISP) standards for email content. Additionally, many third-party email solutions require contract and price negotiations, as well as significant up-front costs.

Amazon SES eliminates these challenges and enables organizations to benefit from the years of experience and sophisticated email infrastructure Amazon.com has built to serve its own large-scale customer base. Using SMTP or a simple API call, an organization can now access a high-quality, scalable email infrastructure to efficiently and inexpensively communicate to their customers. For high email deliverability, Amazon SES uses content filtering technologies to scan an organization's outgoing email messages to help ensure that the content meets ISP standards. The email message is then either queued for sending or routed back to the sender for corrective action. To help organizations further improve the quality of email communications with their customers, Amazon SES provides a built-in feedback loop, which includes notifications of bounce backs, failed and successful delivery attempts, and spam complaints.

Amazon CloudSearch

Amazon CloudSearch is a fully-managed search service in the cloud that allows customers to easily integrate fast and highly scalable search functionality into their applications. With a few clicks in the AWS Management Console, developers simply create a search domain, upload the data they want to make searchable to Amazon CloudSearch, and then the service automatically provisions the technology resources required and deploys a highly tuned search index.

Amazon CloudSearch seamlessly scales as the amount of searchable data increases or as the query rate changes. Developers can change search parameters, fine-tune search relevance, and apply new settings at any time without having to upload the data again.

Amazon CloudSearch enables customers to offload the administrative burden of operating and scaling a search platform. Customers don't have to worry about hardware provisioning, data partitioning, or software patches.

Amazon Elastic Transcoder

Amazon Elastic Transcoder is video transcoding in the cloud. It is designed to be a highly scalable, easy to use and a cost effective way for developers and businesses to convert (or "transcode") video files from their source format into versions that will playback on devices like smartphones, tablets and PCs.

Amazon Elastic Transcoder manages all aspects of the transcoding process for you transparently and automatically. There's no need to administer software, scale hardware, tune performance, or otherwise manage transcoding infrastructure. You simply create a transcoding "job" specifying the location of your source video and how you want it transcoded. Amazon Elastic Transcoder also provides transcoding presets for popular output formats, which means that you don't need to guess about which settings work best on particular devices. All these features are available via service APIs and the AWS Management Console.

Like other Amazon Web Services, there are no contracts or monthly commitments for using Amazon Elastic Transcoder – you simply pay based on the minutes you need to transcode, and resolution of the content transcoded.

Planning Your Next Steps

When you consider your move to the Amazon Web Services cloud, the first thing to do is make sure that your IT plan aligns with your organization's business model. Knowing how you want to take advantage of cloud resources requires understanding your core competencies and identifying the areas that are best served through an external infrastructure.

Next, you have to think through key technology questions. The list of questions will vary depending upon your project and business, but usually includes the following:

- Do you have legacy applications that need greater scalability, reliability, or security than you can afford to maintain in your own environment?
- What are your hardware and bandwidth capacity requirements?
- How will you be prepared to scale up (and down) following deployment?
- How can the cloud advance your IT and business objectives?

As you answer each question, apply the lenses of *flexibility*, *cost effectiveness*, *scalability*, *elasticity*, and *security*. Taking advantage of Amazon Web Services will allow you to focus on your core competencies and leverage the resources and experience Amazon provides.

Getting Started With AWS

1. Read How do I get started with AWS?

2. AWS Whitepapers – http://aws.amazon.com/whitepapers

3. AWS Economics Center – http://aws.amazon.com/economics

4. AWS Security Center – http://aws.amazon.com/security

5. AWS Architecture Center – http://aws.amazon.com/architecture