Amazon EC2 FAQ

Amazon EC2 Windows FAQ

Amazon EC2 Container Service FAQ

AWS Elastic Beanstalk FAQ

AWS Lambda FAQ

Amazon VPC FAQ

Amazon S3 FAQ

Amazon CloudFront FAQ

Amazon EFS FAQ

Amazon Glacier FAQ

AWS Import/Export Snowball FAQ

AWS Storage Gateway FAQ

Amazon RDS FAQ

Amazon Database Migration Service FAQ

Amazon DynamoDB FAQ

Amazon ElastiCache FAQ

Amazon Redshift FAQ

AWS Direct Connect FAQ

Amazon Route 53 FAQ

AWS CodePipeline FAQ

AWS CodeDeploy FAQ

AWS CodeCommit FAQ

Amazon CloudWatch FAQ

AWS CloudFormation FAQ

Amazon CloudTrail FAQ

Amazon Config FAQ

AWS Management Console FAQ

AWS OpsWorks FAQ

Amazon Service Catalog FAQ

AWS Trusted Advisor

AWS Identity and Access Management FAQ

AWS CloudHSM FAQ

AWS Key Management Service FAQ

AWS WAF FAQ

AWS IoT FAQ

Amazon Lumberyard

Amazon GameLift

Amazon Elastic MapReduce FAQ

AWS Data Pipeline FAQ

Amazon Elasticsearch Service FAQ

Amazon Kinesis FAQ

Amazon Machine Learning FAQ

Amazon QuickSight FAQ

Amazon API Gateway FAQ

Amazon AppStream FAQ

Amazon CloudSearch FAQ

Amazon Elastic Transcoder FAQ

Amazon SES FAQ

Amazon SNS FAQ

Amazon SQS FAQTechnical FAQ

Amazon SWF FAQ

Amazon EC2 FAQ

Longer EC2, EBS, and Storage Gateway Resource IDs

Q: What is changing?

EC2 instance and reservation IDs, and volume and snapshot IDs for EBS and Storage Gateway, are changing to a longer format. The transition to longer instance and reservation IDs started in January 2016 and will last through early December 2016, and the transition to longer volume and snapshot IDs started in April 2016 and will last through early December 2016. During this time, you can choose which ID format these resources are assigned, and you can update your management tools and scripts to add support for the longer format. After early December 2016, all newly created instances, reservations, volumes, and snapshots will be required to use the longer ID format.

The new format will only apply to newly created resources; your existing resources won't be affected. Visit the AWS Blog for a step-by-step overview of how to opt in to longer IDs.

Q: Will I need to upgrade to a new version of the AWS SDKs or CLI?

To use the AWS CLI and SDKs with longer IDs, you must upgrade to the following versions:

- PHPv2: Must upgrade to v2.8.27+
- PHPv3: Must upgrade to v3.15.0+
- CLI: Must upgrade to v1.10.2+
- Boto3: Must upgrade to v1.2.1+
- Botocore: Must upgrade to v1.3.24+

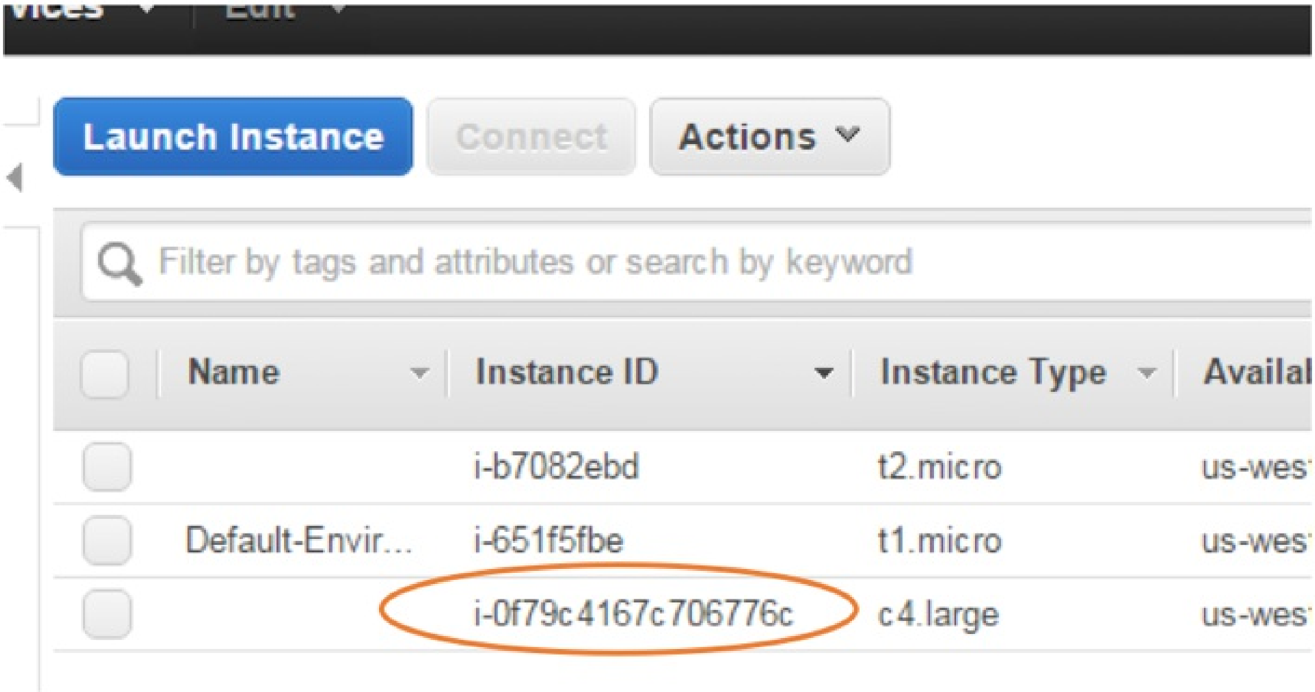
The following SDKs are fully compatible with longer IDs and do not need to be upgraded: PHP v1, Boto v1, Boto v2, Ruby v1, Ruby v2, JavaScript, Java, .NET, AWS Tools for Windows PowerShell, and Go.

For all tools, if you wish to use the new ModifyIdFormat and DescribeIdFormat APIs, you will need to update your tools to receive the new APIs starting in January 2016.

Q: What will the new identifier format look like?

The new identifier format will follow the pattern of the current identifier format, but it will be longer. The new format will be <resource identifier>-<17 characters>, e.g. “i-1234567890abcdef0” for EC2 instances or “snap-1234567890abcdef0” for EBS snapshots.

An example of the new instance ID format in the EC2 Console is shown below.



Name	Instance ID	Instance Type	Availal
	i-b7082ebd	t2.micro	us-wes
Default-Envir...	i-651f5fbe	t1.micro	us-wes
	i-0f79c4167c706776c	c4.large	us-wes

Q: Why is this necessary?

We need to do this given how fast AWS is continuing to grow; we will start to run low on IDs for certain EC2 and EBS resources within a year or so. In order to enable the long-term, uninterrupted creation of new instances, reservations, volumes, and snapshots, we need to introduce a longer ID format for these resources. Additional identifiers might need to expand within the next few years as well.

Q: How does this impact me?

There is a good chance that you won’t need to make any system changes to handle the new format. If you only use the console to manage AWS resources, you might not be impacted at all, but you should still update your settings to use the longer ID format as soon as possible. If you interact with AWS resources via APIs, SDKs, or the AWS CLI, you might be impacted, depending on whether your software makes assumptions about the ID format when validating or persisting resource IDs. If this is the case, you might need to update your systems to handle the new format.

Some failure modes could include:
- If your systems use regular expressions to validate the ID format, you might error if a longer format is encountered.

• If there are expectations about the ID length in your database schemas, you might be unable to store a longer ID.

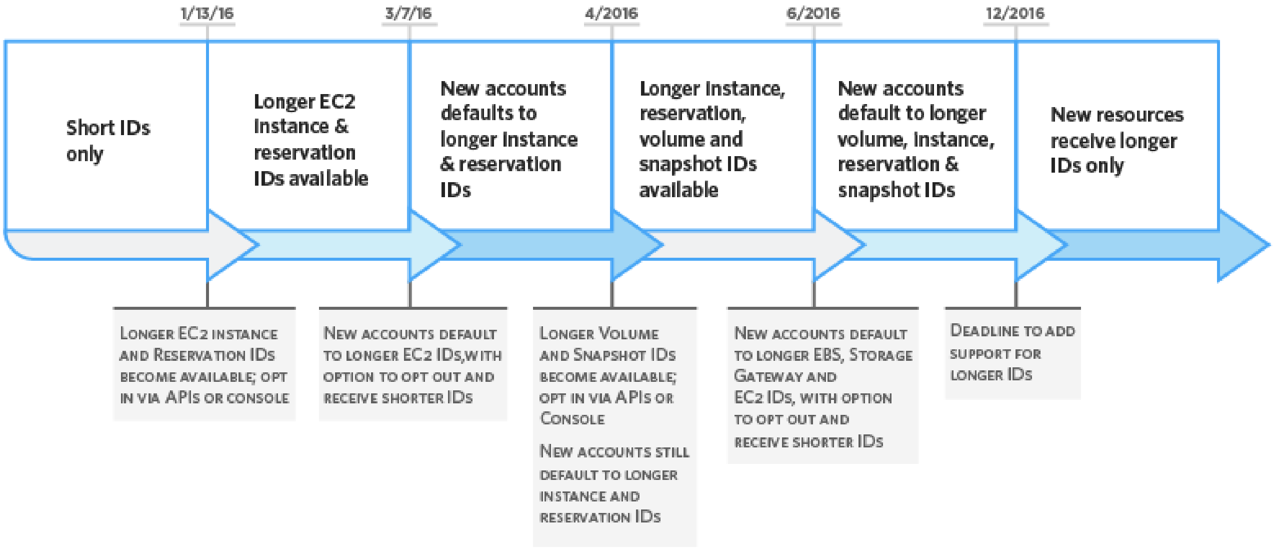
Depending on the tools you are using, you may need to upgrade to newer versions of the AWS CLI and SDKs. See above for a list of affected tools and compatible versions.

Q: Will this affect existing resources?

No; only resources that are created after you opt in to the longer format will be affected. Once a resource has been assigned an ID (long or short), that ID will never change. Any resource created with the old ID format will always retain its shorter ID, and any resource created with the new format will retain its longer ID, even if you opt back out.

Q: When will this happen?

The rollout timeline for longer instance and reservation IDs is shown below.


1/13/16
3/7/16
4/2016
6/2016
12/2016
Short IDs only
Longer EC2 instance & reservation IDs available
New accounts defaults to longer instance & reservation IDs
Longer instance, reservation, volume and snapshot IDs available
New accounts default to longer volume, instance, reservation & snapshot IDs
New resources receive longer IDs only
Longer EC2 instance and Reservation IDs become available; opt in via APIs or console
New accounts default to longer EC2 IDs,with option to opt out and receive shorter IDs
Longer Volume and Snapshot IDs become available; opt in via APIs or Console
New accounts still default to longer instance and reservation IDs
New accounts default to longer EBS, Storage Gateway and EC2 IDs, with option to opt out and receive shorter IDs
Deadline to add support for longer IDs

Starting on January 13, 2016, longer EC2 instance and reservation IDs are available for opt-in via APIs and the console. Between January 2016 and December 2016, all accounts can opt in and out of longer instance and reservation IDs as needed for testing.

Starting on April 28, 2016, new accounts default to longer EC2 instance and reservation IDs in every AWS region except Beijing (China) and AWS GovCloud (US), with the option to request the shorter format if needed.

Longer EBS and Storage Gateway volume and snapshot IDs are available from April 25, 2016 for opt-in via APIs and the Console. New accounts created in June 2016 or later will default to longer snapshot and volume IDs, with the option to opt out if needed.

Early December 2016 is the deadline to add support for longer IDs. The longer IDs transition will occur one region at a time, between December 5, 2016 and December 16, 2016. After that point, the option to switch formats will no longer be available, and all newly created instance, reservation, volume, and snapshot IDs will have the longer format.

Q: Why is the rollout period so long?

We want to give you as much time as possible to test your systems with the new format. A long transition window offers maximum flexibility to test and update your systems incrementally and will help minimize interrupts as you add support for the new format.

Q: What if I prefer to keep receiving the shorter ID format after December 2016?

Unfortunately, this is not possible regardless of your user settings specified.

Q: How does opting in work? And opting out?

Throughout the transition period (January 2016 to December 2016), you can opt to receive longer or shorter IDs by using the APIs or the EC2 Console. ModifyIdFormat sets the format of instance and reservation IDs, and DescribeIdFormat lets you view your ID format settings. Both APIs apply to the user making the call and are region-specific. ID format settings can be modified per IAM user, region, and resource type. Any IAM user without explicit settings will fall back to the settings of the root account. Usually, after you update your ID format settings, it can take a few minutes for the settings to take effect.

If your testing uncovers issues that you need to address, you can opt back out of the new, longer ID format until your systems are prepared to handle longer IDs. This option will be available until December 2016. From December 2016, the new, longer ID format will become mandatory, and the shorter format will no longer be available.

Q: How can I opt in the entire account at once?

Yes, you can opt in using the AWS CLImodify-identity-id-format and describe-identity-id-format and specify the desired ARN and resource type. You will need to do this separately for each resource type (instances, volumes, reservations, and snapshots). To opt in the entire account, you must specify the root account as the Amazon Resource Name (ARN). This will apply changes across the account, and you will not need to set each individual user/role preference. For more information, see the EC2 User Guide or Knowledge Center.

Note: If you opt in the root user, all users/roles launching instances on the account will adopt the root user preference unless their specific user/role (ARN) opt-in preference is already explicitly set. You should only opt in the root user if you are confident that all services using your account supports longer IDs.

Q: How can I opt in across all regions at once?

You can opt in across all regions at once using theLonger-Id-Converter tool. Using this tool you can opt in all regions and all resources. This tool will migrate not only the root or admin account but also all IAM roles/users under the root account across all regions. You can also use this tool to check your opt-in status for your account. For more information about this tool, refer to README file.

Note: If your systems run into an issue after transitioning to longer IDs, you can use the same tool to revert back to using shorter IDs for your account across all regions.

Q: Can I opt in to longer IDs per IAM role?

Yes, you can use the new modify-Identity-id-format and describe-identity-id-format APIs to control and view how different identities are opted in to using longer IDs. You can choose to opt in to longer IDs on a per-account, per-IAM role, or per-IAM user basis. Opting in by IAM user or role can help you test your systems before opting in your entire account. For more information, see the EC2 User Guide.

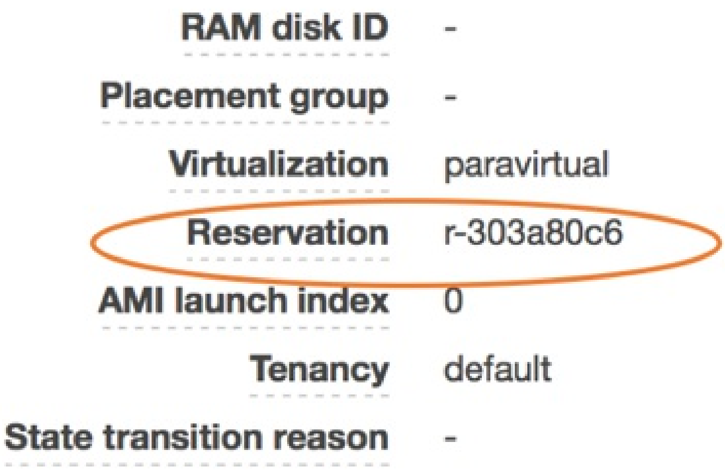
Note: In the 2015-10-01 version of the Amazon EC2 API, if you call describe-id-format or modify-id-format using IAM role credentials, the results apply to the entire AWS account, and not the specific IAM role. In the current version of the Amazon EC2 API, the results will correctly apply to the IAM role only.

Q: What will happen if I take no action?

If you do not opt in to the new format during the transition window, you will be automatically opted in at the final deadline in December 2016. We do not recommend this approach; it is better to add support for the new format during the transition window, which offers the opportunity for controlled testing.

Q: What is a reservation ID? Do reservation IDs only apply to Reserved Instances?

Reservation IDs apply to all instances, and are different from Reserved Instances. Every instance launched by EC2 has a reservation ID. A reservation ID has a one-to-one relationship with an instance launch request, but can be associated with more than one instance if you launch multiple instances using the same launch request. The reservation ID is returned by the DescribeInstances API, and it can be viewed in the EC2 Management Console description of any given instance (see below).



RAM disk ID	-
Placement group	-
Virtualization	paravirtual
Reservation	r-303a80c6
AMI launch index	0
Tenancy	default
State transition reason	-

Q: What best practices do you recommend as I test my systems and add support for the

new ID formats?

If your software can run under multiple distinct AWS accounts, choose (or create) an AWS account to test with. Alternatively, if your software runs under a single AWS account, choose (or create) an IAM user to test with.

Set your chosen account or IAM user to receive longer IDs, test your software, and make any necessary changes. Note that if one IAM user launches an instance with a longer ID, all other users will be able to see the longer ID in subsequent describe calls, regardless of user-specific opt-in settings. Once you are comfortable that your software will operate as expected, you can opt in all of your accounts and / or users. If any unexpected issues arise, you can opt out until the issues are understood and corrected. This testing procedure will be possible until the December 2016 deadline, when all new instances, reservations, volumes, and snapshots will receive the longer ID format.

Once your software is ready for longer IDs, opt in to longer IDs across all of your accounts, regions and resources. When this is complete, you have transitioned to the new format fully and will not need to take further action.

Q: How do I know when I've finished the opt-in process for longer resource IDs?

Once you are done with the testing process described above, opt in to longer IDs across every region and user. Alternatively, you can opt in the root user for every region; this will update the ID format settings for the whole account, as long as no individual IAM users are opted out. You will need to do this separately for each resource type (instances, volumes, reservations, and snapshots). Once this is complete, you are done with the transition process and will not need to make further changes for these resource types. Note that, since existing resources will retain their original IDs, you might see a mix of long IDs (for new resources) and short IDs (for pre-existing resources) when you are done with the opt-in process.

Q: What will be the default ID type for new accounts?

For instances and reservations, accounts created on April 28, 2016 or later will be configured to receive the longer ID format by default in every AWS region except Beijing (China) and AWS GovCloud (US). If you are a new customer, this will make the transition to longer instance and reservation IDs really simple. If you would like your new account to assign the shorter ID format to your resources, then simply reconfigure your account for shorter IDs as described above. This workflow will be necessary until you are ready for your accounts to receive longer IDs.

For volumes and snapshots, accounts created in June 2016 or later will be configured to receive the longer ID format by default, with the option to opt out if necessary until December 2016.

Q: Will you be changing other identifiers?

As AWS continues growing, it's possible that we will need to increase the ID length of other resources in the future.

Q: Does this apply to Spot instances?

Yes; the longer instance and reservation ID formats will apply to all EC2 instance types.

Q: I'm an EC2 Windows customer; is there anything Windows-specific I need to know?

If you use EC2 instance IDs as part of the computer name for your EC2 Windows instances, please note that Windows will automatically truncate the name to 15 characters to adhere to NetBIOS naming conventions. Due to this truncation behavior, you may see duplicate computer names at 15 characters if you're using this naming convention. We recommend choosing a unique naming scheme to avoid complications.

Q: How can I opt in new Auto Scaling instances to longer IDs?

Auto Scaling reflects the setting of the root user. This supersedes what has been configured by the IAM role.

Q: I use AWS through a third-party tool, what do I need to do to handle longer IDs?

We are working with third parties to ensure the best customer experience, but please work with your ISV to determine their level of support for this change prior to turning on the longer ID format in your account.

Q: When will the longer IDs' final transition happen?

The longer IDs transition will occur one region at a time, between December 5, 2016 and December 16, 2016. You can check the scheduled transition date for your each region by using the AWS CLI describe-id-format.

Q: If I opt in to longer IDs and then opt back out during the transition window, what will happen to resources that were created with longer IDs?

Once a resource has been assigned an ID it will not change, so resources that are created with longer IDs will retain the longer IDs regardless of later actions. If you opt in to the longer format, create resources, and then opt out, you will see a mix of long and short resource IDs, even after opting out. The only way to get rid of long IDs will be to delete or terminate the respective resources.

For this reason, exercise caution and avoid creating critical resources with the new format until you have tested your tools and automation.

Q: What should I do if my systems are not working as expected before the final transition, December 16th 2016?

If your systems are not working as expected during the transition period, you can temporarily opt out of longer format IDs and remediate your systems, however your account will automatically be transitioned back to using longer IDs after December 16th, 2016. Regardless of your account settings, all new instances, reservations, volumes, and snapshots will receive the longer format IDs, so it is important for you to test your systems with longer format IDs before the final transition window starts. By testing and opting in earlier, you give yourself valuable time to make modifications to your resources with short resource IDs and you minimize the risk of any impact to your systems.

Q: What will happen if I launch EC2 and EBS resources in multiple regions during the final transition window in December 2016?

Your resources' ID length will depend upon the region you launch your resources. If the region has already transitioned to using longer IDs, resources launched in that region will have longer format IDs; if not, they will have shorter resource IDs. Therefore, during the transition window, you may see a mix of shorter and longer resource IDs; however, after December 16th 2016, all new resources will have longer format IDs in all regions.

Q: If AWS adds new regions during the transition window, will new regions support longer IDs?

Yes. All new regions launching in the second half of 2016 and after will issue longer format instances, reservations, volumes, and snapshot IDs by default for both new and existing accounts.

General

Q: What is 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.

Q: What can I do with Amazon EC2?

Just as Amazon Simple Storage Service (Amazon S3) enables storage in the cloud, Amazon EC2 enables "compute" in the cloud. 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.

Q: How can I get started with Amazon EC2?

To sign up for Amazon EC2, click the “Sign up for This Web Service” button on the Amazon EC2 detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon EC2 sign-up process. After signing up, please refer to the Amazon EC2 documentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for Amazon EC2?

Amazon EC2 registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a couple of minutes and involves receiving a phone call during the registration process and entering a PIN number using the phone key pad.

Q: What can developers now do that they could not before?

Until now, small developers did not have the capital to acquire massive compute resources and ensure they had the capacity they needed to handle unexpected spikes in load. Amazon EC2 enables any developer to leverage Amazon’s own benefits of massive scale with no up-front investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure they have the compute capacity they need to meet their business requirements.

The “Elastic” nature of the service allows developers to instantly scale to meet spikes in traffic or demand. When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Q: How do I run systems in the Amazon EC2 environment?

Once you have set up your account and select or create your AMIs, you are ready to boot your instance. You can start your AMI on any number of On-Demand instances by using the RunInstances API call. You simply need to indicate how many instances you wish to launch. If you wish to run more than 20 On-Demand instances, complete the Amazon EC2 instance request form.

If Amazon EC2 is able to fulfill your request, RunInstances will return success, and we will start launching your instances. You can check on the status of your instances using the DescribeInstances API call. You can also programmatically terminate any number of your instances using the TerminateInstances API call.

If you have a running instance using an Amazon EBS boot partition, you can also use the StopInstances API call to release the compute resources but preserve the data on the boot partition. You can use the StartInstances API when you are ready to restart the associated instance with the Amazon EBS boot partition.

In addition, you have the option to use Spot Instances to reduce your computing costs when you have flexibility in when your applications can run. Read more about Spot Instances for a more detailed explanation on how Spot Instances work.

If you prefer, you can also perform all these actions from theAWS Management Console or through the command line using our command line tools, which have been implemented with this web service API.

Q: What is the difference between using the local instance store and Amazon Elastic Block storage (Amazon EBS) for the root device?

When you launch your Amazon EC2 instances you have the ability to store your root device data on Amazon EBS or the local instance store. By using Amazon EBS, data on the root device will persist independently from the lifetime of the instance. This enables you to stop and restart the instance at a subsequent time, which is similar to shutting down your laptop and restarting it when you need it again.

Alternatively, the local instance store only persists during the life of the instance. This is an inexpensive way to launch instances where data is not stored to the root device. For example, some customers use this option to run large web sites where each instance is a clone to handle web traffic.

Q: How quickly will systems be running?

It typically takes less than 10 minutes from the issue of the RunInstances call to the point where all requested instances begin their boot sequences. This time is dependant on a number of factors including: the size of your AMI, the number of instances you are launching, and how recently you have launched that AMI. Images launched for the first time may take slightly longer to boot.

Q: How do I load and store my systems with Amazon EC2?

Amazon EC2 allows you to set up and configure everything about your instances from your operating system up to your applications. An Amazon Machine Image (AMI) is simply a packaged-up environment that includes all the necessary bits to set up and boot your instance. Your AMIs are your unit of deployment. You might have just one AMI or you might compose your system out of several building block AMIs (e.g., webservers, appservers, and databases). Amazon EC2 provides a number of tools to make creating an AMI easy. Once you create a custom AMI, you will need to bundle it. If you are bundling an image with a root device backed by Amazon EBS, you can simply use the bundle command in the AWS Management Console. If you are bundling an image with a boot partition on the instance store, then you will need to use the AMI Tools to upload it to Amazon S3. Amazon EC2 uses Amazon EBS and Amazon S3 to provide reliable, scalable storage of your AMIs so that we can boot them when you ask us to do so.

Or, if you want, you don't have to set up your own AMI from scratch. You can choose from a number of globally available AMIs that provide useful instances. For example, if you just want a simple Linux server, you can choose one of the standard Linux distribution AMIs.

Q: How do I access my systems?

The RunInstances call that initiates execution of your application stack will return a set of DNS names, one for each system that is being booted. This name can be used to access the system exactly as you would if it were in your own data center. You own that machine while your operating system stack is executing on it.

Q: Is Amazon EC2 used in conjunction with Amazon S3?

Yes, Amazon EC2 is used jointly with Amazon Simple Storage Service (Amazon S3) for instances with root devices backed by local instance storage. By using Amazon S3, developers have access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. In order to execute systems in the Amazon EC2 environment, developers use the tools provided to load their Amazon Machine Images (AMIs) into Amazon S3 and to move them between Amazon S3 and Amazon EC2. See How do I load and store my systems with Amazon EC2?for more information about AMIs.

We expect developers to find the combination of Amazon EC2 and Amazon S3 to be very useful. Amazon EC2 provides cheap, scalable compute in the cloud while Amazon S3 allows users to store their data reliably.

Q: How many instances can I run in Amazon EC2?

You are limited to running up to 20 On-Demand instances, purchasing 20 Reserved Instances, and requesting Spot Instances per your dynamic Spot limit per region. New AWS accounts may start with limits that are lower than the limits described here. Certain instance types are further limited per region as follows:

Instance Type	On-Demand Limit	Reserved Limit	Spot Limit
m4.4xlarge	10	20	Dynamic Spot Limit
m4.10xlarge	5	20	Dynamic Spot Limit
m4.16xlarge	5	20	Dynamic Spot Limit
c4.4xlarge	10	20	Dynamic Spot Limit
c4.8xlarge	5	20	Dynamic Spot Limit
cg1.4xlarge	2	20	Dynamic Spot Limit
hi1.4xlarge	2	20	Dynamic Spot Limit
hs1.8xlarge	2	20	Not offered
cr1.8xlarge	2	20	Dynamic Spot Limit
p2.xlarge	1	20	Dynamic Spot Limit
p2.8xlarge	1	20	Dynamic Spot Limit
p2.16xlarge	1	20	Dynamic Spot Limit
g2.2xlarge	5	20	Dynamic Spot Limit
g2.8xlarge	2	20	Dynamic Spot Limit
r3.4xlarge	10	20	Dynamic Spot Limit
r3.8xlarge	5	20	Dynamic Spot Limit
i2.xlarge	8	20	Dynamic Spot Limit
i2.2xlarge	8	20	Dynamic Spot Limit
i2.4xlarge	4	20	Dynamic Spot Limit
i2.8xlarge	2	20	Dynamic Spot Limit
d2.4xlarge	10	20	Dynamic Spot Limit
d2.8xlarge	5	20	Dynamic Spot Limit

d2.8xlarge	8	20	Dynamic Spot Limit
t2.nano	20	20	Not offered
t2.micro	20	20	Not offered
t2.small	20	20	Not offered
t2.medium	20	20	Not offered
t2.large	20	20	Not offered
All Other Instance Types	20	20	Dynamic Spot Limit

Note that cc2.8xlarge, cg1.4xlarge, hi1.4xlarge, hs1.8xlarge, cr1.8xlarge, G2, D2, and I2 instances are not available in all regions.

If you need more instances, complete the Amazon EC2 instance request form with your use case and your instance increase will be considered. Limit increases are tied to the region they were requested for.

Q: Are there any limitations in sending email from EC2 instances?

Yes. In order to maintain the quality of EC2 addresses for sending email, we enforce default limits on the amount of email that can be sent from EC2 accounts. If you wish to send larger amounts of email from EC2, you can apply to have these limits removed from your account by filling out this form.

Q: How quickly can I scale my capacity both up and down?

Amazon EC2 provides a truly elastic computing environment. Amazon 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. When you need more instances, you simply call RunInstances, and Amazon EC2 will typically set up your new instances in a matter of minutes. Of course, because this is all controlled with web service APIs, your application can automatically scale itself up and down depending on its needs.

Q: What operating system environments are supported?

Amazon EC2 currently supports a variety of operating systems including: Amazon Linux, Ubuntu, Windows Server, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, Fedora, Debian, CentOS, Gentoo Linux, Oracle Linux, and FreeBSD. We are looking for ways to expand it to other platforms.

Q: Does Amazon EC2 use ECC memory?

In our experience, ECC memory is necessary for server infrastructure, and all the hardware underlying Amazon EC2 uses ECC memory.

Q: How is this service different than a plain hosting service?

Traditional hosting services generally provide a pre-configured resource for a fixed amount of time and at a predetermined cost. Amazon EC2 differs fundamentally in the flexibility, control and significant cost savings it offers developers, allowing them to treat Amazon EC2 as their own personal data center with the benefit of Amazon.com’s robust infrastructure.

When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Secondly, many hosting services don’t provide full control over the compute resources being provided. Using Amazon EC2, developers can choose not only to initiate or shut down instances at any time, they can completely customize the configuration of their instances to suit their needs – and change it at any time. Most hosting services cater more towards groups of users with similar system requirements, and so offer limited ability to change these.

Finally, with Amazon EC2, developers enjoy the benefit of paying only for their actual resource consumption – and at very low rates. Most hosting services require users to pay a fixed, up-front fee irrespective of their actual computing power used, and so users risk overbuying resources to compensate for the inability to quickly scale up resources within a short time frame.

Billing

Q: How will I be charged and billed for my use of Amazon EC2?

You pay only for what you use and there is no minimum fee. Pricing is per instance-hour consumed for each instance type. Partial instance-hours consumed are billed as full hours. Data transferred between AWS services in different regions will be charged as Internet Data Transfer on both sides of the transfer. Usage for other Amazon Web Services is billed separately from Amazon EC2.

For EC2 pricing information, please visit thepricing section on the EC2 detail page.

Q: When does billing of my Amazon EC2 systems begin and end?

Billing commences when Amazon EC2 initiates the boot sequence of an AMI instance. Billing ends when the instance terminates, which could occur through a web services command, by running "shutdown -h", or through instance failure. When you stop an instance, we shut it down but don't charge hourly usage for a stopped instance, or data transfer fees, but we do charge for the storage for any Amazon EBS volumes. To learn more, visit the AWS Documentation.

Q: What defines billable EC2 instance-hours?

Instance-hours are billed for any time your instances are in a "running" state. If you no longer wish to be charged for your instance, you must "stop" or "terminate" the instance to avoid being billed for additional instance-hours. Billing starts when an instance transitions into the running state.

Q: If I have two instances in different availability zones, how will I be charged for regional data transfer?

Each instance is charged for its data in and data out. Therefore, if data is transferred between these two instances, it is charged out for the first instance and in for the second instance.

Q. If I have two instances in different regions, how will I be charged for data transfer?

Each instance is charged for its data in and data out at Internet Data Transfer rates. Therefore, if data is transferred between these two instances, it is charged at Internet Data Transfer Out for the first instance and at Internet Data Transfer In for the second instance.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Hardware Information

Q: What kind of hardware will my application stack run on?

Visit Amazon EC2 Pricing for a list of instances available by region.

Q: How do I select the right instance type?

Amazon EC2 instances are grouped into 5 families: General Purpose, Compute Optimized, Memory Optimized, GPU, and Storage Optimized instances. General Purpose Instances have memory to CPU ratios suitable for most general purpose applications and come with fixed performance (M4 and M3 instances) or burstable performance (T2); Compute Optimized instances (C4 and C3 instances) have proportionally more CPU resources than memory (RAM) and are well suited for scale out compute-intensive applications and High Performance

Computing (HPC) workloads; Memory Optimized Instances (R3 instances) offer larger memory sizes for memory-intensive applications, including database and memory caching applications; GPU Compute instances (P2) take advantage of the parallel processing capabilities of NVIDIA Tesla GPUs for high performance parallel computing; GPU Graphics instances (G2) offer high-performance 3D graphics capabilities for applications using OpenGL and DirectX; Storage Optimized Instances include I2 instances that provide very high, low latency, I/O capacity using SSD-based local instance storage for I/O-intensive applications and D2, Dense-storage instances, that provide high storage density and sequential I/O performance for data warehousing, Hadoop and other data-intensive applications. When choosing instance types, you should consider the characteristics of your application with regards to resource utilization (i.e. CPU, Memory, Storage) and select the optimal instance family and instance size.

Q: M1 and M3 standard instances have the same ratio of CPU and memory. When should I use one instance over the other?

M3 instances provide better, more consistent performance than M1 instances for most use-cases. M3 instances also offer SSD-based instance storage that delivers higher I/O performance. M3 instances are also less expensive than M1 instances. For these reasons, we recommend M3 for applications that require general purpose instances with a balance of compute, memory, and network resources. However, if you need more disk storage than what is provided in M3 instances, you may still find M1 instances useful for running your applications.

Q: What is an "EC2 Compute Unit" and why did you introduce it?

Transitioning to a utility computing model fundamentally changes how developers have been trained to think about CPU resources. Instead of purchasing or leasing a particular processor to use for several months or years, you are renting capacity by the hour. Because Amazon EC2 is built on commodity hardware, over time there may be several different types of physical hardware underlying EC2 instances. Our goal is to provide a consistent amount of CPU capacity no matter what the actual underlying hardware.

Amazon EC2 uses a variety of measures to provide each instance with a consistent and predictable amount of CPU capacity. In order to make it easy for developers to compare CPU capacity between different instance types, we have defined an Amazon EC2 Compute Unit. The amount of CPU that is allocated to a particular instance is expressed in terms of these EC2 Compute Units. We use several benchmarks and tests to manage the consistency and predictability of the performance from an EC2 Compute Unit. The EC2 Compute Unit (ECU) provides the relative measure of the integer processing power of an Amazon EC2 instance. Over time, we may add or substitute measures that go into the definition of an EC2 Compute Unit, if we find metrics that will give you a clearer picture of compute capacity.

Q: What is the regional availability of Amazon EC2 instance types?

For a list of all instances and regional availability, visit Amazon EC2 Pricing.

Security

Q: How do I prevent other people from viewing my systems?

You have complete control over the visibility of your systems. The Amazon EC2 security systems allow you to place your running instances into arbitrary groups of your choice. Using the web services interface, you can then specify which groups may communicate with which other groups, and also which IP subnets on the Internet may talk to which groups. This allows you to control access to your instances in our highly dynamic environment. Of course, you should also secure your instance as you would any other server.

Q: Can I get a history of all EC2 API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of all EC2 API calls (including VPC and EBS) made on your account, you simply turn on CloudTrail in the AWS Management Console. For more information, visit the CloudTrail home page.

Q: Where can I find more information about security on AWS?

For more information on security on AWS please refer to ourAmazon Web Services: Overview of Security Processes white paper and to our Amazon EC2 running Windows Security Guide.

Elastic IP

Q: Why am I limited to 5 Elastic IP addresses per region?

Public (IPV4) internet addresses are a scarce resource. There is only a limited amount of public IP space available, and Amazon EC2 is committed to helping use that space efficiently.

By default, all accounts are limited to 5 Elastic IP addresses per region. If you need more the 5 Elastic IP addresses, we ask that you apply for your limit to be raised. We will ask you to think through your use case and help us understand your need for additional addresses. You can apply for more Elastic IP address here. Any increases will be specific to the region they have been requested for.

Q: Why am I charged when my Elastic IP address is not associated with a running instance?

In order to help ensure our customers are efficiently using the Elastic IP addresses, we impose a small hourly charge for each address when it is not associated to a running instance.

Q: Do I need one Elastic IP address for every instance that I have running?

No. You do not need an Elastic IP address for all your instances. By default, every instance comes with a private IP address and an internet routable public IP address. The private address is associated exclusively with the instance and is only returned to Amazon EC2 when the instance is stopped or terminated. The public address is associated exclusively with the instance until it is stopped, terminated or replaced with an Elastic IP address. These IP addresses should be adequate for many applications where you do not need a long lived internet routable end point. Compute clusters, web crawling, and backend services are all examples of applications that typically do not require Elastic IP addresses.

Q: How long does it take to remap an Elastic IP address?

The remap process currently takes several minutes from when you instruct us to remap the Elastic IP until it fully propagates through our system.

Q: Can I configure the reverse DNS record for my Elastic IP address?

Yes, you can configure the reverse DNS record of your Elastic IP address byfilling out this form. Note that a corresponding forward DNS record pointing to that Elastic IP address must exist before we can create the reverse DNS record.

Availability Zones

Q: How isolated are Availability Zones from one another?

Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone.

Q: Is Amazon EC2 running in more than one region?

Yes. Please refer to Regional Products and Services for more details of our product and service availability by region.

Q: How can I make sure that I am in the same Availability Zone as another developer?

We do not currently support the ability to coordinate launches into the same Availability Zone across AWS developer accounts. One Availability Zone name (for example, us-east-1a) in two AWS customer accounts may relate to different physical Availability Zones.

Q: If I transfer data between Availability Zones using public IP addresses, will I be charged twice for Regional Data Transfer (once because it's across zones, and a second time because I'm using public IP addresses)?

No. Regional Data Transfer rates apply if at least one of the following is true, but is only charged once for a given instance even if both are true:

- The other instance is in a different Availability Zone, regardless of which type of address is used.
- Public or Elastic IP addresses are used, regardless of which Availability Zone the other instance is in.

Enhanced Networking

Q: What networking capabilities are included in this feature?

We currently support enhanced networking capabilities using SR-IOV (Single Root I/O Virtualization). SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization compared to traditional implementations. For supported Amazon EC2 instances, this feature provides higher packet per second (PPS) performance, lower inter-instance latencies, and very low network jitter.

Q: Why should I use Enhanced Networking?

If your applications benefit from high packet-per-second performance and/or low latency networking, Enhanced Networking will provide significantly improved performance, consistence of performance and scalability.

Q: How can I enable Enhanced Networking on supported instances?

In order to enable this feature, you must launch an HVM AMI with the appropriate drivers. X1 and m4.16xlarge instances provide the Elastic Network Adapter (ENA) interface (which uses the "ena" Linux driver) for Enhanced Networking. C3, C4, R3, I2, M4 (except m4.16xlarge) and D2 instances use Intel® 82599g Virtual Function Interface (which uses the "ixgbevf" Linux driver). Amazon Linux AMI includes both of these drivers by default. For AMIs that do not contain these drivers, you will need to download and install the appropriate drivers based on the instance types you plan to use. You can use Linux or Windows instructions to enable Enhanced Networking in AMIs that do not include the SR-IOV driver by default. Enhanced Networking is only supported in Amazon VPC.

Q: Do I need to pay an additional fee to use Enhanced Networking?

No, there is no additional fee for Enhanced Networking. To take advantage of Enhanced Networking you need to launch the appropriate AMI on a supported instance type in a VPC.

Q: Why is Enhanced Networking only supported in Amazon VPC?

Amazon VPC allows us to deliver many advanced networking features to you that are not possible in EC2-Classic. Enhanced Networking is another example of a capability enabled by Amazon VPC.

Q: Which instance types support Enhanced Networking?

Currently C3, C4, D2, I2, M4, X1 and R3 instances support Enhanced Networking. X1, P2, and m4.16xlarge instances provide the Elastic Network Adapter (ENA) interface for Enhanced Networking. C3, C4, R3, I2, M4 (except m4.16xlarge) and D2 instances, use Intel® 82599 Virtual Function Interface.

Amazon Elastic Block Storage (EBS)

Q: What happens to my data when a system terminates?

The data stored on a local instance store will persist only as long as that instance is alive. However, data that is stored on an Amazon EBS volume will persist independently of the life of the instance. Therefore, we recommend that you use the local instance store for temporary data and, for data requiring a higher level of durability, we recommend using Amazon EBS volumes or backing up the data to Amazon S3. If you are using an Amazon EBS volume as a root partition, you will need to set the Delete On Terminate flag to "N" if you want your Amazon EBS volume to persist outside the life of the instance.

Q: What kind of performance can I expect from Amazon EBS volumes?

Amazon EBS provides three volume types: General Purpose (SSD) volumes, Provisioned IOPS (SSD) volumes, and Magnetic volumes. These volume types differ in performance characteristics and price, allowing you to tailor your storage performance and cost to the needs of your applications. For more performance infomation see the EBS product details page.

For additional information on Amazon EBS performance, see theAmazon EC2 User Guide’s EBS section.

Q: What is the EBS General Purpose (SSD) volume type?

The EBS General Purpose (SSD) volumes are backed by the same technology found in EBS Provisioned IOPS (SSD) volumes. The EBS General Purpose (SSD) volume type is designed for 99.999% availability, and a broad range of use-cases such as boot volumes, small and medium size databases, and development and test environments. General Purpose (SSD) volumes deliver a ratio of 3 IOPS per GB, offer single digit millisecond latencies, and also have the ability to burst up to 3000 IOPS for short periods.

Q: Which volume type should I choose?

Customers can now choose between three EBS volume types to best meet the needs of their workloads: General Purpose (SSD), Provisioned IOPS (SSD), and Magnetic. General Purpose (SSD) is the new, SSD-backed, general purpose EBS volume type that we recommend as the default choice for customers. General Purpose (SSD) volumes are suitable for a broad range of workloads, including small to medium sized databases, development and test environments, and boot volumes. Provisioned IOPS (SSD) volumes offer storage with consistent and low-latency performance, and are designed for I/O intensive applications such as large relational or NoSQL databases. Magnetic volumes provide the lowest cost per gigabyte of all EBS volume types. Magnetic volumes are ideal for workloads where data is accessed infrequently, and applications where the lowest storage cost is important.

Q: Do you support multiple instances accessing a single volume?

While you are able to attach multiple volumes to a single instance, attaching multiple instances to one volume is not supported at this time.

Q: Will I be able to access my EBS snapshots using the regular Amazon S3 APIs?

No, EBS snapshots are only available through the Amazon EC2 APIs.

Q: Do volumes need to be un-mounted in order to take a snapshot? Does the snapshot need to complete before the volume can be used again?

No, snapshots can be done in real time while the volume is attached and in use. However, snapshots only capture data that has been written to your Amazon EBS volume, which might exclude any data that has been locally cached by your application or OS. In order to ensure consistent snapshots on volumes attached to an instance, we recommend cleanly detaching the volume, issuing the snapshot command, and then reattaching the volume. For Amazon EBS volumes that serve as root devices, we recommend shutting down the machine to take a clean snapshot.

Q: Are snapshots versioned? Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier, and customers can create volumes based on any of their existing snapshots.

Q: What charges apply when using Amazon EBS shared snapshots?

If you share a snapshot, you won't be charged when other users make a copy of your snapshot. If you make a copy of another user's shared volume, you will be charged normal EBS rates.

Q: Can users of my Amazon EBS shared snapshots change any of my data?

Users who have permission to create volumes based on your shared snapshots will first make a copy of the snapshot into their account. Users can modify their own copies of the data, but the data on your original snapshot and any other volumes created by other users from your original snapshot will remain unmodified.

Q: How can I discover Amazon EBS snapshots that have been shared with me?

You can find snapshots that have been shared with you by selecting "Private Snapshots" from the viewing dropdown in the Snapshots section of the AWS Management Console. This section will list both snapshots you own and snapshots that have been shared with you.

Q: How can I find what Amazon EBS snapshots are shared globally?

You can find snapshots that have been shared globally by selecting "Public Snapshots" from the viewing dropdown in the Snapshots section of the AWS Management Console.

Q: Do you offer encryption on Amazon EBS volumes and snapshots?

Yes. EBS offers seamless encryption of data volumes and snapshots. EBS encryption better enables you to meet security and encryption compliance requirements.

Q: How can I find a list of Amazon Public Data Sets?

All information on Public Data Sets is available in our Public Data Sets Resource Center. You can also obtain a listing of Public Data Sets within the AWS Management Console by choosing “Amazon Snapshots” from the viewing dropdown in the Snapshots section.

Q: Where can I learn more about EBS?

You can visit the EBS FAQ page.

Amazon CloudWatch

Q: What is the minimum time interval granularity for the data that Amazon CloudWatch receives and aggregates?

Metrics are received and aggregated at 1 minute intervals.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

You can retrieve metrics data for any Amazon EC2 instance up to 2 weeks from the time you started to monitor it. After 2 weeks, metrics data for an Amazon EC2 instance will not be available if monitoring was disabled for that Amazon EC2 instance. If you want to archive metrics beyond 2 weeks you can do so by calling mon-get-stats command from the command line and storing the results in Amazon S3 or Amazon SimpleDB.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?

Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 2 weeks.

Q: Does the Amazon CloudWatch monitoring charge change depending on which type of Amazon EC2 instance I monitor?

No, the Amazon CloudWatch monitoring charge does not vary by Amazon EC2 instance type.

Q: Why does the graphing of the same time window look different when I view in 5 minute and 1 minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period, the single data point is placed at the 1 minute mark. We recommend using a 1 minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Auto Scaling

Q: Can I scale up my Amazon EC2 capacity fast but scale it down slowly?

Yes. For example, you can define a scale up condition to increase your Amazon EC2 capacity by 10% and a scale down condition to decrease it by 5%.

Q: What happens if a scaling activity causes me to reach my Amazon EC2 limit of instances?

Auto Scaling Service cannot scale past the Amazon EC2 limit of instances that you can run. If you need more Amazon EC2 instances, complete the Amazon EC2 instance request form.

Q: What happens to my Amazon EC2 instances if I delete my Auto Scaling Group?

If you have an Auto Scaling group with running instances and you choose to delete the Auto Scaling group, the instances will be terminated and the Auto Scaling group will be deleted.

Elastic Load Balancing

Q: What load balancing options does the Elastic Load Balancing service offer?

Elastic Load Balancing offers two types of load balancers that both feature high availability, automatic scaling, and robust security. These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request.

Q: When should I use the Classic Load Balancer and when should I use the Application Load Balancer?

The Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while the Application Load Balancer is ideal for applications needing advanced routing capabilities, microservices, and container-based architectures. Please visit Elastic Load

Balancing for more information.

Reserved Instances

Q: What is a Reserved Instance?

Reserved Instances provide you with a discount on usage of EC2 instances, and a capacity reservation when they are applied to a specific Availability Zone, giving you additional confidence that you will be able to launch the instances you have reserved when you need them.

Q: What is the difference between a Reserved Instance and an On-Demand instance?

When an instance is running On-Demand, you are paying the On-Demand rates for it. When a Reserved Instance applies to an instance, you pay the Reserved Instance discounted hourly rate for your instance usage, and a capacity reservation is created for an instance if your Reserved Instance is applied to a specific Availability Zone.

Q: Can you explain the capacity benefit of a Reserved Instance?

When a Reserved Instance applies to a specific Availability Zone, it is reserving instance capacity matching the Reserved Instance configuration. This benefit provides you additional confidence in your ability to launch instances in a specific Availability Zone, when you need them.

Q: Are Reserved Instances actual instances?

No, Reserved Instances are not physical instances, so they don't have to be launched. Reserved Instances are an EC2 offering that provides a discount on your instance usage and a capacity reservation when assigned to a specific Availability Zone.

Q: Do Reserved Instances apply to Spot Instances or instances running on a Dedicated Host?

No, Reserved Instances do not apply to Spot Instances or instances running on Dedicated Hosts. To lower the cost of using Dedicated Hosts, purchase Dedicated Host Reservations.

Q: How do I purchase a Reserved Instance?

You can purchase a Reserved Instance using the AWS Management Console or using the AWS CLI. Visit the Getting Started page to learn more.

Q: How do I purchase a Reserved Instance for a running instance?

You can purchase a Reserved Instance for a running instance by purchasing a Reserved Instance matching the attributes of your running instance. The attributes that need to align are the instance type, region or Availability Zone, tenancy, and platform description. Visit the Getting

Started page to learn more.

Q: When should I purchase a Reserved Instance for a specific Availability Zone?

You should purchase a Reserved Instance for a specific Availability Zone if you need a capacity reservation. Otherwise, you should assign your Reserved Instance to a region to benefit from a broader application of the Reserved Instance rate.

Q: I own Reserved Instances assigned to specific Availability Zones. How do I assign them to a region?

You can assign your Reserved Instances to a region using the EC2 Management Console and modifying the Scope of the Reserved Instance from “Availability Zone” to “Region”. When you are purchasing new Reserved Instances in the AWS console, by default you will see Reserved Instances with a scope of Region.

Q: How does AWS assign my Reserved Instance rate to instance usage in different Availability Zones?

When your Reserved Instance is assigned to a region, AWS applies your Reserved Instance rate to usage on a first-in basis.

Q: Do I control which instances are billed at the lower rate?

No, AWS automatically optimizes which instances are charged at the lower rate to ensure you always pay the lowest amount. For information about hourly billing, and how it applies to Reserved Instances, see Billing Benefits and Payment Options.

Q: Can I reassign my Standard Reserved Instance from one instance type (e.g., c1.xlarge) to another (e.g., m1.large)?

No. A Standard Reserved Instance is associated with a specific instance type for the duration of its term; however, you can change from one instance size (e.g., c3.large) to another (e.g., c3.xlarge) in the same type, if it is a Linux/UNIX Reserved Instance. If you’d like to have flexibility among instance types, we recommend purchasing a Convertible Reserved Instance. Please refer to the Convertible Reserved Instances section of the FAQ for additional information.

Q: When I launch instances and do not specify an Availability Zone, will my Reserved Instance apply to my instance?

If you’ve purchased a Reserved Instance and it’s assigned to a region, your instance can benefit from the Reserved Instance rate. If you’ve assigned your Reserved Instance to a specific Availability Zone and the Availability Zone of your Reserved Instance does not align with the Availability Zone of your instance, the Reserved Instance will not apply to the instance.

Q: How do the payment options impact my bill?

When you purchase Reserved Instances under the All Upfront payment option, you pay for the entire term of the reservation in one upfront payment.

If you have an account with a successful billing history, you can choose the No Upfront option. The entire value of the reservation is spread across every hour in the term and you will be billed for every hour in the term, regardless of usage.

The Partial Upfront payment option is a hybrid of the All Upfront and No Upfront options. You make a small upfront payment, and you are billed a low hourly rate for every hour in the term regardless of usage.

Q: When are Reserved Instances activated?

The billing discount and capacity reservation is activated once your payment has successfully been authorized. You can view the status (pending | active | retired) of your reservations on the "Reserved Instances" page of the Amazon EC2 Console.

Q: Can I use my Reserved Instances with Windows to run a Windows with SQL Standard Server AMI?

Yes. Reservations for instances running Microsoft Windows Server and Microsoft SQL Server are available in every region. To get pricing information and additional details, please visit the Amazon EC2 Running Microsoft Windows Server & SQL Server page.

Q: How do Reserved Instances work with Consolidated Billing?

The account you use to purchase Reserved Instances will receive the capacity reservation. Our system automatically optimizes which instances are charged at the lower rate to ensure that the payer account always pays the lowest amount.

In terms of volume discount tiers, if you leverage Consolidated Billing, AWS will use the aggregate total list price of active reservations across all of your consolidated accounts to determine which volume discount tier to apply. Volume discount tiers are determined at the time of purchase, so you should activate Consolidated Billing prior to purchasing Reserved Instances to ensure that you benefit from the largest possible volume discount tier that your consolidated accounts are eligible to receive.

Q: How do the volume discount tiers work?

When you purchase Reserved Instances in a region, and their values adds up to a value determined by AWS, you automatically receive discounts on your upfront fees and hourly fees for future purchases of Reserved Instances in that region.

These discounts are determined based on the total list value (non-discounted price) of upfront fees for the active reservations you have per region. Your total list value is the sum of all expected payments for a reservation within the term, including both the upfront and recurring hourly payments. The following are the volume discount tiers:

- $0-$500K: Upfront - 0%, Hourly - 0%
- $500K - $4M: Upfront - 5%, Hourly - 5%
- $4M - $10M: Upfront - 10%, Hourly - 10%
- $10M+: Contact Us

When you have active Reserved Instances with a list value totaling more than $500,000 in a single region, you will automatically receive a 5% discount on both upfront and hourly fees for all future purchases in that region. Discounts will continue to apply to new reservations as long as you continue to qualify for this volume discount tier.

To illustrate, let's assume you currently have $400,000 worth of active Reserved Instances in us-east-1. You want to purchase 75 Reserved instances with a list value of $2000 each. That would be a total of $150,000 without any discount tiers.

The first $100,000 of this purchase would be discounted at 0 percent. The remaining $50,000 of this purchase would be discounted by 5 percent, so you would only be charged $47,500 over the term for the purchase, and you would pay discounted hourly fees on those reservations.

To learn more about volume discount tiers, please visit Understanding Reserved Instance Discount Pricing Tiers.

Q: Do Convertible Reserved Instances qualify for Volume Discounts?

No, however the value of each Convertible Reserved Instance that you purchase contributes to your volume discount standing.

Q: How do I calculate the list value of an Reserved Instance?

Here is a sample list value calculation for 3yr Partial Upfront Reserved Instances:

3yr Partial Upfront Volume Discount Value in US-East

	Upfront $	Recurring Hourly $	Recurring Hourly Value	List Value
m3.xlarge	$ 1,345	$ 0.060	$ 1,577	$ 2,922
c3.xlarge	$ 1,016	$ 0.045	$ 1,183	$ 2,199

- Assume 26,280 Hours in a 3yr Term
- Recurring Hourly Value = Recurring Hourly $ * Hours in Term
- List Value = Upfront $ + Recurring Hourly Value

Q: I receive purchasing discounts for my Reserved Instances, will I also receive volume discounts?

No. Discounts based on volume tiers are not cumulative with other discounts for Reserved Instance purchases.

Q: Will the cost of my Reserved Instances change, if my future volume qualifies me for other discount tiers?

Volume discounts are determined at the time of purchase. New purchases will be discounted according to your eligible, volume discount tier. Reserved Instances are billed at the same rate for the duration of their term.

For example, if you have $520K worth of Reserved Instances, and sell reservations worth $50k in the Reserved Instance Marketplace, you would continue to pay the discounted rate for the remaining $470K worth of reservations for the duration of the term. If you have $470K worth of reservations and purchase an additional $50K worth, you would receive a volume tier discount on all Reserved Instances over $500K.

Q: Will Amazon RDS purchases count toward Amazon EC2 volume discount tiers (and vice versa)?

No. Only Amazon EC2 Reserved Instances purchases apply towards the Amazon EC2 volume discount tiers.

Q: What do I need to do at purchase time to receive volume discounts?

No action is required on your part. You will automatically receive volume discounts when you use the existing PurchaseReservedInstance API or EC2 Management Console interface to purchase Reserved Instances. If you purchase more than $10M worth of Reserved Instances, contact us about receiving discounts beyond those that are automatically provided.

Q: How do I determine which volume discount tier applies to me?

To determine your current volume discount tier, please consult the Understanding Reserved Instance Discount Pricing Tiers portion of the Amazon EC2 User Guide.

Q: I have purchased a Reserved Instance for an instance type that is available as an EBS-Optimized instance. Can I re-launch that instance as an EBS-Optimized instance? Do I still get the lowe rate?

If you already own a reservation for an instance type that supports EBS-Optimization, you can re-launch the instance as an EBS-Optimized instance. You will pay the additional hourly charge for EBS-Optimization, in addition to your hourly instance cost.

Convertible Reserved Instances

Q: What is a Convertible Reserved Instance?
A Convertible Reserved Instance is a type of Reserved Instance with attributes that can be changed during the term.

Q: When should I purchase a Convertible Reserved Instance instead of a Standard Reserved Instance?
The Convertible Reserved Instance is useful for customers who can commit to using EC2 instances for a 3-year term in exchange for a significant discount on their EC2 usage, are uncertain about their instance needs in the future, or want to benefit from changes in price.

Q: Can I exchange my Convertible Reserved Instance to benefit from a Convertible Reserved Instance matching a different instance type, operating system, tenancy, or payment option?
Yes, you can select a new instance type, operating system, tenancy, or payment option when you exchange your Convertible Reserved Instances.

Q: Can I transfer a Convertible or Standard Reserved Instance from one region to another?
No, a Reserved Instance is associated with a specific region, which is fixed for the duration of the reservation's term.

Q: How do I change the configuration of a Convertible Reserved Instance?
You can change the configuration of your Convertible Reserved Instance using the EC2 Management Console or the ExchangeReservedInstance API.

Q: Do I need to pay a fee when I exchange my Convertible Reserved Instances?
No, you do not pay a fee when you exchange your Reserved Instances. However may need to pay a one-time true-up charge that accounts for differences in pricing between the Convertible Reserved Instances that you have and the Convertible Reserved Instances that you want.

Q: Does the end date change when I exchange a Convertible Reserved Instance?
No, the end date of the original Reserved Instance is transferred to the Reserved Instances you receive after the exchange.

Q: How do Convertible Reserved Instance exchanges work?
When you exchange one Convertible Reserved Instance for another, EC2 ensures that the total value of the Convertible Reserved Instances is maintained through a conversion. So, if you are converting your Reserved Instance with a total value of $1000 for another Reserved Instance, you will receive a quantity of Convertible Reserved Instances with a value that’s equal to or greater than $1000. You cannot convert your Convertible Reserved Instance for Convertible Reserved Instance(s) of a lesser total value.

Q: Can you define total value?
The total value is the sum of all expected payments that you’d make during the term for the Reserved Instance.

Q: Can you walk me through how the true-up cost is calculated for a conversion between two All Upfront Convertible Reserved Instances?
Sure, let’s say you purchased an All Upfront Convertible Reserved Instance for $1000 upfront, and halfway through the term you decide to change the attributes of the Reserved Instance. Since you’re halfway through the Reserved Instance term, you have $500 left of prorated value remaining on the Reserved Instance. The All Upfront Convertible Reserved Instance that you want to convert into costs $1,200 upfront today. Since you only have half of the term left on your existing Convertible Reserved Instance, there is $600 of value remaining on the desired new Convertible Reserved Instance. The true-up charge that you’ll pay will be the difference in upfront value between original and desired Convertible Reserved Instances, or $100 ($600 - $500).

Q: Can you walk me through a conversion between No Upfront Convertible Reserved Instances?
Unlike conversions between Convertible Reserved Instances with an upfront value, since you’re converting between Reserved Instances without an upfront cost, there will not be a true-up charge. However, the amount you pay on an hourly basis before the exchange will need to be greater than or equal to the amount you pay on a total hourly basis after the exchange.

For example, let’s say you purchased one No Upfront Convertible Reserved Instance (A) with a $0.10/hr rate, and you decide to exchange Convertible Reserved Instance A for another Reserved Instance (B) that costs $0.06/hr. When you convert, you will receive two Reserved Instances of B because the amount that you pay on an hourly basis must be greater than or equal to the amount you’re paying for A on an hourly basis.

Q: Can I customize the number of instances that I receive as a result of a Convertible Reserved Instance exchange?
No, EC2 uses the value of the Convertible Reserved Instances you’re trading in to calculate the minimal number of Convertible Reserved Instances you’ll receive while ensuring the result of the exchange gives you Convertible RIs of equal or greater value.

Q: Are there exchange limits for Convertible Reserved Instances?
No, there are no exchange limits for Convertible Reserved Instances.

Q: Do I have the freedom to choose any instance type when I exchange my Convertible Reserved Instances?
No, you can only exchange into Convertible Reserved Instances that are currently offered by AWS.

Q: Can I upgrade the payment option associated with my Convertible Reserved Instance?
Yes, you can upgrade the payment option associated with your Reserved Instance. For example, you can exchange your No Upfront Reserved Instances for Partial or All Upfront Reserved Instances to benefit from better pricing. You cannot change the payment option from All Upfront to No Upfront, and cannot change from Partial Upfront to No Upfront.

Q: Do Convertible Reserved Instances allow me to benefit from price reductions when they happen?
Yes, you can exchange your Reserved Instances to benefit from lower pricing. For example, if the price of new Convertible Reserved Instances reduces by 10%, you can exchange your Convertible RIs and benefit from the 10% reduction in price.

Reserved Instance Marketplace

Q. What is the Reserved Instance Marketplace?

The Reserved Instance Marketplace is an online marketplace that provides AWS customers the flexibility to sell their Amazon Elastic Compute Cloud (Amazon EC2) Reserved Instances to other businesses and organizations. Customers can also browse the Reserved Instance Marketplace to find an even wider selection of Reserved Instance term lengths and pricing options sold by other AWS customers.

Q. When can I list a Reserved Instance on the Reserved Instance Marketplace?

You can list a Reserved Instance when:

- You've registered as a seller in the Reserved Instance Marketplace.
- You've paid for your Reserved Instance.
- You've owned the Reserved Instance for longer than 30 days.

Q. How will I register as a seller for the Reserved Instance Marketplace?

To register for the Reserved Instance Marketplace, you can enter the registration workflow by selling a Reserved Instance from the EC2 Management Console or setting up your profile from the "Account Settings" page on the AWS portal. No matter the route, you will need to complete the following steps:

1. Start by reviewing the overview of the registration process.
2. Log in to your AWS Account.
3. Enter in the bank account into which you want us to disburse funds. Once you select "Continue", we will set that bank account as the default disbursement option.
4. In the confirmation screen, choose "Continue to Console to Start Listing".

If you exceed $20,000 in sales of Reserved Instances, or plan to sell 50 or more Reserved Instances, you will need to provide tax information before you can list your Reserved Instances. Choose "Continue with Tax Interview". During the tax interview pipeline, you will be prompted to enter your company name, contact name, address, and Tax Identification Number using the TIMS workflow.

Additionally, if you plan to sell Reserved Instances worth more than $50,000 per year you will also need to file a limit increase.

Q. How will I know when I can start selling on the Reserved Instance Marketplace?

You can start selling on the Reserved Instance Marketplace after you have added a bank account through the registration pipeline. Once activation is complete, you will receive a confirmation email. However, it is important to note that you will not be able to receive disbursements until we are able to receive verification from your bank, which may take up to two weeks, depending on the bank you use.

Q. How do I list a Reserved Instance for sale?

To list a Reserved Instance, simply complete these steps in the Amazon EC2 Console:

1. Select the Reserved Instances you wish to sell, and choose "Sell Reserved Instances". If you have not completed the registration process, you will be prompted to register using the registration pipeline.

2. For each Reserved Instance type, set the number of instances you'd like to sell, and the price for the one-time fee you would like to set. Note that you can set the one-time price to different amounts depending on the amount of time remaining so that you don't have to keep adjusting your one-time price if your Reserved Instance doesn't sell quickly. By default you just need to set the current price and we will automatically decrease the one-time price by the same increment each month.

3. Once you have configured your listing, a final confirmation screen will appear. Choose "Sell Reserved Instance".

Q. Which Reserved Instances can I list for sale?

You can list any Reserved Instances that have been active for at least 30 days, and for which we have received payment. Typically, this means that you can list your reservations once they are in the **active** state. It is important to note that if you are an invoice customer, your Reserved Instance can be in the **active** state prior to AWS receiving payment. In this case, your Reserved Instance will not be listed until we have received your payment.

Q. How are listed Reserved Instances displayed to buyers?

Reserved Instances (both third-party and those offered by AWS) that have been listed on the Reserved Instance Marketplace can be viewed in the "Reserved Instances" section of the Amazon EC2 Console. You can also use the DescribeReservedInstancesListings API call.

The listed Reserved Instances are grouped based on the type, term remaining, upfront price, and hourly price. This makes it easier for buyers to find the right Reserved Instances to purchase.

Q. How much of my Reserved Instance term can I list?

You can sell a Reserved Instance for the term remaining, rounded down to the nearest month. For example, if you had 9 months and 13 days remaining, you will list it for sale as a 9-month-term Reserved Instance.

Q. Can I remove my Reserved Instance after I’ve listed it for sale?

Yes, you can remove your Reserved Instance listings at any point until a sale is**pending** (meaning a buyer has bought your Reserved Instance and confirmation of payment is pending).

Q. Which pricing dimensions can I set for the Reserved Instances I want to list?

Using the Reserved Instance Marketplace, you can set an upfront price you’d be willing to accept. You cannot set the hourly price (which will remain the same as was set on the original Reserved Instance), and you will not receive any funds collected from payments associated with the hourly prices.

Q. Can I still use my reservation while it is listed on the Reserved Instance Marketplace?

Yes, you will continue to receive the capacity and billing benefit of your reservation until it is sold. Once sold, any running instance that was being charged at the discounted rate will be charged at the On-Demand rate until and unless you purchase a new reservation, or terminate the instance.

Q. Can I resell a Reserved Instance that I purchased from the Reserved Instance Marketplace?

Yes, you can resell Reserved Instances purchased from the Reserved Instance Marketplace just like any other Reserved Instance.

Q. Are there any restrictions when selling Reserved Instances?

Yes, you must have a US bank account to sell Reserved Instances in the Reserved Instance Marketplace. Support for non-US bank accounts will be coming soon. Also, you may not sell Reserved Instances in the US GovCloud region.

Q. Can I sell Reserved Instances purchased from the public volume pricing tiers?

No, this capability is not yet available.

Q. Is there a charge for selling Reserved Instances on the Reserved Instance Marketplace?

Yes, AWS charges a service fee of 12% of the total upfront price of each Reserved Instance you sell in the Reserved Instance Marketplace.

Q. Can AWS sell subsets of my listed Reserved Instances?

Yes, AWS may potentially sell a subset of the quantity of Reserved Instances that you have listed. For example, if you list 100 Reserved instances, we may only have a buyer interested in purchasing 50 of them. We will sell those 50 instances and continue to list your remaining 50 Reserved Instances until and unless you decide not to list them any longer.

Q. How do buyers pay for Reserved Instances that they've purchased?

Payment for completed Reserved Instance sales are done via ACH wire transfers to a US bank account.

Q. When will I receive my money?

Once AWS has received funds from the customer that has bought your reservation, we will disburse funds via wire transfer to the bank account you specified when you registered for the Reserved Instance Marketplace.

Then, we will send you an email notification letting you know that we’ve wired you the funds. Typically, funds will appear in your account within 3-5 days of when your Reserved Instance was been sold.

Q. If I sell my Reserved Instance in the Reserved Instance Marketplace, will I get refunded for the Premium Support I was charged too?

No, you will not receive a pro-rated refund for the upfront portion of the AWS Premium Support Fee.

Q. Will I be notified about Reserved Instance Marketplace activities?

Yes, you will receive a single email once a day that details your Reserved Instance Marketplace activity whenever you create or cancel Reserved Instance listings, buyers purchase your listings, or AWS disburses funds to your bank account.

Q. What information is exchanged between the buyer and seller to help with the transaction tax calculation?

The buyer’s city, state, zip+4, and country information will be provided to the seller via a disbursement report. This information will enable sellers to calculate any necessary transaction taxes they need to remit to the government (e.g., sales tax, value-added tax, etc.). The legal entity name of the seller will also be provided on the purchase invoice.

Q. Are there any restrictions on the customers when purchasing third-party Reserved Instances?

Yes, you cannot purchase your own listed Reserved Instances, including those in any of your linked accounts (via Consolidated Billing).

Q. Do I have to pay for Premium Support when purchasing Reserved Instances from the Reserved Instance Marketplace?

Yes, if you are a Premium Support customer, you will be charged for Premium Support when you purchase a Reserved Instance through the Reserved Instance Marketplace.

Spot Instances

Q. What is a Spot Instance?

Spot instances are a new way to purchase and consume Amazon EC2 Instances. They allow customers to bid on unused EC2 capacity and run those instances for as long as their bid exceeds the current Spot Price. The Spot Price changes periodically based on supply and demand, and customers whose bids meet or exceed it gain access to the available Spot instances. Spot instances are complementary to On-Demand instances and Reserved Instances, providing another option for obtaining compute capacity.

Q. How is a Spot instance different than an On-Demand instance or Reserved Instance?

Spot instances provide the ability for customers to purchase compute capacity with no upfront commitment, at hourly rates usually lower than the On-Demand rate. Spot instances allow you to specify the maximum hourly price that you are willing to pay to run a particular instance type. Amazon EC2 sets a Spot Price for each instance type in each availability zone, which is the hourly price all customers will pay to run a Spot instance for that given period. The Spot Price fluctuates based on supply and demand for instances, but customers will never pay more than the maximum price they have specified. If the Spot Price moves higher than a customer’s maximum price, the customer’s instance will be shut down by Amazon EC2. Other than those differences, Spot instances perform exactly the same as On-Demand or Reserved Instances. See here for more details on Spot instances.

Q. How do I purchase and start up a Spot instance?

Spot instances can be requested using the EC2 Management Console or Amazon EC2 APIs. To start with the EC2 Management Console:

1. Log in to the EC2 Management Console.
2. Choose "Spot Requests" in the left navigation pane.
3. Choose "Request Spot Instances".
4. Complete the Launch Instance Wizard process, choosing an AMI, region and instance size and type.

5. Enter the number of Spot instances you would like to request, your maximum price, and whether the request is persistent or not.
6. After choosing your key pair and security group(s), you are ready to submit your Spot instance request.

For detail on how to request Spot instances through the Amazon EC2 API, see theAmazon EC2 API Reference.

For a more detailed walk-through of using Spot instances and more information on how to get the most out of Spot instances, see Introduction to Spot Instances.

Q. How many Spot instances can I request?

You are limited to requesting Spot instances per your dynamic Spot limit for each region. Note that not all instance types are available on Spot, and new AWS accounts might start with a lower limit. To learn more about Spot instance limits, please refer to the Amazon EC2 User Guide.

If you would like a higher limit, complete the Amazon EC2 instance request form with your use case and your instance increase will be considered. Limit increases are tied to the region they were requested for.

Q. How can I determine the status of my Spot request?

You can determine the status of your Spot request in the instance provisioning lifecycle by inspecting its Spot Bid Status code and message. By reviewing Spot bid statuses, you can see why your Spot requests state has or has not changed and you can learn how to optimize your Spot requests to get them fulfilled. You can access Spot Bid Status information on the **Spot Instance** page of the EC2 console of the AWS Management Console, as well as through the DescribeSpotInstanceRequests API action and the ec2-describe-spot-instance-requests CLI command. For more information, please visit the Amazon EC2 Developer guide.

Q. Are Spot instances available for all instance families and sizes and in all regions?

Instance types supported in each region are listedhere. Spot instance APIs are available in all regions except the US GovCloud region.

Q. Which operating systems are available as Spot instances?

Linux/UNIX and Windows Server are available. Windows Server with SQL Server is not currently available.

Q. Are there any features or services of Amazon Web Services that are not supported for use with Spot instances?

Amazon DevPay is not supported for use with Spot instances.

Q. Can I use a Spot instance with a paid AMI for third-party software (such as IBM's software packages)?

Not at this time.

Q. Will I be charged if my Spot instance is terminated by Amazon EC2 before the hour is up?

No. If the Spot instance is terminated by Amazon EC2, you will not be charged for a partial hour of usage. However, if you terminate the instance yourself, you will be charged for any hour in which the instance ran.

Q. How often should I expect the Spot price to change?

Amazon EC2 will change the Spot price periodically as new requests are received and as available Spot capacity changes (e.g., due to instance terminations). While the Spot price may change anytime, in general it will change once per hour and in many cases less frequently. We publish the current Spot price and historical prices for Spot instances through the API, and they can also be viewed using the AWS Management Console. This can help you assess the levels and timing of fluctuations in the Spot price over time.

Q. Will all Spot instances started at the same time be charged the same price?

Yes.

Q. Will the price I'm charged for a running Spot instance change during its instance-hour as the Spot price changes?

No. The price per instance-hour for a Spot instance is set at the beginning of each instance-hour for the entire hour. Any changes to the Spot price will not be reflected until the next instance-hour begins.

Q. Where can I see my usage history for Spot instances and see how much I was billed?

The AWS Management Console makes a detailed billing report available which shows Spot instance start and termination times for all instances. Customers can check the billing report against historical Spot prices via the API to verify that the Spot price they were billed is correct.

Q. Why do Spot prices differ across accounts for the same instance type, operating system, and Availability Zone?

To ensure that resources are distributed across Availability Zones for a region, Availability Zones are independently mapped to identifiers for each account. For example, your Availability Zone us-east-1a might not be the same location as us-east-1a for another account. So, Spot prices for the same Availability Zone identifier may be different in different accounts. Note that there's no way for you to coordinate Availability Zones between accounts.

Q. What is a Spot fleet?

A Spot fleet allows you to automatically bid on and manage multiple Spot instances that provide the lowest price per unit of capacity for your cluster or application, like a batch processing job, a Hadoop workflow, or an HPC grid computing job. You can include the instance types that your application can use, and define a target capacity based on your application needs (in units including instances, vCPUs, memory, storage, or network throughput). Spot fleets enable you to launch and maintain the target capacity, and to automatically request resources to replace any that are disrupted or manually terminated. Learn more about Spot fleets.

Q. Is there any additional charge for making Spot fleet requests?

No, there is no additional charge for Spot fleet requests.

Q. What limits apply to a Spot fleet request?

Visit the Spot Fleet Limits section of the Amazon EC2 User Guide to learn about the limits that apply to your Spot fleet request.

Q. What happens if my Spot fleet request tries to launch Spot instances but exceeds my regional Spot request limit?

If your Spot fleet request exceeds your regional Spot instance request limit, individual Spot instance requests will fail with a **Spot request limit exceeded** bid status. Your Spot fleet request's history will show any Spot request limit errors that the fleet request received. Visit the Monitoring Your Spot Fleet section of the Amazon EC2 User Guide to learn how to describe your Spot fleet request's history.

Q. What happens if my Spot fleet request bid price exceeds my Spot bid price limit for

one of the instance types I am requesting?

If your Spot fleet request bid price exceeds your Spot bid price limits, we will submit Spot requests for that instance type at your current Spot bid price limit. Your Spot fleet request's history will show if any of your fleet's instances were affected by your Spot bid price limit. Visit the Monitoring Your Spot Fleet section of the Amazon EC2 User Guide to learn how to describe your Spot fleet request's history.

Q. Are Spot fleet requests guaranteed to be fulfilled?

No. Spot fleet requests allow you to place multiple Spot instance bids simultaneously, and are subject to the same availability and prices as a single Spot instance request. For example, if no resources are available at your Spot fleet request bid price, we may be unable to fulfill your request partially or in full.

Q. Can I submit a multi-Availability Zone fleet request?

Yes, visit the Spot Fleet Examples section of the Amazon EC2 User Guide to learn how to submit a multi-Availability Zone Spot fleet request.

Q. Can I submit a multi-region Spot fleet request?

No, we do not support multi-region fleet requests.

Q. How does Spot fleet allocate resources across the various Spot instance pools specified in the launch specifications?

The RequestSpotFleet API provides two allocation strategies: **lowestPrice** and **diversified**. The lowestPrice strategy allows you to provision your Spot fleet resources in instance pools that provide the lowest price per unit of capacity at the time of the request. The **diversified** strategy allows you to provision your Spot fleet resources across multiple Spot instance pools. This enables you to maintain your fleet's target capacity and increase your application's availability as Spot capacity fluctuates.

Running your application's resources across diverse Spot instance pools also allows you to further reduce your fleet's operating costs over time. Visit the Amazon EC2 User Guide to learn more.

Q. Can I tag a Spot fleet request?

We currently do not support tagging Spot fleet requests.

Q. How can I see which Spot fleet owns my Spot instances?

You can identify the Spot instances associated with your Spot fleet by describing your fleet request. Fleet requests are available for 48 hours after all its Spot instances have been terminated. See the Amazon EC2 User Guide to learn how to describe your Spot fleet request.

Q. Can I modify my Spot fleet request?

Currently, you can only modify the target capacity of your Spot fleet request. You may need to cancel the request and submit a new one to change other request configuration parameters.

Q. Can I specify a different AMI for each instance type that I want to use?

Yes, simply specify the AMI you'd like to use in each launch specification you provide in your Spot fleet request.

Q. Can I use Spot fleet with Elastic Load Balancing, Auto Scaling, or Elastic MapReduce?

No, Elastic Load Balancing, Auto Scaling, or Elastic MapReduce do not directly trigger Spot fleet requests.

Q. Does a Spot fleet request terminate Spot instances when they are no longer running in the lowest priced Spot pools and relaunch them in the lowest priced pools?

No, Spot fleet requests do not automatically terminate and re-launch instances while they are running. However, if you terminate a Spot instance, Spot fleet will replenish it with a new Spot instance in the new lowest priced pool.

Q: Are Spot blocks (Fixed Duration Spot instances) ever interrupted?

Spot blocks are designed not to be interrupted and will run continuously for the duration you select, independent of Spot market price. In rare situations, Spot blocks may be interrupted due to AWS capacity needs. In these cases, we will provide a two-minute warning before we terminate your instance (termination notice), and you will not be charged for the affected instance(s).

Micro Instances

Q. How much compute power do Micro instances provide?

Micro instances provide a small amount of consistent CPU resources and allow you to burst CPU capacity up to 2 ECUs when additional cycles are available. They are well suited for lower throughput applications and web sites that consume significant compute cycles periodically but very little CPU at other times for background processes, daemons, etc. Learn more about use of this instance type.

Q. How does a Micro instance compare in compute power to a Standard Small instance?

At steady state, Micro instances receive a fraction of the compute resources that Small instances do. Therefore, if your application has compute-intensive or steady state needs we recommend using a Small instance (or larger, depending on your needs). However, Micro instances can periodically burst up to 2 ECUs (for short periods of time). This is double the number of ECUs available from a Standard Small instance. Therefore, if you have a relatively low throughput application or web site with an occasional need to consume significant compute cycles, we recommend using Micro instances.

Q. How can I tell if an application needs more CPU resources than a Micro instance is providing?

The CloudWatch metric for CPU utilization will report 100% utilization if the instance bursts so much that it exceeds its available CPU resources during that CloudWatch monitored minute. CloudWatch reporting 100% CPU utilization is your signal that you should consider scaling – manually or via Auto Scaling – up to a larger instance type or scale out to multiple Micro instances.

Q. Are all features of Amazon EC2 available for Micro instances?

Currently Amazon DevPay is not available for Micro instances.

Compute-Optimized Instances

Q. When should I use Compute-optimized instances?

Compute-optimized instances are designed for applications that benefit from high compute power. These applications include high performance front-end fleets, web-servers, batch processing, distributed analytics, high performance science and engineering applications, ad serving, MMO gaming, video-encoding, and distributed analytics.

Q. Can I launch C4 instances as Amazon EBS-optimized instances?

Each C4 instance type is EBS-optimized by default. C4 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on C4 instances, launching a C4 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. How can I use the processor state control feature available on the c4.8xlarge instance?

The c4.8xlarge instance type provides the ability for an operating system to control processor C-states and P-states. This feature is currently available only on Linux instances. You may want to change C-state or P-state settings to increase processor performance consistency, reduce latency, or tune your instance for a specific workload. By default, Amazon Linux provides the highest-performance configuration that is optimal for most customer workloads; however, if your application would benefit from lower latency at the cost of higher single- or dual-core frequencies, or from lower-frequency sustained performance as opposed to bursty Turbo Boost frequencies, then you should consider experimenting with the C-state or P-state configuration options that are available to these instances. For additional information on this feature, see the Amazon EC2 User Guide section on Processor State Control.

Accelerated Computing Instances

Q: What are Accelerated Computing Instances?

Accelerated Computing Instance family is a family of instances which use hardware accelerators, or co-processors, to perform some functions, such as floating point number calculation and graphics processing, more efficiently than is possible in software running on CPUs. Amazon EC2 provides two types of Accelerated Computing Instances – GPU Compute Instances for general-purpose computing and GPU Graphics Instances for graphics intensive applications.

Q. When should I use GPU Graphics and Compute instances?

GPU instances work best for applications with massive parallelism, for example workloads using thousands of threads. Graphics processing is an example with huge computational requirements, where each of the tasks is relatively small, the set of operations performed form a pipeline, and the throughput of this pipeline is more important than the latency of the individual operations. To be able build applications that exploit this level of parallelism one needs GPU device specific knowledge by understanding how to program against various graphics APIs (DirectX, OpenGL) or GPU compute programming models (CUDA, OpenCL).

Q. How are G2 instances different from CG1 instances?

CG1 instances use NVIDIA Tesla GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. CG1 instances provide customers with high bandwidth 10 Gbps networking, double precision floating-point capabilities, and error-correcting code (ECC) memory, making them ideal for High Performance Computing (HPC) applications. G2 instances use NVIDIA GRID GPUs and provide a cost-effective, high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA GRID GPUs also support NVIDIA's fast capture and encode APIs. Example applications include video creation services, 3D visualizations, streaming graphics-intensive applications, and other server-side workloads requiring massive parallel processing power. In addition, Graphics instances can also be used for general purpose computing using CUDA or OpenCL, but are not recommended for network-intensive HPC applications.

Q. How are P2 instances different from G2 instances?

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 20Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory, making them ideal for deep learning, high performance databases, computational fluid dynamics, computational finance, seismic analysis, molecular modeling, genomics, rendering, and other server-side GPU compute workloads. G2 instances use NVIDIA GRID GPUs and provide a cost-effective, high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA GRID GPUs also support NVIDIA's fast capture and encode APIs. Example applications include video creation services, 3D visualizations, streaming graphics-intensive applications, and other server-side graphics workloads.

Q. What APIs and programming models are supported by GPU Graphics and Compute instances?

With the initial driver release, G2 instances support DirectX 9, 10, and 11, OpenGL 4.3, CUDA 5.5, OpenCL 1.1, and DirectCompute. With the latest driver release, CG1 instances support CUDA 5.5, OpenCL 1.1, and DirectCompute. With the latest driver release, P2 instances support CUDA 7.5 and OpenCL 1.2.

Q. Where do I get NVIDIA drivers for CG1, G2 and P2 instances?

There are two methods by which NVIDIA drivers may be obtained. NVIDIA has listings on the AWS Marketplace which offer Amazon Linux AMIs and Windows Server AMIs with the NVIDIA drivers pre-installed. You may also launch 64 bit, HVM AMIs and install the drivers yourself. You must visit the NVIDIA drivers website and search for the NVIDIA Tesla K80 for the P2, NVIDIA GRID K520 for the G2, and the Tesla M2050 for the CG1.

Q. Which AMIs can I use with P2 and G2 instances?

You can currently use Windows Server, SUSE Enterprise Linux, Ubuntu, and Amazon Linux AMIs on P2 and G2 instances. If you want to launch AMIs with operating systems not listed here, contact AWS Customer Support with your request or reach out through EC2 Forums.

Q. Where do I get the NVIDIA GRID SDK?

The NVIDIA GRID SDK is available from NVIDIA directly. Please visit http://www.nvidia.com/object/cloud-get-started.html for information about obtaining the full SDK. NVENC, the frame capture and encoding portion of the GRID SDK, is available on the NVIDIA Developers Zone at https://developer.nvidia.com/nvidia-video-codec-sdk.

Q. Does the use of G2 instances require third-party licenses?

Aside from the NVIDIA drivers and GRID SDK, the use of G2 instances does not necessarily require any third-party licenses. However, you are responsible for determining whether your content or technology used on G2 instances requires any additional licensing. For example, if you are streaming content you may need licenses for some or all of that content. If you are using third-party technology such as operating systems, audio and/or video encoders, and decoders from Microsoft, Thomson, Fraunhofer IIS, Sisvel S.p.A., MPEG-LA, and Coding Technologies, please consult these providers to determine if a license is required. For example, if you leverage the on-board h.264 video encoder on the NVIDIA GRID GPU you should reach out to MPEG-LA for guidance, and if you use mp3 technology you should contact Thomson for guidance..

Q. Why am I unable to see the GPU when using Microsoft Remote Desktop?

When using Remote Desktop, GPUs using the WDDM driver model are replaced with a non-accelerated Remote Desktop display driver. In order to access your GPU hardware, you need to utilize a different remote access tool, such as VNC.

Cluster Instances

Q. What is a Cluster Compute Instance?

Cluster Compute Instances combine high compute resources with a high performance networking for High Performance Compute (HPC) applications and other demanding network-bound applications. Cluster Compute Instances provide similar functionality to other Amazon EC2 instances but have been specifically engineered to provide high performance networking.

Amazon EC2 cluster placement group functionality allows users to group Cluster Compute Instances in clusters – allowing applications to get the low-latency network performance necessary for tightly-coupled node-to-node communication typical of many HPC applications. Cluster Compute Instances also provide significantly increased network throughput both within the Amazon EC2 environment and to the Internet. As a result, these instances are also well suited for customer applications that need to perform network-intensive operations.

Learn more about use of this instance type for HPC applications.

Q. What kind of network performance can I expect when I launch instances in cluster placement group?

The bandwidth an EC2 instance can utilize in a cluster placement group depends on the instance type and its networking performance specification. When launched in a placement group, select EC2 instances can utilize up to 10 Gbps for single-flow and 20 Gbps for multi-flow traffic in each direction (full duplex). Network traffic outside a cluster placement group (e.g. to the Internet) is limited to 5 Gbps (full duplex).

Q. What is a Cluster GPU Instance?

Cluster GPU Instances provide general-purpose graphics processing units (GPUs) with proportionally high CPU and increased network performance for applications benefiting from highly parallelized processing that can be accelerated by GPUs using the CUDA and OpenCL programming models. Common applications include modeling and simulation, rendering and media processing.

Cluster GPU Instances give customers with HPC workloads an option beyond Cluster Compute Instances to further customize their high performance clusters in the cloud for applications that can benefit from the parallel computing power of GPUs.

Cluster GPU Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters – allowing applications to get the low-latency, high bandwidth network performance required for tightly-coupled node-to-node communication typical of many HPC applications.

Learn more about HPC on AWS.

Q. What is a High Memory Cluster Instance?

High Memory Cluster Instances provide customers with large amounts of memory and CPU capabilities per instance in addition to high network capabilities. These instance types are ideal for memory intensive workloads including in-memory analytics systems, graph analysis and many science and engineering applications

High Memory Cluster Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters – allowing applications to get the low-latency, high bandwidth network performance required for tightly-coupled node-to-node communication typical of many HPC and other network intensive applications.

Q. Does use of Cluster Compute and Cluster GPU Instances differ from other Amazon EC2 instance types?

Cluster Compute and Cluster GPU Instances use differs from other Amazon EC2 instance types in two ways.

First, Cluster Compute and Cluster GPU Instances use Hardware Virtual Machine (HVM) based virtualization and run only Amazon Machine Images (AMIs) based on HVM virtualization. Paravirtual Machine (PVM) based AMIs used with other Amazon EC2 instance types cannot be used with Cluster Compute or Cluster GPU Instances.

Second, in order to fully benefit from the available low latency, full bisection bandwidth between instances, Cluster Compute and Cluster GPU Instances must be launched into a cluster placement group through the Amazon EC2 API or AWS Management Console.

Q. What is a cluster placement group?

A cluster placement group is a logical entity that enables creating a cluster of instances by launching instances as part of a group. The cluster of instances then provides low latency, full bisection 10 Gigabit Ethernet bandwidth connectivity between instances in the group. Cluster placement groups are created through the Amazon EC2 API or AWS Management Console.

Q. Are all features of Amazon EC2 available for Cluster Compute and Cluster GPU Instances?

Currently, Amazon DevPay is not available for Cluster Compute or Cluster GPU Instances.

Q. Is there a limit on the number of Cluster Compute or Cluster GPU Instances I can use and/or the size of cluster I can create by launching Cluster Compute Instances or Cluster GPU into a cluster placement group?

There is no limit specific for Cluster Compute Instances. For Cluster GPU Instances, you can launch 2 Instances on your own. If you need more capacity, please complete the Amazon EC2 instance request form (selecting the appropriate primary instance type).

Q. Are there any ways to optimize the likelihood that I receive the full number of instances I request for my cluster via a cluster placement group?

We recommend that you launch the minimum number of instances required to participate in a cluster in a single launch. For very large clusters, you should launch multiple placement groups, e.g. two placement groups of 128 instances, and combine them to create a larger, 256 instance cluster.

Q. Can Cluster GPU and Cluster Compute Instances be launched into a single cluster placement group?

While it may be possible to launch different cluster instance types into a single placement group, at this time we only support homogenous placement groups.

Q. If an instance in a cluster placement group is stopped then started again, will it maintain its presence in the cluster placement group?

Yes. A stopped instance will be started as part of the cluster placement group it was in when it stopped. If capacity is not available for it to start within its cluster placement group, the start will fail.

High I/O Instances

Q. What is a High I/O instance?

High I/O instances use SSD-based local instance storage to deliver very high, low latency, I/O capacity to applications, and are optimized for applications that require tens of thousands of IOPS. Like Cluster instances, High I/O instances can be clustered via cluster placement groups for high bandwidth networking.

Q. Are all features of Amazon EC2 available for High I/O instances?

High I/O instance support all Amazon EC2 features with the exception of Spot Instances. Currently you can only purchase High I/O instances as On-Demand or Reserved Instances.

Q. Is there a limit on the number of High I/O instances I can use?

Currently, you can launch 2 hi1.4xlarge instances by default. If you wish to run more than 2 On-Demand instances, please complete the Amazon EC2 instance request form.

Q. How many IOPS can hi1.4xlarge instances deliver?

Using Linux PV AMIs, High I/O instances can deliver more than 120,000 4K random read IOPS and 10,000-85,000 4K random write IOPS (depending on active LBA span) to applications across 2 * 1 TiB data volumes. For HVM and Windows AMIs, performance will be around 90,000 4K random read IOPS and 9,000-75,000 4K random write IOPS.

Q. What is the sequential throughput of hi1.4xlarge instances?

Sequential throughput on all AMI types (Linux PV, Linux HVM and Windows) is approximately 2 GB/s read and 1.1 GB/s write.

Q. AWS has other database and Big Data offerings. When or why should I use High I/O instances?

High I/O instances are ideal for applications that require access to tens of thousands of low latency IOPS, and can leverage data stores and architectures that manage data redundancy and availability. Example applications are:

- NoSQL databases like Cassandra and MongoDB
- Clustered databases
- OLTP systems

Q. Do High I/O instances provide any failover mechanisms or redundancy?

Like other Amazon EC2 instance types, instance storage on hi1.4xlarge instances persists during the life of the instance. Customers are expected to build resilience into their applications. We recommend using databases and file systems that support redundancy and fault tolerance. Customers should back up data periodically to Amazon S3 for improved data durability.

Q. Do High I/O instances support TRIM?

The TRIM command allows the operating system to inform SSDs which blocks of data are no longer considered in use and can be wiped internally. In the absence of TRIM, future write operations to the involved blocks can slow down significantly. Currently hi1.4xlarge instances do not support TRIM, but TRIM support will be deployed within the next few months. Customers with extremely intensive full LBA random write workloads should plan accordingly. Please note that the current disk provisioning scheme for High I/O instances minimizes the impact of write amplification and most customers will not experience any issues.

Burstable Performance Instances

Q: How are Burstable Performance Instances different?

Amazon EC2 allows you to choose between Fixed Performance Instances (e.g. M3, C3, and R3) and Burstable Performance Instances (e.g. T2). Burstable Performance Instances provide a baseline level of CPU performance with the ability to burst above the baseline. T2 instances are for workloads that don’t use the full CPU often or consistently, but occasionally need to burst.

T2 instances’ baseline performance and ability to burst are governed by CPU Credits. Each T2 instance receives CPU Credits continuously, the rate of which depends on the instance size. T2 instances accrue CPU Credits when they are idle, and use CPU credits when they are active. A CPU Credit provides the performance of a full CPU core for one minute. The following table shows the maximum credit balance and baseline performance for each T2 instance size. Each vCPU of a T2 instance can consume CPU Credits at a maximum rate of 60 per hour when bursting to full core performance.

Model	vCPUs	CPU Credits / hour	Maximum CPU Credit Balance	Baseline CPU Performance
t2.nano	1	3	72	5% of a core
t2.micro	1	6	144	10% of a core
t2.small	1	12	288	20% of a core
t2.medium	2	24	576	40% of a core*
t2.large	2	36	864	60% of a core**

** For the t2.medium, single threaded applications can use 40% of 1 core, or if needed, multithreaded applications can use 20% each of 2 cores.*

***For the t2.large, single threaded applications can use 60% of 1 core, or if needed, multithreaded applications can use 30% each of 2 cores.*

For example, a t2.small instance receives credits continuously at a rate of 12 CPU Credits per hour. This capability provides baseline performance equivalent to 20% of a CPU core. If at any moment the instance does not need the credits it receives, it stores them in its CPU Credit balance for up to 24 hours. If and when your t2.small needs to burst to more than 20% of a core, it draws from its CPU Credit balance to handle this surge seamlessly. Over time, if you find your workload needs more CPU Credits that you have, or your instance does not maintain a positive CPU Credit balance, we recommend either a larger T2 size, such as the t2.medium, or a Fixed Performance Instance type.

Many applications such as web servers, developer environments and small databases don't need consistently high levels of CPU, but benefit significantly from having full access to very fast CPUs when they need them. T2 instances are engineered specifically for these use cases. If you need consistently high CPU performance for applications such as video encoding, high volume websites or HPC applications, we recommend you use Fixed Performance Instances. T2 instances are designed to perform as if they have dedicated high-speed Intel cores available when your application really needs CPU performance, while protecting you from the variable performance or other common side effects you might typically see from over-subscription in other environments.

Q. How do I choose the right Amazon Machine Image (AMI) for my t2.nano instances?

T2.nano, our smallest Burstable Performance Instance size, offers 512 MiB of memory and is designed to offer the full performance of a high frequency Intel CPU core as long as you maintain a CPU credit balance. Your t2.nano maintains a positive credit balance if your workload utilizes less than 5% of the core on average over 24 hours. If your workload uses more than 5% CPU on average, consider a larger t2 instance size, such as the t2.micro. You will want to verify that the minimum memory requirements of your operating system and applications are within 512 MiB. Operating systems with Graphical User Interfaces (GUI) that consume significant memory and CPU, for example Microsoft Windows, might need a t2.micro or larger instance size for many use cases. You can find AMIs suitable for the t2.nano instance type on AWS Marketplace. Windows customers who do not need the GUI can use the Microsoft Windows Server 2012 R2 Core AMI.

Q: When should I choose a Burstable Performance Instance, such as T2?

Workloads ideal for Burstable Performance Instances (e.g., web servers, developer environments, and small databases) don’t use the full CPU often or consistently, but occasionally need to burst. If your application requires sustained high CPU performance, we recommend our Fixed Performance Instances, such as M3, C3, and R3.

Q: How can I see the CPU Credit balance for each T2 instance?

You can see the CPU Credit balance for each T2 instance in EC2 per-Instance metrics in Amazon CloudWatch. T2 instances have two new metrics, CPUCreditUsage and CPUCreditBalance. CPUCreditUsage indicates the amount of CPU Credits used. CPUCreditBalance indicates the balance of CPU Credits.

Q: What happens to CPU performance if my T2 instance is running low on credits (CPU Credit balance is near zero)?

If your T2 instance has a zero CPU Credit balance, performance will remain at baseline CPU performance. For example, the t2.micro provides baseline CPU performance of 10% of a physical CPU core. If your instance’s CPU Credit balance is approaching zero, CPU performance will be lowered to baseline performance over a 15-minute interval.

Q: Does my T2 instance credit balance persist a stop / start?

No, a stopped instance does not retain its previously earned credit balance.

Q: Can T2 instances be purchased as Reserved Instances or Spot Instances?

On-Demand instances and Reserved Instances are the only purchase options available for T2 instances.

Q: How is T2 different from the T1?

Compared to the t1.micro, the t2.micro features better CPU performance, more memory, and lower prices. The T2 family also offers more than one size.

Dense-storage Instances

Q. What is a Dense-storage Instance?

Dense-storage instances are designed for workloads that require high sequential read and write access to very large data sets, such as Hadoop distributed computing, massively parallel processing data warehousing, and log processing applications. The Dense-storage instances offer the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q. How do Dense-storage instances compare to High I/O instances?

High I/O instances (I2) are targeted at workloads that demand low latency and high random I/O in addition to moderate storage density and provide the best price/IOPS across other EC2 instance types. Dense-storage instances (D2) are optimized for applications that require high sequential read/write access and low cost storage for very large data sets and provide the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q. How much disk throughput can Dense-storage instances deliver?

The largest current generation of Dense-storage instances, d2.8xlarge, can deliver up to 3.5 GBps read and 3.1 GBps write disk throughput with a 2 MiB block size. To ensure the best disk throughput performance from your D2 instances on Linux, we recommend that you use the most recent version of the Amazon Linux AMI, or another Linux AMI with a kernel version of 3.8 or later that supports persistent grants - an extension to the Xen block ring protocol that significantly improves disk throughput and scalability.

Q. Do Dense-storage instances provide any failover mechanisms or redundancy?

The primary data storage for Dense-storage instances is HDD-based instance storage. Like all instance storage, these storage volumes persist only for the life of the instance. Hence, we recommend that you build a degree of redundancy (e.g. RAID 1/5/6) or use file systems (e.g. HDFS and MapR-FS) that support redundancy and fault tolerance. You can also back up data periodically to more durable data storage solutions such as Amazon Simple Storage Service (S3) for additional data durability. Please refer to Amazon S3 for reference.

Q. How do Dense-storage instances differ from Amazon EBS?

Amazon EBS offers simple, elastic, reliable (replicated), and persistent block level storage for Amazon EC2 while abstracting the details of the underlying storage media in use. Amazon EC2 instance storage provides directly attached non-persistent, high performance storage building blocks that can be used for a variety of storage applications. Dense-storage instances are specifically targeted at customers who want high sequential read/write access to large data sets on local storage, e.g. for Hadoop distributed computing and massively parallel processing data warehousing.

Q. Can I launch D2 instances as Amazon EBS-optimized instances?

Each D2 instance type is EBS-optimized by default. D2 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on D2 instances, launching a D2 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. Are Dense-storage instances offered in EC2 Classic?

The current generation of Dense-storage instances (D2 instances) can be launched in both EC2-

Classic and Amazon VPC. However, by launching a Dense-storage instance into a VPC, you can leverage a number of features that are available only on the Amazon VPC platform – such as enabling enhanced networking, assigning multiple private IP addresses to your instances, or changing your instances' security groups. For more information about the benefits of using a VPC, see Amazon EC2 and Amazon Virtual Private Cloud (Amazon VPC). You can take steps to migrate your resources from EC2-Classic to Amazon VPC. For more information, see Migrating a Linux Instance from EC2-Classic to a VPC.

Memory Optimized Instances

Q. When should I use Memory-optimized instances?
Memory-optimized instances offer large memory size for memory intensive applications including in-memory applications, in-memory databases, in-memory analytics solutions, High Performance Computing (HPC), scientific computing, and other memory-intensive applications.

Q. When should I use X1 instances?
X1 instances are ideal for running in-memory databases like SAP HANA, big data processing engines like Apache Spark or Presto, and high performance computing (HPC) applications. X1 instances are certified by SAP to run production environments of the next-generation Business Suite S/4HANA, Business Suite on HANA (SoH), Business Warehouse on HANA (BW), and Data Mart Solutions on HANA on the AWS cloud.

Q. What are the key specifications of Intel E7 Haswell processors that power X1 instances?
X1 is the first Amazon EC2 instance type that is powered by four 2.3 GHz Intel® Xeon® E7 8880 v3 (Haswell) processors, which are optimized for enterprise and database workloads. The E7 processors have a high core count to support workloads that scale efficiently on large number of cores. The Intel E7 processors also feature high memory bandwidth and larger L3 caches to boost the performance of in-memory applications. In addition, the Intel E7 processor:
• Enables increased cryptographic performance via the latest Intel AES-NI feature.
• Supports Transactional Synchronization Extensions (TSX) to boost the performance of in-memory transactional data processing.
• Supports Advanced Vector Extensions 2 (Intel AVX2) processor instructions to expand most integer commands to 256 bits.

Q. Do X1 instances enable CPU power management state control?
Yes. You can configure C-states and P-states on both x1.32xlarge and x1.16xlarge. You can use C-states to enable higher turbo frequencies (as much as 3.1 Ghz with one or two core turbo). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

Q: What operating systems are supported on X1 instances?

X1 instances provide high number of vCPUs, which might cause launch issues in some Linux operating systems that have a lower vCPU limit. We strongly recommend that you use the latest AMIs when you launch X1 instances. The following Linux AMIs support launching X1 instances: Amazon Linux AMI 2016.03 (HVM), Ubuntu Server 14.04 LTS (HVM), and Red Hat Enterprise Linux 7.1 (HVM), and SUSE Linux 12 SP1.

AMI support for SAP HANA workloads include: SUSE Linux 12, SUSE Linux 12 SP1, SLES for SAP 12 SP1 (due to kernel requirement of 3.10 or higher). For SAP NetWeaver on AnyDB, the latest RHEL 7.x images are currently supported.

x1.32xlarge will also support Windows Server 2012 R2, 2012 RTM and 2008 R2 64bit (Windows Server 2008 SP2 and older versions will not be supported) and x1.16xlarge will support Windows Server 2012 R2, 2012 RTM, 2008 R2 64bit, 2008 SP2 64bit, and 2003 R2 64bit (Windows Server 32bit versions will not be supported).

Q. What storage options are available for X1 customers?

X1 instances offer SSD based instance store, which is ideal for temporary storage of information such as logs, buffers, caches, temporary tables, temporary computational data, and other temporary content. X1 instance store provides the best I/O performance when you use a Linux kernel that supports persistent grants, an extension to the Xen block ring protocol.
X1 instances are EBS-optimized by default and offer up to 10 Gbps of dedicated bandwidth to EBS volumes. EBS offers multiple volume types to support a wide variety of workloads. For more information see the EC2 User Guide.

Q. How do I build cost-effective failover solution on X1 instances?
You can design simple and cost-effective failover solutions on X1 instances using Amazon EC2 Auto Recovery, an Amazon EC2 feature that is designed to better manage failover upon instance impairment. You can enable Auto Recovery for X1 instances by creating an AWS CloudWatch alarm. Choose the “EC2 Status Check Failed (System)” metric and select the “Recover this instance” action. Instance recovery is subject to underlying limitations, including those reflected in the Instance Recovery Troubleshooting documentation. For more information visit Auto Recovery documentation and Creating Amazon CloudWatch Alarms respectively.

Q. Are there standard SAP HANA reference deployment frameworks available for the X1 instance and the AWS cloud?
You can use the AWS Quick Start reference HANA deployments to rapidly deploy all the necessary HANA building blocks on X1 instances following SAP’s recommendations for high performance and reliability. AWS Quick Starts are modular and customizable, so you can layer additional functionality on top or modify them for your own implementations. For additional information on deploying HANA on AWS, please refer to SAP HANA on AWS Cloud: Quick Start Reference Deployment Guide.

Previous Generation Instances

Q: Why don't I see M1, C1, CC2, HI1, CG1, and HS1 instances on the pricing pages any more?

These have been moved to the Previous Generation Instance page.

Q: Are these Previous Generation instances still being supported?

Yes. Previous Generation instances are still fully supported.

Q: Can I still use/add more Previous Generation instances?

Yes. Previous Generation instances are still available as On-Demand, Reserved Instances, and Spot Instance, from our APIs, CLI, and EC2 Management Console interface.

Q: Are my Previous Generation instances going to be deleted?

No. Your M1, C1, CC2, HI1, CG1, and HS1 instances are still fully functional and will not be deleted because of this change.

Q: Are Previous Generation instances being discontinued soon?

Currently, there are no plans to end of life Previous Generation instances. However, with any rapidly evolving technology the latest generation will typically provide the best performance for the price and we encourage our customers to take advantage of technological advancements.

Q: Will my Previous Generation instances I purchased as a Reserved Instance be affected or changed?

No. Your Reserved Instances will not change, and the Previous Generation instances are not going away.

VM Import/Export

Q. What is VM Import/Export?

VM Import/Export enables customers to import Virtual Machine (VM) images in order to create Amazon EC2 instances. Customers can also export previously imported EC2 instances to create VMs. Customers can use VM Import/Export to leverage their previous investments in building VMs by migrating their VMs to Amazon EC2.

Q. What operating systems are supported?

VM Import/Export currently supports Windows and Linux VMs, includingWindows Server 2003, Windows Server 2003 R2, Windows Server 2008, Windows Server 2012 R1, Red Hat Enterprise Linux (RHEL) 5.1-6.5 (using Cloud Access), Centos 5.1-6.5, Ubuntu 12.04, 12.10, 13.04, 13.10, and Debian 6.0.0-6.0.8, 7.0.0-7.2.0. For more details on VM Import, including supported file formats, architectures, and operating system configurations, please see the VM Import/Export section of the Amazon EC2 User Guide.

Q. What virtual machine file formats are supported?

You can import VMware ESX VMDK images, Citrix Xen VHD images, Microsoft Hyper-V VHD images and RAW images as Amazon EC2 instances. You can export EC2 instances to VMware ESX VMDK, VMware ESX OVA, Microsoft Hyper-V VHD or Citrix Xen VHD images. For a full list of support operating systems, please see What operating systems are supported?.

Q. What is VMDK?

VMDK is a file format that specifies a virtual machine hard disk encapsulated within a single file. It is typically used by virtual IT infrastructures such as those sold by VMware, Inc.

Q. How do I prepare a VMDK file for import using the VMware vSphere client?

The VMDK file can be prepared by calling File-Export-Export to OVF template in VMware vSphere Client. The resulting VMDK file is compressed to reduce the image size and is compatible with VM Import/Export. No special preparation is required if you are using the Amazon EC2 VM Import Connector vApp for VMware vCenter.

Q. What is VHD?

VHD (Virtual Hard Disk) is a file format that that specifies a virtual machine hard disk encapsulated within a single file. The VHD image format is used by virtualization platforms such as Microsoft Hyper-V and Citrix Xen.

Q. How do I prepare a VHD file for import from Citrix Xen?

Open Citrix XenCenter and select the virtual machine you want to export. Under the Tools menu, choose "Virtual Appliance Tools" and select "Export Appliance" to initiate the export task. When the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q. How do I prepare a VHD file for import from Microsoft Hyper-V?

Open the Hyper-V Manager and select the virtual machine you want to export. In the Actions pane for the virtual machine, select "Export" to initiate the export task. Once the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q. Are there any other requirements when importing a VM into Amazon EC2?

The virtual machine must be in a stopped state before generating the VMDK or VHD image. The VM cannot be in a paused or suspended state. We suggest that you export the virtual machine with only the boot volume attached. You can import additional disks using the ImportVolume command and attach them to the virtual machine using AttachVolume. Additionally, encrypted disks (e.g. Bit Locker) and encrypted image files are not supported. You are also responsible for ensuring that you have all necessary rights and licenses to import into AWS and run any software included in your VM image.

Q. Does the virtual machine need to be configured in any particular manner to enable import to Amazon EC2?

Ensure Remote Desktop (RDP) or Secure Shell (SSH) is enabled for remote access and verify that your host firewall (Windows firewall, iptables, or similar), if configured, allows access to RDP or SSH. Otherwise, you will not be able to access your instance after the import is complete. Please also ensure that Windows VMs are configured to use strong passwords for all users including the administrator and that Linux VMs and configured with a public key for SSH access.

Q. How do I import a virtual machine to an Amazon EC2 instance?

You can import your VM images using the Amazon EC2 API tools:

- Import the VMDK, VHD or RAW file via the ec2-import-instance API. The import instance task captures the parameters necessary to properly configure the Amazon EC2 instance properties (instance size, Availability Zone, and security groups) and uploads the disk image into Amazon S3.

- If ec2-import-instance is interrupted or terminates without completing the upload, use ec2-resume-import to resume the upload. The import task will resume where it left off.

- Use the ec2-describe-conversion-tasks command to monitor the import progress and obtain the resulting Amazon EC2 instance ID.

- Once your import task is completed, you can boot the Amazon EC2 instance by specifying its instance ID to the ec2-run-instances API.

- Finally, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3 as it is no longer needed.

Alternatively, if you use the VMware vSphere virtualization platform, you can import your virtual machine to Amazon EC2 using a graphical user interface provided through AWS Management Portal for vCenter. Please refer to Getting Started Guide in AWS Management Portal for vCenter. AWS Management Portal for vCenter includes integrated support for VM Import. Once the portal is installed within vCenter, you can right-click on a VM and select “Migrate to EC2” to create an EC2 instance from the VM. The portal will handle exporting the VM from vCenter, uploading it to S3, and converting it into an EC2 instance for you, with no additional work required. You can also track the progress of your VM migrations within the portal.

Q. How do I export an Amazon EC2 instance back to my on-premise virtualization environment?

You can export your Amazon EC2 instance using the Amazon EC2 CLI tools:

- Export the instance using the ec2-create-instance-export-task command. The export command captures the parameters necessary (instance ID, S3 bucket to hold the exported image, name of the exported image, VMDK, OVA or VHD format) to properly export the instance to your chosen format. The exported file is saved in an S3 bucket that you previously created
- Use ec2-describe-export-tasks to monitor the export progress
- Use ec2-cancel-export-task to cancel an export task prior to completion

Q. Are there any other requirements when exporting an EC2 instance using VM Import/Export?

You can export running or stopped EC2 instances that you previously imported using VM Import/Export. If the instance is running, it will be momentarily stopped to snapshot the boot volume. EBS data volumes cannot be exported. EC2 instances with more than one network interface cannot be exported.

Q. Can I export Amazon EC2 instances that have one or more EBS data volumes attached?

Yes, but VM Import/Export will only export the boot volume of the EC2 instance.

Q. What does it cost to import a virtual machine?

You will be charged standard Amazon S3 data transfer and storage fees for uploading and storing your VM image file. Once your VM is imported, standard Amazon EC2 instance hour and EBS service fees apply. If you no longer wish to store your VM image file in S3 after the import process completes, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3.

Q. What does it cost to export a virtual machine?

You will be charged standard Amazon S3 storage fees for storing your exported VM image file. You will also be charged standard S3 data transfer charges when you download the exported VM file to your on-premise virtualization environment. Finally, you will be charged standard EBS charges for storing a temporary snapshot of your EC2 instance. To minimize storage charges, delete the VM image file in S3 after downloading it to your virtualization environment.

Q. When I import a VM of Windows Server 2003 or 2008, who is responsible for supplying the operating system license?

When you launch an imported VM using Microsoft Windows Server 2003 or 2008, you will be charged standard instance hour rates for Amazon EC2 running the appropriate Windows Server version, which includes the right to utilize that operating system within Amazon EC2. You are responsible for ensuring that all other installed software is properly licensed.

So then, what happens to my on-premise Microsoft Windows license key when I import a VM of Windows Server 2003 or 2008? Since your on-premise Microsoft Windows license key that was associated with that VM is not used when running your imported VM as an EC2 instance, you can reuse it for another VM within your on-premise environment.

Q. Can I continue to use the AWS-provided Microsoft Windows license key after exporting an EC2 instance back to my on-premise virtualization environment?

No. After an EC2 instance has been exported, the license key utilized in the EC2 instance is no longer available. You will need to reactivate and specify a new license key for the exported VM after it is launched in your on-premise virtualization platform.

Q. When I import a VM with Red Hat Enterprise Linux (RHEL), who is responsible for supplying the operating system license?

When you import Red Hat Enterprise Linux (RHEL) VM images, you can use license portability for your RHEL instances. With license portability, you are responsible for maintaining the RHEL licenses for imported instances, which you can do using Cloud Access subscriptions for Red Hat Enterprise Linux. Please contact Red Hat to learn more about Cloud Access and to verify your eligibility.

Q. How long does it take to import a virtual machine?

The length of time to import a virtual machine depends on the size of the disk image and your network connection speed. As an example, a 10 GB Windows Server 2008 SP2 VMDK image takes approximately 2 hours to import when it’s transferred over a 10 Mbps network connection. If you have a slower network connection or a large disk to upload, your import may take significantly longer.

Q. In which Amazon EC2 regions can I use VM Import/Export?

Visit the Region Table page to see product service availability by region.

Q. How many simultaneous import or export tasks can I have?

Each account can have up to five active import tasks and five export tasks per region.

Q. Can I run imported virtual machines in Amazon Virtual Private Cloud (VPC)?

Yes, you can launch imported virtual machines within Amazon VPC.

Q. Can I use the AWS Management Console with VM Import/Export?

No. VM Import/Export commands are available via EC2 CLI and API. You can also use theAWS Management Portal for vCenter to import VMs into Amazon EC2. Once imported, the resulting instances are available for use via the AWS Management Console.

Amazon EC2 Running Microsoft Windows and Other Third-Party Software

Q. Can I use my existing Windows Server license with EC2?

No. Microsoft Windows Server licensing does not currently support using your existing Windows license in Amazon EC2 or any other cloud environment. We encourage you to work with your Microsoft account representative to understand licensing options.

Q. What software licenses can I bring to the Windows environment?

Specific software license terms vary from vendor to vendor. Therefore, we recommend that you check the licensing terms of your software vendor to determine if your existing licenses are authorized for use in Amazon EC2.

Amazon EC2 Running IBM

Q. How am I billed for my use of Amazon EC2 running IBM?

You pay only for what you use and there is no minimum fee. Pricing is per instance-hour consumed for each instance type. Partial instance-hours consumed are billed as full hours. Data transfer for Amazon EC2 running IBM is billed and tiered separately from Amazon EC2. There is no Data Transfer charge between two Amazon Web Services within the same region (i.e. between Amazon EC2 US West and another AWS service in the US West). Data transferred between AWS services in different regions will be charged as Internet Data Transfer on both sides of the transfer.

For Amazon EC2 running IBM pricing information, please visit the pricing section on the Amazon EC2 running IBM detail page.

Q. Can I use Amazon DevPay with Amazon EC2 running IBM?

No, you cannot use DevPay to bundle products on top of Amazon EC2 running IBM at this time.

Service Level Agreement (SLA)

Q. What does your Amazon EC2 Service Level Agreement guarantee?

Our SLA guarantees a Monthly Uptime Percentage of at least 99.95% for Amazon EC2 and Amazon EBS within a Region.

Q. How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for either Amazon EC2 or Amazon EBS (whichever was Unavailable, or both if both were Unavailable) if the Region that you are operating in has an Monthly Uptime Percentage of less than 99.95% during any monthly billing cycle. For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see http://aws.amazon.com/ec2/sla/

Amazon EC2 Windows FAQ

General

What is the relationship between Microsoft and Amazon Web Services?

Amazon Web Services and Microsoft have worked together for many years, starting with AWS launching Windows Server based instances in 2008. AWS is a Gold Certified member of the Microsoft Partner Network and licensed to sell Microsoft software under the Services Provider License Agreement (SPLA). Over the years, AWS and Microsoft have collaborated to make Windows and its associated workloads available in the AWS cloud. Microsoft and AWS have mutual customers running Windows workloads on AWS today, including Dole Foods, Hess Corporation, and Lionsgate. In addition, AWS has released Microsoft-specific technologies that allow users to manage and optimize Windows applications in AWS – such as AWS tools for Windows PowerShell, AWS Management Pack for Microsoft System Center, and AWS Diagnostics for Microsoft Windows Server.

Is Microsoft software supported on AWS?

AWS supports Microsoft software running on AWS. AWS customers have successfully deployed every Microsoft application available in the AWS cloud, including (but not limited to) Microsoft Office, Windows Server, SQL Server, Exchange, SharePoint, Lync, Dynamics, and Remote Desktop Services.

AWS is a member of the Microsoft Partner Network, licensed to resell Microsoft software via the SPLA, and a Microsoft Gold Certified Hosting Partner. AWS also has an active Premier Support agreement with Microsoft.

How does AWS support issues with Microsoft software?

If a problem is identified with a Microsoft product on AWS that expands beyond AWS technologies, then AWS works closely with Microsoft support to provide a coordinated support experience for customers. However, if you have an existing Microsoft Support agreement you can contact Microsoft Support directly, under that agreement. Also, if Microsoft determines that they need to perform infrastructure level debugging, you or Microsoft Support can contact AWS Support to help resolve the issue. Support for Microsoft workloads on Amazon EC2 can be a collaborative effort between you, AWS Support, and Microsoft Support.

What types of Microsoft software can I run on AWS?

You can run many types of Microsoft software on AWS, including but not limited to Microsoft Office, Windows Server, SQL Server, Exchange, SharePoint, Lync, Skype for Business, Microsoft Dynamics products, System Center, BizTalk, and Remote Desktop Services. You can pay for Windows Server and SQL Server licenses directly from AWS to run on Amazon EC2 or Amazon RDS instances. You also have the flexibility to bring your own licenses (BYOL) which allows you to pay Amazon Linux pricing for Amazon EC2 instances instead of paying for license-included instances.

Licensing

What are my licensing options for Microsoft software on Amazon EC2?

On Amazon EC2, you can choose to run instances that include the relevant license fees in their cost (“license included”) or bring in their own licenses (BYOL). For Microsoft software, EC2 allows you to pay for instances that include Windows Server and SQL Server licenses. For all other Microsoft software, customers can bring their own license, subject to Microsoft’s terms.

What is BYOL?

BYOL, or “bring your own license,” is the process you can use to deploy software that you’ve licensed from an ISV on AWS hardware. If you BYOL, you do not pay for instances with licensing included in the cost. Instead you pay the same rate as EC2 instances with Amazon Linux pricing. When you BYOL, you are responsible for managing your own licenses, but Amazon EC2 has features that help you maintain license compliance throughout the lifecycle of your licenses, such as Instance Affinity and targeted placement available through Amazon EC2 Dedicated Hosts.

What is License Mobility?

License Mobility is a benefit available to Microsoft Volume Licensing customers with eligible server applications covered by active Microsoft Software Assurance (SA). License Mobility allows customers to move eligible Microsoft software to third party cloud providers such as AWS for use on EC2 instances with default tenancy. It is important to note that you may not need license mobility if you are using your own licenses on EC2 Dedicated Hosts or EC2 Dedicated Instances For additional details, see the Microsoft License Mobility page on the AWS site.

Do I need to pay for Software Assurance and License Mobility to use my Microsoft licenses in AWS?

No, if you are bringing your own licenses into EC2 Dedicated Hosts or EC2 Dedicated Instances then Software Assurance is not required subject to Microsoft's terms. If you are moving licensed software onto EC2 instances with a default tenancy, Software Assurance is required. You need to have Software Assurance in order to participate in Microsoft's License Mobility program.

If I bring my own license, can I relicense AWS's Microsoft media or do I need to bring in my own media (a.k.a. "bring your own bits")?

No, you must import and license your own media. To get started, you can use theImportImage tool to import your own media. If you are importing media to run on EC2 instances, after the media has been imported, you will see your images in the "My AMIs" console, or you can describe these images using the *DescribeImages* API.

After I import my Microsoft media, do I need to activate my media against my own key management server (KMS)?

Yes, when you launch an instance of your own image, your OS will prompt you to activate the image against your KMS.

How do I know what type of offering to use if I'm bringing my own license?

Please read your licensing terms and conditions and select the AWS model that meets your needs. Generally speaking, there are various products and each have differing levels of BYOL support:

BYOL Licensing Scenarios

License Type	EC2 Dedicated Hosts	EC2 Dedicated Instances	EC2 Multi-Tenant
Windows Server	✓	LI	LI
SQL Server	✓	✓ Only on Windows Server license included EC2 Dedicated Instances	✓ Only if you have licenses with License Mobility and are running on license included Windows Server EC2 instances
MS Office	✓	✓	NA
Windows 7, 8, and 10	✓	✓	NA
MSDN	✓	✓	X
Other	✓ Subject to Microsoft's Terms	✓ Only on Windows Server license included EC2 Dedicated Instances	✓ Only if you have licenses with License Mobility and are running on Windows Server EC2 Instances

✓ = scenario is supported
LI = only offered as license included instances sold by AWS
NA = not applicable
X = not allowed

Under your agreements with Microsoft, you may have a special case to use your licenses in a way that is different than described in the BYOL Licensing Scenario table. If your agreements permit a special case where you have additional rights to use your licenses, please contact your account manager or AWS customer support. For additional questions about Microsoft licensing terms contact Microsoft or your Microsoft reseller.

How do I import my own licensed machine image into AWS?

In order to BYOL of Microsoft software into AWS, you need to use theImportImage tool made available by the EC2 VM Import/Export service. Do not use the ImportInstance tool as it does not support Microsoft BYOL scenarios.

I've read in my licensing terms that certain licenses must be used on infrastructure that’s dedicated for my use. How does Amazon EC2 allow me to meet this requirement if I’m using my own licenses?

Amazon EC2 offers two purchasing options that provide you with dedicated infrastructure: Dedicated Hosts and Dedicated Instances. It is important to note that all BYOL scenarios are supported through the use of Dedicated Hosts, while only certain scenarios are supported by Dedicated Instances. Also, if you bring existing licenses to Dedicated Hosts or Dedicated Instances, then you are using hardware that is fully dedicated to your use and the outsourcing language within the Microsoft Product Terms applies.

For BYOL license scenarios that are server bound (e.g., Windows Server, SQL Server) and require you to license against the number of sockets or physical cores on a dedicated server, you should use Dedicated Hosts.

For licensing scenarios that are VM, CAL, or user bound and do not require you to license against the number of sockets or physical cores on a dedicated server but require you to run on dedicated infrastructure (e.g., Windows 7, Windows 8, Windows 10, BYOL SQL Server) you can use Dedicated Instances.

For more information on Dedicated Hosts, visit the Dedicated Hosts detail page.

For more information on Dedicated Instances, visit the Dedicated Instances detail page.

I've read in my licensing terms that a license cannot move to another region or physical machine for at least 90 days. How does Amazon EC2 help me meet this requirement if I’m using my own licenses?

Instance Affinity (only available through the use of Amazon EC2 Dedicated Hosts) and Dedicated Host targeting helps you to monitor this requirement. When you enable Affinity between an instance and a Dedicated Host, that particular instance will only run on a specific Dedicated Host. Using Dedicated Host targeting, you can launch instances onto a specific Dedicated Host, giving you full control over how your licenses are used. For more information on these features, visit the Dedicated Hosts detail page.

When can I bring my own license using EC2 instances with default tenancy?

Microsoft’s License Mobility Program allows qualifying customers to bring eligible Microsoft software licenses into AWS for use on EC2 instances with default tenancy. The AWS License

Mobility Page is a great place to start the process. If you are planning to take advantage of License Mobility in AWS, you will need to fill out the appropriate License Mobility forms. When you BYOL software under the License Mobility program, you can use these images on EC2 Windows Server license-included instances running on EC2 instances with default tenancy. Windows Server licenses must be purchased from AWS in this scenario.

What is VM Import/Export?

VM Import/Export enables you to easily import virtual machine images from your existing environment to Amazon EC2 instances. This service allows you to leverage your existing investments in the virtual machines that you have built to meet your IT security, configuration management, and compliance requirements by bringing those virtual machines into Amazon EC2 as ready-to-use instances. If you are planning to use your own Microsoft licenses, use the ImportImage tool made available by the VM Import/Export service to import your own Microsoft media.

The VM Import/Export service is available at no additional charge beyond standard usage charges for Amazon EC2 and Amazon S3.

What is EC2's default tenancy?

EC2 Dedicated instances and EC2 Dedicated Hosts provide instance capacity on physical servers that are fully dedicated for your use. Alternatively, EC2 offers instances with a tenancy of ‘default’ which run on physical servers that may host multiple isolated instances from different customers.

What is dedicated infrastructure?

Dedicated infrastructure provides servers that are physically isolated for use by a single customer. Amazon EC2 has two dedicated infrastructure options: Dedicated Hosts and Dedicated Instances. If you bring existing licenses to Dedicated Hosts or Dedicated Instances, then you are using hardware that is fully dedicated to your use. In that case, the outsourcing language within the Microsoft Product Terms applies. When you want to use instances on a Dedicated Host, launch instances with a tenancy of ‘host’. When you want to use Dedicated Instances, launch instances with a tenancy of ‘dedicated’.

What are Amazon EC2 Dedicated Hosts?

A Dedicated Host is a physical EC2 server fully dedicated to your use. With Dedicated Hosts, you have more control over instance placement and gain visibility into the number of sockets and cores installed on a host. You can use these features to leverage your own per-socket or per-core software licenses, including Windows Server and SQL Server. Visit the Dedicated Host detail page for more information.

What are Amazon EC2 Dedicated Instances?

Dedicated instances are Amazon EC2 instances that run on hardware that is dedicated to a single customer. For more information on Dedicated Instances, visit the Dedicated Instance page.

What's the difference between Dedicated Hosts and Dedicated Instances?

Both offerings provide instances that are dedicated to your use. However, Dedicated Hosts provide additional control over your instances and visibility into Host level resources and tooling that allows you to manage software that consumes licenses on a per-core or per-socket basis, such as Windows Server and SQL Server. In addition, AWS Config will keep a record of how your instances use these Dedicated Host resources which will allow you to create your own license usage reports.

Does AWS recommend an EC2 purchasing model if I'm looking to use my own licenses?

Dedicated Hosts supports all BYOL scenarios outlined in this FAQ and it provides customers with more control and visibility over how their instances are placed, which is useful for minimizing risk and licensing costs in a BYOL scenario. Additionally, Dedicated Hosts support per-socket, per-core, VM, and CAL based licenses.

Licensing – Windows Server

Can I buy Windows Server from AWS?

Yes, you can deploy Windows Server on AWS by purchasing Amazon Machine Images (AMIs) with Windows Server pre-installed. If you buy Windows instances from AWS, whether your instances have a tenancy of dedicated or default, the Windows Server license is included in the cost.

With EC2 license-included instances, EC2 manages licensing compliance, you only pay for what you use, you do not need to pay for Software Assurance, and you have the flexibility to upgrade your software when it is made available without additional cost. Also, there is no need to buy additional Windows Server CALs as access is included in the price, along with two remote connections for admin purposes only. If you require more than two connections or need those remote connections for purposes other than admin you may need to bring in Remote Desktop Services CALs.

Can I bring my own Windows Server licenses and use them in EC2?

Yes you can. After you've imported your own Windows Server machine images using the ImportImage tool, you need to launch instances from these machine images on EC2 Dedicated Hosts in order to effectively manage instances and report usage. Microsoft typically requires that you track usage of your licenses against physical resources such as sockets and cores and Dedicated Hosts helps you to do this. Visit the Dedicated Hosts detail page for more information on how to use your own Windows Server licenses on Amazon EC2 Dedicated Hosts.

What are Amazon EC2 Dedicated Hosts?

A Dedicated Host is a physical EC2 server fully dedicated for your use. With Dedicated Hosts, you have more control over instance placement and gain visibility into the number of sockets and cores installed on a host. You can use these features to bring in your own software licenses bound to VMs, sockets, or cores, including Windows Server, SQL Server, and SUSE Enterprise Server. For more information on Dedicated Hosts, visit the Dedicated Hosts detail page.

How do I import and use my own Windows Server license?

You can bring in your own licensed copy of Windows Server media using theImageImport tool made available by the EC2 VM Import/Export service. Once these images are imported, you can find them under the "my AMIs" section in the AWS Management Console or by using the *DescribeImages* API. You can then launch instances from your BYOL machine images onto Dedicated Hosts.

Visit this link for more information on how to bring your own machine images into AWS.

Keep in mind that when you choose to bring in your existing Windows Server licenses, you cannot utilize Windows Server AMIs that you purchase from AWS through license-included instances. You must bring in your own licenses using your own software media.

How do I track usage if I'm bringing my own licenses?

Using AWS Config as the data source and Dedicated Hosts as the platform to run BYOL instances, you can track BYOL usage against physical resources such as sockets and cores. Before you begin launching BYOL instances onto your Dedicated Hosts, ensure AWS Config has been enabled to record Dedicated Host changes. AWS Config keeps track of the configuration changes that occur on a Dedicated Host, including the instances and corresponding IDs of AMIs that ran on a Dedicated Host. These changes are paired with Host level data, such as the Host ID and the number of sockets and physical cores installed on a Dedicated Host. AWS Config will also keep track of instance tags. We recommend that you tag your instances with a meaningful identifier if you would like a human-readable way to identify BYOL instances in the AWS Config output. Visit this page for more information on AWS Config.

How do I determine the number of licenses of Windows Server to bring in?

Each Dedicated Host provides you with the number of sockets (aka physical processors) and physical cores installed on a particular server. Using this information you can calculate the number of Windows server licenses that you need to bring in. Microsoft has published a document that helps customers calculate the number of licenses that are required for Windows Server 2012 for Datacenter and Standard editions.

Visit the Dedicated Hosts detail page for information on the number of instances available per Dedicated Host. On this page you will also find the number of sockets and cores installed on each EC2 Dedicated Host. The instance, socket, and core counts vary by the instance type configuration of the Dedicated Host.

Do I need to have Software Assurance on Windows Server on AWS?

No, if you are using Dedicated Hosts to use your own Windows Server licenses, you do not need to have Software Assurance (SA). Also, if you purchase Windows Server instances from AWS, then there is no need for you to have Software Assurance to cover those Windows Server licenses.

Does License Mobility work with Windows Server?

No, as specified in the Microsoft Product Terms, Windows Server, Windows client, and Microsoft Office are not eligible for License Mobility. Since License Mobility enables the use of licenses on EC2 instances with a default tenancy, License Mobility is not required for licenses used on EC2 Dedicated Hosts. If you choose to use Dedicated Hosts for BYOL scenarios, then you can bring in your own licenses for Windows Server, Windows client, and Microsoft Office without the need for License Mobility.

How can I use my own Windows Server license on EC2 instances with a default tenancy?

You should use your own Windows Server licenses on Dedicated Hosts and you can do this by running instances with a tenancy of ‘host’. You should not use your own Windows Server license on EC2 instances with a default tenancy unless you have approval from Microsoft to do so. If you have negotiated custom terms with Microsoft and have this permission, please contact AWS support or reach out to your account manager.

What is included when I buy Windows Server instances from AWS?

AWS manages the licensing for you; all you need to do is pay for the instances you use. There is also no need to buy additional Windows Server CALs, as access is included in the price. Each instance comes with two remote connections for admin purposes only. If you require more than two connections, or need those connections for purposes other than admin, you may have to bring in additional Remote Desktop Services CALs for use on AWS.

Can I relicense license-included, EC2 Windows Server instances to use my own licenses, pointing at my own KMS server?

No, you cannot relicense existing Windows Server EC2 instances or migrate existing Windows Server EC2 instances over to BYOL VMs. However, if you need to migrate from license-included to BYOL and have applications or OS configurations that need to be migrated, we suggest that you reach out to our partners, such as CloudEndure or AppZero, who may be able to assist with these types of migrations.

Licensing – SQL Server

Can I buy SQL Server from AWS?

Yes, you can buy instances with SQL Server licenses included from AWS to run on either Amazon EC2 or Amazon Relational Database Service (RDS). SQL Server Web Edition, Standard Edition, and Enterprise Edition are available for you to license on both Amazon EC2 and Amazon RDS.

Can I bring in my own SQL Server licenses for use on AWS?

Yes, you can bring in your own licenses (BYOL) on EC2 Dedicated Hosts, EC2 Dedicated Instances with license included Windows Server, or EC2 instances with a default tenancy with License Mobility.

- **EC2 instances with a default tenancy**
  Microsoft's License Mobility through Software Assurance allows qualifying customers to bring in eligible Microsoft software onto AWS for use on EC2 instances with a default tenancy. The AWS License Mobility Page is a great place to start the process. You will need to fill out the appropriate License Mobility forms and file them with Microsoft to ensure that the licenses are able to be brought to AWS. You can use your own SQL Server licenses on top of license-included EC2 Windows Server default tenancy or RDS default tenancy instances.

- **Dedicated**
  The use of Dedicated Hosts allows you use a per-core or per-socket SQL Server licensing model, and you do not need to have access to License Mobility through Software Assurance, which can save you money on licensing costs if you are bringing your own license. If you choose to license your SQL Server licenses against the sockets or cores on a physical machine, you need to use these licenses on Dedicated Hosts. Visit the Dedicated Hosts detail page for more information on how to use your own SQL Server licenses on Amazon EC2 Dedicated Hosts. You can also choose to use EC2 license-included Windows Server Dedicated Instances, where you pay for SQL Server licenses on a VM basis.

Can I use License Mobility with SQL Server?

Yes, license Mobility is a benefit available to Microsoft Volume Licensing customers with eligible server applications (including SQL Server) covered by active Microsoft Software Assurance (SA) contracts. License Mobility allows customers to move eligible Microsoft software to third party cloud providers such as AWS for the end use on EC2 instances with a default tenancy. It is important to note that you may not need license mobility if you are using your own licenses on EC2 Dedicated Hosts or EC2 Dedicated Instances. For additional details, see the Microsoft License Mobility page on the AWS site. Qualifying customers with Software Assurance can bring in their own licenses of SQL Server for use on Amazon EC2 and Amazon RDS instances with a default tenancy.

Do I have to pay for failover SQL Servers for disaster recovery?

If you are bringing your own license and using EC2 Dedicated Hosts or EC2 Dedicated Instances, you do not need to pay for the SQL Servers on these failover servers if these failover servers are passive. For more information on SQL and failover server scenarios, visit this Microsoft SQL Server licensing guide. This pertains only to the SQL Server licenses and not the Windows Server licenses. In all cases you must license Windows Server. However, if you are using SQL Server enabled instances (either your own through License Mobility or as license-included instances), you need to individually license your passive servers running on EC2 Dedicated Hosts or EC2 Dedicated Instances. The following highlights three scenarios for SQL Server failover licensing:

- **Purchasing licenses with Amazon Machine Images (AMIs)**
  If you buy SQL Server licenses from AWS, then you have to pay for failover servers. Microsoft requires that each instance of the server be licensed and that includes disaster recovery failover servers.
- **Bringing existing licenses to Amazon EC2 Dedicated Hosts**
  When you use EC2 Dedicated Hosts or EC2 Dedicated Instances, the outsourcing language

within the Microsoft Product Terms applies. In the case of SQL Server, this means that you may not have to license your failover servers (for SQL Server only – you will still require to license Windows Server) as long as they are passive servers running on EC2 Dedicated Hosts or EC2 Dedicated Instances. See the Microsoft Product Terms or the Microsoft SQL Server licensing guide for more details.

- **Bringing existing licenses with License Mobility**
  If you bring in SQL Server licenses via License Mobility, then you must bring in licenses for the failover servers as well.

How do I know how many SQL Server licenses to bring in?

If you are licensing SQL Server under Microsoft's License Mobility through Software Assurance, the number of licenses required varies based on the instance type, version of SQL Server, and the Microsoft licensing model you choose. To assist you with your virtual core licensing calculations under the Microsoft Product Terms, we provide a table here that shows the number of virtual representations of hardware threads based on instance type.

If you are using Dedicated Hosts, EC2 provides you with the number of physical cores installed on the Dedicated Host. Using this information, you can calculate the number of SQL Server licenses that you need to bring in. For additional information, we recommend referencing Microsoft documentation, such as the licensing guide for SQL server 2014 (see here).

How do I determine the right core factor to license with?

You can determine the core count per server by dividing the number of physical cores on the Dedicated Host by the socket count. This information can be found on the Dedicated Host detail page. You can find the processor types on the EC2 Instance Type detail page.

How do I track usage if I'm bringing my own licenses?

Using AWS Config as the data source you can track configuration changes against physical resources such as sockets and cores. Before you begin launching BYOL instances onto AWS, ensure AWS Config has been enabled to record any changes. AWS Config keeps track of the changes that occur, including the instances and corresponding AMI IDs that ran. These changes are paired with Host level data, such as the Host ID and the number of sockets and physical cores installed. AWS Config will also keep track of instance tags. We recommend that you tag your instances with a meaningful identifier if you would like a human-readable way to identify BYOL instances in your AWS Config logs. Visit this page for more information on AWS Config.

Licensing – MSDN

Can I run MSDN licenses on AWS?

Yes, you can use MSDN licenses on AWS if you bring your own licenses. Microsoft does not allow the use of MSDN on multi-tenant AWS servers. However, if you use Dedicated Instances or Dedicated Hosts and your MSDN licenses are governed by the Microsoft Product Terms, then you can bring your MSDN licenses to AWS. Dedicated hardware is fully dedicated to your use and Microsoft views this as outsourcing which activates different language in the Product Terms.

Can I buy MSDN from AWS?

No, AWS does not sell MSDN licenses.

Can I use MSDN on AWS instances with a default tenancy?

No, Microsoft does not allow MSDN licenses to be utilized on AWS instances with a default tenancy.

Can I use my existing MSDN licenses on EC2 Dedicated Hosts or EC2 Dedicated Instances?

Yes, you can use these licenses on either Dedicated Hosts or Dedicated Instances.

Does License Mobility work with MSDN?

No, MSDN is not included in Microsoft’s License Mobility program.

Licensing – Windows Client (7, 8, 10 etc.)

Can I buy Windows Client from AWS?

No. AWS does not sell any Windows Client operating system licenses on any of our services.

Can I bring in my own Windows Client licenses for use on AWS?

Yes. If you use Dedicated Instances or Dedicated Hosts, then you can bring in your own Windows Client licenses for use on AWS. You may require Software Assurance or Virtual Desktop Access (VDA) in order to utilize the Windows client operating systems such as Windows 7 or Windows 8 on AWS. We recommend you read this Microsoft licensing brief for more information.

Can I use License Mobility with Windows Client?

No, as specified in the Microsoft Product Terms, License Mobility does not apply to Windows Client, Windows Server, or Microsoft Office. Since License Mobility enables the use of specific licenses on EC2 instances with a default tenancy, License Mobility does not apply to licenses that require the use on EC2 Dedicated Hosts or EC2 Dedicated Instances. If you choose to use Dedicated Hosts and BYOL, then you can bring in your own licenses for Windows Client, Windows Server, and Microsoft Office without needing License Mobility.

Licensing – Microsoft Office

Can I bring in my own Office licenses for use on AWS?

Yes, you can BYOL of Microsoft Office for use on EC2 Dedicated Hosts or EC2 Dedicated Instances. If you bring existing licenses to EC2 Dedicated Hosts or EC2 Dedicated Instances, then you are using hardware that is fully dedicated to your use. In that case, the outsourcing language within the Microsoft Product Terms applies. This allows you to bring in Office licenses for use on your own Windows client licenses.

Can I use License Mobility with Microsoft Office?

No, Microsoft does not include Microsoft Office in Microsoft's License Mobility program.

Licensing – Other Microsoft Products (Exchange, SharePoint, Lync, etc.)

Can I buy other Microsoft products from AWS?

No. AWS sells only Windows Server and SQL Server licenses today for use on Amazon EC2.

Can I bring in my own licenses for use on AWS?

Yes. We have many customers that have successfully brought in and deployed licenses on AWS. These deployments include, but are not limited to, Exchange, SharePoint, Lync, Remote Desktop Services, Office, Dynamics products, BizTalk, and System Center.

Customers can choose to use shared EC2 instances and utilize Microsoft's License Mobility program or they can purchase EC2 Dedicated Hosts and utilize physically dedicated hardware.

- **EC2 instances with a default tenancy**
  Microsoft's License Mobility Program allows qualifying customers to bring in eligible Microsoft

software onto AWS default tenancy servers. The AWS License Mobility Page is a great place to start the process. You will need to fill out the appropriate License Mobility forms and file them with Microsoft to ensure that the licenses are able to be imported into AWS.

- **Dedicated**
  If you bring existing licenses to EC2 Dedicated Hosts, then you are using hardware that is physically dedicated to your use. In that case, the outsourcing language within the Microsoft Product Terms applies. Visit the Dedicated Hosts detail page for more information on Dedicated Hosts.

Can I use License Mobility?

Yes. License Mobility is a benefit available to Microsoft Volume Licensing customers with eligible server applications covered by active Microsoft Software Assurance (SA). License Mobility allows customers to move eligible Microsoft software to third party cloud providers such as AWS for the end use on EC2 instances with a default tenancy. It is important to note that you may not need license mobility if you are using your own licenses on EC2 Dedicated Hosts or EC2 Dedicated Instances. For additional details, see the Microsoft License Mobility page on the AWS site. Qualifying customers with Software Assurance can bring in their own licenses of user based products as long as they comply with the terms of the License Mobility program.

Licensing – Other Considerations

Can I use my own SPLA?

We have many customers and partners that have their own SPLA and utilize AWS. Companies that have a signed SPLA with Microsoft are governed by the Services Provider Use Rights (SPUR). The SPUR describes exactly how customers can outsource their infrastructure to AWS. All per core and per processor licenses must be bought from AWS. All user based licenses must be reported on the customer’s own SPLA.

What are Self Hosting rights?

ISVs can choose to utilize self-hosting rights with Microsoft as part of their Enterprise Agreement (EA). This allows them to take advantage of pricing that they have negotiated with Microsoft under their EA. However, Microsoft requires that customers not mix self-hosting rights and SPLA for each application. This means that customers that have self-hosting must purchase Dedicated Hosts from AWS and bring in all Microsoft licenses including Windows Server. Microsoft wrote a licensing brief about self-hosting rights which you can find here.

Can Microsoft BizSpark licenses be used on AWS?

No, at this time new BizSpark licenses cannot be used on AWS. We encourage startups to try AWS Activate, with benefits including usage credits, support, training and more.

Amazon EC2 for Windows Server

How frequently does AWS patch Windows AMIs?

AWS provides updated, fully patched Windows AMIs within 5 business days of Microsoft’s patch Tuesday (second Tuesday of each month).

What happens with previously published AMIs?

AWS deprecates previously published Windows and SQL Server AMIs within 10 business days after a new set of AMIs is published.

How do I know I'm launching the latest AWS published AMI?

When publishing new Windows AMIs, AWS follows a consistent naming scheme. For example, Windows_Server-2012-R2_RTM-English-64Bit-Base-2014.05.20. Look for the date stamp in the AMI name. You find the date stamp (last 8 digits) at the end of the AMI name.

Windows Server 2016

What’s new in Windows Server 2016?

Windows Server 2016 is Microsoft’s newest release of Windows Server. Windows Server 2016 comes loaded with a variety of powerful new features including support for Docker and Windows Containers. The release also features a Nano Server deployment option that boots faster than the Standard Edition and uses a fraction of the disk space. By running Windows Server 2016 on Amazon EC2, users can leverage the performance and elasticity of AWS to get up and running on this new release.

How is AWS supporting Windows Server 2016?

AWS is releasing several new AMIs, including Windows Server 2016, Nano Server, Windows Server 2016 with Containers and Windows Server 2016 with SQL Server 2016.

How is Nano Server different from Windows Server 2016?

Nano Server is optimized to run cloud-hosted applications and containers. Compared to

Windows Server 2016, it starts faster, requires fewer updates, consumes far less disk space, presents less surface area for security threats, and only runs 64-bit applications, tools, and agents. Nano Server has no graphical user interface – all administration is done remotely via PowerShell or WMI.

How is the EC2 Console experience different for Nano Server?

For Nano Server, Get Instance Screenshot and System Log views are supported, however given Nano Server is headless, Connect via RDP is not. Instead, users can administer a running Nano Server instance via PowerShell remoting, via PowerShell CIM sessions over WinRM, or via Windows Remote Management.

Can I create my own images from Windows Server 2016 and Nano Server instances?

Yes, you can create customized AMIs from Windows Server 2016 and Nano Server instances. As a best practice, AWS recommends generalizing an image by running sysprep when creating a new Windows AMI, and this continues to be true for Windows Server 2016. However, sysprep is not included in Nano Server, meaning image generalization is not available when creating a Windows AMI from Nano Server. Alternately, users can customize a Nano Server instance post-launch by using Run Command, which enables configuration via remote command execution.

Are there any other significant changes regarding Windows Server 2016 AMIs?

Windows Server 2016 and Nano Server AMIs feature an all-new version of the SSM agent that replaces the functionality previously supported by the EC2Config service, thereby eliminating the need for EC2Config. With these enhancements, SSM agent now supports a number of advanced settings and launch-time configurations. More details on the new SSM agent in Windows Server 2016 and Nano Server can be found in the User Guide.

How can I run Windows containers?

Launch an instance with the new Windows Server 2016 with Containers AMI. You can find a sample walkthrough in the AWS Blog.

Now that there's a Windows 2016 with Containers AMI, are there plans for Amazon EC2 Container Service (ECS) to support Windows containers?

Yes. Amazon EC2 Container Service (ECS) will support Windows containers by the end of 2016. Please sign up here to receive more information.

What will it cost to run Windows Server 2016?

Windows Server 2016 instances are billed under standard Windows EC2 pricing.

Which EC2 instance types work best with Windows Server 2016?

Microsoft recommends a minimum of 2GB RAM – visit the EC2 Instance Types page to see which instances fit best for your application.

Can I bring my own license (BYOL) for Windows Server 2016?

You can bring your own license to AWS, subject to your licensing terms with Microsoft. UseVM Import to create a Windows Server 2016 AMI from your own copy of Windows Server 2016.

Can I upgrade my Windows Server instance to Windows Server 2016?

Yes, you can upgrade Windows instances to Windows Server 2016. Visit thispage for more details.

What AWS regions support Windows Server 2016?

Windows Server 2016 is available in all AWS regions.

Windows Server 2012 R2

What editions of Windows Server 2012 R2 are available in AMIs?

We will be releasing AMIs with Windows Server 2012 R2 Standard Edition. For details on the differences between the Windows Server Editions, please refer to the Microsoft documentation.

Will it cost more to run Windows Server 2012 R2?

No. Both On-Demand and Reserved instance pricing for Windows Server 2012 R2 is the same as the pricing for earlier versions of Windows Server available on Amazon EC2. You can view the current pricing for Amazon EC2 instances here: http://aws.amazon.com/ec2/pricing.

Which AWS regions are supported?

Windows Server 2012 R2 is available in all AWS regions.

Which Amazon EC2 instance types are supported?

At this time, all Amazon EC2 instance types are supported.

What languages are available?

We support 19 languages with the Windows Server 2012 R2 AMIs. Current list of supported languages: Brazilian Portuguese, Chinese Simplified, Chinese Traditional, Czech, Dutch, English, French, German, Hungarian, Italian, Japanese, Korean, Polish, Russian, Spanish, Swedish, and Turkish.

How do I deploy my applications running Windows Server 2012 R2 to AWS?

You can use AWS Elastic Beanstalk to deploy and manage your applications on Windows Server 2012 R2 in the AWS cloud. Additionally, you can deploy directly to Amazon EC2 instances launched from the EC2 console or the AWS Marketplace. Also, you can use the AWS Toolkit for Visual Studio to get your application deployed and running in a few clicks.

Which SQL Server version/edition and languages are available with Windows Server 2012 R2 AMIs?

The following SQL Server languages, version and editions are available with Windows Server 2012 R2 AMI: English, Japanese and Brazilian Portuguese: SQL Server 2014 (Enterprise (English only), Express, Standard and Web editions).

Windows Server 2012 R2 has two file systems: NTFS and ReFS. Which one should I use?

ReFS was designed for file sharing workloads like sharing content or streaming videos. Windows applications like SQL Server support NTFS and will not install on a ReFS volume.

Can I create a Storage Space using an EBS volume?

Yes. EBS volumes can be used to setup a Storage Pool. The volumes can be formatted as NTFS or ReFS depending upon your application*.

How do I switch to the new Windows Server Start screen?

Move your mouse to the lower left corner, wait for the Start screen and then click to switch into the Start screen.

On previously published Windows Server AMIs I followed the steps as documentedhere to enable enhanced networking. Do I still need to do this for Windows Server 2012 R2 AMIs?

No, you don’t need to do this for the new Windows Server 2012 R2 AMIs. The AMIs provide built-in support for enhanced networking via SR-IOV on R3, C3 and I2 instances.

Microsoft Windows Server 2003

On what date will Microsoft end extended support for Windows Server 2003?

Microsoft extended support for Windows Server 2003 ends on July 14, 2015*.

Can I run my existing Windows Server 2003 instances after the end of extended support?

Yes. You can run Windows Server 2003 and Windows Server 2003 R2 instances on Amazon EC2 after Microsoft extended support ends on July 14, 2015* including instances that are running at that time.

Can I launch new Windows Server 2003 instances after the end of extended support?

You can launch new Windows Server 2003 instances on existing Amazon EC2 instance families after the end of extended support*.

Will I be able to create and launch instances from custom Windows Server 2003 AMIs after the end of extended support?

Yes. You will be able to create custom Windows Server 2003 AMIs and launch instances from those AMIs after July 14, 2015*.

Will I be able to publicly access Windows Server 2003 AMIs from the AWS Console and AWS Marketplace?

Windows Server 2003 AMIs will continue to be published and updated through the August AMI release schedule, and will be removed from the Amazon EC2 quick launch and Marketplace on September 15th, 2015. After September 15th, you will still be able to search for the published AMIs by following the instructions on this page.

Will AWS support Microsoft Windows Server 2003 after July 14, 2015?

Without the ability to receive updates or debugging assistance from Microsoft, our ability to fully resolve issues related to Windows Server 2003 will be restricted to addressing issues which do not require an OS patch. AWS will continue to offer assistance troubleshooting issues with running Windows Server 2003 on Amazon EC2.

Will I be able to import new Windows Server 2003 virtual machines after the end of extended support?

Yes. You can use VM Import to import Windows Server 2003 based VMs after July 14th, 2015.

Will Amazon EC2 continue to provide updated Windows Server 2003 and Windows Server 2003 R2 AMIs?

AWS is unable to provide security and software updates to Windows Server 2003 after extended support ends on July 14th 2015. If Microsoft provides security and software updates to the general public for Windows Server 2003 after July 14, 2015, we will provide them to you via an updated AMI.

Can I build custom AMIs that contain updates provided by Microsoft through a custom support agreement?

Yes. AWS customers can create and launch custom AMIs for their own use, including if those AMIs contain updates resulting from a custom support agreement. AWS customers may not redistribute any custom support updates, however.

Will AWS support Microsoft applications, such as Microsoft SQL Server 2005, running on Windows Server 2003 after the end of extended support?

AWS will continue to offer assistance troubleshooting applications that are still within the Microsoft extended support phase. There is no change in the way applications running on Windows Server are supported. However if the issue requires a patch or OS-level troubleshooting support from Microsoft, the AWS support team may not be able to fully resolve your issue. Please visit the AWS Support page for more details.

Does AWS support in-place OS upgrades for my Windows Server 2003 instances?

Yes, for details on how to perform OS Upgrades on your Amazon EC2 instances, please visit the following page for more details.

Will my scripted references to existing Windows Server 2003 AMIs continue to work after the end of extended support?

On September 15, 2015, the AWS-published Windows Server 2003 AMIs will be removed from the quick start in the instance launch wizard but will still be accessible by following the instructions on this page. The AMI IDs of your custom AMIs will not be changed.

How can I keep up-to-date with Windows Server 2003 security related information?

We encourage you to visit the AWS Security Center to learn about security in the AWS cloud. You can also subscribe to our Security Bulletin RSS Feed to keep abreast of security announcements.

** Following July 14, 2015, Microsoft's extended support for Microsoft Windows Server 2003 will* *end. Because of this, instances running Windows Server 2003 may have a higher risk of failure, security issues, incompatibility, or non-functionality. AWS will continue to offer you use of Windows Server 2003 within Amazon EC2, with an understanding that there may be an increasing number of issues that cannot be diagnosed or resolved, and therefore there is a risk that your Windows Server 2003 instances will lose their functionality entirely.*

Microsoft SharePoint Server

Does AWS offer SharePoint instances?

No, AWS does not offer SharePoint instances at this time.

How can I run SharePoint on AWS?

You can run SharePoint on AWS by deploying eligible licenses with active Software Assurance through Microsoft's License Mobility program. Learn more at http://aws.amazon.com/windows/resources/licensemobility/.

What is Microsoft License Mobility?

Microsoft License Mobility through Software Assurance allows Microsoft customers to more easily move current on-premises Microsoft Server application workloads to Amazon Web Services (AWS), without any additional Microsoft software license fees. This benefit is available to Microsoft Volume Licensing (VL) customers with eligible server applications covered by active Microsoft Software Assurance (SA) contracts. Learn more at http://aws.amazon.com/windows/resources/licensemobility/.

What if I do not have Software Assurance on the Licenses?

Please contact your Microsoft Large Account Reseller (LAR) for options on how to purchased and/or add Software Assurance to existing licenses.

How do SharePoint licenses on AWS work?

One SharePoint license can be assigned to one AWS instance (no max/min size).

How do I use an SQL Instance with SharePoint on AWS?

Customers can run their existing SQL licenses per the License Mobility program or they can run on an AWS SQL instance. For more information on SQL instances running on Amazon EC2, including pricing, please visit http://aws.amazon.com/windows/products/ec2.

AWS Management Pack for Microsoft System Center

What is the AWS Management Pack for Microsoft System Center?

The AWS Management Pack is an extension to Microsoft System Center Operations Manager that enables you to view and monitor your AWS resources directly in the Operations Manager console. This way, you get a single pane of glass to view and monitor your resources, whether they are on-premises or in the AWS cloud.

Which AWS resources can I monitor using the AWS Management Pack?

You can monitor following AWS resources using the AWS Management Pack:

- Amazon EC2 instances (Microsoft Windows and Linux)
- Amazon Elastic Block Store (EBS) volumes
- Elastic Load Balancing
- AWS CloudFormation stacks
- AWS Beanstalk applications

All the default Amazon CloudWatch metrics for these resources—and any Amazon CloudWatch alarms associated with them—are surfaced as performance counters and alerts in Operations Manager.

Which versions of System Center Operations Manager can I use?

The AWS Management Pack is available for “System Center 2012 – Operations Manager” and “System Center Operations Manager 2007 R2”.

Can I monitor AWS resources that are in different AWS regions?

Yes. The management pack gives you a consolidated view of your resources across multiple regions and Availability Zones.

Can I monitor AWS resources that are in Amazon Virtual Private Cloud (VPC)?

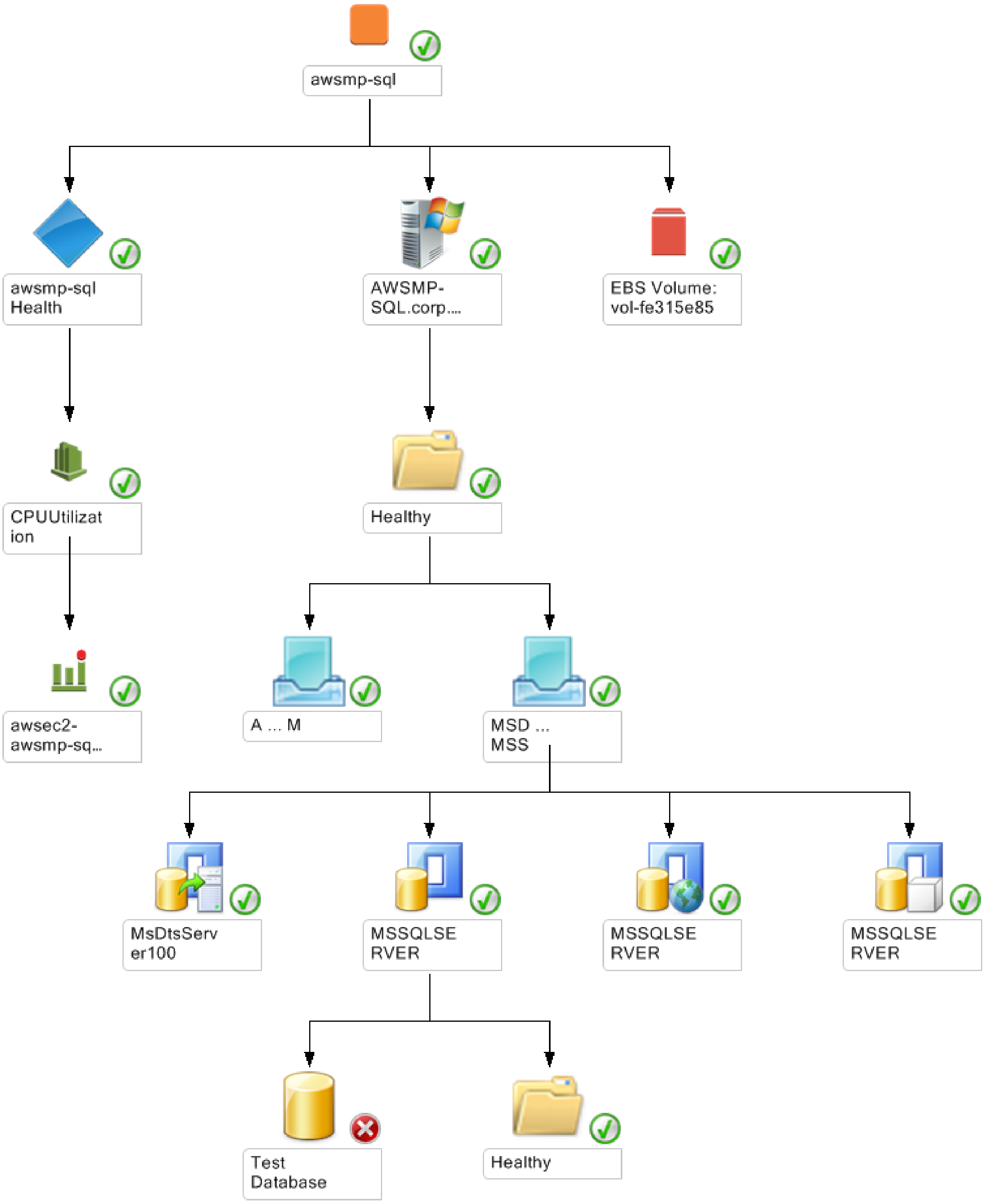
Yes. The management pack gives you a consolidated view of your resources running in Amazon VPC and Amazon EC2.

Can I monitor AWS resources from multiple AWS accounts?

Yes. You can configure the management pack to monitor AWS resources from multiple AWS accounts. Resources from multiple AWS accounts are monitored separately instead of being consolidated in a single view.

Can I monitor applications running within Amazon EC2 instances?

Yes, provided that (a) the Amazon EC2 instances are running Operations Manager Agent, and (b) the application-specific management packs are imported in Operations Manager. This applies to Amazon EC2 instances running Microsoft Windows as well as Linux.



Can I use IAM credentials instead of AWS root account credentials to monitor AWS resources?

Yes. You can configure the AWS Management Pack to use the access key ID and secret access key of a locked-down IAM user instead of using the credentials of a fully-privileged AWS root account.

Can I use my on-premises Operations Manager for the AWS Management Pack?

Yes. You can choose to run Operations Manager either on-premises or in the AWS cloud.

Where can I find more information about the AWS Management Pack?

A comprehensive guide detailing deploying, using, customizing, and troubleshooting the AWS Management Pack is available here.

Can I use my existing System Center Licenses?

Yes, through Microsoft License Mobility.

AWS Systems Manager for Microsoft System Center Virtual Machine Manager

What is AWS Systems Manager for Microsoft System Center Virtual Machine Manager?

AWS Systems Manager for Microsoft SCVMM is a software add-in that lets you administer your AWS resources using SCVMM. You can monitor and manage your EC2 for Windows instances in the AWS Cloud, as well as on-premises virtual machines—from a single console.

What can I do with Systems Manager for SCVMM?

You can list and view EC2 for Windows instances in any region. You can also start, stop, reboot, and terminate instances, as well as connect via RDP.

Which versions of SCVMM can be used with Systems Manager for SCVMM?

You can use AWS Systems Manager with SCVMM 2012 SP1 and later.

Where do I download Systems Manager for SCVMM?

You can download the add-inhere.

How much does Systems Manager for SCVMM cost?

There is no additional cost to download, install or use Systems Manager for SCVMM.

How is this different from the AWS Management Pack for Microsoft System Center?

The AWS Management Pack is used for monitoring and reporting on the performance of EC2 for

Windows instances, whereas AWS Systems Manager lets you start, stop, reboot and terminate instances.

AWS Diagnostics for Microsoft Windows Server

What is the AWS Diagnostics for Microsoft Windows Server?

AWS Diagnostics for Microsoft Windows Server is a standalone executable that can be run on an EC2 Windows Server instance for diagnosis and troubleshooting as well as proactively checking for possible issues. The tool includes a data collector module that collects debug information and packages it in a zip file. It also includes an analyzer module that can analyze the log files collected by the data collector module and parse them based on predefined rule.

When should I use this tool?

If you ever run into issues with your EC2 Windows Server instance and want to do a preliminary check to eliminate some of the known problems, this tool will be a great starting point. It will also relieve the pain point of trying to collect a bunch of different log files manually, especially when you are working with our technical support.

What data will be collected?

Here’s a sample list of data that will be collected. For a comprehensive list, please refer to the user manual

- Network Information, including IP address and route table
- Domain and computer name
- Activation settings, including license status and Key Management Server (KMS) configuration
- Time settings, including current time and time zone
- Installed drivers on the instance
- Windows firewall settings, along with security group rules
- Installed updates
- Mini dump files if Windows has crashed within a week

What are some of the analysis rules?

Here’s a sample list of analysis rules. For a comprehensive list, please refer to the user manual

- Check for activation status and KMS settings
- Check for proper route table entries for metadata and KMS access
- Comparison of AWS Security group rules vs. Windows firewall rules
- Version check for PV driver (Redhat or Citrix)
- Check if the RealTimesUniversal registry key is set
- Default gateway settings if using multiple NICs
- Bug check code in mini dump files

How much does the AWS Diagnostics for Microsoft Windows Server cost?

There is no additional charge for AWS Diagnostics for Microsoft Windows Server.

My EC2 Windows Server just crashed. Can I use this tool to collect log files from the EBS boot volume for that instance?

Yes. In order to do that you will have to detach the volume from the crashed instance and attach it to an instance on which the tool is running. The tool will then parse through the directory structure of the attached EBS volume that you select and will collect the necessary information.

Where can I find more information about the AWS Diagnostic Tool for Microsoft Windows Server?

A comprehensive guide for AWS Diagnostic Tool for Microsoft Windows Server is availablehere.

Other Questions

Will customers have to recreate their environment using other technologies in order to receive support from AWS or Microsoft?

No. Customers can receive support running Microsoft workloads on AWS from both AWS and Microsoft under the customer’s support agreements with AWS or Microsoft without having to recreate their environment using other technologies. In the very rare case a problem could not be duplicated, AWS would work with the customer to recreate the issue in a Microsoft validated environment.

Is AWS SVVP Validated?

AWS does not need to be SVVP validated for customers to be fully supported running Microsoft workloads on AWS. As Microsoft explains: “SVVP does not apply to vendors that are hosting Windows Server or other Microsoft products through the Microsoft Service Provider License Agreement Program (SPLA). Support for SPLA customers is provided under the SPLA agreement by the SPLA hoster.” (see http://www.windowsservercatalog.com/svvp.aspx).

Without SVVP Validation, are Microsoft products fully supported in the AWS environment?

Yes. SVVP validation is not applicable to SPLA providers. Support for SPLA customers is provided under the SPLA agreement by AWS. AWS is fully committed to supporting our customers running Microsoft workloads on AWS.

Amazon EC2 Container Service FAQ

General

Q: What is Amazon EC2 Container Service?
Amazon EC2 Container Service (ECS) is a highly scalable, high performance container management service that supports Docker containers and allows you to easily run applications on a managed cluster of Amazon EC2 instances. Amazon ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure. With simple API calls, you can launch and stop container-enabled applications, query the complete state of your cluster, and access many familiar features like security groups, Elastic Load Balancing, EBS volumes and IAM roles. You can use Amazon ECS to schedule the placement of containers across your cluster based on your resource needs and availability requirements. You can also integrate your own scheduler or third-party schedulers to meet business or application specific requirements.

Q: Why should I use Amazon ECS?
Amazon ECS makes it easy to use containers as a building block for your applications by eliminating the need for you to install, operate, and scale your own cluster management infrastructure. Amazon ECS lets you schedule long-running applications, services, and batch processes using Docker containers. Amazon ECS maintains application availability and allows you to scale your containers up or down to meet your application's capacity requirements. Amazon ECS is integrated with familiar features like Elastic Load Balancing, EBS volumes, VPC, and IAM. Simple APIs let you integrate and use your own schedulers or connect Amazon ECS into your existing software delivery process.

Q: What is the pricing for Amazon ECS?
There is no additional charge for Amazon ECS. You pay for AWS resources (e.g. EC2 instances or EBS volumes) you create to store and run your application. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments.

Q: How is Amazon ECS different from AWS Elastic Beanstalk?
AWS Elastic Beanstalk is an application management platform that helps customers easily deploy and scale web applications and services. It keeps the provisioning of building blocks (e.g., EC2, RDS, Elastic Load Balancing, Auto Scaling, CloudWatch), deployment of applications, and health monitoring abstracted from the user so they can just focus on writing code. You simply specify which container images are to be deployed, the CPU and memory requirements, the port mappings, and the container links. Elastic Beanstalk will automatically handle all the details such as provisioning an Amazon ECS cluster, balancing load, auto-scaling, monitoring, and placing your containers across your cluster.

Elastic Beanstalk is ideal if you want to leverage the benefits of containers but just want the simplicity of deploying applications from development to production by uploading a container image. You can work with Amazon ECS directly if you want more fine-grained control for custom application architectures.

Q: How is Amazon ECS different from AWS Lambda?
Amazon EC2 Container Service is a highly scalable Docker container management service that allows you to run and manage distributed applications that run in Docker containers. AWS Lambda is an event-driven task compute service that runs your code in response to “events” such as changes in data, website clicks, or messages from other AWS services without you having to manage any compute infrastructure.

Using Amazon EC2 Container Service

Q: How do I get started using Amazon ECS?
Visit our Getting Started page for more information on how to start.

Q: Does EC2 Container Service support any other container types?
No. Docker is the only container platform supported by EC2 Container Service at this time.

Q: I want to launch containers. Why do I have to launch Tasks?
Docker encourages you to split your applications up into their individual components, and EC2 Container Service is optimized for this pattern. Tasks allow you to define a set of containers that you would like to be placed together (or part of the same placement decision), their properties, and how they may be linked. Tasks include all the information that EC2 Container Service needs to make the placement decision. To launch a single container, your Task Definition should only include one container definition.

Q: Does Amazon ECS support applications and services?
Yes. The Amazon ECS Service scheduler can manage long-running applications and services. The Service scheduler helps you maintain application availability and allows you to scale your containers up or down to meet your application's capacity requirements. The Service scheduler allows you to distribute traffic across your containers using Elastic Load Balancing. Amazon ECS will automatically register and deregister your containers from the associated load balancer. The Service scheduler will also automatically recover containers that become unhealthy (fail ELB health checks) or stop running to ensure you have the desired number of healthy containers supporting your application. You can scale your application up and down by changing the number of containers you want the service to run. You can update your application by changing its definition or using a new image. The scheduler will automatically start new containers using the new definition and stop containers running the previous version (waiting for the ELB connections to drain if ELB is used).

Q: Does Amazon ECS support dynamic port mapping?
Yes. It is possible to associate a service on Amazon EC2 Container Service (ECS) to an Application Load Balancer for the Elastic Load Balancing (ELB) service. The Application Load Balancer supports a target group that contains a set of instance:ports. You can specify a dynamic port in the ECS task definition which gives the container an unused port when it is scheduled on the EC2 instance. The ECS scheduler will automatically add the task to the Application Load Balancer’s target group using this port.

Q: Does Amazon ECS support batch jobs?
Yes. You can use Amazon ECS Run task to run one or more tasks once. Run task starts the task on an instance that meets the task’s requirements including CPU, memory and ports.

Q: Can I use my own scheduler with Amazon ECS?
Yes. You can use the Describe* APIs to get information about the complete state of your cluster. The APIs return data on all the container instances in a cluster, what tasks they're running, and what resources are still available. With this information, you can use the StartTask API to target specific container instances in your cluster or use a custom scheduler to manage placement based on your requirements.

Q: Can I use my own AMI?
Yes. You can use any AMI that meets the Amazon ECS AMI specification. We recommend starting from the Amazon ECS-enabled Amazon Linux AMI. Partner AMIs compatible with Amazon ECS are also available. You can review the Amazon ECS AMI specification in the documentation.

Q: How can I configure my container instances to pull from Amazon EC2 Container Registry?
Amazon ECR is integrated with Amazon ECS allowing you to easily store, run, and manage container images for applications running on Amazon ECS. All you need to do is specify the

Amazon ECR repository in your Task Definition and attach the AmazonEC2ContainerServiceforEC2Role to your instances. Then Amazon ECS will retrieve the appropriate images for your applications.

Security

Q: How does Amazon ECS isolate containers belonging to different customers?
Amazon ECS schedules containers for execution on customer-controlled Amazon EC2 instances and builds on the same isolation controls and compliance that are available for EC2 customers.

- Your compute instances are located in a Virtual Private Cloud (VPC) with an IP range that you specify. You decide which instances are exposed to the Internet and which remain private.
- Your EC2 instances use an IAM role to access the ECS service.
- Your ECS tasks use an IAM role to access services and resources.
- Security Groups and networks ACLs allow you to control inbound and outbound network access to and from your instances.
- You can connect your existing IT infrastructure to resources in your VPC using industry-standard encrypted IPsec VPN connections.
- You can provision your EC2 resources as Dedicated Instances. Dedicated Instances are Amazon EC2 Instances that run on hardware dedicated to a single customer for additional isolation.

Q: Can I apply additional security configuration and isolation frameworks to my container instances?
Yes. As an Amazon EC2 customer, you have root access to the operating system of your container instances, enabling you to take ownership of the operating system's security settings as well as load and configure additional software components for security capabilities such as monitoring, patch management, log management and host intrusion detection.

Q: Can I operate container instances with different security settings or segregate different tasks across different environments?
Yes. You can configure your different container instances using the tooling of your choice. Amazon ECS allows you to control the placement of tasks in different container instances through the construct of clusters and targeted launches.

Q: Does Amazon ECS support retrieving Docker images from a private or internal source?
Yes. Customers can configure their container instances to access a private Docker image registry within a VPC or a registry that's accessible outside a VPC such as the Amazon EC2 Container Registry.

Q: How do I configure IAM roles for ECS tasks?
You first need to create an IAM role for your task, using the 'Amazon EC2 Container Service Task Role' service role and attaching a policy with the required permissions. When you create a new task definition or a task definition revision you can then specify a role by selecting it form the 'Task Role' drop-down or using the 'taskRoleArn' filed in the JSON format.

AWS Elastic Beanstalk FAQ

General

Q: What is AWS Elastic Beanstalk?
AWS Elastic Beanstalk makes it even easier for developers to quickly deploy and manage applications in the AWS Cloud. Developers simply upload their application, and Elastic Beanstalk automatically handles the deployment details of capacity provisioning, load balancing, auto-scaling, and application health monitoring.

Q: Who should use AWS Elastic Beanstalk?
Those who want to deploy and manage their applications within minutes in the AWS Cloud. You don't need experience with cloud computing to get started. AWS Elastic Beanstalk supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker web applications.

Q: Which languages and development stacks does AWS Elastic Beanstalk support?
AWS Elastic Beanstalk supports the following languages and development stacks:

- Apache Tomcat for Java applications
- Apache HTTP Server for PHP applications
- Apache HTTP Server for Python applications
- Nginx or Apache HTTP Server for Node.js applications
- Passenger or Puma for Ruby applications
- Microsoft IIS 7.5, 8.0, and 8.5 for .NET applications
- Java SE
- Docker

- Go

See Supported Platforms for a complete, up-to-date list of supported language and development stacks.

Q: Will AWS Elastic Beanstalk support other languages?
Yes. AWS Elastic Beanstalk is designed so that it can be extended to support multiple development stacks and programming languages in the future. AWS is working with solution providers on the APIs and capabilities needed to create additional Elastic Beanstalk offerings.

Q: What can developers now do with AWS Elastic Beanstalk that they could not before?
AWS Elastic Beanstalk automates the details of capacity provisioning, load balancing, auto scaling, and application deployment, creating an environment that runs a version of your application. You can simply upload your deployable code (e.g., WAR file), and AWS Elastic Beanstalk does the rest. The AWS Toolkit for Visual Studio and the AWS Toolkit for Eclipse allow you to deploy your application to AWS Elastic Beanstalk and manage it without leaving your IDE. Once your application is running, Elastic Beanstalk automates management tasks–such as monitoring, application version deployment, a basic health check–and facilitates log file access. By using Elastic Beanstalk, developers can focus on developing their application and are freed from deployment-oriented tasks, such as provisioning servers, setting up load balancing, or managing scaling.

Q: How is AWS Elastic Beanstalk different from existing application containers or platform-as-a-service solutions?
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 predetermined by the vendor–with little to no opportunity to take back control over various parts of their application's infrastructure. However, with AWS Elastic Beanstalk, developers retain full control over the AWS resources powering their application. If developers decide they want to manage some (or all) of the elements of their infrastructure, they can do so seamlessly by using Elastic Beanstalk's management capabilities.

Q: What elements of my application can I control when using AWS Elastic Beanstalk?
With AWS Elastic Beanstalk, you can:

- Select the operating system that matches your application requirements (e.g., Amazon Linux or Windows Server 2012 R2)
- Choose from several available database and storage options
- Enable login access to Amazon EC2 instances for immediate and direct troubleshooting
- Quickly improve application reliability by running in more than one Availability Zone
- Enhance application security by enabling HTTPS protocol on the load balancer

- Access built-in Amazon CloudWatch monitoring and getting notifications on application health and other important events
- Adjust application server settings (e.g., JVM settings) and pass environment variables
- Run other application components, such as a memory caching service, side-by-side in Amazon EC2
- Access log files without logging in to the application servers

Q: What are the Cloud resources powering my AWS Elastic Beanstalk application?
AWS Elastic Beanstalk uses proven AWS features and services, such as Amazon EC2, Amazon RDS, Elastic Load Balancing, Auto Scaling, Amazon S3, and Amazon SNS, to create an environment that runs your application. The current version of AWS Elastic Beanstalk uses the Amazon Linux AMI or the Windows Server 2012 R2 AMI.

Q: What kinds of applications are supported by AWS Elastic Beanstalk?
AWS Elastic Beanstalk supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker, and is ideal for web applications. However, due to Elastic Beanstalk's open architecture, non-web applications can also be deployed using Elastic Beanstalk. We expect to support additional application types and programming languages in the future. See Supported Platforms to learn more.

Q: Which operating systems does AWS Elastic Beanstalk use?
AWS Elastic Beanstalk runs on the Amazon Linux AMI and the Windows Server 2012 R2 AMI. Both AMIs are supported and maintained by Amazon Web Services and are designed to provide a stable, secure, and high-performance execution environment for Amazon EC2 Cloud computing.

Q: How quickly will my application start running?
It typically takes a few minutes to create the AWS resources to run your application, measured from the time you upload your application version (e.g., WAR file, ASP.NET files, Node.js files, PHP files, Python files, or Ruby files) to when it is fully deployed and accessible to your users. This time is dependent on a number of factors, including the size of your deployable code and the number of application servers you are deploying.

Q: How quickly will my application get updated?
Deploying new application versions to existing resources (e.g., environments) happens much faster (typically under a minute) and is mostly dependent on the size of the new application version.

Q: How quickly can my application scale up and down?
AWS Elastic Beanstalk provides a truly elastic environment using Auto Scaling. Your application can be configured to automatically scale tens or even hundreds of times based on thresholds, such as CPU utilization or network bandwidth. These thresholds can be easily configured for

your specific application using the Elastic Beanstalk console. With Elastic Beanstalk, you don't have to worry if you will be able to scale quickly to handle peaks in traffic or users, nor if you will be forced to pay for resources that you don't need.

Q: Can I have multiple versions of my application running at the same time?
Yes. AWS Elastic Beanstalk is designed to support multiple running environments, such as one for integration testing, one for pre-production, and one for production. Each environment is independently configured and runs on its own separate AWS resources. Elastic Beanstalk also stores and tracks application versions over time, so an existing environment can be easily rolled back to a prior version or a new environment can be launched using an older version to try and reproduce a customer problem.

Q: How many applications can I run with AWS Elastic Beanstalk?
You can create up to 75 applications and 1,000 application versions. By default, you can run up to 200 environments across all of your applications. If you are also using AWS outside of AWS Elastic Beanstalk, you may not be able to create 10 environments since other limits may be hit sooner. For example, the default AWS account limits allow you to launch up to 20 EC2 instances and create up to 10 elastic load balancers. If you need more resources, complete the AWS Elastic Beanstalk request form, and your request will be promptly evaluated.

Q: Can I use AWS Elastic Beanstalk to deploy applications that must be highly available?
Yes. To do this, you edit your environment configuration settings, select 2 or more instances for Auto Scaling minimum, and set Multiple Availability Zones to “Any 2”. AWS Availability Zones are designed to be physically distinct, fail independently, and be reliable.

Q: What happens if my application stops responding to requests?
AWS Elastic Beanstalk applications are protected against failures in the underlying infrastructure. If an Amazon EC2 instance fails for any reason, AWS Elastic Beanstalk will use Auto Scaling to automatically launch a new instance. Elastic Beanstalk can also detect if your application is not responding on the custom URL even though the underlying infrastructure appears healthy, and will log that as an environment event (e.g., a bad version was deployed) so you can take appropriate action.

Q: Which AWS Regions is AWS Elastic Beanstalk available in?

Please refer to Regional Products and Services for details of Elastic Beanstalk availability by Region.

Q: How do I access AWS Elastic Beanstalk?
You can use the AWS Management Console, the AWS Elastic Beanstalk command line interface (CLI), the AWS Toolkit for Visual Studio, the AWS Toolkit for Eclipse, the AWS Elastic Beanstalk API, or AWS SDKs.

Q: Can I use an integrated development environment, like Eclipse or Microsoft Visual Studio?

Yes. You can use Eclipse and Visual Studio to deploy your application to AWS Elastic Beanstalk. You can use the AWS Toolkit for Eclipse for Java applications and the AWS Toolkit for Visual Studio for .NET applications. The toolkits allow you to develop your application, deploy it to Elastic Beanstalk, and even test it out without having to switch your focus away from your IDE.

Getting Started

Q: How do I sign up for AWS Elastic Beanstalk?

To sign up for AWS Elastic Beanstalk, choose the**Sign Up Now** button on the Elastic Beanstalk detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Elastic Beanstalk process. After signing up, please refer to the AWS Elastic Beanstalk Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for AWS Elastic Beanstalk?

AWS Elastic Beanstalk registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a few minutes and involves receiving an automated phone call during the registration process and entering a PIN number using the phone key pad.

Q: How do I get started after I have signed up?

The best way to get started with AWS Elastic Beanstalk is to work through theAWS Elastic Beanstalk Getting Started Guide, part of our technical documentation. Within a few minutes, you will be able to deploy and use a sample application or upload your own application.

Q: Is there a sample application that I can use to check out AWS Elastic Beanstalk?

Yes. AWS Elastic Beanstalk includes a sample application that you can use to test drive the offering and explore its functionality.

Databases and Storage

Q: Does AWS Elastic Beanstalk store anything in Amazon S3?

Yes. AWS Elastic Beanstalk stores your application files and, optionally, server log files in Amazon S3. If you are using the AWS Management Console, the AWS Toolkit for Visual Studio, or AWS Toolkit for Eclipse, an Amazon S3 bucket will be created in your account for you and the files you upload will be automatically copied from your local client to Amazon S3. Optionally, you may configure Elastic Beanstalk to copy your server log files every hour to Amazon S3. You do this by editing the environment configuration settings.

Q: Can I use Amazon S3 to store application data, like images?

Yes. You can use Amazon S3 for application storage. The easiest way to do this is by including the AWS SDK as part of your application's deployable file. For example, you can include the AWS SDK for Java as part of your application's WAR file.

Q: What database solutions can I use with AWS Elastic Beanstalk?

AWS Elastic Beanstalk does not restrict you to any specific data persistence technology. You can choose to use Amazon Relational Database Service (Amazon RDS) or Amazon DynamoDB, or use Microsoft SQL Server, Oracle, or other relational databases running on Amazon EC2.

Q: How do I set up a database for use with AWS Elastic Beanstalk?

Elastic Beanstalk can automatically provision an Amazon RDS DB instance. The information about connectivity to the DB instance is exposed to your application by environment variables. To learn more about how to configure RDS DB instances for your environment, see the Elastic Beanstalk Developer Guide.

Q: Does this mean I need to modify the application code when moving from test to production?

Not with AWS Elastic Beanstalk. With Elastic Beanstalk, you can specify the connection information in the environment configuration. By extracting the connection string from the application code, you can easily configure different Elastic Beanstalk environments to use different databases.

Security

Q: How do I make my application private?

By default, your application is available publicly at myapp.elasticbeanstalk.com for anyone to access. You can use Amazon VPC to provision a private, isolated section of your application in a virtual network that you define. This virtual network can be made private through specific security group rules, network ACLs, and custom route tables. You can also easily control what other incoming traffic, such as SSH, is delivered or not to your application servers by changing the EC2 security group settings.

Q: Can I run my application inside a Virtual Private Cloud (VPC)?

Yes, you can run your applications in a VPC. For more details, see theAWS Elastic Beanstalk

Developer Guide.

Q: Where can I find more information about security and running applications on AWS?

For more information about security on AWS, please refer to ourAmazon Web Services: Overview of Security Processes document and visit our Security Center.

Q: Is it possible to use Identity & Access Management (IAM) with AWS Elastic Beanstalk?

Yes. IAM users with the appropriate permissions can now interact with AWS Elastic Beanstalk.

Q: Why should I use IAM with AWS Elastic Beanstalk?

IAM allows you to manage users and groups in a centralized manner. You can control which IAM users have access to AWS Elastic Beanstalk, and limit permissions to read-only access to Elastic Beanstalk for operators who should not be able to perform actions against Elastic Beanstalk resources. All user activity within your account will be aggregated under a single AWS bill.

Q: How do I create IAM users?

You can use the IAM console, IAM command line interface (CLI), or IAM API to provision IAM users. By default, IAM users have no access to AWS services until permissions are granted.

Q: How do I grant an IAM user access to AWS Elastic Beanstalk?

You can grant IAM users access to services by using policies. To simplify the process of granting access to AWS Elastic Beanstalk, you can use one of the policy templates in the IAM console to help you get started. Elastic Beanstalk offers two templates: a read-only access template and a full-access template. The read-only template grants read access to Elastic Beanstalk resources. The full-access template grants full access to all Elastic Beanstalk operations, as well as permissions to manage dependent resources, such as Elastic Load Balancing, Auto Scaling, and Amazon S3. You can also use the AWS Policy Generator to create custom policies. For more details, see the AWS Elastic Beanstalk Developer Guide.

Q: Can I restrict access to specific AWS Elastic Beanstalk resources?

Yes. You can allow or deny permissions to specific AWS Elastic Beanstalk resources, such as applications, application versions, and environments.

Q: Who gets billed for the AWS resources that an IAM user creates?

All resources created by IAM users under a root account are owned and billed to the root account.

Q: Who has access to an AWS Elastic Beanstalk environment launched by an IAM user?

The root account has full access to all AWS Elastic Beanstalk environments launched by any IAM user under that account. If you use the Elastic Beanstalk template to grant read-only access to an IAM user, that user will be able to view all applications, application versions, environments, and any associated resources in that account. If you use the Elastic Beanstalk template to grant full access to an IAM user, that user will be able to create, modify, and terminate any Elastic Beanstalk resources under that account.

Q: Can an IAM user access the AWS Elastic Beanstalk console?

Yes. An IAM user can access the AWS Elastic Beanstalk console using their username and password.

Q: Can an IAM user call the AWS Elastic Beanstalk API?

Yes. An IAM user can use their access key and secret key to perform operations using the Elastic Beanstalk API.

Q: Can an IAM user use the AWS Elastic Beanstalk command line interface?

Yes. An IAM user can use their access key and secret key to perform operations using the AWS Elastic Beanstalk command line interface (CLI).

Managed Platform Updates

Q: How can I keep the underlying platform of the environment running my application automatically up-to-date?

You can opt-in to having your AWS Elastic Beanstalk environments automatically updated to the latest version of the underlying platform running your application during a specified maintenance window. Elastic Beanstalk regularly releases new versions of supported platforms (Java, PHP, Ruby, Node.js, Python, .NET, Go, and Docker) with operating system, web and application server, and language and framework updates.

Q: How can I get started with managed platform updates?

To let Elastic Beanstalk automatically manage your platform updates, you must enable managed platform updates in the Configuration tab of the Elastic Beanstalk console or use the EB CLI or API. After you have enabled the feature, you can configure which types of updates to allow and when updates can occur.

Q: What kinds of platform version updates will managed platform updates apply?

AWS Elastic Beanstalk can automatically perform platform updates for new patch and minor platform versions. Elastic Beanstalk will not automatically perform major platform version updates (e.g., Java 7 Tomcat 7 to Java 8 Tomcat 8) because they include backwards incompatible changes and require additional testing. In these cases, you must manually initiate the update.

Q: How does AWS Elastic Beanstalk distinguish between “major,” “minor,” and “patch”

version releases?

AWS Elastic Beanstalk platforms are versioned using this pattern: MAJOR.MINOR.PATCH (e.g., 2.0.0). Each portion is incremented as follows:

- MAJOR version when there are incompatible changes.
- MINOR version when there is additional functionality added in a backward-compatible manner.
- PATCH version when there are backward-compatible bug fixes.

Q: When and how can I perform major version updates?

You can perform major version updates at any time using the AWS Elastic Beanstalk management console, API, or CLI. You have the following options to perform a major version update:

- Apply the update in-place on an existing environment. SeeUpdating Your Elastic Beanstalk Environment's Platform Version.
- Create a clone of an existing environment with the new platform version. SeeClone an Environment to learn more.

Q: How does Elastic Beanstalk apply managed platform updates?

The updates are applied using an immutable deployment mechanism that ensures that no changes are made to the existing environment until a parallel fleet of Amazon EC2 instances, with the updates installed, is ready to be swapped with the existing instances, which are then terminated. In addition, if the Elastic Beanstalk health system detects any issues during the update, traffic is redirected to the existing fleet of instances, ensuring minimal impact to end users of your application.

Q: Will my application be available during the maintenance windows?

Since managed platform updates use an immutable deployment mechanism to perform the updates, your application will be available during the maintenance window and consumers of your application will not be impacted by the update.

Q: What does it cost to use managed platform updates?

There is no additional charge for the managed platform updates feature. You simply pay for the additional EC2 instances necessary to perform the update for the duration of the update.

Q: What is a maintenance window?

A maintenance window is a weekly two-hour-long time slot during which AWS Elastic Beanstalk will initiate platform updates if managed platform updates is enabled and a new version of the platform is available. For example, if you select a maintenance window that begins every Sunday at 2 AM, AWS Elastic Beanstalk will initiate the platform update sometime between 2-4 AM every Sunday. It is important to note that, depending on the configuration of your applications, updates could complete outside of the maintenance window.

The maintenance window is set on a per-environment basis, providing you the option to set different maintenance windows for your various application components or applications. This allows environment updates to be staggered if you do not want multiple pieces of your application to be updated at the same time. If you enable managed platform updates but do not specify a maintenance window, a default weekly 2-hour window will be assigned for your environment. If you want to change when maintenance is performed on your behalf, you can do so by modifying the managed update configuration in the AWS Management Console or by using the UpdateEnvironment API.

Q: How will I be notified of the availability of new platform versions?

You will be notified about the availability of new platform versions through the AWS Management Console, forum announcements, and release notes.

Q: Where can I find details of changes between platform versions?

Details on changes between platform versions can be found on the AWS Elastic Beanstalk Release Notes page.

Q: What operations can I perform on the environment while a managed update is in progress?

The only action available to you while a managed platform update is in-progress is ‘abort’. This will allow you to stop the update immediately and roll back to the previous version.

Q: Which platform version will my environment be updated to if there are multiple new versions released in between maintenance windows?

Your environment will always be updated to the latest version available based on the level (minor plus patch or patch only) you have selected.

Q: Where can I find details of all the managed platform updates that have been performed on my environment?

Details for every managed platform update are available on the events page and are tagged with an event type of “MAINTENANCE.”

Q: How often are platform version updates released?

The number of version releases in a given year varies based on the frequency and content of releases and patches from the language/framework’s vendor or core team, and the outcome of a thorough vetting of these releases and patches by our platform engineering team.

Billing

Q: How much does AWS Elastic Beanstalk cost?

There is no additional charge for AWS Elastic Beanstalk–you pay only for the AWS resources actually used to store and run your application. New AWS customers who are eligible for the AWS Free Tier may deploy an application that runs within the Free Tier using the default settings of Elastic Beanstalk.

Q: How much do the AWS resources powering my application on AWS Elastic Beanstalk cost?

You pay only for what you use, and there is no minimum fee for the use of any AWS resources. For Amazon EC2 pricing information, please visit the pricing section on the EC2 detail page. For Amazon S3 pricing information, please visit the pricing section on the S3 detail page. You can use the AWS simple calculator to estimate your bill for different application sizes.

Q: How do I check how many AWS resources have been used by my application and access my bill?

You can view your charges for the current billing period at any time on the Amazon Web Services web site by logging into your Amazon Web Services account and choosing **Account Activity** under **Your Web Services Account**.

Support

Q: Does AWS Support cover AWS Elastic Beanstalk?

Yes. AWS Support covers issues related to your use of AWS Elastic Beanstalk. For further details and pricing, see the AWS Support page.

Q: What other support options are available?

You can tap into the breadth of existing AWS community knowledge to help you with your development through the AWS Elastic Beanstalk discussion forum.

AWS Lambda FAQ

General

Q: What is AWS Lambda?

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume - there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service - all with zero administration. Just upload your code and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

Q: What events can trigger an AWS Lambda function?

Please see our documentation for a complete list of event sources.

Q: When should I use AWS Lambda versus Amazon EC2?

Amazon Web Services offers a set of compute services to meet a range of needs.

Amazon EC2 offers flexibility, with a wide range of instance types and the option to customize the operating system, network and security settings, and the entire software stack, allowing you to easily move existing applications to the cloud. With Amazon EC2 you are responsible for provisioning capacity, monitoring fleet health and performance, and designing for fault tolerance and scalability. AWS Elastic Beanstalk offers an easy-to-use service for deploying and scaling web applications in which you retain ownership and full control over the underlying EC2 instances. Amazon EC2 Container Service is a scalable management service that supports Docker containers and allows you to easily run distributed applications on a managed cluster of Amazon EC2 instances.

AWS Lambda makes it easy to execute code in response to events, such as changes to Amazon S3 buckets, updates to an Amazon DynamoDB table, or custom events generated by your applications or devices. With Lambda you do not have to provision your own instances; Lambda performs all the operational and administrative activities on your behalf, including capacity provisioning, monitoring fleet health, applying security patches to the underlying compute resources, deploying your code, running a web service front end, and monitoring and logging your code. AWS Lambda provides easy scaling and high availability to your code without additional effort on your part.

Q: What kind of code can run on AWS Lambda?

AWS Lambda offers an easy way to accomplish many activities in the cloud. For example, you can use AWS Lambda to build mobile back-ends that retrieve and transform data from Amazon DynamoDB, handlers that compress or transform objects as they are uploaded to Amazon S3, auditing and reporting of API calls made to any Amazon Web Service, and server-less processing of streaming data using Amazon Kinesis.

Q: What languages does AWS Lambda support?

AWS Lambda supports code written in Node.js (JavaScript), Python, and Java (Java 8 compatible). Your code can include existing libraries, even native ones. Please read our documentation on using Node.js, Python and Java.

Q: Can I access the infrastructure that AWS Lambda runs on?

No. AWS Lambda operates the compute infrastructure on your behalf, allowing it to perform health checks, apply security patches, and do other routine maintenance.

Q: How does AWS Lambda isolate my code?

Each AWS Lambda function runs in its own isolated environment, with its own resources and file system view. AWS Lambda uses the same techniques as Amazon EC2 to provide security and separation at the infrastructure and execution levels.

Q: How does AWS Lambda secure my code?

AWS Lambda stores code in Amazon S3 and encrypts it at rest. AWS Lambda performs additional integrity checks while your code is in use.

Q: What AWS regions are available for AWS Lambda?

Please refer to the AWS Global Infrastructure Region Table.

AWS Lambda Functions

Q: What is an AWS Lambda function?

The code you run on AWS Lambda is uploaded as a “Lambda function”. Each function has associated configuration information, such as its name, description, entry point, and resource requirements. The code must be written in a “stateless” style i.e. it should assume there is no affinity to the underlying compute infrastructure. Local file system access, child processes, and similar artifacts may not extend beyond the lifetime of the request, and any persistent state should be stored in Amazon S3, Amazon DynamoDB, or another Internet-available storage service. Lambda functions can include libraries, even native ones.

Q: Will AWS Lambda reuse function instances?

To improve performance, AWS Lambda may choose to retain an instance of your function and reuse it to serve a subsequent request, rather than creating a new copy. Your code should not assume that this will always happen.

Q: What if I need scratch space on disk for my AWS Lambda function?

Each Lambda function receives 500MB of non-persistent disk space in its own /tmp directory.

Q: Why must AWS Lambda functions be stateless?

Keeping functions stateless enables AWS Lambda to rapidly launch as many copies of the function as needed to scale to the rate of incoming events. While AWS Lambda’s programming model is stateless, your code can access stateful data by calling other web services, such as Amazon S3 or Amazon DynamoDB.

Q: Can I use threads and processes in my AWS Lambda function code?

Yes. AWS Lambda allows you to use normal language and operating system features, such as creating additional threads and processes. Resources allocated to the Lambda function, including memory, execution time, disk, and network use, must be shared among all the threads/processes it uses. You can launch processes using any language supported by Amazon Linux.

Q: What restrictions apply to AWS Lambda function code?

Lambda attempts to impose as few restrictions as possible on normal language and operating system activities, but there are a few activities that are disabled: Inbound network connections are blocked by AWS Lambda, and for outbound connections only TCP/IP sockets are supported, and ptrace (debugging) system calls are restricted. TCP port 25 traffic is also restricted as an anti-spam measure.

Q: How do I create an AWS Lambda function using the Lambda console?

You can author the code for your function using the inline editor in the AWS Lambda console. You can also package the code (and any dependent libraries) as a ZIP and upload it using the AWS Lambda console from your local environment or specify an Amazon S3 location where the ZIP file is located. Uploads must be no larger than 50MB (compressed). You can use the AWS Eclipse plugin to author and deploy Lambda functions in Java and Node.js. If you are using Node.js, you can author the code for your function using the inline editor in the AWS Lambda console. Go to the console to get started.

Q: How do I create an AWS Lambda function using the Lambda CLI?

You can package the code (and any dependent libraries) as a ZIP and upload it using the AWS CLI from your local environment, or specify an Amazon S3 location where the ZIP file is located. Uploads must be no larger than 50MB (compressed). Visit the Lambda Getting Started guide to get started.

Q: How can I manage my AWS Lambda functions?

You can easily list, delete, update, and monitor your Lambda functions using the dashboard in the AWS Lambda console. You can also use the AWS CLI and AWS SDK to manage your Lambda functions. Visit the Lambda Developers Guide to learn more.

Q: How do I monitor an AWS Lambda function?

AWS Lambda automatically monitors Lambda functions on your behalf, reporting real-time metrics through Amazon CloudWatch, including total requests, latency, error rates, and throttled requests. You can view statistics for each of your Lambda functions via the Amazon CloudWatch console or through the AWS Lambda console. You can also call third-party monitoring APIs in your Lambda function. Visit Troubleshooting CloudWatch metrics to learn more. Standard charges for AWS Lambda apply to use Lambda's built-in metrics.

Q: How do I troubleshoot failures in an AWS Lambda function?

AWS Lambda automatically integrates with Amazon CloudWatch logs, creating a log group for each Lambda function and providing basic application lifecycle event log entries, including logging the resources consumed for each use of that function. You can easily insert additional logging statements into your code. You can also call third-party logging APIs in your Lambda function. Visit Troubleshooting Lambda functions to learn more. Amazon CloudWatch Logs rates will apply.

Q: How do I scale an AWS Lambda function?

You do not have to scale your Lambda functions – AWS Lambda scales them automatically on your behalf. Every time an event notification is received for your function, AWS Lambda quickly locates free capacity within its compute fleet and runs your code. Since your code is stateless, AWS Lambda can start as many copies of your function as needed without lengthy deployment and configuration delays. There are no fundamental limits to scaling a function. AWS Lambda will dynamically allocate capacity to match the rate of incoming events.

Q: How are compute resources assigned to an AWS Lambda function?

In the AWS Lambda resource model, you choose the amount of memory you want for your function, and are allocated proportional CPU power and other resources. For example, choosing 256MB of memory allocates approximately twice as much CPU power to your Lambda function as requesting 128MB of memory and half as much CPU power as choosing 512MB of memory. You can set your memory in 64MB increments from 128MB to 1.5GB.

Q: How long can an AWS Lambda function execute?

All calls made to AWS Lambda must complete execution within 300 seconds. The default timeout is 3 seconds, but you can set the timeout to any value between 1 and 300 seconds.

Q: How will I be charged for using AWS Lambda functions?

AWS Lambda is priced on a pay per use basis. Please see theAWS Lambda pricing page for details.

Q: Does AWS Lambda support versioning?

Yes. By default, each AWS Lambda function has a single, current version of the code. Clients of your Lambda function can call a specific version or get the latest implementation. Please read out documentation on versioning Lambda functions.

Q: How long after uploading my code will my AWS Lambda function be ready to call?

Deployment times may vary with the size of your code, but AWS Lambda functions are typically ready to call within seconds of upload.

Q: Can I use my own version of a supported library?

Yes. you can include your own copy of a library (including the AWS SDK) in order to use a different version than the default one provided by AWS Lambda.

Using AWS Lambda to Process AWS Events

Q: What is an event source?

An event source is an AWS service or developer-created application that produces events that trigger an AWS Lambda function to run. Some services publish these events to Lambda by invoking the cloud function directly (for example, Amazon S3). Lambda can also poll resources in other services that do not publish events to Lambda. For example, Lambda can pull records from a Kinesis stream and execute a Lambda function for each message in the stream.

Many other services, such as AWS CloudTrail, can act as event sources simply by logging to Amazon S3 and using S3 bucket notifications to trigger AWS Lambda functions.

Q: What event sources can be used with AWS Lambda?

Please see our documentation for a complete list of event sources.

Q: How are events represented in AWS Lambda?

Events are passed to a Lambda function as an event input parameter. For event sources where events arrive in batches, such as Amazon Kinesis and Amazon DynamoDB Streams, the event parameter may contain multiple events in a single call, based on the batch size you request.To learn more about Amazon S3 event notifications visit Configuring Notifications for Amazon S3 Events. To learn more about Amazon DynamoDB Streams visit the DynamoDB Stream Developers Guide. To learn more about invoking Lambda functions using Amazon SNS, visit the Amazon SNS Developers Guide. For more information on Amazon Cognito events, visit Amazon Cognito. For more information on AWS CloudTrail logs and auditing API calls across AWS services, see AWS CloudTrail.

Q: How do I make an AWS Lambda function respond to changes in an Amazon S3 bucket?

From the AWS Lambda console, you can select a function and associate it with notifications from an Amazon S3 bucket. Alternatively, you can use the Amazon S3 console and configure the bucket’s notifications to send to your AWS Lambda function. This same functionality is also available through the AWS SDK and CLI.

Q: How do I make an AWS Lambda function respond to updates in an Amazon DynamoDB table?

You can trigger a Lambda function on DynamoDB table updates by subscribing your Lambda function to the DynamoDB Stream associated with the table. You can associate a DynamoDB Stream with a Lambda function using the Amazon DynamoDB console, the AWS Lambda console or Lambda's registerEventSource API.

Q: How do I use an AWS Lambda function to process records in an Amazon Kinesis stream?

From the AWS Lambda console, you can select a Lambda function and associate it with an Amazon Kinesis stream owned by the same account. This same functionality is also available through the AWS SDK and CLI.

Q: How does AWS Lambda process data from Amazon Kinesis streams and Amazon DynamoDB Streams?

The Amazon Kinesis and DynamoDB Streams records sent to your AWS Lambda function are strictly serialized, per shard. This means that if you put two records in the same shard, Lambda guarantees that your Lambda function will be successfully invoked with the first record before it is invoked with the second record. If the invocation for one record times out, is throttled, or encounters any other error, Lambda will retry until it succeeds (or the record reaches its 24-hour expiration) before moving on to the next record. The ordering of records across different shards is not guaranteed, and processing of each shard happens in parallel.

Q: How do I use an AWS Lambda function to respond to notifications sent by Amazon Simple Notification Service (SNS)?

From the AWS Lambda console, you can select a Lambda function and associate it with an Amazon SNS topic. This same functionality is also available through the AWS SDK and CLI.

Q: How do I use an AWS Lambda function to respond to emails sent by Amazon Simple Email Service (SES)?

From the Amazon SES Console, you can set up your receipt rule to have Amazon SES deliver your messages to an AWS Lambda function. The same functionality is available through the AWS SDK and CLI.

Q: How do I use an AWS Lambda function to respond to Amazon CloudWatch alarms?

First, configure the alarm to send Amazon SNS notifications. Then from the AWS Lambda console, select a Lambda function and associate it with that Amazon SNS topic. See the Amazon CloudWatch Developer Guide for more on setting up Amazon CloudWatch alarms.

Q: How do I use an AWS Lambda function to respond to changes in user or device data managed by Amazon Cognito?

From the AWS Lambda console, you can select a function to trigger when any datasets associated with an Amazon Cognito identity pool are synchronized. This same functionality is also available through the AWS SDK and CLI. Visit Amazon Cognito for more information on using Amazon Cognito to share and synchronize data across a user's devices.

Q: How can my application trigger an AWS Lambda function directly?

You can invoke a Lambda function using a custom event through AWS Lambda's invoke API. Only the function's owner or another AWS account that the owner has granted permission can invoke the function. Visit the Lambda Developers Guide to learn more.

Q: What is the latency of invoking an AWS Lambda function in response to an event?

AWS Lambda is designed to process events within milliseconds. Latency will be higher immediately after a Lambda function is created, updated, or if it has not been used recently.

Q: What happens if my function fails while processing an event?

For Amazon S3 bucket notifications and custom events, AWS Lambda will attempt execution of your function three times in the event of an error condition in your code or if you exceed a service or resource limit. For ordered event sources that AWS Lambda polls on your behalf, such as Amazon DynamoDB Streams and Amazon Kinesis streams, Lambda will continue attempting execution in the event of a developer code error until the data expires. You can monitor progress through the Amazon Kinesis and Amazon DynamoDB consoles and through the Amazon CloudWatch metrics that AWS Lambda generates for your function. You can also set Amazon CloudWatch alarms based on error or execution throttling rates.

Using AWS Lambda to Build Back-end Services (Mobile & IoT)

Q: How do I create a mobile back-end using AWS Lambda?

You upload the code you want AWS Lambda to execute and then invoke it from your mobile app using the AWS Lambda SDK included in the AWS Mobile SDK. You can make both direct (synchronous) calls to retrieve or check data in real time as well as asynchronous calls. You can also define a custom API using Amazon API Gateway and invoke your Lambda functions through any REST compatible client. To learn more about the AWS Mobile SDK, visit the AWS Mobile SDK page. To learn more about Amazon API Gateway, visit the Amazon API Gateway page.

Q: How do I invoke an AWS Lambda function over HTTPS?

You can invoke a Lambda function over HTTPS by defining a custom RESTful API using Amazon API Gateway. This gives you an endpoint for your function which can respond to REST calls like GET, PUT and POST. Read more about using AWS Lambda with Amazon API Gateway.

Q: How can my AWS Lambda function customize its behavior to the device and app making the request?

When called through the AWS Mobile SDK, AWS Lambda functions automatically gain insight into the device and application that made the call through the ‘context’ object.

Q: How can my AWS Lambda function personalize their behavior based on the identity of the end user of an application?

When your app uses the Amazon Cognito identity, end users can authenticate themselves using a variety of public login providers such as Amazon, Facebook, Google, and other OpenID Connect-compatible services. User identity is then automatically and secured presented to your Lambda function in the form of an Amazon Cognito id, allowing it to access user data from Amazon Cognito, or as a key to store and retrieve data in Amazon DynamoDB or other web services.

Q: How do I create an Alexa skill using AWS Lambda?

AWS Lambda is integrated with the Alexa Skills Kit, a collection of self-service APIs, tools, documentation and code samples that make it easy for you to create voice-driven capabilities (or “skills”) for Alexa. You simply upload the Lambda function code for the new Alexa skill you are creating, and AWS Lambda does the rest, executing the code in response to Alexa voice interactions and automatically managing the compute resources on your behalf. Read the Alexa Skills Kit documentation for more details.

Scalability and Availability

Q: How available are AWS Lambda functions?

AWS Lambda is designed to use replication and redundancy to provide high availability for both the service itself and for the Lambda functions it operates. There are no maintenance windows or scheduled downtimes for either.

Q: Do my AWS Lambda functions remain available when I change my code or its configuration?

Yes. When you update a Lambda function, there will be a brief window of time, typically less than a minute, when requests could be served by either the old or the new version of your function.

Q: Is there a limit to the number of AWS Lambda functions I can execute at once?

No. AWS Lambda is designed to run many instances of your functions in parallel. However, AWS Lambda has a default safety throttle of 100 concurrent executions per account per region. If you wish to submit a request to increase the throttle of 100 concurrent executions you can visit our Support Center, click "Open a new case", and file a service limit increase request.

Q: What happens if my account exceeds the default throttle limit on concurrent executions?

On exceeding the throttle limit, AWS Lambda functions being invoked synchronously will return a throttling error (429 error code). Lambda functions being invoked asynchronously can absorb reasonable bursts of traffic for approximately 15-30 minutes, after which incoming events will be rejected as throttled. In case the Lambda function is being invoked in response to Amazon S3 events, events rejected by AWS Lambda may be retained and retried by S3 for 24 hours. Events from Amazon Kinesis streams and Amazon DynamoDB streams are retried until the Lambda function succeeds or the data expires. Amazon Kinesis and Amazon DynamoDB Streams retain data for 24 hours.

Q: Is the default limit applied on a per function level?

No, the default limit only applies at an account level.

Q: What happens if my Lambda function fails during processing an event?

On failure, Lambda functions being invoked synchronously will respond with an exception. Lambda functions being invoked asynchronously are retried at least 3 times, after which the event may be rejected. Events from Amazon Kinesis streams and Amazon DynamoDB streams are retried until the Lambda function succeeds or the data expires. Kinesis and DynamoDB Streams retain data for 24 hours.

Security and Access Control

Q: How do I allow my AWS Lambda function access to other AWS resources?
You grant permissions to your Lambda function to access other resources using an IAM role. AWS Lambda assumes the role while executing your Lambda function, so you always retain full, secure control of exactly which AWS resources it can use. Visit Setting up AWS Lambda to learn more about roles.

Q: How do I control which Amazon S3 buckets can call which AWS Lambda functions?
When you configure an Amazon S3 bucket to send messages to an AWS Lambda function a resource policy rule will a be created that grants access. Visit the Lambda Developer's Guide to learn more about resource policies and access controls for Lambda functions.

Q: How do I control which Amazon DynamoDB table or Amazon Kinesis stream an AWS

Lambda function can poll?
Access controls are managed through the Lambda function's role. The role you assign to your Lambda function also determines which resource(s) AWS Lambda can poll on its behalf. Visit the Lambda Developer's Guide to learn more.

Q: Can I access resources behind Amazon VPC with my AWS Lambda function?
Yes. You can access resources behind Amazon VPC.

Q: How do I enable and disable the VPC support for my Lambda function?
To enable VPC support, you need to specify one or more subnets in a single VPC and a security group as part of your function configuration. To disable VPC support, you need to update the function configuration and specify an empty list for the subnet and security group. You can change these settings using the AWS APIs, CLI, or AWS Lambda Management Console.

Q: Can a single Lambda function have access to multiple VPCs?
No. Lambda functions provide access only to a single VPC. If multiple subnets are specified, they must all be in the same VPC. You can connect to other VPCs by peering your VPCs.

Q: Can Lambda functions in a VPC also be able to access the internet and AWS Service endpoints?
Lambda functions configured to access resources in a particular VPC will not have access to the internet as a default configuration. If you need access to external endpoints, you will need to create a NAT in your VPC to forward this traffic and configure your security group to allow this outbound traffic.

AWS Lambda Functions in Java

Q: How do I compile my AWS Lambda function Java code?
You can use standard tools like Maven or Gradle to compile your Lambda function. Your build process should mimic the same build process you would use to compile any Java code that depends on the AWS SDK. Run your Java compiler tool on your source files and include the AWS SDK 1.9 or later with transitive dependencies on your classpath. For more details, see our documentation.

Q: What is the JVM environment Lambda uses for execution of my function?
Lambda provides the Amazon Linux build of openjdk 1.8.

AWS Lambda Functions in Node.js

Q: Can I use packages with AWS Lambda?

Yes. You can use NPM packages as well as custom packages. Learn morehere.

Q: Can I execute other programs from within my AWS Lambda function written in Node.js?

Yes. Lambda’s built-in sandbox lets you run batch (“shell”) scripts, other language runtimes, utility routines, and executables. Learn more here.

Q: Is it possible to use native modules with AWS Lambda functions written in Node.js?
Yes. Any statically linked native module can be included in the ZIP file you upload, as well as dynamically linked modules compiled with an rpath pointing to your Lambda function root directory. Learn more here.

Q: Can I execute binaries with AWS Lambda written in Node.js?
Yes. You can use Node.js' child_process command to execute a binary that you've included in your function or any executable from Amazon Linux that is visible to your function. Alternatively several NPM packages exist that wrap command line binaries such as node-ffmpeg. Learn more here.

Q: How do I deploy AWS Lambda function code written in Node.js?
To deploy a Lambda function written in Node.js, simply package your Javascript code and dependent libraries as a ZIP. You can upload the ZIP from your local environment, or specify an Amazon S3 location where the ZIP file is located. For more details, see our documentation.

AWS Lambda Functions in Python

Q: Can I use Python packages with AWS Lambda?

Yes. You can use pip to install any Python packages needed.

Other Topics

Q: Which versions of Amazon Linux, Node.js, Python, JDK, SDKs, and additional libraries does AWS Lambda support?
You can view the list of supported versions here.

Q: Can I change the version of Amazon Linux, JDK or Node.js?
No. AWS Lambda offers a single version of the operating system and language runtime to all

users of the service.

Q: How can I record and audit calls made to the AWS Lambda API?
AWS Lambda is integrated with AWS CloudTrail. AWS CloudTrail can record and deliver log files to your Amazon S3 bucket describing the API usage of your account.

Amazon VPC FAQ

General Questions

Q. What is Amazon Virtual Private Cloud (Amazon VPC)?

Amazon VPC lets you provision a logically isolated section of the Amazon Web Services (AWS) cloud where you can launch AWS resources in a virtual network that you define. 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 also create a hardware Virtual Private Network (VPN) connection between your corporate datacenter and your VPC and leverage the AWS cloud as an extension of your corporate datacenter.

You can easily customize the network configuration for your Amazon VPC. For example, you can create a public-facing subnet for your web servers that have 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.

Q. What are the components of Amazon VPC?

Amazon VPC comprises a variety of objects that will be familiar to customers with existing networks:

- **A Virtual Private Cloud (VPC)**: A logically isolated virtual network in the AWS cloud. You define a VPC’s IP address space from a range you select.

- **Subnet**: A segment of a VPC’s IP address range where you can place groups of isolated resources.

- **Internet Gateway**: The Amazon VPC side of a connection to the public Internet.

- **NAT Gateway**: A highly available, managed Network Address Translation (NAT) service for your resources in a private subnet to access the Internet.

- **Hardware VPN Connection**: A hardware-based VPN connection between your Amazon VPC and your datacenter, home network, or co-location facility.
- **Virtual Private Gateway**: The Amazon VPC side of a VPN connection.
- **Customer Gateway**: Your side of a VPN connection.
- **Router**: Routers interconnect subnets and direct traffic between Internet gateways, virtual private gateways, NAT gateways, and subnets.
- **Peering Connection**: A peering connection enables you to route traffic via private IP addresses between two peered VPCs.
- **VPC Endpoint for S3**: Enables Amazon S3 access from within your VPC without using an Internet gateway or NAT, and allows you to control the access using VPC endpoint policies.

Q. Why should I use Amazon VPC?

Amazon VPC enables you to build a virtual network in the AWS cloud - no VPNs, hardware, or physical datacenters required. You can define your own network space and control how your network, and the Amazon EC2 resources inside your network, is exposed to the Internet. You can also leverage the greatly enhanced security options in Amazon VPC to provide more granular access both to and from the Amazon EC2 instances in your virtual network.

Q. How do I get started with Amazon VPC?

Your AWS resources are automatically provisioned in a ready-to-use default VPC. You can choose to create additional VPCs by going to the Amazon VPC page in the AWS Management Console and selecting "Start VPC Wizard".

You’ll be presented with four basic options for network architectures. After selecting an option, you can modify the size and IP address range of the VPC and its subnets. If you select an option with Hardware VPN Access, you will need to specify the IP address of the VPN hardware on your network. You can modify the VPC to add more subnets or add or remove gateways at any time after the VPC has been created.

The four options are:

1. VPC with a Single Public Subnet Only
2. VPC with Public and Private Subnets
3. VPC with Public and Private Subnets and Hardware VPN Access
4. VPC with a Private Subnet Only and Hardware VPN Access

Billing

Q. How will I be charged and billed for my use of Amazon VPC?

There are no additional charges for creating and using the VPC itself. Usage charges for other Amazon Web Services, including Amazon EC2, still apply at published rates for those resources, including data transfer charges. If you connect your VPC to your corporate datacenter using the optional hardware VPN connection, pricing is per VPN connection-hour (the amount of time you have a VPN connection in the "available" state.) Partial hours are billed as full hours. Data transferred over VPN connections will be charged at standard AWS Data Transfer rates. For VPC-VPN pricing information, please visit the pricing section of the Amazon VPC product page.

Q. What defines billable VPN connection-hours?

VPN connection-hours are billed for any time your VPN connections are in the "available" state. You can determine the state of a VPN connection via the AWS Management Console, CLI, or API. If you no longer wish to use your VPN connection, you simply terminate the VPN connection to avoid being billed for additional VPN connection-hours.

Q. What usage charges will I incur if I use other AWS services, such as Amazon S3, from Amazon EC2 instances in my VPC?

Usage charges for other Amazon Web Services, including Amazon EC2, still apply at published rates for those resources. Data transfer charges are not incurred when accessing Amazon Web Services, such as Amazon S3, via your VPC’s Internet gateway.

If you access AWS resources via your VPN connection, you will incur Internet data transfer charges.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) region is subject to Japanese Consumption Tax. Learn more.

Connectivity

Q. What are the connectivity options for my VPC?

You may connect your VPC to:

- The Internet (via an Internet gateway)

- Your corporate data center using a Hardware VPN connection (via the virtual private gateway)
- Both the Internet and your corporate data center (utilizing both an Internet gateway and a virtual private gateway)
- Other AWS services (via Internet gateway, NAT, virtual private gateway, or VPC endpoints)
- Other VPCs (via VPC peering connections)

Q. How do I connect my VPC to the Internet?

Amazon VPC supports the creation of an Internet gateway. This gateway enables Amazon EC2 instances in the VPC to directly access the Internet.

Q. Are there any bandwidth limitations for Internet gateways? Do I need to be concerned about its availability? Can it be a single point of failure?

No. An Internet gateway is horizontally-scaled, redundant, and highly available. It imposes no bandwidth constraints.

Q. How do instances in a VPC access the Internet?

You can use public IP addresses, including Elastic IP addresses (EIPs), to give instances in the VPC the ability to both directly communicate outbound to the Internet and to receive unsolicited inbound traffic from the Internet (e.g., web servers). You can also use the solutions in the next question.

Q. How do instances without public IP addresses access the Internet?

Instances without public IP addresses can access the Internet in one of two ways:

1. Instances without public IP addresses can route their traffic through a NAT gateway or a NAT instance to access the Internet. These instances use the public IP address of the NAT gateway or NAT instance to traverse the Internet. The NAT gateway or NAT instance allows outbound communication but doesn’t allow machines on the Internet to initiate a connection to the privately addressed instances.

2. For VPCs with a hardware VPN connection or Direct Connect connection, instances can route their Internet traffic down the virtual private gateway to your existing datacenter. From there, it can access the Internet via your existing egress points and network security/monitoring devices.

Q. Can I connect to my VPC using a software VPN?

Yes. You may use a third-party software VPN to create a site to site or remote access VPN connection with your VPC via the Internet gateway.

Q. How does a hardware VPN connection work with Amazon VPC?

A hardware VPN connection connects your VPC to your datacenter. Amazon supports Internet Protocol security (IPsec) VPN connections. Data transferred between your VPC and datacenter routes over an encrypted VPN connection to help maintain the confidentiality and integrity of data in transit. An Internet gateway is not required to establish a hardware VPN connection.

Q. What is IPsec?

IPsec is a protocol suite for securing Internet Protocol (IP) communications by authenticating and encrypting each IP packet of a data stream.

Q. Which customer gateway devices can I use to connect to Amazon VPC?

There are two types of VPN connections that you can create: statically-routed VPN connections and dynamically-routed VPN connections. Customer gateway devices supporting statically-routed VPN connections must be able to:

- Establish IKE Security Association using Pre-Shared Keys
- Establish IPsec Security Associations in Tunnel mode
- Utilize the AES 128-bit or 256-bit encryption function
- Utilize the SHA-1 or SHA-2 (256) hashing function
- Utilize Diffie-Hellman (DH) Perfect Forward Secrecy in "Group 2" mode, or one of the additional DH groups we support
- Perform packet fragmentation prior to encryption

In addition to the above capabilities, devices supporting dynamically-routed VPN connections must be able to:

- Establish Border Gateway Protocol (BGP) peerings
- Bind tunnels to logical interfaces (route-based VPN)
- Utilize IPsec Dead Peer Detection

Q. Which Diffie-Hellman Groups do you support?

We support the following Diffie-Hellman (DH) groups in Phase1 and Phase2.

- Phase1 DH groups 2, 14-18, 22, 23, 24
- Phase2 DH groups 1, 2, 5, 14-18, 22, 23, 24

Q. What customer gateway devices are known to work with Amazon VPC?

The following devices meeting the aforementioned requirements are known to work with hardware VPN connections, and have support in the command line tools for automatic generation of configuration files appropriate for your device:

- Statically-routed VPN connections
  - Cisco ASA 5500 Series version 8.2 (or later) software
  - Cisco ISR running Cisco IOS 12.4 (or later) software
  - Dell SonicWALL Next Generation Firewalls (TZ, NSA, SuperMassive Series) running SonicOS5.8 (or later)
  - Juniper J-Series Service Router running JunOS 9.5 (or later) software
  - Juniper SRX-Series Services Gateway running JunOS 9.5 (or later) software
  - Juniper SSG running ScreenOS 6.1, or 6.2 (or later) software
  - Juniper ISG running ScreenOS 6.1, or 6.2 (or later) software
  - Microsoft Windows Server 2008 R2 (or later) software
  - Yamaha RTX1200 router
- Dynamically-routed VPN connections (requires BGP)
  - Astaro Security Gateway running version 8.3 (or later)
  - Astaro Security Gateway Essential Firewall Edition running version 8.3 (or later)
  - Cisco ISR running Cisco IOS 12.4 (or later) software
  - Dell SonicWALL Next Generation Firewalls (TZ, NSA, SuperMassive Series) running SonicOS5.9 (or later)
  - Fortinet Fortigate 40+ Series running FortiOS 4.0 (or later) software
  - Juniper J-Series Service Router running JunOS 9.5 (or later) software
  - Juniper SRX-Series Services Gateway running JunOS 9.5 (or later) software
  - Juniper SSG running ScreenOS 6.1, or 6.2 (or later) software
  - Juniper ISG running ScreenOS 6.1, or 6.2 (or later) software

- Palo Alto Networks PA Series running PANOS 4.1.2 (or later) software
- Vyatta Network OS 6.5 (or later) software
- Yamaha RTX1200 router

Please note, these sample configurations are for the minimum requirement of AES128, SHA1, and DH Group 2. You will need to modify these sample configuration files to take advantage of AES256, SHA256, or other DH groups.

Q. If my device is not listed, where can I go for more information about using it with Amazon VPC?

We recommend checking the Amazon VPC forum as other customers may be already using your device.

Q. Are there any VPN connection throughput limitations?

Amazon does not enforce any restrictions on VPN throughput. However, other factors, such as the cryptographic capability of your customer gateway, the capacity of your Internet connection, average packet size, the protocol being used (TCP vs. UDP), and the network latency between your customer gateway and the virtual private gateway can affect throughput.

Q. What tools are available to me to help troubleshoot my Hardware VPN configuration?

The DescribeVPNConnection API displays the status of the VPN connection, including the state ("up"/"down") of each VPN tunnel and corresponding error messages if either tunnel is "down". This information is also displayed in the AWS Management Console.

Q. How do I connect a VPC to my corporate datacenter?

Establishing a hardware VPN connection between your existing network and Amazon VPC allows you to interact with Amazon EC2 instances within a VPC as if they were within your existing network. AWS does not perform network address translation (NAT) on Amazon EC2 instances within a VPC accessed via a hardware VPN connection.

Q. Can I NAT my CGW behind a router or firewall?

Yes, you will need to enable NAT-T and open UDP port 4500 on your NAT device.

Q. What IP address do I use for my CGW address?

You will use the public IP address of your NAT device.

Q. How does my connection decide to use NAT-T?

If your device has NAT-T enabled on the tunnel, AWS will use it by default. You will need to open UDP port 4500 or else the tunnel will not establish.

Q. How do I disable NAT-T on my connection?

You will need to disable NAT-T on your device. If you don’t plan on using NAT-T and it is not disabled on your device, we will attempt to establish a tunnel over UDP port 4500. If that port is not open the tunnel will not establish.

Q. I would like to have multiple CGWs behind a NAT, what do I need to do to configure that?

You will use the public IP address of your NAT device for the CGW for each of your connections. You will also need to make sure UDP port 4500 is open.

Q. How many IPsec security associations can be established concurrently per tunnel?

The AWS VPN service is a route-based solution, so when using a route-based configuration you will not run into SA limitations. If, however, you are using a policy-based solution you will need to limit to a single SA, as the service is a route-based solution.

IP Addressing

Q. What IP address ranges can I use within my VPC?

You can address your VPC from any IPv4 address range, including RFC 1918 or publicly routable IP blocks. Publicly routable IP blocks are only reachable via the Virtual Private Gateway and cannot be accessed over the Internet through the Internet gateway. AWS does not advertise customer-owned IP address blocks to the Internet. Additionally, VPCs currently cannot be addressed from IPv6 IP address ranges.

Q. How do I assign IP address ranges to VPCs?

You assign a single Classless Internet Domain Routing (CIDR) IP address block when you create a VPC. Subnets within a VPC are addressed from this range by you. A VPC can be assigned at most one (1) IP address range at any given time; addressing a VPC from multiple IP address ranges is currently not supported. Please note that while you can create multiple VPCs with overlapping IP address ranges, doing so will prohibit you from connecting these VPCs to a common home network via the hardware VPN connection. For this reason we recommend using non-overlapping IP address ranges.

Q. What IP address ranges are assigned to a default VPC?

Default VPCs are assigned a CIDR range of 172.31.0.0/16. Default subnets within a default VPC are assigned /20 netblocks within the VPC CIDR range.

Q. Can I advertise my VPC public IP address range to the Internet and route the traffic through my datacenter, via the hardware VPN, and to my VPC?

Yes, you can route traffic via the hardware VPN connection and advertise the address range from your home network.

Q. How large of a VPC can I create?

Currently, Amazon VPC supports VPCs between /28 (in CIDR notation) and /16 in size. The IP address range of your VPC should not overlap with the IP address ranges of your existing network.

Q. Can I change a VPC's size?

No. To change the size of a VPC you must terminate your existing VPC and create a new one.

Q. How many subnets can I create per VPC?

Currently you can create 200 subnets per VPC. If you would like to create more, pleasesubmit a case at the support center.

Q. Is there a limit on how large or small a subnet can be?

The minimum size of a subnet is a /28 (or 14 IP addresses.) Subnets cannot be larger than the VPC in which they are created.

Q. Can I use all the IP addresses that I assign to a subnet?

No. Amazon reserves the first four (4) IP addresses and the last one (1) IP address of every subnet for IP networking purposes.

Q. How do I assign private IP addresses to Amazon EC2 instances within a VPC?

When you launch an Amazon EC2 instance within a VPC, you may optionally specify the primary private IP address for the instance. If you do not specify the primary private IP address, AWS automatically addresses it from the IP address range you assign to that subnet. You can assign secondary private IP addresses when you launch an instance, when you create an Elastic Network Interface, or any time after the instance has been launched or the interface has been created.

Q. Can I change the private IP addresses of an Amazon EC2 instance while it is running and/or stopped within a VPC?

Primary private IP addresses are retained for the instance's or interface's lifetime. Secondary private IP addresses can be assigned, unassigned, or moved between interfaces or instances at any time.

Q. If an Amazon EC2 instance is stopped within a VPC, can I launch another instance with the same IP address in the same VPC?

No. An IP address assigned to a running instance can only be used again by another instance once that original running instance is in a “terminated” state.

Q. Can I assign IP addresses for multiple instances simultaneously?

No. You can specify the IP address of one instance at a time when launching the instance.

Q. Can I assign any IP address to an instance?

You can assign any IP address to your instance as long as it is:

- Part of the associated subnet's IP address range
- Not reserved by Amazon for IP networking purposes
- Not currently assigned to another interface

Q. Can I assign multiple IP addresses to an instance?

Yes. You can assign one or more secondary private IP addresses to an Elastic Network Interface or an EC2 instance in Amazon VPC. The number of secondary private IP addresses you can assign depends on the instance type. See the EC2 User Guide for more information on the number of secondary private IP addresses that can be assigned per instance type.

Q. Can I assign one or more Elastic IP (EIP) addresses to VPC-based Amazon EC2 instances?

Yes, however, the EIP addresses will only be reachable from the Internet (not over the VPN connection). Each EIP address must be associated with a unique private IP address on the instance. EIP addresses should only be used on instances in subnets configured to route their traffic directly to the Internet gateway. EIPs cannot be used on instances in subnets configured to use a NAT gateway or a NAT instance to access the Internet.

Routing & Topology

Q. What does an Amazon VPC router do?

An Amazon VPC router enables Amazon EC2 instances within subnets to communicate with Amazon EC2 instances in other subnets within the same VPC. The VPC router also enables subnets, Internet gateways, and virtual private gateways to communicate with each other. Network usage data is not available from the router; however, you can obtain network usage statistics from your instances using Amazon CloudWatch.

Q. Can I modify the VPC route tables?

Yes. You can create route rules to specify which subnets are routed to the Internet gateway, the virtual private gateway, or other instances.

Q. Can I specify which subnet will use which gateway as its default?

Yes. You may create a default route for each subnet. The default route can direct traffic to egress the VPC via the Internet gateway, the virtual private gateway, or the NAT gateway.

Q. Does Amazon VPC support multicast or broadcast?

No.

Security & Filtering

Q. How do I secure Amazon EC2 instances running within my VPC?

Amazon EC2 security groups can be used to help secure instances within an Amazon VPC. Security groups in a VPC enable you to specify both inbound and outbound network traffic that is allowed to or from each Amazon EC2 instance. Traffic which is not explicitly allowed to or from an instance is automatically denied.

In addition to security groups, network traffic entering and exiting each subnet can be allowed or denied via network Access Control Lists (ACLs).

Q. What are the differences between security groups in a VPC and network ACLs in a VPC?

Security groups in a VPC specify which traffic is allowed to or from an Amazon EC2 instance. Network ACLs operate at the subnet level and evaluate traffic entering and exiting a subnet. Network ACLs can be used to set both Allow and Deny rules. Network ACLs do not filter traffic between instances in the same subnet. In addition, network ACLs perform stateless filtering while security groups perform stateful filtering.

Q. What is the difference between stateful and stateless filtering?

Stateful filtering tracks the origin of a request and can automatically allow the reply to the request to be returned to the originating computer. For example, a stateful filter that allows inbound traffic to TCP port 80 on a webserver will allow the return traffic, usually on a high numbered port (e.g., destination TCP port 63, 912) to pass through the stateful filter between the client and the webserver. The filtering device maintains a state table that tracks the origin and destination port numbers and IP addresses. Only one rule is required on the filtering device: Allow traffic inbound to the web server on TCP port 80.

Stateless filtering, on the other hand, only examines the source or destination IP address and the destination port, ignoring whether the traffic is a new request or a reply to a request. In the above example, two rules would need to be implemented on the filtering device: one rule to allow traffic inbound to the web server on TCP port 80, and another rule to allow outbound traffic from the webserver (TCP port range 49, 152 through 65, 535).

Q. Within Amazon VPC, can I use SSH key pairs created for instances within Amazon EC2, and vice versa?

Yes.

Q. Can Amazon EC2 instances within a VPC communicate with Amazon EC2 instances not within a VPC?

Yes. If an Internet gateway has been configured, Amazon VPC traffic bound for Amazon EC2 instances not within a VPC traverses the Internet gateway and then enters the public AWS network to reach the EC2 instance. If an Internet gateway has not been configured, or if the instance is in a subnet configured to route through the virtual private gateway, the traffic traverses the VPN connection, egresses from your datacenter, and then re-enters the public AWS network.

Q. Can Amazon EC2 instances within a VPC in one region communicate with Amazon EC2 instances within a VPC in another region?

Yes, they can communicate using public IP addresses, NAT gateway, NAT instances, VPN connections, or Direct Connect connections.

Q. Can Amazon EC2 instances within a VPC communicate with Amazon S3?

Yes. There are multiple options for your resources within a VPC to communicate with Amazon S3. You can use VPC Endpoint for S3, which makes sure all traffic remains within Amazon's network and enables you to apply additional access policies to your Amazon S3 traffic. You can use an Internet gateway to enable Internet access from your VPC and instances in the VPC can communicate with Amazon S3. You can also make all traffic to Amazon S3 traverse the Direct Connect or VPN connection, egress from your datacenter, and then re-enter the public AWS network.

Q. Why can’t I ping the router, or my default gateway, that connects my subnets?

Ping (ICMP Echo Request and Echo Reply) requests to the router in your VPC is not supported. Ping between Amazon EC2 instances within VPC is supported as long as your operating system's firewalls, VPC security groups, and network ACLs permit such traffic.

Q. Can I monitor the network traffic in my VPC?

Yes. You can use the Amazon VPC Flow Logs feature to monitor the network traffic in your VPC.

Amazon VPC & EC2

Q. Within which Amazon EC2 region(s) is Amazon VPC available?

Amazon VPC is currently available in multiple Availability Zones in all Amazon EC2 regions.

Q. Can a VPC span multiple Availability Zones?

Yes.

Q. Can a subnet span Availability Zones?

No. A subnet must reside within a single Availability Zone.

Q. How do I specify which Availability Zone my Amazon EC2 instances are launched in?

When you launch an Amazon EC2 instance you must specify the subnet in which to launch the instance. The instance will be launched in the Availability Zone associated with the specified subnet.

Q. How do I determine which Availability Zone my subnets are located in?

When you create a subnet you must specify the Availability Zone in which to place the subnet. When using the VPC Wizard, you can select the subnet's Availability Zone in the wizard confirmation screen. When using the API or the CLI you can specify the Availability Zone for the subnet as you create the subnet. If you don’t specify an Availability Zone, the default "No Preference" option will be selected and the subnet will be created in an available Availability Zone in the region.

Q. Am I charged for network bandwidth between instances in different subnets?

If the instances reside in subnets in different Availability Zones, you will be charged $0.01 per

GB for data transfer.

Q. When I call DescribeInstances(), do I see all of my Amazon EC2 instances, including those in EC2-Classic and EC2-VPC?

Yes. DescribeInstances() will return all running Amazon EC2 instances. You can differentiate EC2-Classic instances from EC2-VPC instances by an entry in the subnet field. If there is a subnet ID listed, the instance is within a VPC.

Q. When I call DescribeVolumes(), do I see all of my Amazon EBS volumes, including those in EC2-Classic and EC2-VPC?

Yes. DescribeVolumes() will return all your EBS volumes.

Q. How many Amazon EC2 instances can I use within a VPC?

You can run any number of Amazon EC2 instances within a VPC, so long as your VPC is appropriately sized to have an IP address assigned to each instance. You are initially limited to launching 20 Amazon EC2 instances at any one time and a maximum VPC size of /16 (65,536 IPs). If you would like to increase these limits, please complete the following form.

Q. Can I use my existing AMIs in Amazon VPC?

You can use AMIs in Amazon VPC that are registered within the same region as your VPC. For example, you can use AMIs registered in us-east-1 with a VPC in us-east-1. More information is available in the Amazon EC2 Region and Availability Zone FAQ.

Q. Can I use my existing Amazon EBS snapshots?

Yes, you may use Amazon EBS snapshots if they are located in the same region as your VPC. More details are available in the Amazon EC2 Region and Availability Zone FAQ.

Q: Can I boot an Amazon EC2 instance from an Amazon EBS volume within Amazon VPC?

Yes, however, an instance launched in a VPC using an Amazon EBS-backed AMI maintains the same IP address when stopped and restarted. This is in contrast to similar instances launched outside a VPC, which get a new IP address. The IP addresses for any stopped instances in a subnet are considered unavailable.

Q. Can I use Amazon EC2 Reserved Instances with Amazon VPC?

Yes. You can reserve an instance in Amazon VPC when you purchase Reserved Instances. When computing your bill, AWS does not distinguish whether your instance runs in Amazon

VPC or standard Amazon EC2. AWS automatically optimizes which instances are charged at the lower Reserved Instance rate to ensure you always pay the lowest amount. However, your instance reservation will be specific to Amazon VPC. Please see the Reserved Instances page for further details.

Q. Can I employ Amazon CloudWatch within Amazon VPC?

Yes.

Q. Can I employ Auto Scaling within Amazon VPC?

Yes.

Q. Can I launch Amazon EC2 Cluster Instances in a VPC?

Yes. Cluster instances are supported in Amazon VPC, however, not all instance types are available in all regions and Availability Zones.

Default VPCs

Q. What is a default VPC?

A default VPC is a logically isolated virtual network in the AWS cloud that is automatically created for your AWS account the first time you provision Amazon EC2 resources. When you launch an instance without specifying a subnet-ID, your instance will be launched in your default VPC.

Q. What are the benefits of a default VPC?

When you launch resources in a default VPC, you can benefit from the advanced networking functionalities of Amazon VPC (EC2-VPC) with the ease of use of Amazon EC2 (EC2-Classic). You can enjoy features such as changing security group membership on the fly, security group egress filtering, multiple IP addresses, and multiple network interfaces without having to explicitly create a VPC and launch instances in the VPC.

Q. What accounts are enabled for default VPC?

If your AWS account was created after March 18, 2013 your account may be able to launch resources in a default VPC. See this Forum Announcement to determine which regions have been enabled for the default VPC feature set. Also, accounts created prior to the listed dates may utilize default VPCs in any default VPC enabled region in which you’ve not previously launched EC2 instances or provisioned Amazon Elastic Load Balancing, Amazon RDS, Amazon ElastiCache, or Amazon Redshift resources.

Q. How can I tell if my account is configured to use a default VPC?

The Amazon EC2 console indicates which platforms you can launch instances in for the selected region, and whether you have a default VPC in that region. Verify that the region you'll use is selected in the navigation bar. On the Amazon EC2 console dashboard, look for "Supported Platforms" under "Account Attributes". If there are two values, EC2-Classic and EC2-VPC, you can launch instances into either platform. If there is one value, EC2-VPC, you can launch instances only into EC2-VPC. Your default VPC ID will be listed under "Account Attributes" if your account is configured to use a default VPC. You can also use the EC2 DescribeAccountAttributes API or CLI to describe your supported platforms.

Q. Will I need to know anything about Amazon VPC in order to use a default VPC?

No. You can use the AWS Management Console, AWS EC2 CLI, or the Amazon EC2 API to launch and manage EC2 instances and other AWS resources in a default VPC. AWS will automatically create a default VPC for you and will create a default subnet in each Availability Zone in the AWS region. Your default VPC will be connected to an Internet gateway and your instances will automatically receive public IP addresses, just like EC2-Classic.

Q. What are the differences between instances launched in EC2-Classic and EC2-VPC?

See Differences between EC2-Classic and EC2-VPC in the EC2 User Guide.

Q. Do I need to have a VPN connection to use a default VPC?

No. Default VPCs are attached to the Internet and all instances launched in default subnets in the default VPC automatically receive public IP addresses. You can add a VPN connection to your default VPC if you choose.

Q. Can I create other VPCs and use them in addition to my default VPC?

Yes. To launch an instance into nondefault VPCs you must specify a subnet-ID during instance launch.

Q. Can I create additional subnets in my default VPC, such as private subnets?

Yes. To launch into nondefault subnets, you can target your launches using the console or the --subnet option from the CLI, API, or SDK.

Q. How many default VPCs can I have?

You can have one default VPC in each AWS region where your Supported Platforms attribute is set to "EC2-VPC".

Q. What is the IP range of a default VPC?

The default VPC CIDR is 172.31.0.0/16. Default subnets use /20 CIDRs within the default VPC CIDR.

Q. How many default subnets are in a default VPC?

One default subnet is created for each Availability Zone in your default VPC.

Q. Can I specify which VPC is my default VPC?

Not at this time.

Q. Can I specify which subnets are my default subnets?

Not at this time.

Q. Can I delete a default VPC?

Yes. Contact AWS Support if you've deleted your default VPC and want to have it reset.

Q. Can I delete a default subnet?

Yes, but once deleted, it's gone. Your future instance launches will be placed in your remaining default subnet(s).

Q. I have an existing EC2-Classic account. Can I get a default VPC?

The simplest way to get a default VPC is to create a new account in a region that is enabled for default VPCs, or use an existing account in a region you've never been to before, as long as the Supported Platforms attribute for that account in that region is set to "EC2-VPC".

Q. I really want a default VPC for my existing EC2 account. Is that possible?

Yes, however, we can only enable an existing account for a default VPC if you have no EC2-Classic resources for that account in that region. Additionally, you must terminate all non-VPC provisioned Elastic Load Balancers, Amazon RDS, Amazon ElastiCache, and Amazon Redshift resources in that region. After your account has been configured for a default VPC, all future resource launches, including instances launched via Auto Scaling, will be placed in your default VPC. To request your existing account be setup with a default VPC, contact AWS Support. We will review your request and your existing AWS services and EC2-Classic presence to determine if you are eligible for a default VPC.

Q. How are IAM accounts impacted by default VPC?

If your AWS account has a default VPC, any IAM accounts associated with your AWS account use the same default VPC as your AWS account.

Elastic Network Interfaces

Q. Can I attach or detach one or more network interfaces to an EC2 instance while it's running?

Yes.

Q. Can I have more than two network interfaces attached to my EC2 instance?

The total number of network interfaces that can be attached to an EC2 instance depends on the instance type. See the EC2 User Guide for more information on the number of allowed network interfaces per instance type.

Q. Can I attach a network interface in one Availability Zone to an instance in another Availability Zone?

Network interfaces can only be attached to instances residing in the same Availability Zone.

Q. Can I attach a network interface in one VPC to an instance in another VPC?

Network interfaces can only be attached to instances in the same VPC as the interface.

Q. Can I use Elastic Network Interfaces as a way to host multiple websites requiring separate IP addresses on a single instance?

Yes, however, this is not a use case best suited for multiple interfaces. Instead, assign additional private IP addresses to the instance and then associate EIPs to the private IPs as needed.

Q. Will I get charged for an Elastic IP Address that is associated to a network interface but the network interface isn't attached to a running instance?

Yes.

Q. Can I detach the primary interface (eth0) on my EC2 instance?

No. You can attach and detach secondary interfaces (eth1-ethn) on an EC2 instance, but you can't detach the eth0 interface.

Peering Connections

Q. Can I create a peering connection to a VPC in a different region?

No. Peering connections are only available between VPCs in the same region.

Q. Can I peer my VPC with a VPC belonging to another AWS account?

Yes, assuming the owner of the other VPC accepts your peering connection request.

Q. Can I peer two VPCs with matching IP address ranges?

No. Peered VPCs must have non-overlapping IP ranges.

Q. How much do VPC peering connections cost?

There is no charge for creating VPC peering connections, however, data transfer across peering connections is charged. See the Data Transfer section of the EC2 Pricing page for data transfer rates.

Q. Can I use AWS Direct Connect or hardware VPN connections to access VPCs I'm peered with?

No. "Edge to Edge routing" isn't supported in Amazon VPC. Refer to theVPC Peering Guide for additional information.

Q. Do I need an Internet Gateway to use peering connections?

No. VPC peering connections do not require an Internet Gateway.

Q. Is VPC peering traffic within the region encrypted?

No. Traffic between instances in peered VPCs remains private and isolated – similar to how traffic between two instances in the same VPC is private and isolated.

Q. If I delete my side of a peering connection, will the other side still have access to my VPC?

No. Either side of the peering connection can terminate the peering connection at any time. Terminating a peering connection means traffic won't flow between the two VPCs.

Q. If I peer VPC A to VPC B and I peer VPC B to VPC C, does that mean VPCs A and C are peered?

No. Transitive peering relationships are not supported.

Q. What if my peering connection goes down?

AWS uses the existing infrastructure of a VPC to create a VPC peering connection; it is neither a gateway nor a VPN connection, and does not rely on a separate piece of physical hardware. There is no single point of failure for communication or a bandwidth bottleneck.

Q. Are there any bandwidth limitations for peering connections?

Bandwidth between instances in peered VPCs is no different than bandwidth between instances in the same VPC. **Note:** A placement group can span peered VPCs; however, you will not get full-bisection bandwidth between instances in peered VPCs. Read more about Placement Groups.

ClassicLink

Q. What is ClassicLink?

Amazon Virtual Private Cloud (VPC) ClassicLink allows EC2 instances in the EC2-Classic platform to communicate with instances in a VPC using private IP addresses. To use ClassicLink, enable it for a VPC in your account, and associate a Security Group from that VPC with an instance in EC2-Classic. All the rules of your VPC Security Group will apply to communications between instances in EC2-Classic and instances in the VPC.

Q. What does ClassicLink cost?

There is no additional charge for using ClassicLink; however, existing cross Availability Zone data transfer charges will apply. For more information, consult the EC2 pricing page.

Q. How do I use ClassicLink?

In order to use ClassicLink, you first need to enable at least one VPC in your account for ClassicLink. Then you associate a Security Group from the VPC with the desired EC2-Classic instance. The EC2-Classic instance is now linked to the VPC and is a member of the selected Security Group in the VPC. Your EC2-Classic instance cannot be linked to more than one VPC at the same time.

Q. Does the EC2-Classic instance become a member of the VPC?

The EC2-Classic instance does not become a member of the VPC. It becomes a member of the VPC Security Group that was associated with the instance. All the rules and references to the VPC Security Group apply to communication between instances in EC2-Classic instance and resources within the VPC.

Q. Can I use EC2 public DNS hostnames from my EC2-Classic and EC2-VPC instances to address each other, in order to communicate using private IP?

No. The EC2 public DNS hostname will not resolve to the private IP address of the EC2-VPC instance when queried from an EC2-Classic instance, and vice-versa.

Q. Are there any VPCs for which I cannot enable ClassicLink?

Yes. ClassicLink cannot be enabled for a VPC that has a Classless Inter-Domain Routing (CIDR) that is within the 10.0.0.0/8 range, with the exception of 10.0.0.0/16 and 10.1.0.0/16. In addition, ClassicLink cannot be enabled for any VPC that has a route table entry pointing to the 10.0.0.0/8 CIDR space to a target other than "local".

Q. Can traffic from an EC2-Classic instance travel through the Amazon VPC and egress through the Internet gateway, virtual private gateway, or to peered VPCs?

Traffic from an EC2-Classic instance can only be routed to private IP addresses within the VPC. They will not be routed to any destinations outside the VPC, including Internet gateway, virtual private gateway, or peered VPC destinations.

Q. Does ClassicLink affect the access control between the EC2-Classic instance, and other instances that are in the EC2-Classic platform?

ClassicLink does not change the access control defined for an EC2-Classic instance through its existing Security Groups from the EC2-Classic platform.

Q. Will ClassicLink settings on my EC2-Classic instance persist through stop/start cycles?

The ClassicLink connection will not persist through stop/start cycles of the EC2-Classic instance. The EC2-Classic instance will need to be linked back to a VPC after it is stopped and started. However, the ClassicLink connection will persist through instance reboot cycles.

Q. Will my EC2-Classic instance be assigned a new, private IP address after I enable ClassicLink?

There is no new private IP address assigned to the EC2-Classic instance. When you enable ClassicLink on an EC2-Classic instance, the instance retains and uses its existing private IP address to communication with resources in a VPC.

Q: Does ClassicLink allow EC2-Classic Security Group rules to reference VPC Security Groups, or vice versa?

ClassicLink does not allow EC2-Classic Security Group rules to reference VPC Security Groups, or vice versa.

Additional Questions

Q. Can I use the AWS Management Console to control and manage Amazon VPC?

Yes. You can use the AWS Management Console to manage Amazon VPC objects such as VPCs, subnets, route tables, Internet gateways, and IPSec VPN connections. Additionally, you can use a simple wizard to create a VPC.

Q. How many VPCs, subnets, Elastic IP addresses, Internet gateways, customer gateways, virtual private gateways, and VPN connections can I create?

You can have:

- Five Amazon VPCs per AWS account per region
- Two hundred subnets per Amazon VPC
- Five Amazon VPC Elastic IP addresses per AWS account per region
- One Internet gateway per VPC
- Five virtual private gateways per AWS account per region
- Fifty customer gateways per AWS account per region
- Ten IPsec VPN Connections per virtual private gateway

See the VPC User Guide for more information on VPC limits.

Q. Does the Amazon VPC VPN Connection have a Service Level Agreement (SLA)?

Not currently.

Q. Can I obtain AWS Support with Amazon VPC?

Yes. Click here for more information on AWS Support.

Q. Can I use ElasticFox with Amazon VPC?

ElasticFox is no longer officially supported for managing your Amazon VPC. Amazon VPC support is available via the AWS APIs, command line tools, and the AWS Management Console, as well as a variety of third-party utilities.

Amazon S3 FAQ

General

Q: What is Amazon S3?

Amazon S3 is storage for the Internet. It’s a simple storage service that offers software developers a highly-scalable, reliable, and low-latency data storage infrastructure at very low costs.

Q: What can I do with Amazon S3?

Amazon S3 provides a simple web service interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. Using this web service, developers can easily build applications that make use of Internet storage. Since Amazon S3 is highly scalable and you only pay for what you use, developers can start small and grow their application as they wish, with no compromise on performance or reliability.

Amazon S3 is also designed to be highly flexible. Store any type and amount of data that you want; read the same piece of data a million times or only for emergency disaster recovery; build a simple FTP application, or a sophisticated web application such as the Amazon.com retail web site. Amazon S3 frees developers to focus on innovation, not figuring out how to store their data.

Q: How can I get started using Amazon S3?

To sign up for Amazon S3, click the “Sign up for This Web Service” button on theAmazon S3 detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon S3 sign-up process. After signing up, please refer to the Amazon S3 documentation and sample code in the Resource Center to begin using Amazon S3.

Q: What are the technical benefits of Amazon S3?

Amazon S3 was carefully engineered to meet the requirements for scalability, reliability, speed, low-cost, and simplicity that must be met for Amazon’s internal developers. Amazon S3 passes these same benefits onto any external developer. More information about the Amazon S3 design requirements is available on the Amazon S3 detail page.

Q: What can developers do now that they could not before?

Until now, a sophisticated and scalable data storage infrastructure like Amazon’s has been beyond the reach of small developers. Amazon S3 enables any developer to leverage Amazon’s own benefits of massive scale with no up-front investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure their data is quickly accessible, always available, and secure.

Q: What kind of data can I store?

You can store virtually any kind of data in any format. Please refer to theAmazon Web Services Licensing Agreement for details.

Q: How much data can I store?

The total volume of data and number of objects you can store are unlimited. Individual Amazon S3 objects can range in size from a minimum of 0 bytes to a maximum of 5 terabytes. The largest object that can be uploaded in a single PUT is 5 gigabytes. For objects larger than 100 megabytes, customers should consider using the Multipart Upload capability.

Q: What storage classes does Amazon S3 offer?

Amazon S3 offers a range of storage classes designed for different use cases. There are three highly durable storage classes including Amazon S3 Standard for general-purpose storage of frequently accessed data, Amazon S3 Standard - Infrequent Access for long-lived, but less frequently accessed data, and Amazon Glacier for long-term archive. You can learn more about those three storage classes on the Amazon S3 Storage Classes page.

Reduced Redundancy Storage (RRS) is an Amazon S3 storage option that enables customers to reduce their costs by storing noncritical, reproducible data at lower levels of redundancy than Amazon S3's standard storage. You can learn more about Reduced Redundancy Storage on the Reduced Redundancy detail page.

Q: How can I delete large numbers of objects?

You can use Multi-Object Delete to delete large numbers of objects from Amazon S3. This feature allows you to send multiple object keys in a single request to speed up your deletes. Amazon does not charge you for using Multi-Object Delete.

Q: What does Amazon do with my data in Amazon S3?

Amazon will store your data and track its associated usage for billing purposes. Amazon will not otherwise access your data for any purpose outside of the Amazon S3 offering, except when required to do so by law. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: Does Amazon store its own data in Amazon S3?

Yes. Developers within Amazon use Amazon S3 for a wide variety of projects. Many of these projects use Amazon S3 as their authoritative data store, and rely on it for business-critical operations.

Q: How is Amazon S3 data organized?

Amazon S3 is a simple key-based object store. When you store data, you assign a unique object key that can later be used to retrieve the data. Keys can be any string, and can be constructed to mimic hierarchical attributes.

Q: How do I interface with Amazon S3?

Amazon S3 provides a simple, standards-based REST web services interface that is designed to work with any Internet-development toolkit. The operations are intentionally made simple to make it easy to add new distribution protocols and functional layers.

Q: How reliable is Amazon S3?

Amazon S3 gives any developer access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. S3 Standard is designed for 99.99% availability and Standard - IA is designed for 99.9% availability. Both are backed by the Amazon S3 Service Level Agreement.

Q: What data consistency model does Amazon S3 employ?

Amazon S3 buckets in all Regions provide read-after-write consistency for PUTS of new objects and eventual consistency for overwrite PUTS and DELETES.

Learn more

Q: What happens if traffic from my application suddenly spikes?

Amazon S3 was designed from the ground up to handle traffic for any Internet application. Pay-as-you-go pricing and unlimited capacity ensures that your incremental costs don’t change and that your service is not interrupted. Amazon S3’s massive scale enables us to spread load evenly, so that no individual application is affected by traffic spikes.

Q: What is the BitTorrent™ protocol, and how do I use it with Amazon S3?

BitTorrent is an open source Internet distribution protocol. Amazon S3’s bandwidth rates are inexpensive, but BitTorrent allows developers to further save on bandwidth costs for a popular piece of data by letting users download from Amazon and other users simultaneously. Any publicly available data in Amazon S3 can be downloaded via the BitTorrent protocol, in addition to the default client/server delivery mechanism. Simply add the ?torrent parameter at the end of your GET request in the REST API.

Q: Does Amazon S3 offer a Service Level Agreement (SLA)?

Yes. The Amazon S3 SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Q: How can I Increase the number of Amazon S3 buckets that I can provision?
By default, customers can provision up to 100 buckets per AWS account. However, you can increase your Amazon S3 bucket limit by visiting AWS Service Limits.

Regions

Q: Where is my data stored?

You specify a region when you create your Amazon S3 bucket. Within that region, your objects are redundantly stored on multiple devices across multiple facilities. Please refer to Regional Products and Services for details of Amazon S3 service availability by region.

Q: How do I decide which region to store my data in?

There are several factors to consider based on your specific application. You may want to store your data in a region that…

- ...is near to your customers, your data centers, or your other AWS resources in order to reduce data access latencies.
- ...is remote from your other operations for geographic redundancy and disaster recovery purposes.
- ...enables you to address specific legal and regulatory requirements.
- ...allows you to reduce storage costs. You can choose a lower priced region to save money. For S3 pricing information, please visit the S3 pricing page.

Q: I'm not in the US or Europe; can I use Amazon S3?

You can use Amazon S3 regardless of your location. You just have to decide which AWS region(s) you want to store your Amazon S3 data.

Q. Wasn't there a US Standard region?

We renamed the US Standard Region to US East (Northern Virginia) Region to be consistent with AWS regional naming conventions. There is no change to the endpoint and you do not need to make any changes to your application.

Billing

Q: How much does Amazon S3 cost?

With Amazon S3, you pay only for what you use. There is no minimum fee. You can estimate your monthly bill using the AWS Simple Monthly Calculator.

We charge less where our costs are less. Some prices vary across Amazon S3 Regions and are based on the location of your bucket. There is no Data Transfer charge for data transferred within an Amazon S3 Region via a COPY request. Data transferred via a COPY request between Regions is charged at rates specified on the pricing section of the Amazon S3 detail page. There is no Data Transfer charge for data transferred between Amazon EC2 and Amazon S3 within the same Region or for data transferred between the Amazon EC2 Northern Virginia Region and the Amazon S3 US East (Northern Virginia) Region. Data transferred between Amazon EC2 and Amazon S3 across all other Regions (i.e. between the Amazon EC2 Northern California and Amazon S3 US East (Northern Virginia) Regions is charged at rates specified on the pricing section of the Amazon S3 detail page.

For Amazon S3 pricing information, please visit thepricing page.

Q: Why do prices vary depending on which Amazon S3 region I choose?

We charge less where our costs are less. For example, our costs are lower in the US East (Northern Virginia) region than in the US West (Northern California) region.

Q: How will I be charged and billed for my use of Amazon S3?

There are no set-up fees or commitments to begin using the service. At the end of the month, your credit card will automatically be charged for that month's usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site, by logging into your Amazon Web Services account, and clicking "Account Activity" under "Your Web Services Account".

With the AWS Free Usage Tier*, you can get started with Amazon S3 for free in all regions except the AWS GovCloud Region. Upon sign-up, new AWS customers receive 5 GB of Amazon S3 standard storage, 20,000 Get Requests, 2,000 Put Requests, 15GB of data transfer in, and 15GB of data transfer out each month for one year.

Amazon S3 charges you for the following types of usage. Note that the calculations below assume there is no AWS Free Tier in place.

Storage Used:

Amazon S3 storage pricing is summarized on theAmazon S3 Pricing Chart.

The volume of storage billed in a month is based on the average storage used throughout the month. This includes all object data and metadata stored in buckets that you created under your AWS account. We measure your storage usage in “TimedStorage-ByteHrs,” which are added up at the end of the month to generate your monthly charges.

Storage Example:
Assume you store 100GB (107,374,182,400 bytes) of standard Amazon S3 storage data in your bucket for 15 days in March, and 100TB (109,951,162,777,600 bytes) of standard Amazon S3 storage data for the final 16 days in March.

At the end of March, you would have the following usage in Byte-Hours:
Total Byte-Hour usage
= [107,374,182,400 bytes x 15 days x (24 hours / day)] + [109,951,162,777,600 bytes x 16 days x (24 hours / day)] = 42,259,901,212,262,400 Byte-Hours.

Let’s convert this to GB-Months:
42,259,901,212,262,400 Byte-Hours / 1,073,741,824 bytes per GB / 744 hours per month = 52,900 GB-Months

This usage volume crosses three different volume tiers. The monthly storage price is calculated below assuming the data is stored in the US East (Northern Virginia) Region:
1 TB Tier: 1024 GB x $0.0300 = $30.72
1 TB to 50 TB Tier: 50,176 GB (49×1024) x $0.0295 = $1,480.19
50 TB to 450 TB Tier: 1,700 GB (remainder) x $0.0290 = $49.30

Total Storage Fee = $30.72 + $1,480.19 + $49.30 = $1,560.21

Network Data Transferred In:

Amazon S3 Data Transfer In pricing is summarized on the Amazon S3 Pricing Chart.

This represents the amount of data sent to your Amazon S3 buckets. Data Transfer is $0.000 per GB for buckets in the US East (Northern Virginia), US West (Oregon), US West (Northern California), EU (Ireland), EU (Frankfurt), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Sydney), South America (Sao Paulo), and AWS GovCloud (US) Regions.

Network Data Transferred Out:

Amazon S3 Data Transfer Out pricing is summarized on the Amazon S3 Pricing Chart. For Amazon S3, this charge applies whenever data is read from any of your buckets from a location outside of the given Amazon S3 Region.

Data Transfer Out pricing rate tiers take into account your aggregate Data Transfer Out from a given region to the Internet across Amazon EC2, Amazon S3, Amazon RDS, Amazon SimpleDB, Amazon SQS, Amazon SNS and Amazon VPC. These tiers do not apply to Data Transfer Out from Amazon S3 in one AWS region to another AWS region.

Data Transfer Out Example:
Assume you transfer 1TB of data out of Amazon S3 from the US East (Northern Virginia) Region to the Internet every day for a given 31-day month. Assume you also transfer 1TB of data out of an Amazon EC2 instance from the same region to the Internet over the same 31-day month.

Your aggregate Data Transfer would be 62 TB (31 TB from Amazon S3 and 31 TB from Amazon EC2). This equates to 63,488 GB (62 TB * 1024 GB/TB).

This usage volume crosses three different volume tiers. The monthly Data Transfer Out fee is calculated below assuming the Data Transfer occurs in the US East (Northern Virginia) Region:
10 TB Tier: 10,240 GB (10×1024 GB/TB) x $0.120 = $1,228.80
10 TB to 50 TB Tier: 40,960 GB (40×1024) x $0.090 = $3,686.40
50 TB to 150 TB Tier: 12,288 GB (remainder) x $0.070 = $860.16

Total Data Transfer Out Fee = $1,228.80+ $3,686.40 + $860.16= $5,775.36

Data Requests:

Amazon S3 Request pricing is summarized on the Amazon S3 Pricing Chart.

Request Example:
Assume you transfer 10,000 files into Amazon S3 and transfer 20,000 files out of Amazon S3 each day during the month of March. Then, you delete 5,000 files on March 31st.
Total PUT requests = 10,000 requests x 31 days = 310,000 requests
Total GET requests = 20,000 requests x 31 days = 620,000 requests
Total DELETE requests = 5,000×1 day = 5,000 requests

Assuming your bucket is in the US East (Northern Virginia) Region, the Request fees are calculated below:
310,000 PUT Requests: 310,000 requests x $0.005/1,000 = $1.55
620,000 GET Requests: 620,000 requests x $0.004/10,000 = $0.25
5,000 DELETE requests = 5,000 requests x $0.00 (no charge) = $0.00

Data Retrieval:

Amazon S3 data retrieval pricing applies for the Standard – Infrequent Access (Standard - IA) storage class and is summarized on the Amazon S3 Pricing Chart.

Data Retrieval Example:
Assume in one month you retrieve 300GB of Standard - IA, with 100GB going out to the Internet, 100GB going to EC2 in the same AWS region, and 100GB going to CloudFront in the same AWS region.

Your data retrieval fees for the month would be calculated as 300GB x $0.01/GB = $3.00. Note that you would also pay network data transfer fees for the portion that went out to the Internet.

Please see here for details on billing of objects archived to Amazon Glacier.

* * Your usage for the free tier is calculated each month across all regions except the AWS GovCloud Region and automatically applied to your bill – unused monthly usage will not roll over. Restrictions apply; See offer terms for more details.

Q: How am I charged for accessing Amazon S3 through the AWS Management Console?

Normal Amazon S3 pricing applies when accessing the service through the AWS Management Console. To provide an optimized experience, the AWS Management Console may proactively execute requests. Also, some interactive operations result in more than one request to the service.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Security

Q: How secure is my data?

Amazon S3 is secure by default. Only the bucket and object owners originally have access to Amazon S3 resources they create. Amazon S3 supports user authentication to control access to data. You can use access control mechanisms such as bucket policies and Access Control Lists (ACLs) to selectively grant permissions to users and groups of users. You can securely upload/download your data to Amazon S3 via SSL endpoints using the HTTPS protocol. If you need extra security you can use the Server Side Encryption (SSE) option or the Server Side Encryption with Customer-Provide Keys (SSE-C) option to encrypt data stored-at-rest. Amazon S3 provides the encryption technology for both SSE and SSE-C. Alternatively you can use your own encryption libraries to encrypt data before storing it in Amazon S3.

Q: How can I control access to my data stored on Amazon S3?

Customers may use four mechanisms for controlling access to Amazon S3 resources: Identity and Access Management (IAM) policies, bucket policies, Access Control Lists (ACLs) and query string authentication. IAM enables organizations with multiple employees to create and manage multiple users under a single AWS account. With IAM policies, companies can grant IAM users fine-grained control to their Amazon S3 bucket or objects while also retaining full control over everything the users do. With bucket policies, companies can define rules which apply broadly across all requests to their Amazon S3 resources, such as granting write privileges to a subset of Amazon S3 resources. Customers can also restrict access based on an aspect of the request, such as HTTP referrer and IP address. With ACLs, customers can grant specific permissions (i.e. READ, WRITE, FULL_CONTROL) to specific users for an individual bucket or object. With query string authentication, customers can create a URL to an Amazon S3 object which is only valid for a limited time. For more information on the various access control policies available in Amazon S3, please refer to the Access Control topic in the Amazon S3 Developer Guide.

Q: Does Amazon S3 support data access auditing?

Yes, customers can optionally configure Amazon S3 buckets to create access log records for all requests made against it. These access log records can be used for audit purposes and contain details about the request, such as the request type, the resources specified in the request, and the time and date the request was processed.

Q: What options do I have for encrypting data stored on Amazon S3?

You can choose to encrypt data using SSE-S3, SSE-C, SSE-KMS, or a client library such as the Amazon S3 Encryption Client. All four enable you to store sensitive data encrypted at rest in Amazon S3.

SSE-S3 provides an integrated solution where Amazon handles key management and key protection using multiple layers of security. You should choose SSE-S3 if you prefer to have Amazon manage your keys.

SSE-C enables you to leverage Amazon S3 to perform the encryption and decryption of your objects while retaining control of the keys used to encrypt objects. With SSE-C, you don’t need to implement or use a client-side library to perform the encryption and decryption of objects you store in Amazon S3, but you do need to manage the keys that you send to Amazon S3 to encrypt and decrypt objects. Use SSE-C if you want to maintain your own encryption keys, but don’t want to implement or leverage a client-side encryption library.

SSE-KMS enables you to use AWS Key Management Service (AWS KMS) to manage your encryption keys. Using AWS KMS to manage your keys provides several additional benefits. With AWS KMS, there are separate permissions for the use of the master key, providing an additional layer of control as well as protection against unauthorized access to your objects stored in Amazon S3. AWS KMS provides an audit trail so you can see who used your key to access which object and when, as well as view failed attempts to access data from users without permission to decrypt the data. Also, AWS KMS provides additional security controls to support customer efforts to comply with PCI-DSS, HIPAA/HITECH, and FedRAMP industry requirements.

Using an encryption client library, such as the Amazon S3 Encryption Client, you retain control of the keys and complete the encryption and decryption of objects client-side using an encryption library of your choice. Some customers prefer full end-to-end control of the encryption and decryption of objects; that way, only encrypted objects are transmitted over the Internet to Amazon S3. Use a client-side library if you want to maintain control of your encryption keys, are able to implement or use a client-side encryption library, and need to have your objects encrypted before they are sent to Amazon S3 for storage.

For more information on using Amazon S3 SSE-S3, SSE-C, or SSE-KMS, please refer to the topic on Using Encryption in the Amazon S3 Developer Guide.

Q: How does Amazon protect SSE encryption keys?

With SSE, every protected object is encrypted with a unique key. This object key is itself encrypted by a separate master key. A new master key is issued at least monthly. Encrypted data, encryption keys and master keys are stored and secured on separate hosts for multiple layers of protection.

Q: Can I comply with EU data privacy regulations using Amazon S3?

Customers can choose to store all data in the EU by using the EU (Ireland) or EU (Frankfurt) region. It is your responsibility to ensure that you comply with EU privacy laws.

Q: Where can I find more information about security on AWS?

For more information on security on AWS please refer to ourAmazon Web Services: Overview of Security Processes document.

Q: What is an Amazon VPC Endpoint for Amazon S3?

An Amazon VPC Endpoint for Amazon S3 is a logical entity within a VPC that allows connectivity only to S3. The VPC Endpoint routes requests to S3 and routes responses back to the VPC. For more information about VPC Endpoints, read Using VPC Endpoints.

Q: Can I allow a specific Amazon VPC Endpoint access to my Amazon S3 bucket?

You can limit access to your bucket from a specific Amazon VPC Endpoint or a set of endpoints using Amazon S3 bucket policies. S3 bucket policies now support a condition, aws:sourceVpce, that you can use to restrict access. For more details and example policies, read Using VPC Endpoints.

Data Protection

Q: How durable is Amazon S3?

Amazon S3 Standard and Standard - IA are designed to provide 99.999999999% durability of objects over a given year. This durability level corresponds to an average annual expected loss of 0.000000001% of objects. For example, if you store 10,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000,000 years. In addition, Amazon S3 is designed to sustain the concurrent loss of data in two facilities.

As with any environments, the best practice is to have a backup and to put in place safeguards against malicious or accidental users errors. For S3 data, that best practice includes secure access permissions, Cross-Region Replication, versioning and a functioning, regularly tested backup.

Q: How is Amazon S3 designed to achieve 99.999999999% durability?

Amazon S3 Standard and Standard - IA redundantly stores your objects on multiple devices across multiple facilities in an Amazon S3 Region. The service is designed to sustain concurrent device failures by quickly detecting and repairing any lost redundancy. When processing a request to store data, the service will redundantly store your object across multiple facilities before returning SUCCESS. Amazon S3 also regularly verifies the integrity of your data using checksums.

Q: What checksums does Amazon S3 employ to detect data corruption?

Amazon S3 uses a combination of Content-MD5 checksums and cyclic redundancy checks (CRCs) to detect data corruption. Amazon S3 performs these checksums on data at rest and repairs any corruption using redundant data. In addition, the service calculates checksums on all network traffic to detect corruption of data packets when storing or retrieving data.

Q: What is Versioning?

Versioning allows you to preserve, retrieve, and restore every version of every object stored in an Amazon S3 bucket. Once you enable Versioning for a bucket, Amazon S3 preserves existing objects anytime you perform a PUT, POST, COPY, or DELETE operation on them. By default, GET requests will retrieve the most recently written version. Older versions of an overwritten or deleted object can be retrieved by specifying a version in the request.

Q: Why should I use Versioning?

Amazon S3 provides customers with a highly durable storage infrastructure. Versioning offers an additional level of protection by providing a means of recovery when customers accidentally overwrite or delete objects. This allows you to easily recover from unintended user actions and application failures. You can also use Versioning for data retention and archiving.

Q: How do I start using Versioning?

You can start using Versioning by enabling a setting on your Amazon S3 bucket. For more information on how to enable Versioning, please refer to the Amazon S3 Technical Documentation.

Q: How does Versioning protect me from accidental deletion of my objects?

When a user performs a DELETE operation on an object, subsequent simple (un-versioned) requests will no longer retrieve the object. However, all versions of that object will continue to be preserved in your Amazon S3 bucket and can be retrieved or restored. Only the owner of an Amazon S3 bucket can permanently delete a version. You can set Lifecycle rules to manage the lifetime and the cost of storing multiple versions of your objects.

Q: Can I setup a trash, recycle bin, or rollback window on my Amazon S3 objects to recover from deletes and overwrites?

You can use Lifecycle rules along with Versioning to implement a rollback window for your Amazon S3 objects. For example, with your versioning-enabled bucket, you can set up a rule that archives all of your previous versions to the lower-cost Glacier storage class and deletes them after 100 days, giving you a 100 day window to roll back any changes on your data while lowering your storage costs.

Q: How can I ensure maximum protection of my preserved versions?

Versioning's MFA Delete capability, which uses multi-factor authentication, can be used to provide an additional layer of security. By default, all requests to your Amazon S3 bucket require your AWS account credentials. If you enable Versioning with MFA Delete on your Amazon S3 bucket, two forms of authentication are required to permanently delete a version of an object: your AWS account credentials and a valid six-digit code and serial number from an authentication device in your physical possession. To learn more about enabling Versioning with MFA Delete, including how to purchase and activate an authentication device, please refer to the Amazon S3 Technical Documentation.

Q: How am I charged for using Versioning?

Normal Amazon S3 rates apply for every version of an object stored or requested. For example, let's look at the following scenario to illustrate storage costs when utilizing Versioning (let's assume the current month is 31 days long):

1) Day 1 of the month: You perform a PUT of 4 GB (4,294,967,296 bytes) on your bucket.
2) Day 16 of the month: You perform a PUT of 5 GB (5,368,709,120 bytes) within the same bucket using the same key as the original PUT on Day 1.

When analyzing the storage costs of the above operations, please note that the 4 GB object from Day 1 is not deleted from the bucket when the 5 GB object is written on Day 15. Instead, the 4 GB object is preserved as an older version and the 5 GB object becomes the most recently written version of the object within your bucket. At the end of the month:

Total Byte-Hour usage
[4,294,967,296 bytes x 31 days x (24 hours / day)] + [5,368,709,120 bytes x 16 days x (24 hours / day)] = 5,257,039,970,304 Byte-Hours.

Conversion to Total GB-Months
5,257,039,970,304 Byte-Hours x (1 GB / 1,073,741,824 bytes) x (1 month / 744 hours) = 6.581 GB-Months

The storage fee is calculated below assuming data is stored in the US East (Northern Virginia) Region:
0 to 1 TB Tier: 6.581GB x $0.0300 = $0.20

S3 Standard - Infrequent Access

Q: What is S3 Standard - Infrequent Access?

Amazon S3 Standard - Infrequent Access (Standard - IA) is an Amazon S3 storage class for data that is accessed less frequently, but requires rapid access when needed. Standard - IA offers the high durability, throughput, and low latency of Amazon S3 Standard, with a low per GB storage price and per GB retrieval fee. This combination of low cost and high performance make Standard - IA ideal for long-term storage, backups, and as a data store for disaster recovery. The Standard - IA storage class is set at the object level and can exist in the same bucket as Standard, allowing you to use lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: Why would I choose to use Standard - IA?

Standard - IA is ideal for data that is accessed less frequently, but requires rapid access when needed. Standard - IA is ideally suited for long-term file storage, older data from sync and share, backup data, and disaster recovery files.

Q: What performance does S3 Standard - Infrequent Access offer?

S3 Standard - Infrequent Access provide the same performance as S3 Standard storage.

Q: How durable is Standard - IA?

S3 Standard - IA is designed for the same 99.999999999% durability as Standard and Amazon Glacier. Standard - IA is designed for 99.9% availability, and carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle.

Q: How available is Standard - IA?

Designed for 99.9% availability, Standard - IA has a thinner front end that provides one one-hundredth of a percent less availability than S3 Standard. This means that the probability of a request failing and having to be retried is very slightly greater than that of S3 Standard. Standard - IA carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle.

Q: How do I get my data into Standard - IA?

There are two ways to get data into Standard – IA from within S3. You can directly PUT into Standard – IA by specifying STANDARD_IA in the x-amz-storage-class header. You can also set lifecycle policies to transition objects from Standard to Standard - IA.

Q: Are my Standard - IA objects backed with the Amazon S3 Service Level Agreement?

Yes, Standard - IA is backed with theAmazon S3 Service Level Agreement, and customers are eligible for service credits if availability is less than our service commitment in any billing cycle.

Q: How will my latency and throughput performance be impacted as a result of using Standard - IA?

You should expect the same latency and throughput performance as Amazon S3 Standard when using Standard - IA.

Q: How am I charged for using Standard - IA?

Please see the Amazon S3 pricing page for general information about Standard - IA pricing.

Q. What charges will I incur if I change storage class of an object from Standard-IA to Standard with a copy request?

You will incur charges for an Standard-IA copy request and a Standard-IA data retrieval.

Q: Is there a minimum duration for Standard - IA?

Standard - IA is designed for long-lived, but infrequently accessed data that is retained for months or years. Data that is deleted from Standard - IA within 30 days will be charged for a full 30 days. Please see the Amazon S3 pricing page for information about Standard - IA pricing.

Q: Is there a minimum object size for Standard - IA?

Standard - IA is designed for larger objects and has a minimum object size of 128KB. Objects smaller than 128KB in size will incur storage charges as if the object were 128KB. For example, a 6KB object in S3 Standard - IA will incur S3 Standard - IA storage charges for 6KB and an additional minimum object size fee equivalent to 122KB at the S3 Standard - IA storage price. Please see the Amazon S3 pricing page for information about Standard - IA pricing.

Q: Can I tier objects from Standard - IA to Amazon Glacier?

Yes. In addition to using lifecycle policies to migrate objects from Standard to Standard - IA, you can also set up lifecycle policies to tier objects from Standard - IA to Amazon Glacier.

Amazon Glacier

Q: Does Amazon S3 provide capabilities for archiving objects to lower cost storage options?

Yes, Amazon S3 enables you to utilize Amazon Glacier’s extremely low-cost storage service as storage for data archival. Amazon Glacier stores data for as little as $0.007 per gigabyte per month, and is optimized for data that is infrequently accessed and for which retrieval times of several hours are suitable. Examples include digital media archives, financial and healthcare records, raw genomic sequence data, long-term database backups, and data that must be retained for regulatory compliance.

Q: How can I store my data using the Amazon Glacier option?

You can use Lifecycle rules to automatically archive sets of Amazon S3 objects to Amazon Glacier based on lifetime. Use the Amazon S3 Management Console, the AWS SDKs or the Amazon S3 APIs to define rules for archival. Rules specify a prefix and time period. The prefix (e.g. “logs/”) identifies the object(s) subject to the rule. The time period specifies either the number of days from object creation date (e.g. 180 days) or the specified date after which the object(s) should be archived. Any Amazon S3 Standard or Amazon S3 Standard - IA objects which have names beginning with the specified prefix and which have aged past the specified time period are archived to Amazon Glacier. To retrieve Amazon S3 data stored in Amazon Glacier, initiate a restore job via the Amazon S3 APIs or Management Console. Restore jobs typically complete in 3 to 5 hours. Once the job is complete, you can access your data through an Amazon S3 GET object request.

You can use Lifecycle rules for any of your buckets including versioned buckets. You can easily archive your object versions after an elapsed time period (number of days from overwrite/expire).

For more information on using Lifecycle rules for archival, please refer to the Object Archival topic in the Amazon S3 Developer Guide.

Q: Can I use the Amazon S3 APIs or Management Console to list objects that I’ve archived to Amazon Glacier?

Yes, like Amazon S3’s other storage options (Standard or Standard - IA), Amazon Glacier objects stored using Amazon S3’s APIs or Management Console have an associated user-defined name. You can get a real-time list of all of your Amazon S3 object names, including those stored using the Amazon Glacier option, using the Amazon S3 LIST API.

Q: Can I use Amazon Glacier APIs to access objects that I’ve archived to Amazon Glacier?

Because Amazon S3 maintains the mapping between your user-defined object name and Amazon Glacier’s system-defined identifier, Amazon S3 objects that are stored using the Amazon Glacier option are only accessible through the Amazon S3 APIs or the Amazon S3 Management Console.

Q: How can I restore my objects that are archived in Amazon Glacier?

To restore Amazon S3 data stored in Amazon Glacier, initiate a restore request using the Amazon S3 APIs or the Amazon S3 Management Console. Restore requests typically complete in 3 to 5 hours. The restore request creates a temporary copy of your data in RRS while leaving the archived data intact in Amazon Glacier. You can specify the amount of time in days for which the temporary copy is stored in RRS. You can then access your temporary copy from RRS through an Amazon S3 GET request on the archived object.

Q: How long will it take to restore my objects archived in Amazon Glacier?

When processing a restore job, Amazon S3 first retrieves the requested data from Amazon Glacier (which typically takes 3-5 hours), and then creates a temporary copy of the requested data in RRS (which typically takes on the order of a few minutes). You can expect most restore jobs initiated via the Amazon S3 APIs or Management Console to complete in 3-5 hours.

Q: What am I charged for archiving objects in Amazon Glacier?

Amazon Glacier storage is priced from $0.007 per gigabyte per month. Archive and Restore requests are priced from $0.05 per 1,000 requests. For large restores, there is also a restore fee starting at $0.01 per gigabyte. When an archived object is restored, it resides in both RRS and Glacier. You are charged for both RRS and Glacier storage usage for the duration the object remains restored, after which point you are only charged for Glacier storage of the object. There is a pro-rated charge of $0.03 per GB for items that are deleted prior to 90 days. As Amazon Glacier is designed to store data that is infrequently accessed and long lived, these restore and early delete charges will likely not apply to most of you. Standard Amazon S3 rates apply for bandwidth. To learn more, please visit the Amazon S3 detail page.

Q: How is my storage charge calculated for Amazon S3 objects archived to Amazon Glacier?

The volume of storage billed in a month is based on average storage used throughout the month, measured in gigabyte-months (GB-Months). Amazon S3 calculates the object size as the amount of data you stored plus an additional 32 kilobytes of Glacier data plus an additional 8 KB of S3 standard storage data. Amazon Glacier requires an additional 32 kilobytes of data per object for Glacier's index and metadata so you can identify and retrieve your data. Amazon S3 requires 8KB to store and maintain the user-defined name and metadata for objects archived to Amazon Glacier. This enables you to get a real-time list of all of your Amazon S3 objects, including those stored using the Amazon Glacier option, using the Amazon S3 LIST API. For example, if you have archived 100,000 objects that are 1GB each, your billable storage would be:

1.000032 gigabytes for each object x 100,000 objects = 100,003.2 gigabytes of Amazon Glacier storage.
0.000008 gigabytes for each object x 100,000 objects = 0.8 gigabytes of Amazon S3 Standard storage.

If you archive the objects for one month in the US East (Northern Virginia) region, you would be charged:
(100,003.20 GB-Months x $0.0070) + (0.8 GB-Months x $0.0300) = $700.046

Q: How much data can I restore for free?

You can restore up to 5% of the Amazon S3 data stored in Amazon Glacier for free each month. Typically this will be sufficient for backup and archival needs. Your 5% monthly free restore allowance is calculated and metered on a daily prorated basis. For example, if on a given day you have 12 terabytes of Amazon S3 data archived to Amazon Glacier, you can restore up to 20.5 gigabytes of data for free that day (12 terabytes x 5% / 30 days = 20.5 gigabytes, assuming it is a 30 day month).

Q: How will I be charged when restoring large amounts of data from Amazon Glacier?

You can restore up to 5% of your archived data, pro-rated daily, for free each month. For example, if on a given day you have 75 TB of S3 data archived in Amazon Glacier, you can restore up to 128 GB of data for free that day (75 terabytes x 5% / 30 days = 128 gigabytes, assuming it is a 30 day month). In this example, 128 GB is your daily free restore allowance. You are charged a Data Restore fee only if you exceed your daily restore allowance. Let's now look at how this Restore Fee - which is based on your monthly peak billable restore rate - is calculated.

Let’s assume you have 75TB of data archived in Amazon Glacier and you would like to restore 140GB. The data restore fee you pay is determined by how fast you want to restore the data. For example, you can request all the data at once and pay $21.60, or restore it evenly over eight hours, and pay $10.80. If you further spread your restores evenly over 28 hours, your restores would be free because you would be restoring less than 128 GB per day. The more you spread out your restore requests, the lower your peak usage and the lower your cost.

Below we review how to calculate Restore Fees if you archived 75TB data and restored 140 GB in 4 hours, 8 hours and 28 hours respectively.

Example 1: Archiving 75TB of data to Amazon Glacier and restoring 140GB in 4 hours.
First we calculate your peak restore rate. Your peak hourly restore rate each month is equal to the greatest amount of data you restore in any hour over the course of the month. If you initiate several restores in the same hour, these are added together to determine your hourly restore rate. We always assume that a restore request completes in 4 hours for the purpose of calculating your peak restore rate. In this case your peak rate is 140GB/4 hours, which equals 35 GB per hour.

Then we calculate your peak billable restore rate by subtracting the amount of data you get for free from your peak rate. To calculate your free data we look at your daily allowance and divide it by the number of hours in the day that you restored your data. So in this case your free data is 128 GB /4 hours or 32 GB free per hour. This makes your peak billable restore rate as 35 GB/hour – 32 GB/hour which equals 3 GB per hour.

To calculate how much you pay for the month we multiply your peak billable restore rate (3 GB per hour) by the data restore fee ($0.01/GB) by the number of hours in a month (720 hrs). So in this instance you pay 3 GB/Hour * $0.01 * 720 hours, which equals $21.60 to restore 140 GB in 3-5 hours.

Example 2: Archiving 75TB of data to Amazon Glacier and restoring 140GB in 8 hours.
First we calculate your peak restore rate. Again, for the purpose of calculating your restore fee, we always assume restores complete in 4 hours. If you send requests to restore 140GB of data over an 8 hour period, your peak restore rate would then be 140GB / 8 hours = 17.50 GB per hour. (This assumes that your restores start and end in the same day).

Then we calculate your peak billable restore rate by subtracting the amount of data you get for free from your peak rate. To calculate your free data we look at your daily allowance and divide it by the number of hours in the day that you restored your data. So in this case your free data is 128 GB /8 hours or 16 GB free per hour. This makes your billable rate 17.5 GB/hour – 16 GB/hour which equals 1.5 GB/hour. To calculate how much you pay for the month we multiply your peak usage in a single hour (1.5 GB/hour) by the restore fee ($0.01/GB) by the number of hours in a month (720 hrs). So in this instance you pay 1.5 GB/hour * $0.01 * 720 hours, which equals $10.80 to restore 140 GB.

Example 3: Archiving 75TB of data to Amazon Glacier and restoring 140GB in 28 hours.
If you spread your restores over 28 hours, you would no longer exceed your daily free retrieval allowance and would therefore not be charged a Data Restore Fee.

Q: How am I charged for deleting objects from Amazon Glacier that are less than 3 months old?

Amazon Glacier is designed for use cases where data is retained for months, years, or decades. Deleting data that is archived to Amazon Glacier is free if the objects being deleted have been archived in Amazon Glacier for three months or longer. If an object archived in Amazon Glacier is deleted or overwritten within three months of being archived then there will be an early deletion fee. This fee is prorated. If you delete 1GB of data 1 month after uploading it, you will be charged an early deletion fee for 2 months of Amazon Glacier storage. If you delete 1 GB after 2 months, you will be charged for 1 month of Amazon Glacier storage.

Cross-Region Replication

Q: What is Amazon S3 Cross-Region Replication (CRR)?

CRR is an Amazon S3 feature that automatically replicates data across AWS regions. With CRR, every object uploaded to an S3 bucket is automatically replicated to a destination bucket in a different AWS region that you choose. You can use CRR to provide lower-latency data access in different geographic regions. CRR can also help if you have a compliance requirement to store copies of data hundreds of miles apart.

Q: How do I enable CRR?

CRR is a bucket-level configuration. You enable a CRR configuration on your source bucket by specifying a destination bucket in a different region for replication. You can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to enable CRR. Versioning must be turned on for both the source and destination buckets to enable CRR. To learn more, please visit How to Set Up Cross-Region Replicationin the Amazon S3 Developer Guide.

Q: What does CRR replicate to the target bucket?

CRR replicates every object-level upload that you directly make to your source bucket. The metadata and ACLs associated with the object are also part of the replication. Any change to the underlying data, metadata, or ACLs on the object would trigger a new replication to the destination bucket. You can either choose to replicate all objects uploaded to a source bucket or just a subset of objects uploaded by specifying prefixes. Existing data in the bucket prior to enabling CRR is not replicated. You can use S3's COPY API to copy the existing data into your destination bucket. To learn more about CRR please visit How to Set Up Cross-Region Replication in the Amazon S3 Developer Guide.

Q: Can I use CRR with lifecycle rules?

Yes, you can configure separate lifecycle rules on the source and destination buckets. For example, you can configure a lifecycle rule to migrate data from Standard to Standard - IA on the destination bucket or configure a lifecycle rule to archive data into Amazon Glacier.

Q: What is the pricing for CRR?

You pay the Amazon S3 charges for storage, requests, and inter-region data transfer for the replicated copy of data. For example, if you replicate 1,000 1 GB objects (1,000 GB) between regions you will incur a request charge of $0.005 (1,000 requests x $0.005 per 1,000 requests) for replicating 1,000 objects and a charge of $20 ($0.020 per GB transferred x 1,000 GB) for inter-region data transfer. After replication, the 1,000 GB will incur storage charges based on the destination region.

If the source object is uploaded using the multipart upload feature, then it is replicated using the same number of parts and part size. For example, a 100 GB object uploaded using the multipart upload feature (800 parts of 128 MB each) will incur request cost associated with 802 requests (800 Upload Part requests + 1 Initiate Multipart Upload request + 1 Complete Multipart Upload request) when replicated. You will incur a request charge of $0.00401 (802 requests x $0.005 per 1,000 requests) and a charge of $2.00 ($0.020 per GB transferred x 100 GB) for inter-region data transfer. After replication, the 100 GB will incur storage charges based on the destination region.

Please visit the S3 pricing page for more information.

Event Notification

Q1: What are Amazon S3 event notifications?

Amazon S3 event notifications can be sent in response to actions in Amazon S3 like PUTs, POSTs, COPYs, or DELETEs. Notification messages can be sent through either Amazon SNS, Amazon SQS, or directly to AWS Lambda.

Q2: What can I do with Amazon S3 event notifications?

Amazon S3 event notifications enable you to run workflows, send alerts, or perform other actions in response to changes in your objects stored in Amazon S3. You can use Amazon S3 event notifications to set up triggers to perform actions including transcoding media files when they are uploaded, processing data files when they become available, and synchronizing Amazon S3 objects with other data stores. You can also set up event notifications based on object name prefixes and suffixes. For example, you can choose to receive notifications on object names that start with “images/."

Q3: What is included in an Amazon S3 event notification?

For a detailed description of the information included in Amazon S3 event notification messages, please refer to the Configuring Amazon S3 event notifications topic in the Amazon S3 Developer Guide.

Q4: How do I set up Amazon S3 event notifications?

For a detailed description of how to configure event notifications, please refer to theConfiguring Amazon S3 event notifications topic in the Amazon S3 Developer Guide. You can learn more about the AWS messaging services in the Amazon SNS Documentation and the Amazon SQS Documentation.

Q5: What does it cost to use Amazon S3 event notifications?

There are no additional charges from Amazon S3 for event notifications. You pay only for use of Amazon SNS or Amazon SQS to deliver event notifications, or for the cost of running the AWS Lambda function. Visit the Amazon SNS, Amazon SQS, or AWS Lambda pricing pages to view the pricing details for these services.

Static Website Hosting

Q: Can I host my static website on Amazon S3?

Yes, you can host your entire static website on Amazon S3 for an inexpensive, highly available hosting solution that scales automatically to meet traffic demands. Amazon S3 gives you access to the same highly scalable, reliable, fast, inexpensive infrastructure that Amazon uses to run its own global network of web sites. Service availability corresponds to storage class and the service level agreement provides service credits if a customer’s availability falls below our service commitment in any billing cycle. To learn more about hosting your website on Amazon

S3, please see our walkthrough on setting up an Amazon S3 hosted website.

Q: What kinds of websites should I host using Amazon S3 static website hosting?

Amazon S3 is ideal for hosting websites that contain only static content, including html files, images, videos, and client-side scripts such as JavaScript. Amazon EC2 is recommended for websites with server-side scripting and database interaction.

Q: Can I use my own host name with my Amazon S3 hosted website?

Yes, you can easily and durably store your content in an Amazon S3 bucket and map your domain name (e.g. “example.com”) to this bucket. Visitors to your website can then access this content by typing in your website’s URL (e.g., “http://example.com”) in their browser.

Q: Does Amazon S3 support website redirects?

Yes, Amazon S3 provides multiple ways to enable redirection of web content for your static websites. Redirects enable you to change the Uniform Resource Locator (URL) of a web page on your Amazon S3 hosted website (e.g. from www.example.com/oldpage to www.example.com/newpage) without breaking links or bookmarks pointing to the old URL. You can set rules on your bucket to enable automatic redirection. You can also configure a redirect on an individual S3 object.

Q: Is there an additional charge for hosting static websites on Amazon S3?

There is no additional charge for hosting static websites on Amazon S3. The same pricing dimensions of storage, requests, and data transfer apply to your website objects.

Refer to the S3 Pricing page for more information.

Lifecycle Management Policies

Q. What is Lifecycle Management?

S3 Lifecycle management provides the ability to define the lifecycle of your object with a predefined policy and reduce your cost of storage. You can set lifecycle transition policy to automatically migrate Amazon S3 objects to Standard - Infrequent Access (Standard - IA) and/or Amazon Glacier based on the age of the data. You can also set lifecycle expiration policies to automatically remove objects based on the age of the object. You can set a policy for multipart upload expiration, which expires incomplete multipart upload based on the age of the upload.

Q. How do I set up a lifecycle management policy?

You can set up and manage lifecycle policies in the AWS Management Console, S3 REST API, AWS SDKs, or AWS Command Line Interface (CLI). You can specify the policy at the prefix or at the bucket level.

Q: How much does it cost to use lifecycle management?

There is no additional cost to set up and apply lifecycle policies. A transition request is charged per object when an object becomes eligible for transition according to the lifecycle rule. Refer to the S3 Pricing page for pricing information.

Q. What can I do with Lifecycle Management Policies?

As data matures, it can become less critical, less valuable and subject to compliance requirements. Amazon S3 includes an extensive library of policies that help you automate data migration processes. For example, you can set infrequently accessed objects to move into lower cost storage tier (like Standard-Infrequent Access) after a period of time. After another period, it can be moved into Amazon Glacier for archive and compliance, and eventually deleted. These rules can invisibly lower storage costs and simplify management efforts and may be leveraged across the Amazon family of storage services. And these policies also include good stewardship practices to remove objects and attributes that are no longer needed to manage cost and optimize performance.

Q: How can I use Amazon S3's lifecycle policy to lower my Amazon S3 storage costs?

With Amazon S3's lifecycle policies, you can configure your objects to be migrated to Standard - Infrequent Access (Standard - IA), archived to Amazon Glacier, or deleted after a specific period of time. You can use this policy-driven automation to quickly and easily reduce storage costs as well as save time. In each rule you can specify a prefix, a time period, a transition to Standard - IA or Amazon Glacier, and/or an expiration. For example, you could create a rule that archives into Amazon Glacier all objects with the common prefix “logs/” 30 days from creation, and expires these objects after 365 days from creation. You can also create a separate rule that only expires all objects with the prefix “backups/” 90 days from creation. Lifecycle policies apply to both existing and new S3 objects, ensuring that you can optimize storage and maximize cost savings for all current data and any new data placed in S3 without time-consuming manual data review and migration. Within a lifecycle rule, the prefix field identifies the objects subject to the rule. To apply the rule to an individual object, specify the key name. To apply the rule to a set of objects, specify their common prefix (e.g. “logs/”). You can specify a transition action to have your objects archived and an expiration action to have your objects removed. For time period, provide the creation date (e.g. January 31, 2015) or the number of days from creation date (e.g. 30 days) after which you want your objects to be archived or removed. You may create multiple rules for different prefixes.

Learn more.

Q: How can I configure my objects to be deleted after a specific time period?

You can set a lifecycle expiration policy to remove objects from your buckets after a specified number of days. You can define the expiration rules for a set of objects in your bucket through the Lifecycle Configuration policy that you apply to the bucket. Each Object Expiration rule allows you to specify a prefix and an expiration period. The prefix field identifies the objects subject to the rule. To apply the rule to an individual object, specify the key name. To apply the rule to a set of objects, specify their common prefix (e.g. “logs/”). For expiration period, provide the number of days from creation date (i.e. age) after which you want your objects removed. You may create multiple rules for different prefixes. For example, you could create a rule that removes all objects with the prefix “logs/” 30 days from creation, and a separate rule that removes all objects with the prefix “backups/” 90 days from creation.

After an Object Expiration rule is added, the rule is applied to objects that already exist in the bucket as well as new objects added to the bucket. Once objects are past their expiration date, they are identified and queued for removal. You will not be billed for storage for objects on or after their expiration date, though you may still be able to access those objects while they are in queue before they are removed. As with standard delete requests, Amazon S3 doesn’t charge you for removing objects using Object Expiration. You can set Expiration rules for your versioning-enabled or versioning-suspended buckets as well.

Learn more.

Q. Why would I use a lifecycle policy to expire incomplete multipart uploads?

The lifecycle policy that expires incomplete multipart uploads allows you to save on costs by limiting the time non-completed multipart uploads are stored. For example, if your application uploads several multipart object parts, but never commits them, you will still be charged for that storage. This policy lowers your S3 storage bill by automatically removing incomplete multipart uploads and the associated storage after a predefined number of days.

Learn more.

Amazon S3 Transfer Acceleration

Q. What is Transfer Acceleration?

Amazon S3 Transfer Acceleration enables fast, easy, and secure transfers of files over long distances between your client and your Amazon S3 bucket. Transfer Acceleration leverages Amazon CloudFront’s globally distributed AWS Edge Locations. As data arrives at an AWS Edge Location, data is routed to your Amazon S3 bucket over an optimized network path.

Q. How do I get started with Transfer Acceleration?

It's easy to get started with Transfer Acceleration. First,enable Transfer Acceleration on an S3 bucket using the Amazon S3 console, the Amazon S3 API, or the AWS CLI. After Transfer Acceleration is enabled, you can point your Amazon S3 PUT and GET requests to the s3-accelerate endpoint domain name. Your data transfer application must use one of the following two types of endpoints to access the bucket for faster data transfer: <bucketname>.s3-accelerate.amazonaws.com or <bucketname>.s3-accelerate.dualstack.amazonaws.com for the "dual-stack" endpoint. If you want to use standard data transfer, you can continue to use the regular endpoints.

There are certain restrictions on which bucket will work with transfer acceleration. For details, please refer the Amazon S3 developer guide here.

Q. How fast is Transfer Acceleration?

Transfer Acceleration helps you fully utilize your bandwidth, minimize the effect of distance on throughput, and is designed to ensure consistently fast data transfer to Amazon S3 regardless of your client's location. Acceleration primarily depends on your available bandwidth, the distance between the source and destination, and packet loss rates on the network path. Generally, you will see more acceleration when the source is farther from the destination, when there is more available bandwidth, and/or when the object size is bigger.

One customer measured a 50% reduction in their average time to ingest 300 MB files from a global user base spread across the US, Europe, and parts of Asia to a bucket in the Asia Pacific (Sydney) region. Another customer observed cases where performance improved in excess of 500% for users in South East Asia and Australia uploading 250 MB files (in parts of 50MB) to an S3 bucket in the US East (N. Virginia) region.

Try the speed comparison tool to get a preview of the performance benefit from your location!

Q. Who should use Transfer Acceleration?

Transfer Acceleration is designed to optimize transfer speeds from across the world into S3 buckets. If you are uploading to a centralized bucket from geographically dispersed locations, or if you regularly transfer GBs or TBs of data across continents, you may save hours or days of data transfer time.

Q. How secure is Transfer Acceleration?

Transfer Acceleration provides the same security as regular transfers to Amazon S3. All Amazon S3 security features, such as restricting access based on a client's IP address, are supported as well. Transfer Acceleration communicates with clients over standard TCP and does not require firewall changes. No data is ever saved at AWS Edge Locations.

Q. What if Transfer Acceleration isn't faster?

Each time you use Transfer Acceleration to upload an object, we will check whether Transfer

Acceleration is likely to be faster than a regular Amazon S3 transfer. If we determine that Transfer Acceleration is not likely to be faster than a regular Amazon S3 transfer of the same object to the same destination AWS region, we will not charge for that use of Transfer Acceleration for that transfer, and may bypass the Transfer Acceleration system for that upload.

Q. Can I use Transfer Acceleration with multipart uploads?

Yes, Transfer Acceleration supports all bucket level features including multipart upload.

Q. How should I choose between Transfer Acceleration and Amazon CloudFront's PUT/POST?

Transfer Acceleration optimizes the TCP protocol and adds additional intelligence between the client and the S3 bucket, making Transfer Acceleration a better choice if a higher throughput is desired. If you have objects that are smaller than 1GB or if the data set is less than 1GB in size, you should consider using Amazon CloudFront's PUT/POST commands for optimal performance.

Q. How should I choose between Transfer Acceleration and AWS Snowball?

The AWS Import/Export Snowball is ideal for customers moving large batches of data at once. The AWS Snowball has a typical 5-7 days turnaround time. As a rule of thumb, Transfer Acceleration over a fully-utilized 1 Gbps line can transfer up to 75 TBs in the same time. In general, if it will take more than a week to transfer over the Internet, or there are recurring transfer jobs and there is more than 25Mbps of available bandwidth, Transfer Acceleration is a good option. Another option is to use both: perform initial heavy lift moves with an AWS Snowball (or series of AWS Snowballs) and then transfer incremental ongoing changes with Transfer Acceleration.

Q. Can Transfer Acceleration complement AWS Direct Connect?

AWS Direct Connect is a good choice for customers with a private networking requirement or have access to AWS Direct Connect exchanges. Transfer Acceleration is best for submitting data from distributed client locations over the public Internet, or where variable network conditions make throughput poor. Some AWS Direct Connect customers use Transfer Acceleration to help with remote office transfers, where they may suffer from poor Internet performance.

Q. Can Transfer Acceleration complement the AWS Storage Gateway or a 3rd party gateway?

Yes. Storage Gateways are built to extend Amazon S3 into an on-premises environment. Transfer Acceleration may make the connection between Amazon S3 and a gateway perform faster, improving the experience for gateway users.

Q. Can Transfer Acceleration complement 3rd party integrated software?

Yes. Software packages that connect directly into Amazon S3 (read more about storage partner solutions here) can take advantage of Transfer Acceleration when they send their jobs to Amazon S3.

Amazon S3 and IPv6

Q. What is IPv6?

Every server and device connected to the Internet must have a unique address. Internet Protocol Version 4 (IPv4) was the original 32-bit addressing scheme. However, the continued growth of the Internet means that all available IPv4 addresses will be utilized over time. Internet Protocol Version 6 (IPv6) is the new addressing mechanism designed to overcome the global address limitation on IPv4.

Q. What can I do with IPv6?

Using IPv6 support for Amazon S3, applications can connect to Amazon S3 without needing any IPv6 to IPv4 translation software or systems. You can meet compliance requirements, more easily integrate with existing IPv6-based on-premises applications, and remove the need for expensive networking equipment to handle the address translation. You can also now utilize the existing source address filtering features in IAM policies and bucket policies with IPv6 addresses, expanding your options to secure applications interacting with Amazon S3.

Q. How do I get started with IPv6 on Amazon S3?

You can get started by pointing your application to Amazon S3's new "dual-stack"endpoint, which supports access over both IPv4 and IPv6. In most cases, no further configuration is required for access over IPv6, because most network clients prefer IPv6 addresses by default. Your applications may continue to access data through the existing APIs and virtual hosted style (e.g. http://bucket.s3.dualstack.aws-region.amazonaws.com) or path style (e.g. http://s3.dualstack.aws-region.amazonaws.com/bucket) URLs without code changes. When using Amazon S3 Transfer Acceleration, the "dual-stack" endpoint must be of the form http(s)://bucket.s3-accelerate.dualstack.amazonaws.com. However, you must also evaluate your bucket and Identity and Access Management (IAM) policies to ensure you have the appropriate access configured for your new IPv6 addresses. For more information about getting started accessing Amazon S3 over IPv6, see Making Requests to Amazon S3 over IPv6.

Q. If I point to Amazon S3's "dual-stack" endpoint, will I still be able to access Amazon S3's APIs over IPv4?

Yes, you can continue to access Amazon S3 APIs using both IPv6 and IPv4 addresses when connecting to the Amazon S3 "dual-stack" endpoints. You will need to configure your client to prefer IPv4 addresses, which can be an application-level or host-level configuration option for many application runtime languages. Please consult the documentation for the language you are using for your runtime platform for the specific configuration option that prefers IPv4 connections.

Q. Should I expect a change in Amazon S3 performance when using IPv6?

No, you will see the same performance when using either IPv4 or IPv6 with Amazon S3.

Q. Will existing VPC Endpoints continue to work if I point to Amazon S3's "dual-stack" endpoint?

Yes, you can continue using VPC Endpoint to access Amazon S3 over IPv4. If you use the dual-stack endpoint in an IPv4-only VPC, the VPC instances will drop the AAAA record and always access Amazon S3 over IPv4.

Q. If I enable IPv6, will the IPv6 address appear in the Server Access Log?

Yes, IPv6 addresses will now be shown in the Server Access logs if you have the Amazon S3 Server Access logs feature enabled. Any customer tool or software that parses the logs should be updated to handle the new IPv6 address format. Please contact Developer Support if you have any issues with IPv6 traffic impacting your tool or software’s ability to handle IPv6 addresses in Server Access logs.

Q. Do I need to update my bucket and IAM policies?

Yes, if you use policies to grant or restrict access via IP addresses, you will need to update those policies to include the associated IPv6 ranges before you switch to the “dual-stack” endpoint. If your bucket grants or restricts access to specific IAM users, you will also need to have the IAM policy administrator review those users’ IAM policies to ensure they have appropriate access to the associated IPv6 ranges before you switch to the “dual-stack” endpoint. Failure to do so may result in clients incorrectly losing or gaining access to the bucket when they start using IPv6.

Q: What can I do if my clients are impacted by policy, network, or other restrictions in using IPv6 for Amazon S3?

Applications that are impacted by using IPv6 can switch back to the standard IPv4-only endpoints at any time.

Q: Can I use IPv6 with all Amazon S3 features?

No, IPv6 support is not currently available when using Website Hosting and access via BitTorrent. All other features should work as expected when accessing Amazon S3 using IPv6.

Q: Is IPv6 supported in all regions?

You can use IPv6 with Amazon S3 in all commercial AWS Regions except China (Beijing). You can also use IPv6 in the AWS GovCloud (US) region.

Amazon CloudFront FAQ

Q. How do I get started with Amazon CloudFront?

Click the “Create Free Account” button on the Amazon CloudFront detail page. If you choose to use another Amazon Web Service as the origin for the files served through Amazon CloudFront, you must sign up for that service before creating CloudFront distributions.

Q. How do I use Amazon CloudFront?

To use Amazon CloudFront, you:

- For static files, store the definitive versions of your files in one or more origin servers. These could be Amazon S3 buckets. For your dynamically generated content that is personalized or customized, you can use Amazon EC2 – or any other web server – as the origin server. These origin servers will store or generate your content that will be distributed through Amazon CloudFront.

- Register your origin servers with Amazon CloudFront through a simple API call. This call will return a CloudFront.net domain name that you can use to distribute content from your origin servers via the Amazon CloudFront service. For instance, you can register the Amazon S3 bucket “bucketname.s3.amazonaws.com” as the origin for all your static content and an Amazon EC2 instance “dynamic.myoriginserver.com” for all your dynamic content. Then, using the API or the AWS Management Console, you can create an Amazon CloudFront distribution that might return “abc123.cloudfront.net” as the distribution domain name.

- Include the cloudfront.net domain name, or a CNAME alias that you create, in your web application, media player, or website. Each request made using the cloudfront.net domain name (or the CNAME you set-up) is routed to the edge location best suited to deliver the content with the highest performance. The edge location will attempt to serve the request with a local copy of the file. If a local copy is not available, Amazon CloudFront will get a copy from the origin. This copy is then available at that edge location for future requests.

Q. How does Amazon CloudFront provide higher performance?

Amazon CloudFront employs a network of edge locations that cache copies of popular files close to your viewers. Amazon CloudFront ensures that end-user requests are served by the closest edge location. As a result, requests travel shorter distances to request objects, improving performance. For files not cached at the edge locations, Amazon CloudFront keeps persistent connections with your origin servers so that those files can be fetched from the origin servers as quickly as possible. Finally, Amazon CloudFront uses additional optimizations – e.g. wider TCP initial congestion window – to provide higher performance while delivering your content to viewers.

Q. How does Amazon CloudFront lower my costs to distribute content over the Internet?

Like other AWS services, Amazon CloudFront has no minimum commitments and charges you only for what you use. Compared to self-hosting, Amazon CloudFront spares you from the expense and complexity of operating a network of cache servers in multiple sites across the internet and eliminates the need to over-provision capacity in order to serve potential spikes in traffic. Amazon CloudFront also uses techniques such as collapsing simultaneous viewer requests at an edge location for the same file into a single request to your origin server. This reduces the load on your origin servers reducing the need to scale your origin infrastructure, which can bring you further cost savings.

Additionally, if you are using an AWS origin (e.g., Amazon S3, Amazon EC2, etc.), effective December 1, 2014, we are no longer charging for AWS data transfer out to Amazon CloudFront. This applies to data transfer from all AWS regions to all global CloudFront edge locations.

Q. How is Amazon CloudFront different from Amazon S3?

Amazon CloudFront is a good choice for distribution of frequently accessed static content that benefits from edge delivery—like popular website images, videos, media files or software downloads.

Q. How is Amazon CloudFront different from traditional content delivery solutions?

Amazon CloudFront lets you quickly obtain the benefits of high performance content delivery without negotiated contracts or high prices. Amazon CloudFront gives all developers access to inexpensive, pay-as-you-go pricing – with a self-service model. Developers also benefit from tight integration with other Amazon Web Services. The solution is simple to use with Amazon S3, Amazon EC2, and Elastic Load Balancing as origin servers, giving developers a powerful combination of durable storage and high performance delivery. Amazon CloudFront also integrates with Amazon Route 53 and AWS CloudFormation for further performance benefits and ease of configuration.

Q. How will I be charged for my use of Amazon CloudFront?

Amazon CloudFront charges are based on actual usage of the service in four areas: Data Transfer Out, HTTP/HTTPS Requests, Invalidation Requests, and Dedicated IP Custom SSL certificates associated with a CloudFront distribution.

With the AWS Free Usage Tier, you can get started with Amazon CloudFront for free. Upon sign-up, new AWS customers receive 50 GB Data Transfer Out and 2,000,000 HTTP and HTTPS Requests for Amazon CloudFront each month for one year.

- **Data Transfer Out to Internet**
  You will be charged for the volume of data transferred out of the Amazon CloudFront edge locations, measured in GB. If you are using other Amazon Web Services as the origins of your files, you will be charged separately for use of those services, including for storage,

compute hours, GET requests and data transfer out of that service to Amazon CloudFront's edge locations. Usage tiers for data transfer are measured separately for each geographic region. You can see the rates for Amazon CloudFront data transfer to the Internet here.

- **Data Transfer Out to Origin**
  You will be charged for the volume of data transferred out, measured in GB, from the Amazon CloudFront edge locations to your origin (both AWS origins and other origin servers). You can see the rates for Amazon CloudFront data transfer to Origin here.

- **HTTP/HTTPS Requests**
  You will be charged for number of HTTP/HTTPS requests made to Amazon CloudFront for your content. You can see the rates for HTTP/HTTPS requests here.

- **Invalidation Requests**
  You are charged per path in your invalidation request. A path listed in your invalidation request represents the URL (or multiple URLs if the path contains a wildcard character) of the object you want to invalidate from CloudFront cache. You can request up to 1,000 paths each month from Amazon CloudFront at no additional charge. Beyond the first 1,000 paths, you will be charged per path listed in your invalidation requests. You can see the rates for invalidation requests here.

- **Dedicated IP Custom SSL**
  You pay $600 per month for each custom SSL certificate associated with one or more CloudFront distributions using the Dedicated IP version of custom SSL certificate support. This monthly fee is pro-rated by the hour. For example, if you had your custom SSL certificate associated with at least one CloudFront distribution for just 24 hours (i.e. 1 day) in the month of June, your total charge for using the custom SSL certificate feature in June will be (1 day / 30 days) * $600 = $20. To use Dedicated IP Custom SSL certificate support, upload a SSL certificate and use the AWS Management Console to associate it with your CloudFront distributions. If you need to associate more than two custom SSL certificates with your CloudFront distribution, please include details about your use case and the number of custom SSL certificates you intend to use in the CloudFront Limit Increase Form.

Usage tiers for data transfer are measured separately for each geographic region. The prices above are exclusive of applicable taxes, fees, or similar governmental charges, if any exist, except as otherwise noted.

Q: Does your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Q: How am I charged for 304 responses?

A 304 is a response to a conditional GET request and will result in a charge for the HTTP/HTTPS request and the Data Transfer Out to Internet. A 304 response does not contain a message-body; however, the HTTP headers will consume some bandwidth for which you would be charged standard CloudFront data transfer fees. The amount of data transfer depends on the headers associated with your object.

Q. Can I choose to only serve content from less expensive Amazon CloudFront regions?

Yes, "Price Classes" provides you an option to lower the prices you pay to deliver content out of Amazon CloudFront. By default, Amazon CloudFront minimizes end user latency by delivering content from its entire global network of edge locations. However, because we charge more where our costs are higher, this means that you pay more to deliver your content with low latency to end-users in some locations. Price Classes let you reduce your delivery prices by excluding Amazon CloudFront's more expensive edge locations from your Amazon CloudFront distribution. In these cases, Amazon CloudFront will deliver your content from edge locations within the locations in the price class you selected and charge you the data transfer and request pricing from the actual location where the content was delivered.

If performance is most important to you, you don't need to do anything; your content will be delivered by our whole network of locations. However, if you wish to use another Price Class, you can configure your distribution through the AWS Management Console or via the Amazon CloudFront API. If you select a price class that does not include all locations, some of your viewers, especially those in geographic locations that are not in your price class, may experience higher latency than if your content were being served from all Amazon CloudFront locations.

Note that Amazon CloudFront may still occasionally serve requests for your content from an edge location in a location that is not included in your price class. When this occurs, you will only be charged the rates for the least expensive location in your price class.

You can see the list of locations making up each price classhere.

Q. Can I choose to serve content (or not serve content) to specified countries?

Yes, the Geo Restriction feature lets you specify a list of countries in which your users can access your content. Alternatively, you can specify the countries in which your users cannot access your content. In both cases, CloudFront responds to a request from a viewer in a restricted country with an HTTP status code 403 (Forbidden).

Q. How accurate is your GeoIP database?

The accuracy of the IP Address to country lookup database varies by region. Based on recent tests, our overall accuracy for the IP address to country mapping is 99.8%.

Q. Can I serve a custom error message to my end users?

Yes, you can create custom error messages (for example, an HTML file or a .jpg graphic) with your own branding and content for a variety of HTTP 4xx and 5xx error responses. Then you can configure Amazon CloudFront to return your custom error messages to the viewer when your origin returns one of the specified errors to CloudFront.

Q. Where are the edge locations used by Amazon CloudFront located?

Amazon CloudFront uses a global network of edge locations for content delivery. You can see a full list of Amazon CloudFront locations here.

Q. What types of content does Amazon CloudFront support?

Amazon CloudFront supports all files that can be served over HTTP. This includes dynamic web pages, such as HTML or PHP pages, any popular static files that are a part of your web application, such as website images, audio, video, media files or software downloads. For on-demand media files, you can also choose to stream your content using RTMP delivery. Amazon CloudFront also supports delivery of live media over HTTP.

Q. Does Amazon CloudFront support delivery of dynamic content?

Amazon CloudFront supports all files that can be served over HTTP. This includes dynamic web pages, such as HTML or PHP pages, any popular static files that are a part of your web application, such as website images, audio streams, video streams, media files or software downloads. For on-demand media files, you can also choose to stream your content using RTMP delivery. Amazon CloudFront also supports delivery of live media over HTTP.

Q. Is Amazon CloudFront PCI compliant?

Yes, Amazon CloudFront is included in the set of services that are compliant with the Payment Card Industry Data Security Standard (PCI DSS) Merchant Level 1, the highest level of compliance for service providers. Please see our developer's guide for more information.

Q. How does Amazon CloudFront speed up my entire website?

Amazon CloudFront uses standard cache control headers you set on your files to identify static and dynamic content. Delivering all your content using a single Amazon CloudFront distribution helps you make sure that performance optimizations are applied to your entire website or web application. When using AWS origins, you benefit from improved performance, reliability, and ease of use as a result of AWS’s ability to track and adjust origin routes, monitor system health, respond quickly when any issues occur, and the integration of Amazon CloudFront with other AWS services. You also benefit from using different origins for different types of content on a single site – e.g. Amazon S3 for static objects, Amazon EC2 for dynamic content, and custom origins for third-party content – paying only for what you use.

Q. Does Amazon CloudFront work with non-AWS origin servers?

Yes. Amazon CloudFront works with any origin server that holds the original, definitive versions of your content, both static and dynamic. There is no additional charge to use a custom origin.

Q. What types of HTTP requests are supported by Amazon CloudFront?

Amazon CloudFront currently supports GET, HEAD, POST, PUT, PATCH, DELETE and OPTIONS requests.

Q. Does Amazon CloudFront cache POST responses?

Amazon CloudFront does not cache the responses to POST, PUT, DELETE, and PATCH requests – these requests are proxied back to the origin server. You may enable caching for the responses to OPTIONS requests.

Q. What is IPv6?

Every server and device connected to the Internet must have a numeric Internet Protocol (IP) address. As the Internet and the number of people using it grows exponentially, so does the need for IP addresses. IPv6 is a new version of the Internet Protocol that uses a larger address space than its predecessor IPv4. Under IPv4, every IP address is 32 bits long, which allows 4.3 billion unique addresses. An example IPv4 address is 192.0.2.1. In comparison, IPv6 addresses are 128 bits, which allow for approximately three hundred and forty trillion, trillion unique IP addresses. An example IPv6 address is: 2001:0db8:85a3:0:0:8a2e:0370:7334

Q. What can I do with IPv6?

Using IPv6 support for Amazon CloudFront, your applications can connect to Amazon CloudFront edge locations without needing any IPv6 to IPv4 translation software or systems. You can meet the requirements for IPv6 adoption set by governments - including the U.S. Federal government – and benefit from IPv6 extensibility, simplicity in network management, and additional built-in support for security.

Q. Should I expect a change in Amazon CloudFront performance when using IPv6?

No, you will see the same performance when using either IPv4 or IPv6 with Amazon CloudFront.

Q: Are there any Amazon CloudFront features that will not work with IPv6?

All existing features of Amazon CloudFront will continue to work on IPv6, though there are two changes you may need for internal IPv6 address processing before you turn on IPv6 for your distributions.

1. If you have turned on the Amazon CloudFront Access Logs feature, you will start seeing your viewer’s IPv6 address in the “c-ip” field and may need to verify that your log processing systems continue to work for IPv6.

2. When you enable IPv6 for your Amazon CloudFront distribution, you will get IPv6 addresses in the ‘X-Forwarded-For’ header that is sent to your origins. If your origin systems are only able to process IPv4 addresses, you may need to verify that your origin systems continue to

work for IPv6.

Additionally, if you use IP whitelists for Trusted Signers, you should use an IPv4-only distribution for your Trusted Signer URLs with IP whitelists and an IPv4 / IPv6 distribution for all other content. This model sidesteps an issue that would arise if the signing request arrived over an IPv4 address and was signed as such, only to have the request for the content arrive via a different IPv6 address that is not on the whitelist.

To learn more about IPv6 support in Amazon CloudFront, see“IPv6 support on Amazon CloudFront” in the Amazon CloudFront Developer Guide.

Q: Does that mean if I want to use IPv6 at all I cannot use Trusted Signer URLs with IP whitelist?

No. If you want to use IPv6 and Trusted Signer URLs with IP whitelist you should use two separate distributions. You should dedicate a distribution exclusively to your Trusted Signer URLs with IP whitelist and disable IPv6 for that distribution. You would then use another distribution for all other content, which will work with both IPv4 and IPv6.

Q. If I enable IPv6, will the IPv6 address appear in the Access Log?

Yes, your viewer’s IPv6 addresses will now be shown in the “c-ip” field of the access logs, if you have the Amazon CloudFront Access Logs feature enabled. You may need to verify that your log processing systems continue to work for IPv6 addresses before you turn on IPv6 for your distributions. Please contact Developer Support if you have any issues with IPv6 traffic impacting your tool or software’s ability to handle IPv6 addresses in access logs. For more details, please refer to the Amazon CloudFront Access Logs documentation.

Q: Can I disable IPv6 for all my new distributions?

Yes, for both new and existing distributions, you can use the Amazon CloudFront console or API to enable / disable IPv6 per distribution.

Q: Are there any reasons why I would want to disable IPv6?

In discussions with customers, the only common case we heard about was internal IP address processing. When you enable IPv6 for your Amazon CloudFront distribution, in addition to getting an IPv6 address in your detailed access logs, you will get IPv6 addresses in the ‘X-Forwarded-For’ header that is sent to your origins. If your origin systems are only able to process IPv4 addresses, you may need to verify that your origin systems continue to work for IPv6 addresses before you turn on IPv6 for your distributions.

Q: I enabled IPv6 for my distribution but a DNS lookup doesn’t return any IPv6 addresses. What is happening?

Amazon CloudFront has very diverse connectivity around the globe, but there are still certain networks that do not have ubiquitous IPv6 connectivity. While the long term future of the Internet is obviously IPv6, for the foreseeable future every endpoint on the Internet will have IPv4 connectivity. When we find parts of the Internet that have better IPv4 connectivity than IPv6, we will prefer the former.

Q: If I use Route 53 to handle my DNS needs and I created an alias record pointing to an Amazon CloudFront distribution, do I need to update my alias records to enable IPv6?

Yes, you can create Route 53 alias records pointing to your Amazon CloudFront distribution to support both IPv4 and IPv6 by using “A” and “AAAA” record type respectively. If you want to enable IPv4 only, you need only one alias record with type “A”. For details on alias resource record sets, please refer to the Amazon Route 53 Developer Guide.

Q. How do I use HTTP/2?

If you have an existing Amazon CloudFront distribution, you can turn on HTTP/2 using the API or the Management Console. In the Console, go to the “Distribution Configuration” page and navigate to the section “Supported HTTP Versions.” There, you can select "HTTP/2, HTTP/1.1, or HTTP/1.0". HTTP/2 is automatically enabled for all new CloudFront distributions.

Q. What if my origin does not support HTTP/2?

Amazon CloudFront currently supports HTTP/2 for delivering content to your viewers’ clients and browsers. For communication between the edge location and your origin servers, Amazon CloudFront will continue to use HTTP/1.1.

Q. Does Amazon CloudFront support HTTP/2 without TLS?

Not currently. However, most of the modern browsers support HTTP/2 only over an encrypted connection. You can learn more about using SSL with Amazon CloudFront here.

Q. Does Amazon CloudFront support access controls for paid or private content?

Yes, Amazon CloudFront has an optional private content feature. When this option is enabled, Amazon CloudFront will only deliver files when you say it is okay to do so by securely signing your requests. Learn more about this feature by reading the CloudFront Developer Guide.

Q. How can I protect my web applications delivered via CloudFront?

You can integrate your CloudFront distribution with AWS WAF, a web application firewall that helps protect web applications from attacks by allowing you to configure rules based on IP addresses, HTTP headers, and custom URI strings. Using these rules, AWS WAF can block, allow, or monitor (count) web requests for your web application. Please see AWS WAF Developer Guide for more information.

Q. Does Amazon CloudFront support CNAMEs?

Yes. You can add multiple CNAME aliases to each of your distributions. You can see the number of CNAME aliases supported per distribution here. Amazon CloudFront also supports wildcard CNAMEs.

Q. Can I use a CNAME alias I create for my CloudFront distribution to deliver content over HTTPS?

Yes. You can use either the Dedicated IP Custom SSL or SNI Custom SSL feature to deliver content over HTTPS using your own domain name and your own SSL certificate. This gives visitors to your website the security benefits of CloudFront over an SSL connection that uses your own domain name in addition to lower latency and higher reliability. Learn more about the Custom SSL features by visiting the CloudFront Custom SSL detail page and see how to set your CloudFront HTTPS settings by reading the CloudFront Developer Guide.

Q. What is the difference between SNI Custom SSL and Dedicated IP Custom SSL of Amazon CloudFront?

> **Dedicated IP Custom SSL** allocates dedicated IP addresses to serve your SSL content at each CloudFront edge location. Because there is a one to one mapping between IP addresses and SSL certificates, Dedicated IP Custom SSL works with browsers and other clients that do not support SNI. Due to the current IP address costs, Dedicated IP Custom SSL is $600/month prorated by the hour.
>
> **SNI Custom SSL** relies on the SNI extension of the Transport Layer Security protocol, which allows multiple domains to serve SSL traffic over the same IP address by including the hostname viewers are trying to connect to. As with Dedicated IP Custom SSL, CloudFront delivers content from each Amazon CloudFront edge location and with the same security as the Dedicated IP Custom SSL feature. SNI Custom SSL works with most modern browsers, including Chrome version 6 and later (running on Windows XP and later or OS X 10.5.7 and later), Safari version 3 and later (running on Windows Vista and later or Mac OS X 10.5.6. and later), Firefox 2.0 and later, and Internet Explorer 7 and later (running on Windows Vista and later). Older browsers that do not support SNI cannot establish a connection with CloudFront to load the HTTPS version of your content. SNI Custom SSL is available at no additional cost beyond standard CloudFront data transfer and request fees.

Q. What is Server Name Indication?

Server Name Indication (SNI) is an extension of the Transport Layer Security (TLS) protocol. This mechanism identifies the domain (server name) of the associated SSL request so the proper certificate can be used in the SSL handshake. This allows a single IP address to be used across multiple servers. SNI requires browser support to add the server name, and while most modern browsers support it, there are a few legacy browsers that do not. For more details see the SNI section of the CloudFront Developer Guide or the SNI Wikipedia article.

Q. Does CloudFront Integrate with AWS Certificate Manager?

Yes, you can now provision SSL/TLS certificates and associate them with CloudFront distributions within minutes. Simply provision a certificate using the new AWS Certificate Manager (ACM) and deploy it to your CloudFront distribution with a couple of clicks, and let ACM manage certificate renewals for you. ACM allows you to provision, deploy, and manage the certificate with no additional charges.

Note that CloudFront still supports using certificates that you obtained from a third-party certificate authority and uploaded to the IAM certificate store.

Q. Can I point my zone apex (example.com versus www.example.com) at my Amazon CloudFront distribution?

Yes. By using Amazon Route 53, AWS's authoritative DNS service, you can configure an 'Alias' record that lets you map the apex or root (example.com) of your DNS name to your Amazon CloudFront distribution. Amazon Route 53 will then respond to each request for an Alias record with the right IP address(es) for your CloudFront distribution. Route 53 doesn't charge for queries to Alias records that are mapped to a CloudFront distribution. These queries are listed as "Intra-AWS-DNS-Queries" on the Amazon Route 53 usage report.

Q. How does Amazon CloudFront handle query string parameters in the URL?

A query string may be optionally configured to be part of the cache key for identifying objects in the Amazon CloudFront cache. This helps you build dynamic web pages (e.g. search results) that may be cached at the edge for some amount of time.

Q. Can I specify which query parameters to use in the cache key?

Yes, query string whitelisting feature allows you to easily configure Amazon CloudFront to only use certain parameters in the cache key, while still forwarding all the parameters to the origin.

Q. Is there a limit to the number of query parameters that can be whitelisted?

Yes, you can configure Amazon CloudFront to whitelist up to 10 query parameters.

Q. What parameter types are supported?

Amazon CloudFront supports URI query parameters as defined in section 3.4 of RFC3986. Specifically, it supports query parameters embedded in an HTTP GET string after the '?' character, and delimited by the '&' character.

Q. Does CloudFront support gzip compression?

Yes, CloudFront will automatically compress your text or binary data. To use the feature, simply specify within your cache behavior settings that you would like CloudFront to compress objects automatically and ensure that your client adds Accept-Encoding: gzip in the request header (most modern web browser do this by default)..For more information on this feature, please refer to our developer guide.

Q. Can I add or modify request headers forwarded to the origin?

Yes, you can configure Amazon CloudFront to add custom headers, or override the value of existing headers, to requests forwarded to your origin. You can use these headers to help validate that requests made to your origin were sent from CloudFront; you can even configure your origin to only allow requests that contain the custom header values you specify. Additionally, if you use multiple CloudFront distributions with the same origin, you can use custom headers to distinguish origin request made by each different distribution. Finally, custom headers can be used to help determine the right CORS headers returned for your requests. You can configure custom headers via the CloudFront API and the AWS Management Console. There are no additional charges for this feature. For more details on how to set your custom headers, you can read more here.

Q. How does Amazon CloudFront handle HTTP cookies?

Amazon CloudFront supports delivery of dynamic content that is customized or personalized using HTTP cookies. To use this feature, you specify whether you want Amazon CloudFront to forward some or all of your cookies to your custom origin server. Amazon CloudFront then considers the forwarded cookie values when identifying a unique object in its cache. This way, your end users get both the benefit of content that is personalized just for them with a cookie and the performance benefits of Amazon CloudFront. You can also optionally choose to log the cookie values in Amazon CloudFront access logs.

Q. How long will Amazon CloudFront keep my files at the edge locations?

By default, if no cache control header is set, each edge location checks for an updated version of your file whenever it receives a request more than 24 hours after the previous time it checked the origin for changes to that file. This is called the “expiration period.” You can set this expiration period as short as 0 seconds, or as long as you’d like, by setting the cache control headers on your files in your origin. Amazon CloudFront uses these cache control headers to determine how frequently it needs to check the origin for an updated version of that file. For expiration period set to 0 seconds, Amazon CloudFront will revalidate every request with the origin server. If your files don’t change very often, it is best practice to set a long expiration period and implement a versioning system to manage updates to your files.

Q. How do I remove an item from Amazon CloudFront edge locations?

There are multiple options for removing a file from the edge locations. You can simply delete the file from your origin and as content in the edge locations reaches the expiration period defined in each object’s HTTP header, it will be removed. In the event that offensive or potentially harmful material needs to be removed before the specified expiration time, you can use the Invalidation API to remove the object from all Amazon CloudFront edge locations. You can see the charge for making invalidation requests here.

Q. Is there a limit to the number of invalidation requests I can make?

If you're invalidating objects individually, you can have invalidation requests for up to 3,000 objects per distribution in progress at one time. This can be one invalidation request for up to 3,000 objects, up to 3,000 requests for one object each, or any other combination that doesn't exceed 3,000 objects.

If you're using the * wildcard, you can have requests for up to 15 invalidation paths in progress at one time. You can also have invalidation requests for up to 3,000 individual objects per distribution in progress at the same time; the limit on wildcard invalidation requests is independent of the limit on invalidating objects individually. If you exceed this limit, further invalidation requests will receive an error response until one of the earlier request completes.

You should use invalidation only in unexpected circumstances; if you know beforehand that your files will need to be removed from cache frequently, it is recommended that you either implement a versioning system for your files and/or set a short expiration period.

Q. What is streaming? Why would I want to stream my content?

Generally, streaming refers to delivering audio and video to end users on the internet without having to download the media file prior to playback. The protocols used for streaming include proprietary ones such as Adobe’s Real Time Messaging Protocol (RTMP) and those that use HTTP for delivery such as Apple’s HTTP Live Streaming (HLS), Adobe’s HTTP Dynamic Streaming (HDS) and Microsoft’s Smooth Streaming. These protocols are different than the delivery of web pages and other content because streaming protocols deliver content in real time – the viewers watch the bytes as they are delivered. Streaming content has several potential benefits for you and your end-users:

- Streaming can give viewers more control over their viewing experience. For instance, it is easier for a viewer to seek forward in a video using streaming than using traditional download delivery.
- Streaming can give you more control over your content, as no file remains on the viewer's computer when they finish watching a video.
- Streaming can help you reduce your costs, as it only delivers portions of a media file that the viewers actually watch. In contrast, with traditional downloads, frequently the whole media file will be downloaded by the viewers, even if they only watch a portion of the file.

Q. Does Amazon CloudFront support on-demand streaming protocols?

Yes, Amazon CloudFront provides you with multiple options to deliver on-demand content. If you have media files that have been converted to either HLS format or Microsoft Smooth Streaming format prior to storing in Amazon S3 (or a custom origin), you can use an Amazon CloudFront web distribution to stream in that format without having to run any media servers. In addition you can also run a third party streaming server (e.g. Wowza Media Server available on AWS Marketplace) on Amazon EC2 which can convert a media file to the required HTTP streaming format. This server can then be designated as the origin for an Amazon CloudFront web distribution. Another option, if you want to stream using RTMP, is to store your media files on Amazon S3 and use it as the origin for an Amazon CloudFront RTMP distribution.

Q. Does Amazon CloudFront support live streaming to multiple platforms?

Yes. Amazon CloudFront provides you three options to easily and cost-effectively deliver live events over HTTP to multiple platforms:

- **Live Streaming using Wowza Media Server 3.6**: Using Amazon CloudFront with Wowza Media Server combines the benefits of Wowza Media Server with the reliability, scalability, low latency and cost-efficiency of Amazon CloudFront to stream live events to multiple streaming formats, including Apple HTTP Live Streaming (HLS), Adobe HTTP Dynamic Streaming (HDS) and Microsoft Smooth Streaming. We've made this simple for you by creating an AWS CloudFormation template that handles all of the provisioning and sequencing for all the AWS resources you need for this live streaming stack. Amazon CloudFront provides you the scale and flexible pay-as-you-go pricing model, while the use of HTTP protocols for streaming your live event offers your viewers easy access to your live content. Using Amazon CloudFront for live streaming also gives you full control of your Wowza origin server so you can configure it to best work with the specific nature of your event. In addition, you can choose the Amazon EC2 instance type and AWS region that best meet the needs of your live event. A detailed tutorial for setting-up live HTTP streaming using Amazon CloudFront is available here.

- **Live Streaming using Adobe Media Server 5.0**: Amazon CloudFront can be used with Amazon EC2 running Adobe Media Server (AMS 5.0) for live HTTP streaming to both Flash Player and Apple iOS devices. Amazon EC2 (running AMS 5.0) must be configured as the origin for a CloudFront web distribution. Similar to our other live streaming solutions we have setup an AWS CloudFormation template to make it easy for you to setup your pay-as-you-go streaming stack while providing you with full control of the AMS server running in the Amazon EC2 instances provisioned. A detailed tutorial (which also points to the AWS CloudFormation templates) for setting-up live HTTP streaming using CloudFront and AMS 5.0 is available here.

- **Live Streaming using Windows Media Services**: You can also use Amazon CloudFront and Amazon EC2 running Windows Media Services for live streaming. With this solution, you can deliver live media over HTTP to both Microsoft Silverlight clients and Apple iOS devices.

We've made it simple to get started by creating a tutorial and an AWS CloudFormation template to automate the provisioning of AWS resources for your live streaming stack. You only pay for the AWS resources you consume, and have full control over the origin server (Amazon EC2 instance running Windows Media Services) so you can configure additional IIS Live Smooth Streaming functionality.

Q. Does Amazon CloudFront support content coding?

Yes. Amazon CloudFront supports content coding. For more information about how to take advantage of this feature, please see the Developer's Guide.

Q. Can I use Amazon CloudFront if I expect usage peaks higher than 10 Gbps or 15,000 RPS?

Yes. Complete our request for higher limitshere, and we will add more capacity to your account within two business days.

Q: Is there a limit to the number of distributions my Amazon CloudFront account may deliver?

For the current limit on the number of distributions that you can create for each AWS account, see Amazon CloudFront Limits in the Amazon Web Services General Reference. To request a higher limit, please go to the CloudFront Limit Increase Form.

Q: What is the maximum size of a file that can be delivered through Amazon CloudFront?

The maximum size of a single file that can be delivered through Amazon CloudFront is 20 GB. This limit applies to all Amazon CloudFront distributions.

Q: What tools and libraries work with Amazon CloudFront?

There are a variety of tools for managing your Amazon CloudFront distribution and libraries for various programming languages available in our resource center.

Q: Can I get access to request logs for content delivered through Amazon CloudFront?

Yes. When you create or modify a CloudFront distribution, you can enable access logging. When enabled, this feature will automatically write detailed log information in a W3C extended format into an Amazon S3 bucket that you specify. Access logs contain detailed information about each request for your content, including the object requested, the date and time of the request, the edge location serving the request, the client IP address, the referrer, the user agent, the cookie header, and the result type (for example, cache hit/miss/error).

Q: **Does Amazon CloudFront offer ready-to-use reports so I can learn more about my usage, viewers, and content being served?**

Yes. Whether it's receiving detailed cache statistics reports, monitoring your CloudFront usage, seeing where your customers are viewing your content from, or setting near real-time alarms on operational metrics, Amazon CloudFront offers a variety of solutions for your reporting needs. You can access all our reporting options by visiting the Amazon CloudFront Reporting & Analytics dashboard in the AWS Management Console. You can also learn more about our various reporting options by viewing Amazon CloudFront's Reports & Analytics page.

Q: Can I tag my distributions?

Yes. Amazon CloudFront supports cost allocation tagging. Tags make it easier for you to allocate costs and optimize spending by categorizing and grouping AWS resources. For example, you can use tags to group resources by administrator, application name, cost center, or a specific project. To learn more about cost allocation tagging, see Using Cost Allocation Tags. If you are ready to add tags to you CloudFront distributions, see Amazon CloudFront Add Tags page.

Q: Can I get a history of all Amazon CloudFront API calls made on my account for security, operational or compliance auditing?

Yes. To receive a history of all Amazon CloudFront API calls made on your account, you simply turn on AWS CloudTrail in the CloudTrail's AWS Management Console. For more information, visit AWS CloudTrail home page.

Q: Do you have options for monitoring and alarming metrics in real time?

You can monitor, alarm and receive notifications on the operational performance of your Amazon CloudFront distributions within just a few minutes of the viewer request using Amazon CloudWatch. CloudFront automatically publishes six operational metrics, each at 1-minute granularity, into Amazon CloudWatch. You can then use CloudWatch to set alarms on any abnormal patterns in your CloudFront traffic. To learn how to get started monitoring CloudFront activity and setting alarms via CloudWatch, please view our walkthrough in the Amazon CloudFront Developer Guide or simply navigate to theAmazon CloudFront Management Console and select Monitoring & Alarming in the navigation pane.

Q: Can I use the AWS Management Console with Amazon CloudFront?

Yes. You can use the AWS Management Console to configure and manage Amazon CloudFront though a simple, point-and-click web interface. The AWS Management Console supports most of Amazon CloudFront’s features, letting you get Amazon CloudFront’s low latency delivery without writing any code or installing any software. Access to the AWS Management Console is provided free of charge at https://console.aws.amazon.com

Q: Does Amazon CloudFront offer a Service Level Agreement (SLA)?

Yes. The Amazon CloudFront SLA provides for a service credit if a customer’s monthly uptime percentage is below our service commitment in any billing cycle. More information can be found here

Amazon EFS FAQ

General

Q. What is Amazon Elastic File System?

Amazon EFS is a fully-managed service that makes it easy to set up and scale file storage in the Amazon cloud. With a few clicks in the AWS Management Console, you can create file systems that are accessible to Amazon EC2 instances via a file system interface (using standard operating system file I/O APIs) and that support full file system access semantics (such as strong consistency and file locking).

Amazon EFS file systems can automatically scale from gigabytes to petabytes of data without needing to provision storage. Tens, hundreds, or even thousands of Amazon EC2 instances can access an Amazon EFS file system at the same time, and Amazon EFS provides consistent performance to each Amazon EC2 instance. Amazon EFS is designed to be highly durable and highly available. With Amazon EFS, there is no minimum fee or setup costs, and you pay only for the storage you use.

Q. What use cases is Amazon EFS intended for?

Amazon EFS is designed to provide performance for a broad spectrum of workloads and applications, including Big Data and analytics, media processing workflows, content management, web serving, and home directories.

Q. When should I use Amazon EFS vs. Amazon Simple Storage Service (S3) vs. Amazon Elastic Block Store (EBS)?

Amazon Web Services (AWS) offers cloud storage services to support a wide range of storage workloads.

Amazon EFS is a file storage service for use with Amazon EC2. Amazon EFS provides a file system interface, file system access semantics (such as strong consistency and file locking), and concurrently-accessible storage for up to thousands of Amazon EC2 instances.

Amazon EBS is a block level storage service for use with Amazon EC2. Amazon EBS can deliver performance for workloads that require the lowest-latency access to data from a single

EC2 instance.

Amazon S3 is an object storage service. Amazon S3 makes data available through an Internet API that can be accessed anywhere.

Q. Where is my data stored?

Please refer to Regional Products and Services for details of Amazon EFS service availability by region.

Q. How do I get started using Amazon EFS?

To use Amazon EFS, you must have an Amazon Web Services account. If you do not already have an AWS account, you can create one by clicking the “Try the Free Tier” button on the Amazon EFS detail page.

Once you have created an AWS account, please refer to the Amazon EFSGetting Started guide to begin using Amazon EFS. You can create a file system via the AWS Management Console, the AWS Command Line Interface (AWS CLI), and Amazon EFS API (and various language-specific SDKs).

Q. How do I access a file system from an Amazon EC2 instance?

To access your file system, you mount the file system on an Amazon EC2 Linux-based instance using the standard Linux mount command. Once you’ve mounted, you can work with the files and directories in your file system just like you would with a local file system.

Amazon EFS uses the NFSv4.1 protocol. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the Amazon EFS Getting Started guide.

Q. What Amazon EC2 instance types and AMIs work with Amazon EFS?

Amazon EFS is compatible with all Amazon EC2 instance types and is accessible from Linux-based AMIs. You can mix and match the instance types connected to a single file system. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the Amazon EFS Getting Started guide.

Q. How do I manage a file system?

Amazon EFS is a fully-managed service, so all of the file storage infrastructure is managed for you. When you use Amazon EFS, you avoid the complexity of deploying and maintaining complex file system infrastructure. An Amazon EFS file system grows and shrinks automatically as you add and remove files, so you do not need to manage storage procurement or provisioning.

You can administer a file system via the AWS Management Console, the AWS command-line interface (CLI), or the Amazon EFS API (and various language-specific SDKs). The Console, API, and SDK provide the ability to create and delete file systems, configure how file systems are accessed, create and edit file system tags, and display detailed information about file systems.

Q. How do I load data into a file system?

Amazon EFS file systems are mounted on an Amazon EC2 instance, so any data that is accessible to an Amazon EC2 instance can also be read and written to Amazon EFS. To load data that is not currently stored on the Amazon cloud, you can use the same methods you use to transfer files to Amazon EC2 today, such as Secure Copy (SCP). For more information about moving data to the Amazon cloud, please see the Cloud Data Migration page.

Data Protection and Availability

Q. How is Amazon EFS designed to provide high durability and availability?

Every file system object (i.e. directory, file, and link) is redundantly stored across multiple Availability Zones. In addition, a file system can be accessed concurrently from all Availability Zones in the region where it is located, which means that you can architect your application to failover from one AZ to other AZs in the region in order to ensure the highest level of application availability.

Mount targets themselves are designed to be highly available. When designing your application for high availability and failover to other Availability Zones, keep in mind that the IP addresses and DNS for your mount targets in each Availability Zone are static.

Q. How do I back up a file system?

Amazon EFS is designed to be highly durable. If you want to be able to revert to earlier versions of files to undo changes, you can use standard 3rd party backup software.

You can also use AWS Data Pipeline to create regular, automated copies of your file system based on a schedule that you define. For more information and to access an AWS Data Pipeline template provided by Amazon EFS, please see the Amazon EFS Walkthrough: Back Up an EFS File System.

Q. Is my file system accessible directly from outside my VPC?

Amazon EC2 instances within your VPC can access your file system directly, and Amazon EC2-Classic instances can mount a file system via your VPC using ClassicLink.

Scale and Performance

Q. How much data can I store?

Amazon EFS file systems can store petabytes of data. Amazon EFS file systems are elastic, and automatically grow and shrink as you add and remove files. You do not provision file system size or specify a size up front and you pay only for the storage you use.

Q. How many Amazon EC2 instances can connect to a file system?

Amazon EFS supports one to thousands of Amazon EC2 instances connecting to a file system concurrently.

Q. How many file systems can I create?

By default, you can create up to 10 file systems per AWS account per region. You can request to increase your file system limit by visiting AWS Service Limits.

Q. How does Amazon EFS performance compare to that of other storage solutions?

Amazon EFS file systems are distributed across an unconstrained number of storage servers, enabling file systems to grow elastically to petabyte-scale and allowing massively parallel access from Amazon EC2 instances to your data. Amazon EFS's distributed design avoids the bottlenecks and constraints inherent to traditional file servers.

This distributed data storage design means that multi-threaded applications and applications that concurrently access data from multiple Amazon EC2 instances can drive substantial levels of aggregate throughput and IOPS. Big Data and analytics workloads, media processing workflows, content management and web serving are examples of these applications.

The table below compares high-level performance and storage characteristics for Amazon's file and block cloud storage offerings.

	Amazon EFS	**Amazon EBS PIOPS**
Per-operation latency	Low, consistent	Lowest, consistent
Throughput scale	Multiple GBs per second	Single GB per second

Amazon EFS’s distributed nature enables high levels of availability, durability, and scalability. This distributed architecture results in a small latency overhead for each file operation. Due to this per-operation latency, overall throughput generally increases as the average I/O size increases, since the overhead is amortized over a larger amount of data. Amazon EFS's support for highly parallelized workloads (i.e. with consistent operations from multiple threads and multiple EC2 instances) enables high levels of aggregate throughput and IOPS.

Q. What’s the difference between “General Purpose” and “Max I/O” performance modes? Which one should I choose?

“General Purpose” performance mode is appropriate for most file systems, and is the mode selected by default when you create a file system. “Max I/O” performance mode is optimized for applications where tens, hundreds, or thousands of EC2 instances are accessing the file system — it scales to higher levels of aggregate throughput and operations per second with a tradeoff of slightly higher latencies for file operations. For more information, please see the documentation on File System Performance.

Q. How much throughput can a file system support?

The throughput available to a file system scales as a file system grows. Because file-based workloads are typically spiky – requiring high levels of throughput for periods of time and lower levels of throughput the rest of the time – Amazon EFS is designed to burst to allow high throughput levels for periods of time. All file systems deliver a consistent baseline performance of 50 MB/s per TB of storage, all file systems (regardless of size) can burst to 100 MB/s, and file systems larger than 1TB can burst to 100 MB/s per TB of storage. As you add data to your file system, the maximum throughput available to the file system scales linearly and automatically with your storage.

File system throughput is shared across all Amazon EC2 instances connected to a file system. For example, a 1TB file system that can burst to 100MB/s of throughput can drive 100MB/s from a single Amazon EC2 instance, or 10 Amazon EC2 instances can collectively drive 100MB/s. For more information, please see the documentation on File System Performance.

Access Control and Security

Q. How do I control which Amazon EC2 instances can access my file system?

When you create a file system, you create endpoints in your VPC called “mount targets.” Each mount target provides an IP address and a DNS name, and you use this IP address or DNS name in your mount command. Only resources that can access a mount target can access your file system. You can control the network traffic to and from your file system mount targets using VPC security groups.

Q. How do I control who can access my file system?

You can control who can administer your file system usingAWS Identity and Access Management (IAM). You can control access to files and directories with POSIX-compliant user and group-level permissions.

Q. Does Amazon EFS provide encryption of data at rest?

Amazon EFS does not currently provide the option to encrypt data at rest, but we will offer this option soon.

Compatibility

Q. What interoperability and compatibility is there between existing AWS services and Amazon EFS?

Amazon EFS is integrated with a number of other AWS services, including Amazon CloudWatch, AWS CloudFormation, AWS CloudTrail, AWS IAM, and AWS Tagging services.

Amazon CloudWatch allows you to monitor file system activity using metrics. AWS CloudFormation allows you to create and manage file systems using templates.

AWS CloudTrail allows you to record all Amazon EFS API calls in log files.

AWS Identity and Access Management (IAM) allows you to control who can administer your file system. AWS Tagging services allows you to label your file systems with metadata that you define.

Q. What type of locking does Amazon EFS support?

Locking in Amazon EFS follows the NFSv4.1 protocol for advisory locking, and enables your applications to use both whole file and byte range locks.

Q. Are file system names global (like Amazon S3 bucket names)?

Every file system has an automatically generated ID number that is globally unique. You can tag your file system with a name, and these names do not need to be unique.

Pricing and Billing

Q. How much does Amazon EFS cost?

With Amazon EFS, you pay only for the amount of file system storage you use per month in GB. There is no minimum fee and no set-up costs. There are no additional costs for bandwidth or requests. For Amazon EFS pricing information, please visit the pricing section on the Amazon EFS Pricing page.

Q. Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Amazon Glacier FAQ

General

Q: What is Amazon Glacier?

Amazon Glacier is an extremely low-cost storage service that provides secure, durable, and flexible storage for data backup and archival. With Amazon Glacier, customers can reliably store their data for as little as $0.007 per gigabyte per month. Amazon Glacier enables customers to offload the administrative burdens of operating and scaling storage to AWS, so that they don’t have to worry about capacity planning, hardware provisioning, data replication, hardware failure detection and repair, or time-consuming hardware migrations.

Q: How can businesses, government and other organizations benefit from Amazon Glacier?

Amazon Glacier enables any business or organization to easily and cost effectively retain data for months, years, or decades. With Amazon Glacier, customers can now cost effectively retain more of their data for future analysis or reference, and they can focus on their business rather than operating and maintaining their storage infrastructure. Customers seeking compliance storage can deploy compliance controls using Vault Lock to meet regulatory and compliance archiving requirements.

Q: How should I choose between Amazon Glacier and Amazon Simple Storage Service (Amazon S3)?

Amazon S3 is a durable, secure, simple, and fast storage service designed to make web-scale computing easier for developers. Use Amazon S3 if you need low latency or frequent access to your data. Use Amazon Glacier if low storage cost is paramount, your data is rarely retrieved, and data retrieval times of several hours are acceptable.

Amazon S3 now provides a new storage option that enables you to utilize Amazon Glacier's extremely low-cost storage service for data archiving. You can define S3 lifeycycle rules to automatically archive sets of Amazon S3 objects to Amazon Glacier to reduce your storage costs. You can learn more by visiting the Object Lifecycle Management topic in the Amazon S3 Developer Guide.

Q: What kind of data can I store?

You can store virtually any kind of data in any format. You can also deploy compliance storage controls with Vault Lock to store regulatory and compliance archives in an immutable, Write Once Read Many (WORM) format. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: What does Amazon do with my data in Amazon Glacier?

Amazon will store your data and track its associated usage for billing purposes. Amazon will not otherwise access your data for any purpose outside of the Amazon Glacier offering, except if required to do so by law. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: How do I use Amazon Glacier?

Amazon Glacier provides a simple, standards-based REST web services interface as well as Java and .NET SDKs. The AWS Management console can be used to quickly set up Amazon Glacier. Data can then be uploaded and retrieved programmatically. View our documentation for more information on the Glacier APIs and SDKs.

Q: How durable is Amazon Glacier?

Amazon Glacier is designed to provide average annual durability of 99.999999999% for an archive. The service redundantly stores data in multiple facilities and on multiple devices within each facility. To increase durability, Amazon Glacier synchronously stores your data across multiple facilities before returning SUCCESS on uploading archives. Glacier performs regular, systematic data integrity checks and is built to be automatically self-healing.

Getting Started

Q: How is data within Amazon Glacier organized?

You store data in Amazon Glacier as an archive. Each archive is assigned a unique archive ID that can later be used to retrieve the data. An archive can represent a single file or you may choose to combine several files to be uploaded as a single archive. You upload archives into vaults. Vaults are collections of archives that you use to organize your data.

Q: How much data can I store?

There is no maximum limit to the total amount of data that can be stored in Amazon Glacier. Individual archives are limited to a maximum size of 40 terabytes.

Q: What is the minimum amount of data that I can store using Amazon Glacier?

There is no minimum limit to the amount of data that can be stored in Amazon Glacier and individual archives can be from 1 byte to 40 terabytes.

Q: Does the AWS Management Console support Amazon Glacier?

Yes. The AWS Management Console allows you to create and configure vaults, allowing you to easily and quickly setup Glacier. Click here to go the AWS Management Console.

Billing

Q: How much does Amazon Glacier cost?

With Amazon Glacier, storage is priced from $0.007 per gigabyte per month, and you pay for what you use. There are no setup fees, and for most archive use cases your total costs will primarily be made up of your storage cost.

Upload and Retrieve requests are priced from $0.05 per 1,000 requests. Forlarge retrievals, there is also retrieval fee starting at $0.01 per gigabyte. In addition, there is a pro-rated charge of $0.021 per gigabyte for items that are deleted prior to 90 days. As Amazon Glacier is designed to store data that is infrequently accessed and long lived, these charges will likely not apply to most of you.

We charge less where our costs are less. Some prices vary across Amazon Glacier Regions and are based on the location of your vault. There is no Data Transfer charge for data transferred between Amazon EC2 and Amazon Glacier within the same Region. Data transferred between Amazon EC2 and Amazon Glacier across all other Regions (e.g. between the Amazon EC2 Northern California and Amazon Glacier US East North Virginia Regions) will be charged at Internet Data Transfer rates on both sides of the transfer.

To learn more about Glacier pricing, please visit the Glacier pricing page.

Q: How is my storage charge calculated?

The volume of storage billed in a month is based on the average storage used throughout the month, measured in gigabyte-months (GB-Months). The size of each of your archives is calculated as the amount of data you upload plus an additional 32 kilobytes of data for indexing and metadata (e.g. your archive description). This extra data is necessary to identify and retrieve your archive. Here is an example of how to calculate your storage costs using US East (Northern Virginia) Region pricing:

If you upload 100,000 archives that are 1 gigabyte each, your total storage would be:

1.000032 gigabytes for each archive x 100,000 archives = 100,003.20 gigabytes

If you stored the archives for 1 month, you would be charged:

100,003.20 GB-Months x $0.007 = $700.02

If you upload 200,000 archives that are 0.5 gigabytes each, your total storage would be:

0.500032 gigabytes for each archive x 200,000 archives = 100,006.40 gigabytes

If you stored the archives for 1 month, you would be charged:

100,006.40 GB-Months x $0.007 = $700.04

Your storage is measured in “TimedStorage-ByteHrs,” which are added up at the end of the month to generate your monthly charges. For example, if you store an archive that is 1 gigabyte (inclusive of the 32 kilobyte overhead) for one day in the US East (Northern Virginia) Region, your storage usage would be:

1,073,741,824 bytes x 1 day x 24 hours = 25,769,803,776 Byte-Hours

Converting this to GB-Months (assuming a 30 day month) gives:

25,769,803,776 Byte-Hours x (1 GB / 1,073,741,824 bytes) x (1 month / 720 hours) = 0.03 GB-Months

So your storage charge for that day would be:

0.03 GB-Months x $0.007 = $0.00021

To learn more about Glacier pricing and view prices for other regions, please visit theGlacier pricing page.

Q: Why do prices vary depending on which Amazon Glacier Region I choose?

We charge less where our costs are less. For example, our costs are lower in the US East (North Virginia) Region than in the US West (Northern California) Region.

Q: How will I be charged and billed for my use of Amazon Glacier?

There are no setup fees to begin using the service. At the end of the month, your credit card will automatically be charged for that month's usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site, by logging into your Amazon Web Services account, and clicking "Account Activity" under "Your Web Services Account".

Q: How much data can I retrieve for free?

You can retrieve up to 5% of your data stored in Glacier for free each month. Typically this will be sufficient for backup and archival needs. Your 5% monthly free retrieval allowance is calculated and metered on a daily prorated basis. For example, if on a given day you have 12 terabytes of data stored in Glacier, you can retrieve up to 20.5 gigabytes of data for free that day (12 terabytes x 5% / 30 days = 20.5 gigabytes, assuming it is a 30 day month).

Your daily allowance is calculated based on the amount of data you have stored in Glacier, as per your vault inventories. See the Glacier developer guide for more details about vault inventories

Q: How will I be charged when retrieving large amounts of data from Amazon Glacier?

You can retrieve up to 5% of your average monthly storage, pro-rated daily, for free each month. For example, if on a given day you have 75 TB of data stored in Amazon Glacier, you can retrieve up to 128 GB of data for free that day (75 terabytes x 5% / 30 days = 128 GB, assuming it is a 30 day month). In this example, 128 GB is your daily free *retrieval allowance*. Each month, you are only charged a Retrieval Fee if you exceed your daily retrieval allowance. Let's now look at how this Retrieval Fee - which is based on your monthly *peak billable retrieval rate* - is calculated.

Let's assume you are storing 75 TB of data and you would like to retrieve 140 GB. The amount you pay is determined by how fast you retrieve the data. For example, you can request all the data at once and pay $21.60, or retrieve it evenly over eight hours, and pay $10.80. If you further spread your retrievals evenly over 28 hours, your retrievals would be free because you would be retrieving less than 128 GB per day. You can lower your billable retrieval rate and therefore reduce or eliminate your retrieval fees by spreading out your retrievals over longer periods of time.

Below we review how to calculate Retrieval Fees if you stored 75 TB and retrieved 140 GB in 4 hours, 8 hours, and 28 hours respectively.

First we calculate your *peak retrieval rate*. Your peak hourly retrieval rate each month is equal to the greatest amount of data you retrieve in any hour over the course of the month. If you initiate several retrieval jobs in the same hour, these are added together to determine your hourly retrieval rate. We always assume that a retrieval job completes in 4 hours for the purpose of calculating your peak retrieval rate. In this case your peak rate is 140 GB/4 hours, which equals 35 GB per hour.

Then we calculate your *peak billable retrieval rate* by subtracting the amount of data you get for free from your peak rate. To calculate your free data we look at your daily allowance and divide it by the number of hours in the day that you retrieved data. So in this case your free data is 128 GB /4 hours or 32 GB free per hour. This makes your billable retrieval rate 35 GB/hour – 32 GB per hour which equals 3 GB per hour.

To calculate how much you pay for the month we multiply your peak billable retrieval rate (3 GB per hour) by the retrieval fee ($0.01/GB) by the number of hours in a month (720). So in this instance you pay 3 GB/Hour * $0.01 * 720 hours, which equals $21.60 to retrieve 140 GB in 3-5 hours.

First we calculate your peak retrieval rate. Again, for the purpose of calculating your retrieval fee, we always assume retrievals complete in 4 hours. If you request 70GB of data at a time with an interval of at least 4 hours, your peak retrieval rate would then be 70GB / 4 hours = 17.50 GB per hour. (This assumes that your retrievals start and end in the same day).

Then we calculate your peak billable retrieval rate by subtracting the amount of data you get for free from your peak rate. To calculate your free data we look at your daily allowance and divide it by the number of hours in the day that you retrieved data. So in this case your free data is 128 GB /8 hours or 16 GB free per hour. This makes your billable retrieval rate 17.5 GB/hour – 16 GB per hour which equals 1.5 GB/hour. To calculate how much you pay for the month we multiply your peak hourly billable retrieval rate (1.5 GB/hour) by the retrieval fee ($0.01/GB) by the number of hours in a month (720). So in this instance you pay 1.5 GB/hour x $0.01 x 720 hours, which equals $10.80 to retrieve 40 GB.

If you spread your retrievals over 28 hours, you would no longer exceed your daily free retrieval allowance and would therefore not be charged a Retrieval Fee.

As you can see, you are able to significantly reduce, or eliminate, your retrieval fees when longer retrieval periods are suitable, as is often the case for archived data.

To learn more about Glacier pricing, please visit the Glacier pricing page.

Q: How will I be charged when retrieving only a range of an archive?

Range retrievals are priced in precisely the same way as regular retrievals from Amazon Glacier. The amount of data that you specify in your range retrieval requests are summed in order to determine whether your retrievals fall within your daily free retrieval tier. (See How much data can I retrieve for free to learn more). Range retrievals make it even easier for you to retrieve data without paying any retrieval fees. In the event that you do exceed your daily free retrieval tier, it is the range that you request that will determine your retrieval rate. (See How will I be charged when retrieving large amounts of data from Amazon Glacier? to learn more).

Q: How will I be charged for deleting data that is less than 3 months old?

Amazon Glacier is designed for use cases where data is retained for months, years, or decades. Deleting data from Amazon Glacier is free if the archive being deleted has been stored for three months or longer. If an archive is deleted within three months of being uploaded, you will be charged an early deletion fee. In the US East (Northern Virginia) Region, you would be charged a prorated early deletion fee of $0.021 per gigabyte deleted within three months. So if you deleted 1 gigabyte of data 1 month after uploading it, you would be charged a $0.014 early deletion fee. If, instead you deleted 1 gigabyte after 2 months, you would be charged a $0.007 early deletion fee.

To view prices for other regions, visit the Glacier pricing page.

Q: What can I expect the total cost of ownership (TCO) to be?

In a typical archive use case, data is retained for many years with the data often going months without being accessed. When data is retrieved, it is often a small subset of the total data stored. For example, let's assume you upload 1 petabyte of data to Glacier, and each archive is 10 megabytes. If you retain your data for three years, and retrieve up to 10TB a month, retrieving less than your free allowance each day (i.e. less than 3.3 terabytes a day), your monthly TCO over the 3 year period would be $ $7,541 or ~$0.0072 per gigabyte per month (including request charges).

To learn more about Glacier pricing, please visit the Glacier pricing page.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Security

Q: How do I control access to my data?

By default, only you can access your data. In addition, you can control access to your data in Amazon Glacier by using the AWS Identity and Access Management (AWS IAM) service. You simply set up an AWS IAM policy that specifies which users within an account have rights to operations on a given vault.

Q: Is my data encrypted?

Yes, all data in the service will be encrypted on the server side. Amazon Glacier handles key management and key protection for you. Amazon Glacier uses one of the strongest block ciphers available, 256-bit Advanced Encryption Standard (AES-256). 256-bit is the largest key size defined for AES. Customers wishing to manage their own keys can encrypt data prior to uploading it.

Q: Does Amazon Glacier support IAM permissions?

Yes, Glacier will support API-level permissions through AWS Identity and Access Management (IAM) service integration

For more information about IAM, go to:

- AWS Identity and Access Management
- AWS Identity and Access Management Getting Started Guide
- Using AWS Identity and Access Management

Archives and Vaults

Q: What is an archive?

An archive is a durably stored block of information. You store your data in Amazon Glacier as archives. You may upload a single file as an archive, but your costs will be lower if you aggregate your data. TAR and ZIP are common formats that customers use to aggregate multiple files into a single file before uploading to Amazon Glacier. The total volume of data and number of archives you can store are unlimited. Individual Amazon Glacier archives can range in size from 1 byte to 40 terabytes. The largest archive that can be uploaded in a single Upload request is 4 gigabytes. For items larger than 100 megabytes, customers should consider using the Multipart upload capability. Archives stored in Amazon Glacier are immutable, i.e. archives can be uploaded and deleted but cannot be edited or overwritten.

Q: How do I delete archives?

You can delete an archive at any time. You will stop being billed for your archive when your delete request succeeds at which point the archive itself will be inaccessible. Archives that are deleted within 3 months of being uploaded will be charged a deletion fee (see billing section for more details).

Q: How do I upload large archives?

When uploading large archives (100MB or larger), you can use multi-part upload to achieve higher throughput and reliability. Multi-part uploads allow you to break your large archive into smaller chunks that are uploaded individually. Once all the constituent parts are successfully uploaded, they are combined into a single archive.

Q: What is a vault?

A vault is a way to group archives together in Amazon Glacier. You organize your data in Amazon Glacier using vaults. Each archive is stored in a vault of your choice. You may control access to your data by setting vault-level access policies using the AWS Identity and Access Management (IAM) service. You can also attach notification policies to your vaults. These enable you or your application to be notified when data that you have requested for retrieval is ready for download. Click here to learn more about setting up notifications using the Amazon Simple Notification Service (Amazon SNS).

Q: How many vaults can I create?

You can create up to 1,000 vaults per account per region.

Q: How do I effectively manage my Amazon Glacier vaults?

Amazon Glacier allows you to tag your Glacier vaults for easier resource and cost management. Tags are labels that you can define and associate with your vaults, and using tags adds filtering capabilities to operations such as AWS cost reports. For example, you can use tags to allocate Glacier costs and usage across multiple departments in your organization or by any other categorization. You can tag your vaults by using the Glacier Console or the Glacier APIs. For more information see Tagging Your Amazon Glacier Vaults.

Q: How do I delete a vault?

You may delete any Glacier vault that does not contain any archives using the AWS Management Console, the Amazon Glacier APIs or the SDKs. Once a vault has been deleted, you can then re-create a vault with the same name. If your vault contains archives, you must delete all the archives before deleting the vault.

Vault Access Policies

Q: What is a vault access policy?

A vault access policy is a resource-based policy that you can attach directly to your Glacier vault (the resource) to specify who has access to the vault and what actions they can perform on it. To learn more please read Managing Vault Access Policies in the Amazon Glacier developer’s guide.

Q: How are vault access policies different from access control based on AWS Identity and

Access Management (IAM) policies?

Access permissions can be assigned in two ways: asuser-based permissions or as resource-based permissions. Access control based on IAM policies is user-based where you would assign IAM policies to IAM users or groups to control the read, write, and delete permissions on your Glacier vaults. Access control with vault access policies is resource-based where you would attach an access policy directly on a vault to govern access to all users. Vault access policies can make certain use cases simpler. For example, to protect information in a business-critical vault from unintended deletion, you can create a vault access policy that denies delete attempts from all users. This data protection procedure can be accomplished in a matter of minutes in the AWS Management Console without having to audit and revoke delete permissions assigned to users through IAM policies.

Q: Can I use vault access policies to manage cross-account access?

Yes you can. For example, you can grant read-only access on your vault to a business partner in a different AWS account by simply adding that account to the vault’s access policy and specifying that only read activities are allowed.

Q: How does billing work in a cross-account access scenario?

The vault owner’s account will be billed for the charges incurred during cross-account access. For example, Alice (account A) grants Bob (account B) access to Alice’s “movies” vault and allows Bob to upload data. After Bob makes 1000 requests to upload 1GB of data, Alice’s account (account A) will be billed for the 1000 requests as well as the 1GB of data until the data is deleted. Bob’s account (account B) will not incur these charges.

Q: How do I create and manage vault access policies?

You can create and manage vault access policies in the AWS Glacier console or use the vault access APIs in the AWS SDK. To learn more please read Managing Vault Access Policies in the Amazon Glacier developer’s guide.

Q: How many vault access policies can I have?

You can set one vault access policy for each vault. The vault access policy can be used as a single location to view the list of users with vault access and the allowed actions for each user.

Vault Lock

Q: What is Vault Lock?

Vault Lock allows you to easily deploy and enforce compliance controls on individual Glacier vaults via a lockable policy (Vault Lock policy). Once locked, the Vault Lock policy becomes immutable and Glacier will enforce the prescribed controls to help achieve your compliance objectives. To learn more, please read Amazon Glacier Vault Lock in the Amazon Glacier developer’s guide.

Q: What type of compliance controls can I deploy with Vault Lock?

You can deploy a variety of compliance controls in a Vault Lock policy using the AWS Identity and Access Management (IAM) policy language. For example, you can easily set up “Write Once Read Many” (WORM) or time-based records retention for regulatory archives. To learn more, please read Amazon Glacier Vault Lock in the Amazon Glacier developer’s guide.

Q: How does Vault Lock enforce my compliance controls?

Vault Lock enforces your compliance controls via a lockable policy (Vault Lock policy). Once locked, the Vault Lock policy becomes immutable and Glacier will only allow operations on your data that are explicitly permitted by the compliance controls you specified. Vault Lock also ensures that a locked policy cannot be deleted or altered until there are no more archives to protect in the vault. Learn more about Locking a Vault for compliance in the Amazon Glacier developer’s guide.

Q: How is a Vault Lock policy different than a vault access policy?

Both policies govern access controls to your vault, however, a Vault Lock policy can be made immutable and provides strong enforcement for your compliance controls. You can use the Vault Lock policy to deploy regulatory and compliance controls that are typically restrictive and are “set and forget” in nature. In conjunction, you can use the vault access policy to implement access controls that are not compliance related, temporary, and subject to frequent modification. The two policies can be used in tandem to achieve governance and flexibility.

Q: What AWS electronic storage services have been assessed based on financial services regulations?

For customers in the financial services industry, Vault Lock provides added support for broker-dealers who must retain records in a non-erasable and non-rewritable format to satisfy regulatory requirements of SEC Rule 17a-4(f), FINRA Rule 4511, or CFTC Regulation 1.31. You can easily designate the records retention time frame to retain regulatory archives in the original form for the required duration, and also place legal holds to retain data indefinitely until the hold is removed.

Q: What AWS documentation supports the SEC 17a-4(f)(2)(i) and CFTC 1.31(c) requirement for notifying my regulator?

Provide notification to your regulator or “Designated Examining Authority (DEA)” of your choice to use AWS Glacier for electronic storage along with a copy of the Cohasset Assessment. For the purposes of these requirements, AWS is not a designated third party (D3P). Be sure to select a D3P and include this information in your notification to your DEA.

Q: What other controls can be applied with Amazon Glacier Vault Lock?

In certain situations, you may be faced with the need to place a legal hold on your compliance archives for an indefinite period of time. A legal hold can be initiated on a Glacier Vault by creating a vault access policy that denies the use of Glacier's Delete functions if the vault is tagged in a particular way. In addition to time-based retention and legal hold, Glacier Vault Lock can be used to implement a variety of compliance controls which can be made immutable for strong governance, such as enforcing Multifactor Authentication on all data access/read activities to a vault with classified information.

Q: How do I set up Vault Lock?

You can set up Vault Lock in the AWS Glacier console or use the Vault Lock APIs in the AWS SDK. To learn more, please read Getting Started with Amazon Glacier Vault Lock in the Amazon Glacier developer's guide.

Data Retrievals

Q: How can I retrieve data from the service?

You can download data directly from the service using the service's REST API. When you make a request to retrieve data from Glacier, you initiate a retrieval job. Once the retrieval job completes, your data will be available to download for 24 hours. Retrieval jobs typically complete within 3-5 hours.

Q: What operations initiate jobs and why?

To retrieve an archive or a vault inventory, you first initiate a job (Click here for more information about vault inventories). Once you initiate a job, you can call the DescribeJob API to monitor its progress. You can also have notifications automatically sent to you once a job completes. Jobs will typically complete in 3-5 hours. Once a job completes successfully, you can download the data requested or access it using Amazon Elastic Compute Cloud (Amazon EC2).

Q: How long does it take for jobs to complete?

Most jobs will take between 3 to 5 hours to complete.

Q: Can I retrieve part of an archive?

Yes, range retrievals enable you to retrieve a specific range of an archive. Range retrievals are similar to regular retrievals in Amazon Glacier. Both require the initiation of a retrieval job that typically completes within 3-5 hours (See How can I retrieve data? for more information). You can use range retrievals to reduce or eliminate your retrieval fees (See How much data can I retrieve for free?)

When initiating a retrieval job using range retrievals, you provide a byte range that can start at zero (which would be the beginning of your archive), or at any 1MB interval thereafter (e.g. 1MB, 2MB, 3MB, etc). The end of the range can either be the end of your archive or any 1MB interval greater than the beginning of your range.

Q: Why would I retrieve only a range of an archive?

There are several reasons why you might choose to perform a range retrieval. For example, you may have aggregated several files and uploaded them as a single archive. You may then need to retrieve a small selection of those files, in which case you could retrieve only the ranges of the archive that contained the required files. Another reason you could choose to perform a range retrieval is to manage how much data you download from Amazon Glacier in a given period. When data is retrieved from Amazon Glacier, a retrieval job is first initiated, which will typically complete in 3-5 hours. The data retrieved is then available for download for 24 hours. You could therefore retrieve an archive in parts in order to manage the schedule of your downloads. You may also choose to perform range retrievals in order to reduce or eliminate your retrieval fees. If you exceed your free retrieval allowance, you pay a retrieval fee that is based on your peak retrieval rate. Spreading out a retrieval of an archive in smaller parts could therefore allow you reduce your retrieval fees, by reducing your peak retrieval rate. (Click here to learn more about what it costs to retrieve data from Amazon Glacier).

Q: How do I view my jobs?

You can list your ongoing jobs for any of your vaults by calling the ListJobs API. The list of jobs provides information including the job's creation time and date and the job's status (e.g. in-progress, completed successfully, or not in which case reasons for the job not succeeding are provided). The progress of a single job can be tracked by calling the DescribeJob API and providing the corresponding job ID. The status of the job will be returned immediately.

Q: Can I be notified when a job is completed?

Yes. You can optionally configure vaults to send notifications to you or your application when jobs complete. Notifications will be delivered via the Amazon Simple Notification Service (Click here to learn more about Amazon SNS).

Data Retrieval Policies

Q: What are data retrieval policies?

Amazon Glacier data retrieval policies let you define your own data retrieval limits with a few clicks in the AWS console. You can limit retrievals to "Free Tier Only", or if you wish to retrieve more than the free tier, you can specify a "Max Retrieval Rate" to limit your retrieval speed and establish a retrieval cost ceiling. In both cases, Amazon Glacier will not accept retrieval requests that would exceed the retrieval limits you defined. To learn more please read Configuring Data

Retrieval Policies in the Amazon Glacier developer's guide.

Q: How do I set up data retrieval policies?

You can set up data retrieval policies in the Amazon Glacier console or via the Amazon Glacier APIs. To learn more please read Configuring Data Retrieval Policies in the Amazon Glacier developer's guide.

Q: Are data retrieval policies specific to each AWS region?

Yes. You can set one data retrieval policy for each AWS region which will govern all data retrieval activities in the region under your account. Data retrieval policies are region-specific because data retrieval costs vary across AWS regions and the 5% free retrieval tier is also computed based on your storage within the region. Please visit Amazon Glacier Pricing for more information.

Q: Can I use data retrieval policies to "slow down" my retrievals or spread them out?

No, data retrieval policies such as "Free Tier Only" and "Max Retrieval Rate" will not accept a data retrieval request which would exceed your predefined data retrieval limit to help you manage data retrieval cost. Data retrieval policies will not change the 3 to 5 hour data retrieval latency or spread out your retrievals. You can leverage Amazon Glacier's range retrieval feature to spread out retrievals and lower the peak retrieval speed. Learn more.

Q: How is my storage charge calculated if I set a "Max Retrieval Rate" and my retrievals exceed the free tier?

Let's suppose you have 10 GB offree retrieval allowance per day and you set a "Max Retrieval Rate" of 20 GB/hr which shows a data retrieval cost estimate of "$144.00/month or less" in the AWS console (assuming US East region and a 30 day month).

Now let's walk through a few scenarios, assuming this is a new month.

On day 1, you issued a retrieval request for an 8 GB archive. Since 8 GB was less than the free retrieval allowance for the day, your retrieval request was accepted and the data retrieval was free.

On day 2, your coworker accidentally issued a retrieval request for a 100 GB archive by mistake. Because the retrieval rate (based on 4 hour completion) would be 100 GB/4 hours = 25 GB/hr, exceeded the 20 GB/hr "Max Retrieval Rate", the request was rejected and there was no data retrieval charges incurred.

On day 3, you issued a retrieval request for a single 40 GB archive. Since all data retrieval billing assumes the retrievals complete in 4 hours, 40 GB/ 4 hours yielded a retrieval rate of 10 GB/hr which was below the 20 GB/hr "Max Retrieval Rate" you set, so your retrieval request was accepted. Your peak billable retrieval rate for the day was (40 GB – 10 GB free tier) divided by 4 hours which yielded 7.5 GB/hr. Your estimated data retrieval bill at this point would be 7.5 GB/hr

* $0.01 per GB* 720 hours per month = $54 for the month and was below the data retrieval cost estimate of $144.00/month shown in the console.

On day 4, you issued a retrieval request for a 40 GB archive immediately followed by another request for a 44 GB archive. The request for the 40 GB archive was accepted because the retrieval rate (based on 4 hour completion) was 40 GB/4 hours = 10 GB/hr, which was less than the 20 GB/hr “Max Retrieval Rate”. The second request to retrieve a 44 GB archive however, was rejected because while the request alone only yielded 44 GB/4 hours = 11 GB/hr of retrieval rate, the first retrieval request was still in progress. If the second request was accepted, then the combined peak retrieval rate would have been 10 GB/hr + 11 GB/hr = 21 GB/hr and would exceed the 20 GB/hr Max Retrieval Rate you specified. You decided to wait till the next day to retrieve the 44 GB archive.

On day 5, you learned that the 44 GB archive, along with another 36 GB archive both needed to be available as soon as possible for a customer request. This meant that you needed to retrieve both archives at the same time, equivalent to issuing an 80 GB retrieval request that would yield a 20 GB/hr retrieval rate, which was equal to the “Max Retrieval Rate” you set. You issued both requests and they were both accepted. Your billable peak retrieval rate was (80 GB – 10 GB free tier) / 4 hours = 17.5 GB/hr and your estimated data retrieval cost was 17.5 GB/hr * $0.01 per GB* 720 hours per month = $126.00/month, less than the $144.00/month estimate shown in the AWS console based on a 20 GB/hr Max Retrieval Rate. This new estimate overrides the cost estimate of $54/month on day 3. If you incurred no additional data retrieval cost for the rest of the month, your data retrieval cost for the month would be $126.00, again less than the $144.00/month estimate shown in the console.

Data Inventories

Q: Can I see what archives I have stored in Amazon Glacier?

Yes. Although you will need to maintain your own index of data you upload to Amazon Glacier, an inventory of all archives in each of your vaults is maintained for disaster recovery or occasional reconciliation purposes. The vault inventory is updated approximately once a day. You can request a vault inventory as either a JSON or CSV file and will contain details about the archives within your vault including the size, creation date and the archive description (if you provided one during upload). The inventory will represent the state of the vault at the time of the most recent inventory update.

Q: Can I obtain a real time list of my vaults?

Yes, you can list your vaults stored in Amazon Glacier using either the AWS Management Console or by calling the ListVaults API. As well as a list of vault names, you will also be able to see when the vault’s inventory was last updated and a summary of the vault’s contents at that time, as well as the vault’s creation date and creator.

AWS Import/Export Snowball FAQ

General

Q. What is AWS Snowball?

AWS Snowball is a data transport solution that accelerates moving terabytes to petabytes of data into and out of AWS using storage appliances designed to be secure for physical transport. Using Snowball helps to eliminate challenges that can be encountered with large-scale data transfers including high network costs, long transfer times, and security concerns.

Q. How does Snowball work?

AWS Snowball uses secure appliances and the Snowball client to accelerate petabyte-scale data transfers into and out of AWS. You start by using the AWS Management Console to create one or more jobs to request one or multiple Snowball appliances (depending on how much data you need to transfer), and download and install the Snowball client. Once the appliance arrives, connect it to your local network, set the IP address either manually or with DHCP, and use the client to identify the directories you want to copy. The client will automatically encrypt and copy the data to the appliance and notify you when the transfer job is complete. When the transfer is complete and the appliance is ready to be returned, the E Ink shipping label will automatically update to indicate the correct AWS facility to ship to, and you can track the job status by using Amazon Simple Notification Service (Amazon SNS), text messages, or directly in the console.

Q. Who should use Snowball?

Snowball is the right data transfer choice if you need to securely and quickly transfer terabytes to many petabytes of data to AWS. Snowball can also be the right choice if you don’t want to make expensive upgrades to your network infrastructure, if you frequently experience large backlogs of data, if you're located in a physically isolated environment, or if you're in an area where high-bandwidth Internet connections are not available or cost-prohibitive.

Q. How much data can I transfer using Snowball?

You can transfer virtually any amount of data with Snowball, from a few terabytes to many petabytes. You can typically transfer multiple TB of data to each Snowball appliance. You can transfer larger data sets by using multiple Snowballs, either in parallel, or one after another. For example, you can transfer 100 TB of data using two Snowballs in parallel, or you can transfer the data using two Snowballs one after another.

Q. What is the Snowball client?

The Snowball client is software that you install on a local host computer and use to efficiently identify, compress, encrypt, and transfer data from the directories you specify to a Snowball.

Q. How long does it take to transfer my data?

You can use the Snowball client to estimate the time it takes to transfer your data (refer to the user guide for more details). Data transfer speed is affected by a number of factors including local network speed, file size, and the speed at which data can be read from your local servers.

The Snowball client will copy data to the Snowball as fast as conditions allow, (as little as a day to copy 48TB of data, depending on your local environment). End-to-end time to transfer the data into AWS is approximately a week, including the usual shipping and handling time in AWS data centers. You can copy twice that much data in the same amount of time by using two 48TB Snowballs in parallel, or you could copy 80TB of data in two and a half days on a single 80TB Snowball, which would increase your end-to-end time to about a week and a half.

Q. What are the specifications on the Snowball appliance?

Check this Snowball documentation page for the complete list of hardware specs, including interfaces, thermal and power requirements, decibel output, and dimensions.

Q. How long can I have a Snowball for a specific job?

For security purposes, data transfers must be completed within 90 days of a Snowball's preparation.

Q. What network interfaces does Snowball support?

Snowball has 10Gbps network interfaces with RJ45, SFP+ copper, and SFP+ optical network ports.

Q. What is Snowball’s default shipping option? Can I choose expedited shipping?

As a default, Snowball uses two-day shipping by UPS. You can choose expedited shipping if your jobs are time-sensitive.

Snowball Regional Availability

Q. In what regions is Snowball available?

Check the Regional Service Availability pages for the latest information.

Please note that 50TB models are only available in USA regions.

We do not ship Snowballs to several states in the USA. We regret that we cannot provide the service at this time in all states.

Q. May I ship a Snowball between AWS regions?

No. Snowballs are designed to be requested and used within a single AWS region. It may not be requested from one region and returned to another, or used to transfer data out of one region and directly into another. Snowball devices used for imports or exports from an AWS region in the EU may be used with any of the 28 EU countries.

When to Use Snowball

Q. When should I consider using Snowball instead of the Internet?

Snowball is a strong choice for data transfer if you need to securely and quickly transfer terabytes to many petabytes of data to AWS. Snowball can also be the right choice if you don't want to make expensive upgrades to your network infrastructure, if you frequently experience large backlogs of data, if you're located in a physically isolated environment, or if you're in an area where high-speed Internet connections are not available or cost prohibitive.

As a rule of thumb, if it takes more than one week to upload your data to AWS using the spare capacity of your existing Internet connection, then you should consider using Snowball. For example, following the guidelines in the table below, if you have a 100 Mb connection that you can solely dedicate to transferring your data and need to transfer 100 TB of data, it takes more than 100 days to complete data transfer over that connection. You can make the same transfer by using multiple Snowballs in about a week.

Available Internet Connection	Theoretical Min. Number of Days to Transfer 100TB at 80% Network Utilization	When to Consider AWS Snowball?
T3 (44.736Mbps)	269 days	2TB or more
100Mbps	120 days	5TB or more
1000Mbps	12 days	60TB or more

Q. When should I consider using Snowball instead of AWS Direct Connect?

AWS Direct Connect provides you with dedicated, fast connections from your premises to the AWS network. If you need to transfer large quantities of data to AWS on an ongoing basis, AWS Direct Connect might be the right choice.

Snowball can be a strong alternative to Direct Connect if you need to transfer data in large batches or as a one-time transfer, potentially from distributed locations. For these workloads, Snowball can be a simpler, more cost-effective option than setting up a new Direct Connect connection to transfer your data and then terminating the connection upon completion.

Q. When should I consider using Snowball instead of AWS Import/Export Disk?

Snowball provides a faster, simpler, and more cost-effective experience for most use cases when compared to AWS Import/Export Disk.

With Snowball, you don't need to purchase any hardware or write any code to transfer your data. Each Snowball appliance can transfer as much as 80 TB of data and you can use multiple appliances in parallel for larger workloads. Snowball uses tamper-resistant enclosures, 256-bit encryption, and an industry-standard Trusted Platform Module (TPM) that is designed to ensure both security and full chain of custody for your data, and also to reduce management overhead involved with transferring data into or out of AWS.

You can create transfer jobs right from the AWS Management Console. When your transfer is complete and the appliance is ready to be returned, the E Ink shipping label will automatically update to indicate the correct AWS facility to ship to, and you can track the job status by using Amazon SNS, or text messages, or directly in the console.

Security

Q. Does Snowball encrypt my data?

Snowball encrypts all data with 256-bit encryption. You manage your encryption keys by using the AWS Key Management Service (AWS KMS). Your keys are never sent to or stored on the appliance.

Q. Does AWS have a way to tell if the device was tampered with during transit?

In addition to using a tamper-resistant enclosure, Snowball uses an industry-standard Trusted Platform Module (TPM) with a dedicated processor designed to detect any unauthorized modifications to the hardware, firmware, or software. AWS inspects every appliance for any signs of tampering and to verify that no changes were detected by the TPM.

Q. What happens to the data on the appliance when it has been successfully transferred to AWS?

When the data transfer job has been processed and verified, AWS performs a software erasure of the Snowball appliance that follows the National Institute of Standards and Technology (NIST) guidelines for media sanitization.

Q. Is there a way to easily track my data transfer jobs?

Snowball uses an innovative, E Ink shipping label designed to ensure the appliance is automatically sent to the correct AWS facility and which also helps in tracking. When you have completed your data transfer job, you can track it by using Amazon SNS, text messages, and the console.

Q. Can I use AWS Snowball for data with Protected Health Information (PHI)?

Yes. AWS Snowball is a HIPAA-eligible service. If you currently have a Business Associate Agreement (BAA) with AWS, you can begin using Snowball immediately to transfer data into your HIPAA accounts.

Using Snowball to Import Data

Q: How do I get started with Snowball?

To get started with Snowball, visit theGetting Started page.

Q: How do I transfer my data to the Snowball appliance?

When you connect the Snowball appliance to your network and set the IP address using the E Ink display, you'll need to download three things from the AWS Management Console:

1. **Snowball client**: The software tool that is used to transfer data from your on-premises storage to the Snowball appliance. For more information on the Snowball client, see the Tools page.

2. **Job manifest file:** An encrypted metadata file that is used to uniquely identify your data transfer job.

3. **Job manifest unlock code:** A 25-character code to unlock the job manifest file.

When you have downloaded these files, you launch the Snowball client and provide the Snowball appliance's IP address, the manifest file path, and the unlock code. A sample Start command is below:

snowball start -i {Snowball IP} -m path/to/the/job/manifest} -u {unlock code}

After you launch the client and provide this information, the cilent is now connected to the Snowball appliance and is ready for use. Next you'll need to identify the file directories you want to transfer to the appliance and then wait for the transfer to complete. A sample Copy command is below:

snowball cp /path/to/data/on/source/storage/device/directories Snowball/bucketname

Q: What do I do when the data has been transferred to the Snowball appliance?

When the data transfer job is complete, the Snowball appliance's E Ink display automatically updates the return shipping label to indicate the correct AWS facility to ship to. Just drop off the Snowball appliance at the nearest UPS shipping facility and you're all set. You can track the status of your transfer job by using Amazon SNS, or text messages, or directly in the AWS Management Console.

Q: Can I import data from a Hadoop Distributed File System to Snowball?

Yes. You can copy data from a HDFS cluster to Snowball using the Snowball CLI. To learn more, please refer to the Snowball documentation.

Using Snowball to Export Data

Q: What is the Snowball export feature?

Export is a feature of Snowball that enables customers to export terabytes to petabytes of data from Amazon Simple Storage Service (Amazon S3) to on-premises storage.

Q: How do I get my data from AWS with export?

To use Snowball Export simply sign in to the AWS Management Console, choose Snowball, and create an export job. As with an import job, you specify the region and buckets that you want to use. If you don't want to export all of the data from a particular bucket, you can specify a beginning and ending S3 key range sorted in UTF-8 binary order to indicate what data should be exported. The key range that you choose, and all keys between them, are exported. Details on using the console can be found here.

Q: How quickly can I access my exported data?

We typically start exporting your data within 24 hours of receiving your request, and exporting data can take as long as a week. Once the job is complete and the device is ready, we ship it to you using the shipping options selected when you created the job.

Q: Can I pick up the Snowball from your data center so I don't have to wait for shipping?

No. Although you can select one-day shipping, we do have to ship the Snowball to an address that you provide. We don't have a way for you to pick up a Snowball from our data center.

Q: Can I track the export data-writing progress while you prepare my Snowball?

Yes. You can see when we start provisioning a Snowball and get real-time updates as data is written to the device. As with import jobs, you can get notification when the provisioning is complete and when the device has been shipped.

Q: Will AWS encrypt my data before copying it to the Snowball?

Yes. All data that is written is encrypted and the encryption keys for that data are never present on the Snowball.

Q: How do I read my data from the Snowball when I receive the device?

Using the Snowball client, you can copy your data from the Snowball to local storage. The client decrypts your data when it reads it from the Snowball and writes the data to your local storage in the same format as the data was stored in Amazon S3.

Q: How much data can I export?

There is almost no limit the amount of data you can export. If you want to export more data than can fit on one appliance, additional export jobs will be created automatically for you so that all of the data you select can be exported.

Q: Can I retrieve data from more than one bucket?

Yes. You can select as many buckets as you want for export.

Q: How are my Amazon S3 objects mapped to files when I copy them to my local storage?

Each key is copied to your device in a directory tree that starts with the bucket’s name. For example, if the key is “images/orange.jpg” and the bucket is "fruit” then the object is saved to /fruit/images/orange.jpg. Meta data associated with each object is not copied to your storage device.

Q: Can I export data that is in the Amazon Glacier storage class?

No. Before Amazon Glacier data can be exported it needs to be restored to Amazon S3 using the S3 Lifecycle Restore feature.

Q: Do I get a log of what was exported?

Yes. For each job, import or export, a log of the files that were copied and those that could not be copied is generated and available from the Snowball console.

Q: What does it cost to export my data?

In addition to the Snowball Export fees detailed on ourpricing page, you will also be charged all Amazon S3 and Amazon Glacier fees incurred to retrieve your data from those services.

Billing

Q. How much does it cost to transfer data using Snowball?

Each Snowball data transfer job costs a flat fee for device handling and import and export operations at AWS data centers. Snowball is free for use for 10 days at your site. The day that the device is received and the day that the device is shipped are not counted towards these 10 days. Beyond that, a Snowball device costs $15/day for each extra day that it is at your site. There is no cost for transferring data into AWS. Transferring data out of AWS costs $0.03/GB.

The following example illustrates Snowball pricing for an 80 TB model.

Example:

Assume you transfer 60 TB of data into AWS using one Snowball and you keep the Snowball for 14 days (receiving the Snowball from the shipper on day 1 and returning the Snowball to the shipper on day 14).

Service charge for this job:

The service charge for this job is $250.

Extra day charge:

Snowball is free for use for 10 days at your site. The day that the device is received and the day that the device is shipped are not counted toward these 10 days, meaning day 1 and day 14 are free in this case. There are 12 days between day 1 and day 14, and 10 out of 12 days are free. The remaining 2 days are 2 extra days used to transfer your data. The total extra day charge is:

2 days x $15/day = $30

Data transfer:

In this example, you transferred data into AWS, so the data transfer cost is free.

Shipping:

Shipping charges are based on your shipment destination and the shipping option you choose (for example, overnight, or two-day).

Q. How am I charged for Amazon S3 usage?

Snowball will transfer data on your behalf from Snowball appliances to AWS services, such as Amazon S3. Standard AWS service charges apply. In the case of transferring data into S3, standard S3 request and storage charges apply.

Q. Can I purchase a Snowball appliance?

Snowballs are only available on a per-job pay-as-you-go basis, and are not available for purchase.

Workflow Integration Tools

Q. Does the Snowball service support API access?

Yes. The Snowball Job Management API provides programmatic access to the job creation and management features of a Snowball. It is a simple, standards-based REST web service interface, designed to work with any Internet development environment.

Q. What can I do with the Snowball Job Management API?

The API allows partners and customers to build custom integrations to manage the process of requesting Snowballs and communicating job status. The API provides a simple web service interface that you can use to create, list, update, and cancel jobs from anywhere on the web. Using this web service, developers can easily build applications that manage Snowball job workflow. To learn more, please refer to Snowball documentation.

Q. What is the S3 Adapter?

The S3 Adapter provides an S3-compatible interface to the Snowball client for reading and writing data on a Snowball.

Q. What can I do with the S3 Adapter?

The S3 Adapter provides functions to communicate with Snowball, allowing customers to build tools to copy data from file and non-file sources. It includes interfaces to copy data to Snowball with the same encryption that is available through our Snowball command line tool. To learn more, please refer to Snowball documentation.

Q. Why would I use the S3 Adapter rather than the Snowball Client?

The Snowball Client is a turnkey tool that makes it easy to copy file based data to Snowball. Customers who prefer a tighter integration can use the S3 Adapter to easily extend their existing applicaitons and workflows to seamlessly integrate with Snowball.

Q. How is my data secured when I use the S3 Adapter?

The S3 Adapter writes data using the same advanced encryption mechanism that the Snowball Client provides.

Q. Which programming languages does the Snowball S3 Adapter support?

The S3 Adapter communicates over REST which is language-agnostic.

AWS Storage Gateway FAQ

General

Q: What is the AWS Storage Gateway?
The 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 store data to the AWS cloud for scalable and cost-effective storage. The AWS Storage Gateway supports industry-standard storage protocols that work with your existing applications. It provides low-latency performance by maintaining frequently accessed data on-premises while encrypting and storing all of your data in Amazon Simple Storage Service (Amazon S3) or Amazon Glacier.

The AWS Storage Gateway supports three configurations: Gateway-Cached Volumes, Gateway-Stored Volumes, and Gateway-Virtual Tape Library (VTL).

Gateway-Cached Volumes: You can durably and inexpensively store your primary data in Amazon S3, and retain your frequently accessed data locally. Gateway-Cached Volumes provide substantial cost savings on primary storage, minimize the need to scale your storage on-premises, and provide low-latency access to your frequently accessed data. In addition to storing your primary data in Amazon S3 using Gateway-Cached Volumes, you can also take point-in-time snapshots of your Gateway-Cached volume data in Amazon S3, enabling you to make space-efficient versioned copies of your volumes for data protection and various data reuse needs.

Gateway-Stored Volumes: In the event you need low-latency access to your entire data set, you can configure your gateway to store your primary data locally, and asynchronously back up point-in-time snapshots of this data to Amazon S3. Gateway-Stored volumes provide durable and inexpensive off-site backups that you can recover locally or from Amazon EC2 if, for example, you need replacement capacity for disaster recovery.

*Gateway-Virtual Tape Library (VTL):*With Gateway-VTL you can have a limitless collection of virtual tapes. Each virtual tape can be stored in a Virtual Tape Library backed by Amazon S3 or a Virtual Tape Shelf backed by Amazon Glacier. The Virtual Tape Library exposes an industry standard iSCSI interface which provides your backup application with on-line access to the virtual tapes. When you no longer require immediate or frequent access to data contained on a virtual tape, you can use your backup application to move it from its Virtual Tape Library to your Virtual Tape Shelf in order to further reduce your storage costs.

Q: How does the AWS Storage Gateway work?

The AWS Storage Gateway's software appliance is available for download as a virtual machine (VM) image that you install on a host in your datacenter. Once you've installed your gateway and associated it with your AWS Account through our activation process, you can use the AWS Management Console to create either Gateway-Cached or Gateway-Stored storage volumes or Gateway-VTL virtual tape libraries that can be mounted as iSCSI devices by your on-premises applications.

Gateway-Cached volumes allow you to utilize Amazon S3 for your primary data, while retaining some portion of it locally in a cache for frequently accessed data. These volumes minimize the need to scale your on-premises storage infrastructure, while still providing your applications with low-latency access to frequently accessed data. You can create storage volumes up to 32 TB in size and mount them as iSCSI devices from your on-premises application servers. Data written to these volumes is stored in Amazon S3, with only a cache of recently written and recently read data stored locally on your on-premises storage hardware. You can also take point-in-time snapshots of your Gateway-Cached volume data in Amazon S3 in the form of Amazon EBS snapshots, enabling you to make space-efficient versioned copies of your volumes for data protection and various data reuse needs.

Gateway-Stored volumes store your primary data locally, while asynchronously backing up that data to AWS. These volumes provide your on-premises applications with low-latency access to their entire data sets, while providing durable, off-site backups. You can create storage volumes up to 16 TB in size and mount them as iSCSI devices from your on-premises application servers. Data written to your Gateway-Stored volumes is stored on your on-premises storage hardware, and asynchronously backed up to Amazon S3 in the form of Amazon EBS snapshots.

Gateway-Virtual Tape Library (VTL) enables you to seamlessly replace your physical tape infrastructure with a virtual tape infrastructure. Each Gateway-VTL presents your backup application with an industry-standard iSCSI-based Virtual Tape Library (VTL) consisting of a virtual media changer and tape drives. You can create virtual tapes in your Virtual Tape Library using the AWS Management Console. Each Virtual Tape Library can hold up to 1,500 virtual tapes with a maximum aggregate capacity of 1 PB. Virtual tapes are discovered by your backup application using its standard media inventory procedure. Virtual tapes in your Virtual Tape Library are available for immediate access and are backed by Amazon S3. Your backup application can read data from or write data to virtual tapes by mounting them to virtual tape drives using the virtual media changer.

For cost-effective long term retention of data requiring infrequent access, you can use your backup application to move virtual tapes from one or more of your Virtual Tape Libraries to your Virtual Tape Shelf (VTS) that is backed by Amazon Glacier. Your Virtual Tape Shelf is automatically created when you activate your first Gateway-VTL. Virtual tapes that need to be accessed frequently should be stored in a Virtual Tape Library. Data that does not need to be retrieved frequently can be archived to your Virtual Tape Shelf. Access to virtual tapes in your Virtual Tape Library is immediate while virtual tapes in your Virtual Tape Shelf will have to be retrieved and loaded into a Virtual Tape Library before being accessed. You can retrieve virtual tapes from your Virtual Tape Shelf using the AWS Management Console. Virtual tapes retrieved from your Virtual Tape Shelf take about 24 hours to be available and will automatically be loaded into your Virtual Tape Library.

Q: How can I get started using the AWS Storage Gateway?

To get started, sign up for the AWS Storage Gateway by clicking the “Sign Up Now” button on the AWS Storage Gateway detail page. To sign-up, you must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one when you begin the AWS Storage Gateway sign-up process. After you sign up, you can begin setting up and activating your gateway by visiting the AWS Management Console. To learn more, you can also refer to our Getting Started Documentation. We also have a Getting Started Video for Gateway-Cached Volumes.

Q: What are the minimum hardware and software requirements for the AWS Storage Gateway’s VM?

The AWS Storage Gateway VM must be either installed on a host in your datacenter running supported versions of VMware ESXi or Microsoft Hyper-V, or as an AMI running on an EC2 instance. The gateway VM must be deployed with a minimum set of hardware resources.

The AWS Storage Gateway currently supports Microsoft Windows, Red Hat Enterprise Linux, and VMware ESXi, iSCSI initiators.

Please visit the requirements section in the AWS Storage Gateway User Guide for details.

Q: Can I use the AWS Storage Gateway with AWS Direct Connect?

The AWS Storage Gateway efficiently uses your Internet bandwidth to speed up the upload of your on-premises application data to AWS. The AWS Storage Gateway only uploads data that has changed, minimizing the amount of data sent over the Internet. You can also use AWS Direct Connect to further increase throughput and reduce your network costs by establishing a dedicated network connection between your on-premises gateway and AWS.

Q: Can I route my AWS Storage Gateway Internet traffic through a local proxy server?

Yes, the AWS Storage Gateway supports the configuration of a SOCKS proxy between your gateway and AWS. You can specify an IP address and Port number for the host running your proxy, and the AWS Storage Gateway will route all HTTPS traffic through your proxy server.

Q: What type of data reduction does AWS Storage Gateway perform?

The AWS Storage Gateway performs compression of data in-transit and at-rest which can reduce both data transfer and storage charges. All data transfer between the AWS Storage Gateway VM and AWS, and all data stored in AWS, is compressed. In addition the AWS Storage Gateway VM only uploads data that has changed, minimizing the amount of data transferred.

Q: Does the AWS Storage Gateway support bandwidth throttling?

Yes, using the AWS Management Console you can restrict the bandwidth between your gateway and AWS based on a rate that you provide. You can specify individual rates for inbound and outbound traffic.

Longer Volume and Snapshot IDs

Q: Are volume and snapshot ID lengths changing in 2016?

Yes, volume and snapshot IDs will move to a longer format in December 2016. You can opt to receive the longer format from April 2016. Please visit the EC2 FAQ page for more details.

Q: How do I start using longer IDs for volumes or snapshots?

You can opt to receive longer IDs for volumes and snapshots using theAWS Management Console, the AWS Command Line Interface (CLI), the AWS Tools for Windows PowerShell, or though an API function. Opt-in status is can be set per region at the IAM user or account level, and when you opt in newly created volumes and ad hoc snapshots in that region will receive longer IDs. Existing volumes and snapshots will not be affected and will retain their existing short IDs. The longer IDs will be visible in the console, the command line, and API results, for all users even if they have not opted in.

The opt-in status for volumes and snapshots can be changed independently, and applies to both EBS and Storage Gateway volumes and snapshots. For example, if you opt to receive longer snapshot IDs only it will apply to all new EBS snapshots, both those taken from gateway-volumes and EBS volumes. Volume IDs will remain, in the older, short format.

For more detailed instructions on how to manage your opt-in status please visit theAWS Blog.

Q: How do I start using longer IDs for scheduled snapshots?

AWS Storage Gateway enables you tocreate a snapshot schedule for each of your gateway volumes. Scheduled snapshots use the opt-in status of your AWS account root user.

Gateway-Stored and Gateway-Cached Volumes

Q: What is the maximum size of a volume?

Each *gateway-cached volume* can store up to 32 TB of data. Data written to the volume is cached on your on-premises hardware and asynchronously uploaded to AWS for durable storage.

Each *gateway-stored volume* can store up to 16 TB of data. Data written to the volume is stored on your on-premises hardware and asynchronously backed up to AWS for point-in-time snapshots.

For both gateway-cached and gateway-stored volumes, the gateway performs compression of data before it is transferred to AWS and while stored in AWS. This can reduce both data transfer and storage charges. Volume storage is not pre-provisioned; you will be billed for only the amount of data stored on the volume, not the size of the volume you create.

Q: How much volume data can I manage per gateway?

Each *gateway-cached gateway* can support up to 32 volumes for a maximum of 1 PB of data (32 volumes, each 32 TB in size).

Each *gateway-stored gateway* can support up to 32 volumes for a maximum of 512 TB of data (32 volumes, each 16 TB in size).

Q: What performance can I expect from gateway-cached or gateway-stored volumes?

As the AWS Storage Gateway VM sits between your application, Amazon S3, and underlying on-premises storage, the performance you experience will be dependent upon a number of factors, including the speed and configuration of your underlying local disks, the network bandwidth between your iSCSI initiator and gateway VM, the amount of local storage allocated to your gateway VM, and the bandwidth between your gateway VM and Amazon S3.

For gateway-cached volumes, to provide low-latency read access to your on-premises applications, it’s important that you allocate enough local cache disk storage to store your recently accessed data. Our technical documentation provides guidance on how to optimize your environment setup for best performance, including how to properly size your local storage.

Q: Will I be able to access my gateway-stored or gateway-cached volume data using Amazon S3’s APIs?

No, gateway volumes are only accessible from the AWS Storage Gateway and cannot be directly accessed using Amazon S3 APIs. You can take point-in-time snapshots of gateway volumes which are made available in the form of Amazon EBS snapshots.

Snapshots

Q: What are the snapshot limits per gateway?

There are no limits to the number of snapshots or the amount of snapshot data a single gateway can produce.

Q: Why would I use snapshots?

Whether you're using gateway-cached or gateway-stored volumes, you can take point-in-time, incremental snapshots of your volume and store them in Amazon S3 in the form of Amazon EBS snapshots.

For gateway-stored volumes, where your volume data is stored on-premises, snapshots provide durable, off-site backups in Amazon S3. You can create a new gateway-stored volume from a snapshot in the event you need to recover a backup. You can also use a snapshot of your gateway-stored volume as the starting point for a new Amazon EBS volume which you can then attach to an Amazon EC2 instance.

For gateway-cached volumes, where your volume data is already stored in Amazon S3, snapshots can be used to preserve versions of your data, allowing you to revert to a prior version when required or to repurpose a point-in-time version as a new gateway-cached volume. Snapshots can be initiated on a scheduled or ad-hoc basis. When taking a new snapshot, only the data that has changed since your last snapshot is stored. If you have a volume with 100 GB of data, but only 5 GB of data have changed since your last snapshot, only the 5 additional GB of snapshot data will be stored in Amazon S3. When you delete a snapshot, only the data not needed for any other snapshot is removed.

Q: What data will my snapshot contain? How do I know when to take a snapshot to ensure my data is backed up?

Snapshots represent a point-in-time copy of the volume at the time the snapshot is requested. They contain all of the information needed to restore your data (from the time the snapshot was taken) to a new volume. Data written to the volume by your application prior to taking the snapshot, but not yet been uploaded to AWS, will be included in the snapshot.

In practical terms, the snapshot will be assigned an ID and visible in theAWS Management Console and AWS Command Line Interface (CLI) immediately, but will initially be in a PENDING status. Once all data written to the volume prior to the snapshot request has been uploaded from the gateway and into EBS, the status will change to AVAILABLE. At this point the snapshot can be used as the base for a new gateway or EBS volume.

Q: How do I restore a snapshot to a gateway volume?

Using the AWS Management Console, you can create a new gateway volume from a snapshot you've stored in Amazon S3. You can then mount this volume as an iSCSI device to your on-premises application server.

Because gateway-stored volumes store your primary data locally, when creating a new volume from a snapshot, your gateway downloads the data contained within the snapshot to your local hardware, where it becomes the primary data for your new volume.

Because gateway-cached volumes store your primary data in Amazon S3, when creating a new volume from a snapshot, your gateway keeps the snapshot data in Amazon S3 where it becomes the primary data for your new volume.

Q: Can I access my snapshots from within AWS?

Data written to your gateway-stored volumes is stored on your on-premises storage hardware, and asynchronously backed up to Amazon S3 in the form of Amazon EBS snapshots. You can use a snapshot of your gateway-stored volume as the starting point for a new Amazon EBS volume which you can then attach to an Amazon EC2 instance. This allows you to easily mirror data from your on-premises applications to your applications running on Amazon EC2 in the event you require additional on-demand compute capacity for data processing or replacement capacity for disaster recovery purposes.

Q: Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier which can be viewed using the AWS Management Console. You can create AWS Storage Gateway or Amazon EBS volumes based on any of your existing snapshots by specifying this unique identifier.

Q: Do the AWS Storage Gateway’s volumes need to be un-mounted in order to take a snapshot? Does the snapshot need to complete before the volume can be used again?

No, taking snapshots does not require you to un-mount your volumes, nor does it impact your application’s performance. However, snapshots only capture data that has been written to your AWS Storage Gateway volume, which may exclude any data that has been locally buffered by your application or OS.

Q: Can I schedule snapshots of my AWS Storage Gateway volumes?

Yes, you can create a snapshot schedule for each of your gateway-cached and gateway-stored volumes. You can modify both the time the snapshot occurs each day, as well as the frequency (every 1, 2, 4, 8, 12, or 24 hours).

Q: How do I start using longer IDs for scheduled snapshots?

AWS Storage Gateway enables you tocreate a snapshot schedule for each of your gateway volumes. Scheduled snapshots use the opt-in status of your AWS account root user.

For more information longer IDs, please see thegeneral section of this FAQ.

Q: How long does it take to complete a snapshot?

You can take snapshots of your gateway-cached volume in Amazon S3, or your on-premises gateway-stored volume. These snapshots are stored as Amazon EBS snapshots. The time it takes to complete a snapshot is largely dependent upon the size of your volume and the speed of your Internet connection to AWS. The AWS Storage Gateway compresses all data prior to upload, reducing the time to take a snapshot.

Q: Will I be able to access my snapshot data using Amazon S3's APIs?

No, snapshots are only accessible from the AWS Storage Gateway and Amazon EBS and cannot be directly accessed using Amazon S3 APIs.

Gateway-Virtual Tape Library (VTL)

Q: How much data can I store on a virtual tape?

When creating a virtual tape, you can select one of the following sizes: 100 GB, 200 GB, 400 GB, 800 GB, 1.5 TB, and 2.5 TB.

Q: How much data can I store in a Virtual Tape Library?

Each Virtual Tape Library (VTL) can store up to 1,500 virtual tapes with a maximum aggregate capacity of 1 PB.

Q: How much data can I store on a Virtual Tape Shelf?

There is no limit to the amount of data you can store on a Virtual Tape Shelf (VTS).

Q: How do I access my data on virtual tapes?

The virtual tape containing your data must be stored in a Virtual Tape Library before it can be accessed. Access to virtual tapes in your Virtual Tape Library is instantaneous. If the virtual tape containing your data is in your Virtual Tape Shelf, you must first retrieve the virtual tape from your Virtual Tape Shelf. You can retrieve the virtual tape using the AWS Management Console. First select the virtual tape, then choose the Virtual Tape Library into which you want the virtual tape to be loaded. It takes about 24 hours for the retrieved virtual tape to be available in the selected Virtual Tape Library. Once the virtual tape is available in the Virtual Tape Library, you can use your backup application to make use of the virtual tape to restore data.

Q: What backup applications can I use with Gateway-VTL?

Gateway-VTL works with backup and archival applications that use the industry-standard iSCSI-based tape library interface. For a full list of the supported backup applications see the requirements section of the AWS Storage Gateway user guide.

Q: What performance can I expect from a Gateway-VTL?

As the Gateway-VTL sits between your application, AWS, and underlying on-premises storage, the performance you experience will be dependent upon a number of factors, including the speed and configuration of your underlying local disks, processor and memory of the provisioned host, the network bandwidth between your iSCSI initiator and gateway VM, the amount of local storage allocated to the gateway VM, and the bandwidth between the gateway VM and AWS. For Gateway-VTL, to provide predictable write performance to your backup application, it's important that you allocate enough local cache disk storage to durably buffer data that is being uploaded to AWS. Please refer to the technical documentation for guidance on sizing cache.

We also recommend configuring your disks in a RAID (redundant array of independent disks) configuration to improve performance and to protect against disk failures.

Q: Will I be able to access the virtual tapes in my Virtual Tape Library using Amazon S3's APIs? Can I access the virtual tapes in my Virtual Tape Shelf using Amazon Glacier's APIs?

No. You cannot access virtual tape data using Amazon S3's APIs or Amazon Glacier's APIs. However, you can use Gateway-VTL's APIs to manage your Virtual Tape Library and your Virtual Tape Shelf.

Billing

Q: How will I be billed for my use of the AWS Storage Gateway?

You are billed a monthly fee for each of your gateways. This fee is prorated daily. Billing for a gateway begins upon activation and continues until you delete the gateway from the AWS Management Console or via the API.

Q: How will I be billed for storage consumed using AWS Storage Gateway?

Cached volume usage (per GB per month): You are billed for the cached volume data stored in AWS. This fee is prorated daily. You are only billed for the amount of data stored on the volume, not for the capacity of the volume you create.

Stored volume usage (per GB per month): There is no charge for the stored volume data stored locally on your on-premises hardware, or asynchronously backed up to AWS.

Snapshot usage (per GB per month): You are billed for the snapshots your gateway creates in AWS from stored and cached volumes. Snapshots are stored and billed as Amazon EBS snapshots. When taking a new snapshot only the data that has changed since your last snapshot is stored to reduce your storage charges. For more details on snapshots, please visit the Product Details page.

*Virtual Tape Library and Virtual Tape Shelf usage (per GB per month):*You are billed for the virtual tape data you store in AWS. This fee is prorated daily. You are only billed for the portion of virtual tape capacity that you use, not for the size of the virtual tape you create.

In addition, all volume, snapshot, and virtual tape storage is compressed to further reduce your storage charges. For detailed pricing information, please visit the Pricing page

Q: How will I be billed for data transfer to and from AWS?

You are billed for Internet data transfer for each GB downloaded from AWS to your gateway. All data transfer for uploading to AWS is free.

Q: How will I be charged when retrieving data from Virtual Tape Shelf?

You are charged for the peak concurrent data retrieval in a month.

Concurrent data retrieval is calculated when a retrieval from Virtual Tape Shelf (VTS) is initiated. The concurrent data retrieval is the sum of the size of all virtual tape retrievals that were initiated concurrently or were concurrently in progress. If the concurrent data retrieval is more than any prior concurrent data retrieval in the month you are charged for the difference between this new monthly peak and the prior monthly peak. If the concurrent data retrieval is less than or equal to any prior concurrent data retrieval in the month, the retrieval is free.

Example 1: You initiate retrieval of a 100 GB virtual tape from your VTS in US East (Northern Virginia) Region. This is the first retrieval of the month. The prior peak concurrent data retrieval for the month was 0 GB. When the retrieval is initiated the concurrent data retrieved will be 100 GB. 100 GB will be your new monthly peak concurrent data retrieval. The charge for the retrieval will be the difference between the prior and the new peak concurrent data retrieval, i.e., (100 GB – 0 GB) x $0.30/GB = $30.00.

Example 2: You initiate retrieval of one virtual tape containing 500 GB of data from your VTS in US East (Northern Virginia) Region. This is the first retrieval of the month. The prior peak concurrent data retrieval for the month was 0 GB. When the retrieval is initiated the concurrent data retrieved will be 500 GB. 500 GB will be your new monthly peak concurrent data retrieval. You will be charged the difference between the prior and new peak concurrent data retrieval, i.e. (500 GB - 0 GB) x $0.30 / GB = $150. Twelve hours after initiating the retrieval of the 500 GB virtual tape you initiate retrieval of a virtual tape containing 600 GB of data. When the second retrieval is initiated you will have two virtual tape retrievals concurrently in progress as any virtual tape retrieval takes about 24 hours to complete. The concurrent data retrieved will be 500 GB + 600 GB = 1100 GB. The new peak concurrent data retrieval for the month will be 1100 GB and you will now be charged the difference between the new peak concurrent data retrieval and the existing peak concurrent data retrieval, i.e., (1100 GB-500 GB) x $0.30 = $180.

Example 3: You initiate retrieval of one 500 GB virtual tape from the VTS in US East (Northern Virginia) Region. This is the first retrieval of the month. The prior peak concurrent data retrieval for the month was 0 GB. When the retrieval is initiated the concurrent data retrieved will be 500 GB which will be your new peak concurrent data retrieved. The charge for the retrieval will be the difference between the prior and the new peak concurrent data retrieval i.e., (500 GB – 0 GB) x $0.30/GB = $150.00. The next day, you initiate a retrieval of one 500 GB virtual tape after the first tape retrieval is complete. Because 500 GB is equal to the prior peak concurrent data retrieval of 500 GB, the second retrieval will be free.

Q: How will I be charged for deleting data from my Virtual Tape Shelf that is less than 3 months old?

Virtual Tape Shelf is designed for use cases where data is retained for months, years, or decades. Deleting virtual tapes from a Virtual Tape Shelf is free if the virtual tape being deleted has been stored for three months or longer. If a virtual tape is deleted within three months of being archived, you will be charged an early deletion fee. In the US East (Northern Virginia) Region, you would be charged a prorated early deletion fee of $0.03 per GB deleted within three months. For example, if you delete 1 virtual tape containing 1 GB of data 1 month after uploading it, you would be charged a $0.02 early deletion fee. If, instead you delete the same virtual tape after 2 months, you would be charged a $0.01 early deletion fee.

Q: How can I tell how much storage I am going to be billed for?

The Billing and Cost Management console shows an estimate of month-to-date usage for each service, including AWS Storage Gateway volumes and virtual tapes. For a breakdown of usage by individual volume or virtual tape, Detailed Billing Reports enables you to see usage for each resource on a daily basis.

Q: When does each monthly billing cycle begin?

The billing system follows Coordinated Universal Time (UTC). The calendar month begins midnight UTC on the first day of every month.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Security

Q: Does the AWS Storage Gateway encrypt my data?

The AWS Storage Gateway encrypts all data in-transit to and from AWS via SSL.

All volume and snapshot data stored in AWS using Gateway-Stored Volumes and Gateway-Cached Volumes, and all virtual tape data stored in AWS using Gateway-VTL, is encrypted at-rest using Advanced Encryption Standard (AES) 256, a secure symmetric-key encryption standard using 256-bit encryption keys.

Q: What form of iSCSI authentication does the AWS Storage Gateway support?

The AWS Storage Gateway supports authentication between your gateway and iSCSI initiators via CHAP (Challenge-Handshake Authentication Protocol).

Q: Can I get a history of Storage Gateway API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of Storage Gateway API calls made on your account, you simply turn on CloudTrail in the AWS Management Console. For more information visit the AWS CloudTrail details page.

Monitoring and Maintenance

Q: How do I monitor my gateway?

You can use Amazon CloudWatch to monitor the performance metrics for your AWS Storage Gateway, giving you insight into storage, bandwidth, throughput, and latency. These metrics are accessible from the AWS Management Console. Please refer to our technical documentation to learn more how, and to the CloudWatch details and pricing pages.

Q: How does the AWS Storage Gateway manage updates?

When configuring your gateway, you can specify a weekly maintenance schedule. This allows you to control when the AWS Storage Gateway service deploys important updates and software patches to your local gateway. Updates should take only a few minutes to complete.

Support

Q: Does AWS Premium Support cover the AWS Storage Gateway?

Yes, AWS Premium Support covers issues related to your use of the AWS Storage Gateway. Please see the AWS Premium Support detail page for further information and pricing.

Q: What other support options are available?

You can tap into the breadth of existing AWS community knowledge through theAWS Storage Gateway discussion forum.

Amazon RDS FAQ

You can also refer to the FAQs page for each Amazon RDS Database Engine for engine-specific questions.

Amazon Aurora | MySQL | MariaDB | Oracle | PostgreSQL

General

Q: What is Amazon RDS?

Amazon Relational Database Service (Amazon RDS) is a managed 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, MariaDB, Oracle, SQL Server, or PostgreSQL database. This means that the code, applications, and tools you already use today with your existing databases should work seamlessly with Amazon RDS. Amazon RDS automatically patches the database software and backs up your database, storing the backups for a user-defined retention period. You benefit from the flexibility of being able to easily scale the compute resources or storage capacity associated with your relational database instance. In addition, Amazon RDS makes it easy to use replication to enhance database availability, improve data durability, or scale beyond the capacity constraints of a single database instance for read-heavy database workloads. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

Q: What is a database instance (DB Instance)?

You can think of a DB Instance as a database environment in the cloud with the compute and storage resources you specify. You can create and delete DB Instances, define/refine infrastructure attributes of your DB Instance(s), and control access and security via the AWS Management Console, Amazon RDS APIs, and AWS Command Line Interface. You can run one or more DB Instances, and each DB Instance can supportone or more databases or database schemas, depending on engine type.

Q: What does Amazon RDS manage on my behalf?

Amazon RDS manages the work involved in setting up a relational database: from provisioning the infrastructure capacity you request to installing the database software. Once your database is running on its own DB Instance, Amazon RDS automates common administrative tasks, such as performing backups and patching the database software that powers your DB Instance. For optional Multi-AZ deployments, Amazon RDS also manages synchronous data replication across Availability Zones and automatic failover.

Since Amazon RDS provides native database access, you interact with the relational database software as you normally would. This means you're still responsible for managing the database settings that are specific to your application. You'll need to build the relational schema that best fits your use case and are responsible for any performance tuning to optimize your database for your application’s workflow.

Q: When would I use Amazon RDS vs. Amazon EC2 Relational Database AMIs?

Amazon Web Services provides a number of database alternatives for developers. Amazon RDS enables you to run a fully featured relational database while offloading database administration. Using one of our many relational database AMIs on Amazon EC2 allows you to manage your own relational database in the cloud. There are important differences between these alternatives that may make one more appropriate for your use case. See Cloud Databases with AWS for guidance on which solution is best for you.

Q: How do I get started with Amazon RDS?

To sign up for Amazon RDS, you must have an Amazon Web Services account.Create an account if you do not already have one. After you are signed up, please refer to the Amazon RDS documentation, which includes our Getting Started Guide.

Q: How do I create a DB Instance?

DB Instances are simple to create, using either theAWS Management Console, Amazon RDS APIs, or AWS Command Line Interface. To launch a DB Instance using the AWS Management Console, click "RDS," then the "Launch DB Instance" button on the "Instances" tab. From there, you can specify the parameters for your DB instance including DB engine and version, license model, instance type, storage type and amount, and master user credentials.

You also have the ability to change your DB Instance'sbackup retention policy, preferred backup window, and scheduled maintenance window. Alternatively, you can create your DB Instance using the CreateDBInstance API or create-db-instance command.

Q: How do I access my running DB Instance?

Once your DB Instance is available, you can retrieve its endpoint via the DB Instance description in the AWS Management Console, DescribeDBInstances API or describe-db-instances command. Using this endpoint you can construct the connection string required to connect directly with your DB Instance using your favorite database tool or programming language. In order to allow network requests to your running DB Instance, you will need to authorize access. For a detailed explanation of how to construct your connection string and get started, please refer to our Getting Started Guide.

Q: How many DB Instances can I run with Amazon RDS?

By default, customers are allowed to have up to a total of 40 Amazon RDS DB instances. Of those 40, up to 10 can be Oracle or SQL Server DB Instances under the "License Included" model. All 40 can be used for Amazon Aurora, MySQL, MariaDB, Oracle, SQL Server, or PostgreSQL under the "BYOL" model. If your application requires more DB Instances, you can request additional DB Instances via this request form.

Q: How many databases or schemas can I run within a DB Instance?

- RDS for Amazon Aurora: No limit imposed by software
- RDS for MySQL: No limit imposed by software
- RDS for MariaDB: No limit imposed by software
- RDS for Oracle: 1 database per instance; no limit on number of schemas per database

imposed by software

- RDS for SQL Server: 30 databases per instance
- RDS for PostgreSQL: No limit imposed by software

Q: How do I import data to Amazon RDS?

There are a number of simple ways to import data into Amazon RDS, such as with the mysqldump or mysqlimport utilities for MySQL; Data Pump, import/export or SQL Loader for Oracle; Import/Export wizard, full backup files (.bak files) or Bulk Copy Program (BCP) for SQL Server; or pg_dump for PostgreSQL. For more information on data import and export, please refer to the Data Import Guide for MySQL or the Data Import Guide for Oracle or the Data Import Guide for SQL Server or the Data Import Guide for PostgreSQL.

Q: Which relational database engines does Amazon RDS support?

Amazon RDS supports Amazon Aurora, MySQL, MariaDB, Oracle, SQL Server, and PostgreSQL database engines.

Amazon RDS for MySQL currently supports MySQL 5.5, 5.6 and 5.7 (Community Edition) with InnoDB as the default database storage engine. Amazon RDS for MariaDB currently supports MariaDB 10.0 and 10.1. Amazon RDS for Oracle currently supports Oracle Database 11gR2 and 12c. Amazon RDS for SQL Server currently supports 2008 R2, SQL Server 2012 (SP2) and SQL Server 2014. Amazon RDS for PostgreSQL currently supports PostgreSQL 9.3, 9.4 and 9.5.

For information about upgrading a DB Instance to a new DB engine version, refer to theAmazon RDS User Guide.

Q: What is a maintenance window? Will my DB Instance be available during software maintenance?

The Amazon RDS maintenance window is your opportunity to control when DB Instance modifications (such as scaling DB Instance class) and software patching occur, in the event they are requested or required. If a maintenance event is scheduled for a given week, it will be initiated and completed at some point during the maintenance window you identify. Maintenance windows are 30 minutes in duration.

The only maintenance events that require Amazon RDS to take your DB Instance offline are scale compute operations (which generally take only a few minutes from start-to-finish) or required software patching. Required patching is automatically scheduled only for patches that are security and durability related. Such patching occurs infrequently (typically once every few months) and should seldom require more than a fraction of your maintenance window. If you do not specify a preferred weekly maintenance window when creating your DB Instance, a 30 minute default value is assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your DB Instance in the AWS Management Console, the

ModifyDBInstance API or the modify-db-instance command. Each of your DB Instances can have different preferred maintenance windows, if you so choose.

Running your DB Instance as a Multi-AZ deployment can further reduce the impact of a maintenance event. Please refer to the Amazon RDS User Guide for more information on maintenance operations.

Q: Does Amazon RDS provide guidelines for support of new database engine versions, and for deprecating database engine versions that are currently supported?

This statement applies to Amazon RDS for Amazon Aurora, MySQL, MariaDB, Oracle, SQL Server, and PostgreSQL.

Over time, we plan to support additional database versions, both minor and major, for Amazon RDS's engines. The number of new version releases supported in a given year will vary based on the frequency and content of releases and patches from the engine's vendor or core team, and the outcome of a thorough vetting of these releases and patches by our database engineering team. However, as a general guidance, we aim to support new engine versions within 3-5 months of their general availability.

Here is a general statement of Amazon RDS's deprecation policy:

- We intend to support major version releases (e.g., MySQL 5.6) for at least 3 years after they are initially supported by Amazon RDS.
- We intend to support minor versions (e.g., MySQL 5.6.21) for at least 1 year after they are initially supported by Amazon RDS.
- From time to time, we will deprecate major or minor versions. We expect to provide a three-month grace period after the announcement of a deprecation for you to initiate an upgrade to a supported version. At the end of this grace period, an automatic upgrade will be applied to any un-upgraded instances during their scheduled maintenance windows.
- While we strive to meet these guidelines, in some cases we may deprecate specific major or minor versions sooner, such as when there are security issues.

Q: What should I do if my queries seem to be running slow?

For production databases we encourage you to enable Enhanced Monitoring, which provides access to over 50 CPU, memory, file system, and disk I/O metrics. You can enable these features on a per-instance basis and you can choose the granularity (all the way down to 1 second). High levels of CPU utilization can reduce query performance and in this case you may want to consider scaling your DB Instance class. For more information on monitoring your DB Instance, refer to the Amazon RDS User Guide.

If you are using MySQL or MariaDB, you can access the slow query logs for your database to determine if there are slow-running SQL queries and, if so, the performance characteristics of each. You could set the "slow_query_log" DB Parameter and query the mysql.slow_log table to review the slow-running SQL queries. Please refer to the Amazon RDS User Guide to learn more.

If you are using Oracle, you can use the Oracle trace file data to identify slow queries. For more information on accessing trace file data, please refer to Amazon RDS User Guide.

If you are using SQL Server, you can use the client side SQL Server traces to identify slow queries. For information on accessing server side trace file data, please refer to Amazon RDS User Guide.

Q: Why is the pricing for each RDS database engine different?

The pricing for each database engine of RDS varies because our costs are different for each. These costs include many operational components in addition to software licensing. We will continue to work hard to reduce costs and pass on those savings to our customers.

Billing

Q: How will I be charged and billed for my use of Amazon RDS?

You pay only for what you use, and there are no minimum or setup fees. You are billed based on:

- DB Instance hours – Based on the class (e.g. Standard Small, Large, Extra Large) of the DB Instance consumed. Partial DB Instance hours consumed are billed as full hours.
- Storage (per GB per month) – Storage capacity you have provisioned to your DB Instance. If you scale your provisioned storage capacity within the month, your bill will be pro-rated.
- I/O requests per month – Total number of storage I/O requests you have*(for Amazon RDS Magnetic Storage only)*
- Provisioned IOPS per month – Provisioned IOPS rate, regardless of IOPS consumed*(for Amazon RDS Provisioned IOPS (SSD) Storage only)*
- Backup Storage – Backup storage is the storage associated with your automated database backups and any active database snapshots you have taken. Increasing your backup retention period or taking additional database snapshots increases the backup storage consumed by your database. Amazon RDS provides backup storage up to 100% of your provisioned database storage at no additional charge. For example, if you have 10GB-months of provisioned database storage, we will provide up to 10GB-months of backup storage at no additional charge. Based upon our experience as database administrators, the vast majority of

databases require less raw storage for a backup than for the primary data set, meaning that most customers will never pay for backup storage. Backup storage is only free for active DB Instances.

- Data transfer –Internet data transfer in and out of your DB Instance.

For Amazon RDS pricing information, please visit thepricing section on the Amazon RDS product page.

Q: When does billing of my Amazon RDS DB Instances begin and end?

Billing commences for a DB Instance as soon as the DB Instance is available. Billing continues until the DB Instance terminates, which would occur upon deletion or in the event of instance failure.

Q: What defines billable Amazon RDS instance hours?

DB Instance hours are billed for each hour your DB Instance is running in an available state. If you no longer wish to be charged for your DB Instance, you must terminate it to avoid being billed for additional instance-hours. Partial DB Instance hours consumed are billed as full hours.

Q: Why does additional backup storage cost more than allocated DB Instance storage?

The storage provisioned to your DB Instance for your primary data is located within a single Availability Zone. When your database is backed up, the backup data (including transactions logs) is geo-redundantly replicated across multiple Availability Zones to provide even greater levels of data durability. The price for backup storage beyond your free allocation reflects this extra replication that occurs to maximize the durability of your critical backups.

Q: How will I be billed for Multi-AZ DB Instance deployments?

If you specify that your DB Instance should be a Multi-AZ deployment, you will be billed according to the Multi-AZ pricing posted on the Amazon RDS pricing page. Multi-AZ billing is based on:

- Multi-AZ DB Instance Hours – Based on the class (e.g. Small, Large, Extra Large) of the DB Instance consumed. As with standard deployments in a single Availability Zone, partial DB Instance hours consumed are billed as full hours. If you convert your DB Instance deployment between standard and Multi-AZ within a given hour, you will be charged both applicable rates for that hour.

- Provisioned storage (for Multi-AZ DB Instance) – If you convert your deployment between standard and Multi-AZ within a given hour, you will be charged the higher of the applicable storage rates for that hour.

- I/O requests per month – Total number of storage I/O requests you have. Multi-AZ deployments consume a larger volume of I/O requests than standard DB Instance deployments, depending on your database write/read ratio. Write I/O usage associated with database updates will double as Amazon RDS synchronously replicates your data to the standby DB instance. Read I/O usage will remain the same.

- Backup Storage – Your backup storage usage will not change whether your DB Instance is a standard or Multi-AZ deployment. Backups will simply be taken from your standby to avoid I/O suspension on the DB Instance primary.
- Data transfer – You are not charged for the data transfer incurred in replicating data between your primary and standby.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Free Tier

Q: What does the AWS Free Tier for Amazon RDS offer?

The AWS Free Tier for Amazon RDS offer provides free use of Single-AZ Micro DB instances running MySQL, MariaDB, PostgreSQL, Oracle ("Bring-Your-Own-License (BYOL)" licensing model) and SQL Server Express Edition. The free usage tier is capped at 750 instance hours per month. Customers also receive 20 GB of database storage, 10 million I/Os and 20 GB of backup storage for free per month.

Q: For what time period will the AWS Free Tier for Amazon RDS be available to me?

New AWS accounts receive 12 months of AWS Free Tier access. Please see theAWS Free Tier FAQs for more information.

Q: Can I run more than one DB instance under the AWS Free Usage Tier for Amazon RDS?

Yes. You can run more than one Single-AZ Micro DB instance simultaneously and be eligible for usage counted under the AWS Free Tier for Amazon RDS. However, any use exceeding 750 instance hours, across all Amazon RDS Single-AZ Micro DB instances, across all eligible database engines and regions, will be billed at standard Amazon RDS prices.

For example: if you run two Single-AZ Micro DB instances for 400 hours each in a single month, you will accumulate 800 instance hours of usage, of which 750 hours will be free. You will be billed for the remaining 50 hours at the standard Amazon RDS price.

Q: Do I have access to 750 instance hours each of the MySQL, MariaDB, PostgreSQL,

Oracle and SQL Server Micro DB instances under the AWS Free Tier?

No. A customer with access to the AWS Free Tier can use up to 750 instance hours of Micro instances running either MySQL, PostgreSQL, Oracle or SQL Server Express Edition. Any use exceeding 750 instance hours, across all Amazon RDS Single-AZ Micro DB instances, across all eligible database engines and regions, will be billed at standard Amazon RDS prices.

Q: Is the AWS Free Tier for Amazon RDS available in all AWS Regions?

The AWS Free Tier for Amazon RDS is available in all AWS Regions except GovCloud (US).

Q: How am I billed when my instance-hour usage exceeds the Free Tier benefit?

You are billed at standard Amazon RDS prices for instance hours beyond what the Free Tier provides. See the Amazon RDS pricing page for details.

Reserved Instances

Q: What is a Reserved Instance (RI)?

Amazon RDS Reserved Instances give you the option to reserve a DB Instance for a one or three year term and in turn receive a significant discount compared to the On-Demand Instance pricing for the DB Instance. There are three RI payment options -- No Upfront, Partial Upfront, All Upfront -- which enable you to balance the amount you pay upfront with your effective hourly price.

Q: How are Reserved Instances different from On-Demand DB Instances?

Functionally, Reserved Instances and On-Demand DB Instances are exactly the same. The only difference is how your DB Instance(s) are billed: With Reserved Instances, you purchase a one or three year reservation and in return receive a lower effective hourly usage rate (compared with On-Demand DB Instances) for the duration of the term.

Q: How do I purchase and create Reserved Instances?

You can purchase a Reserved Instance in the "Reserved Purchase" section of the AWS Management Console. Alternatively, you can use the Amazon RDS API or AWS Command Line Interface to list the reservations available for purchase then purchase a DB Instance reservation.

Once you have made a reserved purchase, using a Reserved DB Instance is no different than an On-Demand DB Instance. Launch a DB Instance using the same instance class, engine and region for which you made the reservation. As long as your reservation purchase is active, Amazon RDS will apply the reduced hourly rate for which you are eligible to the new DB Instance.

Q: Will there always be reservations available for purchase?

Yes. Reserved Instances are purchased for the Region rather than for the Availability Zone. This means that even if capacity is limited in one Availability Zone, reservations can still be purchased in that Region and used in a different Availability Zone within that Region.

Q: How many Reserved Instances can I purchase?

You can purchase up to 40 Reserved DB Instances. If you wish to run more than 40 DB Instances, please complete the Amazon RDS DB Instance request form.

Q: What if I have an existing DB Instance that I'd like to convert to a Reserved Instance?

Simply purchase a DB Instance reservation with the same DB Instance class, DB Engine and License Model within the same Region as the DB Instance you are currently running and would like to reserve. If the reservation purchase is successful, Amazon RDS will automatically apply your new hourly usage charge to your existing DB Instance.

Q: If I sign up for a Reserved Instance, when does the term begin? What happens to my DB Instance when the term ends?

Pricing changes associated with a Reserved Instance are activated once your request is received while the payment authorization is processed. You can follow the status of your reservation on the AWS Account Activity page or by using the DescribeReservedDBInstances API or describe-reserved-db-instances command. If the one-time payment cannot be successfully authorized by the next billing period, the discounted price will not take effect.

When your reservation term expires, your Reserved Instance will revert to the appropriate On-Demand hourly usage rate for your DB Instance class and Region.

Q: How do I control which DB Instances are billed at the Reserved Instance rate?

The Amazon RDS operations for creating, modifying, and deleting DB Instances do not distinguish between On-Demand and Reserved Instances. When computing your bill, our system will automatically apply your Reservation(s) such that all eligible DB Instances are charged at the lower hourly Reserved DB Instance rate.

Q: If I scale my Reserved Instance class up or down, what happens to my reservation?

Each reservation is associated with the following set of attributes: DB Engine, DB Instance class, Deployment type, License Model and Region. Each reservation can only be applied to a DB Instance with the same attributes for the duration of the term. If you decide to modify any of these attributes of your running DB Instance class before the end of the reservation term, your hourly usage rates for that DB Instance will revert to on demand hourly rates. If you later modify the running DB Instance's attributes to match those of the original reservation, or create a new DB Instance with the same attributes as your original reservation, your reserved pricing will be applied to it until the end of your reservation term.

Q: Can I move a Reserved Instance from one Region or Availability Zone to another?

Each Reserved Instance is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each reservation can, however, be used in any of the available AZs within the associated Region.

Q: Are Reserved Instances available for Multi-AZ Deployments?

Yes. When you call the DescribeReservedDBInstancesOfferings API or describe-reserved-db-instances-offerings command, simply look for the Multi-AZ options listed among the DB Instance configurations available for purchase. If you want to purchase a reservation for a DB Instance with synchronous replication across multiple Availability Zones, specify one of these offerings in your PurchaseReservedDBInstancesOffering call.

Q: Are Reserved Instances available for Read Replicas?

A standard DB Instance reservation can also be applied to a Read Replica, provided the DB Instance class and Region are the same. When computing your bill, our system will automatically apply your Reservation(s), such that all eligible DB Instances are charged at the lower hourly Reserved Instance rate.

Q: Can I cancel a reservation?

No, you cannot cancel your Reserved DB Instance and the one-time payment (if applicable) is not refundable. You will continue to pay for every hour during your Reserved DB Instance term regardless of your usage.

Q: How do the payment options impact my bill?

When you purchase an RI under the All Upfront payment option, you pay for the entire term of the RI in one upfront payment. You can choose to pay nothing upfront by choosing the No Upfront option. The entire value of the No Upfront RI is spread across every hour in the term and you will be billed for every hour in the term, regardless of usage. The Partial Upfront payment option is a hybrid of the All Upfront and No Upfront options. You make a small upfront payment, and you are billed a low hourly rate for every hour in the term regardless of usage.

Hardware and Scaling

Q: How do I determine which initial DB Instance class and storage capacity are appropriate for my needs?

In order to select your initial DB Instance class and storage capacity, you will want to assess your application’s compute, memory and storage needs. For information the about the DB Instance classes available, please refer to the Amazon RDS User Guide.

Q: How do I scale the compute resources and/or storage capacity associated with my Amazon RDS Database Instance?

You can scale the compute resources and storage capacity allocated to your DB Instance with the AWS Management Console (selecting the desired DB Instance and clicking the “Modify” button), the RDS API, or the AWS Command Line Interface. Memory and CPU resources are modified by changing your DB Instance class, and storage available is changed when you modify your storage allocation. Please note that when you modify your DB Instance class or allocated storage, your requested changes will be applied during your specified maintenance window. Alternately, you can use the “apply-immediately” flag to apply your scaling requests immediately. Bear in mind that any other pending system changes will be applied as well.

Monitor the compute and storage resource utilization of your DB Instance, for no additional charge, via Amazon CloudWatch. You can access metrics such as CPU utilization, storage utilization, and network traffic by clicking the “Monitoring” tab for your DB Instance in the AWS Management Console or using the Amazon CloudWatch APIs. To learn more about monitoring your active DB Instances, read the Amazon RDS User Guide.

Please note that for SQL Server, because of the extensibility limitations of striped storage attached to a Windows Server environment, Amazon RDS does not currently support increasing storage. While we plan to support this functionality in the future, we recommend you to provision storage based on anticipated future storage growth. In the interim, if you need to increase the storage of a SQL Server DB Instance, you will need to export the data, create a new DB Instance with increased storage, and import the data into it. Please refer to the data import guide for SQL Server for more information.

Q: What is the hardware configuration for Amazon RDS Storage?

Amazon RDS uses EBS volumes for database and log storage. Depending on the size of storage requested, Amazon RDS automatically stripes across multiple EBS volumes to enhance IOPS performance. For MySQL and Oracle, for an existing DB Instance, you may observe some I/O capacity improvement if you scale up your storage. You can scale the storage capacity allocated to your DB Instance using the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command.

However, for SQL Server, because of the extensibility limitations of striped storage attached to a Windows Server environment, Amazon RDS does not currently support increasing storage.

For more information, see Storage for Amazon RDS.

Q: Will my DB Instance remain available during scaling?

The storage capacity allocated to your DB Instance can be increased while maintaining DB Instance availability. However, when you decide to scale the compute resources available to your DB Instance up or down, your database will be temporarily unavailable while the DB

Instance class is modified. This period of unavailability typically lasts only a few minutes, and will occur during the maintenance window for your DB Instance, unless you specify that the modification should be applied immediately.

Q: How can I scale my DB Instance beyond the largest DB Instance class and maximum storage capacity?

Amazon RDS supports a variety of DB Instance classes and storage allocations to meet different application needs. If your application requires more compute resources than the largest DB Instance class or more storage than the maximum allocation, you can implement partitioning, thereby spreading your data across multiple DB Instances.

Q: What is Amazon RDS General Purpose (SSD) Storage?

Amazon RDS General Purpose (SSD) Storage is suitable for a broad range of database workloads that have moderate I/O requirements. With the baseline of 3 IOPS/GB and ability to burst up to 3,000 IOPS, this storage option provides predictable performance to meet the needs of most applications.

Q: What is Amazon RDS Provisioned IOPS (SSD) Storage?

Amazon RDS Provisioned IOPS (SSD) Storage is an SSD-backed storage option designed to deliver fast, predictable, and consistent I/O performance. With Amazon RDS Provisioned IOPS (SSD) Storage, you specify an IOPS rate when creating a DB Instance, and Amazon RDS provisions that IOPS rate for the lifetime of the DB Instance. Amazon RDS Provisioned IOPS (SSD) Storage is optimized for I/O-intensive, transactional (OLTP) database workloads. For more details, please see the Amazon RDS User Guide.

Q: What is Amazon RDS Magnetic Storage?

Formerly known as Standard storage, Amazon RDS Magnetic Storage is useful for small database workloads where data is accessed less frequently.

Q: How do I choose among the Amazon RDS storage types?

Choose the storage type most suited for your workload.

- High-performance OLTP workloads: Amazon RDS Provisioned IOPS (SSD) Storage
- Database workloads with moderate I/O requirements: Amazon RDS General Purpose (SSD) Storage
- Small database workloads with infrequent I/O: Amazon RDS Magnetic Storage

Q: What are the minimum and maximum IOPS supported by Amazon RDS?

The IOPS supported by Amazon RDS varies by database engine. For more details, please see the Amazon RDS User Guide.

Automated Backups and Database Snapshots

Q: What is the difference between automated backups and DB Snapshots?

Amazon RDS provides two different methods for backing up and restoring your DB Instance(s) automated backups and database snapshots (DB Snapshots).

The automated backup feature of Amazon RDS enables point-in-time recovery of your DB Instance. When automated backups are turned on for your DB Instance, Amazon RDS automatically performs a full daily snapshot of your data (during your preferred backup window) and captures transaction logs (as updates to your DB Instance are made). When you initiate a point-in-time recovery, transaction logs are applied to the most appropriate daily backup in order to restore your DB Instance to the specific time you requested. Amazon RDS retains backups of a DB Instance for a limited, user-specified period of time called the retention period, which by default is one day but can be set to up to thirty five days. You can initiate a point-in-time restore and specify any second during your retention period, up to the Latest Restorable Time. You can use the DescribeDBInstances API to return the latest restorable time for you DB Instance(s), which is typically within the last five minutes. Alternatively, you can find the Latest Restorable Time for a DB Instance by selecting it in the AWS Management Console and looking in the “Description” tab in the lower panel of the Console.

DB Snapshots are user-initiated and enable you to back up your DB Instance in a known state as frequently as you wish, and then restore to that specific state at any time. DB Snapshots can be created with the AWS Management Console, CreateDBSnapshot API, or create-db-snapshot command and are kept until you explicitly delete them.

The snapshots which Amazon RDS performs for enabling automated backups are available to you for copying (using the AWS console or the copy-db-snapshot command) or for the snapshot restore functionality. You can identify them using the "automated" Snapshot Type. In addition, you can identify the time at which the snapshot has been taken by viewing the "Snapshot Created Time" field. Alternatively, the identifier of the "automated" snapshots also contains the time (in UTC) at which the snapshot has been taken.

Please note: When you perform a restore operation to a point in time or from a DB Snapshot, a new DB Instance is created with a new endpoint (the old DB Instance can be deleted if so desired). This is done to enable you to create multiple DB Instances from a specific DB Snapshot or point in time.

Q: Do I need to enable backups for my DB Instance or is it done automatically?

By default and at no additional charge, Amazon RDS enables automated backups of your DB Instance with a 1 day retention period. Free backup storage is limited to the size of your provisioned database and only applies to active DB Instances. For example, if you have 10GB-months of provisioned database storage, we will provide at most 10GB-months of backup storage at no additional charge. If you would like to extend your backup retention period beyond one day, you can do so using the CreateDBInstance API (when creating a new DB Instance) or ModifyDBInstance API (for an existing DB Instance). You can use these APIs to change the RetentionPeriod parameter from 1 to the desired number of days. For more information on automated backups, please refer to the Amazon RDS User Guide.

Q: What is a backup window and why do I need it? Is my database available during the backup window?

The preferred backup window is the user-defined period of time during which your DB Instance is backed up. Amazon RDS uses these periodic data backups in conjunction with your transaction logs to enable you to restore your DB Instance to any second during your retention period, up to the LatestRestorableTime (typically up to the last few minutes). During the backup window, storage I/O may be briefly suspended while the backup process initializes (typically under a few seconds) and you may experience a brief period of elevated latency. There is no I/O suspension for Multi-AZ DB deployments, since the backup is taken from the standby.

Q: Where are my automated backups and DB Snapshots stored and how do I manage their retention?

Amazon RDS DB snapshots and automated backups are stored in S3.

You can use the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended). You can manage your user-created DB Snapshots via the "Snapshots" section of the Amazon RDS Console. Alternatively, you can see a list of the user-created DB Snapshots for a given DB Instance using the DescribeDBSnapshots API or describe-db-snapshots command and delete snapshots with the DeleteDBSnapshot API or delete-db-snapshot command.

Q: What happens to my backups and DB Snapshots if I delete my DB Instance?

When you delete a DB Instance, you can create a final DB Snapshot upon deletion; if you do, you can use this DB Snapshot to restore the deleted DB Instance at a later date. Amazon RDS retains this final user-created DB Snapshot along with all other manually created DB Snapshots after the DB Instance is deleted. Refer to the pricing page for details of backup storage costs.

Automated backups are deleted when the DB Instance is deleted. Only manually created DB Snapshots are retained after the DB Instance is deleted.

Security

Q: What is Amazon Virtual Private Cloud (VPC) and why may I want to use with Amazon RDS?

Amazon VPC lets you create a virtual networking environment in a private, isolated section of the Amazon Web Services (AWS) cloud, where you can exercise complete control over aspects such as private IP address ranges, subnets, routing tables and network gateways. With Amazon VPC, you can define a virtual network topology and customize the network configuration to closely resemble a traditional IP network that you might operate in your own datacenter.

One of the scenarios where you may want to use Amazon RDS in VPC is if you want to run a public-facing web application, while still maintaining non-publicly accessible backend servers in a private subnet. You can create a public-facing subnet for your webservers that has access to the Internet, and place your backend RDS DB Instances in a private-facing subnet with no Internet access. For more information about Amazon VPC, refer to the Amazon Virtual Private Cloud User Guide.

Q: How is using Amazon RDS inside a VPC different from using it on the EC2-Classic platform (non-VPC)?

The basic functionality of Amazon RDS is the same regardless of whether VPC is used. Amazon RDS manages backups, software patching, automatic failure detection, read replicas and recovery whether your DB Instances are deployed inside or outside a VPC.

Amazon RDS DB Instances deployed outside a VPC are assigned an external IP address (to which the Endpoint/DNS Name resolves) that provides connectivity from EC2 or the Internet. In Amazon VPC, Amazon RDS DB instances only have a private IP address (within a subnet that you define). You can configure a VPC to make an Amazon RDS DB instance in that VPC publicly accessible; see the VPC documentation for more information. For more information about the differences between EC2-Classic and EC2-VPC, see the EC2 documentation.

Q: What is a DB Subnet Group and why do I need one?

A DB Subnet Group is a collection of subnets that you may want to designate for your RDS DB Instances in a VPC. Each DB Subnet Group should have at least one subnet for every Availability Zone in a given Region. When creating a DB Instance in VPC, you will need to select a DB Subnet Group. Amazon RDS then uses that DB Subnet Group and your preferred Availability Zone to select a subnet and an IP address within that subnet. Amazon RDS creates and associates an Elastic Network Interface to your DB Instance with that IP address.

Please note that, we strongly recommend you use the DNS Name to connect to your DB Instance as the underlying IP address can change (e.g., during a failover).

For Multi-AZ deployments, defining a subnet for all Availability Zones in a Region will allow Amazon RDS to create a new standby in another Availability Zone should the need arise. You need to do this even for Single-AZ deployments, just in case you want to convert them to Multi-

AZ deployments at some point.

Q: How do I create an Amazon RDS DB Instance in VPC?

For a walk through example of creating a DB Instance in VPC, refer to theAmazon RDS User Guide.

Following are the pre-requisites necessary to create a DB Instances within a VPC:

- You need to have a VPC set up with at least one subnet created in every Availability Zone in the Region you want to deploy your DB Instance. For information on creating Amazon VPC and subnets refer to the Getting Started Guide for Amazon VPC.
- You need to have a DB Subnet Group defined for your VPC.
- You need to have a DB Security Group defined for your VPC (or you can use the default provided).
- In addition, you should allocate adequately large CIDR blocks to each of your subnets so that there are enough spare IP addresses for Amazon RDS to use during maintenance activities including scale compute and failover etc.

Q: How do I control network access to my DB Instance(s)?

Visit the Security Groups section of the Amazon RDS User Guide to learn about the different ways to control access to your DB Instances.

Q: How do I secure Amazon RDS DB Instances running within my VPC?

VPC Security Groups can be used to help secure DB Instances within an Amazon VPC. In addition, network traffic entering and exiting each subnet can be allowed or denied via network Access Control Lists (ACLs). All network traffic entering or exiting your VPC via your IPsec VPN connection can be inspected by your on-premise security infrastructure, including network firewalls, intrusion detection and prevention systems.

Q: How do I connect to an RDS DB Instance in VPC?

DB Instances deployed within a VPC can be accessed by EC2 Instances deployed in the same VPC. If these EC2 Instances are deployed in a public subnet with associated Elastic IPs, you can access the EC2 Instances via the internet.

DB Instances deployed within a VPC can be accessed from the Internet or from EC2 Instances outside the VPC via VPN or bastion hosts that you can launch in your public subnet, or using Amazon RDS's Publicly Accessible option:

- To use a bastion host, you will need to set up a public subnet with an EC2 instance that acts as a SSH Bastion. This public subnet must have an internet gateway and routing rules that allow traffic to be directed via the SSH host, which must then forward requests to the private

IP address of your RDS DB instance.

- To use public connectivity, simply create your DB Instances with the Publicly Accessible option set to yes. With Publicly Accessible active, your DB Instances within a VPC will be fully accessible outside your VPC by default. This means you do not need to configure a VPN or bastion host to allow access to your instances.

You can also set up a VPN Gateway that extends your corporate network into your VPC, and allows access to the RDS DB instance in that VPC. Refer to the Amazon VPC User Guide for more details.

We strongly recommend you use the DNS Name to connect to your DB Instance as the underlying IP address can change (e.g., during failover).

Q: Can I move my existing DB instances outside VPC into my VPC?

If your DB instance is not in a VPC, you can use the AWS Management Console to easily move your DB instance into a VPC. See the Amazon RDS User Guide for more details. You can also take a snapshot of your DB Instance outside VPC and restore it to VPC by specifying the DB Subnet Group you want to use. Alternatively, you can perform a “Restore to Point in Time” operation as well.

Q: Can I move my existing DB instances from inside VPC to outside VPC?

Migration of DB Instances from inside to outside VPC is not supported. For security reasons, a DB Snapshot of a DB Instance inside VPC cannot be restored to outside VPC. The same is true with “Restore to Point in Time” functionality. If you need to move your DB Instance from inside to outside VPC, you will need to export your data from your source DB Instance in your VPC to your target DB Instance deployed outside VPC.

Q: What precautions should I take to ensure that my DB Instances in VPC are accessible by my application?

You are responsible for modifying routing tables and networking ACLs in your VPC to ensure that your DB instance is reachable from your client instances in the VPC.

For Multi-AZ deployments, after a failover, your client EC2 instance and RDS DB Instance may be in different Availability Zones. You should configure your networking ACLs to ensure that cross-AZ communication is possible.

Q: Can I change the DB Subnet Group of my DB Instance?

An existing DB Subnet Group can be updated to add more subnets, either for existing Availability Zones or for new Availability Zones added since the creation of the DB Instance. Removing subnets from an existing DB Subnet Group can cause unavailability for instances if they are running in a particular AZ that gets removed from the subnet group.

At the present time, updating an existing DB Subnet Group does not change the current subnet of the deployed DB instance; an instance-type scale operation is required. Explicitly changing the DB Subnet Group of a deployed DB instance is not currently allowed.

Q: What is an Amazon RDS master user account and how is it different from an AWS account?

To begin using Amazon RDS you will need an AWS developer account. If you do not have one prior to signing up for Amazon RDS, you will be prompted to create one when you begin the sign-up process. A master user account is different from an AWS developer account and used only within the context of Amazon RDS to control access to your DB Instance(s). The master user account is a native database user account which you can use to connect to your DB Instance. You can specify the master user name and password you want associated with each DB Instance when you create the DB Instance. Once you have created your DB Instance, you can connect to the database using the master user credentials. Subsequently, you may also want to create additional user accounts so that you can restrict who can access your DB Instance.

Q: What privileges are granted to the master user for my DB Instance?

For MySQL, the default privileges for the master user include: create, drop, references, event, alter, delete, index, insert, select, update, create temporary tables, lock tables, trigger, create view, show view, alter routine, create routine, execute, trigger, create user, process, show databases, grant option.

For Oracle, the master user is granted the "dba" role. The master user inherits most of the privileges associated with the role. Please refer to the Amazon RDS User Guide for the list of restricted privileges and the corresponding alternatives to perform administrative tasks that may require these privileges.

For SQL Server, a user that creates a database is granted the "db_owner" role. Please refer to the Amazon RDS User Guide for the list of restricted privileges and the corresponding alternatives to perform administrative tasks that may require these privileges.

Q: Is there anything different about user management with Amazon RDS?

No, everything works the way you are familiar with when using a relational database you manage yourself.

Q: Can programs running on servers in my own data center access Amazon RDS databases?

Yes. You have to intentionally turn on the ability to access your database over the internet by configuring Security Groups. You can authorize access for only the specific IPs, IP ranges, or subnets corresponding to servers in your own data center.

Q: Can I encrypt connections between my application and my DB Instance using SSL?

Yes, this option is currently supported for the MySQL, MariaDB, SQL Server, PostgreSQL, and Oracle engines.

Amazon RDS generates an SSL certificate for each DB Instance. Once an encrypted connection is established, data transferred between the DB Instance and your application will be encrypted during transfer.

While SSL offers security benefits, be aware that SSL encryption is a compute-intensive operation and will increase the latency of your database connection. SSL support within Amazon RDS is for encrypting the connection between your application and your DB Instance; it should not be relied on for authenticating the DB Instance itself.

For details on establishing an encrypted connection with Amazon RDS, please visit Amazon RDS's MySQL User Guide, MariaDB User Guide, SQL Server User Guide, PostgreSQL User Guide or Oracle User Guide. To learn more about how SSL works with these engines, you can refer directly to the MySQL documentation, the MariaDB documentation, the MSDN SQL Server documentation, the PostgreSQL documentation, or the Oracle Documentation.

Q: Can I encrypt data at rest on my Amazon RDS databases?

Amazon RDS supports encryption at rest for all database engines, using keys you manage using AWS Key Management Service (KMS). On a database instance running with Amazon RDS encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, read replicas, and snapshots. Encryption and decryption are handled transparently. For more information about the use of KMS with Amazon RDS, see the Amazon RDS User's Guide.

At the present time, encrypting an existing DB Instance is not supported. To use Amazon RDS encryption for an existing database, create a new DB Instance with encryption enabled and migrate your data into it.

Amazon RDS for Oracle and SQL Server support those engines' Transparent Data Encryption technologies. Transparent Data Encryption in Oracle is integrated with AWS CloudHSM, which allows you to securely generate, store, and manage your cryptographic keys in single-tenant Hardware Security Module (HSM) appliances within the AWS cloud. For more information, see the Amazon RDS User's Guide sections on Oracle and SQL Server.

Q: How do I control the actions that my systems and users can take on specific RDS resources?

You can control the actions that your AWS IAM users and groups can take on RDS resources. You do this by referencing the RDS resources in the AWS IAM policies that you apply to your users and groups. RDS resources that can be referenced in an AWS IAM policy includes DB Instances, DB Snapshots, Read Replicas, DB Security Groups, DB Option Groups, DB

Parameter Groups, Event Subscriptions and DB Subnet Groups. In addition, you can tag these resources to add additional metadata to your resources. By using tagging, you can categorize your resources (e.g. "Development" DB Instances, "Production" DB Instances, "Test" DB Instances etc), and write AWS IAM policies that list the permissions (i.e. actions) that can taken on resources with the same tags. For more information, refer to Managing Access to Your Amazon RDS Resources and Databases and Tagging Amazon RDS Resources

Q: I wish to perform security analysis or operational troubleshooting on my RDS deployment. Can I get a history of all RDS API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console home page.

Database Configuration

Q: How do I choose the right configuration parameters for my DB Instance(s)?

By default, Amazon RDS chooses the optimal configuration parameters for your DB Instance taking into account the instance class and storage capacity. However, if you want to change them, you can do so using the AWS Management Console, the Amazon RDS APIs, or the AWS Command Line Interface. Please note that changing configuration parameters from recommended values can have unintended effects, ranging from degraded performance to system crashes, and should only be attempted by advanced users who wish to assume these risks.

Q: What are DB Parameter groups? How are they helpful?

A database parameter group (DB Parameter Group) acts as a “container” for engine configuration values that can be applied to one or more DB Instances. If you create a DB Instance without specifying a DB Parameter Group, a default DB Parameter Group is used. This default group contains engine defaults and Amazon RDS system defaults optimized for the DB Instance you are running. However, if you want your DB Instance to run with your custom-specified engine configuration values, you can simply create a new DB Parameter Group, modify the desired parameters, and modify the DB Instance to use the new DB Parameter Group. Once associated, all DB Instances that use a particular DB Parameter Group get all the parameter updates to that DB Parameter Group.

For more information on configuring DB Parameter Groups, please read theAmazon RDS User Guide.

Q: How can I monitor the configuration of my Amazon RDS resources?

You can use AWS Config to continuously record configurations changes to Amazon RDS DB Instances, DB Subnet Groups, DB Snapshots, DB Security Groups, and Event Subscriptions and receive notification of changes through Amazon Simple Notification Service (SNS). You can also create AWS Config Rules to evaluate whether these RDS resources have the desired configurations.

Multi-AZ Deployments and Read Replicas

Q: What types of replication does Amazon RDS support and when should I use each?

Amazon RDS provides two distinct replication options to serve different purposes.

If you are looking to use replication to increase database availability while protecting your latest database updates against unplanned outages, consider running your DB Instance as a Multi-AZ deployment. When you create or modify your DB Instance to run as a Multi-AZ deployment, Amazon RDS will automatically provision and manage a “standby” replica in a different Availability Zone (independent infrastructure in a physically separate location). In the event of planned database maintenance, DB Instance failure, or an Availability Zone failure, Amazon RDS will automatically failover to the standby so that database operations can resume quickly without administrative intervention. Multi-AZ deployments utilize synchronous replication, making database writes concurrently on both the primary and standby so that the standby will be up-to-date in the event a failover occurs. While our technological implementation for Multi-AZ DB Instances maximizes data durability in failure scenarios, it precludes the standby from being accessed directly or used for read operations. The fault tolerance offered by Multi-AZ deployments make them a natural fit for production environments.

To help you to scale beyond the capacity constraints of a single DB Instance for read-heavy database workloads, Amazon RDS offers Read Replicas. You can create a Read Replica of a given source DB Instance using the AWS Management Console, the RDS API, or the AWS Command Line Interface. Once the Read Replica is created, database updates on the source DB Instance will be propagated to the Read Replica. You can create multiple Read Replicas for a given source DB Instance and distribute your application’s read traffic amongst them.

Read Replicas are supported by Amazon RDS for MySQL and PostgreSQL. Unlike Multi-AZ deployments, Read Replicas for these engines use each's built-in replication technology and are subject to its strengths and limitations. In particular, updates are applied to your Read Replica(s) after they occur on the source DB Instance (“asynchronous” replication), and replication lag can vary significantly. This means recent database updates made to a standard (non Multi-AZ) source DB Instance may not be present on associated Read Replicas in the event of an unplanned outage on the source DB Instance. As such, Read Replicas do not offer the same data durability benefits as Multi-AZ deployments. While Read Replicas can provide some read availability benefits, they and are not designed to improve write availability.

You can use Multi-AZ deployments and Read Replicas in conjunction to enjoy the complementary benefits of each. You can simply specify that a given Multi-AZ deployment is the source DB Instance for your Read Replica(s). That way you gain both the data durability and availability benefits of Multi-AZ deployments and the read scaling benefits of Read Replicas.

Multi-AZ Deployments

Q: What does it mean to run a DB Instance as a Multi-AZ deployment?

When you create or modify your DB Instance to run as a Multi-AZ deployment, Amazon RDS automatically provisions and maintains a synchronous “standby” replica in a different Availability Zone. Updates to your DB Instance are synchronously replicated across Availability Zones to the standby in order to keep both in sync and protect your latest database updates against DB Instance failure. During certain types of planned maintenance, or in the unlikely event of DB Instance failure or Availability Zone failure, Amazon RDS will automatically failover to the standby so that you can resume database writes and reads as soon as the standby is promoted. Since the name record for your DB Instance remains the same, your application can resume database operation without the need for manual administrative intervention. With Multi-AZ deployments, replication is transparent: you do not interact directly with the standby, and it cannot be used to serve read traffic. More information about Multi-AZ deployments is in the Amazon RDS User Guide.

Q: What is an Availability Zone?

Availability Zones are distinct locations within a Region that are engineered to be isolated from failures in other Availability Zones. Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone. Availability Zones within the same Region benefit from low-latency network connectivity.

Q: What do “primary” and “standby” mean in the context of a Multi-AZ deployment?

When you run a DB Instance as a Multi-AZ deployment, the “primary” serves database writes and reads. In addition, Amazon RDS provisions and maintains a “standby” behind the scenes, which is an up-to-date replica of the primary. The standby is “promoted” in failover scenarios. After failover, the standby becomes the primary and accepts your database operations. You do not interact directly with the standby (e.g. for read operations) at any point prior to promotion. If you are interested in scaling read traffic beyond the capacity constraints of a single DB Instance, please see the FAQs on Read Replicas.

Q: What are the benefits of a Multi-AZ deployment?

The chief benefits of running your DB Instance as a Multi-AZ deployment are enhanced database durability and availability. The increased availability and fault tolerance offered by Multi-AZ deployments make them a natural fit for production environments.

Running your DB Instance as a Multi-AZ deployment safeguards your data in the unlikely event of a DB Instance component failure or loss of availability in one Availability Zone. For example, if a storage volume on your primary fails, Amazon RDS automatically initiates a failover to the standby, where all of your database updates are intact. This provides additional data durability relative to standard deployments in a single AZ, where a user-initiated restore operation would be required and updates that occurred after the latest restorable time (typically within the last five minutes) would not be available.

You also benefit from enhanced database availability when running your DB Instance as a Multi-AZ deployment. If an Availability Zone failure or DB Instance failure occurs, your availability impact is limited to the time automatic failover takes to complete. The availability benefits of Multi-AZ also extend to planned maintenance. For example, with automated backups, I/O activity is no longer suspended on your primary during your preferred backup window, since backups are taken from the standby. In the case of patching or DB Instance class scaling, these operations occur first on the standby, prior to automatic fail over. As a result, your availability impact is limited to the time required for automatic failover to complete.

Another implied benefit of running your DB Instance as a Multi-AZ deployment is that DB Instance failover is automatic and requires no administration. In an Amazon RDS context, this means you are not required to monitor DB Instance events and initiate manual DB Instance recovery (via the RestoreDBInstanceToPointInTime or RestoreDBInstanceFromSnapshot APIs) in the event of an Availability Zone failure or DB Instance failure.

Q: Are there any performance implications of running my DB Instance as a Multi-AZ deployments?

You may observe elevated latencies relative to a standard DB Instance deployment in a single Availability Zone as a result of the synchronous data replication performed on your behalf.

Q: When running my DB Instance as a Multi-AZ deployment, can I use the standby for read or write operations?

No, the standby replica cannot serve read requests. Multi-AZ deployments are designed to provide enhanced database availability and durability, rather than read scaling benefits. As such, the feature uses synchronous replication between primary and standby. Our implementation makes sure the primary and the standby are constantly in sync, but precludes using the standby for read or write operations. If you are interested in a read scaling solution, please see the FAQs on Read Replicas.

Q: How do I set up a Multi-AZ DB Instance deployment?

In order to create a Multi-AZ DB Instance deployment, simply click the “Yes” option for “Multi-AZ Deployment” when launching a DB Instance with the AWS Management Console. Alternatively, if you are using the Amazon RDS APIs, you would call the CreateDBInstance API and set the “Multi-AZ” parameter to the value “true.” To convert an existing standard (single AZ) DB Instance to Multi-AZ, modify the DB Instance in the AWS Management Console or use the ModifyDBInstance API and set the Multi-AZ parameter to true.

Q: What happens when I convert my RDS instance from Single-AZ to Multi-AZ?

For the RDS MySQL, MariaDB, PostgreSQL and Oracle database engines, when you elect to convert your RDS instance from Single-AZ to Multi-AZ, the following happens:

- A snapshot of your primary instance is taken
- A new standby instance is created in a different Availability Zone, from the snapshot
- Synchronous replication is configured between primary and standby instances

As such, there should be no downtime incurred when an instance is converted from Single-AZ to Multi-AZ.

Q: What events would cause Amazon RDS to initiate a failover to the standby replica?

Amazon RDS detects and automatically recovers from the most common failure scenarios for Multi-AZ deployments so that you can resume database operations as quickly as possible without administrative intervention. Amazon RDS automatically performs a failover in the event of any of the following:

- Loss of availability in primary Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary
- Storage failure on primary

Note: When operations such as DB Instance scaling or system upgrades like OS patching are initiated for Multi-AZ deployments, for enhanced availability, they are applied first on the standby prior to an automatic failover. As a result, your availability impact is limited only to the time required for automatic failover to complete. Note that Amazon RDS Multi-AZ deployments do not failover automatically in response to database operations such as long running queries, deadlocks or database corruption errors.

Q: Will I be alerted when automatic failover occurs?

Yes, Amazon RDS will emit a DB Instance event to inform you that automatic failover occurred.

You can click the “Events” section of the Amazon RDS Console or use the DescribeEvents API to return information about events related to your DB Instance. You can also use Amazon RDS Event Notifications to be notified when specific DB events occur.

Q: What happens during Multi-AZ failover and how long does it take?

Failover is automatically handled by Amazon RDS so that you can resume database operations as quickly as possible without administrative intervention. When failing over, Amazon RDS simply flips the canonical name record (CNAME) for your DB Instance to point at the standby, which is in turn promoted to become the new primary. We encourage you to follow best practices and implement database connection retry at the application layer.

Failovers, as defined by the interval between the detection of the failure on the primary and the resumption of transactions on the standby, typically complete within one to two minutes. Failover time can also be affected by whether large uncommitted transactions must be recovered; the use of adequately large instance types is recommended with Multi-AZ for best results. AWS also recommends the use of Provisioned IOPS with Multi-AZ instances, for fast, predictable, and consistent throughput performance.

Q: Can I initiate a “forced failover” for my Multi-AZ DB Instance deployment?

Amazon RDS will automatically failover without user intervention under a variety offailure conditions. In addition, Amazon RDS provides an option to initiate a failover when rebooting your instance. You can access this feature via the AWS Management Console or when using the RebootDBInstance API call.

Q: How do I control/configure Multi-AZ synchronous replication?

With Multi-AZ deployments, you simply set the “Multi-AZ” parameter to true. The creation of the standby, synchronous replication, and failover are all handled automatically. This means you cannot select the Availability Zone your standby is deployed in or alter the number of standbys available (Amazon RDS provisions one dedicated standby per DB Instance primary). The standby also cannot be configured to accept database read activity. Learn more about Multi-AZ configurations.

Q: Will my standby be in the same Region as my primary?

Yes. Your standby is automatically provisioned in a different Availability Zone of the*same Region* as your DB Instance primary.

Q: Can I see which Availability Zone my primary is currently located in?

Yes, you can gain visibility into the location of the current primary by using theAWS Management Console or DescribeDBInstances API.

Q: After failover, my primary is now located in a different Availability Zone than my other AWS resources (e.g. EC2 instances). Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same Region. In addition, you may want to consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone failure. Multi-AZ deployments address this need for the database tier without administration on your part.

Q: How do DB Snapshots and automated backups work with my Multi-AZ deployment?

You interact with automated backup and DB Snapshot functionality in the same way whether you are running a standard deployment in a Single-AZ or Multi-AZ deployment. If you are running a Multi-AZ deployment, automated backups and DB Snapshots are simply taken from the standby to avoid I/O suspension on the primary. Please note that you may experience increased I/O latency (typically lasting a few minutes) during backups for both Single-AZ and Multi-AZ deployments.

Initiating a restore operation (point-in-time restore or restore from DB Snapshot) also works the same with Multi-AZ deployments as standard, Single-AZ deployments. New DB Instance deployments can be created with either the RestoreDBInstanceFromSnapshot or RestoreDBInstanceToPointInTime APIs. These new DB Instance deployments can be either standard or Multi-AZ, regardless of whether the source backup was initiated on a standard or Multi-AZ deployment.

Read Replicas

Q: What does it mean to run a DB Instance as a Read Replica?

Read Replicas make it easy to take advantage of supported engines' built-in replication functionality to elastically scale out beyond the capacity constraints of a single DB Instance for read-heavy database workloads. You can create a Read Replica with a few clicks in the AWS Management Console or using the CreateDBInstanceReadReplica API. Once the Read Replica is created, database updates on the source DB Instance will be replicated using a supported engine's native, asynchronous replication. You can create multiple Read Replicas for a given source DB Instance and distribute your application's read traffic amongst them. Since Read Replicas use supported engines' built-in replication, they are subject to its strengths and limitations. In particular, updates are applied to your Read Replica(s) after they occur on the source DB Instance, and replication lag can vary significantly. Read Replicas can be associated with Multi-AZ deployments to gain read scaling benefits in addition to the enhanced database write availability and data durability provided by Multi-AZ deployments.

Q: When would I want to consider using an Amazon RDS Read Replica?

There are a variety of scenarios where deploying one or more Read Replicas for a given source DB Instance may make sense. Common reasons for deploying a Read Replica include:

- Scaling beyond the compute or I/O capacity of a single DB Instance for read-heavy database workloads. This excess read traffic can be directed to one or more Read Replicas.
- Serving read traffic while the source DB Instance is unavailable. If your source DB Instance cannot take I/O requests (e.g. due to I/O suspension for backups or scheduled maintenance), you can direct read traffic to your Read Replica(s). For this use case, keep in mind that the data on the Read Replica may be “stale” since the source DB Instance is unavailable.
- Business reporting or data warehousing scenarios; you may want business reporting queries to run against a Read Replica, rather than your primary, production DB Instance.

Q: Do I need to enable automatic backups on my DB Instance before I can create read replicas?

Yes. Enable automatic backups on your DB Instance before adding Read Replicas, by setting the backup retention period to a value other than 0. Backups must remain enabled for Read Replicas to work.

Q: Which versions of database engines support Amazon RDS Read Replicas?

Amazon RDS for MySQL: DB Instances with MySQL version 5.5 or newer support creation of Read Replicas. Automatic backups must be and remain enabled ou the source DB Instance for Read Replica operations. Automatic backups are supported only for Amazon RDS Read Replicas running MySQL 5.6 and later, not 5.5.

Amazon RDS for PostgreSQL: DB Instances with PostgreSQL version 9.3.5 or newer support creation of Read Replicas. Existing PostgreSQL instances prior to version 9.3.5 need to be upgraded to PostgreSQL version 9.3.5 to take advantage of Amazon RDS Read Replicas.

Amazon RDS for MariaDB: DB Instances with MariaDB 10.0 or newer support creation of Read Replicas. Automatic backups must be and remain enabled on the source DB Instance for Read Replica operations.

Q: How do I deploy a Read Replica for a given DB Instance?

You can create a Read Replica in minutes using the standard CreateDBInstanceReadReplica API or a few clicks on the AWS Management Console. When creating a Read Replica, you can identify it as a Read Replica by specifying a SourceDBInstanceIdentifier. The SourceDBInstanceIdentifier is the DB Instance Identifier of the “source” DB Instance from which you wish to replicate. As with a standard DB Instance, you can also specify the Availability Zone, DB Instance class, and preferred maintenance window. The engine version (e.g., PostgreSQL 9.3.5) and storage allocation of a Read Replica is inherited from the source DB Instance. When you initiate the creation of a Read Replica, Amazon RDS takes a snapshot of your source DB Instance and begins replication. As a result, you will experience a brief I/O suspension on your source DB Instance as the snapshot occurs. The I/O suspension typically lasts on the order of one minute, and is avoided if the source DB Instance is a Multi-AZ deployment (in the case of Multi-AZ deployments, snapshots are taken from the standby). Amazon RDS is also currently working on an optimization (to be released shortly) such that if you create multiple Read Replicas within a 30 minute window, all of them will use the same source snapshot to minimize I/O impact (“catch-up” replication for each Read Replica will begin after creation).

Q: How do I connect to my Read Replica(s)?

You can connect to a Read Replica just as you would connect to a standard DB Instance, using the DescribeDBInstance API or AWS Management Console to retrieve the endpoint(s) for you Read Replica(s). If you have multiple Read Replicas, it is up to your application to determine how read traffic will be distributed amongst them.

Q: How many Read Replicas can I create for a given source DB Instance?

Amazon RDS for MySQL, MariaDB and PostgreSQL currently allow you to create up to five (5) Read Replicas for a given source DB Instance.

Q: Can I create a Read Replica in an AWS Region different from that of the source DB Instance?

Amazon RDS for MySQL, MariaDB and PostgreSQL supports cross-region Read Replicas.

Q: Do Amazon RDS Read Replicas support synchronous replication?

No. Read Replicas in Amazon RDS for MySQL, MariaDB and PostgreSQL are implemented using those engines' native asynchronous replication.

Q: Can I use a Read Replica to enhance database write availability or protect the data on my source DB Instance against failure scenarios?

If you are looking to use replication to increase database write availability and protect recent database updates against various failure conditions, we recommend you run your DB Instance as a Multi-AZ deployment. With Amazon RDS Read Replicas, which employ supported engines' native, asynchronous replication, database writes occur on a Read Replica after they have already occurred on the source DB Instance, and this replication “lag” can vary significantly. In contrast, the replication used by Multi-AZ deployments is synchronous, meaning that all database writes are concurrent on the primary and standby. This protects your latest database updates, since they should be available on the standby in the event a failover is required. In addition, with Multi-AZ deployments replication is fully managed. Amazon RDS automatically monitors for DB Instance failure conditions or Availability Zone failure and initiates automatic failover to the standby if an outage occurs.

Q: Can I create a Read Replica with a Multi-AZ DB Instance deployment as its source?

Yes. Since Multi-AZ DB Instances address a different need than Read Replicas, it makes sense to use the two in conjunction for production deployments and to associate a Read Replica with a

Multi-AZ DB Instance deployment. The “source” Multi AZ-DB Instance provides you with enhanced write availability and data durability, and the associated Read Replica would improve read traffic scalability.

Q: Can I make my Amazon RDS Read Replicas themselves Multi-AZ?

Amazon RDS for MySQL, MariaDB and PostgreSQL do not presently support this.

Q: If my Read Replica(s) use a Multi-AZ DB Instance deployment as a source, what happens if Multi-AZ failover occurs?

In the event of a Multi-AZ failover, any associated and available Read Replicas should automatically resume replication once failover has completed (acquiring updates from the newly promoted primary).

Q: Can I create a Read Replica of another Read Replica?

Amazon RDS for MySQL and MariaDB: You can create a second-tier Read Replica from an existing first-tier Read Replica. By creating a second-tier Read Replica, you may be able to move some of the replication load from the master database instance to a first-tier Read Replica. Please note that a second-tier Read Replica may lag further behind the master because of additional replication latency introduced as transactions are replicated from the master to the first tier replica and then to the second-tier replica.

Amazon RDS for PostgreSQL: Read Replicas of Read Replicas are not currently supported.

Q: Can my Read Replicas only accept database read operations?

Read Replicas are designed to serve read traffic. However, there may be use cases where advanced users wish to complete Data Definition Language (DDL) SQL statements against a Read Replica. Examples might include adding a database index to a Read Replica that is used for business reporting, without adding the same index to the corresponding source DB Instance.

Amazon RDS for MySQL can be configured to permit DDL SQL statements against a Read Replica. If you wish to enable operations other than reads for a given Read Replica, modify the active DB Parameter Group for the Read Replica, setting the “read_only” parameter to “0.”

Amazon RDS for PostgreSQL does not currently support the execution of DDL SQL statements against a Read Replica.

Q: Can I promote my Read Replica into a “standalone” DB Instance?

Yes. Refer to the Amazon RDS User Guide for more details.

Q: Will my Read Replica be kept up-to-date with its source DB Instance?

Updates to a source DB Instance will automatically be replicated to any associated Read Replicas. However, with supported engines' asynchronous replication technology, a Read

Replica can fall behind its source DB Instance for a variety of reasons. Typical reasons include:

- Write I/O volume to the source DB Instance exceeds the rate at which changes can be applied to the Read Replica (this problem is particularly likely to arise if the compute capacity of a Read Replica is less than the source DB Instance)
- Complex or long-running transactions to the source DB Instance hold up replication to the Read Replica
- Network partitions or latency between the source DB Instance and a Read Replica

Read Replicas are subject to the strengths and weaknesses of supported engines' native replication. If you are using Read Replicas, you should be aware of the potential for lag between a Read Replica and its source DB Instance, or “inconsistency”. Click here for guidance on what to do if your Read Replica(s) fall significantly behind its source.

Q: How do I see the status of my active Read Replica(s)?

You can use the standard DescribeDBInstances API to return a list of all the DB Instances you have deployed (including Read Replicas), or simply click on the "DB Instances" tab of the Amazon RDS Console.

Amazon RDS allows you to gain visibility into how far a Read Replica has fallen behind its source DB Instance. The number of seconds that the Read Replica is behind the master is published as an Amazon CloudWatch metric ("Replica Lag") available via the AWS Management Console or Amazon CloudWatch APIs. For Amazon RDS for MySQL, the source of this information is the same as that displayed by issuing a standard "Show Slave Status" MySQL command against the Read Replica. For Amazon RDS for PostgreSQL, you can use the pg_stat_replication view on the source DB Instance to explore replication metrics.

Amazon RDS monitors the replication status of your Read Replicas and updates the Replication State field in the AWS Management console to "Error" if replication stops for any reason (e.g., attempting DML queries on your replica that conflict with the updates made on the master database instance could result in a replication error). You can review the details of the associated error thrown by the MySQL engine by viewing the Replication Error field and take an appropriate action to recover from it. You can learn more about troubleshooting replication issues in the Troubleshooting a Read Replica Problem section of the User Guide for Amazon RDS for MySQL or PostgreSQL.

If a replication error is fixed, the Replication State changes to Replicating.

Q: My Read Replica has fallen significantly behind its source DB Instance. What should I do?

As discussed in the previous questions, “inconsistency” or lag between a Read Replica and its source DB Instance is common with asynchronous replication. If an existing Read Replica has fallen too far behind to meet your requirements, you can delete it and create a new one with the same endpoint by using the same DB Instance Identifier and Source DB Instance Identifier as the deleted Read Replica. Keep in mind that the re-creation process will be counter-productive at small lag levels (e.g. under five minutes of lag), and should be used with prudence (i.e. only when the Read Replica is significantly behind its source DB Instance). Also keep in mind that replica lag may naturally grow and shrink over time, depending on your source DB Instance's steady-state usage pattern.

Scaling the DB Instance class of a Read Replica may reduce replication lag in some cases, particularly if your source DB Instance class is larger than your Read Replica DB Instance class. However, Read Replicas are not guaranteed to work in all cases. There may be scenarios and usage patterns where a Read Replica can never catch up with its source after initial creation, or otherwise remains too far behind its source to meet your use case requirements.

Q: I scaled the compute and/or storage capacity of my source DB Instance. Should I scale the resources for associated Read Replica(s) as well?

For replication to work effectively, we recommend that Read Replicas have as much or more compute and storage resources as their respective source DB Instances. Otherwise replication lag is likely to increase or your Read Replica may run out of space to store replicated updates.

Q: Can DB Snapshots or automated backups be taken of Read Replicas?

No. If you are looking to increase database write availability by taking backups from your Read Replica instead of its source DB Instance, you can accomplish the same objective by running your DB Instance as a Multi-AZ deployment. Backups will then be initiated from the Multi-AZ standby to minimize availability impact.

Q: How do I delete a Read Replica? Will it be deleted automatically if its source DB Instance is deleted?

You can easily delete a Read Replica with a few clicks of the AWS Management Console or by passing its DB Instance identifier to the DeleteDBInstance API.

An Amazon RDS for MySQL or MariaDB Read Replica will stay active and continue accepting read traffic even after its corresponding source DB Instance has been deleted. If you desire to delete the Read Replica in addition to the source DB Instance, you must explicitly do so using the DeleteDBInstance API or AWS Management Console.

If you delete an Amazon RDS for PostgreSQL DB Instance that has Read Replicas, all Read Replicas will be promoted to standalone DB Instances and will be able to accept both read and write traffic. The newly promoted DB Instances will operate independently of one another. If you desire to delete these DB Instances in addition to the original source DB Instance, you must explicitly do so using the DeleteDBInstance API or AWS Management Console.

Q: Can I directly access the event logs for my Database Instance?

With Amazon RDS for MySQL or Amazon RDS for MariaDB, you can use the mysqlbinlog utility to download or stream binary logs from your DB Instance. Amazon RDS for PostgreSQL does not currently provide access to the WAL files for your DB Instance.

Q: How much do Read Replicas cost? When does billing begin and end?

A Read Replica is billed as a standard DB Instance and at the same rates. Click here for more information on DB Instance billing visit this FAQ. Just like a standard DB Instance, the rate per “DB Instance hour” for a Read Replica is determined by the DB Instance class of the Read Replica – please see Amazon RDS detail page for up-to-date pricing. You are not charged for the data transfer incurred in replicating data between your source DB Instance and Read Replica.

Billing for a Read Replica begins as soon as the Read Replica has been successfully created (i.e. when status is listed as “active”). The Read Replica will continue being billed at standard Amazon RDS DB Instance hour rates until you issue a command to delete it.

Q: How does support for Read Replicas vary among the Amazon RDS engines that support this feature?

Read Replicas in both Amazon RDS for PostgreSQL, MySQL, and MariaDB allow you to elastically scale out beyond the capacity constraints of a single DB instance for read-heavy database workloads. There are similarities and differences in the implementations as they leverage native engine features. See the following table for details.

Feature	PostgreSQL	MySQL	MariaDB
Maximum Read Replicas allowed per source DB Instance	5	5	5
Replication method	Asynchronous Physical	Asynchronous Logical	Asynchronous Logical
Must automatic backups be enabled for Read Replica support?	Yes	Yes	Yes
Engine versions for which Read Replicas are available	9.3.5 or later	5.5 or later	10.0 or later
Promotion of Read Replica to a new standalone DB Instance	Supported	Supported	Supported
Creation of Indexes on Read Replica	Currently not supported	Supported	Supported
Creation of Backups of Read Replicas	Currently not supported	Supported	Supported
Chaining of Read Replicas (i.e., Read Replicas of Read Replicas)	Currently not supported	Supported	Supported
Cross-Region Read Replicas	Supported	Supported	Supported

For information about replication support with the Amazon Aurora engine, see theAmazon RDS for Aurora FAQ.

Enhanced Monitoring

Q: What is Enhanced Monitoring for RDS?

Enhanced Monitoring for RDS gives you deeper visibility into the health of your RDS instances. Just turn on the “Enhanced Monitoring” option for your RDS DB Instance and set a granularity and Enhanced Monitoring will collect vital operating system metrics and process information, at the defined granularity.

Q: Which metrics and processes can I monitor in Enhanced Monitoring?

Enhanced Monitoring captures your RDS instance system level metrics such as the CPU, memory, file system and disk I/O among others. The complete list of metrics can be found here.

Q: Which engines are supported by Enhanced Monitoring?

Enhanced Monitoring supports all RDS database engines.

Q: Which instance types are supported by Enhanced Monitoring?

Enhanced Monitoring supports every instance type except t1.micro and m1.small. The software uses a small amount of CPU, memory and I/O and for general purpose monitoring, we recommend switching on higher granularities for instances that are medium or larger. For non-production DB Instances, the default setting for Enhanced Monitoring is “off”, and you have the choice of leaving it disabled or modifying the granularity when it is on.

Q: What information can I view on the RDS dashboard?

You can view all the system metrics and process information for your RDS DB Instances in a graphical format on the console. You can manage which metrics you want to monitor for each instance and customize the dashboard according to your requirements.

Q: Will all the instances in my RDS account sample metrics at the same granularity?

No. You can set different granularities for each DB Instance in your RDS account. You can also choose the instances on which you want to enable Enhanced Monitoring as well as modify the granularity of any instance whenever you want.

Q: How far back can I see the historical metrics on the RDS console?

You can see the performance values for all the metrics for up to 1 hour ago, at a granularity of up to 1 second based on your setting.

Q: How can I visualize the metrics generated by RDS Enhanced Monitoring in CloudWatch?

The metrics from RDS Enhanced Monitoring are delivered into your CloudWatch Logs account. You can create metrics filters in CloudWatch from CloudWatch Logs and display the graphs on the CloudWatch dashboard. For more details, please visit the Amazon CloudWatch page.

Q: When should I use CloudWatch instead of the RDS console dashboard?

You should use CloudWatch if you want to view historical data beyond what is available on the RDS console dashboard. You can monitor your RDS instances in CloudWatch to diagnose the health of your entire AWS stack in a single location. Currently, CloudWatch supports granularities of up to 1 minute and the values will be averaged out for granularities less than that.

Q: Can I set up alarms and notifications based on specific metrics?

Yes. You can create an alarm in CloudWatch that sends a notification when the alarm changes state. The alarm watches a single metric over a time period that you specify, and performs one or more actions based on the value of the metric relative to the specified threshold over a number of time periods. For more details on CloudWatch alarms, please visit the Amazon CloudWatch Developer Guide.

Q: How do I integrate Enhanced Monitoring with my tool that I currently use?

RDS Enhanced Monitoring provides a set of metrics formed as JSON payloads which are delivered into your CloudWatch Logs account. The JSON payloads are delivered at the granularity last configured for the RDS instance.

There are two ways you can consume the metrics via a third-party dashboard or application. Monitoring tools can use CloudWatch Logs Subscriptions to set up a near real time feed for the metrics. Alternatively, you can use filters in CloudWatch Logs to bridge metrics across to CloudWatch to and integrate your application with CloudWatch. Please visit Amazon CloudWatch Documentation for more details.

Q: How can I delete historical data?

Since Enhanced Monitoring delivers JSON payloads into a log in your CloudWatch Logs account, you can control its retention period just like any other CloudWatch Logs stream. The default retention period configured for Enhanced Monitoring in CloudWatch Logs is 30 days. For details on how to change retention settings, please visit Amazon CloudWatch Developer Guide.

Q: What impact does Enhanced Monitoring have on my monthly bills?

Since the metrics are ingested into CloudWatch Logs, your charges will be based on CloudWatch Logs data transfer and storage rates once you exceed CloudWatch Logs free tier. Pricing details can be found here. The amount of information transferred for an RDS instance is directly proportional to the defined granularity for the Enhanced Monitoring feature. Administrators can set different granularities for different instances in their accounts to manage costs.

The approximate volume of data ingested into CloudWatch Logs by Enhanced Monitoring for an instance is as shown below:

Granularity	60 seconds	30 seconds	15 seconds	10 seconds	5 seconds	1 second
Data ingested in CloudWatch Logs* (GB per month)	0.27	0.53	1.07	1.61	3.21	16.07

Amazon Database Migration Service FAQ

FAQs

Q: Will AWS Database Migration Service help me convert my Oracle PL/SQL and SQL Server T-SQL code to Amazon Aurora or MySQL and PostgreSQL stored procedures?

Yes, AWS Database Migration Service already provides a schema conversion tool to help convert Oracle PL/SQL and SQL Server T-SQL code to equivalent code in the Amazon Aurora / MySQL dialect of SQL. When a code fragment cannot be automatically converted to the target language, AWS Database Migration Service will clearly document all locations that require manual input from the application developer. Support for PostgreSQL code conversion targets is coming soon.

Q: How do I get started with AWS Database Migration Service?

Getting started with AWS Database Migration Service is quick and simple. Most data replication tasks can be set up in less than 10 minutes. Visit the AWS Database Migration Service section of the AWS Management Console and enter the Start Migration wizard. Specify your source and target endpoints, select an existing replication instance or create a new one, and accept the default schema mapping rules or define your own transformations. Data replication will start immediately after you complete the wizard.

Q. In addition to one-time data migration, can I use AWS Database Migration Service for continuous data replication?

Yes, you can use AWS Database Migration Service for both one-time data migration into RDS and EC2-based databases as well as for continuous data replication. AWS Database Migration Service will capture changes on the source database and apply them in a transactionally-consistent way to the target. Continuous replication can be done from your data center to the databases in AWS or in the reverse, replicating to a database in your datacenter from a database in AWS. Ongoing continuous replication can also be done between homogenous or heterogeneous databases. For ongoing replication it would be preferable to use Multi-AZ for high-availability.

Q. What sources and targets does AWS Database Migration Service support?

AWS Database Migration Service supports both homogenous and heterogeneous data replication.

Supported database sources include: (1) Oracle, (2) SQL Server, (3) MySQL, (4) Amazon Aurora (5) PostgreSQL and (6) SAP ASE. All sources are supported on-premises, in EC2, and

RDS except Amazon Aurora which is available only in RDS.

RDS SQL Server and RDS Postgres sources are supported in bulk extract mode only; the change data capture mode (CDC) is not yet supported. Amazon Aurora is only available in RDS.

Supported database targets include: (1) Amazon Aurora, (2) Oracle, (3) SQL Server, (4) MySQL, (5) PostgreSQL and (6) SAP ASE. All Oracle, SQL Server, MySQL and Postgres targets are supported on-premises, in EC2 and RDS while SAP ASE is supported only in EC2.

Either the source or the target database (or both) need to reside in RDS or on EC2. Replication between on-premises to on-premises databases is not supported.

Q: Why should I use AWS Database Migration Service instead of my own self-managed replication solution?

AWS Database Migration Service is very easy to use. Replication tasks can be set up in minutes instead of hours or days, compared to the self-managed replication solutions that have to be installed and configured. AWS Database Migration Service monitors for replication tasks, network or host failures, and automatically provisions a host replacement in case of failures that can't be repaired.Users of AWS Database Migration Service don't have to overprovision capacity and invest in expensive hardware and replication software, as they typically have to do with self-managed solutions. With AWS Database Migration Service users can take advantage of on-demand pricing and scale their replication infrastructure up or down, depending on the load. AWS Database Migration Service data replication integrates tightly with the AWS Schema Conversion Tool, simplifying heterogeneous database migration projects.

Q. Can you summarize the database migration steps using AWS Database Migration Service for me?

During a typical simple database migration you will create a target database, migrate the database schema, setup the data replication process, initiate the full load and a subsequent change data capture and apply, and conclude with a switchover of your production environment to the new database once the target database is caught up with the source database.

Q. Are these steps different for continuous data replication?

The only difference is in the last step (the production environment switchover), which is absent for continuous data replication. Your data replication task will run until you change or terminate it.

Q. Can I monitor the progress of a database migration task?

Yes. AWS Database Migration Service has a variety of metrics displayed in the AWS Management Console. It provides an end-to-end view of the data replication process, including diagnostic and performance data for each point in the replication pipeline. AWS Database Migration Service also integrates with other AWS services such as CloudTrail and CloudWatch.

Customers can also leverage the AWS Database Migration Service API and CLI to integrate with their existing tools or build custom monitoring tools to suit their specific needs.

Q. How do I integrate AWS Database Migration Service with other applications?

AWS Database Migration Service provides a provisioning API that allows creating a replication task directly from your development environment, or scripting their creation at scheduled times during the day. The service API and CLI allows developers and database administrators to automate the creation, restart, management and termination of replication tasks.

Q. Can I replicate data from encrypted data sources?

Yes, AWS Database Migration Service can read and write from and to encrypted databases. AWS Database Migration Service connects to your database endpoints on the SQL interface layer. If you use the Transparent Data Encryption features of Oracle or SQL Server, AWS Database Migration Service will be able to extract decrypted data from such sources and replicate it to the target. The same applies to storage-level encryption. As long as AWS Database Migration Service has the correct credentials to the database source, it will be able to connect to the source and propagate data (in decrypted form) to the target. We recommend using encryption-at-rest on the target to maintain the confidentiality of your information. If you use application-level encryption, the data will be transmitted through AWS Database Migration Service as is, in encrypted format, and then inserted into the target database.

Q. Does AWS Database Migration Service migrate the database schema for me?

To quickly migrate a database schema to your target instance you can rely on the Basic Schema Copy feature of AWS Database Migration Service. Basic Schema Copy will automatically create tables and primary keys in the target instance if the target does not already contain tables with the same names. Basic Schema Copy is great for doing a test migration, or when you are migrating databases heterogeneously e.g. Oracle to MySQL or SQL Server to Oracle. Basic Schema Copy will not migrate secondary indexes, foreign keys or stored procedures. When you need to use a more customizable schema migration process (e.g. when you are migrating your production database and need to move your stored procedures and secondary database objects), you can use the AWS Schema Conversion Tool for heterogeneous migrations, or use the schema export tools native to the source engine, if you are doing homogenous migrations like (1) SQL Server Management Studio's Import and Export Wizard, (2) Oracle's SQL Developer Database Export tool or script the export using the dbms_metadata package, (3) MySQL's Workbench Migration Wizard.

Q. Can I copy my replication tasks from one environment to another?

Yes, you can copy the original task, change the source and target endpoints, and AWS Database Migration Service will remap the transformation rules to the new endpoints. This capability is very helpful when you want to promote your finished and tested replication tasks from the integration environment to your pre-prod and, later, to production environments with just a few clicks or one API call. It is also useful when you have many databases with the same schema.

Amazon DynamoDB FAQ

What is DynamoDB

Q: What is Amazon DynamoDB?

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. Amazon DynamoDB enables customers to offload the administrative burdens of operating and scaling distributed databases to AWS, so they don’t have to worry about hardware provisioning, setup and configuration, replication, software patching, or cluster scaling.

Q: What does Amazon DynamoDB manage on my behalf?

Amazon DynamoDB takes away one of the main stumbling blocks of scaling databases, the management of the database software and the provisioning of hardware needed to run it. Customers can deploy a non-relational database in a matter of minutes. DynamoDB automatically partitions and re-partitions your data and provisions additional server capacity as your table size grows or you increase your provisioned throughput. In addition, Amazon DynamoDB synchronously replicates data across three facilities in an AWS Region, giving you high availability and data durability.

Q: What does read consistency mean? Why should I care?

Amazon DynamoDB stores three geographically distributed replicas of each table to enable high availability and data durability. Read consistency represents the manner and timing in which the successful write or update of a data item is reflected in a subsequent read operation of that same item. Amazon DynamoDB exposes logic that enables you to specify the consistency characteristics you desire for each read request within your application.

Q: What is the consistency model of Amazon DynamoDB?

When reading data from Amazon DynamoDB, users can specify whether they want the read to be eventually consistent or strongly consistent:

Eventually Consistent Reads (Default) – the eventual consistency option maximizes your read throughput. However, an eventually consistent read might not reflect the results of a recently completed write. Consistency across all copies of data is usually reached within a second.

Repeating a read after a short time should return the updated data.

Strongly Consistent Reads — in addition to eventual consistency, Amazon DynamoDB also gives you the flexibility and control to request a strongly consistent read if your application, or an element of your application, requires it. A strongly consistent read returns a result that reflects all writes that received a successful response prior to the read.

Q: Does DynamoDB support in-place atomic updates?

Amazon DynamoDB supports fast in-place updates. You can increment or decrement a numeric attribute in a row using a single API call. Similarly, you can atomically add or remove to sets, lists, or maps. View our documentation for more information on atomic updates.

Q: Why is Amazon DynamoDB built on Solid State Drives?

Amazon DynamoDB runs exclusively on Solid State Drives (SSDs). SSDs help us achieve our design goals of predictable low-latency response times for storing and accessing data at any scale. The high I/O performance of SSDs also enables us to serve high-scale request workloads cost efficiently, and to pass this efficiency along in low request pricing.

Q: DynamoDB’s storage cost seems high. Is this a cost-effective service for my use case?

As with any product, we encourage potential customers of Amazon DynamoDB to consider the total cost of a solution, not just a single pricing dimension. The total cost of servicing a database workload is a function of the request traffic requirements and the amount of data stored. Most database workloads are characterized by a requirement for high I/O (high reads/sec and writes/sec) per GB stored. Amazon DynamoDB is built on SSD drives, which raises the cost per GB stored, relative to spinning media, but it also allows us to offer very low request costs. Based on what we see in typical database workloads, we believe that the total bill for using the SSD-based DynamoDB service will usually be lower than the cost of using a typical spinning media-based relational or non-relational database. If you have a use case that involves storing a large amount of data that you rarely access, then DynamoDB may not be right for you. We recommend that you use S3 for such use cases.

It should also be noted that the storage cost reflects the cost of storing multiple copies of each data item across multiple facilities within an AWS Region.

Q: Is DynamoDB only for high-scale applications?

No. DynamoDB offers seamless scaling so you can start small and scale up and down in line with your requirements. If you need fast, predictable performance at any scale then DynamoDB may be the right choice for you.

Getting Started

Q: How do I get started with Amazon DynamoDB?

Click “Sign Up” to get started with Amazon DynamoDB today. From there, you can begin interacting with Amazon DynamoDB using either the AWS Management Console or Amazon DynamoDB APIs. If you are using the AWS Management Console, you can create a table with Amazon DynamoDB and begin exploring with just a few clicks.

Q: What kind of query functionality does DynamoDB support?

Amazon DynamoDB supports GET/PUT operations using a user-defined primary key. The primary key is the only required attribute for items in a table and it uniquely identifies each item. You specify the primary key when you create a table. In addition to that DynamoDB provides flexible querying by letting you query on non-primary key attributes using Global Secondary Indexes and Local Secondary Indexes.

A primary key can either be a single-attribute partition key or a composite partition-sort key. A single attribute partition primary key could be, for example, “UserID”. This would allow you to quickly read and write data for an item associated with a given user ID.

A composite partition-sort key is indexed as a partition key element and a sort key element. This multi-part key maintains a hierarchy between the first and second element values. For example, a composite partition-sort key could be a combination of “UserID” (partition) and “Timestamp” (sort). Holding the partition key element constant, you can search across the sort key element to retrieve items. This would allow you to use the Query API to, for example, retrieve all items for a single UserID across a range of timestamps.

For more information on Global Secondary Indexing and its query capabilities, see the Secondary Indexes section in FAQ.

Q: How do I update and query data items with Amazon DynamoDB?

After you have created a table using the AWS Management Console or CreateTable API, you can use the PutItem or BatchWriteItem APIs to insert items. Then you can use the GetItem, BatchGetItem, or, if composite primary keys are enabled and in use in your table, the Query API to retrieve the item(s) you added to the table.

Q: Does Amazon DynamoDB support conditional operations?

Yes, you can specify a condition that must be satisfied for a put, update, or delete operation to be completed on an item . To perform a conditional operation, you can define a ConditionExpression that is constructed from the following:

- Boolean functions: ATTRIBUTE_EXIST, CONTAINS, and BEGINS_WITH
- Comparison operators: =, <>, , =, BETWEEN, and IN

- Logical operators: NOT, AND, and OR.

You can construct a free-form conditional expression that combines multiple conditional clauses, including nested clauses. Conditional operations allow users to implement optimistic concurrency control systems on DynamoDB. For more information on conditional operations, please see our documentation.

Q: Are expressions supported for key conditions?

Yes, you can specify an expression as part of the Query API call to filter results based on values of primary keys on a table using the KeyConditionExpression parameter.

Q: Are expressions supported for partition and partition-sort keys?

Yes, you can use expressions for both partition and partition-sort keys. Refer to the documentation page for more information on which expressions work on partition and partition-sort keys.

Q: Does Amazon DynamoDB support increment or decrement operations?

Yes, Amazon DynamoDB allows atomic increment and decrement operations on scalar values.

Q: When should I use Amazon DynamoDB vs a relational database engine on Amazon RDS or Amazon EC2?

Today’s web-based applications generate and consume massive amounts of data. For example, an online game might start out with only a few thousand users and a light database workload consisting of 10 writes per second and 50 reads per second. However, if the game becomes successful, it may rapidly grow to millions of users and generate tens (or even hundreds) of thousands of writes and reads per second. It may also create terabytes or more of data per day. Developing your applications against Amazon DynamoDB enables you to start small and simply dial-up your request capacity for a table as your requirements scale, without incurring downtime. You pay highly cost-efficient rates for the request capacity you provision, and let Amazon DynamoDB do the work over partitioning your data and traffic over sufficient server capacity to meet your needs. Amazon DynamoDB does the database management and administration, and you simply store and request your data. Automatic replication and failover provides built-in fault tolerance, high availability, and data durability. Amazon DynamoDB gives you the peace of mind that your database is fully managed and can grow with your application requirements.

While Amazon DynamoDB tackles the core problems of database scalability, management, performance, and reliability, it does not have all the functionality of a relational database. It does not support complex relational queries (e.g. joins) or complex transactions. If your workload requires this functionality, or you are looking for compatibility with an existing relational engine, you may wish to run a relational engine on Amazon RDS or Amazon EC2. While relational database engines provide robust features and functionality, scaling a workload beyond a single relational database instance is highly complex and requires significant time and expertise. As such, if you anticipate scaling requirements for your new application and do not need relational features, Amazon DynamoDB may be the best choice for you.

Q: How does Amazon DynamoDB differ from Amazon SimpleDB?

Which should I use? Both services are non-relational databases that remove the work of database administration. Amazon DynamoDB focuses on providing seamless scalability and fast, predictable performance. It runs on solid state disks (SSDs) for low-latency response times, and there are no limits on the request capacity or storage size for a given table. This is because Amazon DynamoDB automatically partitions your data and workload over a sufficient number of servers to meet the scale requirements you provide. In contrast, a table in Amazon SimpleDB has a strict storage limitation of 10 GB and is limited in the request capacity it can achieve (typically under 25 writes/second); it is up to you to manage the partitioning and re-partitioning of your data over additional SimpleDB tables if you need additional scale. While SimpleDB has scaling limitations, it may be a good fit for smaller workloads that require query flexibility. Amazon SimpleDB automatically indexes all item attributes and thus supports query flexibility at the cost of performance and scale.

Amazon CTO Werner Vogels' DynamoDB blog post provides additional context on the evolution of non-relational database technology at Amazon.

Q: When should I use Amazon DynamoDB vs Amazon S3?

Amazon DynamoDB stores structured data, indexed by primary key, and allows low latency read and write access to items ranging from 1 byte up to 400KB. Amazon S3 stores unstructured blobs and suited for storing large objects up to 5 TB. In order to optimize your costs across AWS services, large objects or infrequently accessed data sets should be stored in Amazon S3, while smaller data elements or file pointers (possibly to Amazon S3 objects) are best saved in Amazon DynamoDB.

Q: Can DynamoDB be used by applications running on any operating system?

Yes. DynamoDB is a fully managed cloud service that you access via API. DynamoDB can be used by applications running on any operating system (e.g. Linux, Windows, iOS, Android, Solaris, AIX, HP-UX, etc.). We recommend using the AWS SDKs to get started with DynamoDB. You can find a list of the AWS SDKs on our Developer Resources page. If you have trouble installing or using one of our SDKs, please let us know by posting to the relevant AWS Forum.

Data Models and APIs

Q: What is the Data Model?

The data model for Amazon DynamoDB is as follows:

Table: A table is a collection of data items – just like a table in a relational database is a collection of rows. Each table can have an infinite number of data items. Amazon DynamoDB is schema-less, in that the data items in a table need not have the same attributes or even the same number of attributes. Each table must have a primary key. The primary key can be a single attribute key or a "composite" attribute key that combines two attributes. The attribute(s) you designate as a primary key must exist for every item as primary keys uniquely identify each item within the table.

Item: An Item is composed of a primary or composite key and a flexible number of attributes. There is no explicit limitation on the number of attributes associated with an individual item, but the aggregate size of an item, including all the attribute names and attribute values, cannot exceed 400KB.

Attribute: Each attribute associated with a data item is composed of an attribute name (e.g. "Color") and a value or set of values (e.g. "Red" or "Red, Yellow, Green"). Individual attributes have no explicit size limit, but the total value of an item (including all attribute names and values) cannot exceed 400KB.

Q: Is there a limit on the size of an item?

The total size of an item, including attribute names and attribute values, cannot exceed 400KB.

Q: Is there a limit on the number of attributes an item can have?

There is no limit to the number of attributes that an item can have. However, the total size of an item, including attribute names and attribute values, cannot exceed 400KB.

Q: What are the APIs?

- CreateTable – Creates a table and specifies the primary index used for data access.
- UpdateTable – Updates the provisioned throughput values for the given table.
- DeleteTable – Deletes a table.
- DescribeTable – Returns table size, status, and index information.
- ListTables – Returns a list of all tables associated with the current account and endpoint.
- PutItem – Creates a new item, or replaces an old item with a new item (including all the attributes). If an item already exists in the specified table with the same primary key, the new item completely replaces the existing item. You can also use conditional operators to replace an item only if its attribute values match certain conditions, or to insert a new item only if that item doesn't already exist.
- BatchWriteItem – Inserts, replaces, and deletes multiple items across multiple tables in a single request, but not as a single transaction. Supports batches of up to 25 items to Put or

Delete, with a maximum total request size of 16 MB.

- UpdateItem – Edits an existing item's attributes. You can also use conditional operators to perform an update only if the item's attribute values match certain conditions.
- DeleteItem – Deletes a single item in a table by primary key. You can also use conditional operators to perform a delete an item only if the item's attribute values match certain conditions.
- GetItem – The GetItem operation returns a set of Attributes for an item that matches the primary key. The GetItem operation provides an eventually consistent read by default. If eventually consistent reads are not acceptable for your application, use ConsistentRead.
- BatchGetItem – The BatchGetItem operation returns the attributes for multiple items from multiple tables using their primary keys. A single response has a size limit of 16 MB and returns a maximum of 100 items. Supports both strong and eventual consistency.
- Query – Gets one or more items using the table primary key, or from a secondary index using the index key. You can narrow the scope of the query on a table by using comparison operators or expressions. You can also filter the query results using filters on non-key attributes. Supports both strong and eventual consistency. A single response has a size limit of 1 MB.
- Scan – Gets all items and attributes by performing a full scan across the table or a secondary index. You can limit the return set by specifying filters against one or more attributes.

Q: What is the consistency model of the Scan operation?

The Scan operation supports eventually consistent and consistent reads. By default, the Scan operation is eventually consistent. However, you can modify the consistency model using the optional ConsistentRead parameter in the Scan API call. Setting the ConsistentRead parameter to true will enable you make consistent reads from the Scan operation. For more information, read the documentation for the Scan operation.

Q: How does the Scan operation work?

You can think of the Scan operation as an iterator. Once the aggregate size of items scanned for a given Scan API request exceeds a 1 MB limit, the given request will terminate and fetched results will be returned along with a LastEvaluatedKey (to continue the scan in a subsequent operation).

Q: Are there any limitations for a Scan operation?

A Scan operation on a table or secondary index has a limit of 1MB of data per operation. After the 1MB limit, it stops the operation and returns the matching values up to that point, and a LastEvaluatedKey to apply in a subsequent operation, so that you can pick up where you left off.

Q: How many read capacity units does a Scan operation consume?

The read units required is the number of bytes fetched by the scan operation, rounded to the nearest 4KB, divided by 4KB. Scanning a table with consistent reads consumes twice the read capacity as a scan with eventually consistent reads.

Q: What data types does DynamoDB support?

DynamoDB supports four scalar data types: Number, String, Binary, and Boolean. Additionally, DynamoDB supports collection data types: Number Set, String Set, Binary Set, heterogeneous List and heterogeneous Map. DynamoDB also supports NULL values.

Q: What types of data structures does DynamoDB support?

DynamoDB supports key-value and document data structures.

Q: What is a key-value store?

A key-value store is a database service that provides support for storing, querying and updating collections of objects that are identified using a key and values that contain the actual content being stored.

Q: What is a document store?

A document store provides support for storing, querying and updating items in a document format such as JSON, XML, and HTML.

Q: Does DynamoDB have a JSON data type?

No, but you can use the document SDK to pass JSON data directly to DynamoDB. DynamoDB's data types are a superset of the data types supported by JSON. The document SDK will automatically map JSON documents onto native DynamoDB data types.

Q: Can I use the AWS Management Console to view and edit JSON documents?

Yes. The AWS Management Console provides a simple UI for exploring and editing the data stored in your DynamoDB tables, including JSON documents. To view or edit data in your table, please log in to the AWS Management Console, choose DynamoDB, select the table you want to view, then click on the “Explore Table” button.

Q: Is querying JSON data in DynamoDB any different?

No. You can create a Global Secondary Index or Local Secondary Index on any top-level JSON element. For example, suppose you stored a JSON document that contained the following information about a person: First Name, Last Name, Zip Code, and a list of all of their friends. First Name, Last Name and Zip code would be top-level JSON elements. You could create an index to let you query based on First Name, Last Name, or Zip Code. The list of friends is not a top-level element, therefore you cannot index the list of friends. For more information on Global

Secondary Indexing and its query capabilities, see the Secondary Indexes section in this FAQ.

Q: If I have nested JSON data in DynamoDB, can I retrieve only a specific element of that data?

Yes. When using the GetItem, BatchGetItem, Query, or Scan APIs, you can define a ProjectionExpression to determine which attributes should be retrieved from the table. Those attributes can include scalars, sets, or elements of a JSON document.

Q. If I have nested JSON data in DynamoDB, can I update only a specific element of that data?

Yes. When updating a DynamoDB item, you can specify the sub-element of the JSON document that you want to update.

Q:What is the Document SDK?

The Document SDK is a datatypes wrapper for JavaScript that allows easy interoperability between JS and DynamoDB datatypes. With this SDK, wrapping for requests will be handled for you; similarly for responses, datatypes will be unwrapped. For more information and downloading the SDK see our GitHub respository here.

Scalability, Availability & Durability

Q: Is there a limit to how much data I can store in Amazon DynamoDB?

No. There is no limit to the amount of data you can store in an Amazon DynamoDB table. As the size of your data set grows, Amazon DynamoDB will automatically spread your data over sufficient machine resources to meet your storage requirements.

Q: Is there a limit to how much throughput I can get out of a single table?

No, you can increase the throughput you have provisioned for your table using UpdateTable API or in the AWS Management Console. DynamoDB is able to operate at massive scale and there is no theoretical limit on the maximum throughput you can achieve. DynamoDB automatically divides your table across multiple partitions, where each partition is an independent parallel computation unit. DynamoDB can achieve increasingly high throughput rates by adding more partitions.

If you wish to exceed throughput rates of 10,000 writes/second or 10,000 reads/second, you must first contact Amazon through this online form.

Q: Does Amazon DynamoDB remain available when I ask it to scale up or down by

changing the provisioned throughput?

Yes. Amazon DynamoDB is designed to scale its provisioned throughput up or down while still remaining available.

Q: Do I need to manage client-side partitioning on top of Amazon DynamoDB?

No. Amazon DynamoDB removes the need to partition across database tables for throughput scalability.

Q: How highly available is Amazon DynamoDB?

The service runs across Amazon’s proven, high-availability data centers. The service replicates data across three facilities in an AWS Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: How does Amazon DynamoDB achieve high uptime and durability?

To achieve high uptime and durability, Amazon DynamoDB synchronously replicates data across three facilities within an AWS Region.

Global Secondary Indexes

Q: What are global secondary indexes?

Global secondary indexes are indexes that contain a partition or partition-and-sort keys that can be different from the table's primary key.

For efficient access to data in a table, Amazon DynamoDB creates and maintains indexes for the primary key attributes. This allows applications to quickly retrieve data by specifying primary key values. However, many applications might benefit from having one or more secondary (or alternate) keys available to allow efficient access to data with attributes other than the primary key. To address this, you can create one or more secondary indexes on a table, and issue Query requests against these indexes.

Amazon DynamoDB supports two types of secondary indexes:

- Local secondary index — an index that has the same partition key as the table, but a different sort key. A local secondary index is "local" in the sense that every partition of a local secondary index is scoped to a table partition that has the same partition key.

- Global secondary index — an index with a partition or a partition-and-sort key that can be different from those on the table. A global secondary index is considered "global" because queries on the index can span all items in a table, across all partitions.

Secondary indexes are automatically maintained by Amazon DynamoDB as sparse objects.

Items will only appear in an index if they exist in the table on which the index is defined. This makes queries against an index very efficient, because the number of items in the index will often be significantly less than the number of items in the table.

Global secondary indexes support non-unique attributes, which increases query flexibility by enabling queries against any non-key attribute in the table.

Consider a gaming application that stores the information of its players in a DynamoDB table whose primary key consists of *UserId* (partition) and *GameTitle* (sort). Items have attributes named *TopScore*, *Timestamp*, *ZipCode*, and others. Upon table creation, DynamoDB provides an implicit index (primary index) on the primary key that can support efficient queries that return a specific user's top scores for all games.

However, if the application requires top scores of users for a particular game, using this primary index would be inefficient, and would require scanning through the entire table. Instead, a global secondary index with GameTitle as the partition key element and TopScore as the sort key element would enable the application to rapidly retrieve top scores for a game.

A GSI does not need to have a sort key element. For instance, you could have a GSI with a key that only has a partition element *GameTitle*. In the example below, the GSI has no projected attributes, so it will just return all items (identified by primary key) that have an attribute matching the *GameTitle* you are querying on.

Q: When should I use global secondary indexes?

Global secondary indexes are particularly useful for tracking relationships between attributes that have a lot of different values. For example, you could create a DynamoDB table with *CustomerID* as the primary partition key for the table and *ZipCode* as the partition key for a global secondary index, since there are a lot of zip codes and since you will probably have a lot of customers. Using the primary key, you could quickly get the record for any customer. Using the global secondary index, you could efficiently query for all customers that live in a given zip code.

To ensure that you get the most out of your global secondary index's capacity, please review our best practices documentation on uniform workloads.

Q: How do I create a global secondary index for a DynamoDB table?

GSIs associated with a table can be specified at any time. For detailed steps on creating a Table and its indexes, see here. You can create a maximum of 5 global secondary indexes per table.

Q: Does the local version of DynamoDB support global secondary indexes?

Yes. The local version of DynamoDB is useful for developing and testing DynamoDB-backed applications. You can download the local version of DynamoDB here.

Q: What are projected attributes?

The data in a secondary index consists of attributes that are projected, or copied, from the table into the index. When you create a secondary index, you define the alternate key for the index, along with any other attributes that you want to be projected in the index. Amazon DynamoDB copies these attributes into the index, along with the primary key attributes from the table. You can then query the index just as you would query a table.

Q: Can a global secondary index key be defined on non-unique attributes?

Yes. Unlike the primary key on a table, a GSI index does not require the indexed attributes to be unique. For instance, a GSI on *GameTitle* could index all items that track scores of users for every game. In this example, this GSI can be queried to return all users that have played the game "TicTacToe."

Q: How do global secondary indexes differ from local secondary indexes?

Both global and local secondary indexes enhance query flexibility. An LSI is attached to a specific partition key value, whereas a GSI spans all partition key values. Since items having the same partition key value share the same partition in DynamoDB, the "Local" Secondary Index only covers items that are stored together (on the same partition). Thus, the purpose of the LSI is to query items that have the same partition key value but different sort key values. For example, consider a DynamoDB table that tracks Orders for customers, where *CustomerId* is the partition key.

An LSI on *OrderTime* allows for efficient queries to retrieve the most recently ordered items for a particular customer.

In contrast, a GSI is not restricted to items with a common partition key value. Instead, a GSI spans all items of the table just like the primary key. For the table above, a GSI on *ProductId* can be used to efficiently find all orders of a particular product. Note that in this case, no GSI sort key is specified, and even though there might be many orders with the same *ProductId*, they will be stored as separate items in the GSI.

In order to ensure that data in the table and the index are co-located on the same partition, LSIs limit the total size of all elements (tables and indexes) to 10 GB per partition key value. GSIs do not enforce data co-location, and have no such restriction.

When you write to a table, DynamoDB atomically updates all the LSIs affected. In contrast, updates to any GSIs defined on the table are eventually consistent.

LSIs allow the Query API to retrieve attributes that are not part of the projection list. This is not supported behavior for GSIs.

Q: How do global secondary indexes work?

In many ways, GSI behavior is similar to that of a DynamoDB table. You can query a GSI using its partition key element, with conditional filters on the GSI sort key element. However, unlike a primary key of a DynamoDB table, which must be unique, a GSI key can be the same for multiple items. If multiple items with the same GSI key exist, they are tracked as separate GSI items, and a GSI query will retrieve all of them as individual items. Internally, DynamoDB will ensure that the contents of the GSI are updated appropriately as items are added, removed or updated.

DynamoDB stores a GSI's projected attributes in the GSI data structure, along with the GSI key and the matching items' primary keys. GSI's consume storage for projected items that exist in the source table. This enables queries to be issued against the GSI rather than the table, increasing query flexibility and improving workload distribution. Attributes that are part of an item in a table, but not part of the GSI key, primary key of the table, or projected attributes are thus not returned on querying the GSI index. Applications that need additional data from the table after querying the GSI, can retrieve the primary key from the GSI and then use either the GetItem or BatchGetItem APIs to retrieve the desired attributes from the table. As GSI's are eventually consistent, applications that use this pattern have to accommodate item deletion (from the table) in between the calls to the GSI and GetItem/BatchItem.

DynamoDB automatically handles item additions, updates and deletes in a GSI when corresponding changes are made to the table. When an item (with GSI key attributes) is added to the table, DynamoDB updates the GSI asynchronously to add the new item. Similarly, when an item is deleted from the table, DynamoDB removes the item from the impacted GSI.

Q: Can I create global secondary indexes for partition-based tables and partition-sort schema tables?

Yes, you can create a global secondary index regardless of the type of primary key the DynamoDB table has. The table's primary key can include just a partition key, or it may include both a partition key and a sort key.

Q: What is the consistency model for global secondary indexes?

GSIs support eventual consistency. When items are inserted or updated in a table, the GSIs are not updated synchronously. Under normal operating conditions, a write to a global secondary index will propagate in a fraction of a second. In unlikely failure scenarios, longer delays may occur. Because of this, your application logic should be capable of handling GSI query results that are potentially out-of-date. Note that this is the same behavior exhibited by other DynamoDB APIs that support eventually consistent reads.

Consider a table tracking top scores where each item has attributes*UserId*, *GameTitle* and *TopScore*. The partition key is *UserId*, and the primary sort key is*GameTitle*. If the application adds an item denoting a new top score for *GameTitle* "TicTacToe" and *UserId* "GAMER123," and then subsequently queries the GSI, it is possible that the new score will not be in the result of the query. However, once the GSI propagation has completed, the new item will start appearing in such queries on the GSI.

Q: Can I provision throughput separately for the table and for each global secondary index?

Yes. GSIs manage throughput independently of the table they are based on. You need to explicitly specify the provisioned throughput for the table and each associated GSI at creation time. You can use the Create Table Wizard of the DynamoDB Console which can assist you in distributing your total throughput among your tables and indexes.

Depending upon on your application, the request workload on a GSI can vary significantly from that of the table or other GSIs. Some scenarios that show this are given below:

- A GSI that contains a small fraction of the table items needs a much lower write throughput compared to the table.
- A GSI that is used for infrequent item lookups needs a much lower read throughput, compared to the table.
- A GSI used by a read-heavy background task may need high read throughput for a few hours per day.

As your needs evolve, you can change the provisioned throughput of the GSI, independently of the provisioned throughput of the table.

Consider a DynamoDB table with a GSI that projects all attributes, and has the GSI key present in 50% of the items. In this case, the GSI's provisioned write capacity units should be set at 50% of the table's provisioned write capacity units. Using a similar approach, the read throughput of the GSI can be estimated. Please see DynamoDB GSI Documentation for more details.

Q: How does adding a global secondary index impact provisioned throughput and storage for a table?

Similar to a DynamoDB table, a GSI consumes provisioned throughput when reads or writes are performed to it. A write that adds or updates a GSI item will consume write capacity units based on the size of the update. The capacity consumed by the GSI write is in addition to that needed for updating the item in the table.

Note that if you add, delete, or update an item in a DynamoDB table, and if this does not result in a change to a GSI, then the GSI will not consume any write capacity units. This happens when an item without any GSI key attributes is added to the DynamoDB table, or an item is updated without changing any GSI key or projected attributes.

A query to a GSI consumes read capacity units, based on the size of the items examined by the query.

Storage costs for a GSI are based on the total number of bytes stored in that GSI. This includes the GSI key and projected attributes and values, and an overhead of 100 bytes for indexing purposes.

Q: Can DynamoDB throttle my application writes to a table because of a GSI's provisioned throughput?

Because some or all writes to a DynamoDB table result in writes to related GSIs, it is possible that a GSI's provisioned throughput can be exhausted. In such a scenario, subsequent writes to the table will be throttled. This can occur even if the table has available write capacity units.

Q: How often can I change provisioned throughput at the index level?

Tables with GSIs have the same daily limits on the number of throughput change operations as normal tables.

Q: How am I charged for DynamoDB global secondary index?

You are charged for the aggregate provisioned throughput for a table and its GSIs by the hour. In addition, you are charged for the data storage taken up by the GSI as well as standard data transfer (external) fees. If you would like to change your GSI's provisioned throughput capacity, you can do so using the DynamoDB Console or the UpdateTable API.

Q: Can I specify which global secondary index should be used for a query?

Yes. In addition to the common query parameters, a GSI *Query command* explicitly includes the name of the GSI to operate against. Note that a query can use only one GSI.

Q: What API calls are supported by a global secondary index?

The API calls supported by a GSI are *Query* and *Scan*. A Query operation only searches index key attribute values and supports a subset of comparison operators. Because GSIs are updated asynchronously, you cannot use the *ConsistentRead* parameter with the query. Please see here for details on using GSIs with queries and scans.

Q: What is the order of the results in scan on a global secondary index?

For a global secondary index, with a partition-only key schema there is no ordering. For global secondary index with partition-sort key schema the ordering of the results for the same partition key is based on the sort key attribute.

Q. Can I change Global Secondary Indexes after a table has been created?

Yes, Global Secondary Indexes can be changed at any time, even after the table has been created.

Q. How can I add a Global Secondary Index to an existing table?

You can add a Global Secondary Indexes through the console or through an API call. On the DynamoDB console, first select the table for which you want to add a Global Secondary Index and click the “Create Index” button to add a new index. Follow the steps in the index creation wizard and select “Create” when done. You can also add or delete a Global Secondary Index using the UpdateTable API call with the GlobalSecondaryIndexes parameter.You can learn more by reading our documentation page.

Q. How can I delete a Global Secondary Index?

You can delete a Global Secondary Index from the console or through an API call. On the DynamoDB console, select the table for which you want to delete a Global Secondary Index. Then, select the “Indexes” tab under “Table Items” and click on the “Delete” button next to delete the index. You can also delete a Global Secondary Index using the UpdateTable API call.You can learn more by reading our documentation page.

Q. Can I add or delete more than one index in a single API call on the same table?

You can only add or delete one index per API call.

Q. What happens if I submit multiple requests to add the same index?

Only the first add request is accepted and all subsequent add requests will fail till the first add request is finished.

Q. Can I concurrently add or delete several indexes on the same table?

No, at any time there can be only one active add or delete index operation on a table.

Q. Should I provision additional throughput to add a Global Secondary Index?

While not required, it is highly recommended that you provision additional write throughput that is separate from the throughput for the index. If you do not provision additional write throughput, the write throughput from the index will be consumed for adding the new index. This will affect the write performance of the index while the index is being created as well as increase the time to create the new index.

Q. Do I have to reduce the additional throughput on a Global Secondary Index once the index has been created?

Yes, you would have to dial back the additional write throughput you provisioned for adding an index, once the process is complete.

Q. Can I modify the write throughput that is provisioned for adding a Global Secondary Index?

Yes, you can dial up or dial down the provisioned write throughput for index creation at any time during the creation process.

Q. When a Global Secondary Index is being added or deleted, is the table still available?

Yes, the table is available when the Global Secondary Index is being updated.

Q. When a Global Secondary Index is being added or deleted, are the existing indexes still available?

Yes, the existing indexes are available when the Global Secondary Index is being updated.

Q. When a Global Secondary Index is being created added, is the new index available?

No, the new index becomes available only after the index creation process is finished.

Q. How long does adding a Global Secondary Index take?

The length of time depends on the size of the table and the amount of additional provisioned write throughput for Global Secondary Index creation. The process of adding or deleting an index could vary from a few minutes to a few hours. For example, let's assume that you have a 1GB table that has 500 write capacity units provisioned and you have provisioned 1000 additional write capacity units for the index and new index creation. If the new index includes all the attributes in the table and the table is using all the write capacity units, we expect the index creation will take roughly 30 minutes.

Q. How long does deleting a Global Secondary Index take?

Deleting an index will typically finish in a few minutes. For example, deleting an index with 1GB of data will typically take less than 1 minute.

Q. How do I track the progress of add or delete operation for a Global Secondary Index?

You can use the DynamoDB console or DescribeTable API to check the status of all indexes associated with the table. For an add index operation, while the index is being created, the status of the index will be “CREATING”. Once the creation of the index is finished, the index state will change from “CREATING” to “ACTIVE”. For a delete index operation, when the request is complete, the deleted index will cease to exist.

Q. Can I get a notification when the index creation process for adding a Global Secondary Index is complete?

You can request a notification to be sent to your email address confirming that the index addition has been completed. When you add an index through the console, you can request a notification on the last step before creating the index. When the index creation is complete, DynamoDB will send an SNS notification to your email.

Q. What happens when I try to add more Global Secondary Indexes, when I already have 5?

You are currently limited to 5 GSIs. The “Add” operation will fail and you will get an error.

Q. Can I reuse a name for a Global Secondary Index after an index with the same name has been deleted?

Yes, once a Global Secondary Index has been deleted, that index name can be used again when a new index is added.

Q. Can I cancel an index add while it is being created?

No, once index creation starts, the index creation process cannot be canceled.

Q: Are GSI key attributes required in all items of a DynamoDB table?

No. GSIs are sparse indexes. Unlike the requirement of having a primary key, an item in a DynamoDB table does not have to contain any of the GSI keys. If a GSI key has both partition and sort elements, and a table item omits either of them, then that item will not be indexed by the corresponding GSI. In such cases, a GSI can be very useful in efficiently locating items that have an uncommon attribute.

Q: Can I retrieve all attributes of a DynamoDB table from a global secondary index?

A query on a GSI can only return attributes that were specified to be included in the GSI at creation time. The attributes included in the GSI are those that are projected by default such as the GSI's key attribute(s) and table's primary key attribute(s), and those that the user specified to be projected. For this reason, a GSI query will not return attributes of items that are part of the table, but not included in the GSI. A GSI that specifies all attributes as projected attributes can be used to retrieve any table attributes. See here for documentation on using GSIs for queries.

Q: How can I list GSIs associated with a table?

The DescribeTable API will return detailed information about global secondary indexes on a table.

Q: What data types can be indexed?

All scalar data types (Number, String, Binary, and Boolean) can be used for the sort key element of the local secondary index key. Set, list, and map types cannot be indexed.

Q: Are composite attribute indexes possible?

No. But you can concatenate attributes into a string and use this as a key.

Q: What data types can be part of the projected attributes for a GSI?

You can specify attributes with any data types (including set types) to be projected into a GSI.

Q: What are some scalability considerations of GSIs?

Performance considerations of the primary key of a DynamoDB table also apply to GSI keys. A GSI assumes a relatively random access pattern across all its keys. To get the most out of secondary index provisioned throughput, you should select a GSI partition key attribute that has a large number of distinct values, and a GSI sort key attribute that is requested fairly uniformly, as randomly as possible.

Q: What new metrics will be available through CloudWatch for global secondary indexes?

Tables with GSI will provide aggregate metrics for the table and GSIs, as well as breakouts of metrics for the table and each GSI.

Reports for individual GSIs will support a subset of the CloudWatch metrics that are supported by a table. These include:

- Read Capacity (Provisioned Read Capacity, Consumed Read Capacity)
- Write Capacity (Provisioned Write Capacity, Consumed Write Capacity)
- Throttled read events
- Throttled write events

For more details on metrics supported by DynamoDB tables and indexes seehere.

Q: Can I auto-scale my tables and indexes in DynamoDB?

While this is not a native function, there are recommended third party libraries located in the Developer Resources section of the DynamoDB web page.

Q: How can I scan a Global Secondary Index?

Global secondary indexes can be scanned via the Console or the Scan API.

To scan a global secondary index, explicitly reference the index in addition to the name of the table you'd like to scan. You must specify the index partition attribute name and value. You can optionally specify a condition against the index key sort attribute.

Q: Will a Scan on Global secondary index allow me to specify non-projected attributes to be returned in the result set?

Scan on global secondary indexes will not support fetching of non-projected attributes.

Q: Will there be parallel scan support for indexes?

Yes, parallel scan will be supported for indexes and the semantics are the same as that for the main table.

Local Secondary Indexes

Q: What are local secondary indexes?

Local secondary indexes enable some common queries to run more quickly and cost-efficiently, that would otherwise require retrieving a large number of items and then filtering the results. It means your applications can rely on more flexible queries based on a wider range of attributes.

Before the launch of local secondary indexes, if you wanted to find specific items within a partition (items that share the same partition key), DynamoDB would have fetched all objects that share a single partition key, and filter the results accordingly. For instance, consider an e-commerce application that stores customer order data in a DynamoDB table with partition-sort schema of customer id-order timestamp. Without LSI, to find an answer to the question “Display all orders made by Customer X with shipping date in the past 30 days, sorted by shipping date”, you had to use the Query API to retrieve all the objects under the partition key “X”, sort the results by shipment date and then filter out older records.

With local secondary indexes, we are simplifying this experience. Now, you can create an index on “shipping date” attribute and execute this query efficiently and just retieve only the necessary items. This significantly reduces the latency and cost of your queries as you will retrieve only items that meet your specific criteria. Moreover, it also simplifies the programming model for your application as you no longer have to write customer logic to filter the results. We call this new secondary index a ‘local’ secondary index because it is used along with the partition key and hence allows you to search locally within a partition key bucket. So while previously you could only search using the partition key and the sort key, now you can also search using a secondary index in place of the sort key, thus expanding the number of attributes that can be used for queries which can be conducted efficiently.

Redundant copies of data attributes are copied into the local secondary indexes you define. These attributes include the table partition and sort key, plus the alternate sort key you define. You can also redundantly store other data attributes in the local secondary index, in order to access those other attributes without having to access the table itself.

Local secondary indexes are not appropriate for every application. They introduce some constraints on the volume of data you can store within a single partition key value. For more information, see the FAQ items below about item collections.

Q: What are Projections?

The set of attributes that is copied into a local secondary index is called a projection. The projection determines the attributes that you will be able to retrieve with the most efficiency. When you query a local secondary index, Amazon DynamoDB can access any of the projected attributes, with the same performance characteristics as if those attributes were in a table of their own. If you need to retrieve any attributes that are not projected, Amazon DynamoDB will automatically fetch those attributes from the table.

When you define a local secondary index, you need to specify the attributes that will be projected into the index. At a minimum, each index entry consists of: (1) the table partition key value, (2) an attribute to serve as the index sort key, and (3) the table sort key value.

Beyond the minimum, you can also choose a user-specified list of other non-key attributes to project into the index. You can even choose to project all attributes into the index, in which case the index replicates the same data as the table itself, but the data is organized by the alternate sort key you specify.

Q: How can I create a LSI?

You need to create a LSI at the time of table creation. It can’t currently be added later on. To create an LSI, specify the following two parameters:

Indexed Sort key – the attribute that will be indexed and queried on.

Projected Attributes – the list of attributes from the table that will be copied directly into the local secondary index, so they can be returned more quickly without fetching data from the primary index, which contains all the items of the table. Without projected attributes, local secondary index contains only primary and secondary index keys.

Q: What is the consistency model for LSI?

Local secondary indexes are updated automatically when the primary index is updated. Similar to reads from a primary index, LSI supports both strong and eventually consistent read options.

Q: Do local secondary indexes contain references to all items in the table?

No, not necessarily. Local secondary indexes only reference those items that contain the indexed sort key specified for that LSI. DynamoDB’s flexible schema means that not all items will necessarily contain all attributes.

This means local secondary index can be sparsely populated, compared with the primary index. Because local secondary indexes are sparse, they are efficient to support queries on attributes that are uncommon.

For example, in the Orders example described above, a customer may have some additional attributes in an item that are included only if the order is canceled (such as CanceledDateTime, CanceledReason). For queries related to canceled items, an local secondary index on either of these attributes would be efficient since the only items referenced in the index would be those that had these attributes present.

Q: How do I query local secondary indexes?

Local secondary indexes can only be queried via the Query API.

To query a local secondary index, explicitly reference the index in addition to the name of the table you’d like to query. You must specify the index partition attribute name and value. You can optionally specify a condition against the index key sort attribute.

Your query can retrieve non-projected attributes stored in the primary index by performing a table fetch operation, with a cost of additional read capacity units.

Both strongly consistent and eventually consistent reads are supported for query using local secondary index.

Q: How do I create local secondary indexes?

Local secondary indexes must be defined at time of table creation. The primary index of the table must use a partition-sort composite key.

Q: Can I add local secondary indexes to an existing table?

No, it’s not possible to add local secondary indexes to existing tables at this time. We are working on adding this capability and will be releasing it in the future. When you create a table with local secondary index, you may decide to create local secondary index for future use by defining a sort key element that is currently not used. Since local secondary index are sparse, this index costs nothing until you decide to use it.

Q: How many local secondary indexes can I create on one table?

Each table can have up to five local secondary indexes.

Q: How many projected non-key attributes can I create on one table?

Each table can have up to 20 projected non-key attributes, in total across all local secondary indexes within the table. Each index may also specifify that all non-key attributes from the primary index are projected.

Q: Can I modify the index once it is created?

No, an index cannot be modified once it is created. We are working to add this capability in the future.

Q: Can I delete local secondary indexes?

No, local secondary indexes cannot be removed from a table once they are created at this time. Of course, they are deleted if you also decide to delete the entire table. We are working on adding this capability and will be releasing it in the future.

Q: How do local secondary indexes consume provisioned capacity?

You don’t need to explicitly provision capacity for a local secondary index. It consumes provisioned capacity as part of the table with which it is associated.

Reads from LSIs and writes to tables with LSIs consume capacity by the standard formula of 1 unit per 1KB of data, with the following differences:

When writes contain data that are relevant to one or more local secondary indexes, those writes are mirrored to the appropriate local secondary indexes. In these cases, write capacity will be consumed for the table itself, and additional write capacity will be consumed for each relevant LSI.

Updates that overwrite an existing item can result in two operations– delete and insert – and thereby consume extra units of write capacity per 1KB of data.

When a read query requests attributes that are not projected into the LSI, DynamoDB will fetch those attributes from the primary index. This implicit GetItem request consumes one read capacity unit per 4KB of item data fetched.

Q: How much storage will local secondary indexes consume?

Local secondary indexes consume storage for the attribute name and value of each LSI's primary and index keys, for all projected non-key attributes, plus 100 bytes per item reflected in the LSI.

Q: What data types can be indexed?

All scalar data types (Number, String, Binary) can be used for the sort key element of the local secondary index key. Set types cannot be used.

Q: What data types can be projected into a local secondary index?

All data types (including set types) can be projected into a local secondary index.

Q: What are item collections and how are they related to LSI?

In Amazon DynamoDB, an item collection is any group of items that have the same partition key, across a table and all of its local secondary indexes. Traditional partitioned (or sharded) relational database systems call these shards or partitions, referring to all database items or rows stored under a partition key.

Item collections are automatically created and maintained for every table that includes local secondary indexes. DynamoDB stores each item collection within a single disk partition.

Q: Are there limits on the size of an item collection?

Every item collection in Amazon DynamoDB is subject to a maximum size limit of 10 gigabytes. For any distinct partition key value, the sum of the item sizes in the table plus the sum of the item sizes across all of that table's local secondary indexes must not exceed 10 GB.

The 10 GB limit for item collections does not apply to tables without local secondary indexes; only tables that have one or more local secondary indexes are affected.

Although individual item collections are limited in size, the storage size of an overall table with local secondary indexes is not limited. The total size of an indexed table in Amazon DynamoDB is effectively unlimited, provided the total storage size (table and indexes) for any one partition key does not exceed the 10 GB threshold.

Q: How can I track the size of an item collection?

DynamoDB's write APIs (PutItem, UpdateItem, DeleteItem, and BatchWriteItem) include an option, which allows the API response to include an estimate of the relevant item collection's size. This estimate includes lower and upper size estimate for the data in a particular item collection, measured in gigabytes.

We recommend that you instrument your application to monitor the sizes of your item collections. Your applications should examine the API responses regarding item collection size, and log an error message whenever an item collection exceeds a user-defined limit (8 GB, for example). This would provide an early warning system, letting you know that an item collection is growing larger, but giving you enough time to do something about it.

Q: What if I exceed the 10GB limit for an item collection?

If a particular item collection exceeds the 10GB limit, then you will not be able to write new items, or increase the size of existing items, for that particular partition key. Read and write operations that shrink the size of the item collection are still allowed. Other item collections in the table are not affected.

To address this problem , you can remove items or reduce item sizes in the collection that has exceeded 10GB. Alternatively, you can introduce new items under a new partition key value to work around this problem. If your table includes historical data that is infrequently accessed, consider archiving the historical data to Amazon S3, Amazon Glacier or another data store.

Q: How can I scan a local secondary index?

To scan a local secondary index, explicitly reference the index in addition to the name of the table you'd like to scan. You must specify the index partition attribute name and value. You can optionally specify a condition against the index key sort attribute.

Your scan can retrieve non-projected attributes stored in the primary index by performing a table fetch operation, with a cost of additional read capacity units.

Q: Will a Scan on a local secondary index allow me to specify non-projected attributes to be returned in the result set?

Scan on local secondary indexes will support fetching of non-projected attributes.

Q: What is the order of the results in scan on a local secondary index?

For local secondary index, the ordering within a collection will be the based on the order of the indexed attribute.

Security and Control

Q: What is DynamoDB Fine-Grained Access Control?

Fine Grained Access Control (FGAC) gives a DynamoDB table owner a high degree of control over data in the table. Specifically, the table owner can indicate *who* (caller) can access *which* items or attributes of the table and perform *what* actions (read / write capability). FGAC is used in concert with AWS Identity and Access Management (IAM), which manages the security credentials and the associated permissions.

Q: What are the common use cases for DynamoDB FGAC?

FGAC can benefit any application that tracks information in a DynamoDB table, where the end user (or application client acting on behalf of an end user) wants to read or modify the table directly, without a middle-tier service. For instance, a developer of a mobile app named *Acme* can use FGAC to track the top score of every *Acme* user in a DynamoDB table. FGAC allows the application client to modify only the top score for the user that is currently running the application.

Q: Can I use Fine Grain Access Control with JSON documents?

Yes. You can use Fine Grain Access Control (FGAC) to restrict access to your data based on top-level attributes in your document. You cannot use FGAC to restrict access based on nested attributes. For example, suppose you stored a JSON document that contained the following information about a person: ID, first name, last name, and a list of all of their friends. You could use FGAC to restrict access based on their ID, first name, or last name, but not based on the list of friends.

Q: Without FGAC, how can a developer achieve item level access control?

To achieve this level of control without FGAC, a developer would have to choose from a few potentially onerous approaches. Some of these are:

1. Proxy: The application client sends a request to a brokering proxy that performs the authentication and authorization. Such a solution increases the complexity of the system architecture and can result in a higher total cost of ownership (TCO).

2. Per Client Table: Every application client is assigned its own table. Since application clients access different tables, they would be protected from one another. This could potentially require a developer to create millions of tables, thereby making database management extremely painful.

3. Per-Client Embedded Token: A secret token is embedded in the application client. The shortcoming of this is the difficulty in changing the token and handling its impact on the stored data. Here, the key of the items accessible by this client would contain the secret token.

Q: How does DynamoDB FGAC work?

With FGAC, an application requests a security token that authorizes the application to access only specific items in a specific DynamoDB table. With this token, the end user application agent can make requests to DynamoDB directly. Upon receiving the request, the incoming request's credentials are first evaluated by DynamoDB, which will use IAM to authenticate the request and determine the capabilities allowed for the user. If the user's request is not permitted, FGAC will prevent the data from being accessed.

Q: How much does DynamoDB FGAC cost?

There is no additional charge for using FGAC. As always, you only pay for the provisioned throughput and storage associated with the DynamoDB table.

Q: How do I get started?

Refer to the Fine-Grained Access Control section of the DynamoDB Developer Guide to learn how to create an access policy, create an IAM role for your app (e.g. a role named AcmeFacebookUsers for a Facebook app_id of 34567), and assign your access policy to the role. The trust policy of the role determines which identity providers are accepted (e.g. Login with Amazon, Facebook, or Google), and the access policy describes which AWS resources can be accessed (e.g. a DynamoDB table). Using the role, your app can now to obtain temporary credentials for DynamoDB by calling the AssumeRoleWithIdentityRequest API of the AWS Security Token Service (STS).

Q: How do I allow users to Query a Local Secondary Index, but prevent them from causing a table fetch to retrieve non-projected attributes?

Some Query operations on a Local Secondary Index can be more expensive than others if they request attributes that are not projected into an index. You an restrict such potentially expensive "fetch" operations by limiting the permissions to only projected attributes, using the "dynamodb:Attributes" context key.

Q: How do I prevent users from accessing specific attributes?

The recommended approach to preventing access to specific attributes is to follow the principle of least privilege, and Allow access to only specific attributes.

Alternatively, you can use a *Deny* policy to specify attributes that are disallowed. However, this is not recommended for the following reasons:

1. With a *Deny* policy, it is possible for the user to discover the hidden attribute names by issuing repeated requests for every possible attribute name, until the user is ultimately denied access.

2. *Deny* policies are more fragile, since DynamoDB could introduce new API functionality in the future that might allow an access pattern that you had previously intended to block.

Q: How do I prevent users from adding invalid data to a table?

The available FGAC controls can determine which items changed or read, and which attributes can be changed or read. Users can add new items without those blocked attributes, and change any value of any attribute that is modifiable.

Q: Can I grant access to multiple attributes without listing all of them?

Yes, the IAM policy language supports a rich set of comparison operations, including StringLike, StringNotLike, and many others. For additional details, please see the IAM Policy Reference.

Q: How do I create an appropriate policy?

We recommend that you use the DynamoDB Policy Generator from the DynamoDB console. You may also compare your policy to those listed in the Amazon DynamoDB Developer Guide to make sure you are following a recommended pattern. You can post policies to the AWS Forums to get thoughts from the DynamoDB community.

Q: Can I grant access based on a canonical user id instead of separate ids for the user based on the identity provider they logged in with?

Not without running a “token vending machine”. If a user retrieves federated access to your IAM role directly using Facebook credentials with STS, those temporary credentials only have information about that user’s Facebook login, and not their Amazon login, or Google login. If you want to internally store a mapping of each of these logins to your own stable identifier, you can run a service that the user contacts to log in, and then call STS and provide them with credentials scoped to whatever partition key value you come up with as their canonical user id.

Q: What information cannot be hidden from callers using FGAC?

Certain information cannot currently be blocked from the caller about the items in the table:

- Item collection metrics. The caller can ask for the estimated number of items and size in bytes of the item collection.
- Consumed throughput The caller can ask for the detailed breakdown or summary of the provisioned throughput consumed by operations.
- Validation cases. In certain cases, the caller can learn about the existence and primary key schema of a table when you did not intend to give them access. To prevent this, follow the principle of least privilege and only allow access to the tables and actions that you intended to allow access to.
- If you deny access to specific attributes instead of whitelisting access to specific attributes, the caller can theoretically determine the names of the hidden attributes if “allow all except for” logic. It is safer to whitelist specific attribute names instead.

Q: Does Amazon DynamoDB support IAM permissions?

Yes, DynamoDB will support API-level permissions through AWS Identity and Access Management (IAM) service integration

For more information about IAM, go to:

- AWS Identity and Access Management
- AWS Identity and Access Management Getting Started Guide
- Using AWS Identity and Access Management

Q: I wish to perform security analysis or operational troubleshooting on my DynamoDB tables. Can I get a history of all DynamoDB API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by AWS CloudTrail enables security analysis, resource change tracking, and compliance auditing. Details about DynamoDB support for CloudTrail can be found here. Learn more about CloudTrail at theAWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console home page.

Pricing

Q: How will I be charged for my use of Amazon DynamoDB?

Each DynamoDB table has provisioned read-throughput and write-throughput associated with it. You are billed by the hour for that throughput capacity if you exceed the free tier.

Please note that you are charged by the hour for the throughput capacity that you provision for your table, whether or not you are sending requests to your table. If you would like to change your table’s provisioned throughput capacity, you can do so using the AWS Management Console or the UpdateTable API.

In addition, DynamoDB also charges for indexed data storage as well as the standard internet data transfer fees

To learn more about DynamoDB pricing, please visit the DynamoDB pricing page.

Q: What are some pricing examples?

Here is an example of how to calculate your throughput costs using US East (Northern Virginia) Region pricing. To view prices for other regions, visit our pricing page.

If you create a table and request 10 units of write capacity and 200 units of read capacity of provisioned throughput, you would be charged:

$0.01 + (4 x $0.01) = $0.05 per hour

If your throughput needs changed and you increased your reserved throughput requirement to 10,000 units of write capacity and 50,000 units of read capacity, your bill would then change to:

(1,000 x $0.01) + (1,000 x $0.01) = $20/hour

To learn more about DynamoDB pricing, please visit the DynamoDB pricing page.

Q: Do your prices include taxes?

For details on taxes, see Amazon Web Services Tax Help.

Q: What is provisioned throughput?

Amazon DynamoDB lets you specify the request throughput you want your table to be able to achieve. Behind the scenes, the service handles the provisioning of resources to achieve the requested throughput rate. Rather than asking you to think about instances, hardware, memory, and other factors that could affect your throughput rate, we simply ask you to provision the throughput level you want to achieve. This is the provisioned throughput model of service.

Amazon DynamoDB lets you specify your throughput needs in terms of units of read capacity and write capacity for your table. During creation of a table, you specify your required read and write capacity needs and Amazon DynamoDB automatically partitions and reserves the appropriate amount of resources to meet your throughput requirements. To decide on the required read and write throughput values, consider the number of read and write data plane API calls you expect to perform per second. If at any point you anticipate traffic growth that may exceed your provisioned throughput, you can simply update your provisioned throughput values via the AWS Management Console or Amazon DynamoDB APIs. You can also reduce the provisioned throughput value for a table as demand decreases. Amazon DynamoDB will remain available while scaling it throughput level up or down.

Q: How does selection of primary key influence the scalability I can achieve?

When storing data, Amazon DynamoDB divides a table into multiple partitions and distributes the data based on the partition key element of the primary key. While allocating capacity resources, Amazon DynamoDB assumes a relatively random access pattern across all primary keys. You should set up your data model so that your requests result in a fairly even distribution of traffic across primary keys. If a table has a very small number of heavily-accessed partition key elements, possibly even a single very heavily-used partition key element, traffic is concentrated on a small number of partitions – potentially only one partition. If the workload is heavily unbalanced, meaning disproportionately focused on one or a few partitions, the operations will not achieve the overall provisioned throughput level. To get the most out of Amazon DynamoDB throughput, build tables where the partition key element has a large number of distinct values, and values are requested fairly uniformly, as randomly as possible. An example of a good primary key is CustomerID if the application has many customers and requests made to various customer records tend to be more or less uniform. An example of a heavily skewed primary key is “Product Category Name” where certain product categories are more popular than the rest.

Q: What is a read/write capacity unit?

How do I estimate how many read and write capacity units I need for my application? A unit of Write Capacity enables you to perform one write per second for items of up to 1KB in size. Similarly, a unit of Read Capacity enables you to perform one strongly consistent read per second (or two eventually consistent reads per second) of items of up to 4KB in size. Larger items will require more capacity. You can calculate the number of units of read and write capacity you need by estimating the number of reads or writes you need to do per second and multiplying by the size of your items (rounded up to the nearest KB).

Units of Capacity required for writes = Number of item writes per second x item size in 1KB blocks

Units of Capacity required for reads* = Number of item reads per second x item size in 4KB blocks

* If you use eventually consistent reads you'll get twice the throughput in terms of reads per second.

If your items are less than 1KB in size, then each unit of Read Capacity will give you 1 strongly consistent read/second and each unit of Write Capacity will give you 1 write/second of capacity. For example, if your items are 512 bytes and you need to read 100 items per second from your table, then you need to provision 100 units of Read Capacity.

If your items are larger than 4KB in size, then you should calculate the number of units of Read Capacity and Write Capacity that you need. For example, if your items are 4.5KB and you want to do 100 strongly consistent reads/second, then you would need to provision 100 (read per second) x 2 (number of 4KB blocks required to store 4.5KB) = 200 units of Read Capacity.

Note that the required number of units of Read Capacity is determined by the number of items being read per second, not the number of API calls. For example, if you need to read 500 items per second from your table, and if your items are 4KB or less, then you need 500 units of Read Capacity. It doesn't matter if you do 500 individual GetItem calls or 50 BatchGetItem calls that each return 10 items.

Q: Will I always be able to achieve my level of provisioned throughput?

Amazon DynamoDB assumes a relatively random access pattern across all primary keys. You should set up your data model so that your requests result in a fairly even distribution of traffic across primary keys. If you have a highly uneven or skewed access pattern, you may not be able to achieve your level of provisioned throughput.

When storing data, Amazon DynamoDB divides a table into multiple partitions and distributes the data based on the partition key element of the primary key. The provisioned throughput associated with a table is also divided among the partitions; each partition's throughput is managed independently based on the quota allotted to it. There is no sharing of provisioned throughput across partitions. Consequently, a table in Amazon DynamoDB is best able to meet the provisioned throughput levels if the workload is spread fairly uniformly across the partition key values. Distributing requests across partition key values distributes the requests across partitions, which helps achieve your full provisioned throughput level.

If you have an uneven workload pattern across primary keys and are unable to achieve your provisioned throughput level, you may be able to meet your throughput needs by increasing your provisioned throughput level further, which will give more throughput to each partition. However, it is recommended that you considering modifying your request pattern or your data model in order to achieve a relatively random access pattern across primary keys.

Q: If I retrieve only a single element of a JSON document, will I be charged for reading the whole item?

Yes. When reading data out of DynamoDB, you consume the throughput required to read the entire item.

Q: What is the maximum throughput I can provision for a single DynamoDB table?

DynamoDB is designed to scale without limits However, if you wish to exceed throughput rates of 10,000 write capacity units or 10,000 read capacity units for an individual table, you must first contact Amazon through this online form. If you wish to provision more than 20,000 write capacity units or 20,000 read capacity units from a single subscriber account you must first contact us using the form described above.

Q: What is the minimum throughput I can provision for a single DynamoDB table?

The smallest provisioned throughput you can request is 1 write capacity unit and 1 read capacity unit.

This falls within the free tier which allows for 25 units of write capacity and 25 units of read capacity. The free tier applies at the account level, not the table level. In other words, if you add up the provisioned capacity of all your tables, and if the total capacity is no more than 25 units of write capacity and 25 units of read capacity, your provisioned capacity would fall into the free tier.

Q: Is there any limit on how much I can change my provisioned throughput with a single request?

You can increase the provisioned throughput capacity of your table by any amount using the UpdateTable API. For example, you could increase your table's provisioned write capacity from 1 write capacity unit to 10,000 write capacity units with a single API call. Your account is still subject to table-level and account-level limits on capacity, as described in our documentation page. If you need to raise your provisioned capacity limits, you can visit our Support Center, click "Open a new case", and file a service limit increase request.

Q: How am I charged for provisioned throughput?

Every Amazon DynamoDB table has pre-provisioned the resources it needs to achieve the throughput rate you asked for. You are billed at an hourly rate for as long as your table holds on to those resources. For a complete list of prices with examples, see the DynamoDB pricing page.

Q: How do I change the provisioned throughput for an existing DynamoDB table?

There are two ways to update the provisioned throughput of an Amazon DynamoDB table. You can either make the change in the management console, or you can use the UpdateTable API call. In either case, Amazon DynamoDB will remain available while your provisioned throughput level increases or decreases.

Q: How often can I change my provisioned throughput?

You can increase your provisioned throughput as often as you want. You can decrease it four times per day. A day is defined according to the GMT time zone. For example, if you decrease the provisioned throughput for your table four times on December 12th, you won't be able to decrease the provisioned throughput for that table again until 12:01am GMT on December 13th.

Keep in mind that you can't change your provisioned throughput if your Amazon DynamoDB table is still in the process of responding to your last request to change provisioned throughput. Use the management console or the DescribeTables API to check the status of your table. If the status is "CREATING", "DELETING", or "UPDATING", you won't be able to adjust the throughput of your table. Please wait until you have a table in "ACTIVE" status and try again.

Q: Does the consistency level affect the throughput rate?

Yes. For a given allocation of resources, the read-rate that a DynamoDB table can achieve is different for strongly consistent and eventually consistent reads. If you request "1,000 read capacity units", DynamoDB will allocate sufficient resources to achieve 1,000 strongly consistent reads per second of items up to 4KB. If you want to achieve 1,000 eventually consistent reads of items up to 4KB, you will need half of that capacity, i.e., 500 read capacity units. For additional guidance on choosing the appropriate throughput rate for your table, see our provisioned throughput guide.

Q: Does the item size affect the throughput rate?

Yes. For a given allocation of resources, the read-rate that a DynamoDB table can achieve does depend on the size of an item. When you specify the provisioned read throughput you would like to achieve, DynamoDB provisions its resources on the assumption that items will be less than 4KB in size. Every increase of up to 4KB will linearly increase the resources you need to achieve the same throughput rate. For example, if you have provisioned a DynamoDB table with 100 units of read capacity, that means that it can handle 100 4KB reads per second, or 50 8KB reads per second, or 25 16KB reads per second, and so on.

Similarly the write-rate that a DynamoDB table can achieve does depend on the size of an item. When you specify the provisioned write throughput you would like to achieve, DynamoDB provisions its resources on the assumption that items will be less than 1KB in size. Every increase of up to 1KB will linearly increase the resources you need to achieve the same throughput rate. For example, if you have provisioned a DynamoDB table with 100 units of write capacity, that means that it can handle 100 1KB writes per second, or 50 2KB writes per second, or 25 4KB writes per second, and so on.

For additional guidance on choosing the appropriate throughput rate for your table, see our provisioned throughput guide.

Q: What happens if my application performs more reads or writes than my provisioned capacity?

If your application performs more reads/second or writes/second than your table’s provisioned throughput capacity allows, requests above your provisioned capacity will be throttled and you will receive 400 error codes. For instance, if you had asked for 1,000 write capacity units and try to do 1,500 writes/second of 1 KB items, DynamoDB will only allow 1,000 writes/second to go through and you will receive error code 400 on your extra requests. You should use CloudWatch to monitor your request rate to ensure that you always have enough provisioned throughput to achieve the request rate that you need.

Q: How do I know if I am exceeding my provisioned throughput capacity?

DynamoDB publishes your consumed throughput capacity as a CloudWatch metric. You can set an alarm on this metric so that you will be notified if you get close to your provisioned capacity.

Q: How long does it take to change the provisioned throughput level of a table?

In general, decreases in throughput will take anywhere from a few seconds to a few minutes, while increases in throughput will typically take anywhere from a few minutes to a few hours.

We strongly recommend that you do not try and schedule increases in throughput to occur at almost the same time when that extra throughput is needed. We recommend provisioning throughput capacity sufficiently far in advance to ensure that it is there when you need it.

Reserved Capacity

Q: What is Reserved Capacity?

Reserved Capacity is a billing feature that allows you to obtain discounts on your provisioned throughput capacity in exchange for:

- A one-time up-front payment
- A commitment to a minimum monthly usage level for the duration of the term of the agreement.

Reserved Capacity applies within a single AWS Region and can be purchased with 1-year or 3-year terms. Every DynamoDB table has provisioned throughput capacity associated with it. When you create or update a table, you specify how much read or write capacity you want it to have. This capacity is what determines the read and write throughput rate that your DynamoDB table can achieve. Reserved Capacity is a billing arrangement and has no direct impact on the performance or capacity of your DynamoDB tables. For example, if you buy 100 write capacity units of Reserved Capacity, you have agreed to pay for that much capacity for the duration of the agreement (1 or 3 years) in exchange for discounted pricing.

Q: How do I buy Reserved Capacity?

Log into the AWS Management Console, go to the DynamoDB console page, and then click on "Reserved Capacity”. This will take you to the "Reserved Capacity Usage" page. Click on "Purchase Reserved Capacity" and this will bring up a form you can fill out to purchase Reserved Capacity. Make sure you have selected the AWS Region in which your Reserved Capacity will be used. After you have finished purchasing Reserved Capacity, you will see purchase you made on the "Reserved Capacity Usage" page.

Q: Can I cancel a Reserved Capacity purchase?

No, you cannot cancel your Reserved Capacity and the one-time payment is not refundable. You will continue to pay for every hour during your Reserved Capacity term regardless of your usage.

Q: What is the smallest amount of Reserved Capacity that I can buy?

The smallest Reserved Capacity offering is 100 capacity units (reads or writes).

Q: Are there APIs that I can use to buy Reserved Capacity?

Not yet. We will provide APIs and add more Reserved Capacity options over time.

Q: Can I move Reserved Capacity from one Region to another?

No. Reserved Capacity is associated with a single Region.

Q: Can I provision more throughput capacity than my Reserved Capacity?

Yes. When you purchase Reserved Capacity, you are agreeing to a minimum usage level and you pay a discounted rate for that usage level. If you provision more capacity than that minimum level, you will be charged at standard rates for the additional capacity.

Q: How do I use my Reserved Capacity?

Reserved Capacity is automatically applied to your bill. For example, if you purchased 100 write capacity units of Reserved Capacity and you have provisioned 300, then your Reserved Capacity purchase will automatically cover the cost of 100 write capacity units and you will pay standard rates for the remaining 200 write capacity units.

Q: What happens if I provision less throughput capacity than my Reserved Capacity?

A Reserved Capacity purchase is an agreement to pay for a minimum amount of provisioned throughput capacity, for the duration of the term of the agreement, in exchange for discounted pricing. If you use less than your Reserved Capacity, you will still be charged each month for that minimum amount of provisioned throughput capacity.

Q: Can I use my Reserved Capacity for multiple DynamoDB tables?

Yes. Reserved Capacity is applied to the total provisioned capacity within the Region in which you purchased your Reserved Capacity. For example, if you purchased 5,000 write capacity units of Reserved Capacity, then you can apply that to one table with 5,000 write capacity units, or 100 tables with 50 write capacity units, or 1,000 tables with 5 write capacity units, etc.

Q: Does Reserved Capacity apply to DynamoDB usage in Consolidated Billing accounts?

Yes. If you have multiple accounts linked with Consolidated Billing, Reserved Capacity units purchased either at the Payer Account level or Linked Account level are shared with all accounts connected to the Payer Account. Reserved capacity will first be applied to the account which purchased it and then any unused capacity will be applied to other linked accounts.

DynamoDB Cross-region Replication

Q: What is a DynamoDB cross-region replication?

DynamoDB cross-region replication allows you to maintain identical copies (called replicas) of a DynamoDB table (called master table) in one or more AWS regions. After you enable cross-region replication for a table, identical copies of the table are created in other AWS regions. Writes to the table will be automatically propagated to all replicas.

Q: When should I use cross-region replication?

You can use cross-region replication for the following scenarios.

- **Efficient disaster recovery:** By replicating tables in multiple data centers, you can switch over to using DynamoDB tables from another region in case a data center failure occurs.

- **Faster reads:** If you have customers in multiple regions, you can deliver data faster by reading a DynamoDB table from the closest AWS data center.
- **Easier traffic management:** You can use replicas to distribute the read workload across tables and thereby consume less read capacity in the master table.
- **Easy regional migration:** By creating a read replica in a new region and then promoting the replica to be a master, you migrate your application to that region more easily.
- **Live data migration:** To move a DynamoDB table from one region to another, you can create a replica of the table from the source region in the destination region. When the tables are in sync, you can switch your application to write to the destination region.

Q: What cross-region replication modes are supported?

Cross-region replication currently supports single master mode. A single master has one master table and one or more replica tables.

Q. How can I set up single master cross-region replication for a table?

You can create cross-region replicas using theDynamoDB Cross-region Replication library.

Q: How do I know when the bootstrapping is complete?

On the replication management application, the state of the replication changes from Bootstrapping to Active.

Q: Can I have multiple replicas for a single master table?

Yes, there are no limits on the number of replicas tables from a single master table. A DynamoDB Streams reader is created for each replica table and copies data from the master table, keeping the replicas in sync.

Q: How much does it cost to set up cross-region replication for a table?

DynamoDB cross-region replication is enabled using the DynamoDB Cross-region Replication Library. While there is no additional charge for the cross-region replication library, you pay the usual prices for the following resources used by the process. You will be billed for:

- Provisioned throughput (Writes and Reads) and storage for the replica tables.
- Data Transfer across regions.
- Reading data from DynamoDB Streams to keep the tables in sync.
- The EC2 instances provisioned to host the replication process. The cost of the instances will depend on the instance type you choose and the region hosting the instances.

Q: In which region does the Amazon EC2 instance hosting the cross-region replication

run?

The cross-region replication application is hosted in an Amazon EC2 instance in the same region where the cross-region replication application was originally launched. You will be charged the instance price in this region.

Q: Does the Amazon EC2 instance Auto Scale as the size and throughput of the master and replica tables change?

Currently, we will not auto scale the EC2 instance. You will need to pick the instance size when configuring DynamoDB Cross-region Replication.

Q: What happens if the Amazon EC2 instance managing the replication fails?

The Amazon EC2 instance runs behind an auto scaling group, which means the application will automatically fail over to another instance. The application underneath uses the Kinesis Client Library (KCL), which checkpoints the copy. In case of an instance failure, the application knows to find the checkpoint and resume from there.

Q: Can I keep using my DynamoDB table while a Read Replica is being created?

Yes, creating a replica is an online operation. Your table will remain available for reads and writes while the read replica is being created. The bootstrapping uses the Scan operation to copy from the source table. We recommend that the table is provisioned with sufficient read capacity units to support the Scan operation.

Q: How long does it take to create a replica?

The time to initially copy the master table to the replica table depends on the size of the master table, the provisioned capacity of the master table and replica table. The time to propagate an item-level change on the master table to the replica table depends on the provisioned capacity on the master and replica tables, and the size of the Amazon EC2 instance running the replication application.

Q: If I change provisioned capacity on my master table, does the provisioned capacity on my replica table also update?

After the replication has been created, any changes to the provisioned capacity on the master table will not result in an update in throughput capacity on the replica table.

Q: Will my replica tables have the same indexes as the master table?

If you choose to create the replica table from the replication application, the secondary indexes on the master table will NOT be automatically created on the replica table. The replication application will not propagate changes made on secondary indices on the master table to replica tables. You will have to add/update/delete indexes on each of the replica tables through the AWS Management Console as you would with regular DynamoDB tables.

Q: Will my replica have the same provisioned throughput capacity as the master table?

When creating the replica table, we recommend that you provision at least the same write capacity as the master table to ensure that it has enough capacity to handle all incoming writes. You can set the provisioned read capacity of your replica table at whatever level is appropriate for your application.

Q: What is the consistency model for replicated tables?

Replicas are updated asynchronously. DynamoDB will acknowledge a write operation as successful once it has been accepted by the master table. The write will then be propagated to each replica. This means that there will be a slight delay before a write has been propagated to all replica tables.

Q: Are there CloudWatch metrics for cross-region replication?

CloudWatch metrics are available for every replication configuration. You can see the metric by selecting the replication group and navigating to the Monitoring tab. Metrics on throughput and number of record processed are available, and you can monitor for any discrepancies in the throughput of the master and replica tables.

Q: Can I have a replica in the same region as the master table?

Yes, as long as the replica table and the master table have different names, both tables can exist in the same region.

Q: Can I add or delete a replica after creating a replication group?

Yes, you can add or delete a replica from that replication group at any time.

Q: Can I delete a replica group after it is created ?

Yes, deleting the replication group will delete the EC2 instance for the group. However, you will have to delete the DynamoDB metadata table.

DynamoDB Triggers

Q. What is DynamoDB Triggers?

DynamoDB Triggers is a feature which allows you to execute custom actions based on item-level updates on a DynamoDB table. You can specify the custom action in code.

Q. What can I do with DynamoDB Triggers?

There are several application scenarios where DynamoDB Triggers can be useful. Some use cases include sending notifications, updating an aggregate table, and connecting DynamoDB tables to other data sources.

Q. How does DynamoDB Triggers work?

The custom logic for a DynamoDB trigger is stored in an AWS Lambda function as code. To create a trigger for a given table, you can associate an AWS Lambda function to the stream (via DynamoDB Streams) on a DynamoDB table. When the table is updated, the updates are published to DynamoDB Streams. In turn, AWS Lambda reads the updates from the associated stream and executes the code in the function.

Q: What does it cost to use DynamoDB Triggers?

With DynamoDB Triggers, you only pay for the number of requests for your AWS Lambda function and the amount of time it takes for your AWS Lambda function to execute. Learn more about AWS Lambda pricing here. You are not charged for the reads that your AWS Lambda function makes to the stream (via DynamoDB Streams) associated with the table.

Q. Is there a limit to the number of triggers for a table?

There is no limit on the number of triggers for a table.

Q. What languages does DynamoDB Triggers support?

Currently, DynamoDB Triggers supports Javascript, Java, and Python for trigger functions.

Q. Is there API support for creating, editing or deleting DynamoDB triggers?

No, currently there are no native APIs to create, edit, or delete DynamoDB triggers. You have to use the AWS Lambda console to create an AWS Lambda function and associate it with a stream in DynamoDB Streams. For more information, see the AWS Lambda FAQ page.

Q. How do I create a DynamoDB trigger?

You can create a trigger by creating an AWS Lambda function and associating the event-source for the function to a stream in DynamoDB Streams. For more information, see the AWS Lambda FAQ page.

Q. How do I delete a DynamoDB trigger?

You can delete a trigger by deleting the associated AWS Lambda function. You can delete an AWS Lambda function from the AWS Lambda console or throughput an AWS Lambda API call. For more information, see the AWS Lambda FAQ and documentation page.

Q. I have an existing AWS Lambda function, how do I create a DynamoDB trigger using this function?

You can change the event source for the AWS Lambda function to point to a stream in DynamoDB Streams. You can do this from the DynamoDB console. In the table for which the stream is enabled, choose the stream, choose the Associate Lambda Function button, and then choose the function that you want to use for the DynamoDB trigger from the list of Lambda functions.

Q. In what regions is DynamoDB Triggers available?

DynamoDB Triggers is available in all AWS regions where AWS Lambda and DynamoDB are available.

DynamoDB Streams

Q: What is DynamoDB Streams?

DynamoDB Streams provides a time-ordered sequence of item-level changes made to data in a table in the last 24 hours. You can access a stream with a simple API call and use it to keep other data stores up-to-date with the latest changes to DynamoDB or to take actions based on the changes made to your table.

Q: What are the benefits of DynamoDB Streams?

Using the DynamoDB Streams APIs, developers can consume updates and receive the item-level data before and after items are changed. This can be used to build creative extensions to your applications built on top of DynamoDB. For example, a developer building a global multi-player game using DynamoDB can use the DynamoDB Streams APIs to build a multi-master topology and keep the masters in sync by consuming the DynamoDB Streams for each master and replaying the updates in the remote masters. As another example, developers can use the DynamoDB Streams APIs to build mobile applications that automatically notify the mobile devices of all friends in a circle as soon as a user uploads a new selfie. Developers could also use DynamoDB Streams to keep data warehousing tools, such as Amazon Redshift, in sync with all changes to their DynamoDB table to enable real-time analytics. DynamoDB also integrates with Elasticsearch using the Amazon DynamoDB Logstash Plugin, thus enabling developers to add free-text search for DynamoDB content.

You can read more about DynamoDB Streams in ourdocumentation.

Q: How long are changes to my DynamoDB table available via DynamoDB Streams?

DynamoDB Streams keep records of all changes to a table for 24 hours. After that, they will be erased.

Q: How do I enable DynamoDB Streams?

DynamoDB Streams have to be enabled on a per-table basis. To enable DynamoDB Streams for an existing DynamoDB table, select the table through the AWS Management Console, choose the Overview tab, click the Manage Stream button, choose a view type, and then click Enable.

For more information, see ourdocumentation.

Q: How do I verify that DynamoDB Streams has been enabled?

After enabling DynamoDB Streams, you can see the stream in the AWS Management Console. Select your table, and then choose the Overview tab. Under Stream details, verify Stream enabled is set to Yes.

Q: How can I access DynamoDB Streams?

You can access a stream available through DynamoDB Streams with a simple API call using the DynamoDB SDK or using the Kinesis Client Library (KCL). KCL helps you consume and process the data from a stream and also helps you manage tasks such as load balancing across multiple readers, responding to instance failures, and checkpointing processed records.

For more information about accessing DynamoDB Streams, see ourdocumentation.

Q: Does DynamoDB Streams display all updates made to my DynamoDB table in order?

Changes made to any individual item will appear in the correct order. Changes made to different items may appear in DynamoDB Streams in a different order than they were received.

For example, suppose that you have a DynamoDB table tracking high scores for a game and that each item in the table represents an individual player. If you make the following three updates in this order:

- Update 1: Change Player 1’s high score to 100 points
- Update 2: Change Player 2’s high score to 50 points
- Update 3: Change Player 1’s high score to 125 points

Update 1 and Update 3 both changed the same item (Player 1), so DynamoDB Streams will show you that Update 3 came after Update 1. This allows you to retrieve the most up-to-date high score for each player. The stream might not show that all three updates were made in the same order (i.e., that Update 2 happened after Update 1 and before Update 3), but updates to each individual player’s record will be in the right order.

Q: Do I need to manage the capacity of a stream in DynamoDB Streams?

No, capacity for your stream is managed automatically in DynamoDB Streams. If you significantly increase the traffic to your DynamoDB table, DynamoDB will automatically adjust the capacity of the stream to allow it to continue to accept all updates.

Q: At what rate can I read from DynamoDB Streams?

You can read updates from your stream in DynamoDB Streams at up to twice the rate of the provisioned write capacity of your DynamoDB table. For example, if you have provisioned enough capacity to update 1,000 items per second in your DynamoDB table, you could read up to 2,000 updates per second from your stream.

Q: If I delete my DynamoDB table, does the stream also get deleted in DynamoDB Streams?

No, not immediately. The stream will persist in DynamoDB Streams for 24 hours to give you a chance to read the last updates that were made to your table. After 24 hours, the stream will be deleted automatically from DynamoDB Streams.

Q: What happens if I turn off DynamoDB Streams for my table?

If you turn off DynamoDB Streams, the stream will persist for 24 hours but will not be updated with any additional changes made to your DynamoDB table.

Q: What happens if I turn off DynamoDB Streams and then turn it back on?

When you turn off DynamoDB Streams, the stream will persist for 24 hours but will not be updated with any additional changes made to your DynamoDB table. If you turn DynamoDB Streams back on, this will create a new stream in DynamoDB Streams that contains the changes made to your DynamoDB table starting from the time that the new stream was created.

Q: Will there be duplicates or gaps in DynamoDB Streams?

No, DynamoDB Streams is designed so that every update made to your table will be represented exactly once in the stream.

Q: What information is included in DynamoDB Streams?

A DynamoDB stream contains information about both the previous value and the changed value of the item. The stream also includes the change type (INSERT, REMOVE, and MODIFY) and the primary key for the item that changed.

Q: How do I choose what information is included in DynamoDB Streams?

For new tables, use the CreateTable API call and specify the ViewType parameter to choose what information you want to include in the stream.
For an existing table, use the UpdateTable API call and specify the ViewType parameter to choose what information to include in the stream.

The ViewType parameter takes the following values:

```
ViewType: {
            { KEYS_ONLY,
        NEW_IMAGE,
        OLD_IMAGE,
        NEW_AND_OLD_IMAGES}
    }
```

The values have the following meaning: KEYS_ONLY: Only the name of the key of items that changed are included in the stream.

- NEW_IMAGE: The name of the key and the item after the update (new item) are included in the stream.
- OLD_IMAGE: The name of the key and the item before the update (old item) are included in the stream.
- NEW_AND_OLD_IMAGES: The name of the key, the item before (old item) and after (new item) the update are included in the stream.

Q: Can I use my Kinesis Client Library to access DynamoDB Streams?

Yes, developers who are familiar with Kinesis APIs will be able to consume DynamoDB Streams easily. You can use the DynamoDB Streams Adapter, which implements the Amazon Kinesis interface, to allow your application to use the Amazon Kinesis Client Libraries (KCL) to access DynamoDB Streams. For more information about using the KCL to access DynamoDB Streams, please see our documentation.

Q: Can I change what type of information is included in DynamoDB Streams?

If you want to change the type of information stored in a stream after it has been created, you must disable the stream and create a new one using the UpdateTable API.

Q: When I make a change to my DynamoDB table, how quickly will that change show up in a DynamoDB stream?

Changes are typically reflected in a DynamoDB stream in less than one second.

Q: If I delete an item, will that change be included in DynamoDB Streams?

Yes, each update in a DynamoDB stream will include a parameter that specifies whether the update was a deletion, insertion of a new item, or a modification to an existing item. For more information on the type of update, see our documentation.

Q: After I turn on DynamoDB Streams for my table, when can I start reading from the stream?

You can use the DescribeStream API to get the current status of the stream. Once the status changes to ENABLED, all updates to your table will be represented in the stream.

You can start reading from the stream as soon as you start creating it, but the stream may not include all updates to the table until the status changes to ENABLED.

Q: What is the Amazon DynamoDB Logstash Plugin for Elasticsearch?

Elasticsearch is a popular open source search and analytics engine designed to simplify real-time search and big data analytics. Logstash is an open source data pipeline that works together with Elasticsearch to help you process logs and other event data. The Amazon DynamoDB Logstash Plugin make is easy to integrate DynamoDB tables with Elasticsearch clusters.

Q: How much does the Amazon DynamoDB Logstash Plugin cost?

The Amazon DynamoDB Logstash Plugin is free to download and use.

Q: How do I download and install the Amazon DynamoDB Logstash Plugin?

The Amazon DynamoDB Logstash Plugin is available on GitHub. Read our documentation page to learn more about installing and running the plugin.

DynamoDB Storage Backend for Titan

Q: What is the DynamoDB Storage Backend for Titan?

The DynamoDB Storage Backend for Titan is a plug-in that allows you to use DynamoDB as the underlying storage layer for Titan graph database. It is a client side solution that implements index free adjacency for fast graph traversals on top of DynamoDB.

Q: What is a graph database?

A graph database is a store of vertices and directed edges that connect those vertices. Both vertices and edges can have properties stored as key-value pairs.

A graph database uses adjacency lists for storing edges to allow simple traversal. A graph in a graph database can be traversed along specific edge types, or across the entire graph. Graph databases can represent how entities relate by using actions, ownership, parentage, and so on.

Q: What applications are well suited to graph databases?

Whenever connections or relationships between entities are at the core of the data you are trying to model, a graph database is a natural choice. Therefore, graph databases are useful for modeling and querying social networks, business relationships, dependencies, shipping movements, and more.

Q: How do I get started using the DynamoDB Storage Backend for Titan?

The easiest way to get started is to launch an EC2 instance running Gremlin Server with the DynamoDB Storage Backend for Titan, using the CloudFormation templates referred to in this documentation page. You can also clone the project from the GitHub repository and start by following the Marvel and Graph-Of-The-Gods tutorials on your own computer by following the instructions in the documentation here. When you're ready to expand your testing or run in production, you can switch the backend to use the DynamoDB service. Please see the AWS documentation for further guidance.

Q: How does the DynamoDB Storage Backend differ from other Titan storage backends?

DynamoDB is a managed service, thus using it as the storage backend for Titan enables you to run graph workloads without having to manage your own cluster for graph storage.

Q: Is the DynamoDB Storage Backend for Titan a fully managed service?

No. The DynamoDB storage backend for Titan manages the storage layer for your Titan workload. However, the plugin does not do provisioning and managing of the client side. For simple provisioning of Titan we have developed a CloudFormation template that sets up DynamoDB Storage Backend for Titan with Gremlin Server; see the instructions available here.

Q: How much does using the DynamoDB Storage Backend for Titan cost?

You are charged the regular DynamoDB throughput and storage costs. There is no additional cost for using DynamoDB as the storage backend for a Titan graph workload.

Q: Does DynamoDB backend provide full compatibility with the Titan feature set on other backends?

A table comparing feature sets of different Titan storage backends is available in the documentation.

Q: Which versions of Titan does the plugin support?

We have released DynamoDB storage backend plugins for Titan versions 0.5.4 and 1.0.0.

Q: I use Titan with a different backend today. Can I migrate to DynamoDB?

Absolutely. The DynamoDB Storage Backend for Titan implements the Titan KCV Store interface so you can switch from a different storage backend to DynamoDB with minimal changes to your application. For full comparison of storage backends for Titan please see our documentation.

Q: I use Titan with a different backend today. How do I migrate to DynamoDB?

You can use bulk loading to copy your graph from one storage backend to the DynamoDB Storage Backend for Titan.

Q: How do I connect my Titan instance to DynamoDB via the plugin?

If you create a graph and Gremlin server instance with the DynamoDB Storage Backend for Titan installed, all you need to do to connect to DynamoDB is provide a principal/credential set to the default AWS credential provider chain. This can be done with an EC2 instance profile, environment variables, or the credentials file in your home folder. Finally, you need to choose a DynamoDB endpoint to connect to.

Q: How durable is my data when using the DynamoDB Storage Backend for Titan?

When using the DynamoDB Storage Backend for Titan, your data enjoys the strong protection of DynamoDB, which runs across Amazon's proven, high-availability data centers. The service replicates data across three facilities in an AWS Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: How secure is the DynamoDB Storage Backend for Titan?

The DynamoDB Storage Backend for Titan stores graph data in multiple DynamoDB tables, thus is enjoys the same high security available on all DynamoDB workloads. Fine-Grained Access Control, IAM roles, and AWS principal/credential sets control access to DynamoDB tables and items in DynamoDB tables.

Q: How does the DynamoDB Storage Backend for Titan scale?

The DynamoDB Storage Backend for Titan scales just like any other workload of DynamoDB. You can choose to increase or decrease the required throughput at any time.

Q: How many vertices and edges can my graph contain?

You are limited by Titan's limits for (2^60) for the maximum number of edges and half as many vertices in a graph, as long as you use the multiple-item model for edgestore. If you use the single-item model, the number of edges that you can store at a particular out-vertex key is limited by DynamoDB's maximum item size, currently 400kb.

Q: How large can my vertex and edge properties get?

The sum of all edge properties in the multiple-item model cannot exceed 400kb, the maximum item size. In the multiple item model, each vertex property can be up to 400kb. In the single-item model, the total item size (including vertex properties, edges and edge properties) can't exceed 400kb.

Q: How many data models are there? What are the differences?

There are two different storage models for the DynamoDB Storage Backend for Titan – single item model and multiple item model. In the single item storage model, vertices, vertex properties, and edges are stored in one item. In the multiple item data model, vertices, vertex properties and edges are stored in different items. In both cases, edge properties are stored in the same items as the edges they correspond to.

Q: Which data model should I use?

In general, we recommend you use the multiple-item data model for the edgestore and graphindex tables. Otherwise, you either limit the number of edges/vertex-properties you can store for one out-vertex, or you limit the number of entities that can be indexed at a particular property name-value pair in graph index. In general, you can use the single-item data model for the other 4 KCV stores in Titan versions 0.5.4 and 1.0.0 because the items stored in them are usually less than 400KB each. For full list of tables that the Titan plugin creates on DynamoDB please see here.

Q: Do I have to create a schema for Titan graph databases?

Titan supports automatic type creation, so new edge/vertex properties and labels will get registered on the fly (see here for details) with the first use. The Gremlin Structure (Edge labels=MULTI, Vertex properties=SINGLE) is used by default.

Q: Can I change the schema of a Titan graph database?

Yes, however, you cannot change the schema of existing vertex/edge properties and labels. For details please see here.

Q: How does the DynamoDB Storage Backend for Titan deal with supernodes?

DynamoDB deals with supernodes via vertex label partitioning. If you define a vertex label as partitioned in the management system upon creation, you can key different subsets of the edges and vertex properties going out of a vertex at different partition keys of the partition-sort key space in the edgestore table. This usually results in the virtual vertex label partitions being stored in different physical DynamoDB partitions, as long as your edgestore has more than one physical partition. To estimate the number of physical partitions backing your edgestore table, please see guidance in the documentation.

Q: Does the DynamoDB Storage Backend for Titan support batch graph operations?

Yes, the DynamoDB Storage Backend for Titan supports batch graph with the Blueprints BatchGraph implementation and through Titan’s bulk loading configuration options.

Q: Does the DynamoDB Storage Backend for Titan support transactions?

The DynamoDB Storage Backend for Titan supports optimistic locking. That means that the DynamoDB Storage Backend for Titan can condition writes of individual Key-Column pairs (in the multiple item model) or individual Keys (in the single item model) on the existing value of said Key-Column pair or Key.

Q: Can I have a Titan instance in one region and access DynamoDB in another?

Accessing a DynamoDB endpoint in another region than the EC2 Titan instance is possible but not recommended. When running a Gremlin Server out of EC2, we recommend connecting to the DynamoDB endpoint in your EC2 instance’s region, to reduce the latency impact of cross-region requests. We also recommend running the EC2 instance in a VPC to improve network performance. The CloudFormation template performs this entire configuration for you.

Q: Can I use this plugin with other DynamoDB features such as update streams and cross-region replication?

You can use Cross-Region Replication with the DynamoDB Streams feature to create read-only replicas of your graph tables in other regions.

DynamoDB CloudWatch Metrics

Q: Does Amazon DynamoDB report CloudWatch metrics?

Yes, Amazon DynamoDB reports several table-level metrics on CloudWatch. You can make operational decisions about your Amazon DynamoDB tables and take specific actions, like setting up alarms, based on these metrics. For a full list of reported metrics, see the *Monitoring DynamoDB with CloudWatch* section of our documentation.

Q: How can I see CloudWatch metrics for an Amazon DynamoDB table?

On the Amazon DynamoDB console, select the table for which you wish to see CloudWatch metrics and then select the Metrics tab.

Q: How often are metrics reported?

Most CloudWatch metrics for Amazon DynamoDB are reported in 1-minute intervals while the rest of the metrics are reported in 5-minute intervals. For more details, see the *Monitoring DynamoDB with CloudWatch* section of our documentation.

Amazon ElastiCache FAQ

Redis

- Features
- Read Replica
- Multi-AZ
- Backup and Restore
- Redis Cluster
- Enhanced Engine

General

The Basics

Q: What is Amazon ElastiCache?

Amazon ElastiCache is a web service that makes it easy to deploy and run Memcached or Redis protocol-compliant server nodes in the cloud. Amazon ElastiCache improves the performance of web applications by allowing you to retrieve information from a fast, managed, in-memory system, instead of relying entirely on slower disk-based databases. The service simplifies and offloads the management, monitoring and operation of in-memory environments, enabling your engineering resources to focus on developing applications. Using Amazon ElastiCache, you can not only improve load and response times to user actions and queries, but also reduce the cost associated with scaling web applications.

Amazon ElastiCache automates common administrative tasks required to operate a distributed in-memory key-value environment. Using Amazon ElastiCache, you can add a caching or in-memory layer to your application architecture in a matter of minutes via a few clicks of the AWS Management Console. Once a cluster is provisioned, Amazon ElastiCache automatically detects and replaces failed nodes, providing a resilient system that mitigates the risk of overloaded databases, which slow website and application load times. Through integration with Amazon CloudWatch monitoring, Amazon ElastiCache provides enhanced visibility into key performance metrics associated with your nodes. Amazon ElastiCache is protocol-compliant with Memcached and Redis, so code, applications, and popular tools that you use today with your existing Memcached or Redis environments will work seamlessly with the service. With the support for clustered configuration in Amazon ElastiCache, you get the benefits of fast, scalable and easy to use managed service that can meet the needs of your most demanding applications. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

Q: What is in-memory caching and how does it help my applications?

The in-memory caching provided by Amazon ElastiCache can be used to significantly improve latency and throughput for many read-heavy application workloads (such as social networking, gaming, media sharing and Q&A portals) or compute-intensive workloads (such as a recommendation engine). In-memory caching improves application performance by storing critical pieces of data in memory for low-latency access. Cached information may include the results of I/O-intensive database queries or the results of computationally-intensive calculations.

Q: Can I use Amazon ElastiCache for use cases other than caching?

A: Yes. ElastiCache for Redis can be used as a primary in-memory key-value data store, providing fast, sub millisecond data performance, high availability and scalability up to 16 nodes plus up to 5 read replicas, each of up to 3.55 TiB of in-memory data. See here for other use cases, such as leaderboards, rate limiting, queues, and chat.

Q: Can I use Amazon ElastiCache through AWS CloudFormation?

AWS CloudFormation simplifies provisioning and management by providing AWS CloudFormation templates for quick and reliable provisioning of the services or applications. AWS CloudFormation provides comprehensive support for Amazon ElastiCache by providing templates to create cluster (both MemCached and Redis) and Replication Groups. The templates are up to date with the latest ElastiCache Redis announcement for clustered Redis configuration and provide flexibility and ease of use to Amazon ElastiCache customers.

Q: What does Amazon ElastiCache manage on my behalf?

Amazon ElastiCache manages the work involved in setting up a distributed in-memory environment, from provisioning the server resources you request to installing the software. Once your environment is up and running, the service automates common administrative tasks such as failure detection and recovery, and software patching. Amazon ElastiCache provides detailed monitoring metrics associated with your nodes, enabling you to diagnose and react to issues very quickly. For example, you can set up thresholds and receive alarms if one of your nodes is overloaded with requests.

Q: What are Amazon ElastiCache nodes, shards and clusters?

A node is the smallest building block of an Amazon ElastiCache deployment. It is a fixed-size chunk of secure, network-attached RAM. Each node runs an instance of the Memcached or Redis protocol-compliant service and has its own DNS name and port. Multiple types of nodes are supported, each with varying amount of associated memory. A Redis shard is a subset of the cluster's keyspace, that can include a primary node and zero or more read-replicas. For more details on Redis deployments see the Redis section below. The shards add up to form a cluster.

Q: Which engines does Amazon ElastiCache support?

Amazon ElastiCache for Memcached currently supports Memcached 1.4.5, 1.4.14 and 1.4.24.

Amazon ElastiCache for Redis currently supports Redis 2.8.21, 2.8.22, 2.8.23, 2.8.24 and 3.2.4.

Q: How do I get started with Amazon ElastiCache?

If you are not already signed up for Amazon ElastiCache, you can click the "Sign Up Now" button on the Amazon ElastiCache detail page and complete the sign-up process. You must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one when you begin the Amazon ElastiCache sign-up process. After you are signed up for ElastiCache, please refer to the Amazon ElastiCache documentation, which includes our Getting Started Guide.

Once you have familiarized yourself with Amazon ElastiCache, you can launch a cluster within minutes by using the AWS Management Console or Amazon ElastiCache APIs.

Q: How do I create a cluster?

Clusters are simple to create, using theAWS Management Console, Amazon ElastiCache APIs, or Command Line Tools. To launch a cluster using the AWS Management Console, click on the "Create" button in either the “Memcached” or “Redis” tab. From there, all you need to specify is your Cluster Identifier, Node Type, and Number of Nodes to create a cluster with the amount of memory you require. Alternatively, you can create your cluster using the CreateCacheCluster API or elasticache-create-cache-cluster command. If you do not specify an Availability Zone when creating a cluster, AWS will place it automatically based upon your memory requirements and available capacity.

Q: What Node Types can I select?

Amazon ElastiCache supports Nodes of the following types:

Current Generation Nodes:

- cache.m3.medium: 2.78 GB
- cache.m3.large: 6.05 GB
- cache.m3.xlarge: 13.3 GB
- cache.m3.2xlarge: 27.9 GB
- cache.m4.large: 6.42 GB
- cache.m4.xlarge: 14.28 GB
- cache.m4.2xlarge: 29.7 GB
- cache.m4.4xlarge: 60.78 GB
- cache.m4.10xlarge: 154.64 GB
- cache.r3.large: 13.5 GB
- cache.r3.xlarge: 28.4 GB
- cache.r3.2xlarge: 58.2 GB
- cache.r3.4xlarge: 118 GB

- cache.r3.8xlarge: 237 GB
- cache.t2.micro: 555 MB
- cache.t2.small: 1.55 GB
- cache.t2.medium: 3.22 GB

Previous Generation Nodes:

- cache.m1.small: 1.3 GB
- cache.m1.medium: 3.35 GB
- cache.m1.large: 7.1 GB
- cache.m1.xlarge: 14.6 GB
- cache.m2.xlarge: 16.7 GB
- cache.m2.2xlarge: 33.8 GB
- cache.m2.4xlarge: 68 GB
- cache.t1.micro: 213 MB
- cache.c1.xlarge: 6.6 GB

Each Node Type above lists the memory available to Memcached or Redis after taking Amazon ElastiCache System Software overhead into account. The total amount of memory in a cluster is an integer multiple of the memory available in each shard. For example, a cluster consisting of ten shards of 6 GB each will provide 60 GB of total memory.

Q: How do I access my nodes?

Once your cluster is available, you can retrieve your node endpoints using the following steps on the AWS Management Console:

- Navigate to the "Amazon ElastiCache" tab.
- Click on the "(Number of) Nodes" link and navigate to the "Nodes" tab.
- Click on the "Copy Node Endpoint(s)" button.

Alternatively, you can use the DescribeCacheClusters API to retrieve the Endpoint list.

You can then configure your Memcached or Redis client with this endpoint list and use your favorite programming language to add or delete data from your ElastiCache Nodes. In order to allow network requests to your nodes, you will need to authorize access. For a detailed explanation to get started, please refer to our Getting Started Guide.

Q: What is a maintenance window? Will my nodes be available during software maintenance?

You can think of the Amazon ElastiCache maintenance window as an opportunity to control when software patching occurs, in the event either are requested or required. If a "maintenance" event is scheduled for a given week, it will be initiated and completed at some point during the 60 minute maintenance window you identify.

Your nodes could incur some downtime during your maintenance window if software patching is scheduled. Please refer to Engine Version Management for more details. Patching can be user requested - for example cache software upgrade, or determined as required (if we identify any security vulnerabilities in the system or caching software). Software patching occurs infrequently (typically once every few months) and should seldom require more than a fraction of your maintenance window. If you do not specify a preferred weekly maintenance window when creating your Cluster, a 60 minute default value is assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your DB Instance in the AWS Management Console or by using the ModifyCacheCluster API. Each of your Clusters can have different preferred maintenance windows, if you so choose.

Billing

Q: How will I be charged and billed for my use of Amazon ElastiCache?

You pay only for what you use and there is no minimum fee. Pricing is per Node-hour consumed for each Node Type. Partial Node-hours consumed are billed as full hours. There is no charge for data transfer between Amazon EC2 and Amazon ElastiCache within the same Availability Zone. While standard Amazon EC2 Regional Data Transfer charges apply when transferring data between an Amazon EC2 instance and an Amazon ElastiCache Node in different Availability Zones of the same Region, you are only charged for the Data Transfer in or out of the Amazon EC2 instance. There is no Amazon ElastiCache Data Transfer charge for traffic in or out of the Amazon ElastiCache Node itself. For more information, please visit the pricing page.

Q: When does billing of my Amazon ElastiCache Nodes begin and end?

Billing commences for a node as soon as the node is available. Billing continues until the node is terminated, which would occur upon deletion.

Q: What defines billable ElastiCache Node hours?

Node hours are billed for any time your nodes are running in an "Available" state. If you no longer wish to be charged for your node, you must terminate it to avoid being billed for additional node hours.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Reserved Nodes

Q: What are Amazon ElastiCache Reserved Nodes?

With Reserved Nodes, you can now make a one-time, up-front payment to create a one or three year reservation to run your node in a specific Region and receive a significant discount off of the ongoing hourly usage charge. There are three Reserved Node types (Light, Medium, and Heavy Utilization Reserved Nodes) that enable you to balance the amount you pay upfront with your effective hourly price.

Q: How are Reserved Nodes different from On-Demand Nodes?

Functionally, Reserved Nodes and On-Demand Nodes are exactly the same. The only difference is how your Node(s) are billed; with Reserved Nodes, you make a one-time up-front payment and receive a lower ongoing hourly usage rate (compared with On-Demand Nodes) for the duration of the term.

Q: How do I purchase and create Reserved Nodes?

You can use the "Purchase Reserved Nodes" option in the AWS Management Console. Alternatively, you can use the API tools to list the reservations available for purchase with the DescribeReservedCacheNodesOfferings API method and then purchase a cache node reservation by calling the PurchaseReservedCacheNodesOffering method.

Creating a Reserved Node is no different than launching an On-Demand Node. You simply specify the node class and Region for which you made the reservation. So long as your reservation purchase was successful, Amazon ElastiCache will apply the reduced hourly rate for which you are eligible to the new node.

Q: Will there always be reservations available for purchase?

Yes. Reserved Nodes are purchased for the Region rather than for the Availability Zone. This means that even if capacity is limited in one Availability Zone, reservations can still be purchased in that Region and used in a different Availability Zone within that Region.

Q: How many Reserved Cache can I purchase?

You can purchase up to 20 Reserved Nodes. If you wish to run more than 20 Nodes please complete the Amazon ElastiCache Node request form.

Q: What if I have an existing node that I'd like to convert to a Reserved Node?

Simply purchase a node reservation with the same node class, within the same region as the node you are currently running and would like to reserve. If the reservation purchase is successful, Amazon ElastiCache will automatically apply your new hourly usage charge to your existing node.

Q: If I sign up for a Reserved Node, when does the term begin? What happens to my node when the term ends?

Pricing changes associated with a Reserved Node are activated once your request is received while the payment authorization is processed. You can follow the status of your reservation on the AWS Account Activity page or by using the DescribeReservedCacheNodes API. If the one-time payment cannot be successfully authorized by the next billing period, the discounted price will not take effect.

When your reservation term expires, your Reserved Node will revert to the appropriate On-Demand hourly usage rate for your node class and region.

Q: How do I control which nodes are billed at the Reserved Node rate?

The Amazon ElastiCache APIs for creating, modifying, and deleting nodes do not distinguish between On-Demand and Reserved Nodes so that you can seamlessly use both. When computing your bill, our system will automatically apply your Reservation(s), such that all eligible nodes are charged at the lower hourly Reserved Cache Node rate.

Q: Can I move a Reserved Node from one Region or Availability Zone to another?

Each Reserved Node is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each reservation can, however, be used in any of the available AZs within the associated Region.

Q: Can I cancel a reservation?

The one-time payment for Reserved Nodes is not refundable. However, you can choose to terminate your node at any time, at which point you will not incur any hourly usage charges if you are using Light and Medium Utilization Reserved Nodes.

Security

Q: How do I control access to Amazon ElastiCache?

When not using VPC, Amazon ElastiCache allows you to control access to your clusters through Cache Security Groups. A Security Group acts like a firewall, controlling network access to your cluster. By default, network access is turned off to your clusters. If you want your applications to access your cluster, you must explicitly enable access from hosts in specific EC2 security groups. This process is called ingress.

To allow network access to your cluster, create a Security Group and link the desired EC2 security groups (which in turn specify the EC2 instances allowed) to it. The Security Group can be associated with your cluster at the time of creation, or using the "Modify" option on the AWS Management Console.

Please note that IP-range based access control is currently not enabled for clusters. All clients to a cluster must be within the EC2 network, and authorized via security groups as described above.

When using VPC, please see here for more information.

Q: Can programs running on servers in my own data center access Amazon ElastiCache?

No. Currently, all clients to an ElastiCache Cluster must be within the Amazon EC2 network, and authorized via security groups as described here.

Q: Can programs running on EC2 instances in a VPC access Amazon ElastiCache?

Yes, EC2 instances in a VPC can access Amazon ElastiCache if the ElastiCache cluster was created within the VPC. Details on how to create an Amazon ElastiCache cluster within a VPC are given here.

Q: What is Amazon Virtual Private Cloud (VPC) and why may I want to use with Amazon ElastiCache?

Amazon VPC lets you create a virtual networking environment in a private, isolated section of the Amazon Web Services (AWS) cloud, where you can exercise complete control over aspects such as private IP address ranges, subnets, routing tables and network gateways. With Amazon VPC, you can define a virtual network topology and customize the network configuration to closely resemble a traditional IP network that you might operate in your own datacenter.

One of the scenarios where you may want to use Amazon ElastiCache in a VPC is if you want to run a public-facing web application, while still maintaining non-publicly accessible backend servers in a private subnet. You can create a public-facing subnet for your webservers that has access to the Internet, and place your backend infrastructure in a private-facing subnet with no Internet access. Your backend infrastructure could include RDS DB Instances and an Amazon ElastiCache Cluster providing the in-memory layer. For more information about Amazon VPC, refer to the Amazon Virtual Private Cloud User Guide.

Q: How do I create an Amazon ElastiCache Cluster in VPC?

For a walk through example of creating an Amazon ElastiCache Cluster in VPC, refer to the Amazon ElastiCache User Guide.

Following are the pre-requisites necessary to create a cluster within a VPC:

- You need to have a VPC set up with at least one subnet. For information on creating Amazon VPC and subnets refer to the Getting Started Guide for Amazon VPC.
- You need to have a Subnet Group defined for your VPC.
- You need to have a VPC Security Group defined for your VPC (or you can use the default provided).
- In addition, you should allocate adequately large CIDR blocks to each of your subnets so that there are enough spare IP addresses for Amazon ElastiCache to use during maintenance activities such as cache node replacement.

Q: How do I create an Amazon ElastiCache Cluster in an existing VPC?

Creating an Amazon ElastiCache Cluster in an existing VPC is the same as that for a newly created VPC. Please see this for more details.

Q: How do I connect to an ElastiCache Node in VPC?

Amazon ElastiCache Nodes, deployed within a VPC, can be accessed by EC2 Instances deployed in the same VPC. If these EC2 Instances are deployed in a public subnet with associated Elastic IPs, you can access the EC2 Instances via the internet.

Amazon ElastiCache Nodes, deployed within a VPC, can never be accessed from the Internet or from EC2 Instances outside the VPC.

We strongly recommend you use the DNS Name to connect to your ElastiCache Node as the underlying IP address can change (e.g., after a cache node replacement).

Q: What is a Subnet Group and why do I need one?

A Subnet Group is a collection of subnets that you must designate for your Amazon ElastiCache Cluster in a VPC. A Subnet Group is created using the Amazon ElastiCache Console. Each Subnet Group should have at least one subnet. Amazon ElastiCache uses the Subnet Group to select a subnet. The IP Addresses from the selected subnet are then associated with the Node Endpoints. Furthermore, Amazon ElastiCache creates and associates Elastic Network Interfaces to nodes with the previously mentioned IP addresses.

Please note that, we strongly recommend you use the DNS Names to connect to your nodes as the underlying IP addresses can change (e.g., after cache node replacement).

Q: Can I change the Subnet Group of my ElastiCache Cluster?

An existing Subnet Group can be updated to add more subnets either for existing Availability Zones are for new Availability Zones added since the creation of the ElastiCache Cluster. However, changing the Subnet Group of a deployed cluster is not currently allowed.

Q: How is using Amazon ElastiCache inside a VPC different from using it outside?

The basic functionality of Amazon ElastiCache remains the same whether VPC is used or not. Amazon ElastiCache manages automatic failure detection, recovery, scaling, auto discovery, and software patching whether your ElastiCache Cluster is inside or outside a VPC.

Similarly, an Amazon ElastiCache Cluster, inside or outside a VPC, is never allowed to be accessed from the Internet. Within a VPC, nodes of an ElastiCache cluster only have a private IP address (within a subnet that you define). Outside of a VPC, the access to the ElastiCache cluster can be controlled using Security Groups as described here.

Q: Can I move my existing ElastiCache Cluster from outside VPC into my VPC?

No, you cannot move an existing Amazon ElastiCache Cluster from outside VPC into a VPC. You will need to create a new Amazon ElastiCache Cluster inside the VPC.

Q: Can I move my existing ElastiCache Cluster from inside VPC to outside VPC?

Currently, direct migration of ElastiCache Cluster from inside to outside VPC is not supported. You will need to create a new Amazon ElastiCache Cluster outside VPC.

Q: How do I control network access to my cluster?

Amazon ElastiCache allows you to control access to your cluster and therefore the nodes using Security Groups in non-VPC deployments. A Security Group acts like a firewall controlling network access to your node. By default, network access is turned off to your nodes. If you want your applications to access your node, you can set your Security Group to allow access from EC2 Instances with specific EC2 Security Group membership or IP ranges. This process is called ingress. Once ingress is configured for a Security Group, the same rules apply to all nodes associated with that Security Group. Security Groups can be configured with the “Security Groups” section of the Amazon ElastiCache Console or using the Amazon ElastiCache APIs.

In VPC deployments, access to your nodes is controlled using the VPC Security Group and the Subnet Group. The VPC Security Group is the VPC equivalent of the Security Group.

Q: What precautions should I take to ensure that my ElastiCache Nodes in VPC are accessible by my application?

You are responsible for modifying routing tables and networking ACLs in your VPC to ensure that your ElastiCache Nodes are reachable from your client instances in the VPC. To learn more see the Amazon ElastiCache Documentation.

Q: Can I use Security Groups to configure the clusters that are part of my VPC?

No, Security Groups are not used when operating in a VPC. Instead they are used in the non VPC settings. When creating a cluster in a VPC you will need to use VPC Security Groups.

Q: Can I associate a regular EC2 security group with a cluster that is launched within a

VPC?

No, you can only associate VPC security groups that are part of the same VPC as your cluster.

Q: Can nodes of an ElastiCache cluster span multiple subnets?

Yes, nodes of an Amazon ElastiCache cluster can span multiple subnets as long as the subnets are part of the same Subnet Group that was associated with the ElastiCache Cluster at creation time.

Parameter Groups

Q: What are Parameter Groups? How are they helpful?

A Parameter Group acts as a "container" for engine configuration values that can be applied to one or more clusters. If you create a cluster without specifying a Parameter Group, a default Parameter Group is used. This default group contains engine defaults and Amazon ElastiCache system defaults optimized for the cluster you are running. However, if you want your cluster to run with your custom-specified engine configuration values, you can simply create a new Parameter Group, modify the desired parameters, and modify the cluster to use the new Parameter Group. Once associated, all clusters that use a particular Parameter Group get all the parameter updates to that Parameter Group. For more information on configuring Parameter Groups, please refer to the Amazon ElastiCache User Guide.

Q: How do I choose the right configuration parameters for my Cluster(s)?

Amazon ElastiCache by default chooses the optimal configuration parameters for your cluster taking into account the Node Type's memory/compute resource capacity. However, if you want to change them, you can do so using our configuration management APIs. Please note that changing configuration parameters from recommended values can have unintended effects, ranging from degraded performance to system crashes, and should only be attempted by advanced users who wish to assume these risks. For more information on changing parameters, please refer to the Amazon ElastiCache User Guide.

Q: How do I see the current setting for my parameters for a given Parameter Group?

You can use the AWS Management Console, Amazon ElastiCache APIs, or Command Line Tools to see information about your Parameter Groups and their corresponding parameter settings.

Memcached

Features

Q: What can I cache using Amazon ElastiCache for Memcached?

You can cache a variety of objects using the service, from the content in persistent data stores (such as Amazon RDS, SimpleDB, or self-managed databases hosted on EC2) to dynamically generated web pages (with Nginx for example), or transient session data that may not require a persistent backing store. You can also use it to implement high-frequency counters to deploy admission control in high volume web applications.

Q: Can I use Amazon ElastiCache for Memcached with an AWS persistent data store such as Amazon SimpleDB or Amazon RDS?

Yes, Amazon ElastiCache is an ideal front-end for data stores like Amazon SimpleDB and Amazon RDS, providing a high-performance middle tier for applications with extremely high request rates and/or low latency requirements.

Q: I use Memcached today. How do I migrate to Amazon ElastiCache?

Amazon ElastiCache is protocol-compliant with Memcached. Therefore, you can use standard Memcached operations like get, set, incr and decr in exactly the same way as you would in your existing Memcached deployments. Amazon ElastiCache supports both the text and binary protocols. It also supports most of the standard stats results, which can also be viewed as graphs via CloudWatch. As a result, you can switch to using Amazon ElastiCache without recompiling or re-linking your applications - the libraries you use will continue to work. To configure the cache servers your application accesses, all you will need to do is to update your application's Memcached config file to include the endpoints of the servers (nodes) we provision for you. You can simply use the "Copy Node Endpoints" option on the AWS Management Console or the "DescribeCacheClusters" API to get a list of the endpoints. As with any migration process, we recommend thorough testing of your new Amazon ElastiCache deployment before completing the cut over from your current solution.

Please note that Amazon ElastiCache currently allows access only from the Amazon EC2 network, so in order to use the service, you should have your application servers in Amazon EC2.

Amazon ElastiCache uses DNS entries to allow client applications to locate servers (nodes). The DNS name for a node remains constant, but the IP address of a node can change over time, for example, when nodes are auto replaced after a failure on a non-VPC installation. See this FAQ for recommendations to deal with node failures.

Configuration and Scaling

Q: How do I select an appropriate Node Type for my application?

Though there is no precise answer for this question, with Amazon ElastiCache, you don't need to worry about getting the number of nodes exactly right, as you can very easily add or remove nodes later. The following two inter-related aspects could be considered for the choice of your initial configuration:

- The total memory required for your data to achieve your target cache-hit rate, and
- The number of nodes required to maintaining acceptable application performance without overloading the database backend in the event of node failure(s).

The amount of memory required is dependent upon the size of your data set and the access patterns of your application. To improve fault tolerance, once you have a rough idea of the total memory required, divide that memory into enough nodes such that your application can survive the loss of one or two nodes. For example, if your memory requirement is 13GB, you may want to use two cache.m3.large nodes instead of using one cache.m3.xlarge node. It is important that other systems such as databases will not be overloaded if the cache-hit rate is temporarily reduced during failure recovery of one or more of nodes. Please refer to the Amazon ElastiCache User Guide for more details.

Q: Can a cluster span multiple Availability Zones?

Yes. When creating a cluster or adding nodes to an existing cluster, you can chose the availability zones for the new nodes. You can either specify the requested amount of nodes in each availability zones or select “spread nodes across zones”. If the cluster is in VPC, nodes can only be placed in availability zones that are part of the selected cache subnet group. For additional details please see ElastiCache VPC documentation.

Q: How many nodes can I run in Amazon ElastiCache?

You can run a maximum of 50 nodes per region. If you need more nodes, please fill in the ElastiCache Limit Increase Request form.

Q: How does Amazon ElastiCache respond to node failure?

The service will detect the node failure and react with the following automatic steps:

- Amazon ElastiCache will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. For VPC installations, ElastiCache will ensure that both the DNS name and the IP address of the node remain the same when nodes are recovered in case of failure. For non-VPC installations, ElastiCache will ensure that the DNS name of a node is unchanged; however, the underlying IP address of the node can change.
- If you associated an SNS topic with your cluster, when the new node is configured and ready

to be used, Amazon ElastiCache will send an SNS notification to let you know that node recovery occurred. This allows you to optionally arrange for your applications to force the Memcached client library to attempt to reconnect to the repaired nodes. This may be important, as some Memcached libraries will stop using a server (node) indefinitely if they encounter communication errors or timeouts with that server.

Q: If I determine that I need more memory to support my application, how do I increase the total memory with Amazon ElastiCache?

You could add more nodes to your existing Memcached Cluster by using the "Add Node" option on "Nodes" tab for your Cache Cluster on the AWS Management Console or calling the ModifyCacheCluster API.

Compatibility

Q: How does Amazon ElastiCache interact with other Amazon Web Services?

Amazon ElastiCache is ideally suited as a front-end for Amazon Web Services like Amazon RDS and Amazon DynamoDB, providing extremely low latency for high performance applications and offloading some of the request volume while these services provide long lasting data durability. The service can also be used to improve application performance in conjunction with Amazon EC2 and EMR.

Q: Is Amazon ElastiCache better suited to any specific programming language?

Memcached client libraries are available for many, if not all of the popular programming languages. For more information on Memcached clients, please see this. If you encounter any issues with specific Memcached clients when using Amazon ElastiCache, please engage us via the Amazon ElastiCache community forum.

Q: What popular Memcached libraries are compatible with Amazon ElastiCache?

Amazon ElastiCache does not require specific client libraries and works with existing Memcached client libraries without recompilation or application re-linking (Memcached 1.4.5 and later); examples include libMemcached (C) and libraries based on it (e.g. PHP, Perl, Python), spyMemcached (Java) and fauna (Ruby).

Auto Discovery

Q: What is Auto Discovery and what can I do with it?

Auto Discovery is a feature that saves developers time and effort, while reducing complexity of their applications. Auto Discovery enables automatic discovery of cache nodes by clients when they are added to or removed from an Amazon ElastiCache cluster. Until now to handle cluster membership changes, developers must update the list of cache node endpoints manually. Depending on how the client application is architected, typically a client initialization, by shutting down the application and restarting it, is needed resulting in downtime. Through Auto Discovery we are eliminating this complexity. With Auto Discovery, in addition to being backwards protocol-compliant with the Memcached protocol, Amazon ElastiCache provides clients with information on cache cluster membership. A client capable of processing the additional information reconfigures itself, without any initialization, to use the most current nodes of an Amazon ElastiCache cluster.

Q: How does Auto Discovery work?

An Amazon ElastiCache cluster can be created with nodes that are addressable via named endpoints. With Auto Discovery the Amazon ElastiCache cluster is also given a unique Configuration Endpoint which is a DNS Record that is valid for the lifetime of the cluster. This DNS Record contains the DNS Names of the nodes that belong to the cluster. Amazon ElastiCache will ensure that the Configuration Endpoint always points to at least one such “target” node. A query to the target node then returns endpoints for all the nodes of the cluster in question. After this, you can connect to the cluster nodes just as before and use the Memcached protocol commands such as get, set, incr and decr. For more details, see here. To use Auto Discovery, you will need an Auto Discovery capable client. Auto Discovery clients for Java and PHP are available for download from the Amazon ElastiCache console. Upon initialization, the client will automatically determine the current members of the Amazon ElastiCache cluster using the Configuration Endpoint. When you make changes to your cache cluster by adding or removing nodes or if a node is replaced upon failure, the Auto Discovery client automatically determines the changes and you do not need to initialize your clients manually.

Q: How can I get started using Auto Discovery?

To get started, download the Amazon ElastiCache Cluster Client by clicking the “Download ElastiCache Cluster Client” link on the Amazon ElastiCache console. Before you can download, you must have an Amazon ElastiCache account; if you do not already have one, you can sign up from the Amazon ElastiCache detail page. After you download the client, you can begin setting up and activating your Amazon ElastiCache cluster by visiting the Amazon ElastiCache console. More details can be found here.

Q: If I continue to use my own Memcached clients with my ElastiCache cluster – will I be able to get this feature?

No, you will not get the Auto Discovery feature with the existing Memcached clients. To use the Auto Discovery feature a client must be able to use a Configuration Endpoint and determine the cluster node endpoints. You may either use the Amazon ElastiCache Cluster Client or extend your existing Memcached client to include the Auto Discovery command set.

Q: What are the minimum hardware / software requirements for Auto Discovery?

To take advantage of Auto Discovery, an Auto Discovery capable client must be used to connect to an Amazon ElastiCache Cluster. Amazon ElastiCache currently supports Auto Discovery capable clients for both Java and PHP. These can be downloaded from the Amazon ElastiCache console. Our customers can create clients for any other language by building upon the popular Memcached clients available.

Q: How do I modify or write my own Memcached client to support auto-discovery?

You can take any Memcached Client Library and add support for Auto Discovery. If you would like to add or modify your own client to enable Auto Discovery, please refer to the Auto Discovery command set documentation.

Q: Are you planning to add support for more languages?

Yes, we are looking at Ruby next and may add more languages after that.

Q: Can I continue to work with my existing Memcached client if I don’t need Auto-discovery?

Yes, Amazon ElastiCache is still Memcached protocol compliant and does not require you to change your clients. However, for taking advantage of auto-discovery feature, we had to enhance the Memcached client capabilities. If you choose to not use the Amazon ElastiCache Cluster Client, you can continue to use your own clients or modify your own client library to understand the auto-discovery command set.

Q: Can I have heterogeneous clients when using Auto Discovery?

Yes, the same Amazon ElastiCache cluster can be connected through an Auto Discovery capable Client and the traditional Memcached client at the same time. Amazon ElastiCache remains 100% Memcached compliant.

Q: Can I stop using Auto Discovery?

Yes, you can stop using Auto Discovery anytime. You can disable Auto Discovery by specifying the mode of operation during the Amazon ElastiCache Cluster client initialization. Also, since Amazon ElastiCache continues to support Memcached 100% you may use any Memcached protocol-compliant client as before.

Engine Version Management

Q: Can I control if and when the engine version powering Amazon ElastiCache Cluster is upgraded to new supported versions?

Amazon ElastiCache allows you to control if and when the Memcached protocol-compliant software powering your cluster is upgraded to new versions supported by Amazon ElastiCache. This provides you with the flexibility to maintain compatibility with specific Memcached versions, test new versions with your application before deploying in production, and perform version upgrades on your own terms and timelines. Version upgrades involve some compatibility risk, thus they will not occur automatically and must be initiated by you. This approach to software patching puts you in the driver's seat of version upgrades, but still offloads the work of patch application to Amazon ElastiCache. You can learn more about version management by reading the FAQs that follow. Alternatively, you can refer to the Amazon ElastiCache User Guide. While Engine Version Management functionality is intended to give you as much control as possible over how patching occurs, we may patch your cluster on your behalf if we determine there is any security vulnerability in the system or cache software.

Q: How do I specify which supported Memcached Version my Cluster should run?

You can specify any currently supported version (minor and/or major) when creating a new cluster. If you wish to initiate an upgrade to a supported engine version release, you can do so using the "Modify" option for your cluster. Simply specify the version you wish to upgrade to via the "Cache Engine Version" field. The upgrade will then be applied on your behalf either immediately (if the "Applied Immediately" option is checked) or during the next scheduled maintenance window for your cluster.

Q: Can I test my cluster against a new version before upgrading?

Yes. You can do so by creating a new cluster with the new engine version. You can point your development/staging application to this cluster, test it and decide whether or not to upgrade your original cluster.

Q: Does Amazon ElastiCache provide guidelines for supporting new Memcached version releases and/or deprecating versions that are currently supported?

Over time, we plan to support additional Memcached versions for Amazon ElastiCache, both major and minor. The number of new version releases supported in a given year will vary based on the frequency and content of the Memcached version releases and the outcome of a thorough vetting of the release by our engineering team. However, as a general guidance, we aim to support new Memcached versions within 3-5 months of their General Availability release.

Q: Which version of the Memcached wire protocol does Amazon ElastiCache support?

Amazon ElastiCache supports the Memcached text and binary protocol of versions 1.4.5, 1.4.14 and 1.4.24 of Memcached.

Redis

Features

Q: What is Amazon ElastiCache for Redis?

Amazon ElastiCache for Redis is a web service that makes it easy to deploy and run Redis protocol-compliant server nodes in the cloud. The service enables the management, monitoring and operation of a Redis node; creation, deletion and modification of the node can be carried out through the ElastiCache console, the command line interface or the web service APIs. Amazon ElastiCache for Redis supports Redis Master / Slave replication.

Q: Is Amazon ElastiCache for Redis protocol-compliant with open source Redis?

Yes, Amazon ElastiCache for Redis is protocol-compliant with open source Redis. Code, applications, drivers and tools a customer uses today with their existing standalone Redis data store will continue to work with ElastiCache for Redis and no code changes will be required for existing Redis deployments migrating to ElastiCache for Redis unless noted. We currently support Redis 2.8.21, 2.8.22, 2.8.23, 2.8.24 and 3.2.4.

Q: What are Amazon ElastiCache for Redis nodes and shards?

An Amazon ElastiCache node is the smallest building block of an ElastiCache for Redis Cluster deployment. Each node supports the Redis protocol with Amazon's enhancements and has its own endpoint and port. Multiple types of nodes are supported, each with varying amount of CPU capability, and memory capacity.

A shard is a collection of one or more nodes that is responsible for a partition of the logical key space. Within a shard, a node may exist in isolation or in a primary/replica relationship with other nodes. If there are multiple nodes within a shard, one of the nodes will take on the read/write primary role and all other nodes will take on a read-only replica role.

Q: Does Amazon ElastiCache for Redis support Redis persistence?

Yes, you can achieve persistence by snapshotting your Redis data using the Backup and Restore feature. Please see here for details.

Q: How can I migrate from Amazon ElastiCache for Memcached to Amazon ElastiCache for Redis and vice versa?

We currently do not support automatically migrating from Memcached to Redis or vice versa. You may, however, use a Memcached client to read from a Memcached cluster and use a Redis client to write to a Redis cluster. Similarly, you may read from a Redis cluster using a Redis client and use a Memcached client to write to a Memcached cluster. Make sure to consider the differences in data format, and cluster configuration between the two engines.

Q: Does Amazon ElastiCache for Redis support Multi-AZ operation?

Yes, with Amazon ElastiCache for Redis you can create a read replica in another AWS

Availability Zone. Upon a failure of the primary node, we will provision a new primary node. In scenarios where the primary node cannot be provisioned, you can decide which read replica to promote to be the new primary. For more details on how to handle node failures see here.

Q: What options does Amazon ElastiCache for Redis provide for node failures?

Amazon ElastiCache for Redis will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. Thus, the DNS name for a Redis node remains constant, but the IP address of a Redis node can change over time. If you have a replication group with one or more read replicas and Multi-AZ is enabled, then in case of primary node failure ElastiCache will automatically detect the failure, select a replica and promote it to become the new primary. It will also propagate the DNS so that you can continue to use the primary endpoint and after the promotion it will point to the newly promoted primary. For more details see the Multi-AZ section of this FAQ. When Redis replication option is selected with Multi-AZ disabled, in case of primary node failure you will be given the option to initiate a failover to a read replica node. The failover target can be in the same zone or another zone. To failback to the original zone, promote the read replica in the original zone to be the primary. You may choose to architect your application to force the Redis client library to reconnect to the repaired Redis server node. This can help as some Redis libraries will stop using a server indefinitely when they encounter communication errors or timeouts.

Q: How does failover work?

For Multi-AZ enabled replication groups, the failover behavior is described at theMulti-AZ section of this FAQ.

If you choose not to enable Multi-AZ, then if Amazon ElastiCache monitors the primary node, and in case the node becomes unavailable or unresponsive, Amazon ElastiCache for Redis will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. Thus, the DNS name for a Redis node remains constant, but the IP address of a Redis node can change over time. However, if the primary node cannot be healed (and your Multi-AZ is disabled) you will have the choice to promote one of the read replicas to be the new primary. See here for how to select a new primary. The DNS record of the primary's endpoint will be updated to point to the promoted read replica node. A read replica node in the original primary's AZ will then be created to be a read replica in the shard and will follow the new primary.

Q: Are my read replicas available during a primary node failure?

Yes, during a primary node failure, the read replicas continue to service requests. After the primary node is restored, either as a healed node or as a promoted read replica, there is a brief period during which the read replicas will not serve any requests as they sync the cache information from the primary.

Q: How do I configure parameters of my Amazon ElastiCache for Redis nodes?

You can configure your Redis installation using a parameter group, which must be specified for a Redis cluster. All read replica clusters use the parameter group of their primary cluster. A Redis parameter group acts as a “container” for Redis configuration values that can be applied to one or more Redis primary clusters. If you create a Redis primary cluster without specifying a parameter group, a default parameter group is used. This default group contains defaults for the node type you plan to run. However, if you want your Redis primary cluster to run with specified configuration values, you can simply create a new cache parameter group, modify the desired parameters, and modify the primary Redis cluster to use the new parameter group.

Q: Can I access Redis through the Amazon ElastiCache console?

Yes, Redis appears as an Engine option in the ElastiCache console. You can create a new Redis cache cluster with the Launch Wizard by choosing the Redis engine. You can also modify or delete an existing Redis cluster using the ElastiCache console.

Q: Can Amazon ElastiCache for Redis clusters be created in an Amazon VPC?

Yes, just as you can create Memcached clusters within a VPC, you can create Redis clusters within a VPC as well. If your account is a VPC by default account, your Redis clusters will be created within the default VPC associated with your account. Using the ElastiCache console, you can specify a different VPC when you create your cluster.

Q: Is Redis password functionality supported in Amazon ElastiCache for Redis?

No, Amazon ElastiCache for Redis does not support Redis passwords. This is because of the inherent limitations of passwords stored in a configuration file. Instead of relying on Redis passwords, ElastiCache for Redis clusters are associated with an EC2 security group, and only clients within this security group have access to the Redis server.

Q. How do I upgrade to a newer engine version?

You can easily upgrade to a newer engine version by using the ModifyCacheCluster or ModifyReplicationGroup APIs and specifying your preferred engine version for the EngineVersion parameter. On the ElastiCache console, you can select a cluster and click “Modify”. In the “Modify” window select your preferred engine version from the available options. The engine upgrade process is designed to make a best effort to retain your existing data and requires Redis replication to succeed. For more details on that see here.

Q. Can I downgrade to an earlier engine version?

No. Downgrading to an earlier engine version is not supported.

Q. How do I scale up to a larger node type?

You can easily scale up to a larger node type by using the ModifyCacheCluster or ModifyReplicationGroup APIs and specifying your preferred node type for the CacheNodeType parameter. On the ElastiCache console, you can select a cache cluster or replication group and click “Modify”. In the “Modify” window select your preferred node type from the available options. The scale up process is designed to make a best effort to retain your existing data and requires Redis replication to succeed. For more details on that see here.

Q. Can I scale down to a smaller node type?

Moving to a smaller node type is currently not supported.

Read Replica

Q: What does it mean to run a Redis node as a Read Replica?

Read Replicas serve two purposes in Redis:

- Failure Handing
- Read Scaling

When you run a node with a Read Replica, the “primary” serves both writes and reads. The Read Replica acts as a “standby” which is “promoted” in failover scenarios. After failover, the standby becomes the primary and accepts your cache operations. Read Replicas also make it easy to elastically scale out beyond the capacity constraints of a single node for read-heavy cache workloads.

Q: When would I want to consider using a Redis read replica?

There are a variety of scenarios where deploying one or more read replicas for a given primary node may make sense. Common reasons for deploying a read replica include:

- Scaling beyond the compute or I/O capacity of a single primary node for read-heavy workloads. This excess read traffic can be directed to one or more read replicas.
- Serving read traffic while the primary is unavailable. If your primary node cannot take I/O requests (e.g. due to I/O suspension for backups or scheduled maintenance), you can direct read traffic to your read replicas. For this use case, keep in mind that the data on the read replica may be “stale” since the primary Instance is unavailable. The read replica can also be used to restart a failed primary warmed up.
- Data protection scenarios; in the unlikely event or primary node failure or that the Availability Zone in which your primary node resides becomes unavailable, you can promote a read replica in a different Availability Zone to become the new primary.

Q: How do I deploy a read replica node for a given primary node?

You can create a read replica in minutes using a CreateReplicationGroup API or a few clicks of the Amazon ElastiCache Management Console. When creating a cluster, you specify the

MasterCacheClusterIdentifier. The MasterCacheClusterIdentifier is the cache cluster Identifier of the “primary” node from which you wish to replicate. You then create the read replica cluster within the shard by calling the CreateCacheCluster API specifying the ReplicationGroupIdentifier and the CacheClusterIdentifier of the master node. As with a standard cluster, you can also specify the Availability Zone. When you initiate the creation of a read replica, Amazon ElastiCache takes a snapshot of your primary node in a shard and begins replication. As a result, you will experience a brief I/O suspension on your primary node as the snapshot occurs. The I/O suspension typically lasts on the order of one minute.

The read replicas are as easy to delete as they are to create; simply use the Amazon ElastiCache Management Console or call the DeleteCacheCluster API (specifying the CacheClusterIdentifier for the read replica you wish to delete).

Q: Can I create both a primary and read replicas at the same time?

Yes. You can create a new cache cluster along with read replicas in minutes using the CreateReplicationGroup API or using the “Create” wizard at the Amazon ElastiCache Management Console and selecting “Multi-AZ Replication”. When creating the cluster, specify an identifier, the total number of desired shard in a cluster a read replicas per shard, along with cahe creation parameters such as node type, engine version, etc. You can also specify the Availability Zone for each shard in the cluster.

Q: How do I connect to my read replica(s)?

You can connect to a read replica just as you would connect to a primary cache node, using the DescribeCacheClusters API or AWS Management Console to retrieve the endpoint(s) for you read replica(s). If you have multiple read replicas, it is up to your application to determine how read traffic will be distributed amongst them.

Q: How many read replicas can I create for a given primary node?

At this time, Amazon ElastiCache allows you to create up to five (5) read replicas for a given primary node.

Q: What happens to read replicas if failover occurs?

In the event of a failover, any associated and available read replicas should automatically resume replication once failover has completed (acquiring updates from the newly promoted read replica).

Q: Can I create a read replica of another read replica?

Creating a read replica of another read replica is not supported.

Q: Can I promote my read replica into a “standalone” primary node?

No, this is not supported. Instead, you may snapshot your ElastiCache for Redis node (you may select the primary or any of the read-replicas). You can then use the snapshot to seed a new ElastiCache for Redis primary.

Q: Will my read replica be kept up-to-date with its primary node?

Updates to a primary node will automatically be replicated to any associated read replicas. However, with Redis’s asynchronous replication technology, a read replica can fall behind its primary cache node for a variety of reasons. Typical reasons include:

- Write I/O volume to the primary cache node exceeds the rate at which changes can be applied to the read replica
- Network partitions or latency between the primary cache node and a read replica

Read replicas are subject to the strengths and weaknesses of Redis replication. If you are using read replicas, you should be aware of the potential for lag between a read replica and its primary cache node, or “inconsistency”. Click here for guidance on how to find out the “inconsistency” of your read replica.

Q: How do I gain visibility into active read replica(s)?

You can use the standard DescribeCacheClusters API to return a list of all the cache clusters you have deployed (including read replicas), or simply click on the "Redis" tab of the Amazon ElastiCache Management Console.

Amazon ElastiCache monitors the replication status of your read replicas and updates the Replication State field to Error if replication stops for any reason. You can review the details of the associated error thrown by the Redis engine by viewing the Replication Error field and take an appropriate action to recover from it. You can learn more about troubleshooting replication issues in the Troubleshooting a Read Replica problem section of the Amazon ElastiCache User Guide. If a replication error is fixed, the Replication State changes to Replicating.

Amazon ElastiCache allows you to gain visibility into how far a read replica has fallen behind its primary through the Amazon CloudWatch metric ("Replica Lag") available via the AWS Management Console or Amazon CloudWatch APIs.

Q: My read replica has fallen significantly behind its primary node. What should I do?

As discussed in the previous questions, “inconsistency” or lag between a read replica and its primary node is common with Redis asynchronous replication. If an existing read replica has fallen too far behind to meet your requirements, you can reboot it. Keep in mind that replica lag may naturally grow and shrink over time, depending on your primary node’s steady-state usage pattern.

Q: How do I delete a read replica? Will it be deleted automatically if its primary node is deleted?

You can easily delete a read replica with a few clicks of the AWS Management Console or by passing its cache cluster identifier to the DeleteCacheCluster API. If you want to delete the read replica in addition to the primary cache node, you must use the DeleteReplicationGroup API or AWS Management Console.

Q: How much do read replicas cost? When does billing begin and end?

A read replica is billed as a standard node and at the same rates. Just like a standard node, the rate per “Node hour” for a read replica is determined by the node class of the read replica – please see Amazon ElastiCache detail page for up-to-date pricing. You are not charged for the data transfer incurred in replicating data between your primary cache node and read replica. Billing for a read replica begins as soon as the read replica has been successfully created (i.e. when status is listed as “active”). The read replica will continue being billed at standard Amazon ElastiCache cache node hour rates until you issue a command to delete it.

Q: What happens during failover and how long does it take?

Initiated failover is supported by Amazon ElastiCache so that you can resume operations as quickly as possible. When failing over, Amazon ElastiCache simply flips the DNS record for your node to point at the read replica, which is in turn promoted to become the new primary. We encourage you to follow best practices and implement cache node connection retry at the application layer. Start-to-finish, failover typically completes within three to six minutes.

Q: Can I create a read replica in another region as my primary?

No. Your read replica may only be provisioned in the same or different Availability Zone of the same Region as your cache node primary.

Q: Can I see which Availability Zone my primary is currently located in?

Yes, you can gain visibility into the location of the current primary by using the AWS Management Console or DescribeCacheClusters API.

After failover, my primary is now located in a different Availability Zone than my other AWS resources (e.g. EC2 instances).

Q: Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same Region. In addition, you may want to consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone failure.

Multi-AZ

Q: What is Multi-AZ for ElastiCache for Redis?

An ElastiCache for Redis shard consists of a primary and up to five read replicas. Redis asynchronously replicates the data from the primary to the read replicas. During certain types of planned maintenance, or in the unlikely event of ElastiCache node failure or Availability Zone failure, Amazon ElastiCache will automatically detect the failure of a primary, select a read replica, and promote it to become the new primary. ElastiCache also propagates the DNS changes of the promoted read replica, so if your application is writing to the primary node endpoint, no endpoint change will be needed.

Q: What are the benefits of using Multi-AZ?

The main benefits of running your ElastiCache for Redis in Multi-AZ mode are enhanced availability and smaller need for administration. If an ElastiCache for Redis primary node failure occurs, the impact on your ability to read/write to the primary is limited to the time it takes for automatic failover to complete. When Multi-AZ is enabled, ElastiCache node failover is automatic and requires no administration. You no longer need to monitor your Redis nodes and manually initiate a recovery in the event of a primary node disruption.

Q: How does Multi-AZ work?

You can use Multi-AZ if you are using ElastiCache for Redis and have a shard consisting of a primary node and one or more read replicas. If the primary node fails, ElastiCache will automatically detect the failure, select one from the available read replicas, and promote it to become the new primary. When cluster_mode parameter is disabled, ElastiCache will propagate the DNS changes of the promoted replica so that your application can keep writing to the primary endpoint. For cluster_mode enabled, ElastiCache will update the node map of the cluster. ElastiCache will also spin up a new node to replace the promoted read replica in the same Availability Zone of the failed primary. In case the primary failed due to temporary Availability Zone disruption, the new replica will be launched once that Availability Zone has recovered.

Q: Can I have replicas in the same Availability Zone as the primary?

Yes. Note that placing both the primary and the replica(s) in the same Availability Zone will not make your ElastiCache for Redis replication group resilient to an Availability Zone disruption.

Q: What events would cause Amazon ElastiCache to fail over to a read replica?

Amazon ElastiCache will failover to a read replica in the event of any of the following:

- Loss of availability in primary’s Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary

Q: When should I use Multi-AZ?

Using Redis replication in conjunction with Multi-AZ provides increased availability and fault tolerance. Such deployments are a natural fit for use in production environments. When running ElastiCache for Redis Cluster with cluster mode enabled, if your shards have one or more read replicas, Multi-AZ will automatically be enabled.

Q: How do I create an ElastiCache for Redis replication group with Multi-AZ enabled?

You can create an ElastiCache for Redis primary and read replicas by clicking "Create" on the ElastiCache Management Console. You can also do so by calling the CreateReplicationGroup API. For existing clusters (Redis 2.8.6, 2.8.19, 2.8.21, 2.8.22, 2.8.23, 2.8.24 and 3.2.4 with cluster_mode=disabled), you can enable Multi-AZ by choosing a cluster and clicking Modify on the ElastiCache Management Console or by using the ModifyReplicationGroup API. Switching a replication group to Multi-AZ is not disruptive to your Redis data and does not interfere your nodes' ability to serve requests.

Q: Which read replica will be promoted in case of primary node failure?

If there are more than one read replicas, the read replica with the smallest asynchronous replication lag to the primary will be promoted.

Q: How much does it cost to use Multi-AZ?

Multi-AZ is free of charge. You only pay for the ElastiCache nodes that you use.

Q: What are the performance implications of Multi-AZ?

ElastiCache currently uses the Redis engine’s native, asynchronous replication and is subject to its strengths and limitations. In particular, when a read replica connects to a primary for the first time, or if the primary changes, the read replica does a full synchronization of the data from the primary, imposing load on itself and the primary. For additional details regarding Redis replication please see here.

Q: What node types support Multi-AZ?

All available node types in ElastiCache support Multi-AZ with one exception. When using Redis 3.x with cluster_mode=disabled, T2 family doesn’t support Multi-AZ.

Q: Will I be alerted when automatic failover occurs?

Yes, Amazon ElastiCache will create an event to inform you that automatic failover occurred. You can use the DescribeEvents API to return information about events related to your ElastiCache node, or click the Events section of the ElastiCache Management Console.

Q: After failover, my primary is now located in a different Availability Zone than my other AWS resources (for example, EC2 instances). Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same region. You may consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone disruption.

Q: Where can I get more information about Multi-AZ?

For more information about Multi-AZ, see ElastiCachedocumentation.

Backup and Restore

Q: What is Backup and Restore?

Backup and Restore is a feature that allows customers to create snapshots of their ElastiCache for Redis clusters. ElastiCache stores the snapshots, allowing users to subsequently use them to restore Redis clusters.

Q: What is a snapshot?

A snapshot is a copy of your entire Redis cluster at a specific moment.

Q: Why would I need snapshots?

Creating snapshots can be useful in case of data loss caused by node failure, as well as the unlikely event of a hardware failure. Another common reason to use backups is for archiving purposes. Snapshots are stored in Amazon S3, which is a durable storage, meaning that even a power failure won’t erase your data.

Q: What can I do with a snapshot?

You can use snapshots to warm start an ElastiCache for Redis cluster with preloaded data.

Q: How does Backup and Restore work?

When a backup is initiated, ElastiCache will take a snapshot of a specified Redis cluster that can later be used for recovery or archiving. You can initiate a backup anytime you choose or set a recurring daily backup with retention period of up to 35 days.

When you choose a snapshot to restore, a new ElastiCache for Redis cluster will be created and populated with the snapshot’s data. This way you can create multiple ElastiCache for Redis clusters from a specified snapshot.

Currently, ElastiCache uses Redis’ native mechanism to create and store an RDB file as the snapshot.

Q: Where are my snapshots stored?

The snapshots are stored in S3.

Q: How can I get started using Backup and Restore?

You can select to use the Backup and Restore feature through theAWS Management Console, through the ElastiCache APIs (CreateCacheCluster, ModifyCacheCluster and ModifyReplicationGroup API’s) and CLI. You can deactivate and reactivate the feature anytime you choose.

Q: How do I specify which Redis cluster and node to backup?

Backup and Restore creates snapshots on a cluster basis. Users can specify which ElastiCache for Redis cluster to backup through the AWS Management Console, CLI or through the CreateSnapshot API. In a Replication Group, you can choose to backup the primary or any of the read-replica clusters. We recommend users enable backup on one of the read-replicas, mitigating any latency effect on the Redis primary.

Q: Does ElastiCache for Memcached support Backup and Restore?

No, snapshots are available only for ElastiCache for Redis.

Q: How can I specify when a backup will take place?

Through the AWS Management Console, CLI or APIs you can specify when to start a single backup or a recurring backup. Users are able to:

- Take a snapshot right now (through “Backup” console button in the "Redis" tab, or CreateSnapshot API)
- Set up an automatic daily backup. The backup will take place during your preferred backup window. You can set that up through Creating/Modifying cluster via console or the CreateCacheCluster, ModifyCacheCluster or ModifyReplicationGroup API’s.

Q: What is a backup window and why do I need it?

The preferred backup window is the user-defined period of time during which your ElastiCache for Redis cluster backup will start. This is helpful if you want to backup at a certain time of day or to refrain from backups during a particularly high-utilization period.

Q: What is the performance impact of taking a snapshot?

While taking a snapshot, you may encounter increased latencies for a brief period at the node. Snapshots use Redis’s built-in BGSAVE and are subject to its strengths and limitations. In particular, the Redis process forks and the parent continues to serve requests while the child saves the data on disk and then exits. The forking increases the memory usage for the duration of the snapshot generation. When this memory usage exceeds that of the available memory of the node, swapping can get triggered, further slowing down the node. For this reason, we recommend generating snapshots on one of the read replicas (instead of the primary). Also, we suggest setting the reserved-memory parameter to minimize swap usage. See here for more details.

Q: Can I create a snapshot from an ElastiCache for Redis read replica?

Yes. Creating a snapshot from a read replica is the best way to backup your data while minimizing performance impact.

Q: In what regions is the Backup and Restore feature available?

Backup and Restore feature is available in all regions where ElastiCache service is available.

Q: Can I export ElastiCache for Redis snapshots to an S3 bucket owned by me?

Yes. You can export your ElastiCache for Redis snapshots to an authorized S3 bucket in the same region as your cluster. For more details on exporting snapshots and setting the required permissions, please refer to this.

Q: Can I copy snapshots from one region to another?

Yes. You must first copy your snapshot into an authorized S3 bucket of your choice in the same region and then use the S3 PUT object- Copy API to copy it to a bucket in another region. For more details on copying S3 objects, please see this.

Q: I have multiple AWS accounts using ElastiCache for Redis. Can I use ElastiCache snapshots from one account to warm start an ElastiCache for Redis cluster in a different one?

Yes. You must first copy your snapshot into an authorized S3 bucket of your choice in the same region and then grant cross-account bucket permissions to the other account. For more details on S3 cross-account permissions, please see this. Finally, specify the S3 location of your RDB file during cluster creation through the Launch Cache Cluster Wizard in the console or through the CreateCacheCluster API.

Q: How much does it cost to use Backup and Restore?

Amazon ElastiCache provides storage space for one snapshot free of charge for each active ElastiCache for Redis cluster. Additional storage will be charged based on the space used by the snapshots with $0.085/GB every month (same price in all regions). Data transfer for using the snapshots is free of charge.

Q: What is the retention period?

Retention period is the time span during which the automatic snapshots are retained. For example, if a retention period is set for 5, a snapshot that was taken today will be retained for 5 days before being deleted. You can choose to copy one or more automatic snapshots to store them as manual so that they won't be deleted after the retention period is over.

Q: How do I manage the retention of my automated snapshots?

You can use the AWS Management Console or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended).

Q: What happens to my snapshots if I delete my ElastiCache for Redis cluster?

When you delete an ElastiCache for Redis cluster, your manual snapshots are retained. You will also have an option to create a final snapshot before the cluster is deleted. Automatic snapshots are not retained.

Q: What nodes types support backup and restore capability?

All ElastiCache for Redis instance node types besides t1.micro and t2 family support backup and restore:

Current Generation Nodes:

- cache.m3.medium
- cache.m3.large
- cache.m3.xlarge
- cache.m3.2xlarge
- cache.m4.large
- cache.m4.xlarge
- cache.m4.2xlarge
- cache.m4.4xlarge
- cache.m4.10xlarge
- cache.r3.large
- cache.r3.xlarge
- cache.r3.2xlarge
- cache.r3.4xlarge
- cache.r3.8xlarge

Previous Generation Nodes:

- cache.m1.small
- cache.m1.medium
- cache.m1.large
- cache.m1.xlarge
- cache.m2.xlarge
- cache.m2.2xlarge
- cache.m2.4xlarge
- cache.c1.xlarge

Q: Can I use my own RDB snapshots stored in S3 to warm start an ElastiCache for Redis cluster?

Yes. You can specify the S3 location of your RDB file during cluster creation through the "Create Cluster" Wizard in the console or through the CreateCacheCluster API.

Q: Can I use the Backup and Restore feature if I am running ElastiCache in a VPC?

Yes.

Redis Cluster

Q: What is ElastiCache for Redis Cluster?

ElastiCache for Redis Cluster allows customers to create and run managed Redis Clusters with multiple shards. It is compatible with open source Redis 3.2 and comes with a number of enhancements for a more stable and robust experience (see the “enhanced engine” section below for additional details on these enhancements).

Q: Why would I need a scale out Redis environment?

There are three main scenarios for running a scale out Redis environment. First, if the total memory size of your Redis data exceeds or is projected to exceed the memory capacity of a single VM. Second, if the write throughput of your application to Redis exceeds the capacity of a single VM. Third, if you would like to spread the data across multiple shards so that any potential issue that comes up with a single node will have a smaller impact on the overall Redis environment.

Q: Why would I run my Redis Cluster workload on Amazon ElastiCache?

Amazon ElastiCache provides a fully managed distributed in-memory Redis environment, from provisioning server resources to installing the engine software and applying any configuration parameters you choose. It uses enhancements to the Redis engine developed by Amazon, which results in a more robust and stable experience (see “enhanced engine” section for more details). Once your Redis environment is up and running, the service automates common administrative tasks such as failure detection and recovery, backups and software patching. It also provides a robust Multi-AZ solution with automatic failover. In case of a failure of one or more primary nodes in your cluster, Amazon ElastiCache will automatically detect the failure and respond by promoting the most up to date replica to primary. This process is automated and does not mandate any manual work on your behalf. Amazon ElastiCache also provides detailed monitoring metrics associated with your ElastiCache nodes, enabling you to diagnose and respond to issues very quickly.

Q: Is ElastiCache for Redis Cluster compatible with open source Redis?

Yes, Amazon ElastiCache for Redis Cluster is compatible with open source Redis 3.2. You can use the open source Redis Cluster clients to access scale-out clusters on ElastiCache for Redis.

Q: Can I modify the number of shards once the cluster is created?

Currently you cannot modify the number of shards in a cluster once it’s created.

Q: What is the upgrade path from current ElastiCache for Redis 2.8.x to ElastiCache for Redis Cluster (version 3.2.4)?

If you are using Redis 3.2 with cluster_mode parameter disabled, you can simply choose the node or cluster you wish to upgrade and modify the engine version. ElastiCache will provision a Redis 3.2.4 cluster and migrate your data to it, while maintaining the endpoint.

If you are using Redis 3.2 with cluster_mode enabled, you can migrate to Redis Cluster by first creating a snapshot of your data using the backup and restore feature. Then, select the created snapshot and click on “Restore Snapshot” to create a Redis 3.2 cluster using the snapshotted data. Finally, update the new endpoint in your client. Note that to use Redis 3.2 in cluster mode you would need to switch to a Redis Cluster client.

Q: Is the pricing for clustered configuration different from non-clustered configuration?

No. Amazon ElastiCache for Redis provides the flexibility of clustered and non-clustered configuration at the same price. Customers can now enjoy enhanced engine functionality within Amazon ElastiCache for Redis and use full feature support for clustered configuration and scalability at the same price.

Q: What is Multi-AZ for ElastiCache for Redis Cluster?

Each shard of an ElastiCache for Redis cluster consists of a primary and up to five read replicas. Redis asynchronously replicates the data from the primary to the read replicas. During certain types of planned maintenance, or in the unlikely event of ElastiCache node failure or Availability Zone failure, Amazon ElastiCache will automatically detect the failure of a primary, select a read-replica, and promote it to become the new primary.

ElastiCache for Redis Cluster provides enhancements and management for Redis 3.x environments. When running an unmanaged Redis environment, in a case of primary node failure, the cluster relies on a majority of masters to determine and execute a failover. If such majority doesn't exist, the cluster will go into failed state, rejecting any further reads and writes. This could lead to major availability impact on the application, as well as requiring human intervention to manually salvage the cluster. ElastiCache for Redis Multi-AZ capability is built to handle any failover case for Redis Cluster with robustness and efficiency.

Q: How is Multi-AZ in ElastiCache for Redis Cluster different than in ElastiCache for Redis versions 2.8.x?

Redis 3.x works with intelligent clients that store a node map with all the cluster nodes' endpoints. During a failover, the client updates the node map with the IP endpoint for the new primary. This provides up to 4x faster failover time than with ElastiCache for Redis 2.8.x.

Q: How does Multi-AZ work for Redis Cluster?

You can use Multi-AZ if you are using an ElastiCache for Redis Cluster with each shard having 1 or more read-replicas. If a primary node of a shard fails, ElastiCache will automatically detect the failure, select one of the available read-replicas, and promote it to become the new primary. The Redis 3.x client will update the promoted replica as primary, no application change is required. ElastiCache will also spin up a new node to replace the promoted read-replica in the same Availability Zone of the failed primary. In case the primary failed due to a temporary Availability Zone failure, the new replica will be launched once that Availability Zone has recovered.

Q: What is a backup in ElastiCache for Redis Cluster?

An ElastiCache for Redis Cluster backup is a series of snapshots of the cluster's shards, stored together to keep a copy of your entire Redis data around a certain time frame.

Q: How is a backup in ElastiCache for Redis Cluster different from a snapshot in ElastiCache for Redis?

Since a non-clustered ElastiCache for Redis environment has a single primary node, a backup is a single file which contains a copy of the Redis data. ElastiCache for Redis Cluster can have one or more shards, thus a backup might contain multiple files.

Q: How do I specify which ElastiCache for Redis nodes to backup in each shard?

You cannot manually specify a node to backup within each shard. When initiating a backup, ElastiCache will automatically select the most up-to-date read replica in each shard and take a snapshot of its data.

Q: How does ElastiCache for Redis Cluster Backup and Restore work?

When a backup is initiated, ElastiCache will take a backup of a specified cluster; that backup can later be used for recovery or archiving. The backup will include a copy of each of the cluster's shards, thus a full backup contains a series of files. You can initiate a backup anytime you choose or set a recurring daily backup with retention period of up to 35 days.

When you choose a backup to restore, a new ElastiCache for Redis cluster will be created and populated with the backup's data.

Currently, ElastiCache uses Redis' native mechanism to create and store an RDB file for each shard as the backup.

Q: Is the backup in ElastiCache for Redis Cluster a point-in-time snapshot?

When you initiate a backup, ElastiCache will trigger backups of all of the shards of your cluster at the same time. In rare cases there might be a need to retake a snapshot of one or more nodes that did not complete successfully the first time. ElastiCache does that automatically and no user intervention is required. But in such a case, while each individual snapshot is a point-in-time representation of the node it was taken from, not all the cluster's snapshots would be taken at the same time.

Q: How can I specify when a backup will take place?

Through the AWS Management Console, CLI or APIs you can specify when to start a single backup or a recurring backup. Users are able to:

- Take a backup right now (through "Create Snapshot" console button or CreateSnapshot API)
- Set up an automatic daily backup. The backup will take place during your preferred backup window. You can set that up through Creating/Modifying cluster via console or the CreateReplicationGroup and ModifyReplicationGroup API's.

Q: Can I use my own RDB snapshots stored in S3 to pre-seed a scale out ElastiCache for Redis Cluster environment?

Yes. You can specify the S3 location of your RDB files during cluster creation through the Create Cluster Wizard in the console or through the CreateReplicationGroup API. ElastiCache will automatically parse the Redis key-space of the RDB snapshot and redistribute it among the shards of the new cluster.

Enhanced Engine

Q: How is the engine within ElastiCache for Redis different from open-source Redis?

The engine within ElastiCache for Redis is fully compatible with open source Redis but also comes with enhancements that improve robustness and stability. Some of the enhancements are:

- More usable memory: You can now safely allocate more memory for your application without risking increased swap usage during syncs and snapshots.
- Improved synchronization: More robust synchronization under heavy load and when recovering from network disconnections. Additionally, syncs are faster as both the primary and replicas no longer use the disk for this operation.
- Smoother failovers: In the event of a failover, your shard now recovers faster as replicas no longer flush their data to do a full re-sync with the primary.

Q: How do I use the enhanced engine?

To use the enhanced engine from the Amazon ElastiCache management console, just select an engine compatible with Redis engine version 2.8.22 or higher when creating a cluster. From that point on you will be using the enhanced engine. You can also use the enhanced engine through the ElastiCache API or AWS CLI by specifying the engine version when running the CreateCacheCluster API.

Q: Do I need to change my application code to use the enhanced engine on ElastiCache?

No. The enhanced engine is fully compatible with open-source Redis, thus you can enjoy its improved robustness and stability without the need to make any changes to your application code.

Q: How much does it cost to use the enhanced engine?

There is no additional charge for using the enhanced engine. As always, you will only be charged for the nodes you use.

Amazon Redshift FAQ

General

Q: What is Amazon Redshift?

Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Customers can start small for just $0.25 per hour with no commitments or upfront costs and scale to a petabyte or more for $1,000 per terabyte per year, less than a tenth of most other data warehousing solutions.

Traditional data warehouses require significant time and resource to administer, especially for large datasets. In addition, the financial cost associated with building, maintaining, and growing self-managed, on-premise data warehouses is very high. Amazon Redshift not only significantly lowers the cost of a data warehouse, but also makes it easy to analyze large amounts of data very quickly.

Amazon Redshift gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections. Queries are distributed and parallelized across multiple physical resources. You can easily scale an Amazon Redshift data warehouse up or down with a few clicks in the AWS Management Console or with a single API call. Amazon Redshift automatically patches and backs up your data warehouse, storing the backups for a user-defined retention period. Amazon Redshift uses replication and continuous backups to enhance availability and improve data durability and can automatically recover from component and node failures. In addition, Amazon Redshift supports Amazon Virtual Private Cloud (Amazon VPC), SSL, AES-256 encryption and Hardware Security Modules (HSMs) to protect your data in transit and at rest.

As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use. Amazon Redshift lets you pay as you go. You can even try Amazon Redshift for free.

Q: What does Amazon Redshift manage on my behalf?

Amazon Redshift manages the work needed to set up, operate, and scale a data warehouse, from provisioning the infrastructure capacity to automating ongoing administrative tasks such as backups, and patching. Amazon Redshift automatically monitors your nodes and drives to help you recover from failures.

Q: How does the performance of Amazon Redshift compare to most traditional databases for data warehousing and analytics?

Amazon Redshift uses a variety of innovations to achieve up to ten times higher performance than traditional databases for data warehousing and analytics workloads:

- *Columnar Data Storage*: Instead of storing data as a series of rows, Amazon Redshift organizes the data by column. Unlike row-based systems, which are ideal for transaction processing, column-based systems are ideal for data warehousing and analytics, where queries often involve aggregates performed over large data sets. Since only the columns involved in the queries are processed and columnar data is stored sequentially on the storage media, column-based systems require far fewer I/Os, greatly improving query performance.

- *Advanced Compression:* Columnar data stores can be compressed much more than row-based data stores because similar data is stored sequentially on disk. Amazon Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores. In addition, Amazon Redshift doesn't require indexes or materialized views and so uses less space than traditional relational database systems. When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.

- *Massively Parallel Processing (MPP):* Amazon Redshift automatically distributes data and query load across all nodes. Amazon Redshift makes it easy to add nodes to your data warehouse and enables you to maintain fast query performance as your data warehouse grows.

Q: How do I get started with Amazon Redshift?

You can sign up and get started within minutes from the Amazon Redshift detail page or via the AWS Management Console. If you don't already have an AWS account, you'll be prompted to create one. Visit our Getting Started Page to see how to try Amazon Redshift for free.

Q: How do I create an Amazon Redshift data warehouse cluster?

You can easily create an Amazon Redshift data warehouse cluster by using the AWS Management Console or the Amazon Redshift APIs. You can start with a single node, 160GB data warehouse and scale all the way to a petabyte or more with a few clicks in the AWS Console or a single API call.

The single node configuration enables you to get started with Amazon Redshift quickly and cost-effectively and scale up to a multi-node configuration as your needs grow. The multi-node configuration requires a leader node that manages client connections and receives queries, and two compute nodes that store data and perform queries and computations. The leader node is provisioned for you automatically and you are not charged for it.

Simply specify your preferred Availability Zone (optional), the number of nodes, node types, a master name and password, security groups, your preferences for backup retention, and other system settings. Once you've chosen your desired configuration, Amazon Redshift will provision the required resources and set up your data warehouse cluster.

Q: What does a leader node do? What does a compute node do?

A leader node receives queries from client applications, parses the queries and develops execution plans, which are an ordered set of steps to process these queries. The leader node then coordinates the parallel execution of these plans with the compute nodes, aggregates the intermediate results from these nodes and finally returns the results back to the client applications.

Compute nodes execute the steps specified in the execution plans and transmit data among themselves to serve these queries. The intermediate results are sent back to the leader node for aggregation before being sent back to the client applications.

Q: What is the maximum storage capacity per compute node? What is the recommended amount of data per compute node for optimal performance?

You can create a cluster using either Dense Storage (DS) nodes or Dense Compute nodes (DC). Dense Storage nodes allow you to create very large data warehouses using hard disk drives (HDDs) for a very low price point. Dense Compute nodes allow you to create very high performance data warehouses using fast CPUs, large amounts of RAM and solid-state disks (SSDs).

Dense Storage (DS) nodes are available in two sizes, Extra Large and Eight Extra Large. The Extra Large (XL) has 3 HDDs with a total of 2TB of magnetic storage, whereas Eight Extra Large (8XL) has 24 HDDs with a total of 16TB of magnetic storage. DS2.8XL has 36 Intel Xeon E5-2676 v3 (Haswell) virtual cores and 244GiB of RAM, and DS2.XL has 4 Intel Xeon E5-2676 v3 (Haswell) virtual cores and 31GiB of RAM. Please see our pricing page for more detail. You can get started with a single Extra Large node, 2TB data warehouse for $0.85 per hour and scale up to a petabyte or more. You can pay by the hour or use reserved instance pricing to lower your price to under $1,000 per TB per year.

Dense Compute (DC) nodes are also available in two sizes. The Large has 160GB of SSD storage, 2 Intel Xeon E5-2670v2 (Ivy Bridge) virtual cores and 15GiB of RAM. The Eight Extra Large is sixteen times bigger with 2.56TB of SSD storage, 32 Intel Xeon E5-2670v2 virtual cores and 244GiB of RAM. You can get started with a single Large node for $0.25 per hour and and scale all the way up to 128 8XL nodes with 326TB of SSD storage, 3,200 virtual cores and 24TiB of RAM.

Amazon Redshift's MPP architecture means you can increase your performance by increasing the number of nodes in your data warehouse cluster. The optimal amount of data per compute node depends on your application characteristics and your query performance needs.

Q: How many nodes can I specify per Amazon Redshift data warehouse cluster?

An Amazon Redshift data warehouse cluster can contain from 1-128 compute nodes, depending on the node type. For details please see our documentation.

Q: How do I access my running data warehouse cluster?

Once your data warehouse cluster is available, you can retrieve its endpoint and JDBC and ODBC connection string from the AWS Management Console or by using the Redshift APIs. You can then use this connection string with your favorite database tool, programming language, or Business Intelligence (BI) tool. You will need to authorize network requests to your running data warehouse cluster. For a detailed explanation please refer to our Getting Started Guide.

Q: When would I use Amazon Redshift vs. Amazon RDS?

Both Amazon Redshift and Amazon RDS enable you to run traditional relational databases in the cloud while offloading database administration. Customers use Amazon RDS databases both for online-transaction processing (OLTP) and for reporting and analysis. Amazon Redshift harnesses the scale and resources of multiple nodes and uses a variety of optimizations to provide order of magnitude improvements over traditional databases for analytic and reporting workloads against very large data sets. Amazon Redshift provides an excellent scale-out option as your data and query complexity grows or if you want to prevent your reporting and analytic processing from interfering with the performance of your OLTP workload.

Q: When would I use Amazon Redshift vs. Amazon Elastic MapReduce (Amazon EMR)?

Amazon Redshift is ideal for large volumes of structured data that you want to persist and query using standard SQL and your existing BI tools. Amazon EMR is ideal for processing and transforming unstructured or semi-structured data to bring in to Amazon Redshift and is also a much better option for data sets that are relatively transitory, not stored for long-term use.

Q: Why should I use Amazon Redshift instead of running my own MPP data warehouse cluster on Amazon EC2?

Amazon Redshift automatically handles many of the time-consuming tasks associated with managing your own data warehouse including:

- *Setup:* With Amazon Redshift, you simply create a data warehouse cluster, define your schema, and begin loading and querying your data. Provisioning, configuration and patching are all managed for you.

- *Data Durability:* Amazon Redshift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3, which is designed for eleven nines of durability. Amazon Redshift mirrors each drive's data to other nodes within your cluster. If a drive fails, your queries will continue with a slight latency increase while Redshift rebuilds your drive from replicas. In case of node failure(s), Amazon Redshift automatically provisions new node(s) and begins restoring data from other drives within the cluster or from Amazon S3. It prioritizes restoring your most frequently queried data so your most frequently executed queries will become performant quickly.

- *Scaling:* You can add or remove nodes from your Amazon Redshift data warehouse cluster with a single API call or via a few clicks in the AWS Management Console as your capacity and performance needs change.

- *Automatic Updates and Patching:* Amazon Redshift automatically applies upgrades and patches your data warehouse so you can focus on your application and not on its administration.

Back to top »

Billing

Q: How will I be charged and billed for my use of Amazon Redshift?

You pay only for what you use, and there are no minimum or setup fees. You are billed based on:

- *Compute node hours* – Compute node hours are the total number of hours you run across all your compute nodes for the billing period. You are billed for 1 unit per node per hour, so a 3-node data warehouse cluster running persistently for an entire month would incur 2,160 instance hours. You will not be charged for leader node hours; only compute nodes will incur charges.
- *Backup Storage* – Backup storage is the storage associated with your automated and manual snapshots for your data warehouse. Increasing your backup retention period or taking additional snapshots increases the backup storage consumed by your data warehouse. There is no additional charge for backup storage up to 100% of your provisioned storage for an active data warehouse cluster. For example, if you have an active Single Node XL data warehouse cluster with 2TB of local instance storage, we will provide up to 2TB-Month of backup storage at no additional charge. Backup storage beyond the provisioned storage size and backups stored after your cluster is terminated are billed at standard Amazon S3 rates.
- *Data transfer* – There is no Data Transfer charge for data transferred to or from Amazon Redshift outside of Amazon VPC. Data Transfer to or from Redshift in Amazon VPC accrues standard AWS data transfer charges.

For Amazon Redshift pricing information, please visit theAmazon Redshift pricing page.

Q: When does billing of my Amazon Redshift data warehouse clusters begin and end?

Billing commences for a data warehouse cluster as soon as the data warehouse cluster is available. Billing continues until the data warehouse cluster terminates, which would occur upon deletion or in the event of instance failure.

Q: What defines billable Amazon Redshift instance hours?

Node usage hours are billed for each hour your data warehouse cluster is running in an available state. If you no longer wish to be charged for your data warehouse cluster, you must terminate it to avoid being billed for additional node hours. Partial node hours consumed are billed as full hours.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with billing address in Japan, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Back to top »

Data Integration and Loading

Q: How do I load data into my Amazon Redshift data warehouse?

You can load data into Amazon Redshift from a range of data sources including Amazon S3, Amazon DynamoDB, Amazon EMR, AWS Data Pipeline and or any SSH-enabled host on Amazon EC2 or on-premises. Amazon Redshift attempts to load your data in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster. For more details on loading data into Amazon Redshift please view our Getting Started Guide.

Q: Can I load data using SQL ‘INSERT' statements?

Yes, clients can connect to Amazon Redshift using ODBC or JDBC and issue 'insert' SQL commands to insert the data. Please note this is slower than using S3 or DynamoDB since those methods load data in parallel to each compute node while SQL insert statements load via the single leader node.

Q: How do I load data from my existing Amazon RDS, Amazon EMR, Amazon DynamoDB, and Amazon EC2 data sources to Amazon Redshift?

You can use our COPY command to load data in parallel directly to Amazon Redshift from Amazon EMR, Amazon DynamoDB, or any SSH-enabled host. Moreover, many ETL companies have certified Amazon Redshift for use with their tools, and a number are offering free trials to help you get started loading your data. Finally, AWS Data Pipeline provides a high performance, reliable, fault tolerant solution to load data from a variety of AWS data sources. You can use AWS Data Pipeline to specify the data source, desired data transformations, and then execute a pre-written import script to load your data into Amazon Redshift.

Q: I have a lot of data for initial loading into Amazon Redshift. Transferring via the Internet would take a long time. How do I load this data?

You can use AWS Import/Export to transfer the data to Amazon S3 using portable storage devices. In addition, you can use AWS Direct Connect to establish a private network connection between your network or datacenter and AWS. You can choose 1Gbit/sec or 10Gbit/sec connection ports to transfer your data.

Back to top »

Security

Q: How does Amazon Redshift keep my data secure?

Amazon Redshift encrypts and keeps your data secure in transit and at rest using industry-standard encryption techniques. To keep data secure in transit, Amazon Redshift supports SSL-enabled connections between your client application and your Redshift data warehouse cluster. To keep your data secure at rest, Amazon Redshift encrypts each block using hardware-accelerated AES-256 as it is written to disk. This takes place at a low level in the I/O subsystem, which encrypts everything written to disk, including intermediate query results. The blocks are backed up as is, which means that backups are encrypted as well. By default, Amazon Redshift takes care of key management but you can choose to manage your keys using your own hardware security modules (HSMs) or manage your keys through AWS Key Management Service.

Q: Can I use Amazon Redshift in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, you can use Amazon Redshift as part of your VPC configuration. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon Redshift data warehouse cluster.

Q: Can I access my Amazon Redshift compute nodes directly?

No. Your Amazon Redshift compute nodes are in a private network space and can only be accessed from your data warehouse cluster's leader node. This provides an additional layer of security for your data.

Back to top »

Availability and Durability

Q: What happens to my data warehouse cluster availability and data durability if a drive

on one of my nodes fails?

Your Amazon Redshift data warehouse cluster will remain available in the event of a drive failure however you may see a slight decline in performance for certain queries. In the event of a drive failure, Amazon Redshift will transparently use a replica of the data on that drive which is stored on other drives within that node. In addition, Amazon Redshift will attempt to move your data to a healthy drive or will replace your node if it is unable to do so. Single node clusters do not support data replication. In the event of a drive failure you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability in the event of individual node failure?

Amazon Redshift will automatically detect and replace a failed node in your data warehouse cluster. The data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB. Amazon Redshift makes your replacement node available immediately and loads your most frequently accessed data from S3 first to allow you to resume querying your data as quickly as possible. Single node clusters do not support data replication. In the event of a drive failure you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability in the event if my data warehouse cluster's Availability Zone (AZ) has an outage?

If your Amazon Redshift data warehouse cluster's Availability Zone becomes unavailable, you will not be able to use your cluster until power and network access to the AZ are restored. Your data warehouse cluster's data is preserved so you can start using your Amazon Redshift data warehouse as soon as the AZ becomes available again. In addition, you can also choose to restore any existing snapshots to a new AZ in the same Region. Amazon Redshift will restore your most frequently accessed data first so you can resume queries as quickly as possible.

Q: Does Amazon Redshift support Multi-AZ Deployments?

Currently, Amazon Redshift only supports Single-AZ deployments. You can run data warehouse clusters in multiple AZ's by loading data into two Amazon Redshift data warehouse clusters in separate AZs from the same set of Amazon S3 input files. In addition, you can also restore a data warehouse cluster to a different AZ from your data warehouse cluster snapshots.

Back to top »

Backup and Restore

Q: How does Amazon Redshift back up my data?

Amazon Redshift replicates all your data within your data warehouse cluster when it is loaded and also continuously backs up your data to S3. Amazon Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3). Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery.

Q: How long does Amazon Redshift retain backups? Is it configurable?

By default, Amazon Redshift retains backups for 1 day. You can configure this to be as long as 35 days.

Q: How do I restore my Amazon Redshift data warehouse cluster from a backup?

You have access to all the automated backups within your backup retention window. Once you choose a backup from which to restore, we will provision a new data warehouse cluster and restore your data to it.

Q: Do I need to enable backups for my data warehouse cluster or is it done automatically?

By default, Amazon Redshift enables automated backups of your data warehouse cluster with a 1-day retention period. Free backup storage is limited to the total size of storage on the nodes in the data warehouse cluster and only applies to active data warehouse clusters. For example, if you have total data warehouse storage of 8TB, we will provide at most 8TB of backup storage at no additional charge. If you would like to extend your backup retention period beyond one day, you can do so using the AWS Management Console or the Amazon Redshift APIS. For more information on automated snapshots, please refer to the Amazon Redshift Management Guide. Amazon Redshift only backs up data that has changed so most snapshots only use up a small amount of your free backup storage.

Q: How do I manage the retention of my automated backups and snapshots?

You can use the AWS Management Console or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended).

Q: What happens to my backups if I delete my data warehouse cluster?

When you delete a data warehouse cluster, you have the ability to specify whether a final snapshot is created upon deletion, which enables a restore of the deleted data warehouse cluster at a later date. All previously created manual snapshots of your data warehouse cluster will be retained and billed at standard Amazon S3 rates, unless you choose to delete them.

Back to top »

Scalability

Q: How do I scale the size and performance of my Amazon Redshift data warehouse cluster?

If you would like to increase query performance or respond to CPU, memory or I/O over-utilization, you can increase the number of nodes within your data warehouse cluster via the AWS Management Console or the ModifyCluster API. When you modify your data warehouse cluster, your requested changes will be applied immediately. Metrics for compute utilization, memory utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon Cloudwatch's custom metric functionality.

Q: Will my data warehouse cluster remain available during scaling?

The existing data warehouse cluster remains available for read operations while a new data warehouse cluster gets created during scaling operations. When the new data warehouse cluster is ready, your existing data warehouse cluster will be temporarily unavailable while the canonical name record of the existing data warehouse cluster is flipped to point to the new data warehouse cluster. This period of unavailability typically lasts only a few minutes, and will occur during the maintenance window for your data warehouse cluster, unless you specify that the modification should be applied immediately. Amazon Redshift moves data in parallel from the compute nodes in your existing data warehouse cluster to the compute nodes in your new cluster. This enables your operation to complete as quickly as possible.

Back to top »

Querying and Analysis

Q: Is Amazon Redshift compatible with my preferred business intelligence software package and ETL tools?

Amazon Redshift uses industry-standard SQL and is accessed using standard JDBC and ODBC drivers. You can download Amazon Redshift custom JDBC and ODBC drivers from the Connect Client tab of our Console. We have validated integrations with popular BI and ETL vendors, a number of which are offering free trials to help you get started loading and analyzing your data. You can also go to the AWS Marketplace to deploy and configure solutions designed to work with Amazon Redshift in minutes.

Monitoring

Q: How do I monitor the performance of my Amazon Redshift data warehouse cluster?

Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon Cloudwatch's custom metric functionality. In addition to CloudWatch metrics, Amazon Redshift also provides information on query and cluster performance via the AWS Management Console. This information enables you to see which users and queries are consuming the most system resources and diagnose performance issues. In addition, you can see the resource utilization on each of your compute nodes to ensure that you have data and queries that are well balanced across all nodes.

Maintenance

Q: What is a maintenance window? Will my data warehouse cluster be available during software maintenance?

You can think of the Amazon Redshift maintenance window as an opportunity to control when data warehouse cluster modifications (such as scaling data warehouse cluster by adding more nodes) and software patching occur, in the event either are requested or required. If a "maintenance" event is scheduled for a given week, it will be initiated and completed at some point during the thirty-minute maintenance window you identify.

Required patching is automatically scheduled only for patches that are security and durability related. Such patching occurs infrequently (typically once every few months). If you do not specify a preferred weekly maintenance window when creating your data warehouse cluster, a default value will be assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your data warehouse cluster in the AWS Management Console or by using the ModifyCluster API. Each of your data warehouse clusters can have different preferred maintenance windows.

Back to top »

AWS Direct Connect FAQ

General Questions

Q. What is AWS Direct Connect?
AWS Direct Connect is a network service that provides an alternative to using the Internet to utilize AWS cloud services.

Q. What can I do with AWS Direct Connect?
Using AWS Direct Connect, data that would have previously been transported over the Internet can now be delivered through a private network connection between AWS and your datacenter or corporate network.

Q. What are the benefits of using AWS Direct Connect and private network connections?
In many circumstances, private network connections can reduce costs, increase bandwidth, and provide a more consistent network experience than Internet-based connections.

Q. Which AWS services can be used with AWS Direct Connect?
All AWS services, including Amazon Elastic Compute Cloud (EC2), Amazon Virtual Private Cloud (VPC), Amazon Simple Storage Service (S3), and Amazon DynamoDB can be used with AWS Direct Connect.

Q. Can I use the same private network connection with Amazon Virtual Private Cloud (VPC) and other AWS services simultaneously?
Yes. Each AWS Direct Connect connection can be configured with one or more virtual interfaces. Virtual interfaces may be configured to access AWS services such as Amazon EC2 and Amazon S3 using public IP space, or resources in a VPC using private IP space.

Q. If I'm using Amazon CloudFront and my origin is in my own data center, can I use AWS Direct Connect to transfer the objects stored in my own data center?
Yes. Amazon CloudFront supports custom origins including origins you run outside of AWS. With AWS Direct Connect, you will pay AWS Direct Connect data transfer rates for origin transfer.

Q. Where is AWS Direct Connect available?
You can find the complete list of Direct Connect locations on the Product Details page.

Q. Can I use AWS Direct Connect if my network is not present at an AWS Direct Connect location?
Yes. APN Partners supporting AWS Direct Connect can help you extend your preexisting data center or office network to an AWS Direct Connect location. Please see APN Partners for more information.

Q. How can I get started with AWS Direct Connect?
Use the AWS Direct Connect tab on the AWS Management Console to create a new connection. Then you will change the region to the region you wish to use. When requesting a connection, you will be asked to select the AWS Direct Connect location you wish to use, the number of ports, and the port speed. You will also have the opportunity to request to have an APN Partner contact you if you need assistance extending your office or data center network to the AWS Direct Connect location.

Q. Can I order a port for AWS GovCloud (US) in the AWS Management Console?
If you wish to order a port to connect to AWS GovCloud (US) you will need to use the AWS GovCloud (US) management console. Details about getting started in the AWS GovCloud (US) region can be found here.

Billing

Q. Are there any setup charges or a minimum service term commitment required to use AWS Direct Connect?
There are no setup charges, and you may cancel at any time. Services provided by APN Partners may have other terms or restrictions that apply.

Q. How will I be charged and billed for my use of AWS Direct Connect?
AWS Direct Connect has two separate charges: port-hours and Data Transfer. Pricing is per port-hour consumed for each port type. Partial port-hours consumed are billed as full hours.

Data Transfer via AWS Direct Connect will be billed in the same month in which the usage occurred. If you have a hosted virtual interface, you will only be charged for the data transferred out of that virtual interface at the applicable Data Transfer rates. The account that owns the port will be charged the port-hour charges. Read more about hosted virtual interfaces here.

For AWS Direct Connect pricing information, please seeAWS Direct Connect pricing. If using an APN partner to facilitate a Direct Connect connection, contact the partner regarding any fees they may charge.

Q. Will regional data transfer be billed at the AWS Direct Connect rate?
No, data transfer between Availability Zones in a region will be billed at the regular regional data transfer rate in the same month in which the usage occurred.

Q. What defines billable port-hours?
Port-hours are billed once the connection between the AWS router and your router is established, or 90 days after you ordered the port, whichever comes first. Port charges will continue to be billed anytime the AWS Direct Connect port is provisioned for your use. If you no longer wish to be charged for your port, please follow the cancellation process detailed in How

do I cancel the AWS Direct Connect service?.

Q. How does AWS Direct Connect work with consolidated billing?
AWS Direct Connect data transfer usage will be aggregated to your master account.

Q. How do I cancel the AWS Direct Connect service?
You can cancel AWS Direct Connect service by deleting your ports from the AWS management console. You should also cancel any service(s) offered by a third party. For example, contact the colocation provider to disconnect any cross-connects to AWS Direct Connect, and/or a network service provider who may be providing network connectivity from your remote locations to the AWS Direct Connect location.

Q: Do your prices include taxes?
Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) region is subject to Japanese Consumption Tax. Learn more.

Technical

Q. What connection speeds are supported by AWS Direct Connect?
1Gbps and 10Gbps ports are available.Speeds of 50Mbps, 100Mbps, 200Mbps, 300Mbps, 400Mbps, and 500Mbps can be ordered from any APN partners supporting AWS Direct Connect. Read more about APN Partners supporting AWS Direct Connect.

Q. Are there limits on the amount of data that I can transfer using AWS Direct Connect?
No. You may transfer any amount of data up to the limit of your selected port speed.

Q. What are the technical requirements for the connection?
AWS Direct Connect supports 1000BASE-LX or 10GBASE-LR connections over singlemode fiber using Ethernet transport. Your device must support 802.1Q VLANs. See the AWS Direct Connect User Guide for more detailed requirements information.

Q. What AWS region(s) can I connect to via this connection?
Each AWS Direct Connect location enables connectivity to the geographically nearest AWS region. You can access all AWS services available in that region.

Direct Connect locations in the US can also access the public endpoints of the other AWS regions using a public virtual interface.

Q. What Availability Zone(s) can I connect to via this connection?
Each AWS Direct Connect location enables connectivity to all Availability Zones within the geographically nearest AWS region.

Q. Are connections to AWS Direct Connect redundant?
Each connection consists of a single dedicated connection between ports on your router and an

Amazon router. We recommend establishing a second connection if redundancy is required. When you request multiple ports at the same AWS Direct Connect location, they will be provisioned on redundant Amazon routers.

Q. Will I lose connectivity if my AWS Direct Connect link fails?
If you have established a second AWS Direct Connect connection, traffic will failover to the second link automatically. We recommend enabling Bidirectional Forwarding Detection (BFD) when configuring your connections to ensure fast detection and failover. If you have configured a back-up IPsec VPN connection instead, all VPC traffic will failover to the VPN connection automatically. Traffic to/from public resources such as Amazon S3 will be routed over the Internet. If you do not have a backup AWS Direct Connect link or a IPsec VPN link, then Amazon VPC traffic will be dropped in the event of a failure. Traffic to/from public resources will be routed over the Internet.

Q. Can I extend one of my VLANs to the AWS Cloud using AWS Direct Connect?
No, VLANs are utilized in AWS Direct Connect only to separate traffic between virtual interfaces.

Q. Does AWS Direct Connect offer a Service Level Agreement (SLA)?
Not at this time.

Q: What are the technical requirements for virtual interfaces to public AWS services such as Amazon EC2 and Amazon S3?
This connection requires the use of the Border Gateway Protocol (BGP) with an Autonomous System Number (ASN) and IP Prefixes. You will need the following information to complete the connection:

- A public or private ASN. If you are using a public ASN, you must own it. If you are using a private ASN, it must be in the 64512 to 65535 range.
- A new unused VLAN tag that you select
- Public IPs (/30) allocated by you for the BGP session

Amazon will advertise public IP prefixes for the region via BGP. Direct Connect customers in the US will receive the public IP prefixes for all US regions. You must advertise public IP prefixes (/30 or smaller) that you own via BGP. For more details, consult the AWS Direct Connect User Guide.

Q: What is an Autonomous System Number (ASN) and do I need one to use AWS Direct Connect?
Autonomous System numbers are used to identify networks that present a clearly defined external routing policy to the Internet. AWS Direct Connect requires an ASN to create a public or private virtual interface. You may use a public ASN which you own, or you can pick any private ASN number between 64512 to 65535 range.

Q. What IP address will be assigned to each end of a virtual interface?
If you are configuring a virtual interface to the public AWS cloud, the IP addresses for both ends of the connection must be allocated from public IP space that you own. If the virtual interface is to a VPC and you choose to have AWS auto-generate the peer IP CIDR, the IP address space for both ends of the connection will be allocated by AWS in the 169.254.0.0/16 range.

Q: Can I connect to the Internet via this connection?
No.

Q: If I have more than one virtual interface attached, can I exchange traffic between the two ports?
Not for public Direct Connect virtual interfaces; but you can exchange traffic between the two ports in the same region if they are connecting to the same VGW.

Q: When creating a virtual interface to work with AWS services using public IP space, what IP prefixes will I receive via BGP?
You will receive all Amazon IP prefixes for the region that you are connecting to. This includes prefixes necessary to reach AWS services, and may include prefixes for other Amazon affiliates, including those of www.amazon.com. For the current list of prefixes advertised by AWS, please download the JSON of AWS IP Address Ranges. Direct Connect customers in the US will receive the public IP prefixes for all US regions. Standard AWS Direct Connect data transfer rates apply for all traffic routed through your AWS Direct Connect connection. Please see the AWS Direct Connect community forum for the additional details in the routing policy of the public virtual interface.

Q. What IP prefixes should I advertise over BGP for virtual interfaces to public AWS services?
You should advertise appropriate public IP prefixes that you own over BGP. Traffic from AWS services destined for these prefixes will be routed over your AWS Direct Connect connection.

Q. Can I locate my hardware next to the equipment that powers AWS Direct Connect?
You can procure rack space within the facility housing the AWS Direct Connect location and deploy your equipment nearby. However, AWS customer equipment cannot be placed within AWS Direct Connect racks or cage areas for security reasons. For more information, contact the APN Partner for the particular facility. Once deployed, you can connect this equipment to AWS Direct Connect using a cross-connect.

Q. How do I enable BFD on my Direct Connect connection?

Asynchronous BFD is automatically enabled for each Direct Connect virtual interface, but will not take effect until it's configured on your router. AWS has set the BFD liveness detection minimum interval to 300, and the BFD liveness detection multiplier to 3.

Q. How do I set up Direct Connect for the AWS GovCloud (US) Region?
See the AWS GovCloud (US) User Guide for detailed instructions on how to set up a Direct

Connect connection for the AWS GovCloud (US) region.

Using AWS Direct Connect with Amazon Virtual Private Cloud

Q. What are the technical requirements for virtual interfaces to VPCs?
This connection requires the use of Border Gateway Protocol (BGP). You will need the following information to complete the connection:

- A public or private ASN. If you are using a public ASN you must own it. If you are using a private ASN, it must be in the 64512 to 65535 range.
- A new unused VLAN tag that you select
- The VPC Virtual Private Gateway (VGW) ID

AWS will allocate private IPs (/30) in the 169.x.x.x range for the BGP session and will advertise the VPC CIDR block over BGP. You can advertise the default route via BGP.

Q. How does AWS Direct Connect differ from an IPSec VPN Connection?
A VPC VPN Connection utilizes IPSec to establish encrypted network connectivity between your intranet and Amazon VPC over the Internet. VPN Connections can be configured in minutes and are a good solution if you have an immediate need, have low to modest bandwidth requirements, and can tolerate the inherent variability in Internet-based connectivity. AWS Direct Connect does not involve the Internet; instead, it uses dedicated, private network connections between your intranet and Amazon VPC.

Q. Can I use AWS Direct Connect and a VPN Connection to the same VPC simultaneously?
Yes. However, only in fail-over scenarios. The Direct Connect path will always be preferred, when established, regardless of AS path prepending.

Q. Can I establish a Layer 2 connection between VPC and my network?
No, Layer 2 connections are not supported.

Amazon Route 53 FAQ

Getting Started

Q. What is a Domain Name System (DNS) Service?

DNS is a globally distributed service that translates human readable names like *www.example.com* into the numeric IP addresses like *192.0.2.1* that computers use to connect to each other. The Internet's DNS system works much like a phone book by managing the mapping between names and numbers. For DNS, the names are domain names *(www.example.com)* that are easy for people to remember and the numbers are IP addresses *(192.0.2.1)* that specify the location of computers on the Internet. DNS servers translate requests for names into IP addresses, controlling which server an end user will reach when they type a domain name into their web browser. These requests are called "queries."

Q. What is Amazon Route 53?

Amazon Route 53 provides highly available and scalable Domain Name System (DNS), domain name registration, and health-checking web services. It is designed to give developers and businesses an extremely reliable and cost effective way to route end users to Internet applications by translating names like *example.com* into the numeric IP addresses, such as *192.0.2.1*, that computers use to connect to each other. You can combine your DNS with health-checking services to route traffic to healthy endpoints or to independently monitor and/or alarm on endpoints. You can also purchase and manage domain names such as *example.com* and automatically configure DNS settings for your domains. Route 53 effectively connects user requests to infrastructure running in AWS – such as Amazon EC2 instances, Elastic Load Balancing load balancers, or Amazon S3 buckets – and can also be used to route users to infrastructure outside of AWS.

Q. What can I do with Amazon Route 53?

With Amazon Route 53, you can create and manage your public DNS records. Like a phone book, Route 53 lets you manage the IP addresses listed for your domain names in the Internet's DNS phone book. Route 53 also answers requests to translate specific domain names like into their corresponding IP addresses like *192.0.2.1*. You can use Route 53 to create DNS records for a new domain or transfer DNS records for an existing domain. The simple, standards-based REST API for Route 53 allows you to easily create, update and manage DNS records. Route 53 additionally offers health checks to monitor the health and performance of your application as well as your web servers and other resources. You can also register new domain names or transfer in existing domain names to be managed by Route 53.

Q. How do I get started with Amazon Route 53?

Amazon Route 53 has a simple web service interface that lets you get started in minutes. Your DNS records are organized into "hosted zones" that you configure with the AWS Management Console or Route 53's API. To use Route 53, you simply:

- Subscribe to the service by clicking on the sign-up button on theservice page.
- If you already have a domain name:

    - Use the AWS Management Console or the*CreateHostedZone* API to create a hosted zone that can store DNS records for your domain. Upon creating the hosted zone, you receive four Route 53 name servers across four different Top-Level Domains (TLDs) to help ensure a high level of availability.
    - Additionally, you can transfer your domain name to Route 53's management via either the AWS Management Console or the API.
- If you don't already have a domain name:
    - Use the AWS Management Console or the API to register your new domain name.
    - Route 53 automatically creates a hosted zone that stores DNS records for your domain. You also receive four Route 53 name servers across four different Top-Level Domains (TLDs) to help ensure a high level of availability.
- Your hosted zone will be initially populated with a basic set of DNS records, including four virtual name servers that will answer queries for your domain. You can add, delete or change records in this set by using the AWS Management Console or by calling the *ChangeResourceRecordSet* API . A list of supported DNS records is availablehere.
- If your domain name is not managed by Route 53, you will need to inform the registrar with whom you registered your domain name to update the name servers for your domain to the ones associated with your hosted zone. If your domain name is managed by Route 53 already, your domain name will be automatically associated with the name servers hosting your zone.

Q. How does Amazon Route 53 provide high availability and low latency?

Route 53 is built using AWS's highly available and reliable infrastructure. The globally distributed nature of our DNS servers helps ensure a consistent ability to route your end users to your application by circumventing any internet or network related issues. Route 53 is designed to provide the level of dependability required by important applications. Using a global anycast network of DNS servers around the world, Route 53 is designed to automatically answer queries from the optimal location depending on network conditions. As a result, the service offers low query latency for your end users.

Q. What are the DNS server names for the Amazon Route 53 service?

To provide you with a highly available service, each Amazon Route 53 hosted zone is served by its own set of virtual DNS servers. The DNS server names for each hosted zone are thus assigned by the system when that hosted zone is created.

Q. What is the difference between a Domain and a Hosted Zone?

A domain is a general DNS concept. Domain names are easily recognizable names for numerically addressed Internet resources. For example, *amazon.com* is a domain. A hosted zone is an Amazon Route 53 concept. A hosted zone is analogous to a traditional DNS zone file; it represents a collection of records that can be managed together, belonging to a single parent domain name. All resource record sets within a hosted zone must have the hosted zone's domain name as a suffix. For example, the *amazon.com* hosted zone may contain records named *www.amazon.com*, and *www.aws.amazon.com*, but not a record named *www.amazon.ca*. You can use the Route 53 Management Console or API to create, inspect, modify, and delete hosted zones. You can also use the Management Console or API to register new domain names and transfer in existing domain names into Route 53's management.

Q. What is the price of Amazon Route 53?

Amazon Route 53 charges are based on actual usage of the service for Hosted Zones, Queries, Health Checks, and Domain Names. For full details, see the Amazon Route 53 pricing page.

You pay only for what you use. There are no minimum fees, no minimum usage commitments, and no overage charges. You can estimate your monthly bill using the AWS Simple Monthly Calculator.

Q. What types of access controls can I set for the management of my Domains on Amazon Route 53?

You can control management access to your Amazon Route 53 hosted zone by using the AWS Identity and Access Management (IAM) service. AWS IAM allows you to control who in your organization can make changes to your DNS records by creating multiple users and managing the permissions for each of these users within your AWS Account. Learn more about AWS IAM here.

Q. I have subscribed for Amazon Route 53 but when I try to use the service it says "The AWS Access Key ID needs a subscription for the service"

When you sign up for a new AWS service, it can take up to 24 hours in some cases to complete activation, during which time you cannot sign up for the service again. If you've been waiting longer than 24 hours without receiving an email confirming activation, this could indicate a problem with your account or the authorization of your payment details. Please contact AWS Customer Service for help.

Q. Does Amazon Route 53 offer a Service Level Agreement (SLA)?

Yes. The Amazon Route 53 SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle. More information can be found here.

Domain Name System (DNS)

Q. Does Amazon Route 53 use an anycast network?

Yes. Anycast is a networking and routing technology that helps your end users' DNS queries get answered from the optimal Route 53 location given network conditions. As a result, your users get high availability and improved performance with Route 53.

Q. Is there a limit to the number of hosted zones I can manage using Amazon Route 53?

Each Amazon Route 53 account is limited to a maximum of 500 hosted zones and 10,000 resource record sets per hosted zone. Complete our request for a higher limit and we will respond to your request within two business days.

Q. How can I import a zone into Route 53?

Route 53 supports importing standard DNS zone files which can be exported from many DNS providers as well as standard DNS server software such as BIND. For newly-created hosted zones, as well as existing hosted zones that are empty except for the default NS and SOA records, you can paste your zone file directly into the Route 53 console, and Route 53 automatically creates the records in your hosted zone. To get started with zone file import, read our walkthrough in the Amazon Route 53 Developer Guide.

Q. Can I create multiple hosted zones for the same domain name?

Yes. Creating multiple hosted zones allows you to verify your DNS setting in a "test" environment, and then replicate those settings on a "production" hosted zone. For example, hosted zone Z1234 might be your test version of *example.com*, hosted on name servers ns-1, ns-2, ns-3, and ns-4. Similarly, hosted zone Z5678 might be your production version of *example.com*, hosted on ns-5, ns-6, ns-7, and ns-8. Since each hosted zone has a virtual set of name servers associated with that zone, Route 53 will answer DNS queries for example.com differently depending on which name server you send the DNS query to.

Q. Does Amazon Route 53 also provide website hosting?

No. Amazon Route 53 is an authoritative DNS service and does not providewebsite hosting. However, you can use Amazon Simple Storage Service (Amazon S3) to host a static website. To host a dynamic website or other web applications, you can use Amazon Elastic Compute Cloud (Amazon EC2), which provides flexibility, control, and significant cost savings over traditional web hosting solutions. Learn more about Amazon EC2 here. For both static and dynamic websites, you can provide low latency delivery to your global end users with Amazon CloudFront. Learn more about Amazon CloudFront here.

Q. Which DNS record types does Amazon Route 53 support?

Amazon Route 53 currently supports the following DNS record types:

- A (address record)
- AAAA (IPv6 address record)
- CNAME (canonical name record)
- MX (mail exchange record)
- NAPTR (name authority pointer record)
- NS (name server record)
- PTR (pointer record)
- SOA (start of authority record)
- SPF (sender policy framework)
- SRV (service locator)
- TXT (text record)
- Additionally, Amazon Route 53 offers ‘Alias’ records (an Amazon Route 53-specific virtual record). Alias records are used to map resource record sets in your hosted zone to Amazon Elastic Load Balancing load balancers, Amazon CloudFront distributions, AWS Elastic Beanstalk environments, or Amazon S3 buckets that are configured as websites. Alias records work like a CNAME record in that you can map one DNS name (example.com) to another ‘target’ DNS name (elb1234.elb.amazonaws.com). They differ from a CNAME record in that they are not visible to resolvers. Resolvers only see the A record and the resulting IP address of the target record.

We anticipate adding additional record types in the future.

Q. Does Amazon Route 53 support wildcard entries? If so, what record types support them?

Yes. To make it even easier for you to configure DNS settings for your domain, Amazon Route 53 supports wildcard entries for all record types. A wildcard entry is a record in a DNS zone that will match requests for any domain name based on the configuration you set. For example, a wildcard DNS record such as **.example.com* will match queries for *www.example.com* and *subdomain.example.com*.

Q. What is the default TTL for the various record types and can I change these values?

The time for which a DNS resolver caches a response is set by a value called the time to live (TTL) associated with every record. Amazon Route 53 does not have a default TTL for any record type. You must always specify a TTL for each record so that caching DNS resolvers can cache your DNS records to the length of time specified through the TTL.

Q. Can I use 'Alias records with my sub-domains?

Yes. You can also use Alias records to map your sub-domains (*www.example.com*, *pictures.example.com*, etc.) to your ELB load balancers, CloudFront distributions, or S3 website buckets.

Q. Are changes to resource record sets transactional?

Yes. A transactional change helps ensure that the change is consistent, reliable, and independent of other changes. Amazon Route 53 has been designed so that changes complete entirely on any individual DNS server, or not at all. This helps ensure your DNS queries are always answered consistently, which is important when making changes such as flipping between destination servers. When using the API, each call to *ChangeResourceRecordSets* returns an identifier that can be used to track the status of the change. Once the status is reported as *INSYNC*, your change has been performed on all of the Route 53 DNS servers.

Q. Can I associate multiple IP addresses with a single record?

Yes. Associating multiple IP addresses with a single record is often used for balancing the load of geographically-distributed web servers. Amazon Route 53 allows you to list multiple IP addresses for an A record and responds to DNS requests with the list of all configured IP addresses.

Q. How quickly will changes I make to my DNS settings on Amazon Route 53 propagate globally?

Amazon Route 53 is designed to propagate updates you make to your DNS records to its world-wide network of authoritative DNS servers within 60 seconds under normal conditions. A change is successfully propagated world-wide when the API call returns an *INSYNC* status listing.

Note that caching DNS resolvers are outside the control of the Amazon Route 53 service and will cache your resource record sets according to their time to live (TTL). The *INSYNC* or *PENDING* status of a change refers only to the state of Route 53's authoritative DNS servers.

Q. Can I see a history of my changes and other operations on my Route 53 resources?

Yes, via AWS CloudTrail you can record and log the API call history for Route 53. Please reference the CloudTrail product page to get started.

Q. Can I use AWS CloudTrail logs to roll back changes to my hosted zones?

No. We recommend that you do not use CloudTrail logs to roll back changes to your hosted zones, because reconstruction of your zone change history using your CloudTrail logs may be incomplete.

Your AWS CloudTrail logs can be used for the purposes of security analysis, resource change tracking, and compliance auditing.

Q. Does Amazon Route 53 support DNSSEC?

Amazon Route 53 does not support DNSSEC for DNS at this time. But Amazon Route 53 allows DNSSEC on domain registration.

Q. Does Amazon Route 53 support IPv6?

Yes. Amazon Route 53 supports both forward (AAAA) and reverse (PTR) IPv6 records. The Amazon Route 53 service itself is also available over IPv6. Recursive DNS resolvers on IPv6 networks can utilize either IPv4 or IPv6 transport in order to make DNS queries over Amazon Route 53.

Q. Can I point my zone apex (example.com versus www.example.com) at my Elastic Load Balancer?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (*example.com*) DNS name to your ELB DNS name (i.e. *elb1234.elb.amazonaws.com*). IP addresses associated with Amazon Elastic Load Balancers can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one or more IP addresses for the load balancer. Queries to Alias records that are mapped to ELB load balancers are free. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my website hosted on Amazon S3?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (*example.com*) DNS name to your Amazon S3 website bucket (i.e. *example.com.s3-website-us-west-2.amazonaws.com*). IP addresses associated with Amazon S3 website endpoints can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one IP address for the bucket. Route 53 doesn't charge for queries to Alias records that are mapped to an S3 bucket that is configured as a website. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my Amazon CloudFront distribution?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (*example.com*) DNS name to your Amazon CloudFront distribution (for example, *d123.cloudfront.net*). IP addresses associated with Amazon CloudFront endpoints vary based on your end user’s location (in order to direct the end user to the nearest CloudFront edge location) and can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with the IP address(es) for the distribution. Route

53 doesn't charge for queries to Alias records that are mapped to a CloudFront distribution. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my AWS Elastic Beanstalk environment?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (*example.com*) DNS name to your AWS Elastic Beanstalk DNS name (i.e. *example.elasticbeanstalk.com*). IP addresses associated with AWS Elastic Beanstalk environments can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one or more IP addresses for the environment. Queries to Alias records that are mapped to AWS Elastic Beanstalk environments are free. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. How can I use Amazon Route 53 with Amazon Simple Storage Service (Amazon S3) and Amazon CloudFront?

For websites delivered via Amazon CloudFront or static websites hosted on Amazon S3, you can use the Amazon Route 53 service to create an Alias record for your domain which points to the CloudFront distribution or S3 website bucket. For S3 buckets not configured to host static websites, you can create a CNAME record for your domain and the S3 bucket name. In all cases, note that you will also need to configure your S3 bucket or your CloudFront distribution respectively with the alternate domain name entry to completely establish the alias between your domain name and the AWS domain name for your bucket or distribution.

For CloudFront distributions and S3 buckets configured to host static websites, we recommend creating an ‘Alias’ record that maps to your CloudFront distribution or S3 website bucket, instead of using CNAMEs. Alias records have two advantages: first, unlike CNAMEs, you can create an Alias record for your zone apex (e.g. example.com, instead of www.example.com), and second, queries to Alias records are free of charge.

Q. Why does the DNS Query Test Tool return a response different than the dig or nslookup commands?

When resource record sets are changed in Amazon Route 53, the service propagates updates you make to your DNS records to its world-wide network of authoritative DNS servers. If you test the record before propagation is complete, you may see an old value when you use the dig or nslookup utilities. Additionally, DNS resolvers on the internet are outside the control of the Amazon Route 53 service and will cache your resource record sets according to their time to live (TTL), which means a dig/nslookup command might return a cached value. You should also make sure that your domain name registrar is using the name servers in your Amazon Route 53 hosted zone. If not, Amazon Route 53 will not be authoritative for queries to your domain.

DNS Routing Policies

Q. Does Amazon Route 53 support Weighted Round Robin (WRR)?

Yes. Weighted Round Robin allows you to assign weights to resource record sets in order to specify the frequency with which different responses are served. You may want to use this capability to do A/B testing, sending a small portion of traffic to a server on which you've made a software change. For instance, suppose you have two record sets associated with one DNS name—one with weight 3 and one with weight 1. In this case, 75% of the time Route 53 will return the record set with weight 3 and 25% of the time Route 53 will return the record set with weight 1. Weights can be any number between 0 and 255.

Q. What is Amazon Route 53's Latency Based Routing (LBR) feature?

LBR (Latency Based Routing) is a new feature for Amazon Route 53 that helps you improve your application's performance for a global audience. You can run applications in multiple AWS regions and Amazon Route 53, using dozens of edge locations worldwide, will route end users to the AWS region that provides the lowest latency.

Q. How do I get started using Amazon Route 53's Latency Based Routing (LBR) feature?

You can start using Amazon Route 53's new LBR feature quickly and easily by using either the AWS Management Console or a simple API. You simply create a record set that includes the IP addresses or ELB names of various AWS endpoints and mark that record set as an LBR-enabled Record Set, much like you mark a record set as a Weighted Record Set. Amazon Route 53 takes care of the rest - determining the best endpoint for each request and routing end users accordingly, much like Amazon CloudFront, Amazon's global content delivery service, does. You can learn more about how to use Latency Based Routing in the Amazon Route 53 Developer Guide.

Q. What is the price for Amazon Route 53's Latency Based Routing (LBR) feature?

Like all AWS services, there are no upfront fees or long term commitments to use Amazon Route 53 and LBR. Customers simply pay for the hosted zones and queries they actually use. Please visit the Amazon Route 53 pricing page for details on pricing for Latency Based Routing queries.

Q. What is Amazon Route 53's Geo DNS feature?

Route 53 Geo DNS lets you balance load by directing requests to specific endpoints based on the geographic location from which the request originates. Geo DNS makes it possible to customize localized content, such as presenting detail pages in the right language or restricting distribution of content to only the markets you have licensed. Geo DNS also lets you balance load across endpoints in a predictable, easy-to-manage way, ensuring that each end-user location is consistently routed to the same endpoint. Geo DNS provides three levels of geographic granularity: continent, country, and state, and Geo DNS also provides a global record which is served in cases where an end user's location doesn't match any of the specific Geo DNS records you have created. You can also combine Geo DNS with other routing types, such as Latency Based Routing and DNS Failover, to enable a variety of low-latency and fault-tolerant architectures. For information on how to configure various routing types, please see the Amazon Route 53 documentation.

Q. How do I get started using Amazon Route 53's Geo DNS feature?

You can start using Amazon Route 53's Geo DNS feature quickly and easily by using either the AWS Management Console or the Route 53 API. You simply create a record set and specify the applicable values for that type of record set, mark that record set as a Geo DNS-enabled Record Set, and select the geographic region (global, continent, country, or state) that you want the record to apply to. You can learn more about how to use Geo DNS in the Amazon Route 53 Developer Guide.

Q. When using Geo DNS, do I need a "global" record? When would Route 53 return this record?

Yes, we strongly recommend that you configure a global record, to ensure that Route 53 can provide a response to DNS queries from all possible locations—even if you have created specific records for each continent, country, or state where you expect your end users will be located. Route 53 will return the value contained in your global record in the following cases:

- The DNS query comes from an IP address not recognized by Route 53's Geo IP database.
- The DNS query comes from a location not included in any of the specific Geo DNS records you have created.

Q. Can I have a Geo DNS record for a continent and different Geo DNS records for countries within that continent? Or a Geo DNS record for a country and Geo DNS records for states within that country?

Yes, you can have Geo DNS records for overlapping geographic regions (e.g., a continent and countries within that continent, or a country and states within that country). For each end user's location, Route 53 will return the most specific Geo DNS record that includes that location. In other words, for a given end user's location, Route 53 will first return a state record; if no state record is found, Route 53 will return a country record; if no country record is found, Route 53 will return a continent record; and finally, if no continent record is found, Route 53 will return the global record.

Q. What is the price for Route 53's Geo DNS feature?

Like all AWS services, there are no upfront fees or long term commitments to use Amazon Route 53 and Geo DNS. Customers simply pay for the hosted zones and queries they actually use. Please visit the Amazon Route 53 pricing page for details on pricing for Geo DNS queries.

Q. What is the difference between Latency Based Routing and Geo DNS?

Geo DNS bases routing decisions on the geographic location of the requests. In some cases, geography is a good proxy for latency; but there are certainly situations where it is not. LatencyBased Routing utilizes latency measurements between viewer networks and AWS datacenters. These measurements are used to determine which endpoint to direct users toward.

If your goal is to minimize end-user latency, we recommend using Latency Based Routing. If you have compliance, localization requirements, or other use cases that require stable routing from a specific geography to a specific endpoint, we recommend using Geo DNS.

DNS Traffic Flow

Q. What is Amazon Route 53 Traffic Flow?

Amazon Route 53 Traffic Flow is an easy-to-use and cost-effective global traffic management service. With Amazon Route 53 Traffic Flow, you can improve the performance and availability of your application for your end users by running multiple endpoints around the world, using Amazon Route 53 Traffic Flow to connect your users to the best endpoint based on latency, geography, and endpoint health. Amazon Route 53 Traffic Flow makes it easy for developers to create policies that route traffic based on the constraints they care most about, including latency, endpoint health, load, and geography. Customers can customize these templates or build policies from scratch using a simple visual policy builder in the AWS Management Console.

Q. What is the difference between a traffic policy and a policy record?

A **traffic policy** is the set of rules that you define to route end users' requests to one of your application's endpoints. You can create a traffic policy using the visual policy builder in the Amazon Route 53 Traffic Flow section of the Amazon Route 53 console. You can also create traffic policies as JSON-formatted text files and upload these policies using the Route 53 API, the AWS CLI, or the various AWS SDKs.

By itself, a traffic policy doesn't affect how end users are routed to your application, because it isn't yet associated with your application's DNS name (such as *www.example.com*). To start using Amazon Route 53 Traffic Flow to route traffic to your application using the traffic policy you've created, you create a **policy record** which associates the traffic policy with the appropriate DNS name within an Amazon Route 53 hosted zone that you own. For example, if you want to use a traffic policy that you've named *my-first-traffic-policy* to manage traffic for your application at *www.example.com*, you will create a policy record for*www.example.com* within your hosted zone *example.com* and choose *my-first-traffic-policy* as the traffic policy.

Policy records are visible in both the Amazon Route 53 Traffic Flow and Amazon Route 53 Hosted Zone sections of the Amazon Route 53 console.

Q. Can I use the same policy to manage routing for more than one DNS name?

Yes. You can reuse a policy to manage more than one DNS name in one of two ways. First, you can create additional policy records using the policy. Note that there is an additional charge for using this method, because you are billed for each policy record that you create.

The second method is to create one policy record using the policy, and then for each additional DNS name that you want to manage using the policy, you create a standard CNAME record pointing at the DNS name of the policy record that you created. For example, if you create a policy record for *example.com*, you can then create DNS records for*www.example.com*, *blog.example.com*, and *www.example.net* with a CNAME value of*example.com* for each record. Note that this method is not possible for records at the zone apex, such as *example.net*, *example.org*, or *example.co.uk* (without www or another subdomain in front of the domain name). For records at the zone apex, you must create a policy record using your traffic policy.

Q. Can I create an Alias record pointing to a DNS name that is managed by a traffic policy?

No, it is not possible to create an Alias record pointing to a DNS name that is being managed by a traffic policy.

Q. Is there a charge for traffic policies that don't have a policy record?

No. We only charge for policy records; there is no charge for creating the traffic policy itself.

Q. How am I billed for using Amazon Route 53 Traffic Flow?

You are billed per policy record. A policy record represents the application of a Traffic Flow policy to a specific DNS name (such as *www.example.com*) in order to use the traffic policy to manage how requests for that DNS name are answered. Billing is monthly and is prorated for partial months. There is no charge for traffic policies that are not associated with a DNS name via a policy record. For details on pricing, see the Amazon Route 53 pricing page.

Private DNS

Q. What is Private DNS?

Private DNS is a Route 53 feature that lets you have authoritative DNS within your VPCs without exposing your DNS records (including the name of the resource and its IP address(es) to the Internet.

Q. Can I use Amazon Route 53 to manage my organization's private IP addresses?

Yes, you can manage private IP addresses within Virtual Private Clouds (VPCs) using Amazon Route 53’s Private DNS feature. With Private DNS, you can create a private hosted zone, and Route 53 will only return these records when queried from within the VPC(s) that you have associated with your private hosted zone. For more details, see the Amazon Route 53 Documentation.

Q. How do I set up Private DNS?

You can set up Private DNS by creating a hosted zone in Route 53, selecting the option to make the hosted zone “private”, and associating the hosted zone with one of your VPCs. After creating the hosted zone, you can associate it with additional VPCs. See the Amazon Route 53 Documentation for full details on how to configure Private DNS.

Q. Do I need connectivity to the outside Internet in order to use Private DNS?

You can resolve internal DNS names from resources within your VPC that do not have Internet connectivity. However, to update the configuration for your Private DNS hosted zone, you need Internet connectivity to access the Route 53 API endpoint, which is outside of VPC.

Q. Can I still use Private DNS if I’m not using VPC?

No. Route 53 Private DNS uses VPC to manage visibility and provide DNS resolution for private DNS hosted zones. To take advantage of Route 53 Private DNS, you must configure a VPC and migrate your resources into it.

Q. Can I use the same private Route 53 hosted zone for multiple VPCs?

Yes, you can associate multiple VPCs with a single hosted zone.

Q. Can I associate VPCs and private hosted zones that I created under different AWS accounts?

Yes, you can associate VPCs belonging to different accounts with a single hosted zone. You can see more details here.

Q. Will Private DNS work across AWS regions?

Yes. DNS answers will be available within every VPC that you associate with the private hosted zone. Note that you will need to ensure that the VPCs in each region have connectivity with each other in order for resources in one region to be able to reach resources in another region. Route 53 Private DNS is supported today in the US East (Northern Virginia), US West (Northern California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), EU (Frankfurt), EU (Ireland), and South America (Sao Paulo) regions.

Q. Can I configure DNS Failover for Private DNS hosted zones?

Yes, it is possible to configure DNS Failover by associating health checks with resource record sets within a Private DNS hosted zone. If your endpoints are within a Virtual Private Cloud (VPC), you have several options to configure health checks against these endpoints. If the endpoints have public IP addresses, then you can create a standard health check against the public IP address of each endpoint. If your endpoints only have private IP addresses, then you cannot create standard health checks against these endpoints. However, you can create metric based health checks, which function like standard Amazon Route 53 health checks except that they use an existing Amazon CloudWatch metric as the source of endpoint health information instead of making requests against the endpoint from external locations.

Q. Can I use Private DNS to block domains and DNS names that I don’t want to be reached from within my VPC?

Yes, you can block domains and specific DNS names by creating these names in one or more Private DNS hosted zones and pointing these names to your own server (or another location that you manage).

Health Checks & DNS Failover

Q. What is DNS Failover?

DNS Failover consists of two components: health checks and failover. Health checks are automated requests sent over the Internet to your application to verify that your application is reachable, available, and functional. You can configure the health checks to be similar to the typical requests made by your users, such as requesting a web page from a specific URL. With DNS failover, Route 53 only returns answers for resources that are healthy and reachable from the outside world, so that your end users are routed away from a failed or unhealthy part of your application.

Q. How do I get started with DNS Failover?

Visit the Amazon Route 53 Developer Guide for details on getting started. You can also configure DNS Failover from within the Route 53 Console.

Q. Does DNS Failover support Elastic Load Balancers (ELBs) as endpoints?

Yes, you can configure DNS Failover for Elastic Load Balancers (ELBs). To enable DNS Failover for an ELB endpoint, create an Alias record pointing to the ELB and set the “Evaluate Target Health” parameter to true. Route 53 creates and manages the health checks for your ELB automatically. You do not need to create your own Route 53 health check of the ELB. You also do not need to associate your resource record set for the ELB with your own health check, because Route 53 automatically associates it with the health checks that Route 53 manages on your behalf. The ELB health check will also inherit the health of your backend instances behind that ELB. For more details on using DNS Failover with ELB endpoints, please consult the Route

53 Developer Guide.

Q. Can I configure a backup site to be used only when a health check fails?

Yes, you can use DNS Failover to maintain a backup site (for example, a static site running on an Amazon S3 website bucket) and fail over to this site in the event that your primary site becomes unreachable.

Q. What DNS record types can I associate with Route 53 health checks?

You can associate any record type supported by Route 53 except SOA and NS records.

Q. Can I health check an endpoint if I don’t know its IP address?

Yes. You can configure DNS Failover for Elastic Load Balancers and Amazon S3 website buckets via the Amazon Route 53 Console without needing to create a health check of your own. For these endpoint types, Route 53 automatically creates and manages health checks on your behalf which are used when you create an Alias record pointing to the ELB or S3 website bucket and enable the "Evaluate Target Health" parameter on the Alias record.

For all other endpoints, you can specify either the DNS name (e.g. www.example.com) or the IP address of the endpoint when you create a health check for that endpoint.

Q. One of my endpoints is outside AWS. Can I set up DNS Failover on this endpoint?

Yes. Just like you can create a Route 53 resource record that points to an address outside AWS, you can set up health checks for parts of your application running outside AWS, and you can fail over to any endpoint that you choose, regardless of location. For example, you may have a legacy application running in a datacenter outside AWS and a backup instance of that application running within AWS. You can set up health checks of your legacy application running outside AWS, and if the application fails the health checks, you can fail over automatically to the backup instance in AWS.

Q. If failover occurs and I have multiple healthy endpoints remaining, will Route 53 consider the load on my healthy endpoints when determining where to send traffic from the failed endpoint?

No, Route 53 does not make routing decisions based on the load or available traffic capacity of your endpoints. You will need to ensure that you have available capacity at your other endpoints, or the ability to scale at those endpoints, in order to handle the traffic that had been flowing to your failed endpoint.

Q. How many consecutive health check observations does an endpoint need to fail to be considered “failed”?

The default is a threshold of three health check observations: when an endpoint has failed three consecutive observations, Route 53 will consider it failed. However, Route 53 will continue to perform health check observations on the endpoint and will resume sending traffic to it once it passes three consecutive observations. You can change this threshold to any value between 1 and 10 observations. For more details, see the Amazon Route 53 Developer Guide.

Q. When my failed endpoint becomes healthy again, how is the DNS failover reversed?

After a failed endpoint passes the number of consecutive health check observations that you specify when creating the health check (the default threshold is three observations), Route 53 will restore its DNS records automatically, and traffic to that endpoint will resume with no action required on your part.

Q. What is the interval between health check observations?

By default, health check observations are conducted at an interval of 30 seconds. You can optionally select a fast interval of 10 seconds between observations.

By checking three times more often, fast interval health checks enable Route 53 to confirm more quickly that an endpoint has failed, shortening the time required for DNS failover to redirect traffic in response to the endpoint's failure.

Fast interval health checks also generate three times the number of requests to your endpoint, which may be a consideration if your endpoint has a limited capacity to serve web traffic. Visit the Route 53 pricing page for details on pricing for fast interval health checks and other optional health check features. For more details, see the Amazon Route 53 Developer Guide.

Q. How much load should I expect a health check to generate on my endpoint (for example, a web server)?

Each heath check is conducted from multiple locations around the world. The number and set of locations is configurable; you can modify the number of locations from which each of your health checks is conducted using the Amazon Route 53 console or API. Each location checks the endpoint independently at the interval that you select: the default interval of 30 seconds, or an optional fast interval of 10 seconds. Based on the current default number of health checking locations, you should expect your endpoint to receive one request every 2-3 seconds on average for standard interval health checks and one or more requests per second for fast-interval health checks.

Q. Do Route 53 health checks follow HTTP redirects?

No. Route 53 health checks consider an HTTP 3xx code to be a successful response, so they don't follow the redirect. This may cause unexpected results for string-matching health checks. The health check searches for the specified string in the body of the redirect. Because the health check doesn't follow the redirect, it never sends a request to the location that the redirect points to and never gets a response from that location. For string matching health checks, we recommend that you avoid pointing the health check at a location that returns an HTTP redirect.

Q. What is the sequence of events when failover happens?

In simplest terms, the following events will take place if a health check fails and failover occurs:

1. Route 53 conducts a health check of your application. In this example, your application fails three consecutive health checks, triggering the following events.

2. Route 53 disables the resource records for the failed endpoint and no longer serves these records. This is the failover step, which causes traffic to begin being routed to your healthy endpoint(s) instead of your failed endpoint.

Q. Do I need to adjust the TTL for my records in order to use DNS Failover?

The time for which a DNS resolver caches a response is set by a value called the time to live (TTL) associated with every record. We recommend a TTL of 60 seconds or less when using DNS Failover, to minimize the amount of time it takes for traffic to stop being routed to your failed endpoint. In order to configure DNS Failover for ELB and S3 Website endpoints, you need to use Alias records which have fixed TTL of 60 seconds; for these endpoint types, you do not need to adjust TTLs in order to use DNS Failover.

Q. What happens if all of my endpoints are unhealthy?

Route 53 can only fail over to an endpoint that is healthy. If there are no healthy endpoints remaining in a resource record set, Route 53 will behave as if all health checks are passing.

Q. Can I use DNS Failover without using Latency Based Routing (LBR)?

Yes. You can configure DNS Failover without using LBR. In particular, you can use DNS failover to configure a simple failover scenario where Route 53 monitors your primary website and fails over to a backup site in the event that your primary site is unavailable.

Q. Can I configure a health check on a site accessible only via HTTPS?

Yes. Route 53 supports health checks over HTTPS, HTTP or TCP.

Q. Do HTTPS health checks validate the endpoint’s SSL certificate?

No, HTTPS health checks test whether it’s possible to connect with the endpoint over SSL and whether the endpoint returns a valid HTTP response code. However, they do not validate the SSL certificate returned by the endpoint.

Q. Do HTTPS health checks support Server Name Indication (SNI)?

Yes, HTTPS health checks support SNI.

Q. How can I use health checks to verify that my web server is returning the correct content?

You can use Route 53 health checks to check for the presence of a designated string in a server response by selecting the “Enable String Matching” option. This option can be used to check a web server to verify that that the HTML it serves contains an expected string. Or, you can create a dedicated status page and use it to check the health of the server from an internal or operational perspective. For more details, see the Amazon Route 53 Developer Guide.

Q. How do I see the status of a health check that I’ve created?

You can view the current status of a health check, as well as details on why it has failed, in the Amazon Route 53 console and via the Route 53 API.

Additionally, each health check’s results are published as Amazon CloudWatch metrics showing the endpoint’s health and, optionally, the latency of the endpoint’s response. You can view a graph of the Amazon CloudWatch metric in the health checks tab of the Amazon Route 53 console to see the current and historical status of the health check. You can also create Amazon CloudWatch alarms on the metric in order to send notifications if the status of the health check changes.

The Amazon CloudWatch metrics for all of your Amazon Route 53 health checks are also visible in the Amazon CloudWatch console. Each Amazon CloudWatch metric contains the Health Check ID (for example, 01beb6a3-e1c2-4a2b-a0b7-7031e9060a6a) which you can use to identify which health check the metric is tracking.

Q. How can I measure the performance of my application’s endpoints using Amazon Route 53?

Amazon Route 53 health checks include an optional latency measurement feature which provides data on how long it takes your endpoint to respond to a request. When you enable the latency measurement feature, the Amazon Route 53 health check will generate additional Amazon CloudWatch metrics showing the time required for Amazon Route 53’s health checkers to establish a connection and to begin receiving data. Amazon Route 53 provides a separate set of latency metrics for each AWS region where Amazon Route 53 health checks are conducted.

Q. How can I be notified if one of my endpoints starts failing its health check?

Because each Route 53 health check publishes its results as a CloudWatch metric, you can configure the full range of CloudWatch notifications and automated actions which can be triggered when the health check value changes beyond a threshold that you specify. First, in either the Route 53 or CloudWatch console, configure a CloudWatch alarm on the health check metric. Then add a notification action and specify the email or SNS topic that you want to publish your notification to. Please consult the Route 53 Developer Guide for full details.

Q: I created an alarm for my health check, but I need to re-send the confirmation email for the alarm's SNS topic. How can I re-send this email?

Confirmation emails can be re-sent from the SNS console.To find the name of the SNS topic associated with the alarm, click the alarm name within the Route 53 console and looking in the box labeled "Send notification to."

Within the SNS console, expand the list of topics, and select the topic from your alarm. Open the "Create Subscription" box and select Email for protocol and enter the desired email address. Clicking "Subscribe" will re-send the confirmation email.

Q. I'm using DNS Failover with Elastic Load Balancers (ELBs) as endpoints. How can I see the status of these endpoints?

The recommended method for setting up DNS Failover with ELB endpoints is to use Alias records with the "Evaluate Target Health" option. Because you don't create your own health checks for ELB endpoints when using this option, there are no specific CloudWatch metrics generated by Route 53 for these endpoints.

You can get metrics on the health of your load balancer in two ways. First, Elastic Load Balancing publishes metrics that indicate the health of the load balancer and the number of healthy instances behind it. For details on configuring CloudWatch metrics for ELB, consult the ELB developer guide. Second, you can create your own health check against the CNAME provided by the ELB, e.g. elb-example-123456678.us-west-2.elb.amazonaws.com. You won't use this health check for DNS Failover itself (because the "Evaluate Target Health" option provides DNS Failover for you), but you can view the CloudWatch metrics for this health check and create alarms to be notified if the health check fails.

For complete details on using DNS Failover with ELB endpoints, please consult theRoute 53 Developer Guide.

Q. For Alias records pointing to Amazon S3 Website buckets, what is being health checked when I set Evaluate Target Health to "true"?

Amazon Route 53 performs health checks of the Amazon S3 service itself in each AWS region. When you enable Evaluate Target Health on an Alias record pointing to an Amazon S3 Website bucket, Amazon Route 53 will take into account the health of the Amazon S3 service in the AWS region where your bucket is located. Amazon Route 53 does not check whether a specific bucket exists or contains valid website content; Amazon Route 53 will only fail over to another location if the Amazon S3 service itself is unavailable in the AWS region where your bucket is located.

Q. What is the cost to use CloudWatch metrics for my Route 53 health checks?

CloudWatch metrics for Route 53 health checks are available free of charge.

Q. Can I configure DNS Failover based on internal health metrics, such as CPU load, network, or memory?

Yes. Amazon Route 53's metric based health checks let you perform DNS failover based on any metric that is available within Amazon CloudWatch, including AWS-provided metrics and custom metrics from your own application. When you create a metric based health check within Amazon

Route 53, the health check becomes unhealthy whenever its associated Amazon CloudWatch metric enters an alarm state.

Metric based health checks are useful to enable DNS failover for endpoints that cannot be reached by a standard Amazon Route 53 health check, such as instances within a Virtual Private Cloud (VPC) that only have private IP addresses. Using Amazon Route 53's calculated health check feature, you can also accomplish more sophisticated failover scenarios by combining the results of metric based health checks with the results of standard Amazon Route 53 health checks, which make requests against an endpoint from a network of checkers around the world. For example, you can create a configuration which fails away from an endpoint if either its public-facing web page is unavailable, or if internal metrics such as CPU load, network in/out, or disk reads show that the server itself is unhealthy.

Q. My web server is receiving requests from a Route 53 health check that I did not create. How can I stop these requests?

Occasionally, Amazon Route 53 customers create health checks that specify an IP address or domain name that does not belong to them. If your web server is getting unwanted HTTP(s) requests that you have traced to Amazon Route 53 health checks, please provide information on the unwanted health check using this form, and we will work with our customer to fix the problem.

Domain Name Registration

Q. Can I register domain names with Amazon Route 53?

Yes. You can use the AWS Management Console or API to register new domain names with Route 53. You can also request to transfer in existing domain names from other registrars to be managed by Route 53. Domain name registration services are provided under our Domain Name Registration Agreement.

Q. What Top Level Domains ("TLDs") do you offer?

Route 53 offers a wide selection of both generic Top Level Domains ("gTLDs": for example, .com and .net) and country-code Top Level Domains ("ccTLDs": for example, .de and .fr). For the complete list, please see the Route 53 Domain Registration Price List.

Q. How can I register a domain name with Route 53?

To get started, log into your account and click on "Domains". Then, click the big blue "Register Domain" button and complete the registration process.

Q. How long does it take to register a domain name?

Depending on the TLD you've selected, registration can take from a few minutes to several hours. Once the domain is successfully registered, it will show up in your account.

Q. How long is my domain name registered for?

The initial registration period is typically one year, although the registries for some top-level domains (TLDs) have longer registration periods. When you register a domain with Amazon Route 53 or you transfer domain registration to Amazon Route 53, we configure the domain to renew automatically. For more information, see Renewing Registration for a Domain in the Amazon Route 53 Developer Guide.

Q. What information do I need to provide to register a domain name?

In order to register a domain name, you need to provide contact information for the registrant of the domain, including name, address, phone number, and email address. If the administrative and technical contacts are different, you need to provide that contact information, too.

Q. Why do I need to provide personal information to register a domain?

ICANN, the governing body for domain registration, requires that registrars provide contact information, including name, address, and phone number, for every domain name registration, and that registrars make this information publicly available via a Whois database. For domain names that you register as an individual (i.e., not as a company or organization), Route 53 provides privacy protection, which hides your personal phone number, email address, and physical address, free of charge. Instead, the Whois contains the registrar's name and mailing address, along with a registrar-generated forwarding email address that third parties may use if they wish to contact you.

Q. Does Route 53 offer privacy protection for domain names I have registered?

Yes, Route 53 provides privacy protection at no additional charge. The privacy protection hides your phone number, email address, and physical address. Your first and last name will be hidden if the TLD registry and registrar allow it. When you enable privacy protection, a Whois query for the domain will contain the registrar's mailing address in place of your physical address, and the registrar's name in place of your name (if allowed). Your email address will be a registrar-generated forwarding email address that third parties may use if they wish to contact you. Domain names registered by companies or organizations are eligible for privacy protection if the TLD registry and registrar allow it.

Q. Where can I find the requirements for specific TLDs?

For a list of TLDs please see theprice list and for the specific registration requirements for each, please see the Amazon Route 53 Developer Guide and our Domain Name Registration Agreement.

Q. What name servers are used to register my domain name?

When your domain name is created we automatically associate your domain with four unique Route 53 name servers, known as a delegation set. You can view the delegation set for your domain in the Amazon Route 53 console. They're listed in the hosted zone that we create for you automatically when you register a domain.

By default, Route 53 will assign a new, unique delegation set for each hosted zone you create. However, you can also use the Route 53 API to create a “reusable delegation set”, which you can then apply to multiple hosted zones that you create. For customers with large numbers of domain names, reusable delegation sets make migration to Route 53 simple, because you can instruct your domain name registrar to use the same delegation set for all your domains managed by Route 53. This feature also makes it possible for you to create “white label” name server addresses such as ns1.example.com, ns2.example.com, etc., which you can point to your Route 53 name servers. You can then use your “white label” name server addresses as the authoritative name servers for as many of your domain names as desired. For more details, see the Amazon Route 53 documentation.

Q. Will I be charged for my name servers?

You will be charged for the hosted zone that Route 53 creates for your domain name, as well as for the DNS queries against this hosted zone that Route 53 serves on your behalf. If you do not wish to be charged for Route 53’s DNS service, you can delete your Route 53 hosted zone. Please note that some TLDs require you to have valid name servers as part of your domain name registration. For a domain name under one of these TLDs, you will need to procure DNS service from another provider and enter that provider’s name server addresses before you can safely delete your Route 53 hosted zone for that domain name.

Q. What is Amazon Registrar, Inc. and what is a registrar of record?

AWS resells domain names that are registered with ICANN-accredited registrars. Amazon Registrar, Inc. is an Amazon company that is accredited by ICANN to register domains. The registrar of record is the “Sponsoring Registrar” listed in the WHOIS record for your domain to indicate which registrar your domain is registered with.

Q. Who is Gandi?

Amazon is a reseller of the registrar Gandi. As the registrar of record, Gandi is required by ICANN to contact the registrant to verify their contact information at the time of initial registration. You MUST verify your contact information if requested by Gandi within the first 15 days of registration in order to prevent your domain name from being suspended. Gandi also sends out reminder notices before the domain comes up for renewal.

Q. Which top-level domains does Amazon Route 53 register through Amazon Registrar and which ones does it register through Gandi?

See our documentation for a list of the domains that you can currently register using Amazon

Route 53. This list includes information about which registrar is the current registrar of record for each TLD that we sell.

Q. Can I transfer my .com and .net domain registrations from Gandi to Amazon?

No. We plan to add this functionality soon.

Q. What is Whois? Why is my information shown in Whois?

Whois is a publicly available database for domain names that lists the contact information and the name servers that are associated with a domain name. Anyone can access the Whois database by using the WHOIS command, which is widely available. It's included in many operating systems, and it's also available as a web application on many websites. The Internet Corporation for Assigned Names and Numbers (ICANN) requires that all domain names have publicly available contact information in case someone needs to get in contact with the domain name holder.

Q. How do I transfer my domain name to Route 53?

To get started, log into your account and click on “Domains”. Then, click the “Transfer Domain” button at the top of the screen and complete the transfer process. Please make sure before you start the transfer process, (1) your domain name is unlocked at your current registrar, (2) you have disabled privacy protection on your domain name (if applicable), and (3) that you have obtained the valid Authorization Code, or “authcode”, from your current registrar which you will need to enter as part of the transfer process.

Q. How do I transfer my existing domain name registration to Amazon Route 53 without disrupting my existing web traffic?

First, you need to get a list of the DNS record data for your domain name, generally available in the form of a “zone file” that you can get from your existing DNS provider. With the DNS record data in hand, you can use Route 53’s Management Console or simple web-services interface to create a hosted zone that can store the DNS records for your domain name and follow its transfer process, which will include such steps as updating the name servers for your domain name to the ones associated with your hosted zone. To complete the domain name transfer process, contact the registrar with whom you registered your domain name and follow its transfer process, which will include steps such as updating the name servers for your domain name to the ones associated with your hosted zone. As soon as your registrar propagates the new name server delegations, the DNS queries from your end users will start to get answered by the Route 53 DNS servers.

Q. How do I check on the status of my transfer request?

You can view the status of domain name transfers in the “Alerts” section on the homepage of the Route 53 console.

Q. What do I do if my transfer wasn’t successful?

You will need to contact your current registrar in order to determine why your transfer failed. Once they have resolved the issue, you can resubmit your transfer request.

Q. How do I transfer my domain name to a different registrar?

In order to move your domain name away from Route 53, you need to initiate a transfer request with your new registrar. They will request the domain name be moved to their management.

Q. Is there a limit to the number of domains I can manage using Amazon Route 53?

Each new Amazon Route 53 account is limited to a maximum of 50 domains. Complete our request form for a higher limitand we will respond to your request within two business days.

Q. Does Amazon Route 53 DNS support DNSSEC?

Amazon Route 53’s DNS services does NOT support DNSSEC at this time. However, our domain name registration service supports configuration of signed DNSSEC keys for domains when DNS service is configured at another provider. More information on configuring DNSSEC for your domain name registration can be found here.

Q. How do I transfer a domain registration that has DNSSEC enabled to Amazon Route 53?

See our documentation for a step-by-step guide on transferring your DNSSEC-enabled domain to Amazon Route 53.

AWS CodePipeline FAQ

General

Q: What is AWS CodePipeline?
AWS CodePipeline is a continuous delivery service that enables you to model, visualize, and automate the steps required to release your software. With AWS CodePipeline, you model the full release process for building your code, deploying to pre-production environments, testing your application and releasing it to production. AWS CodePipeline then builds, tests, and deploys your application according to the defined workflow every time there is a code change. You can integrate partner tools and your own custom tools into any stage of the release process to form an end-to-end continuous delivery solution.

Q: Why should I use AWS CodePipeline?
By automating your build, test, and release processes, AWS CodePipeline enables you to

increase the speed and quality of your software updates by running all new changes through a consistent set of quality checks.

Q: What is continuous delivery?
Continuous delivery is a software development practice where code changes are automatically built, tested, and prepared for a release to production. AWS CodePipeline is a service that helps you practice continuous delivery. Learn more about continuous delivery here.

Concepts

The diagram below represents the concepts discussed in this section.

Q: What is a pipeline?
A pipeline is a workflow construct that describes how software changes go through a release process. You define the workflow with a sequence of stages and actions.

Q: What is a revision?
A revision is a change made to the source location defined for your pipeline. It can include source code, build output, configuration, or data. A pipeline can have multiple revisions flowing through it at the same time.

Q: What is a stage?
A stage is a group of one or more actions. A pipeline can have two or more stages.

Q: What is an action?
An action is a task performed on a revision. Pipeline actions occur in a specified order, in serial or in parallel, as determined in the configuration of the stage. For more information, see Edit a Pipeline and Action Structure Requirements in AWS CodePipeline.

Q: What is an artifact?
When an action runs, it acts upon a file or set of files. These files are called artifacts. These artifacts can be worked upon by later actions in the pipeline. For example, a source action will output the latest version of the code as a source artifact, which the build action will read in. Following the compilation, the build action will upload the build output as another artifact, which will be read by the later deployment actions.

Q: What is a transition?
The stages in a pipeline are connected by transitions, and are represented by arrows in the AWS CodePipeline console. Revisions that successfully complete the actions in a stage will be automatically sent on to the next stage as indicated by the transition arrow. Transitions can be disabled or enabled between stages.

Using AWS CodePipeline

Q: How do I get started with AWS CodePipeline?
You can sign in to theAWS Management Console, create a pipeline, and start using the service. If you want an introduction to AWS CodePipeline, see Getting Started, which includes step-by-step tutorials. Or, see the Pipeline Starter Kit to quickly provision a preconfigured release pipeline with a Jenkins build server using an AWS CloudFormation template.

Q: How do I start a pipeline?
After you create a pipeline, it will automatically trigger a run to release the latest revision of your source code. From then on, every time you make a change to your source location, a new run is triggered. In addition, you can re-run the last revision through a pipeline using the Release Change button in the pipeline console.

Q: How do I stop a pipeline?
To stop a pipeline, you can disable a transition from one stage to another. Once disabled, your pipeline will continue to run revisions through the actions, but it will not promote revisions through the disabled transition to later stages. For more details, see Disable or Enable Transitions in AWS CodePipeline.

Q: Can I edit an existing pipeline?
Yes. You can use the AWS CodePipeline console or AWS CLI to add or remove stages in a pipeline as well as to add, edit, or remove actions in a stage.

Q: Can I create a copy of an existing pipeline?
Yes. You can use the get-pipeline AWS CLI command to get the JSON structure of your existing pipeline. You can then use that JSON and the create-pipeline AWS CLI command to create a new pipeline with the same structure as the existing one.

Q: Can actions run in parallel?
Yes. You can configure one or more actions to run in parallel for any given stage.

Q: What product integrations are available with AWS CodePipeline?
AWS CodePipeline integrates with AWS services such as AWS CodeCommit, Amazon S3, AWS CodeDeploy, AWS Elastic Beanstalk, AWS CloudFormation, AWS OpsWorks, and AWS Lambda. In addition, AWS CodePipeline integrates with a number of partner tools. For details see the product integrations page. Finally, you can write your own custom actions and integrate any existing tool with CodePipeline. For more details on custom actions, see the Create and Add a Custom Action in AWS CodePipeline page.

Q: Can I get a history of AWS CodePipeline API calls?
Yes. To receive a history of AWS CodePipeline API calls made on your account for security analysis and operational troubleshooting purposes, you simply turn on AWS CloudTrail in the

AWS Management Console. For more information, see Logging AWS CodePipeline API calls by Using AWS CloudTrail.

Q: What are the service limits when using AWS CodePipeline?
For information on the service limits, see Limits.

Partners

Q: What do I need to do to integrate with AWS CodePipeline?
If you're interested in becoming an AWS partner who integrates your developer service with AWS CodePipeline, please contact codepipeline-request@amazon.com.

Security

Q: Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodePipeline?
Yes. AWS CodePipeline supports resource-level permissions. You can specify which user can perform what action on a pipeline. For example, you can provide a user read-only access to a pipeline, if you want them to see the pipeline status but not modify the pipeline. You can also set permissions for any stage or action within a pipeline. For more information on using IAM with AWS CodePipeline, see Access Permissions Reference.

Q: Can I enable the pipeline in one AWS account to be accessed by an IAM user in another AWS account?
Yes. You can create an IAM role in the AWS account that owns the pipeline to delegate access to the pipeline and any related resources to an IAM user in another account. For a walkthrough on enabling such a cross account access, see Walkthrough: Delegating Access Across AWS Accounts For Accounts You Own Using IAM Roles and Configure Cross-Account Access to a Pipeline.

Regions

Q: Which regions does AWS CodePipeline support?
Please refer to Regional Products and Services for details of CodePipeline availability by region.

Billing

Q: How much does AWS CodePipeline cost?
For details on AWS CodePipeline cost, see thepricing page.

AWS CodeDeploy FAQ

General

Q: What is AWS CodeDeploy?
AWS CodeDeploy is a service that automates code deployments to any instance, including Amazon EC2 instances and instances running on-premises. AWS CodeDeploy makes it easier for you to rapidly release new features, helps you avoid downtime during deployment, and handles the complexity of updating your applications. You can use AWS CodeDeploy to automate deployments, eliminating the need for error-prone manual operations, and the service scales with your infrastructure so you can easily deploy to one instance or thousands.

Q: Who should use AWS CodeDeploy?
AWS CodeDeploy is designed for developers and administrators who need to deploy applications to any instance, including Amazon EC2 instances and instances running on-premises. It is flexible and can also be used by anyone wanting to update software or run scripts on their instances.

Q: What types of applications can be deployed with AWS CodeDeploy?
AWS CodeDeploy can be used for deploying any type of application. To use AWS CodeDeploy, you specify the files to copy and the scripts to run on each instance during the deployment. AWS CodeDeploy is programming language and architecture agnostic, so you can use scripts for any custom deployment logic.

Q: What operating systems does AWS CodeDeploy support?
AWS CodeDeploy supports a wide variety of operating systems. AWS CodeDeploy provides agents that have been tested on Amazon Linux, Red Hat Enterprise Linux, Ubuntu Server, and Microsoft Windows Server. If you want to use other operating systems, the AWS CodeDeploy agent is available as open source software here. For more information on operating system support, see AWS CodeDeploy Documentation.

Q:Will AWS CodeDeploy work with my existing tool chain?
Yes. AWS CodeDeploy works with a variety of configuration management systems, continuous integration and deployment systems, and source control systems. For more information, see product integrations page.

Q: How is AWS CodeDeploy different from other AWS deployment and management services such as AWS Elastic Beanstalk and AWS OpsWorks?
AWS CodeDeploy is a building block service focused on helping developers deploy and update software on any instance, including Amazon EC2 instances and instances running on-premises. AWS Elastic Beanstalk and AWS OpsWorks are end-to-end application management solutions.

Q: Does AWS CodeDeploy support on-premises instances?
Yes. AWS CodeDeploy supports any instance that can install the CodeDeploy agent and

connect to AWS public endpoints.

Concepts

Q: What is an application?
An application is a collection of software and configuration to be deployed to a group of instances. Typically, the instances in the group run the same software. For example, if you have a large distributed system, the web tier will likely constitute one application and the data tier another application.

Q: What is a revision?
A revision is a specific version of deployable content, such as source code, post-build artifacts, web pages, executable files, and deployment scripts, along with an AppSpec file. The AWS CodeDeploy Agent can access a revision from GitHub or an Amazon S3 bucket.

Q: What is a deployment group?
A deployment group is a set of instances associated with an application that you target for a deployment. You can add instances to a deployment group by specifying a tag, an Auto Scaling group name, or both. You can define multiple deployment groups for an application such as staging and production. For information on tags, see Working with Amazon EC2 Tags in the Console. For more information on deploying to Auto Scaling groups, see Auto Scaling Integration.

Q: What is a deployment configuration?
A deployment configuration is a constraint that determines how a deployment progresses through the instances in a deployment group. You can use a deployment configuration to perform zero-downtime deployments to multi-instance deployment groups. For example, if your application needs at least 50% of the instances in a deployment group to be up and serving traffic, you can specify that in your deployment configuration so that a deployment does not cause downtime. If no deployment configuration is associated with either the deployment or the deployment group, then by default AWS CodeDeploy will deploy to one instance at a time. For more information on deployment configuration, see Instance Health.

Q: What are the parameters that I need to specify for a deployment?
There are three parameters you specify for a deployment:

1. Revision - Specifies what to deploy.
2. Deployment group - Specifies where to deploy.
3. Deployment configuration - An optional parameter that specifies how to deploy.

Q: What is an AppSpec file?

An AppSpec file is a configuration file that specifies the files to be copied and scripts to be executed. The AppSpec file uses the YAML format, and you include it in the root directory of your revision. The AppSpec file is used by the AWS CodeDeploy Agent and consists of two sections. The files section specifies the source files in your revision to be copied and the destination folder on each instance. The hooks section specifies the location (as relative paths starting from the root of the revision bundle) of the scripts to run during each phase of the deployment. Each phase of a deployment is called a deployment lifecycle event. The following is a sample AppSpec file. For more information on an AppSpec file, including all the options that can be specified, see AppSpec File Reference.

```
version: 0.0
os: linux
files:
# You can specify one or more mappings in the files section.
  - source: /
    destination: /var/www/html/WordPress
hooks:
 # The lifecycle hooks sections allows you to specify deployment scripts.
ApplicationStop:
# Step 1: Stop Apache and MySQL if running.
    - location: helper_scripts/stop_server.sh
BeforeInstall:
# Step 2: Install Apache and MySQL.
# You can specify one or more scripts per deployment lifecycle event.
    - location: deploy_hooks/puppet-apply-apache.sh
    - location: deploy_hooks/puppet-apply-mysql.sh
 AfterInstall:
# Step 3: Set permissions.
    - location: deploy_hooks /change_permissions.sh
      timeout: 30
      runas: root
# Step 4: Start the server.
    - location: helper_scripts/start_server.sh
      timeout: 30
      runas: root
```

Q: What are deployment lifecycle events?
A deployment goes through a set of predefined phases called deployment lifecycle events. A deployment lifecycle event gives you an opportunity to run code as part of the deployment. The following table lists the different deployment lifecycle events currently supported, in their order of execution, along with examples of when you may want to use them.

Deployment Lifecycle Event	Description
ApplicationStop	This is the first deployment lifecycle event that occurs even before the revision gets downloaded. The AppSpec file and scripts used for this deployment lifecycle event are from the last successfully deployed revision. You can use the ApplicationStop deployment lifecycle event if you want to gracefully stop the application or remove currently installed packages in preparation of a deployment.
DownloadBundle	During this deployment lifecycle event, the agent copies the revision files to a temporary location on the instance. This deployment lifecycle event is reserved for the agent and cannot be used to run user scripts.
BeforeInstall	You can use the BeforeInstall deployment lifecycle event for preinstall tasks such as decrypting files and creating a backup of the current version.
Install	During this deployment lifecycle event, the agent copies the revision files from the temporary location to the final destination folder. This deployment lifecycle event is reserved for the agent and cannot be used to run user scripts.
AfterInstall	You can use the AfterInstall deployment lifecycle event for tasks such as configuring your application or changing file permissions.
ApplicationStart	You typically use the ApplicationStart deployment lifecycle event to restart services that were stopped during ApplicationStop.
ValidateService	ValidateService is the last deployment lifecycle event and is an opportunity to verify that the deployment completed successfully.

Getting Started

Q: How do I get started with AWS CodeDeploy?
You can sign in to the AWS Management Console and start using AWS CodeDeploy. If you are looking for a quick overview of the service, see Getting Started, which includes a step-by-step

tutorial.

Using AWS CodeDeploy

Q: Are there any prerequisites for using an existing Amazon EC2 instance with AWS CodeDeploy?
The Amazon EC2 instance must be associated with an IAM instance profile and should be running a supported operating system. For more information, see Use an Existing Amazon EC2 Instance.

Q: What are the typical steps to go through for deploying an application using AWS CodeDeploy?
The following diagram shows the typical steps during a deployment. Creating an application and deployment group (see the Concepts section for an explanation of these terms) are typically one-time setup tasks per application. The recurring actions are uploading a revision and deploying it. For a detailed explanation, including step-by-step instructions for each of these tasks, see Deployments.

Q: How can I access AWS CodeDeploy?
You can access AWS CodeDeploy using the AWS Management Console, the AWS Command Line Interface (AWS CLI), the AWS SDKs, and the AWS CodeDeploy APIs.

Q: What changes do I need to make to my code to deploy using AWS CodeDeploy?
You don’t need to make any changes to your code. You simply add a configuration file (called an AppSpec file) in the root directory of your revision bundle that specifies the files to be copied and scripts to be executed.

Q: How can I deploy an application from my source control system using AWS CodeDeploy?
If you are using GitHub, you can deploy a revision in a .zip, .tar, or .tar.gz format from your repository directly to instances. For other source control systems, you can bundle and upload the revision to an Amazon S3 bucket in a .zip, .tar, or .tar.gz format and specify the Amazon S3 location when doing a deployment. If your application needs a build step, make sure that the GitHub repository or the Amazon S3 bucket contains the post-build artifacts. For more information on using GitHub with AWS CodeDeploy, see our product integrations page. For more information on using Amazon S3 for storing revisions, see Push a Revision.

Q: How will AWS CodeDeploy work with my configuration management tool?
You can invoke your configuration management tool from any deployment lifecycle event hook in the AppSpec file. For example, if you have a Chef recipe that you want to run as part of a deployment, you can do so by specifying it in the appropriate deployment lifecycle event hook in the AppSpec file. In addition, you can leverage your configuration management system to install the AWS CodeDeploy agent on instances. For samples that illustrate using AWS CodeDeploy

with configuration management systems such as Chef, Puppet, Ansible, and Saltstack, see our product integrations page.

Q: Can I use AWS CodeDeploy with continuous integration and deployment systems?
Yes. You can integrate AWS CodeDeploy with your continuous integration and deployment systems by calling the public APIs using the AWS CLI or AWS SDKs. You can find prebuilt integrations and samples on our product integrations page.

Q: How do I get my application on the instances that I just added to the deployment group?
Deploy the latest revision to the deployment group for the newly added instances to get your application. Except for Amazon EC2 instances that are launched as part of an Auto Scaling group, AWS CodeDeploy doesn’t automatically deploy the latest revision to newly added instances.

Q: How does AWS CodeDeploy work with Auto Scaling?
You can associate an Auto Scaling group with a deployment group to make sure that newly launched instances always get the latest version of your application. Every time a new Amazon EC2 instance is launched for that Auto Scaling group, it will be first put in a Pending state and a deployment of the last successful revision for that deployment group triggered on that Amazon EC2 instance. If the deployment completes successfully, the state of the Amazon EC2 instance is changed to InService. If that deployment fails, the Amazon EC2 instance is terminated, a new Amazon EC2 instance is launched in Pending state, and a deployment triggered for the newly launched EC2 instance. For more information on Auto Scaling group instance lifecycle events, see Auto Scaling Group Lifecycle.

Q: How do I track the status of a deployment?
You can track the status of a deployment using the AWS Management Console, the AWS Command Line Interface (AWS CLI), the AWS SDKs, and the AWS CodeDeploy APIs.You can see the overall status of a deployment and drill down further to see the status of each instance and the status of each deployment lifecycle event for the instance. You can also see the log entries corresponding to any failure, making it easy to debug deployment issues without having to log into the instance.

Q: Can I stop an in-flight deployment?
Yes. When you stop an in-flight deployment, the AWS CodeDeploy service will instruct the agent on each instance to stop executing additional scripts. To get your application back to a consistent state, you can either redeploy the revision, or deploy another revision.

Q: How do I roll back an application to the previous revision?
To roll back an application to a previous revision, you just need to deploy that revision. AWS CodeDeploy keeps track of the files that were copied for the current revision and removes them before starting a new deployment, so there is no difference between redeploy and roll back. However, you need to make sure that the previous revisions are available for roll back.

Q: Can I use a versioned Amazon S3 bucket to store revisions?
Yes. You can use a versioned Amazon S3 bucket and specify the version ID to uniquely identify a revision.

Q: What are the service limits when using AWS CodeDeploy?
For information on the service limits, see Limits. To increase your service limits, submit a request through the AWS Support Center.

Q: Can I get a history of AWS CodeDeploy API calls made on my account for security analysis and operational troubleshooting purposes?
Yes. To receive a history of AWS CodeDeploy API calls made on your account, you simply turn on AWS CloudTrail in the AWS Management Console.

Security

Q: Can I use AWS CodeDeploy to deploy an application to Amazon EC2 instances running within an Amazon Virtual Private Cloud (VPC)?
Yes, but the AWS CodeDeploy agent installed on the Amazon EC2 instances must be able to access the public AWS CodeDeploy and Amazon S3 service endpoints. For more information, see AWS CodeDeploy Endpoints and Amazon S3 Endpoints.

Q: Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodeDeploy?
Yes. AWS CodeDeploy supports resource-level permissions. For each AWS CodeDeploy resource, you can specify which user has access and to which actions. For example, you can set an IAM policy to let a user deploy a particular application but only list revisions for other applications. You can therefore prevent users from inadvertently making changes to the wrong application. For more information on using IAM with AWS CodeDeploy, see Access Permissions Reference.

Regions

Q: Which regions does AWS CodeDeploy support?
Please refer to Regional Products and Services for details of CodeDeploy availability by region.

Q: How do I deploy an AWS CodeDeploy application to multiple regions?
AWS CodeDeploy performs deployments with AWS resources located in the same region. To deploy an application to multiple regions, define the application in your target regions, copy the application bundle to an Amazon S3 bucket in each region, and then start the deployments using either a serial or parallel rollout across the regions.

Billing

Q: How much does AWS CodeDeploy cost?
There is no additional charge for code deployments to Amazon EC2 instances through AWS CodeDeploy. You pay $0.02 per on-premises instance update using AWS CodeDeploy. Please see the Pricing page for more details.

AWS CodeCommit FAQ

General

Q: What is AWS CodeCommit?
AWS CodeCommit is a secure, highly scalable, managed source control service that hosts private Git repositories. AWS CodeCommit eliminates the need for you to operate your own source control system or worry about scaling its infrastructure. You can use AWS CodeCommit to store anything from code to binaries, and it works seamlessly with your existing Git tools.

Q: What is Git?
Git is an open-source distributed version control system. To work with AWS CodeCommit repositories, you use the Git command line interface (CLI) or any of the available Git clients. To learn more about Git, see the Git documentation. To learn more about using AWS CodeCommit with Git, see Getting Started with AWS CodeCommit.

Q: Who should use AWS CodeCommit?
AWS CodeCommit is designed for software developers who need a secure, reliable, and scalable source control system to store and version their code. In addition, AWS CodeCommit can be used by anyone looking for an easy to use, fully managed data store that is version controlled. For example, IT administrators can use AWS CodeCommit to store their scripts and configurations. Web designers can use AWS CodeCommit to store HTML pages and images.

Q: How is AWS CodeCommit different from other Git-based source control systems?
AWS CodeCommit offers a number of features not offered by other Git source control systems:

- Fully Managed –AWS CodeCommit eliminates the need to host, maintain, backup, and scale your own source control servers.

- Secure –AWS CodeCommit automatically encrypts your files in transit and at rest. AWS CodeCommit is integrated with AWS Identity and Access Management (IAM), allowing you to assign user-specific permissions to your repositories.

- Highly Available – AWS CodeCommit is built on highly scalable, redundant, and durable AWS services such as Amazon S3 and Amazon DynamoDB.

- Scalable - AWS CodeCommit allows you store any number of files and there are no repository size limits.

- Faster Development Lifecycle - AWS CodeCommit keeps your repositories close to your build, staging, and production environments in the AWS cloud. This allows you to increase the speed and frequency of your development lifecycle.

Q: How does AWS CodeCommit compare to a versioned S3 bucket?
AWS CodeCommit is designed for collaborative software development. It manages batches of changes across multiple files, offers parallel branching, and includes version differencing (“diffing”). In comparison, Amazon S3 versioning supports recovering past versions of individual files but doesn’t support tracking batched changes that span multiple files or other features needed for collaborative software development.

Using AWS CodeCommit

Q: How do I get started with AWS CodeCommit?
You can sign in to the AWS Management Console, create a repository, and start working with the repository using Git. If you want an introduction to the service, see Getting Started, which includes a step-by-step tutorial.

Q: How do I create a repository?
You can create a repository from the AWS Management Console or by using the AWS Command Line Interface (AWS CLI), the AWS SDKs, or the AWS CodeCommit APIs.

Q: How do I update files in my repository?
You use Git to work with the repository. For example, you can use the *git clone* command to make a local copy of the AWS CodeCommit repository. Make changes to the local files and use the git commit command when you’re ready to save the changes. Finally, use the *git push* command to upload the changes to the AWS CodeCommit repository. For step-by-step instructions, see Getting Started with AWS CodeCommit.

Q: How do I import my existing repository to AWS CodeCommit?
You can use Git to import any existing Git repository to AWS CodeCommit. For other repositories, such as Subversion and Perforce, you can use a Git importer to first migrate it to a Git repository. For step by step instructions on importing Git repositories, see Migrate an Existing Repository to AWS CodeCommit. For instructions on migrating other repositories to Git, see the Git migration documentation.

Q: What Git operations are currently supported by AWS CodeCommit?
AWS CodeCommit currently supports clone, pull, push and fetch commands.

Q: Does AWS CodeCommit support Git submodules?
Yes. AWS CodeCommit can be used with Git repositories that include submodules.

Q: What are the service limits when using AWS CodeCommit?
For information on the service limits, see Limits.

Q: What is the maximum size for a single file that I can store in CodeCommit?
A single file in a repository cannot be more than 2 GB in size.

Q: How do I backup my repository?
If you have a local copy of the repository from doing a full *git clone*, you can use that to restore data. If you want additional backups, there are multiple ways to do so. One way is to install Git on your backup server and run a scheduled job that uses the *git clone* command to take regular snapshots of your repository. You can use *git pull* instead of *git clone* if you want to copy only the incremental changes. Note that these operations may incur an additional user and/or request charges based on how you setup the backup server and the polling frequency.

Q: How do I restore a deleted AWS CodeCommit repository?
Deleting an AWS CodeCommit repository is a destructive one-way operation that cannot be undone. To restore a deleted repository, you will need to create the repository again and use either a backup or a local copy from a full clone to upload the data. We recommend using IAM policies along with MFA-protection to restrict users who can delete repositories. For more details, see the Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodeCommit? question in the Security section of the FAQ.

Q: How do I manage code reviews with AWS CodeCommit?
For code reviews, you can use any Git-compatible code review system like Review Board.

Q: How do I integrate my continuous integration system with AWS CodeCommit?
Continuous Integration (CI) systems can be configured to use Git to pull code from AWS CodeCommit. For examples on using CI systems with AWS CodeCommit, see our blog post on integrating AWS CodeCommit with Jenkins.

Q: How do I create webhooks using AWS CodeCommit?
In the Amazon Simple Notification Service (SNS) console, you can create a SNS topic with an HTTP endpoint and the desired URL for the webhook. From the AWS CodeCommit console, you can then configure that SNS topic to a repository event using triggers.

Security

Q: Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodeCommit?
Yes. AWS CodeCommit supports resource-level permissions. For each AWS CodeCommit repository, you can specify which users can perform which actions. You can also specify AWS multi-factor authentication (MFA) for a CodeCommit action. This allows you to add an extra level of protection for destructive actions such as deleting repositories. In addition to the AWS

CodeCommit APIs, you can also specify git pull and git push as actions to control access from Git clients. For example, you can create a read-only user for a repository by allowing that user access to git pull but not git push on the repository. For more information on using IAM with AWS CodeCommit, see Access Permissions Reference. For more information on authenticating API access using MFA, see Configuring MFA-Protected API Access.

Q: What communication protocols are supported by AWS CodeCommit?
You can use either the HTTPS or SSH protocols or both to communicate with AWS CodeCommit. To use HTTPS, first install the AWS CLI. The AWS CLI installs a Git credential helper that can be configured with AWS credentials. It automatically signs all HTTPS requests to AWS CodeCommit using the Signature Version 4 signing specification. To use SSH, users create their own public-private key pairs and add their public keys to their IAM users. The private key encrypts the communication with AWS CodeCommit. For step-by-step instructions on setting up HTTPS and SSH access, see the Setting up AWS CodeCommit page.

Q: What ports should I open in my firewall for access to AWS CodeCommit?
You will have to open outbound access to an AWS CodeCommit service endpoint on port 22 (SSH) or port 443 (HTTPS).

Q: How do I encrypt my repository in AWS CodeCommit?
Repositories are automatically encrypted at rest. No customer action is required. AWS CodeCommit uses AWS Key Management Service (KMS) to encrypt repositories. When you create your first repository, an AWS-managed CodeCommit key is created under your AWS account. For details, see Encryption for AWS CodeCommit Repositories.

Q: Can I enable cross-account access to my repository?
Yes. You can create an IAM role in your AWS account to delegate access to a repository to IAM users in other AWS accounts. The IAM users can then configure their AWS CLI to use AWS Security Token Service (STS) and assume the role when running commands. For details see Assuming a Role in the AWS CLI documentation.

Regions

Q: Which regions does AWS CodeCommit support?
Please refer to Regional Products and Services for details of CodeCommit availability by region.

Billing

Q: How much does AWS CodeCommit cost?
AWS CodeCommit costs $1 per active user per month. For every active user, your account receives an additional allowance of 10 GB-month of storage and 2,000 Git requests for that month. Unused allowance for storage and Git requests does not carry over to later months. If

you need more storage or Git requests for your users, additional usage will be charged at $0.06 per GB-month and $0.001 per Git request. Users may store as many Git repositories as they would like. Your usage is calculated each month across all regions and automatically applied to your bill. Please see the pricing page for more details.

Q: What is the definition of an active user in AWS CodeCommit?
An active user is any unique AWS identity (IAM user/role, federated user, or root account) that accesses AWS CodeCommit repositories during the month, either through Git requests or by using the AWS Management Console. A server accessing CodeCommit using a unique AWS identity counts as an active user.

Q: Which Git requests are considered towards the monthly allowance?
A Git request includes any push or pull that transmits repository objects. The request does not count towards your Git request allowance if there is no object transfer due to local and remote branches being up-to-date.

Amazon CloudWatch FAQ

General

Q: What is Amazon CloudWatch?

Amazon CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS. You can use Amazon CloudWatch to collect and track metrics, collect and monitor log files, and set alarms. Amazon CloudWatch can monitor AWS resources such as Amazon EC2 instances, Amazon DynamoDB tables, and Amazon RDS DB instances, as well as custom metrics generated by your applications and services, and any log files your applications generate. You can use Amazon CloudWatch to gain system-wide visibility into resource utilization, application performance, and operational health. You can use these insights to react and keep your application running smoothly.

Q: What can I use to access CloudWatch?

Amazon CloudWatch can be accessed via API, command-line interface, AWS SDKs, and the AWS Management Console.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: What access management policies can I implement for CloudWatch?

Amazon CloudWatch integrates with AWS Identity and Access Management (IAM) so that you can specify which CloudWatch actions a user in your AWS Account can perform. For example, you could create an IAM policy that gives only certain users in your organization permission to use GetMetricStatistics. They could then use the action to retrieve data about your cloud resources.

You can't use IAM to control access to CloudWatch data for specific resources. For example, you can't give a user access to CloudWatch data for only a specific set of instances or a specific LoadBalancer. Permissions granted using IAM cover all the cloud resources you use with CloudWatch. In addition, you can't use IAM roles with the Amazon CloudWatch command line tools.

Q: What is Amazon CloudWatch Logs?

Amazon CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.

With CloudWatch Logs, you can monitor your logs, in near real time, for specific phrases, values or patterns. For example, you could set an alarm on the number of errors that occur in your system logs or view graphs of latency of web requests from your application logs. You can then view the original log data to see the source of the problem. Log data can be stored and accessed indefinitely in highly durable, low-cost storage so you don’t have to worry about filling up hard drives.

Q: What kinds of things can I do with CloudWatch Logs?

CloudWatch Logs is capable of monitoring and storing your logs to help you better understand and operate your systems and applications. You can use CloudWatch Logs in a number of ways. When you use CloudWatch logs, your existing log data is used for monitoring, so no code changes are required. Real time Application and System Monitoring. You can use CloudWatch Logs to monitor applications and systems using log data. For example, CloudWatch Logs can track the number of errors that occur in your application logs and send you a notification whenever the rate of errors exceeds a threshold you specify. CloudWatch Logs uses your log data for monitoring; so, no code changes are required.Long Term Log RetentionYou can use CloudWatch Logs to store your log data indefinitely in highly durable and cost effective storage without worrying about hard drives running out of space. The CloudWatch Logs Agent makes it easy to quickly move both rotated and non rotated log files off of a host and into the log service. You can then access the raw log event data when you need it.

Q: What platforms does the CloudWatch Logs Agent support?

The CloudWatch Logs Agent is supported on Amazon Linux, Ubuntu, CentOS, Red Hat Enterprise Linux, and Windows. This agent will support the ability to monitor individual log files on the host.

Q: Does the CloudWatch Logs Agent support IAM roles?

Yes. The CloudWatch Logs Agent is integrated with Identity and Access Management (IAM) and includes support for both access keys and IAM roles.

Pricing

Q: How much does Amazon CloudWatch cost?

Please see our pricing page for the latest information.

Q: Does the Amazon CloudWatch monitoring charge change depending on which type of Amazon EC2 instance I monitor?

No, the Amazon CloudWatch monitoring charge does not vary by Amazon EC2 instance type.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. Learn more.

AWS Resource and Custom Metrics Monitoring

Q: What can I measure with Amazon CloudWatch Metrics?

Amazon CloudWatch allows you to monitor AWS cloud resources and the applications you run on AWS. Metrics are provided automatically for a number of AWS products and services, including Amazon EC2 instances, EBS volumes, Elastic Load Balancers, Auto Scaling groups, EMR job flows, RDS DB instances, DynamoDB tables, ElastiCache clusters, RedShift clusters, OpsWorks stacks, Route 53 health checks, SNS topics, SQS queues, SWF workflows, and Storage Gateways. You can also monitor custom metrics generated by your own applications and services.

Q: What is the retention period of all metrics?

CloudWatch launched extended retention of metrics in November 1, 2016. The feature enabled storage of all metrics for customers from the previous 14 days to 15 months. CloudWatch Metrics now supports the following three retention schedules:

- 1 minute datapoints are available for 15 days
- 5 minute datapoints are available for 63 days
- 1 hour datapoints are available for 455 days

This means that 1 minute datapoints expire after 15 days, 5 minute data points expire after 63 days, and 1 hour datapoints expire after 455 days. If you need availability of metrics longer than those periods, you can use the GetMetricStatistics API to retrieve the datapoints onto a different storage.

The feature is currently available in US East (N. Virginia), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), S. America (São Paulo), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Seoul), Asia Pacific (Mumbai) and Asia Pacific (Sydney) and will be rolled out to all regions subsequently.

Q: What is the minimum granularity for the data that Amazon CloudWatch receives and aggregates?

The minimum granularity supported by CloudWatch is 1 minute data points. Many metrics are received and aggregated at 1-minute intervals. Some are received at 3-minute or 5-minute intervals.

Depending on the age of data requested, metrics will be available in the granularity defined in the retention schedules defined above. For example, if you request for 1 minute data for a day from 10 days ago, you will receive the 1440 data points. However, if you request for 1 minute data from 5 months back, the UI will automatically change the granularity to 1 hour and the GetMetricStatistics API will not return any output.

Q: Can I delete any metrics?

CloudWatch does not support metric deletion. Metrics expire based on the retention schedules described above.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

No. You can always retrieve metrics data for any Amazon EC2 instance based on the retention schedules described above. However, the CloudWatch console limits the search of metrics to 2 weeks after a metric is last ingested to ensure that the most up to date instances are shown in your namespace.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?
Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 15 months.

Q: Why does the graphing of the same time window look different when I view the metrics in 5 minute and 1 minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period, the single data point is placed at the 1 minute mark. We recommend using a 1 minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Q: What is a Custom Metric?

You can use Amazon CloudWatch to monitor data produced by your own applications, scripts, and services. A custom metric is any metric you provide to Amazon CloudWatch. For example, you can use custom metrics as a way to monitor the time to load a web page, request error rates, number of processes or threads on your instance, or amount of work performed by your application. You can get started with custom metrics by using the PutMetricData API, our sample monitoring scripts for Windows and Linux, CloudWatch collectd plugin, as well as a number of applications and tools offered by AWS partners.

Q: What granularity can I get from a Custom Metric?

The minimum granularity supported by CloudWatch for Custom Metrics is 1 minute data points. Many metrics are received and aggregated at 1-minute intervals. You can send custom metrics to Amazon CloudWatch as frequently as you like, but statistics will be available at according to the retention schedule described above.

Q: When would I use a Custom Metric over having my program emit a log to CloudWatch Logs?

You can monitor your own data using custom metrics, CloudWatch Logs, or both. You may want to use custom metrics if your data is not already produced in log format, for example operating system processes or performance measurements. Or, you may want to write your own application or script, or one provided by an AWS partner. If you want to store and save individual measurements along with additional detail, you may want to use CloudWatch Logs.

Q: What statistics can I view and graph in CloudWatch?

You can retrieve, graph, and set alarms on the following statistical values for Amazon CloudWatch metrics: Average, Sum, Minimum, Maximum, and Sample Count. Statistics can be computed for any time period between 60 seconds and one day.

Log Monitoring

Q: What log monitoring does Amazon CloudWatch provide?

CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.

With CloudWatch Logs, you can monitor your logs, in near real time, for specific phrases, values or patterns. For example, you could set an alarm on the number of errors that occur in your system logs or view graphs of latency of web requests from your application logs. You can then view the original log data to see the source of the problem. Log data can be stored and accessed for up to as long as you need in highly durable, low-cost storage so you don't have to worry about filling up hard drives.

Q: Is CloudWatch Logs available in all regions?

Please refer to Regional Products and Services for details of CloudWatch Logs service availability by region.

Q: How much does CloudWatch Logs cost?

Please see our pricing page for the latest information.

Q: What kinds of things can I do with my logs and Amazon CloudWatch?

CloudWatch Logs is capable of monitoring and storing your logs to help you better understand and operate your systems and applications. When you use CloudWatch Logs with your logs, your existing log data is used for monitoring, so no code change are required. Here are a two examples of what you can do with Amazon CloudWatch and your logs:

Real time Application and System Monitoring: You can use CloudWatch Logs to monitor applications and systems using log data in near real time. For example, CloudWatch Logs can track the number of errors that occur in your application logs and send you a notification whenever the rate of errors exceeds a threshold you specify. Amazon CloudWatch uses your log data for monitoring and consequently it doesn't involve any code changes from you.

Long Term Log Retention: You can use CloudWatch Logs to store your log data for as long as you need in highly durable and cost effective storage without worrying about hard drives running out of space. The CloudWatch Logs Agent makes it easy to quickly move both rotated and non rotated log files off of a host and into the log service. You can then access the raw log event data when you need it.

Q: What types of data can I send to Amazon CloudWatch Logs from my EC2 instances running Microsoft SQL Server and Microsoft Windows Server?

You can configure the EC2Config service to send a variety of data and log files to CloudWatch including: custom text logs, Event (Application, Custom, Security, System) logs, Event Tracing (ETW) logs, and Performance Counter (PCW) data. Learn more about the EC2Config service here.

Q: How frequently does the CloudWatch Logs Agent send data?

The CloudWatch Logs Agent will send log data every five seconds by default and is configurable by the user.

Q: What log formats does CloudWatch Logs support?

CloudWatch Logs can ingest, aggregate and monitor any text based common log data or JSON-formatted logs.

Q: What if I configure the CloudWatch Logs Agent to send non-text log data?

The CloudWatch Logs Agent will record an error in the event it has been configured to report non text log data. This error is recorded in the /var/logs/awslogs.log.

Q: How do I start monitoring my logs with CloudWatch Logs?

You can monitor log events as they are sent to CloudWatch Logs by creating Metric Filters. Metric Filters turn log data into Amazon CloudWatch Metrics for graphing or alarming. Metric Filters can be created in the Console or the CLI. Metric Filters search for and match terms, phrases or values in your log events. When a Metric Filter finds one of the terms, phrases or values in your log events, it counts it in an Amazon CloudWatch Metric that you choose. For example, you can create a Metric Filter to search for and count the occurrence of the word “Error” in your log events. Metric Filters can also extract values from space delimited log events, such as the latency of web requests. You can also use conditional operators and wildcards to create exact matches. The Amazon CloudWatch Console can help you test your patterns before creating Metric Filters.

Q: What is the syntax of Metric Filter patterns?

A Metric Filter pattern can contain search terms or a specification of your common log or JSON event format.

For example, if you want to search for the term Error, the pattern for the metric filter would just be the term Error. Multiple search terms can be included to search for multiple terms. For example, if you wanted to count events which contained the terms Error and Exception you would use the pattern Error Exception. If you wanted to match the term Error Exception exactly, you would put double quotes around the search term, "Error Exception". You can specify as many search terms as you like.

CloudWatch Logs can also be used to extract values from a log event in common log or JSON format. For example, you could track the bytes transferred from your Apache access logs. You can also use conditional operators and wildcards to match and extract the data you are interested in. To use the extraction feature of Metric Filters, log events must be space delimited and use a starting and ending double quote """, or, a starting square brace "[" and a closing square brace "]"square, to enclose fields. Alternatively, they can be JSON-formatted log events. For the full details of the syntax and examples, please see the Developer Guide for Metric Filters.

Q: How do I know that a Metric Filter pattern I specified will match my log events?

CloudWatch Logs lets you test the Metric Filter patterns you want before you create a Metric Filter. You can test your patterns against your own log data that is already in CloudWatch Logs or you can supply your own log events to test. Testing your pattern will show you which log events matched the Metric Filter pattern and, if extracting values, what the extracted value is in the test data. Metric Filter testing is available for use in the console and the CLI.

Q: Can I use regular expressions with my log data?

Amazon CloudWatch Metric Filters does not support regular expressions. To process your log data with regular expressions, consider using Amazon Kinesis and connect the stream with a regular expression processing engine.

Log Management

Q: How do I retrieve my log data?

You can retrieve any of your log data using the CloudWatch Logs console or through the CloudWatch Logs CLI. Log events are retrieved based on the Log Group, Log Stream and time with which they are associated. The CloudWatch Logs API for retrieving log events is GetLogEvents.

Q: How do I search my logs?

You can use the CLI to retrieve your log events and search through them using command line grep or similar search functions.

Q: How long does CloudWatch Logs store my log data?

You can store your log data in CloudWatch Logs for as long as you want. By default, CloudWatch Logs will store your log data indefinitely. You can change the retention for each Log Group at any time.

Alarms

Q: What types of CloudWatch Alarms can be created?

You can create an alarm to monitor any Amazon CloudWatch metric in your account. For example, you can create alarms on an Amazon EC2 instance CPU utilization, Amazon ELB request latency, Amazon DynamoDB table throughput, Amazon SQS queue length, or even the charges on your AWS bill.

Q: What actions can I take from a CloudWatch Alarm?

When you create an alarm, you can configure it to perform one or more automated actions when the metric you chose to monitor exceeds a threshold you define. For example, you can set an alarm that sends you an email, publishes to an SQS queue, stops or terminates an Amazon EC2 instance, or executes an Auto Scaling policy. Since Amazon CloudWatch alarms are integrated with Amazon Simple Notification Service, you can also use any notification type supported by SNS.

Q: What thresholds can I set to trigger a CloudWatch Alarm?

When you create an alarm, you first choose the Amazon CloudWatch metric you want it to monitor. Next, you choose the evaluation period (e.g., five minutes or one hour) and a statistical value to measure (e.g., Average or Maximum). To set a threshold, set a target value and choose whether the alarm will trigger when the value is greater than (>), greater than or equal to (>=), less than (<), or less than or equal to (<=) that value.

Q: My CloudWatch Alarm is constantly in the Alarm state, what did I do wrong?

Alarms continue to evaluate metrics against your chosen threshold, even after they have already triggered. This allows you to view its current up-to-date state at any time. You may notice that one of your alarms stays in the ALARM state for a long time. If your metric value is still in breach of your threshold, the alarm will remain in the ALARM state until it no longer breaches the threshold. This is normal behavior. If you want your alarm to treat this new level as OK, you can adjust the alarm threshold accordingly.

Q: How long can I view my Alarm history?

Alarm history is available for 14 days. To view your alarm history, log in to CloudWatch in the AWS Management Console, choose Alarms from the menu at left, select your alarm, and click the History tab in the lower panel. There you will find a history of any state changes to the alarm as well as any modifications to the alarm configuration.

Dashboards

Q: What is CloudWatch Dashboards?

Amazon CloudWatch Dashboards allow you to create, customize, interact with, and save graphs of AWS resources and custom metrics.

Q: What can I do with CloudWatch dashboards?

You can use CloudWatch Dashboards to monitor your applications and resources to quickly identify issues that might be impacting the health of your applications. You can save and revisit dashboards, add multiple graphs, or add text widgets into a dashboard to embed links and comments. For example, you can include graphs of your resource and application metrics to see when resource health problems might be impacting your applications. You can also view metrics from multiple regions on the same page.

Q: How do I get started with CloudWatch Dashboards?

To get started, visit the Amazon CloudWatch Console and select “Dashboards”. Click the “Create Dashboard” button.

Q: Do the dashboards support auto refresh?

Yes. Dashboards will auto refresh while you have them open.

Q: Can I share my dashboard?

Yes, a dashboard is available to anyone with the correct permissions for the account with the dashboard.

Events

Q: What is CloudWatch Events?

Amazon CloudWatch Events (CWE) is a stream of system events describing changes in your AWS resources. The events stream augments the existing CloudWatch Metrics and Logs streams to provide a more complete picture of the health and state of your applications. You write declarative rules to associate events of interest with automated actions to be taken.

Q: What services emit CloudWatch Events?

Currently, Amazon EC2, Auto Scaling, and AWS CloudTrail are supported. Via AWS CloudTrail, mutating API calls (i.e., all calls except Describe*, List*, and Get*) across all services are visible in CloudWatch Events.

Q: What can I do once an event is received?

When an event matches a rule you've created in the system, you can automatically invoke an AWS Lambda function, relay the event to an Amazon Kinesis stream, notify an Amazon SNS topic, or invoke a built-in workflow.

Q: Can I generate my own events?

Yes. Your applications can emit custom events by using the PutEvents API, with a payload uniquely suited to your needs.

Q: Can I do things on a fixed schedule?

CloudWatch Events is able to generate events on a schedule you set by using the popular Unix cron syntax. By monitoring for these events, you can implement a scheduled application.

Q: What is the difference between CloudWatch Events and AWS CloudTrail?

CloudWatch Events is a near real time stream of system events that describe changes to your AWS resources. With CloudWatch Events, you can define rules to monitor for specific events and perform actions in an automated manner. AWS CloudTrail is a service that records API calls for your AWS account and delivers log files containing API calls to your Amazon S3 bucket or a CloudWatch Logs log group. With AWS CloudTrail, you can look up API activity history related to creation, deletion and modification of AWS resources and troubleshoot operational or security issues.

Q: What is the difference between CloudWatch Events and AWS Config?

AWS Config is a fully managed service that provides you with an AWS resource inventory, configuration history, and configuration change notifications to enable security and governance. Config rules help you determine whether configuration changes are compliant. CloudWatch Events is for reacting in near real time to resource state changes. It doesn’t render a verdict on whether the changes comply with policy or give detailed history like Config/Config Rules do. It is a general purpose event stream.

AWS CloudFormation FAQ

General

Q: What is AWS CloudFormation?

AWS CloudFormation is a service that gives developers and businesses an easy way to create a collection of related AWS resources and provision them in an orderly and predictable fashion.

Q: What can developers now do with AWS CloudFormation that they could not before?

AWS CloudFormation automates and simplifies the task of repeatedly and predictably creating groups of related resources that power your applications. Creating and interconnecting all resources your application needs to run is now as simple as creating a single EC2 or RDS instance.

Q: How is AWS CloudFormation different from AWS Elastic Beanstalk?

These services are designed to complement each other. AWS Elastic Beanstalk provides an environment to easily deploy and run applications in the cloud. It is integrated with developer tools and provides a one-stop experience for you to manage the lifecycle of your applications. AWS CloudFormation is a convenient provisioning mechanism for a broad range of AWS resources. It supports the infrastructure needs of many different types of applications such as existing enterprise applications, legacy applications, applications built using a variety of AWS resources and container-based solutions (including those built using AWS Elastic Beanstalk).

AWS CloudFormation supports Elastic Beanstalk application environments as one of the AWS resource types. This allows you, for example, to create and manage an AWS Elastic Beanstalk–hosted application along with an RDS database to store the application data. In addition to RDS instances, any other supported AWS resource can be added to the group as well.

Q: What new concepts does AWS CloudFormation introduce?

AWS CloudFormation introduces two concepts: The *template*, a JSON or YAML-format, text-based file that describes all the AWS resources you need to deploy to run your application and the *stack*, the set of AWS resources that are created and managed as a single unit when AWS CloudFormation instantiates a template.

Q: How do I get started with AWS CloudFormation?

You can easily access AWS CloudFormation through the AWS Management Console, which gives you a point-and-click, web-based interface to deploy and manage stacks. You can create a new stack from inside the AWS Management Console in a few simple steps:

1. Give the stack a name: Provide a unique name for the stack.
2. Select a template: Select a template from your local file system or from a Amazon S3 URL. This may be one of the sample AWS CloudFormation templates, your own custom template, a template you are managing in a source control repository, or a template you got from a third party.
3. Specify any parameters: If the template allows you to configure the deployment, fill in any parameters or go with the specified defaults.
4. Click "Create": Start the deployment. You can see the current state of the deployment, with all the resource names and stack events in the AWS Management Console.

Q: What resources does AWS CloudFormation support?

To see a complete list of supported AWS resources and their features, visit the Supported AWS Services page in the Release History of the documentation.

AWS CloudFormation custom resources enable management of additional AWS and non-AWS resources.

Q: Can I manage individual AWS resources that are part of an AWS CloudFormation stack?

Yes. AWS CloudFormation does not get in the way; you retain full control of all elements of your infrastructure. You can continue using all your existing AWS and third-party tools to manage your AWS resources.

Q: What are the elements of an AWS CloudFormation template?

AWS CloudFormation templates are JSON or YAML-formatted text files that are comprised of five types of elements:

1. An optional list of template parameters (input values supplied at stack creation time)
2. An optional list of output values (e.g. the complete URL to a web application)
3. An optional list of data tables used to lookup static configuration values (e.g., AMI names)
4. The list of AWS resources and their configuration values
5. A template file format version number

With parameters, you can customize aspects of your template at run time, when the stack is built. For example, the Amazon RDS database size, Amazon EC2 instance types, database and web server port numbers can be passed to AWS CloudFormation when a stack is created. Each parameter can have a default value and description and may be marked as “NoEcho” in order to hide the actual value you enter on the screen and in the AWS CloudFormation event logs. When you create an AWS CloudFormation stack, the AWS Management Console will automatically synthesize and present a pop-up dialog form for you to edit parameter values.

Output values are a very convenient way to present a stack’s key resources (such as the address of an Elastic Load Balancing load balancer or Amazon RDS database) to the user via the AWS Management Console, or the command line tools. You can use simple functions to concatenate string literals and value of attributes associated with the actual AWS resources.

Q: How does AWS CloudFormation choose actual resource names?

You can assign logical names to AWS resources in a template. When a stack is created, AWS CloudFormation binds the logical name to the name of the corresponding actual AWS resource. Actual resource names are a combination of the stack and logical resource name. This allows multiple stacks to be created from a template without fear of name collisions between AWS resources.

Q: Why can’t I name all my resources?

Although AWS CloudFormation allows you to name some resources (such as Amazon S3 buckets), CloudFormation doesn’t allow this for all resources. Naming resources restricts the reusability of templates and results in naming conflicts when an update causes a resource to be replaced. To minimize these issues, CloudFormation will support resource naming on a case by case basis.

Q: Can I install software at stack creation time using AWS CloudFormation?

Yes. AWS CloudFormation provides a set of application bootstrapping scripts that enable you to install packages, files, and services on your EC2 instances by simply describing them in your CloudFormation template. For more details and a how-to see Bootstrapping Applications via

AWS CloudFormation.

Q: Can I use AWS CloudFormation with Chef?

Yes. AWS CloudFormation can be used to bootstrap both the Chef Server and Chef Client software on your EC2 instances. For more details and a how-to see Integrating AWS CloudFormation with Chef.

Q: Can I use AWS CloudFormation with Puppet?

Yes. AWS CloudFormation can be used to bootstrap both the Puppet Master and Puppet Client software on your EC2 instances. For more details and a how-to see Integrating AWS CloudFormation with Puppet.

Q: Does AWS CloudFormation support Amazon EC2 tagging?

Yes. Amazon EC2 resources that support the tagging feature can also be tagged in an AWS template. The tag values can refer to template parameters, other resource names, resource attribute values (e.g. addresses), or values computed by simple functions (e.g., a concatenated a list of strings).

AWS CloudFormation automatically tags Amazon EBS volumes and Amazon EC2 instances with the name of the AWS CloudFormation stack they are part of.

Q: Do I have access to the Amazon EC2 instance, or Auto Scaling Launch Configuration user-data fields?

Yes. You can use simple functions to concatenate string literals and attribute values of the AWS resources and pass them to user-data fields in your template. Please refer to our sample templates to learn more about these easy to use functions.

Q: What happens when one of the resources in a stack cannot be created successfully?

By default, the “automatic rollback on error” feature is enabled. This will cause all AWS resources that AWS CloudFormation created successfully for a stack up to the point where an error occurred to be deleted. This is useful when, for example, you accidentally exceed your default limit of Elastic IP addresses, or you don’t have access to an EC2 AMI you’re trying to run. This feature enables you to rely on the fact that stacks are either fully created, or not at all, which simplifies system administration and layered solutions built on top of AWS CloudFormation.

Q: Can stack creation wait for my application to start up?

Yes. AWS CloudFormation provides a *WaitCondition* resource that acts as a barrier, blocking the creation of other resources until a completion signal is received from an external source such as your application, or management system.

Q: Can I save my data when a stack is deleted?

Yes. AWS CloudFormation allows you to define deletion policies for resources in the template. You can specify that snapshots be created for Amazon EBS volumes or Amazon RDS database instances before they are deleted. You can also specify that a resource should be preserved and not deleted when the stack is deleted. This is useful for preserving Amazon S3 buckets when the stack is deleted.

Q: Can I update my stack after it has been created?

Yes. You can use AWS CloudFormation to modify and update the resources in your existing stacks in a controlled and predictable way. By using templates to manage your stack changes, you have the ability to apply version control to your AWS infrastructure just as you do with the software running on it.

Q: Can I create stacks in a Virtual Private Cloud (VPC)?

Yes. CloudFormation supports creating VPCs, Subnets, Gateways, Route Tables and Network ACLs as well as creating resources such as Elastic IPs, Amazon EC2 Instances, EC2 Security Groups, Auto Scaling Groups, Elastic Load Balancers, Amazon RDS Database Instances and Amazon RDS Security Groups in a VPC.

Getting Started

Q: How do I sign up for AWS CloudFormation?

To sign up for AWS CloudFormation, click**Create Free Account** on the AWS CloudFormation detail page. After signing up, please refer to the AWS CloudFormationdocumentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for AWS CloudFormation?

AWS CloudFormation registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a few minutes and involves receiving an automated phone call during the registration process and entering a PIN number using the phone key pad.

Q: How do I get started after I have signed up?

The best way to get started with AWS CloudFormation is to work through the Getting Started Guide, which is included in our technical documentation. Within a few minutes, you will be able to deploy and use one of our sample templates that illustrate how to create the infrastructure needed to run applications such as Tracks, WordPress, and others.

Q: Are there sample templates that I can use to check out AWS CloudFormation?

Yes, AWS CloudFormation includes sample templates that you can use to test drive the offering and explore its functionality. Our sample templates illustrate how to interconnect and use multiple AWS resources in concert, following best practices for multiple Availability Zone redundancy, scale out, and alarming. To get started, all you need to do is go to the AWS Management Console, click **Create Stack**, and follow the steps to select and launch one of our samples. Once created, select your stack in the console and review the **Template** and **Parameter** tabs to look at the details of the template file used to create the respective stack.

Billing

Q: How much does AWS CloudFormation cost?

There is no additional charge for AWS CloudFormation. You only pay for the AWS resources that are created (e.g., Amazon EC2 instances, Elastic Load Balancing load balancers etc.)

Q: Will I be charged for resources that were rolled back during a failed stack creation attempt?

Yes. Charges for AWS resources created during template instantiation apply irrespective of whether the stack as a whole could be created successfully or not.

Limits and Restrictions

Q: Are there limits to the number of templates or stacks?

There are no limits to the number of templates. Each AWS CloudFormation account is limited to a maximum of 200 stacks. Complete our request for a higher limit here, and we will respond to your request within two business days.

Q: Are there limits to the size of description fields?

Template, Parameter, Output, and Resource description fields are limited to 4096 characters.

Q: Are there limits to the number of parameters or outputs in a template?

You can include up to 60 parameters and 60 outputs in a template.

Regions and Endpoints

Q: What are the AWS CloudFormation service access points in each region?

Endpoints for each region are available in the technicaldocumentation.

Q: What are the AWS regions where AWS CloudFormation is currently available?

Please refer to Regional Products and Services for details of CloudFormation availability by region.

Amazon CloudTrail FAQ

General

Q:What is AWS CloudTrail?

AWS CloudTrail is a web service that records API calls made on your account and delivers log files to your Amazon S3 bucket.

Q:What are the benefits of CloudTrail?

CloudTrail provides visibility into user activity by recording API calls made on your account. CloudTrail records important information about each API call, including the name of the API, the identity of the caller, the time of the API call, the request parameters, and the response elements returned by the AWS service. This information helps you to track changes made to your AWS resources and to troubleshoot operational issues. CloudTrail makes it easier to ensure compliance with internal policies and regulatory standards. For more details, refer to the AWS compliance white paper "Security at scale: Logging in AWS".

Q:Who should turn on CloudTrail?

Customers who need to track changes to resources, answer simple questions about user activity, demonstrate compliance, troubleshoot, or perform security analysis should turn on CloudTrail.

Getting Started

Q: How do I get started with CloudTrail?

The quickest way to get started with CloudTrail is to use the AWS Management Console. You can turn on CloudTrail in few clicks.

Q:How does CloudTrail deliver API call information?

CloudTrail delivers API call information by depositing log files in an Amazon S3 bucket that you choose and configure. Each log file can contain multiple events, and each event represents an API call.

Services and Region Support

Q:What services are supported by CloudTrail?

For a list of services supported by CloudTrail, refer to theCloudTrail documentation.

Q:How are global AWS services supported?

API calls for global AWS services such as AWS IAM and AWS STS are recorded and delivered by CloudTrail along with regional events. By default, CloudTrail delivers API calls for global services in every region.

Q:What regions are supported?

Please refer to Regional Products and Services for details of CloudTrail availability by region.

Q:Are API calls made from the AWS Management Console recorded?

Yes. CloudTrail records API calls made from any client. The AWS Management Console, AWS SDKs, command line tools, and higher level AWS services call AWS APIs, so these calls are recorded.

Q:Where are my log files stored and processed before they are delivered to my Amazon S3 bucket?

API call information for services with regional end points (EC2, RDS etc.) is captured and processed in the same region as to which the API call is made and delivered to the region associated with your Amazon S3 bucket. API call information for services with single end points (IAM, STS etc.) is captured in the region where the end point is located, processed in the region where the CloudTrail trail is configured and delivered to the region associated with your Amazon S3 bucket.

Applying a Trail to all Regions

Q. What is applying a trail to all regions?

Applying a trail to all regions refers to creating the same trail in all regions in a partition. Currently, you can apply a trail to all regions in the **aws** partition that contains the following regions: US East (Northern Virginia), US West (Northern California), US West (Oregon), Europe (Ireland), Europe (Frankfurt), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), and South America (Sao Paulo). For more details on regions and partitions, refer to the Amazon Resource Names and AWS Service Namespaces page.

Q.What are the benefits of applying a trail to all regions?

You can create and manage a trail across all regions in the partition in one API call or few clicks. You will receive a record of API activity made in your AWS account across all regions to one S3 bucket or CloudWatch logs log group. When AWS launches a new region, you will receive the log files containing API activity for the new region without taking any action.

Q. How do I apply a trail to all regions?

In the CloudTrail console, you select yes to apply to all regions in the trail configuration page. If you are using the SDKs or AWS CLI, You set the IsMultiRegionTrail to true.

Q.What happens when I apply a trail to all regions?

Once you apply a trail in all regions, CloudTrail will create a new trail in all regions by replicating the trail configuration. CloudTrail will record and process the log files in each region and will deliver log files containing API activity across all AWS regions to a single S3 bucket and a single CloudWatch Logs log group. If you specified an optional SNS topic, CloudTrail will deliver SNS notifications for all log files delivered to a single SNS topic.

Q. Can I apply an existing trail to all regions?

Yes, you can apply an existing trail to all regions. When you apply an existing trail to all regions, CloudTrail will create a new trail for you in all regions. If you previously created trails in other regions, you can view, edit and delete those trails from the CloudTrail console.

Q. How long will it take for CloudTrail to replicate the trail configuration to all regions?

Typically, it will take less than 30 seconds to replicate the trail configuration to all regions.

Multiple Trails

Q. How many trails can I create in an AWS region?

You can create up to five trails in an AWS region. A trail that applies to all regions exists in each region and is counted as one trail in each region.

Q. What is the benefit of creating multiple trails in an AWS region?

With multiple trails, different stakeholders such as security administrators, software developers and IT auditors can create and manage their own trails. For example, a security administrator can create a trail that applies to all regions and configure encryption using one KMS key. A developer can create a trail that applies to one region for troubleshooting operational issues.

Q. Does CloudTrail support resource level permissions?

Yes, using resource level permissions, you can write granular access control policies to allow or deny access to specific users for a particular trail. For more details, go to CloudTrail documentation.

Security and Expiration

Q:How can I secure my CloudTrail log files?

By default, CloudTrail log files are encrypted using S3 Server Side Encryption (SSE) and placed into your S3 bucket. You can control access to log files by applying IAM or S3 bucket policies. You can add an additional layer of security by enabling S3 Multi Factor Authentication (MFA) Delete on your S3 bucket. For more details on creating and updating a trail, see the CloudTrail documentation.

Q:Where can I download a sample S3 bucket policy and an SNS topic policy?

You can download a sample S3 bucket policy and an SNS topic policy from CloudTrail S3 bucket. You need to update the sample policies with your information before you apply them to your S3 bucket or SNS topic.

Q:How long can I store my activity log files?

You control the retention policies for your CloudTrail log files. By default, log files are stored indefinitely. You can use Amazon S3 object lifecycle management rules to define your own retention policy. For example, you may want to delete old log files or archive them to Amazon Glaicer.

Event Payload, Timeliness and Delivery Frequency

Q:What information is available in an event?

An event contains information about the associated API call: the identity of the caller, the time of the call, the source IP address, the request parameters, and the response elements returned by the AWS service. For more details, see the CloudTrail Event Reference section of the user guide.

Q:How long does it take CloudTrail to deliver an event for an API call?

Typically, CloudTrail delivers an event within 15 minutes of the API call.

Q:How often will CloudTrail deliver log files to my Amazon S3 bucket?

CloudTrail delivers log files to your S3 bucket approximately every 5 minutes. CloudTrail does not deliver log files if no API calls are made on your account.

Q:Can I be notified when new log files are delivered to my Amazon S3 bucket?

Yes. You can turn on Amazon SNS notifications so that you can take immediate action on delivery of new log files.

Q: What happens if CloudTrail is turned on for my account but my Amazon S3 bucket is not configured with the correct policy?

CloudTrail log files are delivered in accordance with the S3 bucket policies that you have in place. If the bucket policies are misconfigured, CloudTrail may not be able to deliver log files.

Log File Aggregation

Q:I have multiple AWS accounts. I would like log files for all the accounts to be delivered to a single S3 bucket. Can I do that?

Yes. You can configure one S3 bucket as the destination for multiple accounts. For detailed instructions, refer to aggregating log files to a single Amazon S3 bucket sectionof the AWS CloudTrail User Guide

Look up API Activity

Q:What use cases can I solve by looking up API activity?

You can look up API activity captured by CloudTrail to troubleshoot operational and security incidents in your AWS account.

Q:Which API activity can I look up for my AWS account?

You can look up API activity related to creation, modification, and deletion of AWS resources in your AWS account for 28 AWS services, including Amazon EC2, Amazon VPC, Amazon RDS and AWS IAM. For a list of services, go to the CloudTrail documentation.

Q:How do I look up API activity captured by CloudTrail?

You can look up API activity captured by CloudTrail using the CloudTrail console, AWS SDKs, and AWS CLI.

Q:How do I look up API activity captured for my account?

If you have already turned on CloudTrail for your account, you do not need to do anything. Simply log on to the CloudTrail console to review the history of API activity for your AWS account. If you haven’t turned on CloudTrail for your account, you simply turn it on and, from that point onward, you can look up captured events.

Q:How far back in time can I look up API activity for my AWS account?

You can look up API activity captured for your AWS account for the last 7 days.

Q:What happens if I stop logging or delete a trail?

If you stop logging or delete a trail, CloudTrail will stop delivering events to your S3 bucket. You will not be able to look up events that occurred after you stopped logging or deleted a trail. You will still be able to look up events that occurred before you stopped logging or deleted a trail for 7 days. If you start logging again, CloudTrail will start delivering events to your S3 bucket, and you will be able to look up events that were captured after you resumed logging.

Q:What filters can I use to look up API activity?

You can specify one of the following attributes: Time range, Event name, User name, Resource name, and Resource type.

Q:In which regions can I look up API activity for my AWS account?

You can look up API activity for your AWS account in these AWS regions: US East (N. Virginia), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), and South America (Sao Paulo).

Integration with CloudWatch Logs

Q: What is CloudTrail integration with CloudWatch Logs?

CloudTrail integration with CloudWatch Logs delivers API activity captured by CloudTrail to a CloudWatch Logs log stream in the CloudWatch Logs log group you specify.

Q:What are the benefits of CloudTrail integration with CloudWatch Logs?

This integration enables you to receive SNS notifications of API activity captured by CloudTrail. For example, you can create CloudWatch alarms to monitor API calls that create, modify and delete Security Groups and Network ACL's. For examples, go to the examples section of the user guide.

Q:How do I turn on CloudTrail integration with CloudWatch Logs?

You can turn on CloudTrail integration with CloudWatch Logs from the CloudTrail console by specifying a CloudWatch Logs log group and an IAM role. You can also use the AWS SDKs or the AWS CLI to turn on this integration.

Q:What happens when I turn on CloudTrail integration with CloudWatch Logs?

After you turn on the integration, CloudTrail continuously delivers API activity to a CloudWatch Logs log stream in the CloudWatch Logs log group you specified. CloudTrail also continues to deliver logs to your Amazon S3 bucket as before.

Q:In which AWS regions is CloudTrail integration with CloudWatch Logs supported?

This integration is supported in the regions where CloudWatch Logs is supported, i.e., the us-east-1 (N.Virginia), us-west-1 (N.California), us-west-2 (Oregon), eu-west-1 (Ireland), ap-northeast-1 (Tokyo), ap-southeast-1 (Singapore), ap-southeast-2 (Sydney) and eu-central-1 (Frankfurt) regions. In the future, this integration will be available in other regions.

Q:How does CloudTrail deliver events containing API activity to my CloudWatch Logs?

CloudTrail assumes the IAM role you specify to deliver API activity to CloudWatch Logs. You limit the IAM role to only the permissions it requires to deliver events to your CloudWatch Logs log stream. To review IAM role policy, go to the user guide of the CloudTrail documentation.

Q:What charges do I incur once I turn on CloudTrail integration with CloudWatch Logs?

After you turn on CloudTrail integration with CloudWatch Logs, you incur standard CloudWatch Logs and CloudWatch charges. For details, go to CloudWatch pricing page.

CloudTrail Log File Encryption using AWS Key

Management Service (KMS)

Q: What is the benefit of CloudTrail log file encryption using Server-side Encryption with KMS?

CloudTrail log file encryption using SSE-KMS allows you to add an additional layer of security to CloudTrail log files delivered to an Amazon S3 bucket by encrypting the log files with a KMS key. By default, CloudTrail will encrypt log files delivered to your Amazon S3 bucket using Amazon S3 server-side encryption.

Q: I have an application that ingests and processes CloudTrail log files. Do I need to make any changes to my application?

With SSE-KMS, Amazon S3 will automatically decrypt the log files so that you do not need to make any changes your application. As always, you need to make sure that your application has appropriate permissions, i.e. Amazon S3 GetObject and KMS Decrypt permissions.

Q. How do I configure CloudTrail log file encryption?

You can use the AWS Management Console, or AWS CLI or the AWS SDKs to configure log file encryption. For detailed instructions, refer to the documentation.

Q:What charges do I incur once I configure encryption using SSE-KMS?

Once you configure encryption using SSE-KMS, you will incur standard AWS KMS charges. For details, go to AWS KMS pricing page.

CloudTrail Log File Integrity Validation

Q.What is CloudTrail log file integrity validation?

CloudTrail log file integrity validation feature allows you to detect whether a CloudTrail log file was unchanged, deleted, or modified since CloudTrail delivered it to the specified Amazon S3 bucket.

Q.What is the benefit of CloudTrail log file integrity validation?

You can use the log file integrity validation as an aid in your IT security and auditing processes.

Q.How do I enable CloudTrail log file integrity validation?

You can enable the CloudTrail log file integrity validation feature from the AWS Management

Console, AWS CLI or AWS SDKs.

Q.What happens once I turn on the log file integrity validation feature?

Once you turn on the log file integrity validation feature, CloudTrail will deliver digest files on an hourly basis. The digest files contain information about the log files that were delivered to your Amazon S3 bucket, hash values for those log files, digital signatures for the previous digest file, and the digital signature for the current digest file in the Amazon S3 metadata section. For more information about digest files, digital signatures and hash values, go to CloudTrail documentation.

Q. Where are the digest files delivered to?

The digest files are delivered to the same Amazon S3 bucket where your log files are delivered to. However, they are delivered to a different folder so that you can enforce granular access control policies. For details, refer to the digest file structure section of the CloudTrail documentation.

Q.How can I validate the integrity of a log file or digest file delivered by CloudTrail?

You can use the AWS CLI to validate that the integrity of log file or digest file. You can also build your own tools to do the validation. For more details on using the AWS CLI for validating the integrity of a log file, refer to the CloudTrail documentation.

Q.I aggregate all my log files across all regions and multiple accounts into one single Amazon S3 bucket. Will the digest files be delivered to the same Amazon S3 bucket?

Yes, CloudTrail will deliver the digest files across all regions and multiple accounts into the same Amazon S3 bucket.

AWS CloudTrail Processing Library

Q: What is AWS CloudTrail Processing Library?

AWS CloudTrail Processing Library is a Java library that makes it easy to build an application that reads and processes CloudTrail log files.You can download CloudTrail Processing Library from GitHub.

Q. What functionality does CloudTrail Processing Library provide?

CloudTrail Processing Library provides functionality to handle tasks such as continuously polling a SQS queue, reading and parsing SQS messages, downloading log files stored in S3, parsing and serializing events in the log file in a fault tolerant manner. For more information, go to the user guide section of the CloudTrail documentation.

Q. What software do I need to start using the CloudTrail Processing Library?

You need aws-java-sdk version 1.9.3 and Java 1.7 or higher.

Partners

Q:How do the AWS partner solutions help me analyze the events recorded by CloudTrail?

Multiple partners offer integrated solutions to analyze CloudTrail log files. These solutions include features like change tracking, troubleshooting, and security analysis. For more information, see the CloudTrail partners section.

Other

Q:What other logging support is available for AWS?

Amazon S3 provides server access logging, which enables logging for requests made against Amazon S3 buckets. Amazon CloudFront provides similar access logging support for CloudFront distributions.

Q:Will turning on CloudTrail impact the performance of my AWS resources, or increase API call latency?

No. Turning on CloudTrail has no impact on performance of your AWS resources or API call latency.

Amazon Config FAQ

General

What is AWS Config?

AWS Config is a fully managed service that provides you with an AWS resource inventory, configuration history, and configuration change notifications to enable security and governance.

With AWS Config you can discover existing AWS resources, export a complete inventory of your AWS resources with all configuration details, and determine how a resource was configured at any point in time. These capabilities enable compliance auditing, security analysis, resource change tracking, and troubleshooting.

What is a Config Rule?

A Config Rule represents desired configurations for a resource and is evaluated against configuration changes on the relevant resources, as recoded by AWS Config. The results of evaluating a rule against the configuration of a resource are available on a dashboard. Using Config Rules, you can assess your overall compliance and risk status from a configuration perspective, view compliance trends over time and pinpoint which configuration change caused a resource to drift out of compliance with a rule.

What are the benefits of AWS Config?

AWS Config makes it easy to track your resource's configuration without the need for up-front investments and avoiding the complexity of installing and updating agents for data collection or maintaining large databases. Once you enable AWS Config, you can view continuously updated details of all configuration attributes associated with AWS resources. You are notified via Amazon Simple Notification Service (SNS) of every configuration change.

How can AWS Config help with audits?

AWS Config gives you access to resource configuration history. You can relate configuration changes with AWS CloudTrail events that possibly contributed to the change in configuration. This information provides you full visibility, right from details, such as "Who made the change?", "From what IP address?" to the effect of this change on AWS resources and related resources. You can use this information to generate reports to aid auditing and assessing compliance over a period of time.

Who should use AWS Config and Config Rules?

Any AWS customer looking to improve their security and governance posture on AWS by continuously evaluating the configuration of their resources would benefit from this capability. Administrators within larger organizations who recommend best practices for configuring resources can codify these rules as Config Rules, and enable self-governance among users. Information Security experts who monitor usage activity and configurations to detect vulnerabilities can benefit from Config Rules. Customers with workloads that need to comply with specific standards (e.g. PCI-DSS or HIPAA) can use this capability to assess compliance of their AWS infrastructure configurations, and generate reports for their auditors. Operators who manage large AWS infrastructure or components that change frequently can also benefit from Config Rules for troubleshooting.Customers who want to track changes to resources configuration, answer questions about resource configurations, demonstrate compliance, troubleshoot or perform security analysis should turn on AWS Config.

Does the service guarantee that my configurations are never out of compliance?

Config Rules provides information about whether your resources are compliant with configuration rules you specify. It will evaluate rules as soon as updated Configuration Items (CIs) for the resource are available within AWS Config. It does not guarantee that resources will be compliant or prevent users from taking non-compliant actions. Further, Config Rules does not automatically snap non-compliant resources back into compliance.

Does the service prevent users from taking non-compliant actions?

Config Rules does not directly affect how end-users consume AWS. It evaluates resource configurations only after a configuration change has been completed and recorded by AWS Config. Config Rules does not prevent the user from making changes that could be non-compliant. To control what a user can provision on AWS and configuration parameters allowed during provisioning, please use AWS Identity and Access Management (IAM) Policies and AWS Service Catalog respectively.

Can rules be evaluated prior to provisioning a resource?

Config Rules evaluates rules after the Configuration Item (CI) for the resource is captured by AWS Config. It does not evaluate rules prior to provisioning a resource or prior to making configuration changes on the resource.

How does AWS Config work with AWS CloudTrail?

AWS CloudTrail records user API activity on your account and allows you to access information about this activity. You get full details about API actions, such as identity of the caller, the time of the API call, the request parameters, and the response elements returned by the AWS service. AWS Config records point-in-time configuration details for your AWS resources as Configuration Items (CIs). You can use a CI to answer “What did my AWS resource look like?” at a point in time. You can use AWS CloudTrail to answer “Who made an API call to modify this resource?” For example, you can use the AWS Management Console for AWS Config to detect security group “Production-DB” was incorrectly configured in the past. Using the integrated AWS CloudTrail information, you can pinpoint which user misconfigured “Production-DB” security group.

Getting Started

How do I get started with this service?

The quickest way to get started with AWS Config is to use the AWS Management Console. You can turn on AWS Config in a few clicks. For additional details, see the Getting Started documentation.

How do I access my resources’ configuration?

You can lookup current and historical resource configuration using the AWS Management Console, AWS Command Line Interface or SDKs.

For additional details, please refer to AWS Config documentation.

Do I turn on AWS Config regionally or globally?

You turn on AWS Config on a per-region basis for your account.

Can AWS Config aggregate data across different AWS accounts?

Yes, you can set up AWS Config to deliver configuration updates from different accounts to one S3 bucket, once the appropriate IAM policies are applied to the S3 bucket. You can also publish notifications to the one SNS Topic, within the same region, once appropriate IAM policies are applied to the SNS Topic.

Is API activity on AWS Config itself logged by AWS CloudTrail?

Yes. All AWS Config API activity, including use of AWS Config APIs to read configuration data, is logged by AWS CloudTrail.

What time and timezones are displayed in the timeline view of a resource? What about daylight savings?

AWS Config displays the time at which Configuration Items (CIs) were recorded for a resource on a timeline. All times are captured in Coordinated Universal Time (UTC). When the timeline is visualized on the management console, the services uses the current time zone (adjusted for daylight savings, if relevant) to display all times in the timeline view.

Config Rules

What is a resource's configuration?

Configuration of a resource is defined by the data included in the Configuration Item (CI) of AWS Config. The initial release of Config Rules makes the CI for a resource available to relevant rules. Config Rules can use this information along with any other relevant information such as other attached resource, business hours, etc. to evaluate compliance of a resource's configuration.

What is a rule?

A rule represents desired Configuration Item (CI) attribute values for resources and are evaluated by comparing those attribute values with CIs recorded by AWS Config. There are two types of rules:

AWS managed rules: AWS managed rules are pre-built and managed by AWS. You simply choose the rule you want to enable, then supply a few configuration parameters to get started.

Learn more »

Customer managed rules: Customer managed rules are custom rules, defined and built by you. You can create a function in AWS Lambda that can be invoked as part of a custom rule and these functions execute in your account. Learn more »

The quickest way to get started with AWS Config is to use the AWS Management Console. You can turn on AWS Config in a few clicks. For additional details, see the Getting Started documentation.

How are rules created?

Rules are typically set up by the AWS account administrator. They can be created by leveraging AWS managed rules – a predefined set of rules provided by AWS or through customer managed rules. With AWS managed rules updates to the rule are automatically applied to any account using that rule. In the customer-managed model, the customer has a full copy of the rule, and executes the rule within his/her own account. These rules are maintained by the customer.

How many rules can I create?

Currently, you can create up to 25 rules per account.

How are rules evaluated?

Any rule can be setup as a change-triggered rule or as a periodic rule. A change-triggered rule is executed when AWS Config records a configuration change for any of the resources specified. Additionally, one of the following must be specified:

> Tag Key:(optional Value): A tag key:value implies any configuration changes recorded for resources with the specified tag key:value will trigger an evaluation of the rule.
>
> Resource type(s): Any configuration changes recorded for any resource within the specified resource type(s) will trigger an evaluation the rule.
>
> Resource ID: Any changes recorded to the resource specified by the resource type and resource ID will trigger an evaluation of the rule.

A periodic rule is triggered at a specified frequency. Available frequencies are 1hr, 3hr, 6hr, 12hr or 24hrs. A periodic rule has a full snapshot of current Configuration Items (CIs) for all resources available to the rule.

What is an evaluation?

Evaluation of a rule determines whether a rule is compliant with a resource at a particular point in time. It is the result of evaluating a rule against the configuration of a resource. Config Rules will capture and store the result of each evaluation. This result will include the resource, rule, time of evaluation and a link to Configuration Item (CI) that caused non-compliance.

What does compliance mean?

A resource is compliant if complies with all rules that apply to it. Otherwise it is noncompliant. Similarly, a rule is compliant if all resources evaluated by the rule comply with the rule. Otherwise it is noncompliant. In some cases, such as when inadequate permissions are available to the rule, an evaluation may not exist for the resource, leading to a state of insufficient data. This state is excluded from determining the compliance status of a resource or rule.

What information does the Config Rules dashboard provide?

The Config Rules dashboard gives you an overview of resources tracked by AWS Config, and a summary of current compliance by resource and by rule. When you view compliance by resource, you can determine if any rule that applies to the resource is currently not compliant. You can view compliance by rule, which tells you if any resource under the purview of the rule is currently non-compliant. Using these summary views, you can dive deeper into the Config timeline view of resources, to determine which configuration parameters changed. Using this dashboard, you can start with an overview and drill into fine-grained views that give you full information about changes in compliance status, and which changes caused non-compliance.

Services and Region Support

What AWS resources types are covered by AWS Config?

At this time, AWS Config monitors configurations for the following resource types.

Amazon EC2

- EC2 Instance
- EC2 Network Interface
- EC2 Security Group
- EC2 Dedicated Hosts
- EC2 Elastic IP (VPC only)

Amazon VPC

- Customer Gateway
- Internet Gateway
- Network ACL
- RouteTable
- Subnets
- VPCs
- VPN Gateway
- VPN Connection

AWS Identity and Access Management (IAM)

- IAM User (includes inline policies)
- IAM Group
- IAM Role
- IAM Managed Policy (Customer-managed only)

Amazon EBS

- EBS Volumes (all types)

AWS CloudTrail

- Trail

What regions is AWS Config available in?

For details on the regions where AWS Config is available, please visit this page:

http://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/

Resource Configuration

What is a configuration item?

A Configuration Item (CI) is the configuration of a resource at a given point-in-time. A CI consists of 5 sections:

1. Basic information about the resource that is common across different resource types (e.g., Amazon Resource Names, tags),

2. Configuration data specific to the resource (e.g., EC2 instance type),

3. Map of relationships with other resources (e.g., EC2::Volume vol-3434df43 is “attached to instance” EC2 Instance i-3432ee3a),

4. AWS CloudTrail event IDs that are related to this state,

5. Metadata that helps you identify information about the CI, such as the version of this CI, and when this CI was captured.

Learn more about configuration items

What are AWS Config relationships and how are they used?

AWS Config takes the relationships among resources into account when recording changes. For example, if a new Amazon EC2 Security Group is associated with an Amazon EC2 Instance, AWS Config records the updated configurations of both the primary resource, the Amazon EC2 Security Group, and related resources, such as the Amazon EC2 Instance, if these resources actually changed.

Does AWS Config record every state a resource has been in?

AWS Config detects change to resource's configuration and records the configuration state that resulted from that change. In cases where several configuration changes are made to a resource in quick succession (e.g. within a span of few minutes), Config will only record the latest configuration of that resource that represents cumulative impact of the set of changes. In these situations, Config will only list the latest change in the *relatedEvents* field of the Configuration Item.This allows users and programs to continue to change infrastructure configurations without having to wait for Config to record intermediate transient states.

Does AWS Config record configuration changes that did not result from API activity on that resource?

Yes, AWS Config will regularly scan configuration of resources for changes that haven't yet been recorded and record these changes. CIs recorded from these scans will not have a *relatedEvent* field in the payload, and only the latest state that is different from state already recorded is picked up.

Pricing

How will I be charged for this service?

With AWS Config, you are charged based on the number Configuration Items (CIs) recorded for supported resources in your AWS account. You are charged only once for recording the CI. There no additional fee for retaining the CI or any up-front commitment. You can stop recording CIs at any time and continue to access the CIs previously recorded. Charges per CI are rolled up into your monthly bill. See pricing details.

If you are using AWS Config Rules, you will be charged based on active Config Rules in that month. When a rule is compared with an AWS resource, the result is recorded as an evaluation. An rule is active if it has one or more evaluations in a month.

Configuration snapshots and configuration history files are delivered to you in the Amazon S3 bucket that you choose, and configuration change notifications are delivered via Amazon Simple Notification Service (SNS). Standard rates for Amazon S3 and Amazon SNS apply. Customer managed rules are authored using AWS Lambda. Standard rates for AWS Lambda apply.

Does the pricing for Config Rules include the costs for AWS Lambda functions?

You can choose from a set of managed rules provided by AWS or you can author your own rules, written as AWS Lambda functions. Managed rules are fully managed and maintained by AWS and you do not pay any additional AWS Lambda charges to run them. Simply enable managed rules, provide any required parameters, and pay a single rate for each AWS Config rule. On the other hand, customer managed rules give you full control by executing these rules as AWS Lambda functions in your account. In addition to monthly charges for an active rule, standard AWS Lambda free tier and function execution rates apply to customer managed rules.

What does shared quota for Config Rules mean?

You receive a quota 20,000 evaluations per active rule per month. For example, if you have 3 Config Rules, you get a quota of 60,000 evaluations for the account. You can choose spread this allowance across the rules in any way.

Do unused evaluations carry over to the next month?

Unused evaluations expire and are reset every billing cycle.

Can you provide breakdown of charges using an example?

Pricing example 1:
AWS Config records each AWS resource and configuration change as a Configuration Item (CI). Assume you record 7,000 CIs/month and have created 5 active rules (2 periodic and 3 change triggered), reporting a combined total of 150 evaluations per day.

AWS Config costs: 7,000 * $0.003 = $21.00
Cost for 5 active rules = 5 * $2.00 = $10.00

Quota for evaluation results = 5 * 20,000 = 100,000
Number of evaluation results used = 150 evaluations * 30 days = 4,500 evaluations/month
Additional charges from evaluation results = $0.0

Total AWS Config monthly charges = $31.00

The service charges you incur depend on the number of CIs recoded by your resources. This depends on the number of resources in your account, and the configuration changes you make to these resources. For an account with several hundred resources, and standard configuration change activity, AWS Config would capture fewer than 3,000 CIs per month, or less than $9 per month.

Pricing Example 2:
Assume you record 50,000 CIs/month and have created 2 active rules, and each of these is evaluated on every CI and report a result results each time.

AWS Config costs: 50,000 * $0.003 = $150.00
Cost for 2 active rules = 2 * $2.00 = $4.00

Quota for evaluation results = 2 * 20,000 = 40,000
Number of evaluation results used = 2 * 50,000 = 100,000
Additional charges from evaluation results = (100,000 – 40,000) = 60,000 * 0.0001 = $6.00

Total AWS Config monthly charges = $150.00 + $4.00+ $6.00 = $160.00

Partner Solutions

What AWS partner solutions are available for AWS Config?

Ecosystem partners such as Splunk, ServiceNow, Evident.IO, CloudCheckr, Redseal Networks and RedHat CloudForms provide offerings that are fully integrated with data from AWS Config. Managed Service Providers, such as 2nd Watch and CloudNexa have also announced integrations with AWS Config. Additionally, with Config Rules, partners such as CloudHealth Technologies, AlertLogic and TrendMicro are providing integrated offerings that can be used by customers. These solutions include capabilities such as change management and security analysis and allow you to visualize, monitor and manage AWS resource configurations.

For more information, see AWS Config partner solutions.

AWS Management Console FAQ

Web Console

Q: What is the AWS Management Console?

The AWS Management Console provides a simple web interface for Amazon Web Services. You can log in using your AWS account name and password. If you've enabled AWS Multi-Factor Authentication, you will be prompted for your device's authentication code.

Q: How do I sign into the Management Console?

You can sign into the management console using your AWS or IAM account credentials at https://console.aws.amazon.com/console/home. For the AWS GovCloud (US) region, you can sign into the management console using your IAM account credentials at https://console.amazonaws-us-gov.com.

Q: Why are you changing the console design?

Our goal is to improve information display, make interactions more consistent, support devices such as tablets, and deliver a customizable experience. You will see these improvements and visual updates rolled out across our services over the coming months.

Q: Can I provide feedback?

Yes! Click the **Feedback** button at the bottom of the console. We're eager to hear about your experience with the new console.

Q: What browsers does the Management Console support?

Important: As of May 1, 2016, the AWS Management Console no longer supports versions of Internet Explorer older than version 11. We recommend migrating to a more recent browser version to ensure the best possible experience and security.

Please contact us if you have any questions.

Browser	Version	Service
Google Chrome	Latest 3 Versions	All services
Mozilla Firefox	Latest 3 Versions	All services
Microsoft Internet Explorer	11	All services
Microsoft Edge	12	All services
Apple Safari	9, 8, 7	All services

Q: When does my session expire?

For security purpose, a login session will expire in 12 hours when you sign into the AWS Management Console with your AWS or IAM account credentials. To resume your work after the session expires, we ask you to click the "**Click login to continue**" button and login again. The duration of federated sessions varies depending on the federation API (GetFederationToken or AssumeRole) and the administrator’s preference. Please go to our Security Blog to learn more about building a secure delegation solution to grant temporary access to your AWS account.

Resource Groups

Q: What is a Resource Group?

A resource group is a collection of resources that share common tags. With the Resource Groups tool, you can create a custom console that organizes and consolidates the information you need based on your project and the resources you use. If you manage resources in multiple regions, you can create a resource group to view resources from different regions on the same screen.

Q: How can I use Resource Groups?

Resource Groups are a simple way to organize and find the resources that you use every day. You can create a resource group for each project, application, or environment that you manage in your AWS account. Since a resource group is simply a collection of resources that share common tags that you have applied to those resources, you can create a resource group for collections of resources that complement the way you use the AWS Management Console. Create resource groups that help you work faster and make you more productive.

To read more about how to use Resource Groups features,click here.

Q: How much does Resource Groups cost?

Resource Groups is free to use. You will not incur any additional charges for creating or using them.

Q: Who can see my resource groups?

Your resource groups are unique to your identity. Each IAM user identity has its own Resource Groups storage, so other identities in your account will not see the resource groups that you create. However, tags on an account's resources are visible to all identities that have permission to view tags in that account.

Q: What permissions do I need to use Resource Groups?

Click here to view the specific permissions required to use Resource Groups.

Q: How can I create a resource group using a tag substring or wildcards in my tag search?

Resource Groups lets you include resources in your resource group by identifying a tag substring. This is similar to appending a wildcard to the beginning and end of a string in the tag value field. To use this feature, begin typing a string in the tag value field of the Resource Group create or edit forms and select the **Contains:** option to find values that contain the characters that you typed. For example, a search using **Contains: Prod** value for the "Name" tag key would return a resource tagged with "FooApp-DB-Prod" as the value in the "Name" tag key. Remember, tag search is case-sensitive.

Q: How many resource groups can I create?

You can create up to 20 resource goups, unless you reach the storage limit. Since Resource Groups can be configured to search for resources based on a number of different parameters (tag key, tag value, resource type, and region), and because each parameter can be of varying lengths, each saved resource group may be a different size. If your groups are simple, you will be able to store up to 20. If your groups are complex, you may receive an error message notifying you that you have reached the storage limit. To create more resource groups, simplify your group configurations to use fewer and smaller parameters.

Q: What resource types are supported in Resource Groups?

Click here to view a list of resource types supported in Resource Groups.

Q: Why can’t I view a list of existing tag keys or values when searching for resources, or why doesn’t tag autocomplete work?

You may have limited permissions to access tag data.Click here to view the specific permissions required to use Resource Groups.

Q: Why can’t I find resources using a resource group?

You may have limited permissions to access tag data.Click here to view the specific permissions required to use Resource Groups. If you have all required permissions to use Resource Groups, check your resource group's configuration and ensure that the resources you are searching for have the correct tags.

Q: The system recognizes me as a federated user. Why are my resource groups shared, or why can't I find my resource groups after I sign out?

For federated users using SAML, resource groups are stored using the value provided for RoleSessionName. This value is configured by your organization's identity provider. More info about SAML federation setup can be found here. This value should be unique to each user authenticating with each role. If RoleSessionName is different for each session, saved resource groups will not be accessible after the session is terminated. If RoleSessionName is the same for each user, saved resource groups will be shared across all identities, and they will not function correctly.

Q: Can I provide feedback?

Yes! Click the **Feedback** button at the bottom of the console. We’re eager to hear about your experience with Resource Groups.

Tag Editor

Q: What is Tag Editor?

Tag Editor is a tool to view and manage tags on your AWS resources, regardless of service or region. Use the tag editor to search for resources by resource type, region, or tag, and then manage the tags applied to those resources.

Q: How can I use Tag Editor?

Click here to read documentation on the Tag Editor’s features and how to use them.

Q: What permissions do I need to use Tag Editor?

Click here to view the specific permissions required to use Tag Editor.

Q: How do tags work?

Tags are words or phrases that act as metadata for organizing your AWS resources. A tag is a key-value pair. In Tag Editor, tag keys are represented as columns, and tag values are strings in the cells of a tag key’s column. Click here to read more about how tags work.

Q: How can I bulk edit tags for multiple resources?

You can use Tag Editor to edit tags for multiple resources in multiple regions at once. In Tag Editor, select the checkboxes for each of the resources you want to edit, and then click **Edit tags for selected**. Follow the on-screen prompts to manage tags on these resources.

Q: How can I search for resources by tag substring or use wildcards in my tag search?

Tag Editor lets you search for resources by tag substring. This is similar to appending a wildcard to the beginning and end of a string. To use this feature, begin typing a string in the tag value field of the Tag Editor search form and select the **Contains:** option to find values that contain the characters that you typed. For example, a search using **Contains: Prod** value for the "Name" tag key would return a resource tagged with "FooApp-DB-Prod" as the value in the "Name" tag key. Remember, tag search is case-sensitive.

Q: How can I use the Tag Editor to search for resources that do not have a particular tag key applied or that have an empty value?

Tag Editor lets you search for resources that do not have a particular tag key applied. You can also search for resources that have a tag key applied with an empty (blank) tag value. To search for untagged resources, select the appropriate tag key, and then select "Not tagged" or "Empty value" in the tag value search field.

Q: What resource types can be tagged using the Tag Editor?

Tag Editor supports all resource types that support tags. You can view all of the services that support tags here.

Q: Why can’t I search for resources by tag?

You may have limited permissions to access tag data.Click here to view the specific permissions required to use Tag Editor.

Q: Why can’t I add, remove, or modify a tag key or value?

You may have limited permissions to access tag data.Click here to view the specific permissions required to use Tag Editor.

Q: Can I provide feedback?

Yes! Click the **Feedback** button at the bottom of the console. We’re eager to hear about your experience with the Tag Editor.

AWS Console Mobile App

Q: How do I sign in?

The app supports several authentication methods, including owner/root credentials, IAM user credentials, and AWS access keys. An owner account is the AWS login that created the account. An IAM user is an identity that has been created by an administrator through the IAM service. Note that IAM users need to also provide their account alias, which can be found at the top of the web console sign-in screen. AWS access keys are used to sign programmatic requests that the app makes to AWS.

For security reasons, we recommend that you secure your device with a passcode and that you follow an AWS best practice by creating and using an IAM user's credentials to log in to the app. If you lose your device, an IAM user can be deactivated to prevent unauthorized access. Root accounts cannot be deactivated.

Click here to learn more about the different types of AWS security credentials.

Q: Where can I download the app?

Download the app from Amazon Appstore, Google Play, or iTunes.

Q: What services are supported?

The app supports Elastic Compute Cloud (EC2), Amazon S3, Elastic Load Balancing (ELB), Amazon Route 53, Amazon Relational Database Service (RDS), Auto Scaling, AWS Elastic Beanstalk, Amazon DynamoDB, AWS CloudFormation, AWS OpsWorks, and CloudWatch. The mobile app does not support the AWS GovCloud (US) region. For a full description, see the AWS Console Mobile App page. We plan to add new features to the mobile app. Tell us what you need using the feedback link in the app.

Q: Is MFA supported?

Yes. We recommend using either a hardware MFA device or a virtual MFA on a separate mobile device for the greatest level of account protection.

Q: Can I create resources?

You cannot create resources in the current version. However, we’re considering this for future releases. Please use the feedback link in the app's menu to tell us what you need.

Q: Can I download S3 objects?

You can use the app to generate a pre-signed URL for an S3 object. A pre-signed URL grants time-limited permission to download the object. Read more about pre-signed URLs here.

In order to open a pre-signed URL for an S3 object in your device's browser, use the app to navigate to the S3 object's detail page and tap "View in browser". Your device configuration will determine what actions are possible with the object.

Q: Can I view my current AWS usage charges?

Yes, you can view your current usage charges in the app. Simply visit yourBilling Preferences page and select the checkbox to Receive Billing Alerts. In order to view usage charges, your identity must have permission to view CloudWatch.

Q: What time period and services does the Service Health section cover?

The Service Health section covers all AWS services in all regions for the previous 36 hours.

Q: What versions of Android are supported?

iOS 5.0+ and Android 2.3+ are supported.

Q: Does the app support tablets?

The app is currently optimized for iOS and Android phones, but it works on iPad and Android tablet devices.

Q: What Android app permissions are required?

- Network communication
- Full internet access (used to contact Amazon Web Services)
- View network state (used to gracefully handle network loss)

Q: My app is having trouble. What should I do?

From your phone's home screen press the menu and select settings. From the settings options, go to Apps, select AWS Console, and press the **clear data** button. The next time you start the app, the app will be reset.

Q: I lost my phone. What should I do?

We strongly recommend that in addition to signing out of the app when you have completed your tasks and using a password lock on your phone, you use an IAM user to manage AWS on your phone. If you lose your phone, you can remove the IAM user's access.

Q: Can I provide feedback?

Yes! Click the **Feedback** button in the app's menu. We're eager to hear about your experience with the app.

AWS OpsWorks FAQ

General

Q: What is AWS OpsWorks?

AWS OpsWorks is a flexible configuration management solution with automation tools that enable you to model and control your applications and their supporting infrastructure. AWS OpsWorks makes it easy to manage the complete application lifecycle, including resource provisioning, configuration management, application deployment, software updates, monitoring, and access control. AWS OpsWorks is designed for IT administrators and ops-minded developers who want an easy way to manage applications of nearly any scale and complexity without sacrificing control. With AWS OpsWorks you can create a logical architecture, provision resources based on that architecture, deploy your applications and all supporting software and packages in your chosen configuration, and then operate and maintain the application through lifecycle stages such as auto-scaling events and software updates.

Q: Who should use AWS OpsWorks?

System administrators and ops-minded developers who are looking for a powerful end-to-end configuration management solution should consider AWS OpsWorks. AWS OpsWorks is targeted at DevOps users who want better management and automation tools to help them customize and control their environments. An AWS OpsWorks user typically values:

Control. AWS OpsWorks makes it easy to model all the components of your application and then configure any aspect of your application and its supporting infrastructure. With support for scripted changes using Chef recipes (learn more here) at defined stages in the application lifecycle, you have fine-grained control of your application and its interaction with related components. Your recipes can be stored with your source code, making it easy to track changes. From one-time deployments to auto scaled growth, your application will reflect your settings through its complete lifecycle.

Automation. Instead of manual steps, you specify how to scale, maintain, and deploy your applications and AWS OpsWorks performs the tasks for you. For example, AWS OpsWorks can set up instances to host your apps based on the exact configurations that you specify (code to deploy, etc.), scale your apps using load-based or time-based auto scaling, and maintain the health of your apps by detecting and replacing failed instances. AWS OpsWorks uses Chef recipes to start new app server instances, configure app server software, and deploy apps. You can also apply your own Chef recipes to make changes to your database and monitoring infrastructure.

Q: What can users do with AWS OpsWorks that they could not do before?

AWS OpsWorks delivers a solution that lets you:

Model and support any application. You can deploy your application in the configuration you choose on Amazon Linux, Ubuntu, RHEL, and Windows. AWS OpsWorks lets you model your application with layers. Layers define how to configure a set of resources that are managed together. For example, you might define a web layer for your application that consists of Amazon EC2 instances, Amazon EBS volumes, and Elastic IP addresses. You can also define the software configuration for each layer, including installation scripts and initialization tasks. When an instance is added to a layer, AWS OpsWorks automatically applies the specified configuration. Because AWS OpsWorks supports Chef recipes (visit here for details), you can leverage hundreds of community-built configurations such as PostgreSQL, Nginx, and Solr. For example, you can create an application that consists of multiple Python apps installed on Django connected to a CouchDB database.

Automate tasks. AWS OpsWorks enables you to automate management actions so that they are performed automatically and reliably. You can benefit from automatic failover, package management, Elastic Load Balancing configuration, and automatic rule-based or time-based instance scaling. Common tasks automatically handled for you, and you can also extend and customize that automation. AWS OpsWorks supports continuous configuration through lifecycle events that automatically update your instances' configuration to adapt to environment changes, such as auto scaling events. With AWS OpsWorks there is no need to log in to several machines and manually update your configuration. Whenever your environment changes, AWS OpsWorks updates your configuration.

Q: What kinds of applications are supported by AWS OpsWorks?

AWS OpsWorks supports a wide variety of application architectures, from simple web applications to highly complex custom applications.

Q: How can I access AWS OpsWorks?

AWS OpsWorks is available through the AWS Management Console, AWS SDKs, and the AWS Command Line Interface.

Q: What regions does AWS OpsWorks support?

Please refer to Regional Products and Services for details of OpsWorks availability by region.

Q: How is AWS OpsWorks different than AWS CloudFormation?

AWS OpsWorks and AWS CloudFormation both support application modeling, deployment, configuration, management, and related activities. Both support a wide variety of architectural patterns, from simple web applications to highly complex applications. AWS OpsWorks and AWS CloudFormation differ in abstraction level and areas of focus.

AWS CloudFormation is a building block service that enables customers to provision and manage almost any AWS resource via a JSON-based domain specific language. AWS CloudFormation focuses on providing foundational capabilities for the full breadth of AWS, without prescribing a particular model for development and operations. Customers define templates and use them to provision and manage AWS resources, operating systems and application code.

In contrast, AWS OpsWorks is a higher level service that focuses on providing highly productive and reliable DevOps experiences for IT administrators and ops-minded developers. To do this, AWS OpsWorks employs a configuration management model based on concepts such as stacks and layers, and provides integrated experiences for key activities like deployment, monitoring, auto-scaling, and automation. Compared to AWS CloudFormation, AWS OpsWorks supports a narrower range of application-oriented AWS resource types including Amazon EC2 instances, Amazon EBS volumes, Elastic IPs, and Amazon CloudWatch metrics.

Q: How is AWS OpsWorks different than AWS Elastic Beanstalk?

AWS OpsWorks is a configuration management platform while AWS Elastic Beanstalk is an application management platform.

AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker. Customers upload their code and Elastic Beanstalk automatically does the rest.

AWS OpsWorks and AWS Elastic Beanstalk both automate operations but serve different needs and purposes. AWS Elastic Beanstalk is designed for developers who want to deploy web applications without worrying about operations. Developers simply upload their code and Elastic Beanstalk automatically handles the deployment, from capacity provisioning, load balancing, auto-scaling to application health monitoring. The application will be ready to use without any infrastructure or resource configuration work on the developer’s part.

In contrast, AWS OpsWorks is an integrated configuration management platform for IT administrators and DevOps engineers who want a high degree of customization and control over operations. AWS OpsWorks users leverage Chef recipes to automate operations like software configurations, package installations, database setups, server scaling, and code deployment.

Q: Can I manage resources created by AWS OpsWorks using other service consoles or CLIs?

While all the resources used by OpsWorks to build your environment are visible in their respective services, in general you should manage the resources exclusively through OpsWorks to avoid side effects with OpsWorks automation. There are a few exceptions, however:

- AWS OpsWorks will create and delete volumes for instances based on the configuration specified in your layer. If you want to snapshot the volumes, or manage volumes you have chosen to retain when an instance is terminated, you can do this directly with the Amazon EC2 service.
- AWS OpsWorks provides security groups for your layers with defaults that match the ports required for that layer. If you want to customize the security group or create new security groups, you can do this directly with the EC2 service.
- AWS OpsWorks uses key pairs that enable you to ssh into your instances. If you want to create or manage key pairs, you can do this directly with the Amazon EC2 service.
- AWS OpsWorks creates default AWS IAM policies for the users you add to your stack. If you want finer grained permissions, you will need to add those permissions to the user in IAM.
- AWS OpsWorks sends metrics from all your resources to Amazon CloudWatch. If you want to view these metrics or set alarms, you can do this directly with Amazon CloudWatch.

Q: Are AWS OpsWorks instances the same as Amazon EC2 instances?

An AWS OpsWorks instance defines an Amazon EC2 instance and its relationship with related resources, such as its availability zone, type, and associated volumes. An AWS OpsWorks instance may be represented by many Amazon EC2 instances over its lifetime, but only one at a time. This lets AWS OpsWorks consistently bind resources such as volumes and Elastic IPs with the Amazon EC2 instance when it is started.

Q: Are there any limits to AWS OpsWorks?

By default, you can create up to 40 Stacks, and each stack can hold up to 40 layers, 40 instances, and 40 apps. You should also be aware of other AWS limits. For example, the default AWS account limits allow you to launch up to 20 Amazon EC2 instances. If you need more resources, complete a request form and your request will be evaluated promptly. You can also install a limited number of packages through each OpsWorks layer. If you want to install a larger number of packages, use a custom cookbook to install packages.

Q: Does AWS OpsWorks support Windows Server?

Yes. AWS OpsWorks supports Windows Server 2012 R2

Q: Does AWS OpsWorks support on-premises servers?

Yes. AWS OpsWorks supports all Linux machines that can install the OpsWorks agent and have connection to AWS public endpoints.

Q: What network requirements do my servers have to have to work with AWS OpsWorks?

Your servers will just need to be able to connect to AWSpublic endpoints.

Getting Started

Q: How do I sign up for AWS OpsWorks?

To sign up for AWS OpsWorks, click the Sign Up Now button on the OpsWorks detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the AWS OpsWorks process.

Q: How do I get started with AWS OpsWorks?

The best way to get started with AWS OpsWorks is to work through the AWS OpsWorks Getting Started Guide (Linux | Windows), part of our technical documentation. Within a few minutes, you will be able to deploy and use your application.

Application Configuration and Management

Q: What elements of my application can I control when using AWS OpsWorks?

An AWS OpsWorks stack defines the configuration of your entire application: the load balancers, server software, database, etc. You control every part of the stack by building layers that define the software packages deployed to your instances and other configuration details such as Elastic IPs and security groups. You can also deploy your software onto layers by identifying the repository and optionally using Chef recipes to automate everything Chef can do, such as creating directories and users, configuring databases, etc. You can use OpsWorks' built-in automation to scale your application and automatically recover from instance failures. You can control who can view and manage the resources that are used by your application, including ssh access to the instances that your application uses.

Q: What software versioning and revision control systems does AWS OpsWorks support?

AWS OpsWorks can retrieve the code you want to deploy from common version control systems like Git and Subversion as well as HTTP and private or public S3 bundles. For example, you can deploy a specific version of your application by adding the version or branch from your Git repository into your OpsWorks app definition. You can also use Chef recipes to deploy your apps from anywhere you like using rsync or scp.

Q: What operating systems does AWS OpsWorks support?

AWS OpsWorks currently supports Amazon Linux, Ubuntu 12.04 LTS, Ubuntu 14.04 LTS, and Windows Server 2012 R2.

Q: Does AWS OpsWorks support Microsoft Windows?

Yes. AWS OpsWorks supports Windows Server 2012 R2.

Q: Can I use AWS OpsWorks to deploy applications that are highly available?

Yes. If your application supports horizontal scaling, you can create instances in multiple availability zones, and your load balancer will route traffic among your instances. If any instance fails, OpsWorks' auto healing can replace it. If your application uses other techniques to achieve availability goals, such as a database with an active and a passive node, you can use Chef recipes to configure it.

Q: How do I model my application in AWS OpsWorks?

AWS OpsWorks provides three concepts to model your application:

A Stack is the highest-level management unit. A stack contains the set of Amazon EC2 instances and instance blueprints, called layers, used to launch and manage these instances. Applications, user permissions, and other resources are scoped and controlled in the context of the Stack. For example, you might create a stack for your development web application that includes a front-end load balancer, the PHP servers, your PHP apps, and the MySQL database. You can also create a stack for your production web application with a similar configuration by cloning the development stack.

A Layer is a blueprint for how to setup and configure an instance and related resources such as volumes, Elastic IPs, and can automatically take care of infrastructure configuration like SSL settings. You can also define the software configuration for each layer, including installation scripts, initialization tasks, and packages. For example, if you use a Ruby layer, OpsWorks can install not only Rails, but also all the Gems your application requires. Layers also include lifecycle events that let you automate configuration actions in response to changes in an instance's status (see "What are Lifecycle Events" for details) using Chef recipes (see "What is Chef and how does AWS OpsWorks use it" for details). Layers can include time- or load-based auto scaling to handle demand peaks without manual interaction.

An App is software downloaded from a repository (e.g., Git, S3) and deployed to a layer. You can use the deploy lifecycle event to automate configuration steps such as connecting your application to a database. OpsWorks supports the ability to deploy multiple apps per stack and per layer.

Q: What is Chef and how does AWS OpsWorks use it?

Chef is an open source framework sponsored by Chef Software, Inc. that automates how applications are configured, deployed, and managed through the use of code. AWS OpsWorks uses Chef recipes to deploy and configure software components on Amazon EC2 instances. Chef has a rich ecosystem with hundreds of Cookbooks that can be used with AWS OpsWorks such as PostgreSQL, Nginx, and Solr.

Q: What are lifecycle events?

AWS OpsWorks creates events that correspond to lifecycle stages. These events can be used to trigger Chef recipes on each instance to perform specific configuration tasks. OpsWorks leverages Chef recipes to perform basic management for each event based on the type of layer. You can also create custom recipes to script any configuration change that your application needs for a specific lifecycle event. The following lifecycle events are supported:

Setup is sent to the instance when it is instantiated or successfully booted. For example, you could trigger a Chef recipe for a Rails application server that installs dependencies like Apache, Ruby, Passenger, and Ruby on Rails.

Configure notifies all instances whenever the state of the stack changes. For example, when a new instance is successfully added to an application server layer, the configure event triggers a Chef recipe that updates the OpsWorks Load Balancer layer configuration to reflect the added application server instance.

Deploy is triggered whenever an application is deployed. For example, you could trigger a Chef recipe for a Rails application server that executes the tasks needed to check out and download your application and tells Passenger to reload it.

Undeploy is sent when you delete an application. For example, the undeploy event can trigger a custom Chef recipe that specifies any cleanup steps that need to be run, such as deleting database tables.

Shutdown is sent to an instance 45 seconds before actually stopping the instance. For example, the shutdown event can trigger a custom Chef recipe that shuts down services.

Q: Does AWS OpsWorks support existing Chef cookbooks?

Yes. You can use existing Chef recipes. For more information, see thedocumentation.

Q: How do I create Chef cookbooks and recipes?

Probably the easiest way to get started is to use existing Chef recipes. There is a rich ecosystem of public repositories containing Chef cookbooks with recipes that can run with little to no modification. The OpsWorks Getting Started Guide also includes an example Chef recipe and describes how it works.

Q: Can I use my own AMIs?

Yes. You can use your own AMIs or customize the AMIs OpsWorks supports using Chef scripts to install agents and other software that you require. Using your own Windows AMIs is not currently supported.

Q: Can I use Amazon EC2 user data to customize instance setup?

No. Instance setup is done exclusively through Chef recipes.

Q: What load balancing options does AWS OpsWorks support?

OpsWorks supports Elastic Load Balancing, HAProxy using community Chef recipes, or any load balancer you choose to install on your EC2 instances using a custom layer and Chef recipes. This gives you rich customization options and fine-grained control of your application’s load balancer.

Q: What automatic instance scaling options does AWS OpsWorks support?

OpsWorks supports automatic time and load-based instance scaling to adapt the number of running instances to match your load. With load-based auto scaling, you can set thresholds for CPU, memory, or load to define when additional instances will be started. Once the load spike is gone and your down-scaling thresholds are met, OpsWorks will shut down the additional instances. With time-based auto scaling, you can define at what time of the day instances will be started and stopped. The instances in your auto scaling pool can vary in size, letting you scale gradually or quickly, and can be configured for multiple availability zones to improve reliability. OpsWorks does not support EC2 Auto Scaling at this time.

Q: What monitoring and alarming options does AWS OpsWorks support?

OpsWorks sends all of your instance and volume metrics to CloudWatch, making it easy to view graphs and set alarms to help you troubleshoot and take automated action based on the state of your resources. You can also see the thirteen 1-minute metrics OpsWorks collects (including CPU, memory, and load) from your instances in the OpsWorks Monitoring view.

Q: What databases does AWS OpsWorks support?

You can use AWS services such as Amazon RDS or use Chef recipes to install databases such as MySQL, Cassandra, or MongoDB. This gives you rich customization options and fine-grained control of your application’s database.

Q: Does AWS OpsWorks support tags?

OpsWorks automatically tags all resources with the name of the stack and layer that they are associated with. You can use these tags with Cost Allocation Reports to organize and track your AWS costs using tagging. To learn more about Cost Allocation and tagging, please visit AWS Account Billing.

Security

Q: Can I run my application inside an Amazon Virtual Private Cloud (VPC)?

Yes. See the OpsWorks documentation for more information.

Q: Is it possible to use AWS Identity & Access Management (IAM) with AWS OpsWorks?

Yes, OpsWorks supports IAM users, permissions, and roles. You can designate permissions by user, including view, deploy, and manage. You can also specify which users can ssh directly into instances. OpsWorks support for IAM roles lets you give a user access to OpsWorks without having to give access to dependent services like EC2. For example, you can explicitly deny a user the ability to perform EC2 actions, but the user can still control EC2 instances through OpsWorks if they have OpsWorks permissions to deploy or manage stack resources. This lets you prevent an OpsWorks user from inadvertently stopping an instance from the EC2 console.

Q: Can I manage what ports are open on my instances?

AWS OpsWorks provides a standard set of built-in security groups — one for each layer — which are associated with layers by default. The stack's Use OpsWorks security groups setting allows you to instead provide your own custom security groups. With this option, you must create appropriate EC2 security groups and associate a security group with each layer that you create. However, you can still manually associate a built-in security group with a layer on creation; custom security groups are required only for those layers that need custom settings. For more information on security groups, see Amazon EC2 Security Groups. Note that OpsWorks requires connectivity outbound from the EC2 instance on port 443 to configure your instance.

Q: What does AWS OpsWorks run on the instance?

OpsWorks uses an agent on the instance to perform configuration tasks and provide heartbeat health status. The agent runs as an unprivileged user on the operating system. Every instance also has a user that is used for deployments. This user doesn't have any login rights or access rights apart from deployment.

Q: Where can I find more information about security and running applications on AWS?

For more information about AWS security please refer to our Amazon Web Services: Overview of Security Processes document and visit our Security Center.

Billing

Q: How much does AWS OpsWorks cost?

You pay for on-premises servers supported by AWS OpsWorks by the hour; there are no minimum fees and no upfront commitments. The pricing for each on-premises server on which you install the OpsWorks agent is $0.02 per hour.

There is no additional charge for Amazon EC2 instances supported by AWS OpsWorks. You pay for AWS resources (e.g. EC2 instances, EBS volumes, Elastic IP addresses) created using OpsWorks in the same manner as if you created them manually. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments.

Q: How do I check how many AWS resources have been used by my application and access my bill?

You can view your charges for the current billing period at any time on the Amazon Web Services web site by logging into your Amazon Web Services account and clicking Account Activity under Your Web Services Account. OpsWorks automatically tags all resources with the name of the stack and layer that they are associated with. You can use these tags with Cost Allocation Reports to organize and track your AWS costs using tagging.

Amazon Service Catalog FAQ

General

Q: What is AWS Service Catalog?

AWS Service Catalog allows IT administrators to create, manage, and distribute catalogs of approved products to end users, who can then access the products they need in a personalized portal. Administrators can control which users have access to each product to enforce compliance with organizational business policies. Administrators can also setup adopted roles so that End users only require IAM access to AWS Service Catalog in order to deploy approved resources. AWS Service Catalog allows your organization to benefit from increased agility and reduced costs because end users can find and launch only the products they need from a catalog that you control.

Q: Who should use AWS Service Catalog?

AWS Service Catalog was developed for organizations, IT teams, and managed service providers (MSPs) that need to centralize policies. It allows IT administrators to vend and manage AWS resource and services. For large organizations, it provides a standard method of provisioning cloud resources for thousands of users. It is also suitable for small teams, where front-line development managers can provide and maintain a standard dev/test environment.

Q: How do I get started with AWS Service Catalog?

In the AWS Management Console, choose AWS Service Catalog in Management Tools. In the AWS Service Catalog console, administrators can create portfolios, add products, and grant users permissions to use them with just a few clicks. End users logged into the AWS Service Catalog console can see and launch the products that administers have created for them.

Q: What can end users to do with AWS Service Catalog that they could not do before?

End users have a simple portal in which to discover and launch products that comply with organizational policies and budget constraints.

Q: What is a portfolio?

A portfolio is a collection of products, with configuration information that determines who can use those products and how they can use them. Administrators can create a customized portfolio for each type of user in an organization and selectively grant access to the appropriate portfolio. When an administrator adds a new version of a product to a portfolio, that version is automatically available to all current portfolio users. The same product can be included in multiple portfolios. Administrators also can share portfolios with other AWS accounts and allow the administrators of those accounts to extend the portfolios by applying additional constraints. By using portfolios, permissions, sharing, and constraints, administrators can ensure that users are launching products that are configured properly for the organization’s needs.

Q: What is a product?

A product is a service or application for end users. A catalog is a collection of products that the administrator creates, adds to portfolios, and provides updates for using AWS Service Catalog. A product can comprise one or more AWS resources, such as Amazon Elastic Compute Cloud (Amazon EC2) instances, storage volumes, databases, monitoring configurations, and networking components. It can be a single compute instance running AWS Linux, a fully configured multitier web application running in its own environment, or anything in between.

Administrators distribute products to end users in portfolios. Administrators create catalogs of products by importing AWS CloudFormation templates. These templates define the AWS resources that the product needs to work, the relationships between components, and the parameters that the end user chooses when launching the product to configure security groups, create key pairs, and perform other customizations.

An end user with access to a portfolio can use the AWS Management Console to find a standard dev/test environment product, for example, in the form of an AWS CloudFormation template, then manage the resulting resources using the AWS CloudFormation console. For information about creating a product, see “How do I create a product?” in the Administrator FAQ.

Q: Is AWS Service Catalog a regionalized service?

Yes. AWS Service Catalog is fully regionalized, so you can control the regions in which data is stored. Portfolios and products are a regional construct which will need to be created per region and are only visible/usable on the regions in which they were created.

Q: In which regions is AWS Service Catalog available?

We support the US East (N. Virginia), US West (Oregon), EU (Ireland) and Asia Pacific (Tokyo and Singapore) regions.

Q: Are APIs available? Can I use the CLI to access AWS Service Catalog?

Yes, APIs are available but only for the End User Actions and these are enabled through the CLI. Administrator actions such as portfolio and product management are not available through API at this time. You can find more information in the AWS Service Catalog documentation or download the latest AWS SDK or CLI.

IT Administrator

Q: How do I create a portfolio?

You create portfolios in the AWS Service Catalog console. For each portfolio, you specify the name, a description, and owner.

Q: How do I create a product?

To create a product, you first create an AWS CloudFormation template by using an existing AWS CloudFormation template or creating a custom template. Next, you use the AWS Service Catalog console to upload the template and create the product. When creating products, you can provide additional information for the product listing, including a detailed product description, version information, support information, and tags.

Q: Why would I use tags with a portfolio?

Tags are useful for identifying and categorizing AWS resources that are provisioned by end users. You can also use tags in AWS Identity and Access Management (IAM) policies to allow or deny access to IAM users, groups, and roles or to restrict operations that can be performed by IAM users, groups, and roles. When you add tags to your portfolio, the tags are applied to all instances of resources provisioned from products in the portfolio.

Q: How do I make a portfolio available to my users?

You publish portfolios that you've created or that have been shared with you to make them available to IAM users in the AWS account. To publish a portfolio, you add IAM users, groups, or roles to the portfolio from the AWS Service Catalog console by navigating to the portfolio details page. When you add users to a portfolio, they can browse and launch any of the products in the portfolio. Typically, you create multiple portfolios with different products and access permissions customized for specific types of end users. For example, a portfolio for a development team will likely contain different products from a portfolio targeted at the sales and marketing team. A single product can be published to multiple portfolios with different access permissions and provisioning policies.

Q: Can I share my portfolio with other AWS accounts?

Yes. You can share your portfolios with users in one or more other AWS accounts. When you share your portfolio with other AWS accounts, you retain ownership and control of the portfolio. Only you can make changes, such as adding new products or updating products. You, and only you, can also "unshare" your portfolio at any time. Any products, or stacks, currently in use will continue to run until the stack owner decides to terminate them.

To share your portfolio, you specify the account ID you want to share with, and then send the Amazon Resource Number (ARN) of the portfolio to that account. The owner of that account can create a link to this shared portfolio, and then assign IAM users from that account to the portfolio. To help end users with discovery, you can curate a directory of portfolios.

Q: Can I customize the experience for end users when they use a product?

Yes. You can tailor a product's user experience for specific end users. The AWS CloudFormation template contains input parameters that drive the user experience. You can define business-level input parameters (such as "How many users do you need to support?" or "Are you going to store PII data?") or infrastructure-level input parameters (such as "Which Amazon EC2 instance type?") depending on the user. When the AWS CloudFormation template is deployed, the user is asked these questions and can select from a constrained list of answers for each question. Depending on the answers, the template may be deployed using different Amazon Elastic Compute Cloud (EC2) instances and different AWS resources.

Q: Can I create a product from an existing Amazon EC2 AMI?

Yes. You can use an existing Amazon EC2 AMI to create a product by wrapping it in an AWS CloudFormation template.

Q: Can I use products from the AWS Marketplace?

Yes. You can subscribe to a product in the AWS Marketplace and use the Amazon EC2 AMI for the product to create an AWS Service Catalog product. To do that, you wrap the subscribed product in an AWS CloudFormation template. For more details on how to package your AWS Marketplace products, please see click here.

Q: How do I control access to portfolios and products?

To control access to portfolios and products, you assign IAM users, groups, or roles on the Portfolio details page. Providing access allows users to see the products that are available to them in the AWS Service Catalog console.

Q: Can I provide a new version of a product?

Yes. You can create new product versions in the same way you create new products. When a new version of a product is published to a portfolio, end users can choose to launch the new version. They can also choose to update their running stacks to this new version. AWS Service Catalog does not automatically update products that are in use when an update becomes available.

Q: Can I provide a product and retain full control over the associated AWS resources?

Yes. You have full control over the AWS accounts and roles used to provision products. To provision AWS resources, you can use either the user's IAM access permissions or your pre-defined IAM role. To retain full control over the AWS resources, you specify a specific IAM role at the product level. AWS Service Catalog uses the role to provision the resources in the stack.

Q: Can I restrict the AWS resources that users can provision?

Yes. You can define rules that limit the parameter values that a user enters when launching a product. These rules are called template constraints because they constrain how the AWS CloudFormation template for the product is deployed. You use a simple editor to create template constraints, and you apply them to individual products.

AWS Service Catalog applies constraints when provisioning a new product or updating a product that is already in use. It always applies the most restrictive constraint among all constraints applied to the portfolio and the product. For example, consider a scenario where the product allows all EC2 instances to be launched and the portfolio has two constraints: one that allows all non-GPU type EC2 instances to be launched and one that allows only t1.micro and m1.small EC2 instances to be launched. For this example, AWS Service Catalog applies the second, more restrictive constraint (t1.micro and m1.small).

End User

Q: How do I find out which products are available?

You can see which products are available by logging in to the AWS Service Catalog console and searching the portal for products that meet your needs, or you can navigate to the full product list page. You can sort to find the product that you want.

For each product, you can view a Product details page that displays information about the product, including the version, whether a newer version of the product is available, a description, support information, and tags associated with the product. The Product details page might also indicate whether the product will be provisioned using your access permissions (Self) or an administrator-specified role (role-arn). &nbsp;

Q: How do I deploy a product?

When you find a product that meets your requirements in the portal, choose Launch. You will be guided through a series of questions about how you plan to use the product. The questions might be about your business needs or your infrastructure requirements (such as “Which EC2 instance type?”). When you have provided the required information, you’ll see the product in the AWS Service Catalog console. While the product is being provisioned, you will see that it is “in progress.” After provisioning is complete, you will see “complete” and information, such as endpoints or Amazon Resource Names (ARNs), that you can use to access the product.

Q: Can I see which products I am using?

Yes. You can see which products you are using in the AWS Service Catalog console. You can see all of the stacks that are in use, along with the version of the product used to create them.

How do I update my products when a new version becomes available?

When a new version of a product is published, you can use the Update Stack command to use that version. If you are currently using a product for which there is an update, it continues to run until you close it, at which point you can choose to use the new version.

Q: How do I monitor the health of my products?

You can see the products that you are using and their health state in the AWS Service Catalog console.

AWS Trusted Advisor

Service Limits Check Questions

Q. What service limits do you check?

The following table shows the limits that Trusted Advisor checks.

Service	Limits
Amazon Elastic Compute Cloud (Amazon EC2)	Elastic IP addresses (EIPs) On-Demand instances Reserved Instances - purchase limit (monthly)
Amazon Elastic Block Store (Amazon EBS)	Active volumes Active snapshots Provisioned IOPS Provisioned IOPS (SSD) volume storage (GiB) General Purpose (SSD) volume storage (GiB) Magnetic volume storage (GiB)
Amazon Relational Database Service (Amazon RDS)	DB instances DB parameter groups DB security groups DB snapshots per user Max auths per security group Read replicas per master Storage quota (GiB)
Amazon Simple Email Service (Amazon SES)	Daily sending quota
Amazon Virtual Private Cloud (Amazon VPC)	Elastic IP addresses (EIPs) Internet gateways VPCs
Auto Scaling	Auto Scaling groups Launch configurations
Elastic Load Balancing (ELB)	Active load balancers
Identity and Access Management (IAM)	Users Groups Roles Instance profiles Server certificates

We are constantly working on including more services in this check. Yourfeedback is really helpful to us.

Q. What are the default service limits?

For a list of the default service limits and instructions for requesting increases, seeAWS Service Limits.

Q. How can I get the Service Limit data with command-line tools?

This AWS Command Line Interface command displays all of the resources in your account that Trusted Advisor has flagged as approaching or reaching the limit, sorted by limit name.

```
aws support describe-trusted-advisor-check-result --language en --c
heck-id eW7HH0l7J9 --query 'result.sort_by(flaggedResources[?status
!=`ok`],&metadata[2])[].metadata' --output table
```

Here is sample output from the command. The fourth column shows the limit amount, and the fifth column shows the current usage.

The filtering and sorting in the --query option requires AWS CLI version 1.3.0 or later. If you use the Windows Command Prompt window, enclose the --query parameter in double quotes (") instead of single quotes (').

To display all resources, replace [?status!=`ok`] with [ ]. To sort by a different column, change the number in &metadata[2]. For example, to sort by region, use &metadata[0].

```
-----------------------------------------------------------------
------------
|                      DescribeTrustedAdvisorCheckResult
         |
+-----------+------+-----------------------------------+----+----
+----------+
|  us-east-1|  ELB |  Active Load Balancers            | 20 | 18 |
Yellow  |
|  us-east-1|  EC2 |  Elastic IP addresses (EIPs)      |  5 |  5 |
Yellow  |
|  us-west-1|  EC2 |  Elastic IP addresses (EIPs)      |  5 |  4 |
Yellow  |
|  us-east-1|  EC2 |  VPC Elastic IP addresses (EIPs)  |  5 |  4 |
Yellow  |
|  us-west-2|  EC2 |  VPC Elastic IP addresses (EIPs)  |  5 |  5 |
Yellow  |
+----------+-----+------------------------------+----+----
+----------+
```

Reserved Instance Optimization Check Questions

Q. What data set are you using to make a Reserved Instance recommendation?

We calculate the recommendation based on the usage in the last completed calendar month. For example, if it is the 25th of April, the recommendation is based on data from March 1 to March 31.

Q. Does the recommendation consider volume discounts?

No, the recommendation uses standard pricing. Actual results may vary on discounted pricing tiers. We recommend contacting your sales representative by completing the AWS Sales & Business Development form to review a more detailed optimization plan if you are receiving volume discounts.

Q. I just purchased a new Reserved Instance. Why isn't it showing up in the recommendation?

New Reserved Instance purchases are updated on a daily basis. Refresh the check 24 hours after you make your purchase to see the new recommendation. Also note that the check does not include third-party Reserved Instances purchased from the Reserved Instance Marketplace.

Q. How do you calculate the optimized number of Reserved Instances?

Our system analyzes the hourly usage history during the previous calendar month across all consolidated accounts. The system calculates the number of running instances in each Availability Zone and for each type of instance. An hourly cost is determined by aggregating the cost of all instances that ran the previous month, whether they ran as On-Demand or as a Reserved Instance. In addition to the hourly usage charges, the system calculates a fixed charge by amortizing the one-time upfront fees for each Reserved Instance already purchased.

By adding the aggregated hourly charges and the amortized upfront fees, the system is able to determine your baseline cost for the month. The system then incorporates the hourly and amortized upfront costs for additional Partial Upfront Reserved Instances, and the amortized upfront costs of any existing Reserved Instances into the calculation. Given the baseline cost based on the previous usage, and the costs for adding additional Partial Upfront Reserved Instances, the system uses a simple gradient descent algorithm to determine the number of Partial Upfront Reserved Instances that would result in the lowest overall cost.

Q. How do you amortize the cost of existing Reserved Instances?

The upfront fee for each Reserved Instance is amortized over the period of the Reserved Instance. In simple terms, if the upfront fee was $1200, and the term length was one year, the system will divide $1200 by 12 months, resulting in a cost of $100 per month.

Q. I have many accounts, and the Availability Zones are different for each one. How do you account for that?

We normalize all Availability Zones across all Consolidated Billing accounts and reflect the values using the primary payer account mapping.

Q. Do you include other Reserved Instance types in the recommendation?

Only Partial Upfront Reserved Instances are recommended by this check. However, hourly usage charges and amortized upfront fees for other Reserved Instance types are included in the calculation.

Q. Why are there separate sections for 1 year and 3 year Reserved Instances?

Customers have a choice between buying 1 year and 3 year term Reserved Instances from AWS. This check assumes you will purchase Reserved Instances for either 1 year or 3 year terms, not both. As a result, recommendations for purchasing additional 1 year or 3 year term Reserved Instances are not additive across both term lengths, so recommendations are called out separately.

To illustrate: In a recommendation for three additional 1 year Reserved Instances or four additional 3 year Reserved Instances, we are recommending the purchase of three or four Reserved Instances respectively, not a total of seven additional Reserved Instances.

Q. Are all instance types included in the recommendation?

Recommendations are available for Amazon Linux/UNIX and Windows Reserved Instances. The calculation excludes usage and recommendations for Red Hat Enterprise Linux, SUSE Linux Enterprise, Amazon RDS, Amazon ElastiCache, and others.

Q. I use Spot instances. Do you include Spot rates in the calculation?

Due to the variability of the Spot instance market, the system uses on-demand rates when calculating the optimized number of Reserved Instances.

Q. I have third-party Reserved Instances from the Reserved Instance Marketplace. Do you include those in the results?

No; only Reserved Instances offered directly by AWS are included. If you haveReserved Instances from third-party sellers, those Reserved Instances are not accounted for by this check.

Q. Does the recommendation include any money I make if I sell my existing Reserved Instances to purchase the recommended Partial Upfront Reserved Instances?

The system does not include any money that could result from the sale of existing Reserved Instances when calculating the optimal number of Partial Upfront Reserved Instances.

AWS Identity and Access Management FAQ

IAM Home

General

Q: What is AWS Identity and Access Management (IAM)?
You can use AWS IAM to securely control individual and group access to your AWS resources. You can create and manage user identities ("IAM users") and grant permissions for those IAM users to access your resources. You can also grant permissions for users outside of AWS (federated users).

Q: How do I get started with IAM?

After you've signed up for AWS, you can create users and groups and assign them permissions to access your AWS resources. You can use the IAM console (for web-based access), the AWS Command Line Interface (CLI, for command line access), or the API or SDKs (for programmatic access). To grant permissions, you create policy documents that you attach to users, groups, or other entities. See the video, Getting Started with AWS IAM.

Q: What problems does IAM solve?

IAM makes it easy to provide multiple users secure access to your AWS resources. IAM enables you to:

- Manage IAM users and their access: You can create users in AWS's identity management system, assign users individual security credentials (such as access keys, passwords, multi-factor authentication devices), or request temporary security credentials to provide users access to AWS services and resources. You can specify permissions to control which operations a user can perform.
- Manage access for federated users: You can request security credentials with configurable expirations for users who you manage in your corporate directory, allowing you to provide your employees and applications secure access to resources in your AWS account without creating an IAM user account for them. You specify the permissions for these security credentials to control which operations a user can perform.

Q: Who can use IAM?

Any AWS customer can use IAM. The service is offered at no additional charge. You will be charged only for the use of other AWS services by your users.

Q: What is a user?

A user is a unique identity recognized by AWS services and applications. Similar to a login user in an operating system like Windows or UNIX, a user has a unique name and can identify itself using familiar security credentials such as a password or access key. A user can be an individual, system, or application requiring access to AWS services. IAM supports users (referred to as "IAM users") managed in AWS's identity management system, and it also enables you to grant access to AWS resources for users managed outside of AWS in your corporate directory (referred to as "federated users").

Q: What can a user do?

A user can place requests to web services such as Amazon S3 and Amazon EC2. A user's ability to access web service APIs is under the control and responsibility of the AWS account under which it is defined. You can permit a user to access any or all of the AWS services that have been integrated with IAM and to which the AWS account has subscribed. If permitted, a user has access to all of the resources under the AWS account. In addition, if the AWS account has access to resources from a different AWS account, its users may be able to access data under those AWS accounts. Any AWS resources created by a user are under control of and paid for by its AWS account. A user cannot independently subscribe to AWS services or control resources.

Q: How do users call AWS services?
Users can make requests to AWS services using security credentials. Explicit permissions govern a user's ability to call AWS services. By default, users have no ability to call service APIs on behalf of the account.

Q: How do I get started with IAM?
To start using IAM, you must subscribe to at least one of the AWS services that is integrated with IAM. You then can create and manage users, groups, and permissions via IAM APIs, the AWS CLI, or the IAM console, which gives you a point-and-click, web-based interface. You can also use the AWS Policy Generator to create policies.

IAM User Management

Q: How are IAM users managed?
IAM supports multiple methods to:

- Create, delete, and list IAM users.
- Manage group membership.
- Manage user security credentials.
- Assign permissions.

You can create and manage users, groups, and permissions via IAM APIs, theAWS CLI, or the IAM console, which gives you a point-and-click, web-based interface. You can also use the AWS Policy Generator and AWS Policy Simulator to create and test policies.

Q: What is a group?
A group is a collection of IAM users. Manage group membership as a simple list:

- Add users to or remove them from a group.
- A user can belong to multiple groups.
- Groups cannot belong to other groups.
- Groups can be granted permissions using access control policies. This makes it easier to manage permissions for a collection of users, rather than having to manage permissions for each individual user.

- Groups do not have security credentials, and cannot access web services directly; they exist solely to make it easier to manage user permissions. For details, see Working with Groups and Users.

Q: What kinds of security credentials can IAM users have?
IAM users can have any combination of credentials that AWS supports, such as an AWS access key, X.509 certificate, SSH key, password for web app logins, or an MFA device. This allows users to interact with AWS in any manner that makes sense for them. An employee might have both an AWS access key and a password; a software system might have only an AWS access key to make programmatic calls; IAM users might have a private SSH key to access AWS CodeCommit repositories; and an outside contractor might have only an X.509 certificate to use the EC2 command-line interface. For details, see Temporary Security Credentials in the IAM documentation.

Q: Which AWS services support IAM users?
You can find the complete list of AWS services that support IAM users in the AWS Services That Work with IAM section of the IAM documentation. AWS plans to add support for other services over time.

Q: Can I enable and disable user access?
Yes. You can enable and disable an IAM user's access keys via the IAM APIs, AWS CLI, or IAM console. If you disable the access keys, the user cannot programmatically access AWS services.

Q: Who is able to manage users for an AWS account?
The AWS account holder can manage users, groups, security credentials, and permissions. In addition, you may grant permissions to individual users to place calls to IAM APIs in order to manage other users. For example, an administrator user may be created to manage users for a corporation—a recommended practice. When you grant a user permission to manage other users, they can do this via the IAM APIs, AWS CLI, or IAM console.

Q: Can I structure a collection of users in a hierarchical way, such as in LDAP?
Yes. You can organize users and groups under paths, similar to object paths in Amazon S3—for example /mycompany/division/project/joe.

Q: Can I define users regionally?
Not initially. Users are global entities, like an AWS account is today. No region is required to be specified when you define user permissions. Users can use AWS services in any geographic region.

Q: How are MFA devices configured for IAM users?
You (the AWS account holder) can order multiple MFA devices. You can then assign these devices to individual IAM users via the IAM APIs, AWS CLI, or IAM console.

Q: What kind of key rotation is supported for IAM users?

User access keys and X.509 certificates can be rotated just as they are for an AWS account's root access identifiers. You can manage and rotate programmatically a user's access keys and X.509 certificates via the IAM APIs, AWS CLI, or IAM console.

Q: Can IAM users have individual EC2 SSH keys?
Not in the initial release. IAM does not affect EC2 SSH keys or Windows RDP certificates. This means that although each user has separate credentials for accessing web service APIs, they must share SSH keys that are common across the AWS account under which users have been defined.

Q: Where can I use my SSH keys?

Currently, IAM users can use their SSH keys only with AWS CodeCommit to access their repositories.

Q: Do IAM user names have to be email addresses?
No, but they can be. User names are just ASCII strings that are unique within a given AWS account. You can assign names using any naming convention you choose, including email addresses.

Q: Which character sets can I use for IAM user names?
You can only use ASCII characters for IAM entities.

Q: Are user attributes other than user name supported?
Not at this time.

Q: How are user passwords set?
You can set an initial password for an IAM user via the IAM console, AWS CLI, or IAM APIs. User passwords never appear in clear text after the initial provisioning, and are never displayed or returned via an API call. IAM users can manage their passwords via the **My Password** page in the IAM console. Users access this page by selecting the **Security Credentials** option from the drop-down list in the upper right corner of the AWS Management Console.

Q: Can I define a password policy for my user's passwords?
Yes, you can enforce strong passwords by requiring minimum length or at least one number. You can also enforce automatic password expiration, prevent re-use of old passwords, and require a password reset upon the next AWS sign-in. For details, see Setting an Account Policy Password for IAM Users.

Q: Can I set usage quotas on IAM users?
No. All limits are on the AWS account as a whole. For example, if your AWS account has a limit of 20 Amazon EC2 instances, IAM users with EC2 permissions can start instances up to the limit. You cannot limit what an individual user can do.

IAM Role Management

Q: What is an IAM role?
An IAM role is an IAM entity that defines a set of permissions for making AWS service requests. IAM roles are not associated with a specific user or group. Instead, trusted entities *assume* roles, such as IAM users, applications, or AWS services such as EC2.

Q: What problems do IAM roles solve?
IAM roles allow you to delegate access with defined permissions to trusted entities without having to share long-term access keys. You can use IAM roles to delegate access to IAM users managed within your account, to IAM users under a different AWS account, or to an AWS service such as EC2.

Q: How do I get started with IAM roles?
You create a role in a way similar to how you create a user—name the role and attach a policy to it. For details, see Creating IAM Roles.

Q: How do I assume an IAM role?
You assume an IAM role by calling the AWS Security Token Service (STS) AssumeRole APIs (in other words, AssumeRole, AssumeRoleWithWebIdentity, and AssumeRoleWithSAML). These APIs return a set of temporary security credentials that applications can then use to sign requests to AWS service APIs.

Q: How many IAM roles can I assume?
There is no limit to the number of IAM roles you can assume, but you can only act as one IAM role when making requests to AWS services.

Q: Who can use IAM roles?
Any AWS customer can use IAM roles.

Q: How much do IAM roles cost?
IAM roles are free of charge. You will continue to pay for any resources a role in your AWS account consumes.

Q: How are IAM roles managed?
You can create and manage IAM roles via the IAM APIs, AWS CLI, or IAM console, which gives you a point-and-click, web-based interface.

Q: What is the difference between an IAM role and an IAM user?
An IAM user has permanent long-term credentials and is used to directly interact with AWS services. An IAM role does not have any credentials and cannot make direct requests to AWS services. IAM roles are meant to be assumed by authorized entities, such as IAM users, applications, or an AWS service such as EC2.

Q: When should I use an IAM user, IAM group, or IAM role?

An IAM user has permanent long-term credentials and is used to directly interact with AWS services. An IAM group is primarily a management convenience to manage the same set of permissions for a set of IAM users. An IAM role is an AWS Identity and Access Management (IAM) entity with permissions to make AWS service requests. IAM roles cannot make direct requests to AWS services; they are meant to be assumed by authorized entities, such as IAM users, applications, or AWS services such as EC2. Use IAM roles to delegate access within or between AWS accounts.

Q: Can I add an IAM role to an IAM group?
Not at this time.

Q: How many policies can I attach to an IAM role?

For inline policies: You can add as many inline policies as you want to a user, role, or group, but the total aggregate policy size (the sum size of all inline policies) per entity cannot exceed the following limits:

- User policy size cannot exceed 2,048 characters.
- Role policy size cannot exceed 10,240 characters.
- Group policy size cannot exceed 5,120 characters.

For managed policies: You can add up to 10 managed policies to a user, role, or group. The size of each managed policy cannot exceed 5,120 characters.

Q: How many IAM roles can I create?
You are limited to 250 IAM roles under your AWS account. If you need more roles, submit the IAM limit increase request formwith your use case, and we will consider your request.

Q: To which services can I attach an IAM role make service calls?
Your application can make requests to all AWS services that support role sessions.

Q: What is IAM roles for EC2 instances?
IAM roles for EC2 instances enables your applications running on EC2 to make requests to AWS services such as Amazon S3, Amazon SQS, and Amazon SNS without you having to copy AWS access keys to every instance. For details, see IAM Roles for Amazon EC2.

Q: What are the features of IAM roles for EC2 instances?

IAM roles for EC2 instances provides the following features:

- AWS temporary security credentials to use when making requests from running EC2 instances to AWS services.
- Automatic rotation of the AWS temporary security credentials.
- Granular AWS service permissions for applications running on EC2 instances.

Q: What problem does IAM roles for EC2 instances solve?
IAM roles for EC2 instances simplifies management and deployment of AWS access keys to EC2 instances. Using this feature, you associate an IAM role with an instance. Then your EC2 instance provides the temporary security credentials to applications running on the instance, and the applications can use these credentials to make requests securely to the AWS service resources defined in the role.

Q: How do I get started with IAM roles for EC2 instances?
To understand how roles work with EC2 instances, you need to use the IAM console to create a role, launch an EC2 instance that uses that role, and then examine the running instance. You can examine the instance metadata to see how the role credentials are made available to an instance. You can also see how an application that runs on an instance can use the role. For more details, see How Do I Get Started?

Q: Can I use the same IAM role on multiple EC2 instances?
Yes.

Q: Can I change the IAM role on a running EC2 instance?
No. At this time, you cannot change the IAM role on a running EC2 instance. You can change the permissions on the IAM role associated with a running instance, and the updated permissions take effect almost immediately.

Q: Can I associate an IAM role with an already running EC2 instance?
No. You can only associate an IAM role while launching an EC2 instance.

Q: Can I associate an IAM role with an Auto Scaling group?

Yes. You can add an IAM role as an additional parameter in an Auto Scaling launch configuration and create an Auto Scaling group with that launch configuration. All EC2 instances launched in an Auto Scaling group that is associated with an IAM role are launched with the role as an input parameter. For more details, see What Is Auto Scaling? in the *Auto Scaling Developer Guide*.

Q: Can I associate more than one IAM role with an EC2 instance?
No. You can only associate one IAM role with an EC2 instance at this time.

Q: What happens if I delete an IAM role that is associated with a running EC2 instance?
Any application running on the instance that is using the role will be denied access immediately.

Q: Can I control which IAM roles an IAM user can associate with an EC2 instance?
Yes. For details, see Permissions Required for Using Roles with Amazon EC2.

Q: Which permissions are required to launch EC2 instances with an IAM role?
You must grant an IAM user two distinct permissions to successfully launch EC2 instances with roles:

- Permission to launch EC2 instances.
- Permission to associate an IAM role with EC2 instances.

For details, see Permissions Required for Using Roles with Amazon EC2.

Q: Who can access the access keys on an EC2 instance?
Any local user on the instance can access the access keys associated with the IAM role.

Q: How do I use the IAM role with my application on the EC2 instance?
If you develop your application with the AWS SDK, the AWS SDK automatically uses the AWS access keys that have been made available on the EC2 instance. If you are not using the AWS SDK, you can retrieve the access keys from the EC2 instance metadata service. For details, see Using an IAM Role to Grant Permissions to Applications Running on Amazon EC2 Instances.

Q: How do I rotate the temporary security credentials on the EC2 instance?
The AWS temporary security credentials associated with an IAM role are automatically rotated multiple times a day. New temporary security credentials are made available no later than five minutes before the existing temporary security credentials expire.

Q: Can I use IAM roles for EC2 instances with any instance type or Amazon Machine Image?
Yes. IAM roles for EC2 instances also work in Amazon Virtual Private Cloud (VPC), with spot and reserved instances.

Permissions

Q: How do permissions work?

Access control policies are attached to users, groups, and roles to assign permissions to AWS resources. By default, IAM users, groups, and roles have no permissions; users with sufficient permissions must use a policy to grant the desired permissions.

Q: How do I assign permissions using a policy?

To set permissions, you can create and attach policies using the AWS Management Console, the IAM API, or the AWS CLI. Users who have been granted the necessary permissions can create policies and assign them to IAM users, groups, and roles.

Q: What are managed policies?

Managed policies are IAM resources that express permissions using the IAM policy language. You can create, edit, and manage separately from the IAM users, groups, and roles to which they are attached. After you attach a managed policy to multiple IAM users, groups, or roles, you can update that policy in one place and the permissions automatically extend to all attached entities. Managed policies are managed either by you (these are called customer managed policies) or by AWS (these are called AWS managed policies). For more information about managed policies, see Managed Policies and Inline Policies.

Q: How do I assign commonly used permissions?

AWS provides a set of commonly used permissions that you can attach to IAM users, groups, and roles in your account. These are called AWS managed policies. One example is read-only access for Amazon S3. When AWS updates these policies, the permissions are applied automatically to the users, groups, and roles to which the policy is attached. AWS managed policies automatically appear in the **Policies** section of the IAM console. When you assign permissions, you can use an AWS managed policy or you can create your own customer managed policy. Create a new policy based on an existing AWS managed policy, or define your own.

Q: How do group-based permissions work?

Use IAM groups to assign the same set of permissions to multiple IAM users. A user can also have individual permissions assigned to them. The two ways to attach permissions to users work together to set overall permissions.

Q: What is the difference between assigning permissions using IAM groups and assigning permissions using managed policies?

Use IAM groups to collect IAM users and define common permissions for those users. Use managed policies to share permissions across IAM users, groups, and roles. For example, if you want a group of users to be able to launch an Amazon EC2 instance, and you also want the role on that instance to have the same permissions as the users in the group, you can create a managed policy and assign it to the group of users and the role on the Amazon EC2 instance.

Q: How are IAM policies evaluated in conjunction with Amazon S3, Amazon SQS, Amazon SNS, and AWS KMS resource-based policies?

IAM policies are evaluated together with the service’s resource-based policies. When a policy of any type grants access (without explicitly denying it), the action is allowed. For more information about the policy evaluation logic, see IAM Policy Evaluation Logic.

Q: Can I use a managed policy as a resource-based policy?

Managed policies can only be attached to IAM users, groups, or roles. You cannot use them as resource-based policies.

Q: How do I set granular permissions using policies?

Using policies, you can specify several layers of permission granularity. First, you can define specific AWS service actions you wish to allow or explicitly deny access to. Second, depending on the action, you can define specific AWS resources the actions can be performed on. Third, you can define conditions to specify when the policy is in effect (for example, if MFA is enabled or not).

Q: How can I easily remove unnecessary permissions?

To help you determine which permissions are needed, the IAM console now displaysservice last accessed data that shows the hour when an IAM entity (a user, group, or role) last accessed an AWS service. Knowing if and when an IAM entity last exercised a permission can help you remove unnecessary permissions and tighten your IAM policies with less effort.

Q: Can I grant permissions to access or change account-level information (for example, payment instrument, contact email address, and billing history)?

Yes, you can delegate the ability for an IAM user or a federated user to view AWS billing data and modify AWS account information. For more information about controlling access to your billing information, see Controlling Access.

Q: Who can create and manage access keys in an AWS account?

Only the AWS account owner can manage the access keys for the root account. The account owner and IAM users or roles that have been granted the necessary permissions can manage access keys for IAM users.

Q: Can I grant permissions to access AWS resources owned by another AWS account?
Yes. Using IAM roles, IAM users and federated users can access resources in another AWS account via the AWS Management Console, the AWS CLI, or the APIs. See Manage IAM Roles for more information.

Q: What does a policy look like?

The following policy grants access to add, update, and delete objects from a specific folder, example_folder, in a specific bucket, example_bucket.

```
{
  "Version":"2012-10-17",
  "Statement":[
    {
      "Effect":"Allow",
      "Action":[
        "s3:PutObject",
        "s3:GetObject",
        "s3:GetObjectVersion",
        "s3:DeleteObject",
        "s3:DeleteObjectVersion"
      ],
      "Resource":"arn:aws:s3:::example_bucket/example_folder/*"
    }
  ]
}
```

Policy Simulator

Q: What is the IAM policy simulator?
The IAM policy simulator is a tool to help you understand, test, and validate the effects of your access control policies.

Q: What can the policy simulator be used for?
You can use the policy simulator in several ways. You can test policy changes to ensure they have the desired effect before committing them to production. You can validate existing policies attached to users, groups, and roles to verify and troubleshoot permissions. You can also use the policy simulator to understand how IAM policies and resource-based policies work together to grant or deny access to AWS resources.

Q: Who can use the policy simulator?
The policy simulator is available to all AWS customers.

Q: How much does the policy simulator cost?
The policy simulator is available at no extra cost.

Q: How do I get started?
Go to https://policysim.aws.amazon.com, or click the link on the IAM console under “Additional Information.” Specify a new policy or choose an existing set of policies from a user, group, or role that you’d like to evaluate. Then select a set of actions from the list of AWS services, provide any required information to simulate the access request, and run the simulation to determine whether the policy allows or denies permissions to the selected actions and resources. To learn more about the IAM policy simulator, watch our Getting Started video or see the documentation.

Q: What kinds of policies does the IAM policy simulator support?
The policy simulator supports testing of newly entered policies and existing policies attached to users, groups, or roles. In addition, you can simulate whether resource-level policies grant access to a particular resource for Amazon S3 buckets, Amazon Glacier vaults, Amazon SNS topics, and Amazon SQS queues. These are included in the simulation when an Amazon Resource Name (ARN) is specified in the **Resource** field in **Simulation Settings** for a service that supports resource policies.

Q: If I change a policy in the policy simulator, do those changes persist in production?
No. To apply changes to production, copy the policy that you’ve modified in the policy simulator

and attach it to the desired IAM user, group, or role.

Q: Can I use the policy simulator programmatically?
Yes. You can use the policy simulator using the AWS SDKs or AWS CLI in addition to the policy simulator console. Use the iam:SimulatePrincipalPolicy API to programmatically test your existing IAM policies. To test the effects of new or updated policies that are not yet attached to a user, group, or role, call the iam:SimulateCustomPolicy API.

Signing In

Q: How does an IAM user sign in?
An IAM user must sign in using the account's sign-in URL, which will direct them to a page where they can enter their IAM user name and password. This sign-in URL is located on the dashboard of the IAM console. The AWS account's system administrator must communicate the sign-in URL to the IAM user.

Q: What is an AWS account alias?
The account alias is a name you define to make it more convenient to identify your account. You can create an alias using the IAM APIs, AWS Command Line Tools, or the IAM console. You can have one alias per AWS account.

Q: Do IAM users always have to use the direct link?
IAM users must use the account-specific URL for the first sign-in. Thereafter, the account-specific URL is stored as a cookie in the user's browser. This allows a user to return to http://aws.amazon.com and simply click **Sign In to the Console**. If the user clears his browser cookies or uses a different browser, he must use the account-specific URL once again.

Q: Which AWS sites can IAM users access?
IAM users can sign in to the following AWS sites:

- AWS Management Console
- AWS Forums
- AWS Support Center
- AWS Marketplace

Q: Can IAM users sign in to other Amazon.com properties with their credentials?
No. Users created with IAM are recognized only by AWS services and applications.

Q: Is there an authentication API to verify IAM user sign-ins?
No. There is no programmatic way to verify user sign-ins.

Q: Can users SSH to EC2 instances using their AWS user name and password?
No. User security credentials created with IAM are not supported for direct authentication to customer EC2 instances. Managing EC2 SSH credentials is the customer's responsibility within the EC2 console.

Temporary Security Credentials

Q: What are temporary security credentials?
Temporary security credentials consist of the AWS access key ID, secret access key, and security token. Temporary security credentials are valid for a specified duration and for a specific set of permissions. Temporary security credentials are sometimes simply referred to as *tokens*. Tokens can be requested for IAM users or for federated users you manage in your own corporate directory. For more information, see Common Scenarios for Temporary Credentials.

Q: What are the benefits of temporary security credentials?
Temporary security credentials allow you to:

- Extend your internal user directories to enable federation to AWS, enabling your employees and applications to securely access AWS service APIs without needing to create an AWS identity for them.
- Request temporary security credentials for an unlimited number of federated users.
- Configure the time period after which temporary security credentials expire, offering improved security when accessing AWS service APIs through mobile devices where there is a risk of losing the device.

Q: How can I request temporary security credentials for federated users?
You can call the GetFederationToken, AssumeRole, AssumeRoleWithSAML, or AssumeRoleWithWebIdentity STS APIs.

Q: How can IAM users request temporary security credentials for their own use?
IAM users can request temporary security credentials for their own use by calling the AWS STS GetSessionToken API. The default expiration for these temporary credentials is 12 hours; the minimum is 15 minutes, and the maximum is 36 hours.

You can also use temporary credentials withMulti-Factor Authentication (MFA)-Protected API Access.

Q: How can I use temporary security credentials to call AWS service APIs?
If you're making direct HTTPS API requests to AWS, you can sign those requests with the temporary security credentials that you get from AWS Security Token Service (AWS STS). To do this, do the following:

- Use the access key ID and secret access key that are provided with the temporary security credentials the same way you would use long-term credentials to sign a request. For more information about signing HTTPS API requests, see Signing AWS API Requests in the AWS General Reference.
- Use the session token that is provided with the temporary security credentials. Include the session token in the "x-amz-security-token" header. See the following example request.
    - For Amazon S3, via the "x-amz- security-token" HTTP header.
    - For other AWS services, via the SecurityToken parameter.

Q: Which AWS services accept temporary security credentials?
For a list of supported services, see AWS Services That Work with IAM

Q: What is the maximum size of the access policy that I can specify when requesting temporary security credentials (either GetFederationToken or AssumeRole)?
The policy plaintext must be 2048 bytes or shorter. However, an internal conversion compresses it into a packed binary format with a separate limit.

Q: Can a temporary security credential be revoked prior to its expiration?
No. When requesting temporary credentials, we recommend the following:

- When creating temporary security credentials, set the expiration to a value that is appropriate for your application.
- Because root account permissions cannot be restricted, use an IAM user and not the root account for creating temporary security credentials. You can revoke permissions of the IAM user that issued the original call to request it. This action almost immediately revokes privileges for all temporary security credentials issued by that IAM user

Q: Can I reactivate or extend the expiration of temporary security credentials?
No. It is a good practice to actively check the expiration and request a new temporary security credential before the old one expires. This rotation process is automatically managed for you when temporary security credentials are used in roles for EC2 instances.

Q: Are temporary security credentials supported in all regions?
Customers can request tokens from AWS STS endpoints in all regions, including AWS GovCloud (US) and China (Beijing) regions. Temporary credentials from AWS GovCloud (US) and China (Beijing) can be used only in the region from which they originated. Temporary credentials requested from any other region such as US East (N. Virginia) or EU (Ireland) can be used in all regions except AWS GovCloud (US) and China (Beijing).

Q: Can I restrict the use of temporary security credentials to a region or a subset of regions?

No. You cannot restrict the temporary security credentials to a particular region or subset of regions, except the temporary security credentials from AWS GovCloud (US) and China (Beijing), which can be used only in the respective regions from which they originated.

Q: What do I need to do before I can start using an AWS STS endpoint?

AWS STS endpoints are active by default in all regions and you can start using them without any further actions.

Q: What happens if I try to use a regional AWS STS endpoint that has been deactivated for my AWS account?

If you attempt to use a regional AWS STS endpoint that has been deactivated for your AWS account, you will see an **AccessDenied** exception from AWS STS with the following message: "AWS STS is not activated in this region for account: *AccountID*. Your account administrator can activate AWS STS in this region using the IAM console."

Q: What permissions are required to activate or deactivate AWS STS regions from the Account Settings page?

Only users with at least iam:* permissions can activate or deactivate AWS STS regions from the **Account Settings** page in the IAM console. Note that the AWS STS endpoints in US East (N. Virginia), AWS GovCloud (US), and China (Beijing) regions are always active and cannot be deactivated.

Q: Can I use the API or CLI to activate or deactivate AWS STS regions?

No. There is no API or CLI support at this time to activate or deactivate AWS STS regions. We plan to provide API and CLI support in a future release.

Identity Federation

Q: What is identity federation?

AWS Identity and Access Management (IAM) supports identity federation for delegated access to the AWS Management Console or AWS APIs. With identity federation, external identities are granted secure access to resources in your AWS account without having to create IAM users. These external identities can come from your corporate identity provider (such as Microsoft Active Directory or from the AWS Directory Service) or from a web identity provider (such as Amazon Cognito, Login with Amazon, Facebook, Google, or any OpenID Connect-compatible provider).

Q: What are federated users?

Federated users (external identities) are users you manage outside of AWS in your corporate directory, but to whom you grant access to your AWS account using temporary security credentials. They differ from IAM users, which are created and maintained in your AWS account.

Q: Do you support SAML?
Yes, AWS supports the Security Assertion Markup Language (SAML) 2.0.

Q: What SAML profiles does AWS support?
The AWS single sign-on (SSO) endpoint supports the IdP-initiated HTTP-POST binding WebSSO SAML Profile. This enables a federated user to sign in to the AWS Management Console using a SAML assertion. A SAML assertion can also be used to request temporary security credentials using the AssumeRoleWithSAML API. For more information, see About SAML 2.0-Based Federation.

Q: Can federated users access AWS APIs?
Yes. You can programmatically request temporary security credentials for your federated users to provide them secure and direct access to AWS APIs. We have provided a sample application that demonstrates how you can enable identity federation, providing users maintained by Microsoft Active Directory access to AWS service APIs. For more information, see Using Temporary Security Credentials to Request Access to AWS Resources.

Q: Can federated users access the AWS Management Console?
Yes. There are a couple ways to achieve this. One way is by programmatically requesting temporary security credentials (such as GetFederationToken or AssumeRole) for your federated users and including those credentials as part of the sign-in request to the AWS Management Console. After you have authenticated a user and granted them temporary security credentials, you generate a sign-in token that is used by the AWS single sign-on (SSO) endpoint. The user's actions in the console are limited to the access control policy associated with the temporary security credentials. For more details, see Creating a URL that Enables Federated Users to Access the AWS Management Console (Custom Federation Broker).

Alternatively, you can post a SAML assertion directly to AWS sign-in (https://signin.aws.amazon.com/saml). The user's actions in the console are limited to the access control policy associated with the IAM role that is assumed using the SAML assertion. For more details, see Enabling SAML 2.0 Federated Users to Access the AWS Management Console.

Using either approach allows a federated user to access the console without having to sign in with a user name and password. We have provided a sample application that demonstrates how you can enable identity federation, providing users maintained by Microsoft Active Directory access to the AWS Management Console.

Q: How do I control what a federated user is allowed to do when signed in to the console?
When you request temporary security credentials for your federated user using an AssumeRole API, you can optionally include an access policy with the request. The federated user's privileges are the intersection of permissions granted by the access policy passed with the request and the access policy attached to the IAM role that was assumed. The access policy passed with the request cannot elevate the privileges associated with the IAM role being assumed. When you

request temporary security credentials for your federated user using the GetFederationToken API, you must provide an access control policy with the request. The federated user's privileges are the intersection of the permissions granted by the access policy passed with the request and the access policy attached to the IAM user that was used to make the request. The access policy passed with the request cannot elevate the privileges associated with the IAM user used to make the request. These federated user permissions apply to both API access and actions taken within the AWS Management Console.

Q: What permissions does a federated user need to use the console?
A user requires permissions to the AWS service APIs called by the AWS Management Console. Common permissions required to access AWS services are documented in Using Temporary Security Credentials to Request Access to AWS Resources.

Q: How do I control how long a federated user has access to the AWS Management Console?
Depending on the API used to create the temporary security credentials, you can specify a session limit between 15 minutes and 36 hours (for GetFederationToken and GetSessionToken) and between 15 minutes and 12 hours (for AssumeRole* APIs), during which time the federated user can access the console. When the session expires, the federated user must request a new session by returning to your identity provider, where you can grant them access again. Learn more about setting session duration.

Q: What happens when the identity federation console session times out?
The user is presented with a message stating that the console session has timed out and that they need to request a new session. You can specify a URL to direct users to your local intranet web page where they can request a new session. You add this URL when you specify an Issuer parameter as part of your sign-in request. For more information, see Enabling SAML 2.0 Federated Users to Access the AWS Management Console.

Q: How many federated users can I give access to the AWS Management Console?
There is no limit to the number of federated users who can be given access to the console.

Q: What is web identity federation?

Web identity federation allows you to create AWS-powered mobile apps that use public identity providers (such as Amazon Cognito, Login with Amazon, Facebook, Google, or any OpenID Connect-compatible provider) for authentication. With web identity federation, you have an easy way to integrate sign-in from public identity providers (IdPs) into your apps without having to write any server-side code and without distributing long-term AWS security credentials with the app.

For more information about web identity federation and to get started, seeAbout Web Identity Federation.

Q: How do I enable web identity federation with accounts from public IdPs?

For best results, use Amazon Cognito as your identity broker for almost all web identity federation scenarios. Amazon Cognito is easy to use and provides additional capabilities such as anonymous (unauthenticated) access, and synchronizing user data across devices and providers. However, if you have already created an app that uses web identity federation by manually calling the AssumeRoleWithWebIdentity API, you can continue to use it and your apps will still work.

Here are the basic steps to enable identify federation using one of the supported web IdPs:

1. Sign up as a developer with the IdP and configure your app with the IdP, who gives you a unique ID for your app.
2. If you use an IdP that is compatible with OIDC, create an identity provider entity for it in IAM.
3. In AWS, create one or more IAM roles.
4. In your application, authenticate your users with the public IdP.
5. In your app, make an unsigned call to the AssumeRoleWithWebidentity API to request temporary security credentials.
6. Using the temporary security credentials you get in the AssumeRoleWithWebidentity response, your app makes signed requests to AWS APIs.
7. Your app caches the temporary security credentials so that you do not have to get new ones each time the app needs to make a request to AWS.

For more detailed steps, see Using Web Identity Federation APIs for Mobile Apps.

Q: How does identity federation using AWS Directory Service differ from using a third-party identity management solution?

If you want your federated users to be able to access only the AWS Management Console, using AWS Directory Service provides similar capabilities compared to using a third-party identity management solution. End users are able to sign in using their existing corporate credentials and access the AWS Management Console. Because AWS Directory Service is a managed service, customers do not need to set up or manage federation infrastructure, but rather need to create an AD Connector directory to integrate with their on-premises directory. If you are interested in providing your federated users access to AWS APIs, use a third-party offering, or deploy your own proxy server.

Billing

Q: Does AWS Billing provide aggregated usage and cost breakdowns by user?
No, this is not currently supported.

Q: Does the IAM service cost anything?
No, this is a feature of your AWS account provided at no additional charge.

Q: Who pays for usage incurred by users under an AWS Account?
The AWS account owner controls and is responsible for all usage, data, and resources under the account.

Q: Is billable user activity logged in AWS usage data?
Not currently. This is planned for a future release.

Q: How does IAM compare with Consolidated Billing?
IAM and Consolidated Billing are complementary features. Consolidated Billing enables you to consolidate payment for multiple AWS accounts within your company by designating a single paying account. The scope of IAM is not related to Consolidated Billing. A user exists within the confines of an AWS account and does not have permissions across linked accounts. For more details, see Paying Bills for Multiple Accounts Using Consolidated Billing.

Q: Can a user access the AWS accounts billing information?
Yes, but only if you let them. In order for IAM users to access billing information, you must first grant access to the Account Activity or Usage Reports. See Controlling Access.

Additional Questions

Q: What happens if a user tries to access a service that has not yet been integrated with IAM?
The service returns an “Access denied” error.

Q: Are IAM actions logged for auditing purposes?
Yes. You can log IAM actions, STS actions, and AWS Management Console sign-ins by activating AWS CloudTrail. To learn more about AWS logging, see AWS CloudTrail.

Q: Is there any distinction between people and software agents as AWS entities?
No, both of these entities are treated like users with security credentials and permissions. However, people are the only ones to use a password in the AWS Management Console.

Q: Do users work with AWS Support Center and Trusted Advisor?
Yes, IAM users have the ability to create and modify support cases as well as use Trusted Advisor.

Q: Are there any default quota limits associated with IAM?
Yes, by default your AWS account has initial quotas set for all IAM-related entities. For details

see Limitations on IAM Entities and Objects.

These quotas are subject to change. If you require an increase, you can access the Service Limit Increase form via the Contact Us page, and choose **IAM Groups and Users** from the **Limit Type** drop-down list.

Multi-Factor Authentication

Q. What is AWS MFA?
AWS multi-factor authentication (AWS MFA) provides an extra level of security that you can apply to your AWS environment. You can enable AWS MFA for your AWS account and for individual AWS Identity and Access Management (IAM) users you create under your account.

Q. How does AWS MFA work?
AWS MFA uses an authentication device that continually generates random, six-digit, single-use authentication codes. There are two primary ways to authenticate using an AWS MFA device:

- AWS Management Console users: When a user with MFA enabled signs in to an AWS website, they are prompted for their user name and password (the first factor–what they know), and an authentication code from their AWS MFA device (the second factor–what they have). All AWS websites that require sign-in, such as the AWS Management Console, fully support AWS MFA. You can also use AWS MFA together with Amazon S3 secure delete for additional protection of your S3 stored versions.

- AWS API users: You can enforce MFA authentication by adding MFA restrictions to your IAM policies. To access APIs and resources protected in this way, developers can request temporary security credentials and pass optional MFA parameters in their AWS Security Token Service (STS) API requests (the service that issues temporary security credentials). MFA-validated temporary security credentials can be used to call MFA-protected APIs and resources.

Q. How do I get AWS MFA?
You follow two easy steps:

Get an authentication device. You have three options:

- You can purchase a hardware device that is compatible with AWS MFA from Gemalto, a third party provider.

- You can install a virtual AWS MFA compatible application on a device such as your smartphone.

- You can sign up for the preview of SMS MFA, which allows you to use the text messaging functionality of your mobile phone to receive security codes (*available only for IAM users*).

Visit the MFA page for details on how to acquire a hardware or virtual MFA device and how to set up SMS MFA.

After you have the authentication device, you must activate it. You activate an AWS MFA device for your AWS account or your IAM users in the IAM Console. You can also use the IAM CLI to activate it for an IAM user.

Q. Is there a fee associated with using AWS MFA?
AWS does not charge any additional fees for using AWS MFA with your AWS account. However, if you want to use a physical authentication device then you will need to purchase an authentication device that is compatible with AWS MFA from Gemalto, a third party provider. For more details, please visit Gemalto's website.

Q. Can I have multiple authentication devices active for my AWS account?
Yes. Each IAM user can have its own authentication device. However, each identity (IAM user or root account) can be associated with only one authentication device.

Q. Can I use my authentication device with multiple AWS accounts?
No. The authentication device or mobile phone number is bound to an individual AWS identity (IAM user or root account). If you have a TOTP-compatible application installed on your smartphone, you can create multiple virtual MFA devices on the same smartphone. Each one of the virtual MFA devices is bound to a single identity, just like a hardware device. If you dissociate (deactivate) the authentication device, you can then reuse it with a different AWS identity. The authentication device cannot be used by more than one identity simultaneously.

Q. I already have a hardware authentication device from my place of work or from another service I use, can I re-use this device with AWS MFA?
No. AWS MFA relies on knowing a unique secret associated with your authentication device in order to support its use. Because of security constraints that mandate such secrets never be shared between multiple parties, AWS MFA cannot support the use of your existing hardware authentication device. Only a compatible hardware authentication device purchased from Gemalto can be used with AWS MFA.

Purchasing an MFA Device

Q. I'm having a problem with an order for an authentication device using the third-party provider Gemalto's website. Where can I get help?
Gemalto's customer service can assist you.

Q. I received a defective or damaged authentication device from the third party provider Gemalto. Where can I get help?
Gemalto's customer service can assist you.

Q. I just received an authentication device from the third party provider Gemalto. What should I do?
You simply need to activate the authentication device to enable AWS MFA for your AWS account. See the IAM console to perform this task.

Provisioning a Virtual MFA Device

Q. What is a virtual MFA device?
A virtual MFA device is an entry created in a TOTP compatible software application that can generate six-digit authentication codes. The software application can run on any compatible computing device, such as a smartphone.

Q. What are the differences between a virtual MFA device and physical MFA devices?
Virtual MFA devices use the same protocols as the physical MFA devices. Virtual MFA devices are software based and can run on your existing devices such as smartphones. Most virtual MFA applications also allow you to enable more than one virtual MFA device, which makes them more convenient than physical MFA devices.

Q. Which virtual MFA applications can I use with AWS MFA?
You can use applications that generate TOTP-compliant authentication codes, such as the Google Authenticator application, with AWS MFA. You can provision virtual MFA devices either automatically by scanning a QR code with the device's camera or by manual seed entry in the virtual MFA application.

Visit the MFA page for a list of supported virtual MFA applications.

Q. What is a QR code?
A QR code is a two-dimensional barcode that is readable by dedicated QR barcode readers and most smartphones. The code consists of black squares arranged in larger square patterns on a white background. The QR code contains the required security configuration information to provision a virtual MFA device in your virtual MFA application.

Q. How do I provision a new virtual MFA device?
You can configure a new virtual MFA device in the IAM console for your IAM users as well as for your AWS root account. You can also use the aws iam create-virtual-mfa-device command in the AWS CLI or the CreateVirtualMFADevice API to provision new virtual MFA devices under your account. The aws iam create-virtual-mfa-device and the CreateVirtualMFADevice API return the required configuration information, called a seed, to configure the virtual MFA device in your AWS MFA compatible application. You can either grant your IAM users the permissions to call this API directly or perform the initial provisioning for them.

Q. How should I handle and distribute the seed material for virtual MFA devices?
You should treat seed material like any other secret (for example the AWS secret keys and

passwords).

Q. How can I enable an IAM user to manage virtual MFA devices under my account?
Grant the IAM user the permission to call the CreateVirtualMFADevice API. You can use this API to provision new virtual MFA devices.

Setting up SMS MFA

Q. How do I begin using the SMS option during the preview?
To sign up for the preview, you must visit the MFA page and register by clicking the SMS MFA sign-up button. After acceptance into the preview, which typically occurs within one or two business days, you receive a confirmation email with instructions about how to set up SMS MFA. You can then navigate to the IAM console and enable SMS MFA for an IAM user. The process involves entering a phone number for each IAM user. Then, when the IAM user signs in to the AWS Management Console, the user receives a 6-digit security code via a standard SMS text message and must enter it during sign-in.

Q. Can I use SMS MFA with root accounts during the preview?
No. Support for the SMS MFA option is limited to IAM users during the preview.

Q. Can I use SMS MFA when assuming temporary security credentials from AWS STS?
No. In the preview, you cannot use SMS MFA when assuming temporary security credentials from AWS STS.

Enabling AWS MFA Devices

Q. Where do I enable AWS MFA?
You can enable AWS MFA for an AWS account and your IAM users in the IAM console, the AWS CLI, or by calling the AWS API.

Q. What information do I need to activate a hardware or virtual authentication device?
If you are activating the MFA device with the IAM console then you only need the device. If you are using the AWS CLI or the IAM API then you need the following:

1. The serial number of the authentication device. The format of the serial number depends on whether you are using a hardware device or a virtual device:

   - Hardware MFA device: The serial number is on the bar-coded label on the back of the device.
   - Virtual MFA device: The serial number is the Amazon Resource Name (ARN) value returned when you run the iam-virtualmfadevicecreate command in the AWS CLI or call the CreateVirtualMFADevice API.

2. Two consecutive authentication codes displayed by the authentication device.

Q. My authentication device seems to be working normally, but I am not able to activate it. What should I do?
Please contact us for help.

Using AWS MFA

Q. If I enable AWS MFA for my AWS root account or my IAM users, do they always need to use an authentication code to sign in to the AWS Portal or AWS Management Console?
Yes. The AWS account and your IAM users must have their MFA device with them any time they need to sign in any AWS site.

If the authentication device associated with the AWS root account is damaged, lost, stolen, or stops working, you can contact us for help with disabling AWS MFA for the root account. This allows you to temporarily sign in to AWS using just the user name and password for the AWS account.

With virtual and hardware MFA, if your IAM users lose or damage their authentication device or if it is stolen or stops working, you can disable AWS MFA yourself using the IAM console or the AWS CLI. With SMS MFA, there is no disruption to MFA if you acquire a new mobile phone that retains the same phone number.

Q. If I enable AWS MFA for my AWS root account or IAM users, do they always need to enter an MFA code to directly call AWS APIs?
No, it’s optional. However, you must enter an MFA code if you plan to call APIs that are secured by MFA-protected API access.

If you are calling AWS APIs using access keys for your AWS root account or IAM user, you do not need to enter an MFA code. For security reasons, we recommend that you remove all access keys from your AWS root account and instead call AWS APIs with the access keys for an IAM user that has the required permissions.

Q. How do I sign in to the AWS Portal and AWS Management Console using my authentication device?
Follow these two steps:

If you are signing in as an AWS root account, sign in as usual with your user name and password when prompted. To sign in as an IAM user, use the account-specific URL and provide your user name and password when prompted.

On the next page, enter the six-digit authentication code that appears on your authentication device.

Q. Does AWS MFA affect how I access AWS Service APIs?
AWS MFA changes the way IAM users access AWS Service APIs only if the account administrator(s) choose to enable MFA-protected API access. Administrators may enable this

feature to add an extra layer of security over access to sensitive APIs by requiring that callers authenticate with an AWS MFA device. For more information, see the MFA-protected API access documentation in more detail.

Other exceptions include S3 PUT bucket versioning, GET bucket versioning, and DELETE object APIs, which allow you to require MFA authentication to delete or change the versioning state of your bucket. For more information see the S3 documentation discussing Configuring a Bucket with MFA Delete in more detail.

For all other cases, AWS MFA does not currently change the way you access AWS service APIs.

Q. Can I use a given authentication code more than once?
No. For security reasons, you can use each authentication code only once.

Q. I was recently asked to resync my authentication device because my authentication codes were being rejected. Should I be concerned?
No, this can happen occasionally. AWS MFA relies on the clock in your authentication device being in sync with the clock on our servers. Sometimes, these clocks can drift apart. If this happens, when you use the authentication device to sign in to access secure pages on the AWS website or the AWS Management Console, AWS automatically attempts to resync the authentication device by requesting that you provide two consecutive authentication codes (just as you did during activation).

Q. My authentication device seems to be working normally, but I am not able to use it to sign in to the AWS Management Console. What should I do?
We suggest you try re-syncing the authentication device using this link if your MFA device protects your AWS root account credentials (requires sign-in), or this link for your IAM user's credentials. If you already tried to resync and are still having trouble signing in, please contact us for help.

Q. My authentication device is lost, damaged, or stolen, and now I can’t sign in to the AWS Management Console. What should I do?
If the authentication device is associated with an AWS root account, follow these steps:

Contact us for help with disabling AWS MFA so you can temporarily access the AWS Management Console using just your user name and password.

Be sure to change your Amazon password in case an attacker has stolen your authentication device and might also have your current password.

Purchase a new authentication device from the third party providerGemalto using their website or provision a new virtual MFA device under your account using the IAM console.

When you complete the preceding steps, use the IAM console to activate the new authentication device to reenable AWS MFA for your AWS account.

If the authentication device is associated with an IAM user, you can use the IAM console, AWS CLI, or AWS API to remove the MFA device for the IAM user. If you are using SMS MFA for the IAM user, there is no disruption to MFA if you acquire a new mobile phone that retains the same phone number.

Q. My physical authentication device has stopped working and now I can’t sign in to the AWS Portal or AWS Management Console. What should I do?
If the physical authentication device is associated with an AWS root account, follow these steps:

Contact us for help with disabling AWS MFA so you can temporarily access the AWS Management Console using just your user name and password.

Contact the third party provider Gemalto for further assistance with the authentication device.

When you have another authentication device, come back to the AWS website and activate the authentication device to reenable AWS MFA for your AWS account, just as before.

If the authentication device is associated with an IAM user, the user should contact the person who provided the IAM user name and password.

Q. How do I disable AWS MFA?
To disable AWS MFA for your AWS account, you can deactivate your authentication device using the Security Credentials page. To disable AWS MFA for your IAM users, you need to use the IAM console or the AWS CLI.

Q. Can I use AWS MFA in GovCloud?
Yes, you can use AWS virtual MFA in GovCloud. AWS does not currently support hardware MFA devices in GovCloud.

MFA-protected API access

Q. What is MFA-protected API access?
MFA-protected API access is optional functionality that lets account administrators enforce additional authentication for customer-specified APIs by requiring that users provide a second authentication factor in addition to a password. Specifically, it enables administrators to include conditions in their IAM policies that check for and require MFA authentication for access to selected APIs. Users making calls to those APIs must first get temporary credentials that indicate the user entered a valid MFA code.

Q. What problem does MFA-protected API access solve?
Previously, customers could require MFA for access to the AWS Management Console, but could not enforce MFA requirements on developers and applications interacting directly with AWS service APIs. MFA-protected API access ensures that IAM policies are universally

enforced regardless of access path. As a result, you can now develop your own application that uses AWS and prompts the user for MFA authentication before calling powerful APIs or accessing sensitive resources.

Q. How do I get started with MFA-protected API access?
You can get started in two simple steps:

1. Assign an MFA device to your IAM users. You can purchase a hardware key fob, use SMS MFA with any SMS-compatible mobile phone, or download a free TOTP-compatible application for your smart phone, tablet, or computer. See the MFA detail page for more information on AWS MFA devices.

2. Enable MFA-protected API access by creating permission policies for the IAM users and/or IAM groups from which you want to require MFA authentication. To learn more about access policy language syntax, see the access policy language documentation.

Q. How do developers and users access APIs and resources secured with MFA-protected API access?
Developers and users interact with MFA-protected API access both in the AWS Management Console and at the APIs.

In the AWS Management Console, any MFA-enabled IAM user must authenticate with their device to sign in. Users that do not have MFA do not receive access to MFA-protected APIs and resources.

At the API level, developers can integrate AWS MFA into their applications to prompt users to authenticate using their assigned MFA devices before calling powerful APIs or accessing sensitive resources. Developers enable this functionality by adding optional MFA parameters (serial number and MFA code) to requests to obtain temporary security credentials (such requests are also referred to as “session requests”). If the parameters are valid, temporary security credentials that indicate MFA status are returned. See the temporary security credentials documentation for more information.

Q. Who can use MFA-protected API access?
MFA-protected API access is available for free to all AWS customers.

Q. Which services does MFA-protected API access work with?
MFA-protected API access is supported by all AWS services that support temporary security credentials. For a list of supported services, see AWS Services that Work with IAMand review the column labeled Supports temporary security credentials.

Q. What happens if a user provides incorrect MFA device information when requesting temporary security credentials?
The request to issue temporary security credentials fails. Temporary security credential requests that specify MFA parameters must provide the correct serial number of the device linked to the

IAM user as well as a valid MFA code.

Q. Does MFA-protected API access control API access for AWS root accounts?
No, MFA-protected API access only controls access for IAM users. Root accounts are not bound by IAM policies, which is why we recommend that you create IAM users to interact with AWS service APIs rather than use AWS root account credentials.

Q. Do users have to have an MFA device assigned to them in order to use MFA-protected API access?
Yes, a user must first be assigned a unique virtual, hardware, or SMS MFA device.

Q. Is MFA-protected API access compatible with S3 objects, SQS queues, and SNS topics?
Yes.

Q. How does MFA-protected API access interact with existing MFA use cases such as S3 MFA Delete?
MFA-protected API access and S3 MFA Delete do not interact with each other. S3 MFA Delete currently does not support temporary security credentials. Instead, calls to the S3 MFA Delete API must be made using long-term access keys.

Q. Does MFA-protected API access work in the GovCloud (US) region?
Yes.

Q. Does MFA-protected API access work for federated users?
Customers cannot use MFA-protected API access to control access for federated users. The GetFederatedSession API does not accept MFA parameters. Since federated users can't authenticate with AWS MFA devices, they are unable to access resources designated using MFA-protected API access.

Q. Can I use SMS MFA when assuming temporary security credentials form AWS STS?

During the preview, you cannot use SMS MFA when assuming temporary security credentials from AWS STS.

AWS CloudHSM FAQ

General

Q: What is AWS CloudHSM?
The AWS CloudHSM service helps you meet corporate, contractual and regulatory compliance requirements for data security by using dedicated Hardware Security Module (HSM) appliances within the AWS cloud. AWS and AWS Marketplace partners offer a variety of solutions for protecting sensitive data within the AWS platform, but for some applications and data subject to contractual or regulatory mandates for managing cryptographic keys, additional protection may

be necessary. CloudHSM complements existing data protection solutions and allows you to protect your encryption keys within HSMs that are designed and validated to government standards for secure key management. CloudHSM allows you to securely generate, store and manage cryptographic keys used for data encryption in a way that keys are accessible only by you.

Q: What is a Hardware Security Module (HSM)?

A Hardware Security Module (HSM) is a hardware appliance that provides secure key storage and cryptographic operations within a tamper-resistant hardware device. HSMs are designed to securely store cryptographic key material and use the key material without exposing it outside the cryptographic boundary of the appliance.

Q: What can I do with CloudHSM?

You can use the CloudHSM service to support a variety of use cases and applications, such as database encryption, Digital Rights Management (DRM), Public Key Infrastructure (PKI), authentication and authorization, document signing, and transaction processing. You can read about several common use cases for CloudHSM in this AWS Security blog post.

Q: What types of HSMs are available?

As part of the service, AWS currently provides Luna SA 7000 HSM appliances from SafeNet, Inc., with version 5 of the Luna SA software.

Q: How does CloudHSM work?

When you use the AWS CloudHSM service you receive dedicated single tenant access to each CloudHSM appliance. Each appliance appears as a network resource in your Virtual Private Cloud (VPC). You, not Amazon, initialize and manage the HSM partitions on the HSM. As part of provisioning, you receive administrator credentials for the appliance, and may create an HSM partition on the appliance. After creating an HSM partition, you can configure a client on your EC2 instance that allows your applications to use the APIs provided by the HSM.

The cryptographic partition is a logical and physical security boundary that restricts access to your keys, so only you control your keys and operations performed by the HSM. Amazon administrators will manage and monitor the health of the HSM appliance, but do not have access to the cryptographic partition. Your applications use standard cryptographic APIs, in conjunction with HSM client software installed on the application instance, to send cryptographic requests to the HSM. The client software transparently sets up a secure channel to the HSM appliance using credentials that you create and sends requests on this channel, and the HSM performs the operations and returns the results over the secure channel. The client then returns the result to the application through the cryptographic API.

Q: I don’t currently have a VPC. Can I still use AWS CloudHSM?

No. To protect and isolate your CloudHSM from other Amazon customers, CloudHSM must be provisioned inside a VPC. Creating a VPC is easy. Please see the VPC Getting Started Guide for more information.

Q: Does my application need to reside in the same VPC as the CloudHSM instance?

No, but the server or instance on which your application and the HSM client is running must have network (IP) reachability to the HSM. You can establish network connectivity from your application to the HSM in many ways, including operating your application in the same VPC, with VPC peering, with a VPN connection, or with Direct Connect. Please see the VPC Peering Guide and VPC User Guide for more details.

Q: Does CloudHSM work with on-premises HSMs?

Yes. The software and firmware versions of your on-premises HSMs must match those of the CloudHSM instances. You can connect CloudHSM instances in your VPC to your datacenter using the VPN capability built into VPC or with AWS Direct Connect.

Q: How can my application use CloudHSM?

SafeNet has integrated and tested the Luna SA HSM with a number of commercial software solutions. Examples include Oracle Database 11g, Microsoft SQL Server 2008 and 2012, SafeNet Virtual KeySecure, and Apache web server SSL termination with private keys stored in the HSM. Please see the CloudHSM User Guide for a complete list of supported applications and links to the technical application notes that explain how configure your applications with CloudHSM.

If you are developing your own custom application, your application can use the standard APIs supported by the Luna SA HSM, including PKCS#11, Microsoft CAPI/CNG and Java JCA/JCE (Java Cryptography Architecture/Java Cryptography Extensions). The SafeNet documentation provides a complete list of supported APIs. Please refer to the CloudHSM User Guide for code samples and help with getting started.

Q: Can I use CloudHSM to store keys or encrypt data used by other AWS services?
Yes. You can write custom applications and integrate them with CloudHSM, or you can leverage one of the third party encryption solutions available from AWS Technology Partners. Examples include EBS volume encryption and S3 object encryption and key management. Please see the CloudHSM User Guide for a list of supported applications and links to technical application notes that describe third party solutions that work with CloudHSM.

Q: Can other AWS services use CloudHSM to store and manage keys (for example Amazon S3 or Amazon Redshift)?
Amazon Relational Database Service (RDS) for Oracle Database and Amazon Redshift can be configured to store master keys in CloudHSM instances. Please refer to the relevant Amazon RDS documentation or Amazon Redshift documentation for more details. Over time we may integrate CloudHSM with other AWS services. If this is of interest to you, please let us know.

Q: Can CloudHSM be used to perform personal identification number (PIN) block

translation or other cryptographic operations used with debit payment transactions?

The Luna SA HSM is a general purpose HSM that is not capable of supporting these operations.

Getting CloudHSM Service

Q: Where is CloudHSM available?

CloudHSM is available today in the US East (Northern Virginia), US East (Ohio), US West (Northern California), US West (Oregon), EU (Ireland), EU (Frankfurt), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Sydney) and AWS GovCloud (US) regions. If you are interested in using CloudHSM in any other regions, please Contact Us. You can find availability information for AWS services on the AWS Products and Services by Region page.

Q: How do I get started with CloudHSM?

You can provision a CloudHSM instance with a few API calls through the AWS SDK, API or via the CloudHSM Command Line Interface Tools. To learn more, please see the CloudHSM User Guide for information about getting started with the CLI Tools, the CloudHSM Developer Guide for information about the API, or the Tools for Amazon Web Services page for more information about the SDK. If you want to start with a free trial, see the CloudHSM Free Trial page for more information.

Q: How long does it take to get CloudHSM service?

Requests for HSM appliances typically can be satisfied within 15 minutes through the AWS API, SDK, or CloudHSM CLI Tools.

Q: How do I terminate CloudHSM service?

Before ending service, AWS requires you to delete all your cryptographic key material from the HSM appliance. After you delete (zeroize) all of your key material, you can use the CloudHSM API, SDK, or CLI Tools to stop using the service, or you can contact us for assistance. Please refer to the CloudHSM User Guide for further instructions.

Billing

Q: How will I be charged and billed for my use of the AWS CloudHSM service?

You will be charged an upfront fee for each CloudHSM instance you launch, and an hourly fee for each hour thereafter until you terminate the instance. Amazon reserves the right to charge for network data transfers in and out of a CloudHSM that exceed 5000 GB per month. For more information, please visit the AWS CloudHSM pricing page.

Q: Is there a Free Tier for the CloudHSM service?

No, but you may be eligible for a CloudHSM Free Trial.

Q: Do I need to purchase any licenses for the HSM to use the CloudHSM service?

No. The CloudHSM service includes everything you need to allow you to connect as many as 800 clients to the HSM and use 20 partitions on the HSM.

Provisioning and Operations

Q: Are there any prerequisites for signing up for CloudHSM?

Yes. In order to start using CloudHSM there are a few prerequisites, including a Virtual Private Cloud (VPC) in the region where you want CloudHSM service. It's easy to configure a VPC and a complete CloudHSM test environment using a CloudFormation template provided by AWS. Refer to the CloudHSM User Guide for more details.

Q: How much capacity do I need?

CloudHSM instances are not rate limited; they run at the full rated capacity of the HSM appliance. You should evaluate your cryptographic workload and compare it against the performance and scale characteristics of the Luna SA appliance, described on the SafeNet Luna SA Product Page.

Q: Which versions of CloudHSM firmware and software are supported?

Two versions of the SafeNet appliance software and firmware are currently supported: 5.1.5/6.2.1 (FIPS validated) and 5.3.5/6.10.2 (FIPS candidate). Customers requiring other versions to support their application should contact AWS Support.

Q: Can I upgrade the firmware or software of the HSM?

Maintaining the appliance firmware and software is the responsibility of the customer. Upgrade instructions for supported versions can be found in the CloudHSM Upgrade Guide.

Q: How many HSM appliances will I need?

AWS strongly recommends that you use at least two appliances, and that each appliance is in a high availability (HA) configuration as described in the CloudHSM User Guide. You can use the SafeNet Luna client to load balance across two or more HSMs.

Q: Who is responsible for key durability?

AWS does not have Security Officer or Cloning Domain credentials for any HSMs that you are using. These credentials are needed to perform backups or to configure high availability. Therefore you are solely responsible for the durability of the key material on the HSMs that you are using.

Q: How do I set up a high availability (HA) configuration?

You can find more information about high availability configuration in theCloudHSM User Guide. The CloudHSM CLI Tools are designed to simplify configuration and operations using high availability HSM configurations.

Q: How many HSMs can be connected in an HA group?

At this time, the maximum number of HSMs in an HA group is sixteen.

Q: Can I back up the contents of a CloudHSM?

Yes. For security reasons, the HSM is configured in the factory to allow the contents of the HSM to be duplicated only to another HSM. The HSM also requires that you configure the same cloning domain on the source and target HSM devices (your ownership and control of the cloning domain credential gives you and only you the ability to clone the contents of the HSM). You can clone the contents of your CloudHSM to another SafeNet Luna SA using a high availability (HA) configuration or to a SafeNet Luna Backup HSM. The CloudHSM CLI Tools are designed to simplify cloning and other operations using high availability HSM configurations.

Q: Do I have to purchase the SafeNet Luna Backup HSM in order to use the CloudHSM service?

No. The Luna Backup HSM is optional.

Q: Where can I purchase the SafeNet Luna Backup HSM?

You can purchase it directly from SafeNet using the following link: http://www.safenet-inc.com/request-information/

Q: Is there an SLA for CloudHSM?

At the present time, there is no SLA for CloudHSM.

Security

Q: Do I share the HSM instance with other AWS customers?

No. As part of the service you receive dedicated single tenant access to the HSM appliance.

Q: How does AWS manage the HSM without having access to my encryption keys?

Separation of duties and role-based access control is inherent in the design of the SafeNet Luna SA HSM. AWS has administrative credentials to the appliance, but these credentials can only be used to manage the appliance, not the HSM partitions on the appliance. AWS uses these credentials to monitor and maintain the health and availability of the appliance. You can use syslog and SNMP to monitor the health and availability of an HSM appliance that you are using.

AWS controls availability of the appliance but is unable to access or use your keys. For instance, AWS can remove your network access to the appliance, or can re-initialize the appliance, which will result in destruction of your keys. However, AWS cannot extract your keys or cause the appliance to perform cryptographic operations using your keys.

AWS is not involved in the creation and management of the key material stored within an HSM. You control the HSM partitions and must perform these tasks. In Luna SA terminology, AWS has **Admin** credentials to the HSM appliance, but never has**Security Officer** or **Partition** credentials. Details about these roles and more information about the Luna SA HSM appliance are available in the topic E - Concepts in the Luna SA documentation.

Q: Can I monitor the HSM appliance?

Yes. The appliance generates logs that can be monitored via syslog. You may use your own syslog endpoint to monitor appliance logins, NTLS connections, environmental conditions, etc. AWS also monitors the appliance for health and availability.

Q: What happens if someone tampers with the HSM appliance?

The SafeNet Luna SA appliance has both physical and logical tamper detection and response mechanisms that trigger key deletion (zeroization) of the appliance and generate event logs. The HSM is designed to detect tampering if the physical barrier of the HSM appliance is breached. In addition, after three unsuccessful attempts to access an HSM partition with HSM Admin credentials, the HSM appliance erases its HSM partitions. If the HSM detects a tampering attempt, it stops responding for approximately ten minutes and then restarts. After restarting, the HSM generates a local syslog event and if configured for remote syslog monitoring, it sends a syslog message. For more information, see the SafeNet Luna SA documentation.

Q: What happens in case of failure?

Amazon monitors and maintains the appliance and network for availability and error conditions. If an appliance fails or loses network connectivity, an AWS engineer will investigate. If the outage is short (for example, a transient network event), then service will be restored as soon as possible. If the outage is anticipated to be long (for example, a hardware failure on an HSM appliance), then AWS will either notify you so you can provision a new CloudHSM instance yourself, or provision a replacement instance and notify you that it is ready, so you can migrate your workload to the new appliance.

If you previously used the CLI Tools to configure HA partition groups, you can clone all of the keys from one HSM to another using the ‘clone-hsm’ command. You can check the health of an individual HSM using the CloudHSM API, SDK, or CLI Tools, and you can check the overall health of the service at any time using the AWS Service Health Dashboard.

Q: Could I lose my keys if a single HSM appliance fails?

Yes. It is possible to lose your keys if the CloudHSM instance that you are using fails and you are not using two or more CloudHSM instances, or a combination of a CloudHSM and an on-premises HSM, in a high availability mode. Amazon strongly recommends that you use two or more CloudHSM instances, in separate Availability Zones, in a high availability mode in order to avoid loss of cryptographic keys.

Q: Can Amazon recover my keys if I lose my credentials to the appliance?

No. Amazon does not have access to your keys or credentials and therefore has no way to recover your keys if you lose your credentials.

Q: How do I know that I can trust CloudHSM appliances?

The Luna SA is designed to meet Federal Information Processing Standard (FIPS) 140-2 and Common Criteria EAL4+ standards. You can find more information about Luna SA regulatory compliance and third party validation on the Luna SA product page.

SafeNet has documented a process that allows you to confirm that you are communicating directly with an HSM appliance. Please refer to How Do I Know That I'm Communicating with an HSM? for more information.

Q: How do I operate a CloudHSM in *FIPS 140-2* mode?

The appliance can be operated in FIPS 140-2 Level 2 mode by disabling non-FIPS-compliant algorithms and enabling password authentication in the HSM policy when you create the HSM partition. Please see the Luna SA documentation for full details on this procedure.

Q: Does the CloudHSM service support FIPS 140-2 Level 3?

No. The Luna SA as it is configured for the CloudHSM service with password-based authentication does not support FIPS 140-2 Level 3.

Q: Does the CloudHSM appliance (Luna SA) meet any of the requirements for FIPS 140-2 Level 3?

Yes. The SafeNet Luna SA meets the physical security, EMI/EMC, and design assurance requirements for FIPS 140-2 Level 3. For more information, please refer to the Non-proprietary Security Policy for Luna® PCI-e Cryptographic Module.

Q: How can I securely distribute an HSM partition credential to my instances?

Please refer to the following AWS Security Blog post which describesUsing IAM roles to distribute non-AWS credentials to your EC2 instances.

Q: Can I get a history of all CloudHSM API calls made from my account?

Yes. AWS CloudTrail records AWS API calls for your account. The AWS API call history produced by CloudTrail lets you perform security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the CloudTrail home page, and turn it on via CloudTrail's AWS Management Console.

Q: Which events are not logged in CloudTrail?

CloudTrail does not include any of the HSM device or access logs, but you can collect logs from the HSM using syslog.

Compliance

Q: Which AWS compliance initiatives include CloudHSM?

CloudHSM is included in the AWS Payment Card Industry (PCI) Data Security Standard (DSS) (PCI-DSS), Service Organization Control (SOC) 1, SOC 2, and SOC 3 audits. Please refer to the AWS Compliance site for more information about these compliance programs, and to learn more about the security controls in place for CloudHSM.

Q: How can I request compliance reports that include CloudHSM in scope?

You can request compliance reports through your Business Development representative. If you don’t have one, you can request one here.

Performance and Capacity

Q: How many cryptographic operations per second are supported?

AWS encourages you to measure the performance parameters that are important for your applications before deploying production applications that have performance dependencies on the CloudHSM service. You can also review SafeNet’s performance results. Please contact us if you have specific questions.

Q: How many keys can be stored on a CloudHSM instance?

As configured for the CloudHSM service, the Luna SA HSM has 2 MB of key and object storage. CloudHSM applications can typically store approximately 14,000 symmetric keys, 1,200 RSA 2048 key pairs, or between 4000 and 6000 ECDSA key pairs, depending on which curve is used. The number of keys your application is able to store depends on the application and how much additional space is consumed with metadata for each key. AWS encourages you to test and measure capacity parameters that are important for your application.

Q: How many simultaneous client sessions are supported?

The SafeNet Luna SA is designed to support 800 simultaneous client connections. AWS encourages you to test and measure capacity and performance parameters that are important for your application.

AWS CloudHSM for Amazon RDS Oracle TDE

Q: What is Transparent Data Encryption (TDE) and how is it relevant to Amazon RDS?

Transparent Data Encryption (TDE) is a feature of Oracle Database for encrypting the data in a database without the need for users to manage the encryption key. Amazon Relational Database Service (Amazon RDS) for Oracle supports TDE for Oracle Database 11g Enterprise Edition.

Q: What is CloudHSM for Amazon RDS Oracle TDE?

CloudHSM for Amazon RDS Oracle TDE enables Transparent Data Encryption, a standard feature of Oracle 11g, for encrypting the database in a way that is transparent to your applications, while creating and storing the master encryption key on CloudHSM devices that you control.

Q: What can I do with CloudHSM for Amazon RDS Oracle TDE?

You can encrypt your Amazon RDS Oracle database using TDE with a master encryption key created and stored in CloudHSM appliances that you control. The Amazon RDS database instance cannot start unless you provide access to the master key that is created and stored in the HSM hardware. Storing the master encryption key in a third-party validated HSM that you control can help you meet strict regulatory and compliance requirements for strong key protection.

Q: How do I get started with CloudHSM for Amazon RDS Oracle TDE?

You can use a CloudFormation template provided by AWS to configure the prerequisites for CloudHSM, or you can configure these prerequisites manually. Then create two or three CloudHSM instances using the CLI Tools. With a couple more CLI Tools commands, initialize and configure the CloudHSM instances with a high availability (HA) configuration. Finally, create an Amazon RDS database instance and configure it to use the HSM HA group that you created and provide your HSM partition credential to Amazon RDS. Refer to the Amazon RDS User Guide for more details.

Q: How can I make sure the TDE master key is available to Amazon RDS?

AWS recommends using three HSMs in a high availability configuration with Amazon RDS. You can use the CloudHSM CLI Tools to configure a high availability (HA) group of CloudHSM appliances.

Q: Which database engines does CloudHSM for Amazon RDS Oracle TDE work with?

Oracle Database 11g Enterprise Edition is supported. Please contact us if you are interested in using CloudHSM with another database engine or with a different version of the Oracle database.

Q: How many database instances can share a single CloudHSM partition?

Storing master keys from different database instances on the same partition is supported by Oracle and RDS, so you are limited only by the storage space on the HSM.

CloudHSM for Amazon RDS Oracle TDE Security

Q: How is my data protected with CloudHSM for Amazon RDS Oracle?

CloudHSM for Amazon RDS Oracle TDE uses the Oracle TDE feature to encrypt your database. Rather than storing the master encryption key in the Oracle software wallet, the master key is stored in an HSM. The Oracle database documentation provides more details about the operation of Oracle TDE.

Q: How can I change (rotate) the database master encryption key?

AWS automatically rotates the master encryption key once per year. AWS can also rotate the master key by request. Rotating the master key creates a new key and retains the old keys in the HSM, which consumes key storage space on the HSM. Storage space on the HSM is very limited and excessive key rotations could exhaust the storage capacity of the HSM.

Q: When are master keys created in the HSM?

A new master key is created when creating a new database instance (including restoring a database from a backup) if an option group with the *TDE-with-CloudHSM* option enabled is applied to the instance, and when AWS rotates the master key.

Q: Can Amazon recover my keys if I lose my credentials to the HSM?

No. Amazon does not have access to your keys or credentials and therefore has no way to recover your keys if you lose your credentials, and this could result in unrecoverable data loss.

Command Line Interface Tools

Q: What are the CloudHSM Command Line Interface (CLI) Tools?

The CloudHSM CLI tools simplify and centralize CloudHSM administration. They make it easier for you, acting as the HSM Security Officer, to configure and manage the HSM. The tools also work in conjunction with and use the CloudHSM API to make it easier to configure your application to work with several HSMs in a high availability configuration.

Q: What can I do with the CloudHSM Command Line Interface Tools?

The tools make it easy to set up your application to use the HA and load balancing features of the SafeNet Luna client software. The tools also centralize configuration and other operations that previously required you to log in to each HSM and type Luna shell commands. For example, with a few CLI Tools commands you can create a CloudHSM instance, initialize the HSM, create a group of HSM partitions across multiple HSMs, generate the HSM client configuration, register HSM clients with the HSM, and distribute certificates between the client and HSM.

Q: How can I download and get started with the CloudHSM Command Line Interface Tools?

You'll find instructions in the CloudHSM User Guide.

Q: Do the CloudHSM CLI Tools provide AWS with access to the contents of the HSM?

No. The CLI tools are Python scripts that connect directly to the HSM via SSH and execute Luna shell commands on the HSM on your behalf. When you run a command, you provide your HSM credentials to allow the script to connect to the HSM via SSH and configure it, but this does not provide your credentials to AWS.

For example, to initialize an HSM with the ***initialize-hsm*** command, you specify the HSM object identifier (ARN), the HSM security officer (SO) password, and several other parameters on the command line. The tool uses the ARN to look up the IP address of the HSM using the CloudHSM ***DescribeHSM*** API call, and then it connects to the HSM via SSH and issues Luna shell commands to initialize the HSM. The credentials you provide are used by the script, and are not shared with AWS. The CLI Tools source code is available under the Apache License v2 open source license, so you can review the code.

Q: Can the CloudHSM CLI Tools and API be used to configure several HSMs in a High Availability (HA) configuration?

Yes. The CloudHSM Command Line Interface Tools and API are designed to be used together to simplify the process of configuring and maintaining several HSMs in a high availability (HA) group.

The CloudHSM API uses an abstraction called a high availability partition group (HAPG) configuration object to simplify this configuration. You must use an HAPG configuration object when you are using CloudHSM for Amazon RDS, and you can optionally use it with your applications that use CloudHSM appliances in an HA configuration.

Q: Can I use HA Partition Groups across regions?

No. HA Partition Groups can only be configured for HSMs in the same region.

Q: On what operating systems can I use the CloudHSM Command Line Tools?

Amazon Linux. Please let us know if there are other operating systems on which you would like to use the tools.

Q: What are the network connectivity requirements for using the CloudHSM Command Line Interface Tools?

The Amazon Linux instance on which you use the CLI tools must have network reachability to your CloudHSM instances and to the CloudHSM API endpoint on the public Internet.

Are the CloudHSM Command Line Tools included in the AWS Command Line Interface?

Not at this time. The CloudHSM CLI Tools must be downloaded and installed separately.

CloudHSM API & SDK

Q: What can I do with the CloudHSM API & SDK?

You can create, modify, and delete CloudHSM instances, and get the status of your CloudHSM instances. What you can do with the API is limited to operations that AWS can perform with its HSM appliance administrator credentials. You control the security officer credentials which provide access to the contents of the HSM, so the API cannot access the contents of the HSM that are restricted to the security officer. To learn more, please see the CloudHSM Developer Guide for information about the API, or the Tools for Amazon Web Services page for more information about the SDK.

Support and Maintenance

Q: How is routine maintenance performed on HSM appliances?

AWS' routine maintenance procedures for HSM appliances are designed to avoid simultaneous downtime in multiple AZs in the same region.

AWS monitors and maintains the HSM appliances, and may correct minor configuration issues related to availability of the appliance. Such operations should not interfere with your use of the HSM appliance. If a management operation must be performed which could disrupt service (for example, if a security patch must be installed or the device must be rebooted), then AWS will usually attempt to contact you in advance to notify you of the pending change. AWS will not perform routine maintenance on HSM appliances in multiple AZs within the same region within the same 24-hour period.

In unforeseen circumstances, it is possible that AWS might perform emergency maintenance without prior notice. AWS will try to avoid this situation, as well as situations where emergency maintenance is performed within the same 24-hour period on HSM appliances in multiple AZs in the same region.

AWS strongly recommends that you use two or more CloudHSM instances in separate Availability Zones, and configured for high availability.

Q: I lost my SSH key for a CloudHSM instance, how can I reset it?

Contact AWS Support.

Q: I am having a problem with CloudHSM. What do I do?

Contact AWS Support.

AWS Key Management Service FAQ

KMS

General

What is AWS Key Management Service (KMS)?

AWS KMS is a managed encryption service that enables you to easily encrypt your data. AWS KMS provides a highly available key storage, management, and auditing solution for you to encrypt your data across AWS services and within your own applications.

Why should I use AWS KMS?

If you are a developer who needs to encrypt data in your applications, you should use the AWS SDKs with AWS KMS support to easily use and protect encryption keys. If you’re an IT administrator looking for a scalable key management infrastructure to support your developers and their growing number of applications, you should use AWS KMS to reduce your licensing costs and operational burden. If you’re responsible for proving data security for regulatory or compliance purposes, you should use AWS KMS to verify that data is encrypted consistently across the applications where it is used and stored.

How do I get started with AWS KMS?

The easiest way is to get started using AWS KMS is to check the box to encrypt your data within supported AWS services and use the default keys that are created in your account for each service. If you want further controls over the management of these keys, you can create keys in AWS KMS and assign them to be used in the supported AWS services when creating encrypted resources as well as use them directly within your own applications. AWS KMS can be accessed from the “Encryption Keys” section of the AWS Identity and Access Management (IAM) console for web-based access, and the AWS KMS Command Line Interface or AWS Software Development Kit for programmatic access. Visit the Getting Started page to learn more.

In what Regions is KMS available?

Availability is listed on our global Products and Services by Region page.

What key management features are available in AWS KMS?

You can perform the following key management functions in AWS KMS:

- Create keys with a unique alias and description
- Import your own keys
- Define which IAM users and roles can manage keys
- Define which IAM users and roles can use keys to encrypt and decrypt data
- Choose to have AWS KMS automatically rotate your keys on an annual basis
- Temporarily disable keys so they cannot be used by anyone
- Re-enable disabled keys
- Delete keys that you no longer use
- Audit use of keys by inspecting logs in AWS CloudTrail

How does AWS KMS work?

AWS KMS allows you to centrally manage and securely store your keys. You can generate keys in KMS or import them from your key management infrastructure. These keys can be used from within your applications and supported AWS services to protect your data, but the key never leaves KMS AWS. You submit data to AWS KMS to be encrypted, or decrypted, under keys that you control. You set usage policies on these keys that determine which users can use them to encrypt and decrypt data. All requests to use these keys are logged in AWS CloudTrail so you can understand who used which key when.

Where is my data encrypted if I use AWS KMS?

You can use AWS KMS to help encrypt data locally in your own applications or have it encrypted within a supported AWS service. You can use an AWS SDK with AWS KMS support to do the encryption wherever your applications run. You can also request a supported AWS service to encrypt your data as it is being stored. AWS CloudTrail provides access logs to allow you to audit how your keys were used in either situation.

Which AWS cloud services are integrated with AWS KMS?

AWS Key Management Service is seamlessly integrated with several other AWS services to make encrypting data in those services as easy as checking a box and selecting the master key you want to use. See the Product Details page for the list of AWS services currently integrated with KMS. All use of your keys within integrated services appears in AWS CloudTrail logs. See the AWS KMS Developer's Guide for more information on how integrated services use AWS KMS.

How do AWS cloud services use my keys to encrypt data?

AWS cloud services integrated with AWS KMS use a method called envelope encryption to protect your data. Envelope encryption is an optimized method for encrypting data that uses two different keys. A data key is generated and used by the AWS service to encrypt each piece of data or resource. The data key is encrypted under a master key that you define in AWS KMS. The encrypted data key is then stored by the AWS service. When you need your data decrypted by the AWS service, the encrypted data key is passed to AWS KMS and decrypted under the master key that was originally encrypted under so the service can then decrypt your data.

Why use envelope encryption? Why not just send data to AWS KMS to encrypt directly?

While AWS KMS does support sending data less than 4 KB to be encrypted, envelope encryption can offer significant performance benefits. When you encrypt data directly with KMS it must be transferred over the network. Envelope encryption reduces the network load for your application or AWS cloud service. Only the request and fulfillment of the data key through KMS must go over the network. Since the data key is always stored in encrypted form, it is easy and safe to distribute that key where you need it to go without worrying about it being exposed. Encrypted data keys are sent to AWS KMS and decrypted under master keys to ultimately allow you to decrypt your data. The data key is available directly in your application without having to send the entire block of data to AWS KMS and suffer network latency.

What's the difference between a key I create vs. default master keys created for me for use within AWS cloud services?

You have the option of selecting a specific master key to use when you want an AWS service to encrypt data on your behalf. A default master key specific to each service is created in your account as a convenience the first time you try to create an encrypted resource. This key is managed by AWS KMS but you can always audit its use in AWS CloudTrail. You can alternately create a customer master key in AWS KMS that you can then use in your own applications or from within a supported AWS service. AWS will update the policies on default master keys as needed to enable new features in supported services automatically. AWS does not modify policies on keys you create.

Why should I create a customer master key?

Creating a key in AWS KMS gives you more control than you have with default service master keys. When you create a customer master key, you can choose to use key material generated by KMS on your behalf or import your own key material, define an alias, a description, and opt-in to have the key automatically rotated once per year if it backed by key material generated by KMS. You also can define permissions on the key to control who can use and manage the key. Management and usage activity related to the key is available for audit in AWS CloudTrail.

Can I import keys into KMS?

Yes. You can import a copy of your key from your own key management infrastructure to KMS and use it with any integrated AWS service or from within your own applications.

When would I use an imported key?

You can use an imported key to get greater control over the creation, lifecycle management, and durability of your key in KMS. Imported keys are designed to help you meet your compliance requirements which may include the ability to generate or maintain a secure copy of the key in your infrastructure, and the ability to delete the imported copy of the key on demand from AWS infrastructure once you no longer need the key.

What type of keys can I import?

You can import 256-bit symmetric keys.

How is the key that I import into KMS protected in transit?

During the import, your key must be wrapped by a KMS-provided public key using one of the two RSA PKCS#1 schemes. This ensures that your encrypted key can only be decrypted by KMS.

What’s the difference between a key I import vs. a key generated for me by KMS?

There are two main differences between a key that you import vs. a key created for you by KMS:

1. You must securely maintain a copy of your imported keys in your key management infrastructure so that you can re-import them at any time. AWS ensures the availability, security, and durability of keys generated by KMS on your behalf until you schedule the keys for deletion.

2. You may set an expiration period for an imported key to automatically delete the key from KMS after the expiration period. You may also delete an imported key on demand without deleting the underlying customer master key. Further, you can manually disable or delete a customer master key with an imported key at any time. A key generated by KMS can only be disabled or scheduled for deletion, it cannot have an expiration time placed on it.

Can I rotate my keys?

Yes. You can choose to have KMS automatically rotate keys generated by KMS on your behalf every year. Automatic key rotation is not supported for imported keys. If you choose to import keys to KMS, you can manually rotate them whenever you want.

Do I have to re-encrypt my data after keys in AWS KMS are rotated?

If you choose to have KMS automatically rotate keys generated by KMS on your behalf, you don’t have to re-encrypt your data. AWS KMS keeps previous versions of keys to use for decryption of data encrypted under an old version of a key. All new encryption requests against a key in AWS KMS are encrypted under the newest version of the key.

If you manually rotate your keys, you may have to re-encrypt your data depending on your application’s configuration.

Can I delete a key from AWS KMS?

Yes. You can schedule a customer master key and associated metadata that you created in KMS for deletion, with a configurable waiting period from 7 to 30 days. This waiting period allows you to verify the impact of deleting a key on your applications and users that depend on it. The default waiting period is 30 days. You can cancel the deletion during the waiting period. The key cannot be used if it is scheduled for deletion until you cancel the deletion during the waiting period. The key gets deleted at the end of the configurable waiting period if you don’t cancel the deletion. Once a key gets deleted, you can no longer use it. All data protected under a deleted master key is inaccessible.

For customer master keys with imported key material, you can delete the key material without deleting the customer master key id or metadata in two ways. First, you can delete your imported key material on demand without a waiting period. Second, at the time of importing the key material into the customer master key, you may define an expiration time for how long AWS can use your imported key material before it is deleted. You can re-import your key material into the customer master key if you need to use it again.

What should I do if my imported key material has expired or I accidentally deleted it?

You can re-import your copy of the key material with a valid expiration period to KMS under the original customer master key so it can be used.

Can I be alerted that I need to re-import the key?

Yes. Once you import your key to a customer master key, you will receive an Amazon

CloudWatch Metric every few minutes that counts down the time to expiration of the imported key. You will also receive an Amazon CloudWatch Event once the imported key under your customer master key expires. You can build logic that acts on these metrics or events and automatically re-imports the key with a new expiration period to avoid an availability risk.

Can I use AWS KMS to help manage encryption of data outside of AWS cloud services?

Yes. AWS KMS is supported in AWS SDKs, AWS Encryption SDK, and the Amazon S3 Encryption Client to facilitate encryption of data within your own applications wherever they run. AWS SDK in the Java, Ruby, .NET, and PHP platforms support AWS KMS APIs. Visit the Developing on AWS website for more information.

Is there a limit to the number of keys I can create in AWS KMS?

You can create up to 1000 customer master keys per account per region. As both enabled and disabled customer master keys count towards the limit, we recommend deleting disabled keys that you no longer use. Default master keys created on your behalf for use within supported AWS services do not count against this limit. There is no limit to the number of data keys that can be derived using a master key and used in your application or by AWS services to encrypt data on your behalf. You may request a limit increase for customer master keys by visiting the AWS Support Center.

Billing

How will I be charged and billed for my use of AWS KMS?

With AWS KMS, you pay only for what you use, there is no minimum fee. There are no set-up fees or commitments to begin using the service. At the end of the month, your credit card will automatically be charged for that month’s usage.

You are charged for all customer master keys you create, and for API requests made to the service each month above a free tier.

For current pricing information, please visit the AWS KMS pricing page.

Is there a free tier?

Yes. With the AWS Free Usage Tier you can get started with AWS KMS for free in all regions. Default master keys created on your behalf are free to store in your account. There is a free tier for usage as well that provides a free number of requests to AWS KMS each month. For current information on pricing, including the free tier, please visit the AWS KMS pricing page.

Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. You can learn more here.

Security

Who can use and manage my keys in AWS KMS?

AWS KMS enforces usage and management policies that you define. You choose to allow AWS Identity and Access Management (IAM) users and roles from your account or other accounts to use and manage your keys.

Can AWS employees access my keys in AWS KMS?

AWS KMS is designed so that no one has access to your master keys. The service is built on systems that are designed to protect your master keys with extensive hardening techniques such as never storing plaintext master keys on disk, not persisting them in memory, and limiting which systems can connect to the device. All access to update software on the service is controlled by a multi-level approval process that is audited and reviewed by an independent group within Amazon.

More details about these security controls can be found in theAWS KMS Cryptographic Details whitepaper. In addition, you can request a copy of the Service Organization Controls (SOC) report available from AWS Compliance to learn more about security controls AWS uses to protect your data and master keys.

Can I use KMS to help me comply with the encryption and key management requirements in the Payment Card Industry Data Security Standard (PCI DSS 3.1)?

Yes. KMS has been validated as having the functionality and security controls to help you meet the encryption and key management requirements (primarily referenced in sections 3.5 and 3.6 of the PCI DSS 3.1).

For more details on PCI DSS compliant services in AWS, you can read thePCI DSS FAQs.

How does AWS KMS secure the data keys I export and use in my application?

You can request that AWS KMS generate data keys that can be returned for use in your own application. The data keys are encrypted under a master key you define in AWS KMS so that you can safely store the encrypted data key along with your encrypted data. Your encrypted data key (and therefore your source data) can only be decrypted by users with permissions to use the original master key used in encrypting the data key.

What length of keys does AWS KMS generate?

Master keys in AWS KMS are 256-bits in length. Data keys can be generated at 128-bit or 256-bit lengths and encrypted under a master key you define. AWS KMS also provides the ability to generate random data of any length you define suitable for cryptographic use.

Can I export a master key from AWS KMS and use it in my own applications?

No. Master keys are created and used only within AWS KMS to help ensure their security, enable your policies to be consistently enforced, and provide a centralized log of their use.

What geographic region are my keys stored in?

Keys are only stored and used in the region in which they are created. They cannot be transferred to another region. For example; keys created in the EU-Central (Frankfurt) region are only stored and used within the EU-Central (Frankfurt) region.

How can I tell who used or changed the configuration of my keys in AWS KMS?

Logs in AWS CloudTrail will show requests on your master keys, including both management requests (e.g. create, rotate, disable, policy edits) and cryptographic requests (e.g. encrypt/decrypt). Turn on AWS CloudTrail in your account to view these logs.

How does AWS KMS compare to AWS CloudHSM?

AWS CloudHSM provides you with a dedicated hardware device installed in your Amazon Virtual Private Cloud (VPC) that provides a FIPS 140-2 Level 2 validated single-tenant HSM to store and use your keys. You have total control over your keys and the application software that uses them with CloudHSM.

AWS KMS allows you to control the encryption keys used by your applications and supported AWS services in multiple regions around the world from a single console. Centralized management of all your keys in AWS KMS lets you enforce who can use your keys, when they get rotated, and who can manage them. AWS KMS integration with AWS CloudTrail gives you the ability to audit the use of your keys to support your regulatory and compliance activities.

AWS WAF FAQ

General

1. What is AWS WAF?

AWS WAF is a web application firewall that helps protect web applications from attacks by allowing you to configure rules that allow, block, or monitor (count) web requests based on conditions that you define. These conditions include IP addresses, HTTP headers, HTTP body, URI strings, SQL injection and cross-site scripting.

2. How does AWS WAF block or allow traffic?

As Amazon CloudFront receives requests for your web sites, it forwards those requests to AWS WAF for inspection against your rules. Once a request meets a condition defined in your rules, AWS WAF instructs Amazon CloudFront to either block or allow the request based on the action you define.

3. How does AWS WAF protect my web site or application?

AWS WAF is tightly integrated with Amazon CloudFront, AWS's CDN that AWS customers commonly use to deliver content for their websites and applications. Your rules run in all AWS Edge Locations, located around the world close to your end users. This means security doesn't come at the expense of performance. Blocked requests are stopped before they reach your web servers.

4. Can I use AWS WAF to protect web sites not hosted in AWS?

Yes, AWS WAF is integrated with Amazon CloudFront, which supports custom origins outside of AWS.

5. What types of attacks can AWS WAF help me to stop?

AWS WAF helps protects your website from common attack techniques like SQL injection and Cross-Site Scripting (XSS). In addition, you can create rules that can block attacks from specific user-agents, bad bots, or content scrapers. See the AWS WAF Developer Guide for examples.

6. Can I get a history of all AWS WAF API calls made on my account for security, operational or compliance auditing?

Yes. To receive a history of all AWS WAF API calls made on your account, you simply turn on AWS CloudTrail in the CloudTrail's AWS Management Console. For more information, visit AWS CloudTrail home page or visit the AWS WAF Developer Guide.

7. Does AWS WAF support IPv6?

Yes, support for IPv6 allows the AWS WAF to inspect HTTP/S requests coming from both IPv6 and IPv4 addresses.

8. Does IPSet match condition for an AWS WAF Rule support IPv6?

Yes, you can setup new IPv6 match condition(s) for new and existing WebACLs, as per the documentation.

9. Can I expect to see IPv6 address appear in the AWS WAF sampled requests where applicable?
Yes. The sampled requests will show the IPv6 address where applicable.

10. Can I use IPv6 with all AWS WAF features?
Yes. You will be able to use all the existing features for traffic both over IPv6 and IPv4 without any discernable changes to performance, scalability or availability of the service.

AWS WAF Configuration

1. Can I configure custom error pages?
Yes, you can configure CloudFront to present a custom error page when requests are blocked. Please see the CloudFront Developer Guide for more information

2. How long does it take AWS WAF to propagate my rules?
After an initial setup, adding or changing to rules typically takes around a minute to propagate worldwide.

3. How can I see if my rules are working?
AWS WAF includes two different ways to see how your website is being protected: one-minute metrics are available in CloudWatch and Sampled Web Requests are available in the AWS WAF API or management console. These allow you to see which requests were blocked, allowed, or counted and what rule was matched on a given request (i.e., this web request was blocked due to an IP address condition, etc.). For more information see the AWS WAF Developer Guide.

4. How can I test my rules?
AWS WAF allows you to configure a “count” action for rules, which counts the number of web requests that meet your rule conditions. You can look at the number of counted web requests to estimate how many of your web requests would be blocked or allowed if you enable the rule.

5. How long are Real-Time Metrics and Sampled Web Requests stored?
Real-Time Metrics are stored in Amazon CloudWatch. Using Amazon CloudWatch you can configure the time period in which you want to expire events. Sampled Web Requests are stored for up to 2 hours.

6. Can AWS WAF inspect HTTPS traffic?
Yes. AWS WAF helps protect applications running on Amazon CloudFront and can inspect web requests transmitted over HTTP or HTTPS.

AWS IoT FAQ

Introduction

Q: What is AWS IoT?

AWS IoT is a managed cloud platform that lets connected devices easily and securely interact with cloud applications and other devices. AWS IoT can support billions of devices and trillions of messages, and can process and route those messages to AWS endpoints and to other devices reliably and securely. With AWS IoT, your applications can keep track of and communicate with all your devices, all the time, even when they aren't connected.

AWS IoT makes it easy to use AWS services like AWS Lambda, Amazon Kinesis, Amazon S3, Amazon Machine Learning, Amazon DynamoDB, Amazon CloudWatch, AWS CloudTrail, and Amazon Elasticsearch Service with built-in Kibana integration, to build IoT applications that gather, process, analyze and act on data generated by connected devices, without having to manage any infrastructure.

Q: What does AWS IoT offer?

Connectivity between devices and the AWS cloud. First, with AWS IoT you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. The AWS IoT service supports standard communication protocols (HTTP, MQTT, and WebSockets are supported currently). Communication is secured using TLS.

> **Connectivity between devices and the AWS cloud.** First, with AWS IoT you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. The AWS IoT service supports standard communication protocols (HTTP, MQTT, and WebSockets are supported currently). Communication is secured using TLS.
>
> **Processing data sent from connected devices**. Secondly, with AWS IoT you can continuously ingest, filter, transform, and route the data streamed from connected devices. You can take actions based on the data and route it for further processing and analytics.
>
> **Application interaction with connected devices.** Finally, the AWS IoT service accelerates IoT application development. It serves as an easy to use interface for applications running in the cloud and on mobile devices to access data sent from connected devices, and send data and commands back to the devices.

Q: How does AWS IoT work?

Connected devices, such as sensors, actuators, embedded devices, smart appliances, and wearable devices, connect to AWS IoT over HTTPS, WebSockets, or secure MQTT. Included in

AWS IoT is a **Device Gateway** that allows secure, low-latency, low-overhead, bi-directional communication between connected devices and your cloud and mobile applications.

The AWS IoT service also contains a **Rules Engine** which enables continuous processing of data sent by connected devices. You can configure rules to filter and transform the data. You also configure rules to route the data to other AWS services such as DynamoDB, Kinesis, Lambda, SNS, SQS, CloudWatch, Elasticsearch Service with built-in Kibana integration, as well as to non-AWS services, via Lambda for further processing, storage, or analytics.

There is also a **Device Registry** where you can register and keep track of devices connected to AWS IoT, or devices that may connect in the future. **Device Shadows** in the AWS IoT service enable cloud and mobile applications to query data sent from devices and send commands to devices, using a simple REST API, while letting AWS IoT handle the underlying communication with the devices. The shadows accelerate application development by providing a uniform interface to devices, even when they use one of the several IoT communication and security protocols with which the applications may not be compatible. Shadows also accelerate application development by providing an always available interface to devices even when the connected devices are constrained by intermittent connectivity, limited bandwidth, limited computing ability or limited power.

Communication with AWS IoT is secure. The service requires all of its clients (connected devices, server applications, mobile applications, or human users) to use strong authentication (X.509 certificates, AWS IAM credentials, or 3rd party authentication via AWS Cognito). All communication is encrypted. AWS IoT also offers fine-grained authorization to isolate and secure communication among authenticated clients.

Similar to other AWS services, users can access AWS IoT via the AWS Management Console and the CLI. Applications can access AWS IoT easily with the AWS SDKs available for several programming languages. AWS IoT further simplifies development and operations of IoT applications by integrating with Amazon CloudWatch.

To simplify the development of code running on connected devices, AWS IoT provides open-source device SDKs for C, Node.js, and the Arduino Yún platform. AWS IoT has also partnered with hardware manufacturers to make the AWS IoT Device SDKs available on several IoT, embedded OS, and micro-controller platforms.

Q: Which AWS regions is AWS IoT service available in?

AWS IoT is currently available in the following AWS regions:

- US East (N. Virginia)
- US West (Oregon)
- EU (Ireland)
- EU (Frankfurt)

- Asia Pacific (Sydney)
- Asia Pacific (Seoul)
- Asia Pacific (Tokyo)
- Asia Pacific (Singapore)

You can use AWS IoT regardless of your geographic location, as long as you have access to one of the above AWS regions.

Q: How do I get started with using AWS IoT?

Use the AWS IoT console or refer to the Quickstart section of our developer guide to test drive the AWS IoT service in minutes.

Also, take a look at the AWS-powered Starter Kits provided by our partners.

Refer to the AWS IoT documentation for further details.

Accessing AWS IoT

Q: What are the ways for accessing AWS IoT?

You can use the AWS Management Console, the AWS SDKs, the AWS CLI, and the AWS IoT APIs to access the AWS IoT service. Connected devices can use the AWS IoT Device SDKs to simplify the communication with the AWS IoT service.

The AWS IoT APIs and commands are largely divided into control plane operations and data plane operations. The control plane operations enable you to do tasks such as configuring security, registering devices, configuring rules for routing data, and setting up logging. The data plane operations enable you to ingest data from connected devices into AWS IoT with low latency and high throughput rate at a large scale.

Q: What communication and authentication protocols does AWS IoT support?

For control plane operations, AWS IoT supports HTTPS. For data plane operations, AWS IoT supports HTTPS, WebSockets, and secure MQTT – a protocol often used in IoT scenarios.

HTTPS and WebSockets requests sent to AWS IoT are authenticated using AWS IAM or AWS Cognito, both of which support the AWS SigV4 authentication. If you are using the AWS SDKs or the AWS CLI, the SigV4 authentication is taken care of for you under the hood. HTTPS requests can also be authenticated using X.509 certificates. MQTT messages to AWS IoT are authenticated using X.509 certificates.

With AWS IoT you can use AWS IoT generated certificates, as well as those signed by your preferred Certificate Authority (CA).

Q: Can devices that are NOT directly connected to the Internet access AWS IoT?

Yes, via a physical hub. Devices connected to a private IP network and devices using non-IP radio protocols such as ZigBee or Bluetooth LE can access AWS IoT as long as they have a physical hub as an intermediary between them and AWS IoT for communication and security.

Q: How should applications access AWS IoT?

Applications connecting to AWS IoT largely fall in two categories: 1. companion apps and 2. server applications. Companion apps are mobile or client-side browser applications that interact with connected devices via the cloud. A mobile app that lets a consumer remotely unlock a smart lock in the consumer's house is an example of a companion app. Server applications are designed to monitor and control a large number of connected devices at once. An example of a server application would be a fleet management website that plots thousands of trucks on a map in real-time.

AWS IoT enables both companion apps and server applications to access connected devices via uniform, RESTful APIs. Applications also have the option to use pub/sub to communicate directly with the connected devices.

Typically the companion apps would authenticate using end-user identities which are managed either by your own identity store or a third party identity provider such as Facebook and Login with Amazon. For companion apps, use Amazon Cognito, which integrates with several identity providers. Cognito identities can be authorized to access AWS IoT, and their access can be restricted only to the resources relevant to them. For example, as a connected washing machine manufacturer, you can authorize your consumers to access your AWS IoT information pertaining only to their individual washing machines.

Server applications (such as a mapping application running on Amazon EC2) can use IAM roles to access AWS IoT.

Q: Can I get a history of AWS IoT API calls made on my account for security analysis and operational troubleshooting purposes?

Yes, to receive a history of AWS IoT API calls made on your account, you simply turn on CloudTrail in the AWS Management Console.

Device Gateway

Q: What is the AWS IoT Device Gateway?

The Device Gateway forms the backbone of communication between connected devices and the cloud capabilities such as the AWS IoT Rules Engine, Device Shadows, and other AWS and 3rd-party services.

The Device Gateway supports the pub/sub messaging pattern, which enables scalable, low-latency, and low-overhead communication. It is particularly useful for IoT scenarios where billions of devices are expected to communicate frequently and with minimal delay. Pub/sub involves clients publishing messages on logical communication channels called ‘topics’ and clients subscribing to topics to receive messages. The device gateway enables the communication between publishers and subscribers. Traditionally, organizations have had to provision, operate, scale, and maintain their own servers as device gateways to take advantage of pub/sub. AWS IoT service has eliminated this barrier by providing the AWS IoT device gateway.

The Device Gateway scales automatically with your usage, without any operational overhead for you. AWS IoT supports secure communication with the device gateway, AWS-account level isolation, as well as fine-grained authorization within an AWS account. The device gateway currently supports publish and subscribe over secure MQTT and WebSockets, as well as publish over HTTPS.

Q: What is MQTT?

MQTT is a lightweight pub/sub protocol, designed to minimize network bandwidth and device resource requirements. MQTT also supports secure communication using TLS. MQTT is often used in IoT use cases. MQTT v3.1.1 is an OASIS standard, and the AWS IoT device gateway supports most of the MQTT specification.

Rules Engine

Q: What is the AWS IoT Rules Engine?

The AWS IoT Rules Engine enables continuous processing of inbound data from devices connected to the AWS IoT service. You can configure rules in the Rules Engine in an intuitive, SQL-like syntax to automatically filter and transform inbound data. You can further configure rules to route data from the AWS IoT service to several other AWS services as well as your own or 3rd party services.
Here are just a few example use cases of rules:
• Filtering and transforming incoming messages and storing them as time series data in DynamoDB.
• Sending a push notification via SNS when the data from a sensor crosses a certain threshold.
• Saving a firmware file to S3
• Processing messages simultaneously from a multitude of devices using Kinesis
• Invoke Lambda to do custom processing on incoming data
• Sending a command to a group of devices with an automated republish

Q: How are the rules defined and triggered?

An AWS IoT rule consists of two main parts:

- A SQL statement that specifies the pub/sub topics to apply the rule on, data transformation to perform, if any, and the condition under which the rule should be executed. The rule is applied on every message published on the specified topics.

- An actions list that defines the actions to take when the rule is executed, that is, when an incoming message matches the condition specified in the rule.

Rule definitions use a JSON-based schema. You can directly edit the JSON or use the rules editor in the AWS Management Console.
As an example, here is a rule for saving temperature data from a sensor to DynamoDB whenever the temperature is above 50:

```
{
    "sql": "SELECT * from 'iot/tempSensors/#' WHERE temp > 50",
    "description": "Rule to save sensor data when temperature is about 50",
    "actions": [
    {
        "dynamoDB": {
        "tableName": "HighTempTable",
        "roleArn": "arn:aws:iam::your-aws-account-id:role/dynamoPut",
        "hashKeyField": "key",
        "hashKeyValue": "${topic(3)}",
        "rangeKeyField": "timestamp",
        "rangeKeyValue": "${timestamp()}"
    }
   }
 ]
}
```

Sensors in this example are publishing on their topics under “iot/tempSensors/”. The first line of the rule defines the SQL SELECT statement used to query on the “iot/tempSensors/#” topic. It contains a WHERE clause that extracts the value of a ‘temp’ field in the message’s payload and checks if it passes the condition ‘greater than 50’. If the condition is met, the data is stored in the specified DynamoDB table. The example uses built-in functions for tasks such as traversing the message payload and getting current time.

Q: Where can I learn more about rules?

You can learn more about rule here:AWS IoT Rules documentation

Device Registry and Device Shadows

Q: What is the AWS IoT Device Registry and what should I use it for?

IoT scenarios can range from a small number of mission-critical devices to large fleets of devices. The AWS IoT Device Registry allows you to organize and track those devices. You can maintain a logical handle in the Device Registry for every device you are connecting to AWS IoT. Each device in the Device Registry can be uniquely identified and can have metadata such as model numbers, support contact, and certificates associated with it. You can search for connected devices in the Device Registry based on the metadata.

Q: What is a Thing Type?

Thing Types allow you to effectively manage your catalogue of devices by defining common characteristics for devices that belong to the same device category. In addition, a Thing associated with a Thing Type can now have up to 50 attributes including 3 searchable attributes.

Q: What is Simplified Permission Management?

This feature allows you to easily manage permission policies for a large number of devices by using variables that reference Registry or X.509 certificate properties. The integration of Registry and Certificate properties with device policies offers the benefits listed below:

- You can now reference Device Registry properties in device permission policies. Referencing device properties defined in the Device Registry allows your policies to reflect any changes made in the Device Registry. For example, by referencing the Thing Attribute named “building-address” as a variable in the policy, devices will automatically inherit a new set of permissions when they move buildings.

- You can share a single generic policy for multiple devices. A generic policy can be shared among the same category of devices instead of creating a unique policy per device. For example, a policy that references the “serial-number” as a variable, can be attached to all the devices of the same model. When devices of the same serial number connect, policy variables will be automatically substituted by their serial-number.

Q: What are the Device Shadows?

The Device Shadows enable cloud and mobile applications to easily interact with the connected devices registered in AWS IoT. A Device Shadow in AWS IoT contains properties of a connected device. You can define any set of properties applicable to your use case. For example, for a smart light bulb, you might define ‘on-or-off’, ‘color’, and ‘brightness’ as the properties. The connected device is expected to report the actual values of those properties, which are stored in the Device Shadow. Applications get and update the properties simply by using a RESTful API provided by the AWS IoT service. The AWS IoT service and the AWS IoT Device SDKs take care of synchronizing property values between the connected device and its shadow in AWS IoT.

Q: Do I have to use Device Registry and Device Shadows?

You can have applications communicate directly to the connected devices using the Device Gateway and/or the Rules Engine in AWS IoT. However, we recommend using the Device Registry and Device Shadows since they offer richer and more structured development and management experience that lets you focus on the unique value you want to create for your customers rather than having to focus on the underlying communication and synchronization between the connected devices and the cloud.

Q: What is the lifecycle of a device and its Shadow in AWS IoT?

- You register a device (such as a light bulb) in the Device Registry.
- You program connected device to publish a set of its property values or 'state ("I am ON and my color is RED") to the AWS IoT service.
- The last reported state is stored in the device's Shadow in AWS IoT.
- An application (such as a mobile app controlling the light bulb) uses a RESTful API to query AWS IoT for the last reported state of the light bulb, without the complexity of communicating directly with the light bulb.
- When a user wants to change the state (such as turning the light bulb from ON to OFF), the application uses a RESTful API to request an update, i.e. sets a 'desired' state for the device in AWS IoT. AWS IoT takes care of synchronizing the desired state to the device.
- The application gets notified when the connected device updates its state to the desired state.

Q: Where can I learn more about Device Registry and Device Shadows?

For more information on the Device Registry, seeAWS IoT Device Registry. For more information on Shadows, see AWS IoT Device Shadows.

Security and Access Control

Q: Can I configure fine-grained authorization in AWS IoT?

Yes. Similar to other AWS services, in AWS IoT you have fine-grained control over the set of API actions each identity is authorized to invoke. In addition, you have fine-grained control over the pub/sub topics that an identity can publish or subscribe to, as well as over the devices and shadows in the Device Registry that an identity can access.

Q: Where can I learn more about Security and Access Control in AWS IoT?

For more information, see AWS IoT Security and Identity.

Q: What is Just-in-time registration of certificates?

Just-in-time registration (JITR) of device certificates expands on the"Use Your Own Certificate" feature launched in April 2016 by simplifying the process of enrolling devices with AWS IoT. Prior to support for JITR, the device enrollment process required two steps: first, registering the Certificate Authority (CA) certificate to AWS IoT, then individually registering the device certificates that were signed by the CA. Now, with JITR you can complete the second step by auto-registering device certificates when devices connect to AWS IoT for the first time. This saves time spent on registering device certificates and allows devices to remain off-line during the manufacturing process. To further automate IoT device provisioning, you can create an AWS IoT rule with a Lambda action that activates the certificates and attaches policies. For more information, visit the Internet of Things Blog on AWS or Developer Documentation.

AWS IoT Device SDK

Q: What is the AWS IoT Device SDK?

The AWS IoT Device SDKs simplify and accelerate the development of code running on connected devices (micro-controllers, sensors, actuators, smart appliances, wearable devices, etc.). First, devices can optimize the memory, power, and network bandwidth consumption by using the Device SDKs. At the same time, Device SDKs enable highly secure, low-latency, and low-overhead communication with built-in TLS, WebSockets, and MQTT support. The Device SDKs also accelerate IoT application development by supporting higher level abstractions such as synchronizing the state of a device with its shadow in the AWS IoT service.

AWS IoT Device SDKs are freely available as open-source projects. For more details visit our Device SDK page.

Q: Which programming languages and hardware platforms does the AWS IoT Device SDK support?

AWS currently offers the AWS IoT Device SDKs for C and Node.js languages, as well as for the Arduino Yún platform.

In addition, several hardware manufacturers have partnered with AWS to make the AWS IoT Device SDKs available on their respective platforms. You can find out more about the hardware platforms on our Getting Started page.

Lastly, AWS IoT Device SDKs are open-source. You can port them to the languages and hardware platforms of your choice if they are not supported already.

Q: Should I use AWS IoT Device SDK or the AWS SDKs?

The AWS IoT Device SDK complements the AWS SDKs. IoT projects often involve code running on micro-controllers and other resource-constrained devices. However, IoT projects often include application running in the cloud and on mobile devices that interact with the micro-controllers/resource-constrained devices. AWS IoT Device SDKs are designed to be used on the micro-controllers/resource-constrained devices, while the AWS SDKs are designed for cloud and mobile applications.

Q: Where can I learn more about AWS IoT Device SDK?

For more information on the AWS IoT Device SDKs, seeAWS IoT Device SDKs.

Billing

Q: Is the AWS IoT service available in AWS Free Tier?

Yes. As part of the AWS Free Tier, AWS IoT offers 250,000 messages per month at no charge, for the first 12 months.

Q: How much does AWS IoT service cost?

Please visit our pricing page for information.

Amazon Lumberyard

General

Q. What is Amazon Lumberyard?
Amazon Lumberyard is a free, crossplatform AAA game engine deeply integrated with AWS and Twitch – with full source code provided. Whether you are a major studio, an indie developer, a student, or a hobbyist, Lumberyard provides a growing set of tools to create the highest-quality games, connect your games to the vast compute and storage of the AWS Cloud, and engage fans on Twitch. Lumberyard helps developers build beautiful worlds, make realistic characters, and create stunning real-time effects. With Lumberyard's visual scripting tool, even non-technical game developers can add cloud-connected features to a game in minutes (such as a community news feed, daily gifts, or server-side combat resolution) through a drag-and-drop GUI interface. Lumberyard is also integrated with Amazon GameLift, a new AWS service for deploying, operating, and scaling session-based multiplayer games. With Amazon GameLift, Amazon Lumberyard developers can quickly scale high-performance game servers up and down to meet player demand, without any additional engineering effort or upfront costs.

Amazon Lumberyard is free, with no seat licenses, royalties, or subscriptions required. With Amazon Lumberyard, developers only pay standard AWS fees for the AWS services they choose to use. With Amazon GameLift, you simply pay for the standard AWS fees for Amazon EC2, Amazon EBS, and data transfer you actually use, plus a small fee per Daily Active User.

Q. How does Amazon make money with Lumberyard?

Lumberyard is free, including source. We make money when you use other AWS services. We built Lumberyard to make it faster and easier to build fantastic live, multiplayer, community-driven games – which naturally connect to the cloud to provide these features to players. However, there is no requirement to connect your game to the cloud. There are also no seat fees, subscription fees, or requirements to share revenue. You pay only for the infrastructure resources you choose to use. For full licensing details, see our Licensing FAQs below.

Q. Is Amazon Game Studios using Lumberyard to build games?

Yes, all of our games are built with Lumberyard. Check them out on the Amazon Game Studios website.

Q. Is Lumberyard based on CryEngine?

Lumberyard is made up of proven technology from CryEngine, AWS, Twitch, and Double Helix. We've hired some of the best game technologists in the world, who have already made over 996 additions, fixes, and improvements to Lumberyard. For example, we've integrated a brand new networking layer, GridMate, so your engineers can more easily build low-latency multiplayer games with large numbers of players. We've introduced Cloud Canvas, which enables your engineers and technical designers with little to no backend experience to build live online game features, such as community news feeds, sharing scores, and server-side combat resolution, in minutes using Lumberyard's visual scripting system. We've also integrated Lumberyard with Amazon GameLift, so you can deploy, scale, and operate session-based multiplayer games. We've built a new component entity system so that you can easily populate and define the behaviors of the game world by creating entities and defining their behavior by adding components using drag-and-drop workflows in the Lumberyard Editor, and added a new code generation system to allow you to annotate your C++ code and generate the code you need. We've advanced the engine to include support for mobile devices, including support for Metal. We've created a new launcher and project configurator so your team can get set up without engineering help. We've also created new workflows so your artists can iterate faster and create higher-quality content, including a new particle effects editor, new FBX mesh importer, 2D/UI editor, and cross-platform asset pipeline. Please see our full release notes of additions, fixes, and improvements to learn more, and tune in to our GameDev Blog for more news on what we are working on.

Q. Do I really get source code access to Lumberyard?

Yes. Access to full C++ source code is included with the download of Lumberyard.

Q. What kind of support is available for Lumberyard?

All Lumberyard customers have access to documentation, tutorials, forums, and samples and assets. Additional support for Lumberyard is available via AWS Premium Support plans.

Q. Can I use Lumberyard for non-game purposes, such as architecture, simulations, and animated movies?

Yes, please do.

Q. What are the system requirements for building a game with the Lumberyard Editor and tools?
We recommend you use a PC with Windows 7 64-bit, 8+ GB RAM, 60 GB of storage, a 3GHz+ quad-core processor, and a 2+ GB DX11+ compatible video card. Windows 10 64-bit is also supported.

Q. What device platforms does Lumberyard support?
Lumberyard currently supports PC, Xbox One, PlayStation 4, iOS (iPhone 5S+), and Android (Samsung Note 4 and equivalents). Support for broader mobile hardware is coming soon, along with additional support for Mac and Linux. Please note that Sony and Microsoft only permit developers who have passed their screening process to develop games for their console platforms.

Q. Does Lumberyard support VR?
Yes. We currently support Oculus Rift, HTC Vive, and OSVR. Because VR is a rapidly evolving area, we've built Lumberyard's VR support to be modular, meaning you can add support for new HMDs without writing code, helping you support new HMDs as they are released. To find out more about our modular VR system, check out our blog here.

Q. How do I get started with Xbox and PlayStation game development?
If you are a licensed Microsoft Xbox developer, please e-mail your Name, Studio name, and the licensed e-mail address to lumberyard-consoles@amazon.com. If you are a licensed Sony PlayStation developer, please visit SCE DevNet. Under the Middleware Directory click "Confirm Status" for Amazon Lumberyard.

Lumberyard and AWS

Q. If I build a single-player game that uses no cloud connectivity, do I have to pay to use the engine?
No, in this case you would pay us nothing.

Q. Do I need an AWS account to use Lumberyard?
No, but by downloading or using Lumberyard, you agree to the AWS Customer Agreement and Lumberyard Service Terms. If you want to use Amazon GameLift or Lumberyard’s Cloud Canvas to build connected gameplay features, you or someone from your game team needs to register for an AWS account and provision services to your account.

Q. Do I have to run my game on AWS?
No. If you own and operate your own private servers, you do not need to use AWS. You also don’t need to use AWS if your game does not use any servers. For example, if you release a free-standing single-player or local-only multiplayer game, you pay us nothing.

Q. Is there a surcharge or other additional fee over and above AWS service rates for Lumberyard customers?
No.

Q. How do I authorize my team of developers to use Cloud Canvas and AWS via the Lumberyard Editor?
To enable team members to access AWS through Cloud Canvas, you first need to create an IAM user for your team members and generate access keys and secret keys. Your team members can enter these keys in the Lumberyard Editor's Credentials Manager under the AWS menu. For more information, please see the Cloud Canvas and IAM documentation.

Q. Can I grant certain team members permissions or restrictions to access specific AWS services in Cloud Canvas?
Yes, Cloud Canvas lets you configure permissions so your development, test, and release resources can have different access restrictions. Cloud Canvas creates AWS IAM Managed Policies, which can be used to grant access to the AWS IAM Users and Groups you choose. You can customize the permissions by editing your Cloud Canvas configuration in the Lumberyard Editor and update the user and group assignments in the AWS Management Console as needed.

Q. Which AWS services are available in Cloud Canvas?
Cloud Canvas enables you to use DynamoDB, S3, Cognito, SQS, SNS, and Lambda via the Lumberyard Flow Graph visual scripting tool.

Mods

Q. Can I include Lumberyard's tools so my players can build mods for my game?
Yes. Your right to redistribute Lumberyard in your game includes the right to redistribute pieces of the development environment in your game too. A list of redistributable components is included in the documentation. These rights also apply to companion products that you make available to end users to modify and create derivative works of your game. If you want to distribute Lumberyard components in source code form, please contact us.

Licensing

Q. What are the license terms for Lumberyard?
Your use of Lumberyard is governed by the AWS Customer Agreement and Lumberyard Service Terms.

Q. Do I have to sell my Lumberyard game on Amazon?
No, you can sell your game wherever you'd like. Of course, we'd love to see your game on

Amazon, and you can find information about publishing PC and Mac games on Amazon here and publishing on the Amazon Appstore here.

Q. Can I take Lumberyard and make my own game engine and distribute it?
No. While you may maintain an internal version of Lumberyard that you have modified, you may not distribute that modified version in source code form, or as a freestanding game engine to third parties. You also may not use Lumberyard to distribute your own game engine, to make improvements to another game engine, or otherwise compete with Lumberyard or Amazon GameLift.

Q. Is Lumberyard "open source"?
No. We make the source code available to enable you to fully customize your game, but your rights are limited by the Lumberyard Service Terms. For example, you may not publicly release the Lumberyard engine source code, or use it to release your own game engine.

Q: Can I make plugins or tools for Lumberyard?
Yes. However, please note that if your plugin incorporates Lumberyard code, then you must follow the requirements related to distributing Lumberyard Materials in the Service Terms, for example, to not distribute Lumberyard Materials in source code form. If your plugin merely calls Lumberyard functions or APIs as part of its operation, then the distribution requirements would not apply.

Q: Can I redistribute source code modifications to Lumberyard?
Yes, you can redistribute up to 50 lines of source code, on forums (including the official Lumberyard contribution forum), or elsewhere. You may also share modifications with your contractors and publishers working on your game, as set out in the Service Terms. Otherwise, you may not release Lumberyard engine source code.

Q. Can my Lumberyard game connect to services like Steamworks, Xbox Live, PSN, Apple Game Center, Google Play Games, or console social services?
Yes. Your game may read and write data to platform services and public third-party game services for player save state, identity, social graph, matchmaking, chat, notifications, achievements, leaderboards, advertising, player acquisition, in-game purchasing, analytics, and crash reporting.

Q. Can my game use an alternate web service instead of AWS?
No. If your game servers use a non-AWS alternate web service, we obviously don't make any money, and it's more difficult for us to support future development of Lumberyard. By "alternate web service" we mean any non-AWS web service that is similar to or can act as a replacement for Amazon EC2, Amazon Lambda, Amazon DynamoDB, Amazon RDS, Amazon S3, Amazon EBS, Amazon EC2 Container Service, or Amazon GameLift. You can use hardware you own and operate for your game servers.

Q. Is it okay for me to use my own servers?

Yes. You can use hardware you own and operate for your game.

Q. Can I use the game assets that are included with Lumberyard in my game?
Yes. Lumberyard includes asset packs that you can use in your games and prototypes. We also provide additional high-fidelity assets and samples that you may find useful on our website.

Q: Can I redistribute assets from Lumberyard or Lumberyard sample projects?
Not on their own. You can modify and/or redistribute Lumberyard sample assets such as audio, textures, meshes, animations, game data files, and scripts as part of your game. But you can't, for example, resell Lumberyard assets in an asset store.

Q. Can I use Lumberyard in a way not permitted by the Service Terms?
Please contact us if you would like to use Lumberyard in a way that is not permitted by the Lumberyard Service Terms.

Q. Does Lumberyard support integrations with third-party middleware?
Yes. Lumberyard is already integrated with popular middleware, including Perforce, Wwise, Substance, and more. If you're a middleware provider interested in integrating with Lumberyard, please contact us.

Registration

Q. Where can I tell you about my Lumberyard game?
Please register your Lumberyard project here before it is released.

Other

Q. How do I submit feedback or suggestions?
Please visit our GameDev Forums or email us your feedback.

Q. I'd love to join your team. Are you hiring?
Yes, our team is growing, and we'd love to hear from you if you're interested in joining our team. Check out our careers page to learn more.

Amazon GameLift

GameLift

General

Q. What is Amazon GameLift?

Amazon GameLift is a managed service for deploying, operating, and scaling session-based multiplayer game servers in the cloud, with no upfront costs. You can deploy your first game server in the cloud in just minutes, saving up to thousands of hours in upfront software development and lowering the technical risks that often cause developers to cut cloud-based multiplayer features from their designs. Built on AWS's proven computing environment, Amazon GameLift lets you scale high performance game servers up and down to meet player demand. You pay only for the capacity you use, so you can get started whether you're working on a new game idea or running a game with millions of players.

Q. How does Amazon GameLift help me?

In the past, most multiplayer games used peer-to-peer networking to connect players over the Internet. At the start of each multiplayer session, the game would select a player's computer to host the server. All of the other players would connect to this server. This approach resulted in several problems (e.g., vulnerability to cheating and hacking, unpredictable performance of player hardware, and difficulty in updating live game data for all players simultaneously, which top game developers solved by designing their multiplayer games to run on cloud-based servers they provided for players to use.

Running multiplayer game servers in the cloud has many benefits for players and developers. Cloud-based games have fewer connection failures due to firewall and proxy configurations, give players a consistent gameplay experience, run compute-intensive games on more robust hardware, allow developers to make game changes more quickly by updating servers under their control, and better protect their players and game economy from cheating by using servers to validate transactions.

Designing, building, deploying, and operating a game's multiplayer backend is difficult. Every game backend must be designed to handle rapidly fluctuating player traffic, including the potential for hundreds of thousands or millions of players to arrive on launch day. Designing and building reliable software to deploy and automatically scale servers is technically risky, expensive, and time consuming. It can take teams of specialized engineers thousands of hours to design and build a highly scalable, reliable, and secure system. Many developers also struggle to forecast the right amount of servers to use for their game. If they deploy too few servers, players have poor experiences as they face wait times or unreliable gameplay. If developers deploy too many servers, they spend money on compute resources they do not use.

Amazon GameLift makes this simple for session-based multiplayer games. We have done the hard work already, so you can use Amazon GameLift to deploy, operate, and scale your game servers without the time, cost, and risks associated with building the software yourself.

Q. What types of game genres are recommended for use with Amazon GameLift?

Amazon GameLift is designed to support session-based games with game loops that begin and end within a specified time period. Typically, these are multiplayer games in genres like first person shooters, MOBAs, fighting, racing, or sports. Amazon GameLift is not designed to support games with persistent worlds that never reset, such as MMOs or sandbox games. Amazon GameLift is also not designed to support asynchronous turn-based multiplayer designs that are often used for mobile or social games.

Q. Does Amazon GameLift work for latency-intolerant games, such as first-person shooters? Does Amazon GameLift add latency to my game?
Amazon GameLift is designed to work for latency-intolerant games and Amazon Lumberyard's multiplayer networking layer is optimized for latency-intolerant gameplay. Amazon GameLift introduces no additional latency during gameplay.

Prior to being redirected to a specific game server for gameplay, players initiate a play-slot reservation with your game on Amazon GameLift. This reservation and redirection process occurs once per play session and may take several seconds from the time the player initiates the request to join a game. Once Amazon GameLift connects a player to a game server for gameplay, all player-to-server communication is done directly between your game client and servers. Your player's latency during gameplay will depend on the player's Internet connection and distance to the game server. Amazon GameLift supports regions in North America, Europe, and Asia, so you can choose where to best deploy game servers for your players.

Q. How do I make suggestions or give feedback?
Please visit our GameDev Forums, or contact us.

Billing

Q. How much does Amazon GameLift cost?
You pay for the AWS compute, storage and bandwidth capacity you actually use, plus $1.50 per 1,000 Daily Active Users. Please see our pricing page for more information.

Q. When does billing of my Amazon GameLift games begin and end?
Billing begins when you allocate GameLift capacity in the Amazon GameLift console and Amazon GameLift launches your game's server binary for the first time on each instance. Billing concludes when you de-allocate capacity. Partial instance-hours consumed are billed as full hours. Amazon GameLift also offers a Free Tier so that you can get started for free.

Q. How does Amazon GameLift measure Daily Active Users (DAU)?
Amazon GameLift counts a uniquely identified user as active when he begins a play session on Amazon GameLift. DAU is measured from 00:00:00 GMT (UTC), and is aggregated across all of your game's fleets in all regions. To accurately count DAU, you need to provide a unique, non-personally identifiable player ID for each user. For more information about generating identifiers for your players, please refer to the Working with Player Sessions and Player IDssection of the Amazon GameLift Developer Guide.

Development

Q. How do I get started with Amazon GameLift?
First, download Amazon Lumberyard or the Amazon GameLift Server SDK. After you have created an Amazon Lumberyard game server build, you can upload the server build to Amazon GameLift via the AWS command line interface. Using the Amazon GameLift console, you can create a fleet to then configure your game client so your players can connect and play. For more details, please see the Amazon GameLift Developer Guide.

Q. Is there a sample game I can use to test Amazon GameLift?
Yes. The Amazon Lumberyard download includes a sample multiplayer project called MultiplayerProject for you to use to evaluate and test. You can select MultiplayerProject using the Lumberyard Project Configurator. The Amazon GameLift tutorials provide a step-by-step guide using this sample game.

Q. Can I use Amazon GameLift with any game engine?
Amazon GameLift currently only supports the Amazon Lumberyard engine.

Q. What platforms and languages are supported by GameLift?
The Amazon GameLift SDK for C++ supports game servers that run onWindows Server 2012 R2 or Amazon Linux.

Game clients and game services (such as matchmaking or player authentication) can use the AWS SDK to communicate with the GameLift service and connect players to games. The AWS SDK provides support for Amazon GameLift in C++, Java, .NET, Go, Python, Ruby, PHP, and JavaScript.

Q. Do I have to authenticate players in order to access my Amazon GameLift game servers?
No, you are not required to authenticate players. You can connect your game client to your servers on Amazon GameLift without authenticating, provided your game uses AWS credentials.

Any requests to Amazon GameLift to create, search for, or join game sessions require AWS credentials. To learn more about how to setup and configure AWS credentials, go to AWS Identity and Access Management.

Amazon Cognito can provide temporary, limited-privilege credentials for accessing AWS resources without player authentication. In addition, with Amazon Cognito you can create unique end user identifiers for accessing AWS cloud services by using public login providers such as Amazon, Facebook, Twitter, Google, and any OpenID Connect compatible provider, or by using your own user identity system.

Q. How do I debug my Amazon GameLift game?
To help debug your game, Amazon GameLift provides two tools. Amazon GameLift lets you

automatically collect and store up to seven days of server logs generated by your game. You can pull down logs for specific game sessions to either debug your server or better understand player behavior. Amazon GameLift also lets you modify open ports and protocols on your fleet so that you can connect remote debugging or profiling tools to your running servers. You can modify your fleet's port settings either using the AWS command line interface, or through the Amazon GameLift console.

Q. Are logs collected in real-time?
Fleet-level aggregated data is generally available in the Amazon GameLift console within ten minutes of collection and server-level aggregated data is generally available within five minutes of collection.

Q: How can I help players find game sessions to join?
Amazon GameLift provides a search capability that helps you filter and sort game sessions by characteristics meaningful to your players. You can filter and sort game sessions on attributes like game session age, current player count, maximum players count, open slots, or your own custom game properties.

Q. Can I get a history of Amazon GameLift API calls made on my account for security analysis and operational troubleshooting purposes?
Yes. To receive a history of Amazon GameLift API calls made on your account, you simply turn on CloudTrail in the AWS Management Console.

Q. What kind of support is available?
Every customer has access to documentation, tutorials, and forums. Additional support is available AWS Premium Support packages.

Instances and Fleets

Q. Which EC2 instance types does Amazon GameLift support?
Amazon GameLift supports many different EC2 On-Demand instance types running Windows Server 2012 R2 and Amazon Linux. Please see our pricing page for a full list of compatible instances. If you are interested in using EC2 Reserved or Spot Instances, please contact us.

Q. What server operating systems are currently supported?
Windows Server 2012 R2 and Amazon Linux.

Q: How many EC2 instances can I run in Amazon GameLift?
Amazon GameLift is limited by the number of instances available to your AWS account. If you need more Amazon EC2 instances for use with Amazon GameLift, complete the instance request form in the Amazon GameLift console.

Q: How many server processes can I run on an EC2 instance?

Multiple. The number of server processes depends on the performance requirements of your game servers and the instance type you choose for your fleet. When you set up a fleet, you will select an instance type and configure the fleet to concurrently run an optimum number of server processes. Running more processes on fewer instances can help you decrease costs. You can also configure your fleet to run multiple server builds or game configurations on each instance.

Q. How do I add or remove instances from my Amazon GameLift game?
You can increase or decrease the available instances using the Amazon GameLift page in the Amazon GameLift console.

Q. How quickly can I add or remove new instances?
Amazon GameLift utilizes Amazon EC2, which provides a truly elastic computing environment. Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can provision one, hundreds or even thousands of server instances simultaneously. You can control the quantity of servers from the fleet details page in the Amazon GameLift console. Amazon EC2 always strives to have enough On-Demand capacity available to meet your needs, but during periods of very high demand, it is possible that you might not be able to launch specific On-Demand instance types in specific Availability Zones for short periods of time.

With Amazon GameLift's autoscaling feature, you can minimize any lag in providing extra capacity to meet player demand-or conversely, quickly scale down to avoid paying for capacity you don't need. To enable autoscaling, you can define a set of rules that are based on real-time measures of game server capacity and player demand. For example, you could tell Amazon GameLift to scale down whenever the number of unused instances exceeds a certain number for a period of time or scale up whenever the number of available instances falls below a pre-set threshold. For more information, see the Autoscaling section of the GameLift Documentation.

Q. How do I select the right instance type for my game?
The right instance type depends on your game's server performance and the number of server processes you plan to run concurrently on each instance. The computational complexity of your game, optimization of your game and network code, and maximum number of players are the main drivers for the size of the instance that you will need. One of the advantages of Amazon GameLift is that you only pay for what you use, which makes it convenient and inexpensive to test the performance of your game on different instance families and types. By letting you select an instance type from a dropdown menu before you launch your fleet, Amazon GameLift makes it easy to test different EC2 instance types to see which instance best matches your particular game's performance.

Q. How can I find out more about the performance of EC2 instances?
For details, see the Amazon EC2 FAQ.

Q. What happens to my data when an EC2 instance terminates?
When you terminate an EC2 instance, any data that has been stored on the ephemeral (local to the EC2 instance) storage is lost. Prior to termination, Amazon GameLift stores any data written

to the logs path to S3. These logs are accessible up to seven days after the instance has been terminated. These logs are accessible from the Amazon GameLift page in the Amazon GameLift console.

Q. What is a fleet?
A fleet is a set of EC2 instances running one build containing your game servers. You can accommodate changes in player demand by increasing or decreasing the number of EC2 instances in your fleet. A fleet is configured to use a certain EC2 instance type, to deploy a build, and to run concurrently a specified number of processes for each game server in the build. For example, if you have a free and a premium server version, you might include both in your game build and then configure your fleet to run 2 free server processes and 4 premium server processes per EC2 instance.

Q. Is there a limit to the number of fleets I can create?
Yes, by default, you are limited to 20 fleets per region. However, if you need more fleets, please contact us to request an increase to your limit.

Q. How long does it take to create a fleet?
It typically takes less than an hour to create the resources necessary to run your game, measured from the time you upload your game binary to when it is fully deployed and accessible to your players. This time is dependent on the size of your binary and the number of instances you are deploying.

Q. Does Amazon GameLift allow me to update a live production fleet, or revert to a previous fleet if there is a problem?
Yes, Amazon GameLift makes updating production fleets simple with its alias feature. An alias enables you to direct traffic to fleets without having to change the client end-point descriptor. After creating a new production fleet, you can edit an alias to point from an older fleet to this newer fleet, routing all connecting players to the new fleet.

Similarly, if you discover an issue with a fleet (e.g., you find an issue in your game code), you can edit an alias to redirect traffic from the new fleet to an older fleet.

Q. Which regions is Amazon GameLift available in?
Amazon GameLift is currently running in the following regions: US East (N. Virginia), US West (Oregon), EU West (Ireland), and Asia Pacific (Tokyo). Please refer to the AWS Global Infrastructure Region Table for the current information on product and service availability by region.

Storage

Q. What is the build catalog?
The build catalog is a record of all of your server builds that have been uploaded to Amazon

GameLift. Builds in ready state are available for fleet creation at any time.

Q. Is there a limit to how many builds I can store in the build catalog?
The build catalog can store the maximum of 1,000 builds or 100GB of storage.

Q. How much storage does Amazon GameLift provide on each instance?
Amazon GameLift includes 50GB EBS General Purpose (SSD) Volume for each instance.

Operational Limits

Q. Is there a maximum number of allowed instances per fleet?
No, you are only limited by your AWS account limits.

Q. Is there a maximum number of players per game session supported?
Yes, the maximum number of players per game session is 200.

Q. Is there a maximum number of players per instance supported?
No, the maximum number of players per instance is primarily dictated by your game design and game code.

Q. Is there a maximum number of players per fleet supported?
No.

Other

Q. What is an AMI?
An Amazon Machine Image (AMI) is a supported and maintained image provided by AWS for use on Amazon EC2. Amazon GameLift uses Windows 2012 R2 and Amazon Linux to run your game server. An AMI is designed to provide a stable, secure, and high performance execution environment for applications running on Amazon EC2. It also includes packages that enable easy integration with AWS, including launch configuration tools and many popular AWS libraries and tools. AWS provides ongoing security and maintenance updates to all instances running the Amazon AMI.

Q. Can anyone access the Amazon GameLift console?
No. Only your authorized users with the necessary AWS credentials can access the Amazon GameLift console. You can use AWS Identity and Access Management (IAM) to securely share AWS credentials across a team. Please see Creating IAM Policies for Amazon GameLift for more details.

Q. Where can I find more information about security and running applications on AWS?
For more information about securing AWS resources, see the AWS Security Center.

Amazon Elastic MapReduce FAQ

General

Q: What is Amazon EMR?

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 utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).

Q: What can I do with Amazon EMR?

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, 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.

Amazon EMR is ideal for problems that necessitate the fast and efficient processing of large amounts of data. The web service interfaces allow you to build processing workflows, and programmatically monitor progress of running clusters. In addition, you can use the simple web interface of the AWS Management Console to launch your clusters and monitor processing-intensive computation on clusters of Amazon EC2 instances.

Q: Who can use Amazon EMR?

Anyone who requires simple access to powerful data analysis can use Amazon EMR. You don’t need any software development experience to experiment with several sample applications available in the Developer Guide and on the AWS Big Data Blog.

Q: What can I do with Amazon EMR that I could not do before?

Amazon EMR significantly reduces the complexity of the time-consuming set-up, management. and tuning of Hadoop clusters or the compute capacity upon which they sit. You can instantly spin up large Hadoop clusters which will start processing within minutes, not hours or days. When your cluster finishes its processing, unless you specify otherwise, it will be automatically terminated so you are not paying for resources you no longer need.

Using this service you can quickly perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial analysis, scientific simulation, and bioinformatics research.

As a software developer, you can also develop and run your own more sophisticated applications, allowing you to add functionality such as scheduling, workflows, monitoring, or other features.

Q: What is the data processing engine behind Amazon EMR?

Amazon EMR uses Apache Hadoop as its distributed data processing engine. Hadoop is an open source, Java software framework that supports data-intensive distributed applications running on large clusters of commodity hardware. Hadoop implements a programming model named “MapReduce,” where the data is divided into many small fragments of work, each of which may be executed on any node in the cluster. This framework has been widely used by developers, enterprises and startups and has proven to be a reliable software platform for processing up to petabytes of data on clusters of thousands of commodity machines.

Q: What is an Amazon EMR cluster?

Amazon EMR historically referred to an Amazon EMR cluster (and all processing steps assigned to it) as a "cluster". Every cluster or cluster has a unique identifier that starts with "j-".

Q: What is a cluster step?

A cluster step is a user-defined unit of processing, mapping roughly to one algorithm that manipulates the data. A step is a Hadoop MapReduce application implemented as a Java jar or a streaming program written in Java, Ruby, Perl, Python, PHP, R, or C++. For example, to count the frequency with which words appear in a document, and output them sorted by the count, the first step would be a MapReduce application which counts the occurrences of each word, and the second step would be a MapReduce application which sorts the output from the first step based on the counts.

Q: What are different cluster states?

STARTING – The cluster provisions, starts, and configures EC2 instances.
BOOTSTRAPPING – Bootstrap actions are being executed on the cluster.
RUNNING – A step for the cluster is currently being run.
WAITING – The cluster is currently active, but has no steps to run.
TERMINATING - The cluster is in the process of shutting down.
TERMINATED - The cluster was shut down without error.
TERMINATED_WITH_ERRORS - The cluster was shut down with errors.

Q: What are different step states?

PENDING – The step is waiting to be run.
RUNNING – The step is currently running.

COMPLETED – The step completed successfully.
CANCELLED – The step was cancelled before running because an earlier step failed or cluster was terminated before it could run.
FAILED – The step failed while running.

Q: What are some EMR best practices?

If you are running EMR in production you should specify an AMI version, Hive version, Pig version, etc. to make sure the version does not change unexpectedly (e.g. when EMR later adds support for a newer version). If your cluster is mission critical, only use Spot instances for task nodes because if the Spot price increases you may lose the instances. In development, use logging and enable debugging to spot and correct errors faster. If you are using GZIP, keep your file size to 1–2 GB because GZIP files cannot be split. Click here to download the white paper on Amazon EMR best practices.

Return to Top >>

Launching a Cluster

Q: How can I access Amazon EMR?

You can access Amazon EMR by using the AWS Management Console, Command Line Tools, SDKS, or the EMR API.

Q: How can I launch a cluster?

You can launch a cluster through the AWS Management Console by filling out a simple cluster request form. In the request form, you specify the name of your cluster, the location in Amazon S3 of your input data, your processing application, your desired data output location, and the number and type of Amazon EC2 instances you'd like to use. Optionally, you can specify a location to store your cluster log files and SSH Key to login to your cluster while it is running. Alternatively, you can launch a cluster using the RunJobFlow API or using the 'create' command in the Command Line Tools.

Q: How can I get started with Amazon EMR?

To sign up for Amazon EMR, click the "Sign Up Now" button on the Amazon EMR detail page http://aws.amazon.com/elasticmapreduce. You must be signed up for Amazon EC2 and Amazon S3 to access Amazon EMR; if you are not already signed up for these services, you will be prompted to do so during the Amazon EMR sign-up process. After signing up, please refer to the Amazon EMR documentation, which includes our Getting Started Guide – the best place to get going with the service.

Q: How can I terminate a cluster?

At any time, you can terminate a cluster via the AWS Management Console by selecting a cluster and clicking the “Terminate” button. Alternatively, you can use the TerminateJobFlows API. If you terminate a running cluster, any results that have not been persisted to Amazon S3 will be lost and all Amazon EC2 instances will be shut down.

Q: Does Amazon EMR support multiple simultaneous cluster?

Yes. At any time, you can create a new cluster, even if you’re already running one or more clusters.

Q: How many clusters can I run simultaneously?

You can start as many clusters as you like. You are limited to 20 instances across all your clusters. If you need more instances, complete the Amazon EC2 instance request form and your use case and instance increase will be considered. If your Amazon EC2 limit has been already raised, the new limit will be applied to your Amazon EMR clusters.

Return to Top >>

Developing

Q: Where can I find code samples?

Check out the sample code in theseArticles and Tutorials.

Q: How do I develop a data processing application?

You can develop a data processing job on your desktop, for example, using Eclipse or NetBeans plug-ins such as IBM MapReduce Tools for Eclipse (http://www.alphaworks.ibm.com/tech/mapreducetools). These tools make it easy to develop and debug MapReduce jobs and test them locally on your machine. Additionally, you can develop your cluster directly on Amazon EMR using one or more instances.

Q: What is the benefit of using the Command Line Tools or APIs vs. AWS Management Console?

The Command Line Tools or APIs provide the ability to programmatically launch and monitor progress of running clusters, to create additional custom functionality around clusters (such as sequences with multiple processing steps, scheduling, workflow, or monitoring), or to build value-added tools or applications for other Amazon EMR customers. In contrast, the AWS Management Console provides an easy-to-use graphical interface for launching and monitoring your clusters directly from a web browser.

Q: Can I add steps to a cluster that is already running?

Yes. Once the job is running, you can optionally add more steps to it via the AddJobFlowSteps API. The AddJobFlowSteps API will add new steps to the end of the current step sequence. You may want to use this API to implement conditional logic in your cluster or for debugging.

Q: Can I run a persistent cluster?

Yes. Amazon EMR clusters that are started with the –alive flag will continue until explicitly terminated. This allows customers to add steps to a cluster on demand. You may want to use this to debug your application without having to repeatedly wait for cluster startup. You may also use a persistent cluster to run a long-running data warehouse cluster. This can be combined with data warehouse and analytics packages that runs on top of Hadoop such as Hive and Pig.

Q: Can I be notified when my cluster is finished?

You can sign up for up Amazon SNS and have the cluster post to your SNS topic when it is finished. You can also view your cluster progress on the AWS Management Console or you can use the Command Line, SDK, or APIs get a status on the cluster.

Q: What programming languages does Amazon EMR support?

You can use Java to implement Hadoop custom jars. Alternatively, you may use other languages including Perl, Python, Ruby, C++, PHP, and R via Hadoop Streaming. Please refer to the Developer’s Guide for instructions on using Hadoop Streaming.

Q: What OS versions are supported with Amazon EMR?

At this time Amazon EMR supports Debian/Squeeze in 32 and 64 bit modes.

Q: Can I view the Hadoop UI while my cluster is running?

Yes. Please refer to the Hadoop UI section in the Developer’s Guide for instructions on how to access the Hadoop UI.

Q: Does Amazon EMR support third-party software packages?

Yes. The recommended way to install third-party software packages on your cluster is to use Bootstrap Actions. Alternatively you can package any third party libraries directly into your Mapper or Reducer executable. You can also upload statically compiled executables using the Hadoop distributed cache mechanism.

Q: Which Hadoop versions does Amazon EMR support?

For the latest versions supported by Amazon EMR, please reference the documentation.

Q: Can I use a data processing engine other than Hadoop?

Yes, some EMR customers use Spark and Shark (In-memory mapreduce and datawarehousing) as their processing engine. See this article for instructions on how to do this.

Q: Does Amazon contribute Hadoop improvements to the open source community?

Yes. Amazon EMR is active with the open source community and contributes many fixes back to the Hadoop source.

Q: Does Amazon EMR update the version of Hadoop it supports?

Amazon EMR periodically updates its supported version of Hadoop based on the Hadoop releases by the community. Amazon EMR may choose to skip some Hadoop releases.

Q: How quickly does Amazon EMR retire support for old Hadoop versions?

Amazon EMR service retires support for old Hadoop versions several months after deprecation. However, Amazon EMR APIs are backward compatible, so if you build tools on top of these APIs, they will work even when Amazon EMR updates the Hadoop version it's using.

Return to Top >>

Debugging

Q: How can I debug my cluster?

You first select the cluster you want to debug, then click on the "Debug" button to access the debug a cluster window in the AWS Management Console. This will enable you to track progress and identify issues in steps, jobs, tasks, or task attempts of your clusters. Alternatively you can SSH directly into the Amazon Elastic Compute Cloud (Amazon EC2) instances that are running your cluster and use your favorite command-line debugger to troubleshoot the cluster.

Q: What is the cluster debug tool?

The cluster debug tool is a part of the AWS Management Console where you can track progress and identify issues in steps, jobs, tasks, or task attempts of your clusters. To access the cluster debug tool, first select the cluster you want to debug and then click on the "Debug" button.

Q: How can I enable debugging of my cluster?

To enable debugging you need to set "Enable Debugging" flag when you create a cluster in the AWS Management Console. Alternatively, you can pass the --enable-debugging and --log-uri flags in the Command Line Client when creating a cluster.

Q: Where can I find instructions on how to use the debug a cluster window?

Please reference the AWS Management Console section of theDeveloper's Guide for instructions on how to access and use the debug a cluster window.

Q: What types of clusters can I debug with the debug a cluster window?

You can debug all types of clusters currently supported by Amazon EMR including custom jar, streaming, Hive, and Pig.

Q: Why do I have to sign-up for Amazon SimpleDB to use cluster debugging?

Amazon EMR stores state information about Hadoop jobs, tasks and task attempts under your account in Amazon SimpleDB. You can subscribe to Amazon SimpleDB here.

Q: Can I use the cluster debugging feature without Amazon SimpleDB subscription?

You will be able to browse cluster steps and step logs but will not be able to browse Hadoop jobs, tasks, or task attempts if you are not subscribed to Amazon SimpeDB.

Q: Can I delete historical cluster data from Amazon SimpleDB?

Yes. You can delete Amazon SimpleDB domains that Amazon EMR created on your behalf. Please reference the Amazon SimpleDB documentation for instructions.

Return to Top >>

Managing Data

Q: How do I get my data into Amazon S3?

You can use Amazon S3 APIs to upload data to Amazon S3. Alternatively, you can use many open source or commercial clients to easily upload data to Amazon S3.

Q: How do I get logs for completed clusters?

Hadoop system logs as well as user logs will be placed in the Amazon S3 bucket which you specify when creating a cluster.

Q: Do you compress logs?

No. At this time Amazon EMR does not compress logs as it moves them to Amazon S3.

Q: Can I load my data from the internet or somewhere other than Amazon S3?

Yes. Your Hadoop application can load the data from anywhere on the internet or from other AWS services. Note that if you load data from the internet, EC2 bandwidth charges will apply. Amazon EMR also provides Hive-based access to data in DynamoDB.

Return to Top >>

Billing

Q: Can Amazon EMR estimate how long it will take to process my input data?

No. As each cluster and input data is different, we cannot estimate your job duration.

Q: How much does Amazon EMR cost?

As with the rest of AWS, you pay only for what you use. There is no minimum fee and there are no up-front commitments or long-term contracts. Amazon EMR pricing is in addition to normal Amazon EC2 and Amazon S3 pricing.

For Amazon EMR pricing information, please visit EMR's pricing page.

Amazon EC2, Amazon S3 and Amazon SimpleDB charges are billed separately. Pricing for Amazon EMR is per instance-hour consumed for each instance type, from the time cluster began processing until it is terminated. Each partial instance-hour consumed will be billed as a full hour. For additional details on Amazon EC2 Instance Types, Amazon EC2 Spot Pricing, Amazon EC2 Reserved Instances Pricing, Amazon S3 Pricing, or Amazon SimpleDB Pricing, follow the links below:

Amazon EC2 Instance Types

Amazon EC2 Reserved Instances Pricing

Amazon EC2 Spot Instances Pricing

Amazon S3 Pricing

Amazon SimpleDB Pricing

Q: When does billing of my Amazon EMR cluster begin and end?

Billing commences when Amazon EMR starts running your cluster. You are only charged for the resources actually consumed. For example, let’s say you launched 100 Amazon EC2 Standard Small instances for an Amazon EMR cluster, where the Amazon EMR cost is an incremental $0.015 per hour. The Amazon EC2 instances will begin booting immediately, but they won’t necessarily all start at the same moment. Amazon EMR will track when each instance starts and will check it into the cluster so that it can accept processing tasks.

In the first 10 minutes after your launch request, Amazon EMR either starts your cluster (if all of your instances are available) or checks in as many instances as possible. Once the 10 minute mark has passed, Amazon EMR will start processing (and charging for) your cluster as soon as 90% of your requested instances are available. As the remaining 10% of your requested instances check in, Amazon EMR starts charging for those instances as well.

So, in the above example, if all 100 of your requested instances are available 10 minutes after you kick off a launch request, you’ll be charged $1.50 per hour (100 * $0.015) for as long as the cluster takes to complete. If only 90 of your requested instances were available at the 10 minute mark, you’d be charged $1.35 per hour (90 * $0.015) for as long as this was the number of instances running your cluster. When the remaining 10 instances checked in, you'd be charged $1.50 per hour (100 * $0.015) for as long as the balance of the cluster takes to complete.

Each cluster will run until one of the following occurs: you terminate the cluster with the TerminateJobFlows API call (or an equivalent tool), the cluster shuts itself down, or the cluster is terminated due to software or hardware failure. Partial instance hours consumed are billed as full hours.

Q: Where can I track my Amazon EMR, Amazon EC2 and Amazon S3 usage?

You can track your usage in theBilling & Cost Management Console.

Q: How do you calculate the Normalized Instance Hours displayed on the console ?

On the AWS Management Console, every cluster has a Normalized Instance Hours column that displays the approximate number of compute hours the cluster has used. Normalized Instance Hours are hours of compute time based on the standard of 1 hour of m1.small usage = 1 hour normalized compute time. The following table outlines the normalization factor used to calculate normalized instance hours for the various instance sizes:

For example, if you run a 10-node r3.8xlarge cluster for an hour, the total number of Normalized Instance Hours displayed on the console will be 640 (10 (number of nodes) x 64 (normalization factor) x 1 (number of hours tthat the cluster ran) = 640).

This is an approximate number and should not be used for billing purposes. Please refer to the Billing & Cost Management Console for billable Amazon EMR usage. Note that we recently changed the normalization factor to accurately reflect the weights of the instances, and the normalization factor does not affect your monthly bill.

Q: Does Amazon EMR support Amazon EC2 On-Demand, Spot, and Reserved Instances?

Yes. Amazon EMR seamlessly supports On-Demand, Spot, and Reserved Instances.Click here to learn more about Amazon EC2 Reserved Instances. Click here to learn more about Amazon EC2 Spot Instances.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Return to Top >>

Security

Q: How do I prevent other people from viewing my data during cluster execution?

Amazon EMR starts your instances in two Amazon EC2 security groups, one for the master and another for the slaves. The master security group has a port open for communication with the service. It also has the SSH port open to allow you to SSH into the instances, using the key specified at startup. The slaves start in a separate security group, which only allows interaction with the master instance. By default both security groups are set up to not allow access from external sources including Amazon EC2 instances belonging to other customers. Since these are security groups within your account, you can reconfigure them using the standard EC2 tools or dashboard. Click here to learn more about EC2 security groups.

Q: How secure is my data?

Amazon S3 provides authentication mechanisms to ensure that stored data is secured against unauthorized access. Unless the customer who is uploading the data specifies otherwise, only that customer can access the data. Amazon EMR customers can also choose to send data to Amazon S3 using the HTTPS protocol for secure transmission. In addition, Amazon EMR always uses HTTPS to send data between Amazon S3 and Amazon EC2. For added security, customers may encrypt the input data before they upload it to Amazon S3 (using any common data compression tool); they then need to add a decryption step to the beginning of their cluster when Amazon EMR fetches the data from Amazon S3.

Q: Can I get a history of all EMR API calls made on my account for security or compliance auditing?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console.

Return to Top >>

Regions & Availability Zones

Q: How does Amazon EMR make use of Availability Zones?

Amazon EMR launches all nodes for a given cluster in the same Amazon EC2 Availability Zone. Running a cluster in the same zone improves performance of the jobs flows because it provides a higher data access rate. By default, Amazon EMR chooses the Availability Zone with the most available resources in which to run your cluster. However, you can specify another Availability

Zone if required.

Q: In what Regions is this Amazon EMR available?

For a list of the supported Amazon EMR AWS regions, please visit theAWS Region Table for all AWS global infrastructure.

Q: Which Region should I select to run my clusters?

When creating a cluster, typically you should select the Region where your data is located.

Q: Can I use EU data in a cluster running in the US region and vice versa?

Yes you can. If you transfer data from one region to the other you will be charged bandwidth charges. For bandwidth pricing information, please visit the pricing section on the EC2 detail page.

Q: What is different about the AWS GovCloud (US) region?

The AWS GovCloud (US) region is designed for US government agencies and customers. It adheres to US ITAR requirements. In GovCloud, EMR does not support spot instances or the enable-debugging feature. The EMR Management Console is not yet available in GovCloud.

Return to Top >>

Managing your Cluster

Q: How does Amazon EMR use Amazon EC2 and Amazon S3?

Customers upload their input data and a data processing application into Amazon S3. Amazon EMR then launches a number of Amazon EC2 instances as specified by the customer. The service begins the cluster execution while pulling the input data from Amazon S3 using S3N protocol into the launched Amazon EC2 instances. Once the cluster is finished, Amazon EMR transfers the output data to Amazon S3, where customers can then retrieve it or use as input in another cluster.

Q: How is a computation done in Amazon EMR?

Amazon EMR uses the Hadoop data processing engine to conduct computations implemented in the MapReduce programming model. The customer implements their algorithm in terms of map() and reduce() functions. The service starts a customer-specified number of Amazon EC2 instances, comprised of one master and multiple slaves. Amazon EMR runs Hadoop software on these instances. The master node divides input data into blocks, and distributes the processing of the blocks to the slave node. Each slave node then runs the map function on the data it has been allocated, generating intermediate data. The intermediate data is then sorted and partitioned and sent to processes which apply the reducer function to it. These processes also run on the slave nodes. Finally, the output from the reducer tasks is collected in files. A single “cluster” may involve a sequence of such MapReduce steps.

Q: How reliable is Amazon EMR?

Amazon EMR manages an Amazon EC2 cluster of compute instances using Amazon’s highly available, proven network infrastructure and datacenters. Amazon EMR uses industry proven, fault-tolerant Hadoop software as its data processing engine. Hadoop splits the data into multiple subsets and assigns each subset to more than one Amazon EC2 instance. So, if an Amazon EC2 instance fails to process one subset of data, the results of another Amazon EC2 instance can be used.

Q: How quickly will my cluster be up and running and processing my input data?

Amazon EMR starts resource provisioning of Amazon EC2 On-Demand instances almost immediately. If the instances are not available, Amazon EMR will keep trying to provision the resources for your cluster until they are provisioned or you cancel your request. The instance provisioning is done on a best-efforts basis and depends on the number of instances requested, time when the cluster is created, and total number of requests in the system. After resources have been provisioned, it typically takes fewer than 15 minutes to start processing.

In order to guarantee capacity for your clusters at the time you need it, you may pay a one-time fee for Amazon EC2 Reserved Instances to reserve instance capacity in the cloud at a discounted hourly rate. Like On-Demand Instances, customers pay usage charges only for the time when their instances are running. In this way, Reserved Instances enable businesses with known instance requirements to maintain the elasticity and flexibility of On-Demand Instances, while also reducing their predictable usage costs even further.

Q: Which Amazon EC2 instance types does Amazon EMR support?

Amazon EMR supports 12 EC2 instance types including Standard, High CPU, High Memory, Cluster Compute, High I/O, and High Storage. Standard Instances have memory to CPU ratios suitable for most general-purpose applications. High CPU instances have proportionally more CPU resources than memory (RAM) and are well suited for compute-intensive applications. High Memory instances offer large memory sizes for high throughput applications. Cluster Compute instances have proportionally high CPU with increased network performance and are well suited for High Performance Compute (HPC) applications and other demanding network-bound applications. High Storage instances offer 48 TB of storage across 24 disks and are ideal for applications that require sequential access to very large data sets such as data warehousing and log processing. See the EMR pricing page for details on available instance types and pricing per region.

Q: How do I select the right Amazon EC2 instance type?

When choosing instance types, you should consider the characteristics of your application with regards to resource utilization and select the optimal instance family. One of the advantages of Amazon EMR with Amazon EC2 is that you pay only for what you use, which makes it convenient and inexpensive to test the performance of your clusters on different instance types and quantity. One effective way to determine the most appropriate instance type is to launch several small clusters and benchmark your clusters.

Q: How do I select the right number of instances for my cluster?

The number of instances to use in your cluster is application-dependent and should be based on both the amount of resources required to store and process your data and the acceptable amount of time for your job to complete. As a general guideline, we recommend that you limit 60% of your disk space to storing the data you will be processing, leaving the rest for intermediate output. Hence, given 3x replication on HDFS, if you were looking to process 5 TB on m1.xlarge instances, which have 1,690 GB of disk space, we recommend your cluster contains at least (5 TB * 3) / (1,690 GB * .6) = 15 m1.xlarge core nodes. You may want to increase this number if your job generates a high amount of intermediate data or has significant I/O requirements. You may also want to include additional task nodes to improve processing performance. See Amazon EC2 Instance Types for details on local instance storage for each instance type configuration.

Q: How long will it take to run my cluster?

The time to run your cluster will depend on several factors including the type of your cluster, the amount of input data, and the number and type of Amazon EC2 instances you choose for your cluster.

Q: If the master node in a cluster goes down, can Amazon EMR recover it?

No. If the master node goes down, your cluster will be terminated and you'll have to rerun your job. Amazon EMR currently does not support automatic failover of the master nodes or master node state recovery. In case of master node failure, the AWS Management console displays "The master node was terminated" message which is an indicator for you to start a new cluster. Customers can instrument check pointing in their clusters to save intermediate data (data created in the middle of a cluster that has not yet been reduced) on Amazon S3. This will allow resuming the cluster from the last check point in case of failure.

Q: If a slave node goes down in a cluster, can Amazon EMR recover from it?

Yes. Amazon EMR is fault tolerant for slave failures and continues job execution if a slave node goes down. In the current version, Amazon EMR does not automatically provision another node to take over failed slaves.

Q: Can I SSH onto my cluster nodes?

Yes. You can SSH onto your cluster nodes and execute Hadoop commands directly from there.

If you need to SSH into a slave node, you have to first SSH to the master node, and then SSH into the slave node.

Q: Can I use Microsoft Windows instances with Amazon EMR?

At this time, Amazon EMR supports Debian/Lenny in 32 and 64 bit modes. We are always listening to customer feedback and will add more capabilities over time to help our customers solve their data crunching business problems.

Q: What is Amazon EMR Bootstrap Actions?

Bootstrap Actions is a feature in Amazon EMR that provides users a way to run custom set-up prior to the execution of their cluster. Bootstrap Actions can be used to install software or configure instances before running your cluster. You can read more about bootstrap actions in EMR's Developer Guide.

Q: How can I use Bootstrap Actions?

You can write a Bootstrap Action script in any language already installed on the cluster instance including Bash, Perl, Python, Ruby, C++, or Java. There are several pre-defined Bootstrap Actions available. Once the script is written, you need to upload it to Amazon S3 and reference its location when you start a cluster. Please refer to the “Developer’s Guide”: http://docs.amazonwebservices.com/ElasticMapReduce/latest/DeveloperGuide/ for details on how to use Bootstrap Actions.

Q: How do I configure Hadoop settings for my cluster?

The EMR default Hadoop configuration is appropriate for most workloads. However, based on your cluster’s specific memory and processing requirements, it may be appropriate to tune these settings. For example, if your cluster tasks are memory-intensive, you may choose to use fewer tasks per core and reduce your job tracker heap size. For this situation, a pre-defined Bootstrap Action is available to configure your cluster on startup. See the Configure Memory Intensive Bootstrap Action in the Developer’s Guide for configuration details and usage instructions. An additional predefined bootstrap action is available that allows you to customize your cluster settings to any value of your choice. See the Configure Hadoop Bootstrap Action in the Developer’s Guide for usage instructions.

Q: Can I modify the number of slave nodes in a running cluster?

Yes. Slave nodes can be of two types: (1) core nodes, which both host persistent data using Hadoop Distributed File System (HDFS) and run Hadoop tasks and (2) task nodes, which only run Hadoop tasks. While a cluster is running you may increase the number of core nodes and you may either increase or decrease the number of task nodes. This can be done through the API, Java SDK, or though the command line client. Please refer to the Resizing Running clusters section in the Developer’s Guide for details on how to modify the size of your running cluster.

Q: When would I want to use core nodes versus task nodes?

As core nodes host persistent data in HDFS and cannot be removed, core nodes should be reserved for the capacity that is required until your cluster completes. As task nodes can be added or removed and do not contain HDFS, they are ideal for capacity that is only needed on a temporary basis.

Q: Why would I want to modify the number of slave nodes in my running cluster?

There are several scenarios where you may want to modify the number of slave nodes in a running cluster. If your cluster is running slower than expected, or timing requirements change, you can increase the number of core nodes to increase cluster performance. If different phases of your cluster have different capacity needs, you can start with a small number of core nodes and increase or decrease the number of task nodes to meet your cluster's varying capacity requirements.

Q: Can I automatically modify the number of slave nodes between cluster steps?

Yes. You may include a predefined step in your workflow that automatically resizes a cluster between steps that are known to have different capacity needs. As all steps are guaranteed to run sequentially, this allows you to set the number of slave nodes that will execute a given cluster step.

Q: How can I allow other IAM users to access my cluster?

To create a new cluster that is visible to all IAM users within the EMR CLI: Add the --visible-to-all-users flag when you create the cluster. For example: elastic-mapreduce --create --visible-to-all-users. Within the Management Console, simply select "Visible to all IAM Users" on the Advanced Options pane of the Create cluster Wizard.

To make an existing cluster visible to all IAM users you must use the EMR CLI. Use --set-visible-to-all-users and specify the cluster identifier. For example: elastic-mapreduce --set-visible-to-all-users true --jobflow j-xxxxxxx. This can only be done by the creator of the cluster.

To learn more, see the Configuring User Permissions section of the EMR Developer Guide.

Return to Top >>

Tagging your Cluster

Q: What Amazon EMR resources can I tag?

You can add tags to an active Amazon EMR cluster. An Amazon EMR cluster consists of Amazon EC2 instances, and a tag added to an Amazon EMR cluster will be propagated to each active Amazon EC2 instance in that cluster. You cannot add, edit, or remove tags from terminated clusters or terminated Amazon EC2 instances which were part of an active cluster.

Q: Does Amazon EMR tagging support resource-based permissions with IAM Users?

No, Amazon EMR does not support resource-based permissions by tag. However, it is important to note that propagated tags to Amazon EC2 instances behave as normal Amazon EC2 tags. Therefore, an IAM Policy for Amazon EC2 will act on tags propagated from Amazon EMR if they match conditions in that policy.

Q: How many tags can I add to a resource?

You can add up to ten tags on an Amazon EMR cluster.

Q: Do my Amazon EMR tags on a cluster show up on each Amazon EC2 instance in that cluster? If I remove a tag on my Amazon EMR cluster, will that tag automatically be removed from each associated EC2 instance?

Yes, Amazon EMR propagates the tags added to a cluster to that cluster's underlying EC2 instances. If you add a tag to an Amazon EMR cluster, it will also appear on the related Amazon EC2 instances. Likewise, if you remove a tag from an Amazon EMR cluster, it will also be removed from its associated Amazon EC2 instances. However, if you are using IAM policies for Amazon EC2 and plan to use Amazon EMR's tagging functionality, you should make sure that permission to use the Amazon EC2 tagging APIs CreateTags and DeleteTags is granted.

Q: How do I get my tags to show up in my billing statement to segment costs?

Select the tags you would like to use in your AWS billing reporthere. Then, to see the cost of your combined resources, you can organize your billing information based on resources that have the same tag key values.

Q: How do I tell which Amazon EC2 instances are part of an Amazon EMR cluster?

An Amazon EC2 instance associated with an Amazon EMR cluster will have two system tags:

- aws:elasticmapreduce:instance-group-role=CORE
  - Key = instance-group role ; Value = [CORE or TASK]
- aws:elasticmapreduce:job-flow-id=j-12345678
  - Key = job-flow-id ; Value = [JobFlowID]

Q: Can I edit tags directly on the Amazon EC2 instances?

Yes, you can add or remove tags directly on Amazon EC2 instances that are part of an Amazon EMR cluster. However, we do not recommend doing this, because Amazon EMR’s tagging system will not sync the changes you make to an associated Amazon EC2 instance directly. We recommend that tags for Amazon EMR clusters be added and removed from the Amazon EMR console, CLI, or API to ensure that the cluster and its associated Amazon EC2 instances have the correct tags.

Return to Top >>

Using EBS Volumes

Q: What can I do now that I could not do before?

Most EC2 instances have fixed storage capacity attached to an instance, known as an "instance store". You can now add EBS volumes to the instances in your Amazon EMR cluster, allowing you to customize the storage on an instance. The feature also allows you to run Amazon EMR clusters on EBS-Only instance families such as the M4 and C4.

Q: What are the benefits of adding EBS volumes to an instance running on Amazon EMR?

You will benefit by adding EBS volumes to an instance in the following scenarios:

1. Your processing requirements are such that you need a large amount of HDFS (or local) storage that what is available today on an instance. With support for EBS volumes, you will be able to customize the storage capacity on an instance relative to the compute capacity that the instance provides. Optimizing the storage on an instance will allow you to save costs.

2. You are running on an older generation instance family (such as the M1 and M2 family) and want to move to latest generation instance family but are constrained by the storage available per node on the next generation instance types. Now you can use any of the new generation instance type and add EBS volumes to optimize the storage. Internal benchmarks indicate that you can save cost and improve performance by moving from an older generation instance family (M1 or M2) to a new generation one (M4, C4 & R3). The Amazon EMR team recommends that you run your application to arrive at the right conclusion.

3. You want to use or migrate to the next-generation EBS-Only M4 and C4 family.

Q: Can I persist my data on an EBS volume after a cluster is terminated?

Currently, Amazon EMR will delete volumes once the cluster is terminated. If you want to persist data outside the lifecycle of a cluster, consider using Amazon S3 as your data store.

Q: What kind of EBS volumes can I attach to an instance?

Amazon EMR allows you to use different EBS Volume Types: General Purpose SSD (GP2), Magnetic and Provisioned IOPS (SSD).

Q: What happens to the EBS volumes once I terminate my cluster?

Amazon EMR will delete the volumes once the EMR cluster is terminated.

Q: Can I use an EBS with instances that already have an instance store?

Yes, You can add EBS volumes to instances that have an instance store.

Q: Can I attach and EBS volume to a running cluster?

No, currently you can only add EBS volumes when launching a cluster.

Q: Can I snapshot volumes from a cluster?

The EBS API allows you to Snapshot a cluster. However, Amazon EMR currently does not allow you to restore from a snapshot.

Q: Can I use encrypted EBS volumes?

No, encrypted volumes are not supported in the current release.

Q: What happens when I remove an attached volume from a running cluster?

Removing an attached volume from a running cluster will be treated as a node failure. Amazon EMR will replace the node and the EBS volume with each of the same.

Return to Top >>

Using Hive

Q: What is Apache Hive?

Hive is an open source datawarehouse and analytics package that runs on top of Hadoop. Hive is operated by a SQL-based language called Hive QL that allows users to structure, summarize, and query data sources stored in Amazon S3. Hive QL goes beyond standard SQL, adding first-class support for map/reduce functions and complex extensible user-defined data types like Json and Thrift. This capability allows processing of complex and even unstructured data sources such as text documents and log files. Hive allows user extensions via user-defined functions written in Java and deployed via storage in Amazon S3.

Q: What can I do with Hive running on Amazon EMR?

Using Hive with Amazon EMR, you can implement sophisticated data-processing applications with a familiar SQL-like language and easy to use tools available with Amazon EMR. With Amazon EMR, you can turn your Hive applications into a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence tasks.

Q: How is Hive different than traditional RDBMS systems?

Traditional RDBMS systems provide transaction semantics and ACID properties. They also allow tables to be indexed and cached so that small amounts of data can be retrieved very quickly. They provide for fast update of small amounts of data and for enforcement of referential integrity constraints. Typically they run on a single large machine and do not provide support for executing map and reduce functions on the table, nor do they typically support acting over complex user defined data types.

In contrast, Hive executes SQL-like queries using MapReduce. Consequently, it is optimized for doing full table scans while running on a cluster of machines and is therefore able to process very large amounts of data. Hive provides partitioned tables, which allow it to scan a partition of a table rather than the whole table if that is appropriate for the query it is executing.

Traditional RDMS systems are best for when transactional semantics and referential integrity are required and frequent small updates are performed. Hive is best for offline reporting, transformation, and analysis of large data sets; for example, performing click stream analysis of a large website or collection of websites.

One of the common practices is to export data from RDBMS systems into Amazon S3 where offline analysis can be performed using Amazon EMR clusters running Hive.

Q: How can I get started with Hive running on Amazon EMR?

The best place to start is to review our written or video tutorial located here http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2862

Q: Are there new features in Hive specific to Amazon EMR?

Yes. There are four new features which make Hive even more powerful when used with Amazon EMR, including:

a/ The ability to load table partitions automatically from Amazon S3. Previously, to import a partitioned table you needed a separate alter table statement for each individual partition in the table. Amazon EMR a now includes a new statement type for the Hive language: “alter table recover partitions.” This statement allows you to easily import tables concurrently into many clusters without having to maintain a shared meta-data store. Use this functionality to read from tables into which external processes are depositing data, for example log files.

b/ The ability to specify an off-instance metadata store. By default, the metadata store where Hive stores its schema information is located on the master node and ceases to exist when the cluster terminates. This feature allows you to override the location of the metadata store to use, for example a MySQL instance that you already have running in EC2.

c/ Writing data directly to Amazon S3. When writing data to tables in Amazon S3, the version of Hive installed in Amazon EMR writes directly to Amazon S3 without the use of temporary files. This produces a significant performance improvement but it means that HDFS and S3 from a Hive perspective behave differently. You cannot read and write within the same statement to the same table if that table is located in Amazon S3. If you want to update a table located in S3, then create a temporary table in the cluster’s local HDFS filesystem, write the results to that table, and then copy them to Amazon S3.

d/ Accessing resources located in Amazon S3. The version of Hive installed in Amazon EMR allows you to reference resources such as scripts for custom map and reduce operations or additional libraries located in Amazon S3 directly from within your Hive script (e.g., add jar s3://elasticmapreduce/samples/hive-ads/libs/jsonserde.jar).

Q: What types of Hive clusters are supported?

There are two types of clusters supported with Hive: interactive and batch. In an interactive mode a customer can start a cluster and run Hive scripts interactively directly on the master node. Typically, this mode is used to do ad hoc data analyses and for application development. In batch mode, the Hive script is stored in Amazon S3 and is referenced at the start of the cluster. Typically, batch mode is used for repeatable runs such as report generation.

Q: How can I launch a Hive cluster?

Both batch and interactive clusters can be started from AWS Management Console, EMR command line client, or APIs. Please refer to the Using Hive section in the Developer's Guide for more details on launching a Hive cluster.

Q: When should I use Hive vs. PIG?

Hive and PIG both provide high level data-processing languages with support for complex data types for operating on large datasets. The Hive language is a variant of SQL and so is more accessible to people already familiar with SQL and relational databases. Hive has support for partitioned tables which allow Amazon EMR clusters to pull down only the table partition relevant to the query being executed rather than doing a full table scan. Both PIG and Hive have query plan optimization. PIG is able to optimize across an entire scripts while Hive queries are optimized at the statement level.

Ultimately the choice of whether to use Hive or PIG will depend on the exact requirements of the application domain and the preferences of the implementers and those writing queries.

Q: What version of Hive does Amazon EMR support?

Amazon EMR supports multiple versions of Hive, including version 0.11.0.

Q: Can I write to a table from two clusters concurrently

No. Hive does not support concurrently writing to tables. You should avoid concurrently writing to the same table or reading from a table while you are writing to it. Hive has non-deterministic behavior when reading and writing at the same time or writing and writing at the same time.

Q: Can I share data between clusters?

Yes. You can read data in Amazon S3 within a Hive script by having 'create external table' statements at the top of your script. You need a create table statement for each external resource that you access.

Q: Should I run one large cluster, and share it amongst many users or many smaller clusters?

Amazon EMR provides a unique capability for you to use both methods. On the one hand one large cluster may be more efficient for processing regular batch workloads. On the other hand, if you require ad-hoc querying or workloads that vary with time, you may choose to create several separate cluster tuned to the specific task sharing data sources stored in Amazon S3.

Q: Can I access a script or jar resource which is on my local file system?

No. You must upload the script or jar to Amazon S3 or to the cluster's master node before it can be referenced. For uploading to Amazon S3 you can use tools including s3cmd, jets3t or S3Organizer.

Q: Can I run a persistent cluster executing multiple Hive queries?

Yes. You run a cluster in a manual termination mode so it will not terminate between Hive steps. To reduce the risk of data loss we recommend periodically persisting all of your important data in Amazon S3. It is good practice to regularly transfer your work to a new cluster to test you process for recovering from master node failure.

Q: Can multiple users execute Hive steps on the same source data?

Yes. Hive scripts executed by multiple users on separate clusters may contain create external table statements to concurrently import source data residing in Amazon S3.

Q: Can multiple users run queries on the same cluster?

Yes. In the batch mode, steps are serialized. Multiple users can add Hive steps to the same cluster, however, the steps will be executed serially. In interactive mode, several users can be logged on to the same cluster and execute Hive statements concurrently.

Q: Can data be shared between multiple AWS users?

Yes. Data can be shared using standard Amazon S3 sharing mechanism described here http://docs.amazonwebservices.com/AmazonS3/latest/index.html?S3_ACLs.html

Q: Does Hive support access from JDBC?

Yes. Hive provides JDBC drive, which can be used to programmatically execute Hive statements. To start a JDBC service in your cluster you need to pass an optional parameter in the Amazon EMR command line client. You also need to establish an SSH tunnel because the security group does not permit external connections.

Q: What is your procedure for updating packages on EMR AMIs?

We run a select set of packages from Debian/stable including security patches. We will upgrade a package whenever it gets upgraded in Debian/stable. The “r-recommended” package on our image is up to date with Debian/stable (http://packages.debian.org/search?keywords=r-recommended).

Q: Can I update my own packages on EMR clusters?

Yes. You can use Bootstrap Actions to install updates to packages on your clusters.

Q: Can I process DynamoDB data using Hive?

Yes. Simply define an external Hive table based on your DynamoDB table. You can then use Hive to analyze the data stored in DynamoDB and either load the results back into DynamoDB or archive them in Amazon S3. For more information please visit our Developer Guide.

Return to Top >>

Using Impala

Q: What is Impala?

Impala is an open source tool in the Hadoop ecosystem for interactive, ad hoc querying using SQL syntax. Instead of using MapReduce, it leverages a massively parallel processing (MPP) engine similar to that found in traditional relational database management systems (RDBMS). With this architecture, you can query your data in HDFS or HBase tables very quickly, and leverage Hadoop’s ability to process diverse data types and provide schema at runtime. This lends Impala to interactive, low-latency analytics. In addition, Impala uses the Hive metastore to hold information about the input data, including the partition names and data types. Also, Impala on Amazon EMR requires AMIs running Hadoop 2.x or greater. Click here to learn more about Impala.

Q: What can I do with Impala running on Amazon EMR?

Similar to using Hive with Amazon EMR, leveraging Impala with Amazon EMR can implement sophisticated data-processing applications with SQL syntax. However, Impala is built to perform faster in certain use cases (see below). With Amazon EMR, you can use Impala as a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence. Here are three use cases:

- Use Impala instead of Hive on long-running clusters to perform ad hoc queries. Impala reduces interactive queries to seconds, making it an excellent tool for fast investigation. You could run Impala on the same cluster as your batch MapReduce workflows, use Impala on a long-running analytics cluster with Hive and Pig, or create a cluster specifically tuned for Impala queries.

- Use Impala instead of Hive for batch ETL jobs on transient Amazon EMR clusters. Impala is faster than Hive for many queries, which provides better performance for these workloads. Like Hive, Impala uses SQL, so queries can easily be modified from Hive to Impala.
- Use Impala in conjunction with a third party business intelligence tool. Connect a client ODBC or JDBC driver with your cluster to use Impala as an engine for powerful visualization tools and dashboards.

Both batch and interactive Impala clusters can be created in Amazon EMR. For instance, you can have a long-running Amazon EMR cluster running Impala for ad hoc, interactive querying or use transient Impala clusters for quick ETL workflows.

Q: How is Impala different than traditional RDBMSs?

Traditional relational database systems provide transaction semantics and database atomicity, consistency, isolation, and durability (ACID) properties. They also allow tables to be indexed and cached so that small amounts of data can be retrieved very quickly, provide for fast updates of small amounts of data, and for enforcement of referential integrity constraints. Typically, they run on a single large machine and do not provide support for acting over complex user defined data types. Impala uses a similar distributed query system to that found in RDBMSs, but queries data stored in HDFS and uses the Hive metastore to hold information about the input data. As with Hive, the schema for a query is provided at runtime, allowing for easier schema changes. Also, Impala can query a variety of complex data types and execute user defined functions. However, because Impala processes data in-memory, it is important to understand the hardware limitations of your cluster and optimize your queries for the best performance.

Q: How is Impala different than Hive?

Impala executes SQL queries using a massively parallel processing (MPP) engine, while Hive executes SQL queries using MapReduce. Impala avoids Hive's overhead from creating MapReduce jobs, giving it faster query times than Hive. However, Impala uses significant memory resources and the cluster's available memory places a constraint on how much memory any query can consume. Hive is not limited in the same way, and can successfully process larger data sets with the same hardware. Generally, you should use Impala for fast, interactive queries, while Hive is better for ETL workloads on large datasets. Impala is built for speed and is great for ad hoc investigation, but requires a significant amount of memory to execute expensive queries or process very large datasets. Because of these limitations, Hive is recommended for workloads where speed is not as crucial as completion. Click here to view some performance benchmarks between Impala and Hive.

Q: Can I use Hadoop 1?

No, Impala requires Hadoop 2, and will not run on a cluster with an AMI running Hadoop 1.x.

Q: What instance types should I use for my Impala cluster?

For the best experience with Impala, we recommend using memory-optimized instances for your cluster. However, we have shown that there are performance gains over Hive when using standard instance types as well. We suggest reading our Performance Testing and Query Optimization section in the Amazon EMR Developer's Guide to better estimate the memory resources your cluster will need with regards to your dataset and query types. The compression type, partitions, and the actual query (number of joins, result size, etc.) all play a role in the memory required. You can use the EXPLAIN statement to estimate the memory and other resources needed for an Impala query.

Q: What happens if I run out of memory on a query?

If you run out of memory, queries fail and the Impala daemon installed on the affected node shuts down. Amazon EMR then restarts the daemon on that node so that Impala will be ready to run another query. Your data in HDFS on the node remains available, because only the daemon running on the node shuts down, rather than the entire node itself. For ad hoc analysis with Impala, the query time can often be measured in seconds; therefore, if a query fails, you can discover the problem quickly and be able to submit a new query in quick succession.

Q: Does Impala support user defined functions?

Yes, Impala supports user defined functions (UDFs). You can write Impala specific UDFs in Java or C++. Also, you can modify UDFs or user-defined aggregate functions created for Hive for use with Impala. For information about Hive UDFs, click here.

Q: Where is the data stored for Impala to query?

Impala queries data in HDFS or in HBase tables. If you are storing your data in Amazon S3, we recommend you follow our Amazon EMR Best Practices Whitepaper for methods to transfer your data to HDFS.

Q: Can I run Impala and MapReduce at the same time on a cluster?

Yes, you can set up a multitenant cluster with Impala and MapReduce. However, you should be sure to allot resources (memory, disk, and CPU) to each application using YARN on Hadoop 2.x. The resources allocated should be dependent on the needs for the jobs you plan to run on each application.

Q: Does Impala support ODBC and JDBC drivers?

While you can use ODBC drivers, Impala is also a great engine for third-party tools connected through JDBC. You can download and install the Impala client JDBC driver from http://elasticmapreduce.s3.amazonaws.com/libs/impala/1.2.1/impala-jdbc-1.2.1.zip. From the client computer where you have your business intelligence tool installed, connect the JDBC driver to the master node of an Impala cluster using SSH or a VPN on port 21050. For more information, see Open an SSH Tunnel to the Master Node.

Return to Top >>

Using Pig

Q: What is Apache Pig?

Pig is an open source analytics package that runs on top of Hadoop. Pig is operated by a SQL-like language called Pig Latin, which allows users to structure, summarize, and query data sources stored in Amazon S3. As well as SQL-like operations, Pig Latin also adds first-class support for map/reduce functions and complex extensible user defined data types. This capability allows processing of complex and even unstructured data sources such as text documents and log files. Pig allows user extensions via user-defined functions written in Java and deployed via storage in Amazon S3.

Q: What can I do with Pig running on Amazon EMR?

Using Pig with Amazon EMR, you can implement sophisticated data-processing applications with a familiar SQL-like language and easy to use tools available with Amazon EMR. With Amazon EMR, you can turn your Pig applications into a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence tasks.

Q: How can I get started with Pig running on Amazon EMR?

The best place to start is to review our written or video tutorial located here http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2735&categoryID=269

Q: Are there new features in Pig specific to Amazon EMR?

Yes. There are three new features which make Pig even more powerful when used with Amazon EMR, including:

a/ Accessing multiple filesystems. By default a Pig job can only access one remote file system, be it an HDFS store or S3 bucket, for input, output and temporary data. EMR has extended Pig so that any job can access as many file systems as it wishes. An advantage of this is that temporary intra-job data is always stored on the local HDFS, leading to improved perfomance.

b/ Loading resources from S3. EMR has extended Pig so that custom JARs and scripts can come from the S3 file system, for example "REGISTER s3:///my-bucket/piggybank.jar"

c/ Additional Piggybank function for String and DateTime processing. These are documented here http://developer.amazonwebservices.com/connect/entry.jspa?externalID=2730.

Q: What types of Pig clusters are supported?

There are two types of clusters supported with Pig: interactive and batch. In an interactive mode a customer can start a cluster and run Pig scripts interactively directly on the master node. Typically, this mode is used to do ad hoc data analyses and for application development. In batch mode, the Pig script is stored in Amazon S3 and is referenced at the start of the cluster. Typically, batch mode is used for repeatable runs such as report generation.

Q: How can I launch a Pig cluster?

Both batch and interactive clusters can be started from AWS Management Console, EMR command line client, or APIs.

Q: What version of Pig does Amazon EMR support?

Amazon EMR supports multiple versions of Pig, including 0.11.1.

Q: Can I write to a S3 bucket from two clusters concurrently

Yes, you are able to write to the same bucket from two concurrent clusters.

Q: Can I share input data in S3 between clusters?

Yes, you are able to read the same data in S3 from two concurrent clusters.

Q: Can data be shared between multiple AWS users?

Yes. Data can be shared using standard Amazon S3 sharing mechanism described here http://docs.amazonwebservices.com/AmazonS3/latest/index.html?S3_ACLs.html

Q: Should I run one large cluster, and share it amongst many users or many smaller clusters?

Amazon EMR provides a unique capability for you to use both methods. On the one hand one large cluster may be more efficient for processing regular batch workloads. On the other hand, if you require ad-hoc querying or workloads that vary with time, you may choose to create several separate cluster tuned to the specific task sharing data sources stored in Amazon S3.

Q: Can I access a script or jar resource which is on my local file system?

No. You must upload the script or jar to Amazon S3 or to the cluster’s master node before it can be referenced. For uploading to Amazon S3 you can use tools including s3cmd, jets3t or S3Organizer.

Q: Can I run a persistent cluster executing multiple Pig queries?

Yes. You run a cluster in a manual termination mode so it will not terminate between Pig steps. To reduce the risk of data loss we recommend periodically persisting all important data in Amazon S3. It is good practice to regularly transfer your work to a new cluster to test you process for recovering from master node failure.

Q: Does Pig support access from JDBC?

No. Pig does not support access through JDBC.

Return to Top >>

Using HBase

Q: What is Apache HBase?

HBase is an open source, non-relational, distributed database modeled after Google's BigTable. It was developed as part of Apache Software Foundation's Hadoop project and runs on top of Hadoop Distributed File System(HDFS) to provide BigTable-like capabilities for Hadoop. HBase provides you a fault-tolerant, efficient way of storing large quantities of sparse data using column-based compression and storage. In addition, HBase provides fast lookup of data because data is stored in-memory instead of on disk. HBase is optimized for sequential write operations, and it is highly efficient for batch inserts, updates, and deletes. HBase works seamlessly with Hadoop, sharing its file system and serving as a direct input and output to Hadoop jobs. HBase also integrates with Apache Hive, enabling SQL-like queries over HBase tables, joins with Hive-based tables, and support for Java Database Connectivity (JDBC).

Q: Are there new features in HBase specific to Amazon EMR?

With Amazon EMR you can back up HBase to Amazon S3 (full or incremental, manual or automated) and you can restore from a previously created backup. Learn more about HBase and EMR.

Q: Which versions of HBase are supported on Amazon EMR?

Amazon EMR supports HBase 0.94.7 and HBase 0.92.0. To use HBase 0.94.7 you must specify AMI version 3.0.0. If you are using the CLI you must use version 2013-10-07 or later.

Return to Top >>

Kinesis Connector

Q: What does EMR Connector to Kinesis enable?

The connector enables EMR to directly read and query data from Kinesis streams. You can now perform batch processing of Kinesis streams using existing Hadoop ecosystem tools such as Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.

Q: What does the EMR connector to Kinesis enable that I couldn’t have done before?

Reading and processing data from a Kinesis stream would require you to write, deploy and maintain independent stream processing applications. These take time and effort. However, with this connector, you can start reading and analyzing a Kinesis stream by writing a simple Hive or Pig script. This means you can analyze Kinesis streams using SQL! Of course, other Hadoop ecosystem tools could be used as well. You don't need to developed or maintain a new set of processing applications.

Q: Who will find this functionality useful?

The following types of users will find this integration useful:

- Hadoop users who are interested in utilizing the extensive set of Hadoop ecosystem tools to analyze Kinesis streams.
- Kinesis users who are looking for an easy way to get up and running with stream processing and ETL of Kinesis data.
- Business analysts and IT professionals who would like to perform ad-hoc analysis of data in Kinesis streams using familiar tools like SQL (via Hive) or scripting languages like Pig.

Q: What are some use cases for this integration?

The following are representative use cases are enabled by this integration:

- Streaming Log Analysis: You can analyze streaming web logs to generate a list of top 10 error type every few minutes by region, browser, and access domains.
- Complex Data Processing Workflows: You can join Kinesis stream with data stored in S3, Dynamo DB tables, and HDFS. You can write queries that join clickstream data from Kinesis with advertising campaign information stored in a DynamoDB table to identify the most effective categories of ads that are displayed on particular websites.
- Ad-hoc Queries: You can periodically load data from Kinesis into HDFS and make it available as a local Impala table for fast, interactive, analytic queries.

Q: What EMR AMI version do I need to be able to use the connector?

You need to use EMR's AMI version 3.0.4 and later.

Q: Is this connector a stand-alone tool?

No, it is a built in component of the Amazon distribution of Hadoop and is present on EMR AMI versions 3.0.4 and later. Customer simply needs to spin up a cluster with AMI version 3.0.4 or later to start using this feature.

Q: What data format is required to allow EMR to read from a Kinesis stream?

The EMR Kinesis integration is not data format specific. You can read data in any format. Individual Kinesis records are presented to Hadoop as standard records that can be read using any Hadoop MapReduce framework. Individual frameworks like Hive, Pig and Cascading have built in components that help with serialization and deserialization, making it easy for developers to query data from many formats without having to implement custom code. For example, in Hive users can read data from JSON files, XML files and SEQ files by specifying the appropriate Hive SerDe when they define a table. Pig has a similar component called Loadfunc/Evalfunc and Cascading has a similar component called a Tap. Hadoop users can leverage the extensive ecosystem of Hadoop adapters without having to write format specific code. You can also implement custom deserialization formats to read domain specific data in any of these tools.

Q: How do I analyze a Kinesis stream using Hive in EMR?

Create a table that references a Kinesis stream. You can then analyze the table like any other table in Hive. Please see our tutorials for page more details.

Q: Using Hive, how do I create queries that combine Kinesis stream data with other data source?

First create a table that references a Kinesis stream. Once a Hive table has been created, you can join it with tables mapping to other data sources such as Amazon S3, Amazon Dynamo DB, and HDFS. This effectively results in joining data from Kinesis stream to other data sources.

Q: Is this integration only available for Hive?

No, you can use Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.

Q: How do I setup scheduled jobs to run on a Kinesis stream?

The EMR Kinesis input connector provides features that help you configure and manage scheduled periodic jobs in traditional scheduling engines such as Cron. For example, you can develop a Hive script that runs every N minutes. In the configuration parameters for a job, you can specify a **Logical Name** for the job. The Logical Name is a label that will inform the EMR Kinesis input connector that individual instances of the job are members of the same periodic schedule. The Logical Name allows the process to take advantage of iterations, which are explained next.

Since MapReduce is a batch processing framework, to analyze a Kinesis stream using EMR, the continuous stream is divided in to batches. Each batch is called an **Iteration**. Each Iteration is assigned a number, starting with 0. Each Iteration’s boundaries are defined by a start sequence number and end sequence number. Iterations are then processed sequentially by EMR.

In the event of an attempt’s failure, the EMR Kinesis input connector will re-try the iteration within the Logical Name from the known start sequence number of the iteration. This functionality ensures that successive attempts on the same iteration will have precisely the same input records from the Kinesis stream as the previous attempts. This guarantees idempotent (consistent) processing of a Kinesis stream.

You can specify Logical Names and Iterations as runtime parameters in your respective Hadoop tools. For example, in the tutorial section “Running queries with checkpoints”, the code sample shows a scheduled Hive query that designates a Logical Name for the query and increments the iteration with each successive run of the job.

Additionally, a sample cron scheduling script is provided in thetutorials.

Q: Where is the metadata for Logical Names and Iterations stored?

The metadata that allows the EMR Kinesis input connector to work in scheduled periodic workflows is stored in Amazon DynamoDB. You must provision an Amazon Dynamo DB table and specify it as an input parameter to the Hadoop Job. It is important that you configure appropriate IOPS for the table to enable this integration. Please refer to the getting started tutorial for more information on setting up your Amazon Dynamo DB table.

Q: What happens when an iteration processing fails?

Iterations identifiers are user-provided values that map to specific boundary (start and end sequence numbers) in a Kinesis stream. Data corresponding to these boundaries is loaded in the Map phase of the MapReduce job. This phase is managed by the framework and will be automatically re-run (three times by default) in case of job failure. If all the retries fail, you would still have options to retry the processing starting from last successful data boundary or past data boundaries. This behavior is controlled by providing kinesis.checkpoint.iteration.no parameter during processing. Please refer to the getting started tutorial for more information on how this value is configured for different tools in the Hadoop ecosystem.

Q: Can I run multiple queries on the same iteration?

Yes, you can specify a previously run iteration by setting the kinesis.checkpoint.iteration.no parameter in successive processing. The implementation ensures that successive runs on the same iteration will have precisely the same input records from the Kinesis stream as the previous runs.

Q: What happens if records in an Iteration expire from the Kinesis stream?

In the event that the beginning sequence number and/or end sequence number of an iteration belong to records that have expired from the Kinesis steam, the Hadoop job will fail. You would need to use a different Logical Name to process data from the beginning of the Kinesis stream.

Q: Can I push data from EMR into Kinesis stream?

No. The EMR Kinesis connector currently does not support writing data back into a Kinesis stream.

Q: Does the EMR Hadoop input connector for Kinesis enable continuous stream processing?

The Hadoop MapReduce framework is a batch processing system. As such, it does not support continuous queries. However there is an emerging set of Hadoop ecosystem frameworks like Twitter Storm and Spark Streaming that enable to developers build applications for continuous stream processing. A Storm connector for Kinesis is available at on GitHub here and you can find a tutorial explaining how to setup Spark Streaming on EMR and run continuous queries here.

Additionally, developers can utilize the Kinesis client library to develop real-time stream processing applications. You can find more information on developing custom Kinesis applications in the Kinesis documentation here.

Q: Can I specify access credential to read a Kinesis stream that is managed in another AWS account?

Yes. You can read streams from another AWS account by specifying the appropriate access credentials of the account that owns the Kinesis stream. By default, the Kinesis connector utilizes the user-supplied access credentials that are specified when the cluster is created. You can override these credentials to access streams from other AWS Accounts by setting the kinesis.accessKey and kinesis.secretKey parameters. The following examples show how to set the kinesis.accessKey and kinesis.secretKey parameters in Hive and Pig.

Code sample for Hive:

```
...
STORED BY
'com.amazon.emr.kinesis.hive.KinesisStorageHandler'
TBLPROPERTIES(
"kinesis.accessKey"="AwsAccessKey",
"kinesis.secretKey"="AwsSecretKey",
);
```

Code sample for Pig:

```
…
raw_logs = LOAD 'AccessLogStream' USING com.amazon.emr.kinesis.pig.Kin
esisStreamLoader('kinesis.accessKey=AwsAccessKey', 'kinesis.secretKey=AwsSecretKey'
) AS (line:chararray);
```

Q: Can I run multiple parallel queries on a single Kinesis Stream? Is there a performance impact?

Yes, a customer can run multiple parallel queries on the same stream by using separate logical names for each query. However, reading from a shard within a Kinesis stream is subjected to a rate limit of of 2MB/sec. Thus, if there are N parallel queries running on the same stream, each one would get roughly (2/N) MB/sec egress rate per shard on the stream. This may slow down the processing and in some cases fail the queries as well.

Q: Can I join and analyze multiple Kinesis streams in EMR?

Yes, for example in Hive, you can create two tables mapping to two different Kinesis streams and create joins between the tables.

Q: Does the EMR Kinesis connector handle Kinesis scaling events, such as merge and split events?

Yes. The implementation handles split and merge events. The Kinesis connector ties individual Kinesis shards (the logical unit of scale within a Kinesis stream) to Hadoop MapReduce map tasks. Each unique shard that exists within a stream in the logical period of an Iteration will result in exactly one map task. In the event of a shard split or merge event, Kinesis will provision new unique shard Ids. As a result, the MapReduce framework will provision more map tasks to read from Kinesis. All of this is transparent to the user.

Q: What happens if there is there are periods of "silence" in my stream?

The implementation allows you to configure a parameter called kinesis.nodata.timeout. For example, consider a scenario where kinesis.nodata.timeout is set to 2 minutes and you want to run a Hive query every 10 minutes. Additionally, consider some data has been written to the stream since the last iteration (10 minutes ago). However, currently no new records are arriving, i.e. there is a silence in the stream. In this case, when the current iteration of the query launches, the Kinesis connector would find that no new records are arriving. The connector will keep polling the stream for 2 minutes and if no records arrive for that interval then it will stop and process only those records that were already read in the current batch of stream. However, if new records start arriving before kinesis.nodata.timeout interval is up, then the connector will wait for an additional interval corresponding to a parameter called kinesis.iteration.timeout. Please look at the tutorials to see how to define these parameters.

Q: How do I debug a query that continues to fail in each iteration?

In the event of a processing failure, you can utilize the same tools they currently do when debugging Hadoop Jobs. Including the Amazon EMR web console, which helps identify and access error logs. More details on debugging an EMR job can be found here.

Q: What happens if I specify a DynamoDB table that I don't have access to?

The job would fail and the exception would show up in error logs for the job.

Q: What happens if job doesn't fail but checkpointing to DynamoDB fails?

The job would fail and the exception would show up in error logs for the job.

Q: How do I maximize the read throughput from Kinesis stream to EMR?

Throughput from Kinesis stream increases with instance size used and record size in the Kinesis stream. We recommend that you use m1.xlarge and above for both master and core nodes for this feature.

Return to Top >>

AWS Data Pipeline FAQ

General

Q: What is AWS Data Pipeline?

AWS Data Pipeline is a web service that makes it easy to schedule regular data movement and data processing activities in the AWS cloud. AWS Data Pipeline integrates with on-premise and cloud-based storage systems to allow developers to use their data when they need it, where they want it, and in the required format. AWS Data Pipeline allows you to quickly define a dependent chain of data sources, destinations, and predefined or custom data processing activities called a pipeline. Based on a schedule you define, your pipeline regularly performs processing activities such as distributed data copy, SQL transforms, MapReduce applications, or custom scripts against destinations such as Amazon S3, Amazon RDS, or Amazon DynamoDB. By executing the scheduling, retry, and failure logic for these workflows as a highly scalable and fully managed service, Data Pipeline ensures that your pipelines are robust and highly available.

Q: What can I do with AWS Data Pipeline?

Using AWS Data Pipeline, you can quickly and easily provision pipelines that remove the development and maintenance effort required to manage your daily data operations, letting you focus on generating insights from that data. Simply specify the data sources, schedule, and processing activities required for your data pipeline. AWS Data Pipeline handles running and monitoring your processing activities on a highly reliable, fault-tolerant infrastructure. Additionally, to further ease your development process, AWS Data Pipeline provides built-in activities for common actions such as copying data between Amazon Amazon S3 and Amazon RDS, or running a query against Amazon S3 log data.

Q: How is AWS Data Pipeline different from Amazon Simple Workflow Service?

While both services provide execution tracking, retry and exception-handling capabilities, and the ability to run arbitrary actions, AWS Data Pipeline is specifically designed to facilitate the specific steps that are common across a majority of data-driven workflows – inparticular, executing activities after their input data meets specific readiness criteria, easily copying data between different data stores, and scheduling chained transforms. This highly specific focus means that its workflow definitions can be created rapidly and with no code or programming knowledge.

Q: What is a pipeline?

A pipeline is the AWS Data Pipeline resource that contains the definition of the dependent chain of data sources, destinations, and predefined or custom data processing activities required to execute your business logic.

Q: What is a data node?

A data node is a representation of your business data. For example, a data node can reference a specific Amazon S3 path. AWS Data Pipeline supports an expression language that makes it easy to reference data which is generated on a regular basis. For example, you could specify that your Amazon S3 data format is s3://example-bucket/my-logs/logdata-#{scheduledStartTime('YYYY-MM-dd-HH')}.tgz.

Q: What is an activity?

An activity is an action that AWS Data Pipeline initiates on your behalf as part of a pipeline. Example activities are EMR or Hive jobs, copies, SQL queries, or command-line scripts.

Q: What is a precondition?

A precondition is a readiness check that can be optionally associated with a data source or activity. If a data source has a precondition check, then that check must complete successfully before any activities consuming the data source are launched. If an activity has a precondition, then the precondition check must complete successfully before the activity is run. This can be useful if you are running an activity that is expensive to compute, and should not run until specific criteria are met.

Q: What is a schedule?

Schedules define when your pipeline activities run and the frequency with which the service expects your data to be available. All schedules must have a start date and a frequency; for example, every day starting Jan 1, 2013, at 3pm. Schedules can optionally have an end date, after which time the AWS Data Pipeline service does not execute any activities. When you associate a schedule with an activity, the activity runs on it. When you associate a schedule with a data source, you are telling the AWS Data Pipeline service that you expect the data to be updated on that schedule. For example, if you define an Amazon S3 data source with an hourly schedule, the service expects that the data source contains new files every hour.

Back to top »

Functionality

Q: Does Data Pipeline supply any standard Activities?

Yes, AWS Data Pipeline provides built-in support for the following activities:

- CopyActivity: This activity can copy data between Amazon S3 and JDBC data sources, or run a SQL query and copy its output into Amazon S3.
- HiveActivity: This activity allows you to execute Hive queries easily.
- EMRActivity: This activity allows you to run arbitrary Amazon EMR jobs.
- ShellCommandActivity: This activity allows you to run arbitrary Linux shell commands or programs.

Q: Does AWS Data Pipeline supply any standard preconditions?

Yes, AWS Data Pipeline provides built-in support for the following preconditions:

- DynamoDBDataExists: This precondition checks for the existence of data inside a DynamoDB table.
- DynamoDBTableExists: This precondition checks for the existence of a DynamoDB table.
- S3KeyExists: This precondition checks for the existence of a specific AmazonS3 path.
- S3PrefixExists: This precondition checks for at least one file existing within a specific path.
- ShellCommandPrecondition: This precondition runs an arbitrary script on your resources and checks that the script succeeds.

Q: Can I supply my own custom activities?

Yes, you can use the ShellCommandActivity to run arbitrary Activity logic.

Q: Can I supply my own custom preconditions?

Yes, you can use the ShellCommandPrecondition to run arbitrary precondition logic.

Q: Can you define multiple schedules for different activities in the same pipeline?

Yes, simply define multiple schedule objects in your pipeline definition file and associate the desired schedule to the correct activity via its schedule field. This allows you to define a pipeline in which, for example, log files are stored in Amazon S3 each hour to drive generation of an aggregate report one time per day.

Q: What happens if an activity fails?

An activity fails if all of its activity attempts return with a failed state. By default, an activity retries three times before entering a hard failure state. You can increase the number of automatic retries to 10; however, the system does not allow indefinite retries. After an activity exhausts its attempts, it triggers any configured onFailure alarm and will not try to run again unless you manually issue a rerun command via the CLI, API, or console button.

Q: How do I add alarms to an activity?

You can define Amazon SNS alarms to trigger on activity success, failure, or delay. Create an alarm object and reference it in the onFail,onSuccess, or onLate slots of the activity object.

Q: Can I manually rerun activities that have failed?

Yes. You can rerun a set of completed or failed activities by resetting their state to SCHEDULED. This can be done by using the Rerun button in the UI or modifying their state in the command line or API. This will immediately schedule a of re-check all activity dependencies, followed by the execution of additional activity attempts. Upon subsequent failures, the Activity will perform the original number of retry attempts.

Q: On what resources are activities run?

AWS Data Pipeline activities are run on compute resources that you own. There are two types of compute resources: AWS Data Pipeline–managed and self-managed. AWS Data Pipeline–managed resources are Amazon EMR clusters or Amazon EC2 instances that the AWS Data Pipeline service launches only when they're needed. Resources that you manage are longer running and can be any resource capable of running the AWS Data Pipeline Java-based Task Runner (on-premise hardware, a customer-managed Amazon EC2 instance, etc.).

Q: Will AWS Data Pipeline provision and terminate AWS Data Pipeline-managed compute resources for me?

Yes, compute resources will be provisioned when the first activity for a scheduled time that uses those resources is ready to run and those instances will be terminated when the final activity that uses the resources has completed successfully or failed.

Q: Can multiple compute resources be used on the same pipeline?

Yes, simply define multiple cluster objects in your definition file and associate the cluster to use for each activity via its runsOn field. This allows pipelines to combine AWS and on-premise resources, or to use a mix of instance types for their activities – for example, you may want to use a t1.micro to execute a quick script cheaply, but later on the pipeline may have an Amazon EMR job that requires the power of a cluster of larger instances.

Q: Can I execute activities on on-premise resources, or AWS resources that I manage?

Yes. To enable running activities using on-premise resources, AWS Data Pipeline supplies a Task Runner package that can be installed on your on-premise hosts. This package continuously polls the AWS Data Pipeline service for work to perform. When it’s time to run a particular activity on your on-premise resources, for example, executing a DB stored procedure or a database dump, AWS Data Pipeline will issue the appropriate command to the Task Runner. In order to ensure that your pipeline activities are highly available, you can optionally assign multiple Task Runners to poll for a given job. This way, if one Task Runner becomes unavailable, the others will simply pick up its work.

Q: How do I install a Task Runner on my on-premise hosts?

You can install the Task Runner package on your on-premise hosts using the following steps:

1. Download the AWS Task Runner package.
2. Create a configuration file that includes your AWS credentials.
3. Start the Task Runner agent via the following command:
   java -jar TaskRunner-1.0.jar --config ~/credentials.json --workerGroup=[myWorkerGroup]
4. When defining activities, set the activity to run on [myWorkerGroup] in order to dispatch them to the previously installed hosts.

Back to top »

Getting Started

Q: How can I get started with AWS Data Pipeline?

To get started with AWS Data Pipeline, simply visit theAWS Management Console and go to the AWS Data Pipeline tab. From there, you can create a pipeline using a simple graphical editor.

Q: What can I do with AWS Data Pipeline?

With AWS Data Pipeline, you can schedule and manage periodic data-processing jobs. You can use this to replace simple systems which are current managed by brittle, cron-based solutions, or you can use it to build complex, multi-stage data processing jobs.

Q: Are there Sample Pipelines that I can use to try out AWS Data Pipeline?

Yes, there are sample pipelines in ourdocumentation. Additionally, the console has several pipeline templates that you can use to get started.

Back to top »

Limits

Q: How many pipelines can I create in AWS Data Pipeline?

By default, your account can have 100 pipelines.

Q: Are there limits on what I can put inside a single pipeline?

By default, each pipeline you create can have 100 objects.

Q: Can my limits be changed?

Yes. If you would like to increase your limits, simplycontact us.

Back to top »

Billing

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Back to top »

Amazon Elasticsearch Service FAQ

General

Q: What is Amazon Elasticsearch Service?

Amazon Elasticsearch Service is a managed service that makes it easy to deploy, operate, and scale Elasticsearch clusters in the AWS Cloud.

Q: Which Elasticsearch version does Amazon Elasticsearch Service support?

Amazon Elasticsearch Service currently supports Elasticsearch versions 2.3 and 1.5.

Q: What is an Amazon Elasticsearch domain?

Amazon Elasticsearch domains are Elasticsearch clusters created using the Amazon Elasticsearch Service console, API, or CLI. Each domain is an Elasticsearch cluster in the cloud with the compute and storage resources you specify. You can create and delete domains, define/refine infrastructure attributes of your domains, and control access and security. You can run one or more Amazon Elasticsearch domains.

Q: What does Amazon Elasticsearch Service manage on my behalf?

Amazon Elasticsearch Service manages the work involved in setting up a domain, from provisioning infrastructure capacity you request to installing the Elasticsearch software. Once your domain is running, Amazon Elasticsearch Service automates common administrative tasks, such as performing backups, monitoring instances and patching software that powers your Amazon Elasticsearch instance. Amazon Elasticsearch Service is integrated with Amazon CloudWatch to produce metrics that provide information about the state of the domains. Amazon Elasticsearch Service also offers options to modify your domain instance and storage settings to simplify the task of managing your domain based on your application needs.

Q: Does Amazon Elasticsearch Service support the open source Elasticsearch engine APIs?

Amazon Elasticsearch Service will support the following Elasticsearch APIs, so code, applications, and popular tools that you're already using with your current Elasticsearch environments will work seamlessly: /_alias, /_aliases, /_all, /_analyze, /_bulk, /_cat, /_cluster/health, /_cluster/settings, /_cluster/stats, /_count, /_flush, /_mapping, /_mget, /_msearch, /_nodes, /_plugin/kibana, /_plugin/kibana3, /_percolate, /_refresh, /_search, /_snapshot, /_stats, /status, /_template.

Return to Top >>

Setup and Configuration

Q: Can I create and modify my Amazon Elasticsearch domain through the Amazon Elasticsearch Service console?

Yes. You can create a new Amazon Elasticsearch domain with the Domain Creation Wizard in the console. While creating a new domain you can specify the number of instances, instance types, and EBS volumes you want allocated to your domain. You can also modify or delete existing Amazon Elasticsearch domains using the console.

Q: Can I use CloudFormation Templates to provision Amazon ES domains?

Yes. AWS CloudFormation supports Amazon ES. For more information, see theCloudFormation Template Reference documentation.

Q: Does Amazon Elasticsearch Service support configuring dedicated master nodes for each domain?

Yes. You can configure the dedicated master setting for an Amazon Elasticsearch domain. When choosing a dedicated master setup, you will be able to also set up the instance type and instance count for the dedicated master configuration.

Q: Can I create multiple Elasticsearch indices within a single Amazon Elasticsearch domain?

Yes. You can create multiple Elasticsearch indices within the same Amazon Elasticsearch domain. Elasticsearch will manage storing of each index and any associated replica among the instances allocated to the domain.

Q: Does Amazon Elasticsearch Service support integration with Logstash?

Yes. Amazon Elasticsearch Service supports integration with Logstash. You can set up your Amazon Elasticsearch domain as the backend store for all logs coming through your Logstash implementation. You can set up access control on your Amazon Elasticsearch domain to either use request signing to authenticate calls from your Logstash implementation, or use resource based IAM policies to include IP addresses of instances running your Logstash implementation.

Q: Does Amazon Elasticsearch Service support integration with Kibana?

Yes. Amazon Elasticsearch Service includes built in support for Kibana that is deployed with your Amazon Elasticsearch domain. Currently Amazon Elasticsearch Service supports Kibana versions 3 and 4. Kibana 3 is supported only in Elasticsearch 1.5.

Q: Can I create custom reports with the Kibana installation included with Amazon Elasticsearch Service?

Yes. Kibana supports creating and saving custom reports through the user interface. For more information on using Kibana, refer to Kibana documentation from Elastic.co.

Q: What plugins are available by default with Amazon Elasticsearch Service?

Amazon Elasticsearch Service comes prepackaged with the following list of plugins available from the Elasticsearch community:
Kibana 3 (supported in Elasticsearch 1.5 only)
Kibana 4
jetty
cloud-aws
kuromoji
icu

Q: Which instance types are available with Amazon Elasticsearch Service?

Amazon Elasticsearch Service will support the following instance types:
T2.micro
T2.small
T2.medium
M3.medium
M3.large
M3.xlarge
M3.2xlarge
R3.large

R3.xlarge
R3.2xlarge
R3.4xlarge
R3.8xlarge
I2.xlarge
I2.2xlarge

Q: What storage options are available with Amazon Elasticsearch Service?

Customers will be able to choose between local on-instance storage available with the provisioned Amazon Elasticsearch instances or EBS volumes to store their Elasticsearch indices. During domain creation, if a customer selects the EBS storage option, the Domain Creation Wizard will prompt the customer to specify the type and size of EBS volume to be allocated to the domain. Customers will be able to modify the EBS settings after domain creation as well to increase or decrease the size and modify the volume type as needed.

Q: What types of EBS volumes does Amazon Elasticsearch Service support?

Customers will be able to choose between Magnetic, General Purpose, and Provisioned IOPS EBS volumes.

Q: Is there a limit on the amount of EBS storage that can be allocated to an Amazon Elasticsearch domain?

Yes. Amazon Elasticsearch Service supports 1 EBS volume (max size of 512GB) per instance associated with a cluster. With a maximum of 20 data nodes allowed per Amazon Elasticsearch cluster, customers can allocate about 10 TB of storage to a single Amazon Elasticsearch domain.

Return to Top >>

Administration

Q: How do I control access to Amazon Elasticsearch Service?

Customers will be able to set up IAM policies to control access to their Amazon Elasticsearch domains and sub resources like indices within the domains. IAM policies can also be set up to control access to the control plane API for operations like creating and scaling clusters and data plane API for operations like uploading documents and executing Elasticsearch requests.

Q: Can programs running on servers in my own data center access my Amazon Elasticsearch domains?

Yes. You can set up IAM policies to allow programs running on servers outside of AWS to access your Amazon Elasticsearch domains. You can use either IP based access policies or resource based access policies with signed requests to access the Amazon Elasticsearch domains. Click here for more information about signed requests.

Q: How can I migrate data from my existing Elasticsearch cluster to my new Amazon Elasticsearch domain?

To migrate data from an existing Elasticsearch cluster you should create a snapshot of an existing Elasticsearch cluster, and store the snapshot in your Amazon S3 bucket. Then you can create a new Amazon Elasticsearch domain and load data from the snapshot into the newly created Amazon Elasticsearch domain using the Elasticsearch restore API.

Q: How can I scale an Amazon Elasticsearch domain?

Amazon Elasticsearch Service allows you to control the scaling of your Amazon Elasticsearch domains using the console, API, and CLI. You can scale your Amazon Elasticsearch domain by adding, removing, or modifying instances or storage volumes depending on your application needs. Amazon Elasticsearch Service is integrated with Amazon CloudWatch to provide metrics about the state of your Amazon Elasticsearch domains to enable you to make appropriate scaling decisions for your domains.

Q: Does scaling my Amazon Elasticsearch domain require downtime?

No. Scaling your Amazon Elasticsearch domain by adding or modifying instances, and storage volumes is an online operation that does not require any downtime.

Q: What options does Amazon Elasticsearch Service provide for node failures?

Amazon Elasticsearch Service automatically detects node failures and replaces the node. The service will acquire new instances, and will then redirect Elasticsearch requests and document updates to the new instances. In the event that the node cannot be replaced, customers will be able to use any snapshots they have of their cluster to restart the domain with preloaded data.

Q: Does Amazon Elasticsearch Service support cross-zone replication?

Yes. Customers can enable Zone Awareness for their Amazon Elasticsearch domains either at domain creation time or by modifying a live domain. When Zone Awareness is enabled, Amazon Elasticsearch Service will distribute the instances supporting the domain across two different Availability Zones. Then, if replication is enabled in the Elasticsearch engine, Elasticsearch will allocate replicas of the domain across these different instances enabling cross-zone replication.

Q: Does Amazon Elasticsearch Service expose any performance metrics through Amazon CloudWatch?

Yes. Amazon Elasticsearch Service exposes several performance metrics through Amazon CloudWatch including number of nodes, cluster health, searchable documents, EBS metrics (if applicable), CPU, memory and disk utilization for data and master nodes. Please refer to the service documentation for a full listing of available CloudWatch metrics.

Q: I wish to perform security analysis or operational troubleshooting of my Amazon Elasticsearch Service deployment. Can I get a history of all the Amazon Elasticsearch Service API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by AWS CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about AWS CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console home page.

Q: What is a snapshot?

A snapshot is a copy of your Amazon Elasticsearch domain at a moment in time.

Q: Why would I need snapshots?

Creating snapshots can be useful in case of data loss caused by node failure, as well as the unlikely event of a hardware failure. You can use snapshots to recover your Amazon Elasticsearch domain with preloaded data or to create a new Amazon Elasticsearch domain with preloaded data. Another common reason to use backups is for archiving purposes. Snapshots are stored in Amazon S3.

Q: Does Amazon Elasticsearch Service provide automated snapshots?

Yes. By default, Amazon Elasticsearch Service will automatically create daily snapshots of each Amazon Elasticsearch domain. The daily snapshots are setup to occur between midnight and 1AM UTC. Customers will also be able to modify the timing of the automated snapshot to better suit their needs.

Q: Can I change the default settings for the automated daily snapshot provided by Amazon Elasticsearch Service?

Yes. You will be able to change the timing of the automated daily snapshot to suit your application schedule.

Q: How long are the automated daily snapshots stored by Amazon Elasticsearch Service?

Amazon Elasticsearch Service will retain the last 14 days worth of automated daily snapshots.

Q: Is there a charge for the automated daily snapshots?

There is no additional charge for the automated daily snapshots. The snapshots are stored for free in an Amazon Elasticsearch Service S3 bucket and will be made available for node recovery purposes.

Q: Can I create additional snapshots of my Amazon Elasticsearch domains as needed?

Yes. You can use the Elasticsearch snapshot API to create additional manual snapshots in addition to the daily-automated snapshots created by Amazon Elasticsearch Service. The manual snapshots are stored in your S3 bucket and will incur relevant Amazon S3 usage charges.

Q: Can snapshots created by the manual snapshot process be used to recover a domain in the event of a failure?

Yes. Customers can create a new Amazon Elasticsearch domain and load data from the snapshot into the newly created Amazon Elasticsearch domain using the Elasticsearch restore API.

Q: What happens to my snapshots when I delete my Amazon Elasticsearch domain?

The daily snapshots retained by Amazon Elasticsearch Service will be deleted as part of domain deletion. Before deleting a domain, you should consider creating a snapshot of the domain in your own S3 buckets using the manual snapshot process. The snapshots stored in your S3 bucket will not be affected if you delete your Amazon Elasticsearch domain.

Return to Top >>

Pricing

Q: How will I be charged and billed for my use of Amazon Elasticsearch Service?

You pay only for what you use, and there are no minimum or setup fees. You are billed based on:

- Amazon Elasticsearch instance hours – Based on the class (e.g. Standard Small, Large, Extra Large) of the Amazon Elasticsearch instance consumed. Partial Amazon Elasticsearch instance hours consumed are billed as full hours.
- Storage (per GB per month) – EBS Storage capacity you have provisioned to your Amazon Elasticsearch instance. If you scale your provisioned storage capacity within the month, your bill will be pro-rated.
- Provisioned IOPS per month – EBS Provisioned IOPS rate, regardless of IOPS consumed (for Amazon Elasticsearch Service Provisioned IOPS (SSD) Storage only).
- Data transfer – Regular AWS data transfer charges apply.

Please refer to the Amazon Elasticsearch Servicepricing page for detailed pricing information.

Q: When does billing of my Amazon Elasticsearch domain begin and end?

Billing commences for an Amazon Elasticsearch instance as soon as the instance is available.

Billing continues until the Amazon Elasticsearch instance terminates, which would occur upon deletion or in the event of instance failure.

Q: What defines billable instance hours for Amazon Elasticsearch Service?

Amazon Elasticsearch instance hours are billed for each hour your instance is running in an available state. If you no longer wish to be charged for your Amazon Elasticsearch instance, you must delete the domain to avoid being billed for additional instance hours. Partial Amazon Elasticsearch instance hours consumed are billed as full hours.

Return to Top >>

Amazon Kinesis FAQ

General

Q: What is Amazon Kinesis Streams?

Amazon Kinesis Streams enables you to build custom applications that process or analyze streaming data for specialized needs. You can continuously add various types of data such as clickstreams, application logs, and social media to an Amazon Kinesis stream from hundreds of thousands of sources. Within seconds, the data will be available for your Amazon Kinesis Applications to read and process from the stream.

Q: What does Amazon Kinesis Streams manage on my behalf?

Amazon Kinesis Streams manages the infrastructure, storage, networking, and configuration needed to stream your data at the level of your data throughput. You do not have to worry about provisioning, deployment, ongoing-maintenance of hardware, software, or other services for your data streams. In addition, Amazon Kinesis Streams synchronously replicates data across three facilities in an AWS Region, providing high availability and data durability.

Q: What can I do with Amazon Kinesis Streams?

Amazon Kinesis Streams is useful for rapidly moving data off data producers and then continuously processing the data, be it to transform the data before emitting to a data store, run real-time metrics and analytics, or derive more complex data streams for further processing. The following are typical scenarios for using Amazon Kinesis Streams:

- Accelerated log and data feed intake: Instead of waiting to batch up the data, you can have your data producers push data to an Amazon Kinesis stream as soon as the data is produced, preventing data loss in case of data producer failures. For example, system and application logs can be continuously added to a stream and be available for processing within seconds.

- Real-time metrics and reporting: You can extract metrics and generate reports from Amazon Kinesis stream data in real-time. For example, your Amazon Kinesis Application can work on metrics and reporting for system and application logs as the data is streaming in, rather than wait to receive data batches.
- Real-time data analytics: With Amazon Kinesis Streams, you can run real-time streaming data analytics. For example, you can add clickstreams to your Amazon Kinesis stream and have your Amazon Kinesis Application run analytics in real-time, enabling you to gain insights out of your data at a scale of minutes instead of hours or days.
- Complex stream processing: You can create Directed Acyclic Graphs (DAGs) ofAmazon Kinesis Applications and data streams. In this scenario, one or more Amazon Kinesis Applications can add data to another Amazon Kinesis stream for further processing, enabling successive stages of stream processing.

Q: How do I use Amazon Kinesis Streams?

After you sign up for Amazon Web Services, you can start using Amazon Kinesis Streams by:

- Creating an Amazon Kinesis stream through either Amazon Kinesis Management Console or *CreateStream* operation.
- Configuring your data producers to continuously add data to your stream.
- Building your Amazon Kinesis Applications to read and process data from your stream, using either Amazon Kinesis API or Amazon Kinesis Client Library (KCL).

Q: What are the limits of Amazon Kinesis Streams?

The throughput of an Amazon Kinesis stream is designed to scale without limits via increasing the number of shards within a stream. However, there are certain limits you should keep in mind while using Amazon Kinesis Streams:

- By default, Records of a stream are accessible for up to 24 hours from the time they are added to the stream. You can raise this limit to up to 7 days by enabling extended data retention.
- The maximum size of a data blob (the data payload before Base64-encoding) within one record is 1 megabyte (MB).
- Each shard can support up to 1000 PUT records per second.

For more information about other API level limits, seeAmazon Kinesis Streams Limits.

Q: How does Amazon Kinesis Streams differ from Amazon SQS?

Amazon Kinesis Streams enables real-time processing of streaming big data. It provides ordering of records, as well as the ability to read and/or replay records in the same order to multiple Amazon Kinesis Applications. The Amazon Kinesis Client Library (KCL) delivers all records for a given partition key to the same record processor, making it easier to build multiple applications reading from the same Amazon Kinesis stream (for example, to perform counting, aggregation, and filtering).

Amazon Simple Queue Service (Amazon SQS) offers a reliable, highly scalable hosted queue for storing messages as they travel between computers. Amazon SQS lets you easily move data between distributed application components and helps you build applications in which messages are processed independently (with message-level ack/fail semantics), such as automated workflows.

Q: When should I use Amazon Kinesis Streams, and when should I use Amazon SQS?

We recommend Amazon Kinesis Streams for use cases with requirements that are similar to the following:

- Routing related records to the same record processor (as in streaming MapReduce). For example, counting and aggregation are simpler when all records for a given key are routed to the same record processor.
- Ordering of records. For example, you want to transfer log data from the application host to the processing/archival host while maintaining the order of log statements.
- Ability for multiple applications to consume the same stream concurrently. For example, you have one application that updates a real-time dashboard and another that archives data to Amazon Redshift. You want both applications to consume data from the same stream concurrently and independently.
- Ability to consume records in the same order a few hours later. For example, you have a billing application and an audit application that runs a few hours behind the billing application. Because Amazon Kinesis Streams stores data for up to 7 days, you can run the audit application up to 7 days behind the billing application.

We recommend Amazon SQS for use cases with requirements that are similar to the following:

- Messaging semantics (such as message-level ack/fail) and visibility timeout. For example, you have a queue of work items and want to track the successful completion of each item independently. Amazon SQS tracks the ack/fail, so the application does not have to maintain a persistent checkpoint/cursor. Amazon SQS will delete acked messages and redeliver failed messages after a configured visibility timeout.
- Individual message delay. For example, you have a job queue and need to schedule individual

jobs with a delay. With Amazon SQS, you can configure individual messages to have a delay of up to 15 minutes.

- Dynamically increasing concurrency/throughput at read time. For example, you have a work queue and want to add more readers until the backlog is cleared. With Amazon Kinesis Streams, you can scale up to a sufficient number of shards (note, however, that you'll need to provision enough shards ahead of time).
- Leveraging Amazon SQS's ability to scale transparently. For example, you buffer requests and the load changes as a result of occasional load spikes or the natural growth of your business. Because each buffered request can be processed independently, Amazon SQS can scale transparently to handle the load without any provisioning instructions from you.

Key Amazon Kinesis Streams Concepts

Q: What is a shard?

Shard is the base throughput unit of an Amazon Kinesis stream. One shard provides a capacity of 1MB/sec data input and 2MB/sec data output. One shard can support up to 1000 PUT records per second. You will specify the number of shards needed when you create a stream. For example, you can create a stream with two shards. This stream has a throughput of 2MB/sec data input and 4MB/sec data output, and allows up to 2000 PUT records per second. You can monitor shard-level metrics in Amazon Kinesis Streams and add or remove shards from your stream dynamically as your data throughput changes by resharding the stream.

Q: What is a record?

A record is the unit of data stored in an Amazon Kinesis stream. A record is composed of a sequence number, partition key, and data blob. Data blob is the data of interest your data producer adds to a stream. The maximum size of a data blob (the data payload before Base64-encoding) is 1 megabyte (MB).

Q: What is a partition key?

Partition key is used to segregate and route records to different shards of a stream. A partition key is specified by your data producer while adding data to an Amazon Kinesis stream. For example, assuming you have a stream with two shards (shard 1 and shard 2). You can configure your data producer to use two partition keys (key A and key B) so that all records with key A are added to shard 1 and all records with key B are added to shard 2.

Q: What is a sequence number?

A sequence number is a unique identifier for eachrecord. Sequence number is assigned by Amazon Kinesis when a data producer calls *PutRecord* or *PutRecords* operation to add data to an Amazon Kinesis stream. Sequence numbers for the same partition key generally increase over time; the longer the time period between *PutRecord* or *PutRecords* requests, the larger the sequence numbers become.

Creating Amazon Kinesis Streams

Q: How do I create an Amazon Kinesis stream?

After you sign up for Amazon Web Services, you can create an Amazon Kinesis stream through either Amazon Kinesis Management Console or *CreateStream* operation.

Q: How do I decide the throughput of my Amazon Kinesis stream?

The throughput of an Amazon Kinesis stream is determined by the number ofshards within the stream. Follow the steps below to estimate the initial number of shards your stream needs. Note that you can dynamically adjust the number of shards within your stream via resharding.

1. Estimate the average size of therecord written to the stream in kilobytes (KB), rounded up to the nearest 1 KB. (*average_data_size_in_KB*)
2. Estimate the number of records written to the stream per second. (*number_of_records_per_second*)
3. Decide the number of Amazon Kinesis Applications consuming data concurrently and independently from the stream. (*number_of_consumers*)
4. Calculate the incoming write bandwidth in KB (*incoming_write_bandwidth_in_KB*), which is equal to the *average_data_size_in_KB* multiplied by the *number_of_records_per_seconds*.
5. Calculate the outgoing read bandwidth in KB (*outgoing_read_bandwidth_in_KB*), which is equal to the *incoming_write_bandwidth_in_KB* multiplied by the *number_of_consumers*.

You can then calculate the initial number of shards (*number_of_shards*) your stream needs using the following formula:

number_of_shards = max (*incoming_write_bandwidth_in_KB*/1000, *outgoing_read_bandwidth_in_KB*/2000)

Q: What is the minimum throughput I can request for my Amazon Kinesis stream?

The throughput of an Amazon Kinesis stream scales by unit ofshard. One single shard is the smallest throughput of a stream, which provides 1MB/sec data input and 2MB/sec data output.

Q: What is the maximum throughput I can request for my Amazon Kinesis stream?

The throughput of an Amazon Kinesis stream is designed to scale without limits. By default, each account can provision 10 shards per region. You can use theAmazon Kinesis Streams Limits form to request more than 10 shards within a single region.

Q: How can record size affect the throughput of my Amazon Kinesis stream?

A shard provides 1MB/sec data input rate and supports up to 1000 PUT records per sec. Therefore, if the record size is less than 1KB, the actual data input rate of a shard will be less than 1MB/sec, limited by the maximum number of PUT records per second.

Adding Data to Amazon Kinesis Streams

Q: How do I add data to my Amazon Kinesis stream?

You can add data to an Amazon Kinesis stream via*PutRecord* and *PutRecords* operations, Amazon Kinesis Producer Library (KPL), or Amazon Kinesis Agent.

Q: What is the difference between*PutRecord* and *PutRecords*?

PutRecord operation allows a single data record within an API call and*PutRecords* operation allows multiple data records within an API call. For more information about *PutRecord* and *PutRecords* operations, see *PutRecord* and *PutRecords*.

Q: What is Amazon Kinesis Producer Library (KPL)?

Amazon Kinesis Producer Library (KPL) is an easy to use and highly configurable library that helps you put data into an Amazon Kinesis stream. KPL presents a simple, asynchronous, and reliable interface that enables you to quickly achieve high producer throughput with minimal client resources.

Q: What programming languages or platforms can I use to access Amazon Kinesis API?

Amazon Kinesis API is available in Amazon Web Services SDKs. For a list of programming languages or platforms for Amazon Web Services SDKs, see Tools for Amazon Web Services.

Q: What programming language is Amazon Kinesis Producer Library (KPL) available in?

Amazon Kinesis Producer Library (KPL)'s core is built with C++ module and can be compiled to work on any platform with a recent C++ compiler. The library is currently available in a Java interface. We are looking to add support for other programming languages.

Q: What is Amazon Kinesis Agent?

Amazon Kinesis Agent is a pre-built Java application that offers an easy way to collect and send data to your Amazon Kinesis stream. You can install the agent on Linux-based server environments such as web servers, log servers, and database servers. The agent monitors certain files and continuously sends data to your stream. For more information, see Writing with Agents.

Q: What platforms do Amazon Kinesis Agent support?

Amazon Kinesis Agent currently supports Amazon Linux or Red Hat Enterprise Linux.

Q: Where do I get Amazon Kinesis Agent?

You can download and install Amazon Kinesis Agent using the following command and link:

On Amazon Linux: sudo yum install –y aws-kinesis-agent

On Red Hat Enterprise Linux: sudo yum install –y https://s3.amazonaws.com/streaming-data-agent/aws-kinesis-agent-latest.amzn1.noarch.rpm

From GitHub: awlabs/amazon-kinesis-agent

Q: How do I use Amazon Kinesis Agent?

After installing Amazon Kinesis Agent on your servers, you configure it to monitor certain files on the disk and then continuously send new data to your Amazon Kinesis stream. For more information, see Writing with Agents.

Q: What happens if the capacity limits of an Amazon Kinesis stream are exceeded while the data producer adds data to the stream?

The capacity limits of an Amazon Kinesis stream are defined by the number ofshards within the stream. The limits can be exceeded by either data throughput or the number of PUT records. While the capacity limits are exceeded, the put data call will be rejected with a *ProvisionedThroughputExceeded* exception. If this is due to a temporary rise of the stream's input data rate, retry by the data producer will eventually lead to completion of the requests. If this is due to a sustained rise of the stream's input data rate, you should increase the number of shards within your stream to provide enough capacity for the put data calls to consistently succeed. In both cases, Amazon CloudWatch metrics allow you to learn about the change of the stream's input data rate and the occurrence of *ProvisionedThroughputExceeded* exceptions.

Q: What data is counted against the data throughput of an Amazon Kinesis stream during a *PutRecord* or *PutRecords* call?

Your data blob, partition key, and stream name are required parameters of a*PutRecord* or *PutRecords* call. The size of your data blob (before Base64 encoding) and partition key will be counted against the data throughput of your Amazon Kinesis stream, which is determined by the number of shards within the stream.

Reading and Processing Data from Amazon Kinesis Streams

Q: What is an Amazon Kinesis Application?

An Amazon Kinesis Application is a data consumer that reads and processes data from an Amazon Kinesis stream. You can build your applications using either Amazon Kinesis API or Amazon Kinesis Client Library (KCL).

Q: What is Amazon Kinesis Client Library (KCL)?

Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET is a pre-built library that helps you easily build Amazon Kinesis Applications for reading and processing data from an Amazon Kinesis stream. KCL handles complex issues such as adapting to changes in stream volume, load-balancing streaming data, coordinating distributed services, and processing data with fault-tolerance. KCL enables you to focus on business logic while building applications.

Q: What is Amazon Kinesis Connector Library?

Amazon Kinesis Connector Library is a pre-built library that helps you easily integrate Amazon Kinesis Streams with other AWS services and third-party tools. Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET is required for using Amazon Kinesis Connector Library. The current version of this library provides connectors to Amazon DynamoDB, Amazon Redshift, Amazon S3, and Elasticsearch. The library also includes sample connectors of each type, plus Apache Ant build files for running the samples.

Q: What is Amazon Kinesis Storm Spout?

Amazon Kinesis Storm Spout is a pre-built library that helps you easily integrate Amazon Kinesis Streams with Apache Storm. The current version of Amazon Kinesis Storm Spout fetches data from Amazon Kinesis stream and emits it as tuples. You will add the spout to your Storm topology to leverage Amazon Kinesis Streams as a reliable, scalable, stream capture, storage, and replay service.

Q: What programming language are Amazon Kinesis Client Library (KCL), Amazon Kinesis Connector Library, and Amazon Kinesis Storm Spout available in?

Amazon Kinesis Client Library (KCL) is currently available in Java, Python, Ruby, Node.js, and .NET. Amazon Kinesis Connector Library and Amazon Kinesis Storm Spout are currently available in Java. We are looking to add support for other programming languages.

Q: Do I have to use Amazon Kinesis Client Library (KCL) for my Amazon Kinesis Application?

No, you can also use Amazon Kinesis API to build your Amazon Kinesis Application. However, we recommend using Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js |

.NET if applicable because it performs heavy-lifting tasks associated with distributed stream processing, making it more productive to develop applications.

Q: How does Amazon Kinesis Client Library (KCL) interact with an Amazon Kinesis Application?

Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET acts as an intermediary between Amazon Kinesis Streams and your Amazon Kinesis Application. KCL uses the *IRecordProcessor* interface to communicate with your application. Your application implements this interface, and KCL calls into your application code using the methods in this interface.

For more information about building application with KCL, seeDeveloping Consumer Applications for Amazon Kinesis Using the Amazon Kinesis Client Library.

Q: What is a worker and a record processor generated by Amazon Kinesis Client Library (KCL)?

An Amazon Kinesis Application can have multiple application instances and a worker is the processing unit that maps to each application instance. A record processor is the processing unit that processes data from a shard of an Amazon Kinesis stream. One worker maps to one or more record processors. One record processor maps to one shard and processes records from that shard.

At startup, an application calls into Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET to instantiate a worker. This call provides KCL with configuration information for the application, such as the stream name and AWS credentials. This call also passes a reference to an *IRecordProcessorFactory* implementation. KCL uses this factory to create new record processors as needed to process data from the stream. KCL communicates with these record processors using the *IRecordProcessor* interface.

Q: How does Amazon Kinesis Client Library (KCL) keep tracking data records being processed by an Amazon Kinesis Application?

Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET automatically creates an Amazon DynamoDB table for each Amazon Kinesis Application to track and maintain state information such as resharding events and sequence number checkpoints. The DynamoDB table shares the same name with the application so that you need to make sure your application name doesn't conflict with any existing DynamoDB tables under the same account within the same region.

All workers associated with the same application name are assumed to be working together on the same Amazon Kinesis stream. If you run an additional instance of the same application code, but with a different application name, KCL treats the second instance as an entirely separate application also operating on the same stream.

Please note that your account will be charged for the costs associated with the Amazon DynamoDB table in addition to the costs associated with Amazon Kinesis Streams.

For more information about how KCL tracks application state, seeTracking Amazon Kinesis Application state.

Q: How can I automatically scale up the processing capacity of my Amazon Kinesis Application using Amazon Kinesis Client Library (KCL)?

You can create multiple instances of yourAmazon Kinesis Application and have these application instances run across a set of Amazon EC2 instances that are part of an Auto Scaling group. While the processing demand increases, an Amazon EC2 instance running your application instance will be automatically instantiated. Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET will generate a worker for this new instance and automatically move record processors from overloaded existing instances to this new instance.

Q: Why does *GetRecords* call return empty result while there is data within my Amazon Kinesis stream?

One possible reason is that there is no record at the position specified by the currentshard iterator. This could happen even if you are using TRIM_HORIZON as shard iterator type. An Amazon Kinesis stream represents a continuous stream of data. You should call *GetRecords* operation in a loop and the record will be returned when the shard iterator advances to the position where the record is stored.

Q: What is *ApproximateArrivalTimestamp* returned in GetRecords operation?

Each record includes a value called *ApproximateArrivalTimestamp.* It is set when the record is successfully received and stored by Amazon Kinesis. This timestamp has millisecond precision and there are no guarantees about the timestamp accuracy. For example, records in a shard or across a stream might have timestamps that are out of order.

Q: What happens if the capacity limits of an Amazon Kinesis stream are exceeded while Amazon Kinesis Application reads data from the stream?

The capacity limits of an Amazon Kinesis stream are defined by the number ofshards within the stream. The limits can be exceeded by either data throughput or the number of read data calls. While the capacity limits are exceeded, the read data call will be rejected with a *ProvisionedThroughputExceeded* exception. If this is due to a temporary rise of the stream's output data rate, retry by the Amazon Kinesis Application will eventually lead to completions of the requests. If this is due to a sustained rise of the stream's output data rate, you should increase the number of shards within your stream to provide enough capacity for the read data calls to consistently succeed. In both cases, Amazon CloudWatch metrics allow you to learn about the change of the stream's output data rate and the occurrence of *ProvisionedThroughputExceeded* exceptions.

Managing Amazon Kinesis Streams

Q: How do I change the throughput of my Amazon Kinesis stream?

There are two ways to change the throughput of your stream. You can use the UpdateShardCount API or the AWS Management Console to scale the number of shards in a stream, or you can change the throughput of an Amazon Kinesis stream by adjusting the number of shards within the stream (resharding).

Q: How long does it take to change the throughput of my Amazon Kinesis stream using UpdateShardCount or the AWS Management Console?

Typical scaling requests should take a few minutes to complete. Larger scaling requests will take longer than smaller ones.

Q: What are the limitations of UpdateShardCount?

For information about limitations of UpdateShardCount, see the*Amazon Kinesis Streams Service API Reference*.

Q: Does Amazon Kinesis Streams remain available when I change the throughput of my Amazon Kinesis stream using UpdateShardCount or via resharding?

Yes. You can continue adding data to and reading data from your Amazon Kinesis stream while you use UpdateShardCount or reshard to change the throughput of the stream.

Q: What is resharding?

Resharding is the process used to scale your stream using a series of shard splits or merges. In a shard split, a single shard is divided into two shards, which increases the throughput of the stream. In a shard merge, two shards are merged into a single shard, which decreases the throughput of the stream. For more information, see Resharding a Stream in the *Amazon Kinesis Streams developer guide*.

Q: How often can I and how long does it take to change the throughput of my Amazon Kinesis stream by resharding it?

A resharding operation such as shard split or shard merge takes a few seconds. You can only perform one resharding operation at a time. Therefore, for an Amazon Kinesis stream with only one shard, it takes a few seconds to double the throughput by splitting one shard. For a stream with 1000 shards, it takes 30K seconds (8.3 hours) to double the throughput by splitting 1000 shards. We recommend increasing the throughput of your stream ahead of the time when extra throughput is needed.

Q: How do I change the data retention period of my Amazon Kinesis stream?

Amazon Kinesis stores your data for up to 24 hours by default. You can raise data retention period to up to 7 days by enabling extended data retention.

For more information about changing data retention period, seeChanging Data Retention Period.

Q: How do I monitor the operations and performance of my Amazon Kinesis stream?

Amazon Kinesis Streams Management Console displays key operational and performance metrics such as throughput of data input and output of your Amazon Kinesis streams. Amazon Kinesis Streams also integrates with Amazon CloudWatch so that you can collect, view, and analyze CloudWatch metrics for your streams and shards within those streams. For more information about Amazon Kinesis Streams metrics, see Monitoring Amazon Kinesis Streams with Amazon CloudWatch.

Please note that all stream-level metrics are free of charge. All enabled shard-level metrics are charged at Amazon CloudWatch Pricing.

Q: How do I manage and control access to my Amazon Kinesis stream?

Amazon Kinesis Streams integrates with AWS Identity and Access Management (IAM), a service that enables you to securely control access to your AWS services and resources for your users. For example, you can create a policy that only allows a specific user or group to add data to your Amazon Kinesis stream. For more information about access management and control of your stream, see Controlling Access to Amazon Kinesis Streams Resources using IAM.

Q: How do I log API calls made to my Amazon Kinesis stream for security analysis and operational troubleshooting?

Amazon Kinesis integrates with Amazon CloudTrail, a service that records AWS API calls for your account and delivers log files to you. For more information about API call logging and a list of supported Amazon Kinesis API operations, see Logging Amazon Kinesis API calls Using Amazon CloudTrail.

Q: How do I effectively manage my Amazon Kinesis streams and the costs associated with these streams?

Amazon Kinesis Streams allows you to tag your Amazon Kinesis streams for easier resource and cost management. A tag is a user-defined label expressed as a key-value pair that helps organize AWS resources. For example, you can tag your streams by cost centers so that you can categorize and track your Amazon Kinesis Streams costs based on cost centers. For more information about Amazon Kinesis Streams tagging, see Tagging Your Amazon Kinesis Streams.

Q: How can I describe how I'm utilizing my shard limit?

You can understand how you're utilizing your shard limit for an account using theDescribeLimits API. The DescribeLimits API will return the shard limit and the number of open shards in your account. If you need to raise your shard limit, please request a limit increase.

Pricing and Billing

Q: Is Amazon Kinesis Streams available in AWS Free Tier?

No. Amazon Kinesis Streams is not currently available in AWS Free Tier. AWS Free Tier is a program that offers free trial for a group of AWS services. For more details about AWS Free Tier, see AWS Free Tier.

Q: How much does Amazon Kinesis Streams cost?

Amazon Kinesis Streams uses simple pay as you go pricing. There is neither upfront cost nor minimum fees and you only pay for the resources you use. The costs of Amazon Kinesis Streams has two core dimensions and one optional dimension:

- Hourly Shard cost determined by the number of shards within your Amazon Kinesis stream.
- PUT Payload Unit cost determined by the number of 25KB payload units that your data producers add to your stream.
- Optional Extended Data Retention cost determined by number of shard hours incurred by your stream.

For more information about Amazon Kinesis Streams costs, see Amazon Kinesis Streams Pricing.

Q: Does my PUT Payload Unit cost change by using *PutRecords* operation instead of *PutRecord* operation?

PUT Payload Unit charge is calculated based on the number of 25KB payload units added to your Amazon Kinesis stream. PUT Payload Unit cost is consistent when using *PutRecords* operation or *PutRecord* operation.

Q: Am I charged for shards in "CLOSED" state?

A shard could be in "CLOSED" state after resharding. You will not be charged for shards in "CLOSED" state.

Q: Other than Amazon Kinesis Streams costs, are there any other costs that might incur to my Amazon Kinesis Streams usage?

If you use Amazon EC2 for running your Amazon Kinesis Applications, you will be charged for Amazon EC2 resources in addition to Amazon Kinesis Streams costs.

Amazon Kinesis Client Library (KCL) uses Amazon DynamoDB table to track state information of record processing. If you use KCL for you Amazon Kinesis Applications, you will be charged for Amazon DynamoDB resources in addition to Amazon Kinesis Streams costs.

If you enable Enhanced Shard-Level Metrics, you will be charged for Amazon CloudWatch cost associated with enabled shard-level metrics in addition to Amazon Kinesis Streams costs.

Please note that the above are three common but not exhaustive cases.

Amazon Machine Learning FAQ

General

Q: What is Amazon Machine Learning?

Amazon Machine Learning is a machine service that allows you to easily build predictive applications, including fraud detection, demand forecasting, and click prediction. Amazon Machine Learning uses powerful algorithms that can help you create machine learning models by finding patterns in existing data, and using these patterns to make predictions from new data as it becomes available. The AWS Management Console and API provide data and model visualization tools, as well as wizards to guide you through the process of creating machine learning models, measuring their quality and fine-tuning the predictions to match your application requirements. Once the models are created, you can get predictions for your application by using the simple API, without having to implement custom prediction generation code or manage any infrastructure. Amazon Machine Learning is highly scalable and can generate billions of predictions, and serve those predictions in real-time and at high throughput. With Amazon Machine Learning there is no setup cost and you pay as you go, so you can start small and scale as your application grows.

Q: What can I do with Amazon Machine Learning?

You can use Amazon Machine Learning to create a wide variety of predictive applications. For example, you can use Amazon Machine Learning to help you build applications that flag suspicious transactions, detect fraudulent orders, forecast demand, personalize content, predict user activity, filter reviews, listen to social media, analyze free text, and recommend items.

Q: What is machine learning?

Machine learning (ML) is a technology that helps you use historical data to make informed business decisions. ML algorithms discover patterns in data and construct mathematical models using these patterns. Then, you can use the models to make predictions on future data. For example, one possible application of machine learning is detecting fraudulent transactions based on examples of both successful and failed past purchases.

Q: How do I get started with Amazon Machine Learning?

The best way to get started with Amazon Machine Learning is to follow thetutorial in the *Amazon Machine Learning Developer Guide.* The tutorial guides you through creating a machine learning model from a sample dataset, evaluating this model, and using it to create predictions. After completing the tutorial, you can use Amazon Machine Learning to create your own ML models. For more information, see the Amazon Machine Learning Developer Guide and the Amazon Machine Learning API Reference.

Q: What is training data?

Training data is used to create machine learning models. It consists of known data points from the past. You can use Amazon Machine Learning to extract patterns from this data, and use them to build machine learning models.

Q: What is the target attribute?

The target attribute is a special attribute in the training data that contains the information that Amazon Machine Learning attempts to predict. For example, let's say you want to build a model that predicts whether a transaction is fraudulent or not. Your training data contains metadata on a past transaction that has a target attribute of "1" if the transaction was ultimately declined by the bank, or "0" otherwise. You use Amazon Machine Learning to discover patterns that connect the target attribute with the transaction metadata (all other attributes). You use ML models based on these patterns to make a prediction without the target attribute present. In this example, it means predicting whether a transaction is fraudulent based on its metadata, before knowing whether the bank will reject it or not.

Q: What algorithm does Amazon Machine Learning use to generate models?

Amazon Machine Learning currently uses an industry-standard logistic regression algorithm to generate models.

Q: In which AWS regions is Amazon Machine Learning available?

For a list of the supported Amazon Machine Learning AWS regions, please visitthe AWS Region Table for all AWS global infrastructure. Also for more information, see Regions and Endpoints in the *AWS General Reference.*

Q: What is the service availability of Amazon Machine Learning?

Amazon Machine Learning is designed for high availability. There are no maintenance windows or scheduled downtimes. The API for model training, evaluation, and batch prediction runs in Amazon's proven, high-availability data centers, with service stack replication configured across three facilities in each AWS region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: What security measures does Amazon Machine Learning have?

Amazon Machine Learning ensures that ML models and other system artifacts are encrypted in transit and at rest. Requests to the Amazon Machine Learning API and console are made over a secure (SSL) connection. You can use AWS Identity and Access Management (AWS IAM) to control which IAM users have access to specific Amazon Machine Learning actions and resources.

Creating Models

Q: Where do I store my data?

You can use Amazon Machine Learning to read your data from three data stores: (a) one or more files in Amazon S3, (b) results of an Amazon Redshift query, or (c) results of an Amazon Relational Database Service (RDS) query when executed against a database running with the MySQL engine. Data from other products can usually be exported into CSV files in Amazon S3, making it accessible to Amazon Machine Learning. For detailed instructions for configuring permissions that enable Amazon Machine Learning to access the supported data stores, see the Amazon Machine Learning Developer Guide.

Q: Are there limits to the size of the dataset I can use for training?

Amazon Machine Learning can train models on datasets up to 100 GB in size.

Q: How do I know if my data has errors?

You can use Amazon Machine Learning to detect data formatting errors. The data insights feature of the Amazon Machine Learning service console helps you find deeper errors within your data—for example, fields that are empty or contain unexpected values. Amazon Machine Learning will be able to train ML models and generate accurate predictions in the presence of a small number of both kinds of data errors, enabling your requests to succeed even if some data observations are invalid or incorrect.

Q: What do I do if my data is incomplete or some information is missing?

It is always best to ensure that your data is as complete and accurate as possible. The learning algorithms of Amazon Machine Learning tolerates small amounts of incomplete or missing information without it adversely affecting model quality; as the number of mistakes increases, the resulting model quality will be degraded. Amazon Machine Learning stops processing your model training request if the number of records that fail processing is greater than either 10,000 or 10% of all records in the dataset, whichever comes first.

To correct incomplete or missing information, you need to return to the master datasource and either correct the data in that source, or exclude the observations with incomplete or missing information from the datasets used to train Amazon Machine Learning models. For example, if you find that some rows in an Amazon Redshift table contain invalid values, you can modify the query used to select data for Amazon Machine Learning to exclude these rows.

Q: How do I know if my model is giving accurate predictions?

Amazon Machine Learning includes powerful model evaluation features. You can use Amazon Machine Learning to compute an industry-standard evaluation metric for any of your models, helping you understand these models' predictive quality. You can also use Amazon Machine Learning to ensure that the model evaluation is unbiased by choosing to withhold a part of the training data for evaluation purposes, ensuring that the model is never evaluated with data points that were seen at the training time. The Amazon Machine Learning service console provides powerful, easy-to-use tools to explore and understand the results of model evaluations.

Q: How do I tune my model if it isn't giving the results I want?

The best way to increase a model's quality is by using more and higher-quality data to train it. Adding more observations, adding additional types of information (features), and transforming your data to optimize the learning process (feature engineering) are all great ways to improve the model's predictive accuracy. You can use Amazon Machine Learning to create many prototype models, and you can use the built-in data processors of Amazon Machine Learning to make several common types of feature engineering as simple as editing a line in the built-in "recipe" language. Additionally, Amazon Machine Learning can automatically create a suggested data transformation recipe based on your data when you create a new datasource object pointing to your data—this recipe will be automatically optimized based on your data contents.

Amazon Machine Learning also provides several parameters for tuning the learning process: (a) target size of the model, (b) the number of passes to be made over the data, and (c) the type and amount of regularization to apply to the model. The default settings of Amazon Machine Learning works well for many real-world ML tasks, but can be adjusted as needed by using either the service console or API.

Finally, one important aspect of model tuning to consider is how predictions generated by your ML model are interpreted by your application, to align them optimally with the business goals. Amazon Machine Learning helps you adjust the interpretation cut-off score for binary classification models, enabling you to make an informed trade-off between different kinds of mistakes that a trained model can make. For example, some applications are very tolerant of false positive errors, but false negative errors are highly undesirable—the Amazon Machine Learning service console helps you adjust the score cut-off to align with this requirement. For more information, see Evaluating ML Models in the *Amazon Machine Learning Developer Guide*.

Q: Can I export my models out of Amazon Machine Learning?

No.

Q: Can I import existing models into Amazon Machine Learning?

No.

Q: Does Amazon Machine Learning need to make a permanent copy of my data to create machine learning models?

No. Amazon Machine Learning need only read-access to your data to find and extract the patterns within it, and store them within ML models. ML models are not copies of your data. When accessing data stored in Amazon Redshift or Amazon RDS, Amazon Machine Learning will export the query results into an S3 location of your choice, and then read these results from S3. You will retain full ownership of this temporary data copy, and will be able to remove it after the Amazon Machine Learning operation is completed.

Back to Top >>

Generating Predictions

Q: Once my model is ready, how do I get predictions for my applications?

You can use Amazon Machine Learning to retrieve predictions in two ways: using the batch API or real-time API. The batch API is used to request predictions for a large number of input data records—it works offline, and returns all the predictions at once. The real-time API is used to request predictions for individual input data records, and returns the predictions immediately. The real-time API can be used at high throughput, generating multiple predictions at the same time in response to parallel requests.

Any ML model built with Amazon Machine Learning can be used through either the batch API or real-time API—the choice is yours, and depends only on your application's requirements. You typically use the batch API for applications that operate on bulk data records, and the real-time API for interactive web, mobile and desktop applications.

Q: How fast can the Amazon Machine Learning real-time API generate predictions?

Most real-time prediction requests return a response within 100 MS, making them fast enough for interactive web, mobile, or desktop applications. The exact time it takes for the real-time API to generate a prediction varies depending on the size of the input data record, and the complexity of the data processing "recipe" associated with the ML model that is generating the predictions

Q: How many concurrent real-time API requests does Amazon Machine Learning support?

Each ML model that is enabled for real-time predictions is assigned an endpoint URL. By default, you can request up to 200 transactions per second (TPS) from any real-time prediction endpoint. Contact customer support if this limit is not sufficient for your application's needs.

Q: How quickly can Amazon Machine Learning return batch predictions?

The batch prediction API is fast and efficient. The time it takes to return the batch prediction results depends on several factors, including (a) the size of the input data, (b) the complexity of the data processing “recipe” associated with the ML model that is generating the predictions, and (c) the number of other batch jobs (data processing, model training, evaluation, and other batch processing requests) that are simultaneously running in your account, among others. By default, Amazon Machine Learning executes up to five batch jobs simultaneously. Contact customer support if this limit is not sufficient for your application’s needs.

Q: How can I monitor how my predictions are performing?

Monitoring your prediction performance takes two primary forms: (a) monitoring the volume of batch and real-time prediction traffic, and (b) monitoring the quality of the predictive models.

You can monitor the volume of your prediction traffic by consulting the Amazon CloudWatch metrics that are published by Amazon Machine Learning into your CloudWatch account. For each ML model ID that has received either batch or real-time predictions during the monitoring period, Amazon Machine Learning will publish the number of data records for which predictions were successfully generated, and the number of ML records that failed parsing, resulting in no prediction being generated.

To monitor the quality of your ML model over time, an industry best practice is to regularly capture a random sample of data records that have been submitted by your application for prediction, obtain true answers (also known as “targets”), and then use Amazon Machine Learning to create an evaluation of the resulting dataset. Amazon Machine Learning will compute a model quality metric by comparing the targets with the predictions being generated. If you find that the quality of the metrics is decreasing over time, it is likely an indicator that you need to train a new model with new data points, as the data that was originally used to train a model is no longer matching the real world. For example, if you use your ML model to detect fraudulent transactions, you might find that its quality drops over time because new methods of transaction fraud, not known at the time of model training, have appeared. You can counter this trend by training a new ML model, with examples of the latest fraudulent transactions, enabling Amazon Machine Learning to discover the patterns that identify these transactions, among others.

Amazon QuickSight FAQ

General

Q: What is Amazon QuickSight?

Amazon QuickSight is a very fast, easy-to-use, cloud-powered business analytics service that makes it easy for all employees within an organization to build visualizations, perform ad-hoc analysis, and quickly get business insights from their data, anytime, on any device. Upload CSV and Excel files; connect to SaaS applications like Salesforce; access on-premises databases like

SQL Server, MySQL, and PostgreSQL; and seamlessly discover your AWS data sources such as Amazon Redshift, Amazon RDS, Amazon Aurora, and Amazon S3. QuickSight enables organizations to scale their business analytics capabilities to hundreds of thousands of users, and delivers fast and responsive query performance by using a robust in-memory engine (SPICE).

Q: How is Amazon QuickSight different from traditional Business Intelligence (BI) solutions?

Traditional BI solutions often require teams of data engineers to spend months building complex data models before generating a report. They typically lack interactive ad-hoc data exploration and visualization, limiting users to canned reports and pre-selected queries. Traditional BI solutions also require significant up-front investment in complex and costly hardware and software, and then customers to invest in even more infrastructure to maintain fast query performance as database sizes grow. This cost and complexity makes it difficult for companies to enable analytics solutions across their organizations. Amazon QuickSight has been designed to solve these problems by bringing the scale and flexibility of the AWS Cloud to business analytics. Unlike traditional BI or data discovery solutions, getting started with Amazon QuickSight is simple and fast. When you log in, Amazon QuickSight seamlessly discovers your data sources in AWS services such as Amazon Redshift, Amazon RDS, and Amazon Simple Storage Service (Amazon S3). You can connect to any of the data sources discovered by Amazon QuickSight and get insights from this data in minutes. You can choose for Amazon QuickSight to keep the data in SPICE up-to-date as the data in the underlying sources change. SPICE supports rich data discovery and business analytics capabilities to help customers derive valuable insights from their data without worrying about provisioning or managing infrastructure. Organizations pay a low monthly fee for each Amazon QuickSight user, eliminating the cost of long-term licenses. With Amazon QuickSight, organizations can deliver rich business analytics functionality to all employees without incurring a huge cost upfront.

Q: What is SPICE?

Amazon QuickSight is built with "SPICE" – a Super-fast, Parallel, In-memory Calculation Engine. Built from the ground up for the cloud, SPICE uses a combination of columnar storage, in-memory technologies enabled through the latest hardware innovations and machine code generation to run interactive queries on large datasets and get rapid responses. SPICE supports rich calculations to help you derive valuable insights from your analysis without worrying about provisioning or managing infrastructure. Data in SPICE is persisted until it is explicitly deleted by the user. SPICE also automatically replicates data for high availability and enables QuickSight to scale to hundreds of thousands of users who can all simultaneously perform fast interactive analysis across a wide variety of AWS data sources.

Q: How can I get started with Amazon QuickSight?

To get started, sign up to Amazon Quicksight and get 1 user and 1GB of SPICE capacity for free. If you're already signed up, you can access QuickSight at https://quicksight.aws.amazon.com.

Mobile and Web Access

Q: Can I use Amazon QuickSight on my mobile device?

The iPhone app for Amazon QuickSight lets you access your data anywhere, and explore analyses, stories, and dashboards. Look for Android and tablet support, as well as viewing and creating annotations in offline mode in the future. You can also use any of the modern browsers running on laptops or desktops or access Amazon QuickSight or using a web browser on any mobile device.

Q: On which browsers is Amazon QuickSight supported?

Amazon QuickSight supports the latest versions of Mozilla Firefox, Chrome, Safari, Internet Explorer version 10 and above and Edge.

Data Management

Q: Which data sources does Amazon QuickSight support?

You can connect to AWS data sources including Amazon RDS, Amazon Aurora, Amazon Redshift and Amazon S3. You can also upload Excel spreadsheets or flat files (CSV, TSV, CLF, and ELF), connect to on-premises databases like SQL Server, MySQL and PostgreSQL and import data from SaaS applications like Salesforce.

Q: Can I connect Amazon QuickSight to my Amazon EC2 or on-premises database?

Yes. In order to connect Amazon QuickSight to an Amazon EC2 or on-premises database, you need to add the Amazon QuickSight IP range to the authorized list (whitelist) in your hosted database.

Q: How do I upload my data files into Amazon QuickSight?

You can upload XLSX, CSV, TSV, CLF, XLF data files directly from Amazon QuickSight website. You can also upload them to an Amazon S3 bucket and point Amazon QuickSight to the Amazon S3 object.

Q: How do I access my data in AWS data sources?

Amazon QuickSight seamlessly discovers your AWS data sources that are available in your account with your approval. You can immediately start browsing the data and building visualizations. You can also explicitly connect to other AWS data sources that are not in your account or in a different region by providing connection details for those sources.

Q: My source data is not in a clean format. How do I format and transform the data before visualizing?

Amazon QuickSight lets you prepare data that is not ready for visualization. Select the "Edit/Preview Data" button in the connection dialog. Amazon QuickSight supports various functions to format and transform your data. You can alias data fields and change data types. You can subset your data using built in filters and perform database join operations using drag and drop. You can also create calculated fields using mathematical operations and built-in functions such conditional statements, string, numerical and date functions.

Q: How much data can I analyze with Amazon QuickSight?

With Amazon QuickSight you don't need to worry about scale. You can seamlessly grow your data from a few hundred megabytes to many terabytes of data without managing any infrastructure.

User Management

Q: How do I add manage access to Amazon QuickSight?

When you create a new Amazon QuickSight account, you have administrative privileges by default. If you are invited to become an Amazon QuickSight user, whoever invites you assigns you either an ADMIN or a USER role. If you have an ADMIN role, you can create and delete user accounts, purchase annual subscriptions and SPICE capacity in addition to using the service.

To create a user account, you send an email invitation to the user via an in-application interface, and then the user completes the account creation by specifying a password and signing in.

Visualization and Analysis

Q: How do I create an analysis with Amazon QuickSight?

Creating an analysis is simple. Amazon QuickSight seamlessly discovers data in popular AWS data repositories within your AWS account. Simply point Amazon QuickSight to one of the discovered data sources. To connect to another AWS data source that is not in your AWS account or in a different region, you can provide the connection details of the source. Then, select a table and start analyzing your data. You can also upload spreadsheets and CSV files and use Amazon QuickSight to analyze your files. To create a visualization, start by selecting the data fields you want to analyze, or drag the fields directly on to the visual canvas, or a combination of both actions. Amazon QuickSight will automatically select the appropriate visualization to display based on the data you've selected.

Q: How does Amazon QuickSight select the right visualization to use for my data?

Amazon QuickSight has an innovative technology called AutoGraph that allows it to select the most appropriate visualizations based on the properties of the data, such as cardinality and data type. The visualization types are chosen to best reveal the data and relationships in an effective way.

Q: How do I create a dashboard?

Dashboards are a collection of visualizations, tables, and other visual displays arranged and visible together. With Amazon QuickSight, you can compose a dashboard within an analysis by arranging the layouts and size of visualizations and then publish the dashboard to an audience within your organization.

Q: What types of visualizations are supported in Amazon QuickSight?

Amazon QuickSight supports assorted visualizations that facilitate different analytical approaches:

- Comparison and distribution
  - Bar charts (several assorted variants)
- Changes over time
  - Line graphs
  - Area line charts
- Correlation
  - Scatter plots
  - Heat maps
- Aggregation
  - Pie graphs
  - Tree maps
- Tabular
  - Pivot tables

Q: What is a suggested visualization? How does Amazon QuickSight generate suggestions?

Amazon QuickSight comes with a built-in suggestion engine that provides you with suggested visualizations based on the properties of the underlying datasets. Suggestions serve as possible first or next-steps of an analysis and removes the time-consuming task of interrogating and understanding the schema of your data. As you work with more specific data, the suggestions will update to reflect the next steps appropriate to your current analysis.

Q: What are Stories?

Stories are guided tours through specific views of an analysis. They are used to convey key points, a thought process, or the evolution of an analysis for collaboration. You can construct them in Amazon QuickSight by capturing and annotating specific states of the analysis. When readers of the story click on an image in the story, they are then taken into the analysis at that point, where they can explore on their own.

Q: What type of calculations does Amazon QuickSight enable?

You can perform typical arithmetic and comparison functions; conditional functions such as if,then; and date, numeric, and string calculations.

Q: How can I get sample data to explore in QuickSight?

For your convenience, sample analyses are automatically generated when you create an account in Amazon QuickSight. The raw data can also be downloaded from the links below:

- Business overview
- People overview
- Sales pipeline
- Web and marketing analytics

Security and Access

Q: How is data transmitted to Amazon QuickSight?

You have several options to get your data into Amazon QuickSight: file upload, connect to AWS data sources, connect to external data stores over JDBC/ODBC, or through other API-based data store connectors.

Q: Can I choose the AWS region to connect to hosted or on-premises databases over JDBC/ODBC?

Yes. For better performance and user interactivity, customers are advised to use the region where your data is stored. The Amazon QuickSight auto discovery feature detects data sources only within the AWS region of the Amazon QuickSight endpoint to which you are connected. For a list of the supported Amazon QuickSight AWS regions, please visit the AWS Region Table for all AWS global infrastructure.

Q: Does Amazon QuickSight support multi-factor authentication?

Yes. You can enable multi-factor authentication (MFA) for your AWS account via the AWS Management console.

Q: How do I connect my VPC to Amazon QuickSight?

If your VPC has been set up with public connectivity, you can add Amazon QuickSight’s IP address range to your database instances’ security group rules to enable traffic flow into your VPC and database instances.

Sharing

Q: How do I share an analysis, dashboard, or story in Amazon QuickSight?

You can share an analysis, dashboard, or story using the share icon from the QuickSight service interface. You will be able to select the recipients (email address, username or group name), permission levels, and other options before sharing the content with others.

Upgrading from Standard to Enterprise Edition

Q: Can I switch between Amazon QuickSight Enterprise Edition and Standard Edition after signing up?

Yes, you can switch between Standard and Enterprise Editions at any point in time. To do so, you will have to unsubscribe from Standard Edition and then follow the subscription flow to sign up for the Enterprise Edition (or vice versa).

Q: Will my data and users be migrated when I switch editions between Standard and Enterprise?

At this point we do not support migration of data and users between versions. Once you complete the switch to either version, you will be required to re-invite users to your account and re-provision datasets that you might require.

Amazon API Gateway FAQ

General

Q: What is Amazon API Gateway?

Amazon API Gateway is a fully managed service that makes it easy for developers to publish, maintain, monitor, and secure APIs at any scale. With a few clicks in the AWS Management Console, you can create an API that acts as a “front door” for applications to access data, business logic, or functionality from your back-end services, such as applications running on Amazon Elastic Compute Cloud (Amazon EC2), code running on AWS Lambda, or any web application. Amazon API Gateway handles all of the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, authorization and access control, monitoring, and API version management. Amazon API Gateway has no minimum fees or startup costs. You pay only for the API calls you receive and the amount of data transferred out.

Q: Why use Amazon API Gateway?

Amazon API Gateway provides developers with a simple, flexible, fully managed, pay-as-you-go service that handles all aspects of creating and operating robust APIs for application back ends. With Amazon API Gateway, you can launch new services faster and with reduced investment so you can focus on building your core business services. Amazon API Gateway was built to help you with several aspects of creating and managing APIs:

> 1) **Metering.** API Gateway helps you define plans that meter and restrict third-party developer access to your APIs. You can define a set of plans, configure throttling, and quota limits on a per API key basis. API Gateway automatically meters traffic to your APIs and lets you extract utilization data for each API key.
>
> 2) **Security.** API Gateway provides you with multiple tools to authorize access to your APIs and control service operation access. Amazon API Gateway allows you to leverage AWS administration and security tools, such as AWS Identity and Access Management (IAM) and Amazon Cognito, to authorize access to your APIs. Amazon API Gateway can verify signed API calls on your behalf using the same methodology AWS uses for its own APIs. Using custom authorizers written as AWS Lambda functions, API Gateway can also help you verify incoming bearer tokens, removing authorization concerns from your backend code.
>
> 3) **Resiliency.** Amazon API Gateway helps you manage traffic with throttling so that backend operations can withstand traffic spikes. Amazon API Gateway also helps you improve the performance of your APIs and the latency your end users experience by caching the output of API calls to avoid calling your backend every time.
>
> 4) **Operations Monitoring.** After an API is published and in use, API Gateway provides you with a metrics dashboard to monitor calls to your services. The Amazon API Gateway

dashboard, through integration with Amazon CloudWatch, provides you with backend performance metrics covering API calls, latency data and error rates. You can enable detailed metrics for each method in your APIs and also receive error, access or debug logs in CloudWatch Logs**.**

5) **Lifecycle Management.** After an API has been published, you often need to build and test new versions that enhance or add new functionality. Amazon API Gateway lets you operate multiple API versions and multiple stages for each version simultaneously so that existing applications can continue to call previous versions after new API versions are published.

6) **Designed for Developers**. Amazon API Gateway allows you to quickly create APIs and assign static content for their responses to reduce cross-team development effort and time-to-market for your applications. Teams who depend on your APIs can begin development while you build your backend processes.

Q: How do I get started with Amazon API Gateway?

You can quickly and easily create a custom API using Amazon API Gateway. For a simple “Hello World” example, follow these steps:

1. Go to the Amazon API Gateway console.

2. Select an existing REST API or create a new one by entering a name for the API.

3. On the REST API tree view, click “Create Resource”.

4. Choose a name for your resource, such as “cars”.

5. With the new resource selected, click the button to create a new method and select the HTTP verb associated with the method (for example, GET).

6. Select the integration type (for example, HTTP Proxy), and enter the URL the Amazon API Gateway should call.

7. Define how requests and responses are transformed using a mapping template, or accept the default settings to pass all of the request and response data through without applying any transformation.

8. Configure the method’s security settings.

9. Deploy your new API to a stage.

10. From the **Stage management** page, set up caching and throttling.

11. On the **Client Platforms** tab in the Amazon API Gateway console, click the button to download the Android, iOS SDK, or JavaScript library that contains helper methods to call your sayHello operation. The SDK library makes calling your APIs similar to calling a local method. The client SDK automatically handles retries, informing the developer of network or other fault conditions. The SDK library includes the logic necessary to authenticate the client application to your APIs.

12. Integrate the downloaded SDK into your mobile application. Write the code to invoke your custom API. For example, to invoke the getCar(int carId) API in an iOS application:

```
–(void)getSampleCar

{

NSString *response = [MyServiceClient getCar:1323];

NSLog( @”Response was [%@]”, response );

}
```

13. Run your application.

Q: Can I create HTTPS endpoints?

Yes, all of the APIs created with Amazon API Gateway expose HTTPS endpoints only. Amazon API Gateway does not support unencrypted (HTTP) endpoints. By default, Amazon API Gateway assigns an internal domain to the API that automatically uses the Amazon API Gateway certificate. When configuring your APIs to run under a custom domain name, you can provide your own certificate for the domain.

Q: What data types can I use with Amazon API Gateway?

APIs built on Amazon API Gateway can accept any payloads sent over HTTP. Typical data formats include JSON, XML, query string parameters, and request headers. You can declare any content type for your API’s responses, and then use the transform templates to change the back-end response into your desired format.

Q: With what backends can Amazon API Gateway communicate?

Amazon API Gateway can execute AWS Lambda functions in your account, or call HTTP endpoints hosted on AWS Elastic Beanstalk, Amazon EC2, and also non-AWS hosted HTTP based operations that are accessible via the public Internet. API Gateway also allows you to specify a mapping template to generate static content to be returned, helping you mock your APIs before the backend is ready. You can also integrate API Gateway with other AWS services directly – for example, you could expose an API method in API Gateway that sends data directly to Amazon Kinesis.

Q: For which client platforms can Amazon API Gateway generate SDKs?

API Gateway generates custom SDKs for mobile app development with Android and iOS, and for web app development with JavaScript. Once an API and its models are defined in API

Gateway, you can use the AWS console or the API Gateway APIs to generate and download a client SDK.

Q: In which AWS regions is Amazon API Gateway available?

Please refer to Regional Products and Services for details of Amazon API Gateway service availability by region.

Q: What can I manage through the Amazon API Gateway console?

Through the Amazon API Gateway console, you can define the REST API and its associated resources and methods, manage the API lifecycle, generate client SDKs and view API metrics. You can also use the API Gateway console to define your APIs' usage plans, manage developers' API keys, and configure throttling and quota limits. All of the same actions are available through the API Gateway APIs.

Q: What is a REST API?

In Amazon API Gateway, a REST API is a group of resources and methods, or*endpoints*. REST APIs can be deployed to different stages and cloned to new versions.

Q: What is a resource?

A resource is a typed object that is part of your API's domain. Each resource may have associated a data model, relationships to other resources, and can respond to different methods. You can also define resources as variables to intercept requests to multiple child resources.

Q: What is a method?

Each resource within a REST API can support one or more of the standard HTTP methods. You define which verbs should be supported for each resource (GET, POST, PUT, PATCH, DELETE, HEAD, OPTIONS) and their implementation. For example, a GET to the cars resource should return a list of cars. To connect all methods within a resource to a single backend endpoint, API Gateway also supports a special "ANY" method.

Q: What is an usage plan?
Usage plans help you declare plans for third-party developers that restrict access only to certain APIs, define throttling and request quota limits, and associate them with API keys. You can also extract utilization data on an per-API key basis to analyze API usage and generate billing documents. For example, you can create a basic, professional, and enterprise plans – you can configure the basic usage plan to only allow 1,000 requests per day and a maximum of 5 requests per second (RPS).

Q: What is the Amazon API Gateway API lifecycle?

With Amazon API Gateway, each REST API can have multiple stages. Stages are meant to help with the development lifecycle of an API -- for example, after you've built your APIs and you deploy them to a development stage, or when you are ready for production, you can deploy them to a production stage.

Q: What is a stage?

In Amazon API Gateway, stages are similar to tags. They define the path through which the deployment is accessible. For example, you can define a development stage and deploy your cars API to it. The resource will be accessible at https://www.myapi.com/dev/cars. You can also set up custom domain names to point directly to a stage, so that you don't have to use the additional path parameter. For example, if you pointed myapi.com directly to the development stage, you could access your cars resource at https://www.myapi.com/cars. Stages can be configured using variables that can be accessed from your API configuration or mapping templates.

Q: What are stage variables?
Stage variables let you define key/value pairs of configuration values associated with a stage. These values, similarly to environment variables, can be used in your API configuration. For example, you could define the HTTP endpoint for your method integration as a stage variable, and use the variable in your API configuration instead of hardcoding the endpoint – this allows you to use a different endpoint for each stage (e.g. dev, beta, prod) with the same API configuration. Stage variables are also accessible in the mapping templates and can be used to pass configuration parameters to your Lambda or HTTP backend.

Q: What if I mistakenly deployed to a stage?

Amazon API Gateway saves the history of your deployments. At any point, using the Amazon API Gateway APIs or the console, you can roll back a stage to a previous deployment.

Q: Can I run multiple versions of the same REST API?

Yes. Amazon API Gateway gives you the ability to clone an existing API. When you are ready to start working on the next major version of your API, you will be able to keep working on your version 1 and version 2 APIs simultaneously.

Q: Can I use my Swagger API definitions?

Yes. You can use our open sourceSwagger importer tool to import your Swagger API definitions into Amazon API Gateway. With the Swagger importer tool you can create and deploy new APIs as well as update existing ones.

Security and Authorization

Q: How do I authorize access to my APIs?

With Amazon API Gateway, you can optionally set your API methods to require authorization. When setting up a method to require authorization you can leverage AWS Signature Version 4 or custom authorizers to support your own bearer token auth strategy.

Q: How does AWS Signature Version 4 work?

You can use AWS credentials -- access and secret keys – to sign requests to your service and authorize access like other AWS services. The signing of an Amazon API Gateway API request is managed by the custom API Gateway SDK generated for your service. You can retrieve temporary credentials associated with a role in your AWS account using Amazon Cognito.

Q: What is a custom authorizer?

Custom authorizers are AWS Lambda functions. With custom request authorizers, you will be able to authorize access to APIs using a bearer token auth strategy such as OAuth. When an API is called, API Gateway checks if a custom authorizer is configured, API Gateway then calls the Lambda function with the incoming authorization token. You can use Lambda to implement various authorization strategies (e.g. JWT verification, OAuth provider callout) that return IAM policies which are used to authorize the request. If the policy returned by the authorizer is valid, API Gateway will cache the policy associated with the incoming token for up to 1 hour.

Q: Can Amazon API Gateway generate API keys for distribution to third-party developers?

Yes. API Gateway can generate API keys and associate them with an usage plan. Calls received from each API key are monitored and included in the Amazon CloudWatch Logs you can enable for each stage. However, we do not recommend you use API keys for authorization. You should use API keys to monitor usage by third-party developers and leverage a stronger mechanism for authorization, such as signed API calls or OAuth.

Q: How can I address or prevent API threats or abuse?

Amazon API Gateway supports throttling settings for each method in your APIs. You can set a standard rate limit and a burst rate limit per second for each method in your REST APIs. Further, Amazon API Gateway automatically protects your backend systems from distributed denial-of-service (DDoS) attacks, whether attacked with counterfeit requests (Layer 7) or SYN floods (Layer 3).

Q: Can Amazon API Gateway work within an Amazon VPC?

No. Amazon API Gateway endpoints are always public to the Internet. Proxy requests to backend operations also need to be publicly accessible on the Internet. However, you can generate a client-side SSL certificate in Amazon API Gateway to verify that requests to your backend systems were sent by API Gateway using the public key of the certificate.

Q: Can I verify that it is API Gateway calling my backend?

Yes. Amazon API Gateway can generate a client-side SSL certificate and make the public key of that certificate available to you. Calls to your backend can be made with the generated certificate, and you can verify calls originating from Amazon API Gateway using the public key of the certificate.

Q: Can I use AWS CloudTrail with Amazon API Gateway?

Yes. Amazon API Gateway is integrated with AWS CloudTrail to give you a full auditable history of the changes to your REST APIs. All API calls made to the Amazon API Gateway APIs to create, modify, delete, or deploy REST APIs are logged to CloudTrail in your AWS account.

Management, Metrics and Logging

Q: How can I monitor my Amazon API Gateway APIs?

Amazon API Gateway logs API calls, latency, and error rates to Amazon CloudWatch in your AWS account. The metrics are also available through the Amazon API Gateway console in a REST API dashboard. API Gateway also meters utilization by third-party developers, the data is available in the API Gateway console and through the APIs.

Q: Can I set up alarms on the Amazon API Gateway metrics?

Yes, Amazon API Gateway sends logging information and metrics to Amazon CloudWatch. You can utilize the Amazon CloudWatch console to set up custom alarms.

Q: How can I set up metrics for Amazon API Gateway?

By default, Amazon API Gateway monitors traffic at a REST API level. Optionally, you can enable detailed metrics for each method in your REST API from the deployment configuration APIs or console screen. Detailed metrics are also logged to Amazon CloudWatch and will be charged at the CloudWatch rates.

Q: Can I determine which version of the API my customers are using?

Yes. Metric details are specified by REST API and stage. Additionally, you can enable metrics for each method in your REST API.

Q: Does Amazon API Gateway provide logging support?

Yes. Amazon API Gateway integrates with Amazon CloudWatch Logs. You can optionally enable logging for each stage in your API. For each method in your REST APIs, you can set the verbosity of the logging, and if full request and response data should be logged.

Q: How quickly are logs available?

Logs, alarms, error rates and other metrics are stored in Amazon CloudWatch and are available near real time.

Throttling and Caching

Q: How can I protect my backend systems and applications from traffic spikes?

Amazon API Gateway provides throttling at multiple levels including global and by service call. Throttling limits can be set for standard rates and bursts. For example, API owners can set a rate limit of 1,000 requests per second for a specific method in their REST APIs, and also configure Amazon API Gateway to handle a burst of 2,000 requests per second for a few seconds. Amazon API Gateway tracks the number of requests per second. Any requests over the limit will receive a 429 HTTP response. The client SDKs generated by Amazon API Gateway retry calls automatically when met with this response.

Q: Can I throttle individual developers calling my APIs?
Yes. With usage plans you can set throttling limits for individual API keys.

Q: How does throttling help me?

Throttling ensures that API traffic is controlled to help your backend services maintain performance and availability.

Q: At which levels can Amazon API Gateway throttle inbound API traffic?

Throttling rate limits can be set at the method level. You can edit the throttling limits in your method settings through the Amazon API Gateway APIs or in the Amazon API Gateway console.

Q: How are throttling rules applied?
First. API Gateway checks against your AWS account limit. If the traffic is below the set account limit, API Gateway checks the limit you have set on a stage or method. If the traffic is below the stage limit, then API Gateway applies the usage plans limits you set on a per-API key basis.

Q: Does Amazon API Gateway provide API result caching?

Yes. You can add caching to API calls by provisioning an Amazon API Gateway cache and specifying its size in gigabytes. The cache is provisioned for a specific stage of your APIs. This improves performance and reduces the traffic sent to your back end. Cache settings allow you to control the way the cache key is built and the time-to-live (TTL) of the data stored for each method. Amazon API Gateway also exposes management APIs that help you invalidate the cache for each stage.

Q: What happens if a large number of end users try to invoke my API simultaneously?

If caching is not enabled and throttling limits have not been applied, then all requests will pass through to your backend service until the account level throttling limits are reached. If throttling limits are in place, then Amazon API Gateway will shed the necessary amount of requests and send only the defined limit to your back-end service. If a cache is configured, then Amazon API

Gateway will return a cached response for duplicate requests for a customizable time, but only if under configured throttling limits. This balance between the backend and client ensures optimal performance of the APIs for the applications that it supports. Requests that are throttled will be automatically retried by the client-side SDKs generated by Amazon API Gateway. By default, Amazon API Gateway does not set any cache on your API methods.

Q: How do APIs scale?

Amazon API Gateway acts as a proxy to the backend operations that you have configured. Amazon API Gateway will automatically scale to handle the amount of traffic your API receives. Amazon API Gateway does not arbitrarily limit or throttle invocations to your backend operations and all requests that are not intercepted by throttling and caching settings in the Amazon API Gateway console are sent to your backend operations.

Billing

Q: How am I charged for using Amazon API Gateway?

Amazon API Gateway rates are $3.50 per million API calls, plus the cost of data transfer out, in gigabytes. If you choose to provision a cache for your API, hourly rates apply. Please see the API Gateway Pricing pages for details on data transfer and caching costs.

Q: Who pays for Amazon API Gateway API calls generated by third-party developers?

The API owner is charged for the calls to their APIs on API Gateway.

Q: If an API response is served by cached data, is it still considered an API call for billing purposes?

Yes. API calls are counted equally for billing purposes whether the response is handled by your backend operations or the Amazon API Gateway caching operation.

Amazon AppStream FAQ

General

Q: What is Amazon AppStream?

Amazon AppStream enables you to stream your existing Windows applications from the cloud, reaching more users on more devices, without code modifications. With Amazon AppStream, your application will be deployed and rendered on AWS infrastructure and the output is streamed to mass-market devices, such as personal computers, tablets, and mobile phones. Because your application is running in the cloud, it can scale to handle vast computational and storage needs, regardless of the devices your customers are using. Amazon AppStream provides an

SDK for streaming your application from the cloud. You can integrate your own custom clients, subscriptions, identity, and storage solution with AppStream to build a custom streaming solution that meets the needs of your business.

Q: What are the benefits of streaming over rendering content locally?

Interactively streaming your application from the cloud provides several benefits:

Remove Device Constraints: You can leverage the compute power of AWS to deliver experiences that wouldn’t normally be possible due to the GPU, CPU, memory or physical storage constraints of local devices.

Multi-Platform Support: You can write your application once and stream it to multiple device platforms. To support a new device, just write a small client to connect to your application.

Easy Updates: Because your application is centrally managed by Amazon AppStream, updating your application is as simple as providing a new version of your application to Amazon AppStream. That's all you need to do to immediately upgrade all of your customers without any action on their part.

Instant On: Streaming your application with AppStream lets your customers start using your application or game immediately, without the delays associated with large file downloads and time-consuming installations.

Improved Security: Unlike traditional boxed software and digital downloads, where your application is available for theft or reverse engineering, Amazon AppStream stores your application binary securely in AWS datacenters.

Q: Do some applications work better with Amazon AppStream than others?

Many types of applications work well as streaming applications: CAD, 3D modeling, simulation, games, video and photo-editing software, medical imaging, and life sciences applications. These applications benefit most from streaming because the application runs on the vast computational resources of AWS, yet your customers can interact with the application using low-powered devices, with very little noticeable change in application performance.

You can also consider a hybrid scenario, in which you stream part of an application from Amazon AppStream and host part of the application natively on the device. For example, a game could stream a portion of the visuals, such as rendering a detailed background and render character animations natively on the device.

Applications that have extremely low tolerance for latency are not recommended for streaming. Examples include first person shooters or player vs. player fighting games.

Any application that can benefit from the additional CPU, GPU, memory, or storage available in AWS can use Amazon AppStream. You can stream an existing application or game, invent new types of applications for Amazon AppStream, or build an ecosystem of streamed applications. You should design your input and response model around the expected latency. All platforms, whether console or PC, experience unexpected latency in some of their components, and creating applications for Amazon AppStream is no different. You can read more about AppStream use cases here.

Q: Can my Amazon AppStream applications run offline?

No. Amazon AppStream requires a sustained Internet connection to stream your application. You can, however, provide your customers an off-line experience by hosting part of your application on Amazon AppStream and running part of your application natively on the device. When your customers are connected to the Internet, your application can benefit from the additional resources AppStream offers.

Q: What are the minimum bandwidth requirements for AppStream?

A minimum of 3 Mbps is the minimum recommended bandwidth for running streaming applications. The STX protocol has been optimized to deliver high-quality streaming sessions in congested network environments.

Getting Started

Q: How do I get started with Amazon AppStream?

You can begin using AppStream by visiting the AWS Management console and following the simple instructions.

Q: What does Amazon AppStream manage on my behalf?

Amazon AppStream launches and manages AWS resources to host your application, deploys your application on those resources, scales your application to meet client demand, and dynamically adjusts the video that is streamed to a client to match the network conditions experienced by that client.

Q: How is streamed video quality managed?

Amazon AppStream's STX Protocol measures and adapts to changing network conditions to maintain a fluid experience. If the network conditions are such that Amazon AppStream is unable to deliver a quality video stream, the protocol will inform your client application so you can respond with the best experience for your customers.

Q: How does the Amazon AppStream Host capture the video stream?

When you deploy your application through the AWS Management Console, you will be prompted to provide the file that launches the application. AppStream will use the launcher file to automatically capture the video stream from the corresponding application. Alternatively, you can modify your application to render video to APIs provided by the Amazon AppStream Application and Client SDKs instead of a local render target. For more information, see Stream Video to a Client in the Amazon AppStream Developer Guide.

Q: How does the Amazon AppStream server capture Audio?

Amazon AppStream provides two ways to stream audio: you can modify your application to send audio to APIs provided by the Amazon AppStream Application and Client SDKs or you can configure Amazon AppStream to automatically capture system audio. For more information, see Stream Audio to a Client in the Amazon AppStream Developer Guide.

Q: How are user inputs captured, mapped and sent to the Amazon AppStream application?

AppStream example clients provide default mappings for mouse, keyboard and touch events that can be used to interactive with your streaming application. Your application should be capable of processing these events through the standard Windows message pump.

You can also use APIs provided by the Amazon AppStream Application and Client SDKs from your client application to stream user input back to your application hosted on Amazon AppStream. Amazon AppStream also includes support for HID-compliant controllers such as the Xbox 360 controller for Microsoft Windows and MOGA Android Game Controller. The Amazon AppStream Application and Client SDKs also provide an API to capture raw input, which you can use to handle other types of input such as accelerometer or global positioning system (GPS) data. For more information, see Build a Client Application and Receive Raw User Input From a Client in the Amazon AppStream Developer Guide.

Q: What else do I need to consider when building my Application for Amazon AppStream?

Because your application is no longer hosted on the client device, there are several things to consider:

Persistent storage: Your streaming application is hosted on a virtual server that does not persist data after the client session ends. It is your responsibility to save any state information your application needs to save between sessions. This information can be persisted to external stores such as Amazon S3 or Amazon DynamoDB. Traditionally log collection is difficult to gather from native client applications, but since your application runs in AWS you can collect these directly and learn about your customers' experiences more easily.

Hybrid applications: You need to consider whether you will you stream the entire application, or run part of it natively on the client device.

User authorization: You need to build and manage an authentication, authorization, and entitlements system for your users.

Network availability: You need to think about how you want your client application to handle situations where a network connection is degraded or unavailable. There are many ways to handle this, but the best mechanism will depend on your use case and your customers.

For more information, see the Amazon AppStream Developer's Guide.

Q: While streaming my application using AppStream, can I log application performance data?

You can log performance data for your streaming application using the AWS Management Console. With AppStream tools, you can collect different logs pertaining to your application streaming session and submit a zip file containing these logs to any specified S3 bucket in your account. The log types are (i) custom logs generated by your application (ii) standard streams (stdin, stdout) (iii) Windows mini dumps in case of application crashes (iv) utilization logs that specify CPU, Memory, and Disk Consumption. Logs will be available in your specified S3 bucket after the end of a streaming session. Learn more about setting up log collection for your application.

Platform Support

Q: What client operating systems are supported?

Amazon AppStream allows you to build client applications for Kindle FireOS, Android 2.3 (Gingerbread) and higher, Chrome version 37 and higher (for Chrome desktop browsers and Chromebooks), iOS 7.0 and higher, Mac OS X Mountain Lion (10.8.5) or higher, and Windows 7 and higher. For more information on building clients for these operating systems, see Build a Client Application in the Amazon AppStream Developer Guide.

Q: What server operating system is supported?

Amazon AppStream supports streaming applications developed for Windows Server 2008 R2. Windows Server 2008 is a 64-bit operating system. You can add support for 32-bit applications by using the WoW64 extensions. If your application has other dependencies, such as the .NET framework, include those dependencies in your application installer. For more information on building streaming applications for Amazon AppStream, see Build an Amazon AppStream Streaming Application in the Amazon AppStream Developer Guide.

Q: What regions will my application be streamed from?

AppStream is available in US East (N.Virginia) and Asia Pacific (Tokyo). AppStream will stream the application from the region in which it has been deployed to.

Q: What EC2 instance types does Amazon AppStream support?

Amazon AppStream is available on the g2.2xlarge instance type. Additional instance types will be added in later releases. Please send us feedback to suggest new instance types for Amazon AppStream. For more information about EC2 instance types, see Instance Types in the Amazon Elastic Compute Cloud User Guide.

Q. Does use of AppStream require third-party licenses?

AppStream utilizes the H.264/AVC video format for encoding streamed video, and the open-source Opus audio format for encoding streamed audio. The operation of your content, including the transmission of internet video of your content and your distribution of any associated video decoder in your client application, may require that you obtain license rights from third parties. In addition, AppStream components, including client binary components, include certain open source packages which carry attribution requirements. For more information, see Build an Amazon AppStream Application in the Amazon AppStream Developer Guide.

Scalability, Capacity Management and Monitoring

Q: How does Amazon AppStream scale?

Amazon AppStream uses Amazon Elastic Cloud Compute (EC2) and Auto Scaling to launch virtual servers running your application and to adjust the number of servers to match the demand for client connections. Each client session runs on a separate virtual server. You can use Amazon AppStream to specify capacity needs, and then the service automatically scales your streamed application and connects customers' devices to it.

Q: How do I monitor my application using Amazon AppStream?

The Amazon AppStream console displays information about your application's current capacity, such as the number of current client sessions and the number of unused application servers available for connections.

Q: Can I use other AWS services from my Amazon AppStream application?

Yes. You can call other AWS services from your application by using the AWS SDK or by sending HTTP query requests.

Streaming

Q: What is the Amazon AppStream STX Protocol?

The Amazon AppStream STX Protocol is a proprietary protocol used to stream high quality application video over varying network conditions. It streams video and audio encoded using standard H.264 encoding over the UDP transport layer. The protocol also captures user input and sends it over TCP back to the application being streamed from the cloud. Network conditions are constantly measured during this process and information is sent back to the encoder on the server. The server dynamically responds by altering the video and audio encoding in real-time to produce a high quality stream for a wide variety of applications and network conditions.

Unlike other remote-access services which provide access to the operating system, an end user uses your client software to connect directly to the AppStream application. This cuts latency and improves performance.

Q: What is latency and how does it affect my application?

Latency is the time it takes your application to react to user input, such as a mouse click, key press or touch gesture. The ability of humans to perceive latency varies by individual and by the type of application they are using. For example, applications that have few user interactions or which do not redraw often can tolerate more latency than applications that continuously respond to user interaction, such as a driving game. When you host your application on Amazon AppStream, the service adds a small amount of latency as part of the process to capture and encode the video stream. This process is highly optimized to minimize this latency. In addition, the Amazon AppStream STX Protocol and encoding process has been tuned to minimize latency over a variety of network conditions.

Any application running over the Internet is subject to latency introduced by network conditions: the distance between Internet servers, local network configurations, switch hardware, and the connection technology (wired, wireless, or 4G cellular). To mitigate the effects of latency introduced by network conditions, the Amazon AppStream STX Protocol dynamically adjusts the data streamed to a client to match the network conditions experienced by that client.

Latency is also determined by the region from which your application is streamed. You can provide the best experience for your users by deploying the application in a region closest to them. Know more about the region in which AppStream is available. Another source of latency is the client device. This latency depends on factors such as whether video is rendered by hardware or software and the latency of the input device (keyboard, mouse, or touchscreen.)

The total latency experienced by your customers depends on all of these factors. For many applications, the overall latency of an application streamed over Amazon AppStream is low enough to provide a great customer experience.

The Amazon AppStream team is constantly evaluating ways to reduce total latency and improve the performance of applications hosted on Amazon AppStream.

Q: How is the quality of streaming sessions measured?

There are four main variables to consider when measuring the streaming quality for your end user: (1) the video fidelity, (2) the audio fidelity, (3) the round-trip latency between user inputs and rendered outputs, and (4) the synchronization between audio and video (A/V sync). Each of these quality variables are affected by conditions of the streaming ecosystem.

For example, the video quality is affected by frame resolution (the number of pixels rendered on the client) and frame rate (the number of times per second that a new frame is rendered on the client), while audio quality is affected by the digital sampling rate and the number of audio channels. Both video and audio quality are affected by digital compression (the format and the amount) and the data packet loss.

In order to monitor the streaming quality of each AppStream session, and provide feedback to your users about the quality, AppStream provides QoS (quality of service) APIs to monitor the PSNR (peak signal-to-noise ratio) of the video and streaming latency.

Security

Q: How do I authenticate users to my Amazon AppStream application?

You control who has access to your streaming application by building an entitlement service. Your entitlement service should authenticate users, either with custom logic or by using a service such as Login with Amazon. After your entitlement service has authenticated a user, it calls into Amazon AppStream to entitle a new client session. The entitlement service then returns an entitlement URL to the client that the client uses to connect to your application.

The Amazon AppStream SDK provides a sample implementation of an entitlement service and a CloudFormation template that you can use to deploy the service on AWS. For more information about building an entitlement service, see Build an Entitlement Service in the Amazon AppStream Developer Guide.

Q: Who can access the management console for my Amazon AppStream application?

You can use Amazon Identity and Access Management (IAM) to add users to your AWS account and grant them access to view and manage your Amazon AppStream application. For more information about using IAM, see What is IAM? in AWS Identity and Access Management.

Q: Where can I find more information about security and running applications on AWS?

For more information about securing AWS resources, see theAmazon Web Services: Overview of Security Processes whitepaper and the AWS Security Center.

Q: How is the data from my streamed application encrypted to the client?

The streamed video and user inputs are not encrypted at this time. You can encrypt communication to and from the entitlement service.

Billing

Q: How much does Amazon AppStream cost?

You will be charged for each hour that your customers stream your application, with no long-term commitments. This frees you from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs. Your costs include all of the compute resources needed to run your streaming application, applicable Windows license costs, and bandwidth used by the stream.

Usage time is calculated at the end of each individual streaming session run by your customers. You are billed each month for the total usage time of all streaming sessions combined (rounded to the second).

For example, if one user streams a session for 45 minutes and 30 seconds, and another user streams a session for 120 minutes and 20 seconds, the total amount billed will be for 165 minutes and 50 seconds, which is equivalent to 2.764 hours. At $0.830/hr, these two sessions will incur a charge of $2.29. View AppStream pricing.

Amazon CloudSearch FAQ

General

Q: What is Amazon CloudSearch?

Amazon CloudSearch is a fully-managed service in the AWS Cloud that makes it easy to set up, manage, and scale a search solution for your website or application.

Q: What are the benefits of running a managed search service like Amazon CloudSearch over running my own search service on EC2?

Amazon CloudSearch provides several benefits over running your own self-managed search service including easy configuration, auto scaling for data and traffic, self-healing clusters, and high availability with Multi-AZ. With a few clicks in the AWS Management Console, you can create a search domain and upload the data you want to make searchable, and Amazon CloudSearch automatically provisions the required resources and deploys a highly tuned search index.

Q: What is a search engine?

A search engine makes it possible to search large collections of mostly textual data items (called documents) to quickly find the best matching results. Search requests are usually a few words of unstructured text, such as "matt damon movies". The returned results are usually ranked with the best matching, or most relevant, items listed first (the ones that are most "about" the search words).

Documents may be completely unstructured, or they can contain multiple fields that can optionally be searched individually. For example, a search service for movies might have documents with fields for title, director, actor, description, and reviews. Results returned by a search engine are typically proxies for the underlying documents, such as URLs that reference particular web pages. However, the search service can also return the actual contents of individual fields.

Q: What benefits does Amazon CloudSearch offer?

Amazon CloudSearch is a fully managed search service that automatically scales with the volume of data and complexity of search requests to deliver fast and accurate results. Amazon CloudSearch lets customers add search capability without needing to manage hosts, traffic and data scaling, redundancy, or software packages. Users pay low hourly rates only for the resources consumed. Amazon CloudSearch can offer significantly lower total cost of ownership compared to operating and managing your own search environment.

Q: Can Amazon CloudSearch be used with a storage service?

A search service and a storage service are complementary. A search service requires that your documents already be stored somewhere, whether it's in files of a file system, data in Amazon S3, or records in an Amazon DynamoDB or Amazon RDS instance. The search service is a rapid retrieval system that makes those items searchable with sub-second latencies through a process called indexing.

Q: Can Amazon CloudSearch be used with a database?

Search engines and databases are not mutually exclusive - in fact, they are often used together. If you already have a database that contains structured data, you might want to use a search engine to intelligently filter and rank the database contents using search keywords as relevance criteria.

A search service can be used to index and search both structured and unstructured data. Content can come from multiple sources and can include database fields along with files in a variety of formats, web pages, and so on. A search service can support customizable result ranking as well as special search features such as using facets for filtering that are not available in databases.

Q: What regions is Amazon CloudSearch available in?

Amazon CloudSearch is available in the following AWS regions: US East (Northern Virgina), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), South America (Sao Paulo) and Asia Pacific (Singapore, Tokyo, Sydney).

About the 2013-01-01 API

Q: What new features does Amazon CloudSearch support?

With this latest release Amazon CloudSearch supports several new search and administration features. The key new features include:

- Language support:
  - 34 languages, plus "multiple" to handle mixed language fields
  - Per-field language configuration
  - Language-specific text analysis
  - Multiple levels of algorithmic stemming are available for many languages, including "none"
- Enhanced search features:
  - Suggestions
  - Highlighting
  - Geospatial search
  - New data types: date, double, 64 bit signed int, latlon
  - Sloppy phrase search
  - Term boosting
  - Enhanced range searching for all field types
  - Support for multiple query parsers: simple, structured, lucene, dismax
  - Query parser configuration options
- Administration features:
  - High availability option
  - IAM integration
  - User configurable scaling
- Available in additional AWS Regions: Asia Pacific (Tokyo), Asia Pacific (Sydney), and South America (Sao Paulo)

Q: Does Amazon CloudSearch still support dictionary stemming?

Yes. The new version of Amazon CloudSearch supports dictionary stemming in addition to algorithmic stemming.

Q: Does the new version of Amazon CloudSearch use Apache Solr?

Yes. The latest version of Amazon CloudSearch has been modified to use Apache Solr as the underlying text search engine. Amazon CloudSearch now provides several popular search engine features available with Apache Solr in addition to the managed search service experience that makes it easy to set up, operate, and scale a search domain.

Q: Can I access the new version of Amazon CloudSearch through the console?

Yes. You can access the new version of Amazon CloudSearch through the console. If you are a current Amazon CloudSearch customer with existing search domains, you have the option to select which version of Amazon CloudSearch you want to use when creating new search domains. New customers will use the new version of Amazon CloudSearch by default and will not have access to the 2011-01-01 version.

Q: What data types does the new version of Amazon CloudSearch support?

Amazon CloudSearch supports two types of text fields, text and literal. Text fields are processed according to the language configured for the field to determine individual words that can serve as matches for queries. Literal fields are not processed and must match exactly, including case. CloudSearch also supports four numeric types: int, double, date, and latlon. Int fields hold 64-bit, signed integer values. Double fields hold double-width floating point values. Date fields hold dates specified in UTC (Coordinated Universal Time) according to IETF RFC3339: yyyy-mm-ddT00:00:00Z. Latlon fields contain a location stored as a latitude and longitude value pair.

Q: Will my existing search domains created with the 2011-02-01 version of Amazon CloudSearch continue to work?

Yes. Existing search domains created with the 2011-02-01 version of Amazon CloudSearch will continue to work.

Q: Will I be able to use the new features on my existing search domains created with the 2011-01-01 version of Amazon CloudSearch?

No. Existing search domains created with the 2011-01-01 version of Amazon CloudSearch will not have access to the features available in the new version. To access the new features you will have to create a new search domain using the 2013-01-01 version of Amazon CloudSearch.

Q: How can I migrate my applications built using the 2011-01-01 version of Amazon CloudSearch to the new version of Amazon CloudSearch?

To use the new version of Amazon CloudSearch you need to recreate existing domains using the new version of Amazon CloudSearch and re-upload your data. For more information, see Migrating to the 2013-01-01 API in the Amazon CloudSearch Developer Guide.

Q: Will AWS continue to support the 2011-02-01 version of Amazon CloudSearch?

Yes. AWS will continue support for the 2011-02-01 version of Amazon CloudSearch.

Q: Can I create new search domains using the 2011-02-01 version of Amazon CloudSearch?

Current Amazon CloudSearch customers who have existing 2011-02-01 domains will be able to choose whether their new domains use the 2011-02-01 API or the new 2013-01-01 API. Search domains created by new customers will automatically be created with the 2013-01-01 API.

Q: Can I take advantage of the free trial offer with the new version of Amazon CloudSearch?

New customers will still be able to take advantage of the free trial offer available with Amazon CloudSearch. See the Amazon CloudSearch Free Trial page for details.

Getting Started

Q: How do I get started with Amazon CloudSearch?

To sign up for Amazon CloudSearch, click the **Create Free Account** button on the Amazon CloudSearch detail page and complete the sign-up process. You must have an Amazon Web Services account. If you do not already have one, you will be prompted to create an AWS account when you begin the Amazon CloudSearch sign-up process.

After you have signed up, select **Amazon CloudSearch** from the AWS Management Console. Using the Amazon CloudSearch console you can quickly create a search domain, configure your search fields, upload sample data, and send search queries to your search domain. You can also use the AWS SDKs and the CLI to perform these operations.

For more information, see the Getting Started tutorial in the Amazon CloudSearch Developer Guide.

Q: Do the AWS SDKs support Amazon CloudSearch?

Yes, the AWS SDKs for Java, Ruby, Python, .Net, PHP, and Node.js provide support for CloudSearch. Using the AWS SDKs you can quickly create a search domain, configure your search fields, upload data, and send search queries to your search domain.

Q: Does the AWS CLI support Amazon CloudSearch?

Yes, the AWS CLI provides support for CloudSearch. Using the AWS CLI you can quickly create a search domain, configure your search fields, upload data, and send search queries to your search domain.

Q: Can I still use the Amazon CloudSearch CLTs?

Yes, the Amazon CloudSearch CLTs will continue to work.

Search Domains, Data, and Indexing

Q: What is a search domain and how do I create one?

A search domain is a data container and a set of services that make the data searchable. These services include:

- A document service that allows you upload data to your domain for indexing.
- A search service that allows you to perform search requests against your indexed data.
- A configuration service for controlling your domain's behavior (including relevance ranking).

You can create, manage, and delete search domains using the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do I upload documents to my search domain?

You upload documents to your domain using the AWS Management Console, AWS SDKs, or AWS CLI.

Q: Do my documents need to be in a particular format?

To make your data searchable, you need to format your data in JSON or XML. Each item that you want to be able to receive as a search result is represented as a document. Every document has a unique document ID and one or more fields that contain the data that you want to search and return in results. Amazon CloudSearch generates a search index from your document data according to the index fields configured for the domain. As your data changes, you submit updates to add or delete documents from your index.

Q: How do I create document batches formatted for Amazon CloudSearch?

To create document batches that describe your data, you create JSON or XML text files that specify:

- The operation type: add or delete
- A unique identifier
- The actual fields and their data

The following example shows a single document batch formatted in JSON:

```
[
    {

       "fields" : {

         "directors" : [

             "Francis Lawrence"

          ],

          "release_date" : "2013-11-11T00:00:00Z",

          "genres" : [

             "Action",

             "Adventure",

             "Sci-Fi",

             "Thriller"

          ],

          "image_url" : "http://ia.media-imdb.com/images/M/MV5xMzzA
x._V1_SX400_.jpg",

          "plot" : "Katniss Everdeen and Peeta Mellark become targe
ts of the Capitol after their victory in the 74th Hunger Games spar
ks a rebellion in the Districts of Panem.",

          "title" : "The Hunger Games: Catching Fire",

          "rank" : 4,

          "running_time_secs" : 8760,

          "actors" : [

             "Jennifer Lawrence",
             "Josh Hutcherson",

             "Liam Hemsworth"

           ],

          "year" : 2013

       },

       "id" : "tt1951264",

       "type" : "add"

    }

]
```

Note that numeric values such as the year are not enclosed in quotes, and that values in a multi-value field such as genres are listed in a JSON array.

To make this data available to Amazon CloudSearch, you can save it to a file and upload it using the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do my documents get indexed?

Documents are automatically indexed when you upload them to your search domain. You can also explicitly re-index your documents when you make configuration changes by sending an IndexDocuments request.

Q: When do I need to re-index my domain?

Certain configuration options, such as adding a new index field or updating your stemming or stopword dictionaries, are not available until your domain is re-indexed. When you have made changes that require indexing, the domain’s status will indicate that it needs to be indexed. You can initiate indexing from the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do I send search requests to my search domain?

Every search domain has a REST-based search service with a unique URL (search endpoint) that accepts search requests for its document set. You can send search requests from the AWS

Management Console, AWS SDKs, or AWS CLI.

Q: Can a search domain span multiple Availability Zones?

Yes. If you enable the Multi-AZ option, Amazon CloudSearch deploys additional instances in a second availability zone in the same Region. For more information, see Configuring Availability Options in the Amazon CloudSearch Developer Guide.

Q: Can I move a search domain from one region to another?

At this time, there is no way to automatically migrate a search domain from one region to another. You will need to create a new domain in the target region, configure the domain and upload your data, then delete the original domain.

Q: How do I delete my search domain?

To delete a search domain, click on Delete Domain button in the Amazon CloudSearch console. You can also delete domains through the AWS SDKs or AWS CLI.

Q: How do I delete documents from my search domain?

To delete documents you specify a delete operation in your batch upload that contains the ID of the document you want to remove.

You can submit data updates through the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do I empty my search domain?

If you wish to maintain your domain’s endpoints, you can send a delete for each document that is in your domain.

Q: Why is my domain in the "Processing" state?

A domain can be in one of three different states: “processing,” “active,” or “reindexing.” Normally, your domain will be in the “active” state, which indicates that no changes are currently being made, that the domain can be queried and updated, and that all previous changes are currently visible in the search results.

When a domain needs to be re-indexed, Amazon CloudSearch needs to rebuild the index entirely. However, the domain does not enter the “processing” state until you initiate reindexing. During this stage, the domain can still be queried and updated, but the configuration changes won't be visible in search results until indexing is completed, and the domain's status changes back to “active.”

You can also continue to upload document batches to your domain. However, if you submit a large volume of updates while your domain is in the “processing” state, it can increase the amount of time it takes for the updates to be applied to your search index. If this becomes an issue, slow down your update rate until the domain returns to the “active” state.

Best Practices

Q: What are the best practices for bootstrapping data into CloudSearch?

After you’ve launched your domain, the next step is loading your data into Amazon CloudSearch. You’ll likely need to upload a single large dataset, and then make smaller updates or additions as new data comes in. The following guidelines will help make bootstrapping your initial data into CloudSearch quick and easy.

1. Use the curl-v command line tool when preparing your script

During the upload of a dataset, the script you’ve written reads your data and uses it to create JSON or XML documents. We recommend preparing this script in advance, and using curl or another simple command line tool to see if you’re able to upload the documents that the script creates. The “-v” option in curl often provides more detailed information about syntax problems than the AWS SDK or Boto, which both suppress errors for production purposes. Curl displays more detailed error messages, which helps identify the sources of any issues.

2. Use the UTF-8 character code

Make sure that all data is formatted in the UTF-8 character code format, and that any bad Unicode characters have been removed before uploading to CloudSearch. Illegal characters will cause the document upload to fail.

3. Batch your documents

Batching your documents is perhaps the most important step in data bootstrapping. Submitting documents to CloudSearch individually is not only inefficient, but also leads to preventable errors.

A document batch is simply a collection of add and delete operations that represent the documents you want to add, update, or delete from your domain. Batches are described in either JSON or XML, and when you upload them to a domain, the data is indexed automatically, according to the domain's indexing options. Since you’re billed for the total number of document batches uploaded to your search domain, it’s more cost-effective to upload your data in batches of 5 MB, the maximum allowed per upload. You can also upload batches in parallel to reduce the amount of time it takes to upload your data.

4. Pre-scale

It’s also important to pre-scale your data before uploading it to CloudSearch. Pre-scaling involves selecting the appropriate instance type for the amount of data you wish to upload.

Choosing an instance with enough capacity to handle the size of your upload can help prevent errors and a high replication count. Although replication can help decrease search response time, it doesn’t increase the size of the data pipe or address core problems in data uploads.

CloudSearch will automatically scale up to larger instances as you send more data. Still, pre-selecting the appropriate instance type saves time later in the bootstrapping process, as scaling from one instance to another tends to be a slower process. Below is a sample script to pre-scale the domain for boostrapping and to restore the instance type after data is loaded.

Pre-scale before bootstrapping:

```
aws cloudsearch update-scaling-parameters --domain-name foo --scaling-parameters DesiredInstanceType=search.m3.2xlarge

aws cloudsearch index-documents --domain-name foo
```

Restore after data loading:

```
aws cloudsearch update-scaling-parameters --domain-name foo --scaling-parameters DesiredInstanceType=search.m1.small

aws cloudsearch index-documents --domain-name foo
```

Q: What are some ways to avoid 504 errors?

If you’re seeing 504 errors or high replication counts, try moving to larger instance type. For example, if you’re having problems with m3.large, move up to m3.xlarge. If you continue to get 504 errors even after pre-scaling, start batching the data and increase the delay between retries.

Q: What are the best practices to accelerate domain configuration and re-indexing?

When you change the configuration options of your search domain, you must rebuild your search index for those changes to take effect in search results. Rebuilding the index can take 30 to 60 minutes whether you make one configuration change at a time or several configuration changes at once. Even if your domain has only a small number of documents, re-indexing takes this time because of the processing and provisioning necessary to build the index and distribute it. Therefore, you should plan your configuration changes ahead of time, make all of your changes at once, and then re-index your domain. The same applies when setting up a new domain - plan your configuration before you set it up so that you can index only once and get up and running in the shortest time possible.

Some domain changes require re-indexing while others just require re-deploying the existing index. Redeploying the domain takes 10 to 15 minutes compared to 30-60 minutes for re-indexing. During re-deployment, CloudSearch creates new nodes, deploys the index on them, and shuts down the old nodes. Your domain status changes to “Processing” during re-deployment. When re-indexing is needed, your domain status changes to “Needs Indexing,” followed by “Processing” once you have initiated indexing. Once the new index is created, your domain is re-deployed. The following table summarizes which changes require re-indexing followed by re-deployment and which changes require just re-deployment. Understanding this will help you better plan your configuration changes.

Change	Needs re-indexing	Needs re-deployment
Multi-AZ	No	Yes
Index fields	Yes	Yes
Index field options	Yes	Yes
Instance type	Yes	Yes
Partition count	Yes	Yes
Replication count	No	Yes
Suggesters	Yes	Yes
Expressions	No	Yes
Analysis schemes	Yes	Yes

Search Features

Q: What search features does Amazon CloudSearch provide?

Amazon CloudSearch provides features to index and search both structured data and plain text, including faceted search, free text search, Boolean search expressions, customizable relevance ranking, query time rank expressions, field weighting, searching and sorting of results using any field, and text processing options including tokenization, stopwords, stemming and synonyms. It also provides near real-time indexing for document updates. New features include:

- Autocomplete suggestions
- Highlighting
- Geospatial search
- New data types: date, double, 64 bit signed int, LatLon
- Dynamic fields
- Index field statistics
- Sloppy phrase search
- Term boosting
- Enhanced range searching for all field types
- Search filters that don't affect relevance
- Support for multiple query parsers: simple, structured, lucene, dismax
- Query parser configuration options

Q: What is faceting?

Faceting allows you to categorize your search results into refinements on which the user can further search. For example, a user might search for "umbrellas", and facets allow you to group the results by price, such as $0-$10, $10-$20, $20-$40, and so on. Amazon CloudSearch also allows for result counts to be included in facets, so that each refinement has a count of the number of documents in that group. The example could then be: $0-$10 (4 items), $10-$20 (123 items), $20-$40 (57 items), and so on.

Q: What languages does Amazon CloudSearch support?

Amazon CloudSearch currently supports 34 languages: Arabic (ar), Armenian (hy), Basque (eu), Bulgarian (bg), Catalan (ca), simplified Chinese (zh-Simp), traditional Chinese (zh-Trad), Czech (cs), Danish (da), Dutch (nl), English (en), Finnish (fi), French (fr), Galician (gl), German (de), Greek (el), Hebrew (he), Hindi (hi), Hungarian (hu), Indonesian (id), Irish (ga), Italian (it), Japanese (ja), Korean (ko), Latvian (la), Norwegian (no), Persian (fa), Portuguese (pt), Romanian (ro), Russian (ru), Spanish (es), Swedish (sv), Thai (th), and Turkish (tr). In addition, Amazon CloudSearch supports a Multiple (mul) option for fields that contain mixed languages.

Q: Does Amazon CloudSearch support geospatial search?

Yes, Amazon CloudSearch has a native type to support latitude and longitude (latlon), so that you can easily implement geographically-based searching and sorting. For more information, see Searching and Ranking Results by Geographic Location in the Amazon CloudSearch Developer Guide.

Performance

Q: How quickly will my uploaded documents become searchable?

Documents uploaded to a search domain typically become searchable within seconds to a few minutes.

Q: How many search requests can I send to my search domain?

There is no intrinsic limit on the number of search requests that can be sent to a search domain.

Q: What factors affect the latency of my search requests?

Your search requests are typically processed within a few hundred milliseconds, frequently much faster. Latency is affected by many factors including the time it takes for your request and responses to travel between your own application and your search domain, the complexity of your search request, and how heavily you are using your search domain.

Q: What makes one search request more complex than another?

Amazon CloudSearch is designed to efficiently process a wide range of search requests very quickly. Search requests vary in complexity depending on the expressions that determine which documents match and additional criteria that determine how closely each document matches. Search requests that match a large number of documents take longer to process than those that match very few documents. Search requests that compute complex expressions take longer to process than those that rank using a simple criteria such as a single field. To help you understand the difference in complexity between Search requests, the time it took to process the request is returned as part of the response.

Q: Where should I run my search application to minimize communication time with my search domain?

Applications hosted in the same AWS Region as your search domain will experience the fastest communication times.

Scaling

Q: What is a search instance?

A search instance is a single search engine in the cloud that indexes documents and responds to search requests. It has a finite amount of RAM and CPU resources for indexing data and processing requests.

Q: What is a search partition?

A search partition is the portion of your data which fits on a single search instance. A search domain can have one or more search partitions, and the number of search partitions can change as your documents are indexed.

Q: How does my search domain scale to meet my application needs?

Search domains scale in two dimensions: data and traffic. As your data volume grows, you need more (or larger) Search instances to contain your indexed data, and your index is partitioned among the search instances. As your request volume or request complexity increases, each Search Partition must be replicated to provide additional CPU for that Search Partition. For example, if your data requires three search partitions, you will have 3 search instances in your search domain. As your traffic increases beyond the capacity of a single search instance, each partition is replicated to provide additional CPU capacity, adding an additional three search instances to your search domain. Further increases in traffic will result in additional replicas, to a maximum of 5, for each search partition.

Q: How much data can I upload to my search domain?

The number of partitions you need depends on your data and configuration, so the maximum data you can upload is the data set that when your search configuration is applied results in 10 search partitions. When you exceed your search partition limit, your domain will stop accepting uploads until you delete documents and re-index your domain. If you need more than 10 search partitions, please contact us.

Q: Do I need to select the number and type of search instances for my search domain?

CloudSearch is a fully managed search service that automatically scales your search domain and selects the number and type of search instances. All search instances in a given search domain are of the same type and this type can change over time as your data or traffic grows.

You can also configure scaling options for an Amazon CloudSearch domain to:

- Increase the upload capacity
- Speed up search requests

- Increase the search capacity
- Improve fault tolerance

Q: What instance types does Amazon CloudSearch support?

Amazon CloudSearch supports the following instance types:

- Small Search Instance
- Large Search Instance
- Extra Large Search Instance
- Double Extra Large Search Instance

Q: How do I find out the number and type of search instances in my search domain?

You can find out the number and type of search instances in your search domain by using the AWS Management Console, AWS SDKs, or AWS CLI. The number and type of search instances change over time and automatically scale up and down according to your indexable data and search traffic.

Q: How quickly does my search domain scale to accommodate changes in data and traffic?

Search domains typically react to increases in traffic changes within minutes. Changes in data volume or a reduction in traffic might take longer but you can accelerate this process by invoking an IndexDocuments operation. If you are about to upload a large amount of data or expect a surge in query traffic, you can prescale your domain by setting the desired instance type and replication count. For more information, see Configuring Scaling Options in the Amazon CloudSearch Developer Guide.

Q: Does Amazon CloudSearch support Multi-AZ deployments?

Yes. Amazon CloudSearch supports Multi-AZ deployments. When you enable the Multi-AZ option, Amazon CloudSearch provisions and maintains extra instances for your search domain in a second Availability Zone to ensure high availability. Updates are automatically applied to the instances in both Availability Zones. Search traffic is distributed across all of the instances and the instances in either zone are capable of handling the full load in the event of a failure.

Q: How does the new Multi-AZ feature work? Will my system experience any downtime in the event of a failure?

When the Multi-AZ option is enabled, Amazon CloudSearch instances in either zone are capable of handling the full load in the event of a failure. If there's service disruption or the instances in one zone become degraded, Amazon CloudSearch routes all traffic to the other Availability Zone. Redundant instances are restored in a separate Availability Zone without any administrative intervention or disruption in service.

Some inflight queries might fail and will need to be retried. Updates sent to the search domain are stored durably and will not be lost in the event of a failure.

Q: Can a search domain be deployed in more than 2 Availability Zones?

No. The maximum number of Availability Zones a domain can be deployed in is two.

Q: Can I modify the Multi-AZ configuration on my search domain?

Yes. You can turn the Multi-AZ configuration on and off for your search domains. The service is not interrupted when this setting is changed.

Q: Can I choose which Availability Zones my search domain is deployed in?

No. At this time Amazon CloudSearch automatically chooses an alternate Availability Zone in the same Region.

Q: Can I choose the instance type my domain uses?

Yes. With the latest release, Amazon CloudSearch enables you to specify the desired instance type for your domain. If necessary, Amazon CloudSearch will scale your domain up to a larger instance type, but will never scale back to a smaller instance type.

Q: What is the fastest way to get my data into CloudSearch?

By default, all domains start out on a small search instance. If you need to upload a large amount of data, you should prescale your domain to a larger instance type. For more information, see Bulk Uploads in the Amazon CloudSearch Developer Guide.

Q: How do I know which instance type I should choose for my initial setup?

For datasets of less than 1 GB of data or fewer than one million 1 KB documents, start with the default settings of a single small search instance. For larger data sets consider pre-warming the domain by setting the desired instance type. For data sets up to 8 GB, start with a large search instance. For datasets between 8 GB and 16 GB, start with an extra large search instance. For datasets between 16 GB and 32 GB, start with a double extra large search instance. Contact us if you need more upload capacity or have more than 500 GB to index.

Security

Q: What additional security features are available with the new version of Amazon CloudSearch?

With the latest release, Amazon CloudSearch now provides IAM integration for the configuration service and all search domain services. You can control access to specific Amazon CloudSearch actions and require request authentication for all requests. Requests are authenticated using Signature Version 4 signing.

Q: How do I upload my data to Amazon CloudSearch securely?

You send us your data using a secure and encrypted SSL connection by using HTTPS instead of HTTP when you connect to Amazon CloudSearch.

Q: My data is already encrypted. Can I just send you the encrypted data and the encryption key?

We do not support user-generated encryption keys. You will need to decrypt the data and upload it using HTTPS.

Q: Do you support encrypted search results?

Yes. We support HTTPS for all Amazon CloudSearch requests.

Q: How can I prevent specific users from accessing my search domain?

Amazon CloudSearch supports IAM integration for the configuration service and all search domain services. You can grant users full access to Amazon CloudSearch, restrict their access to specific domains, and allow or deny access to specific actions.

Pricing

Q: How will I be charged and billed for my use of Amazon CloudSearch?

There are no set-up fees or commitments to begin using the service. Following the end of the month, your credit card will automatically be charged for that month's usage. You can view your charges for the current billing period at any time on the AWS web site by logging into your Amazon Web Services account and clicking **Account Activity** under Your Web Services Account.

Q: How much does it cost to use Amazon CloudSearch?

There are no changes to the pricing structure for Amazon CloudSearch at this time. For detailed pricing information, see Amazon CloudSearch Pricing.

Q: Is a free trial available for Amazon CloudSearch?

Yes, a free trial is available for new CloudSearch customers. For more information, seeAmazon CloudSearch 30 Day Free Trial.

Q: How much does it cost to use the new version of Amazon CloudSearch?

There are no changes to the pricing structure for Amazon CloudSearch at this time. See the Pricing page for more information.

Q: Are there any cost savings to using the new version of Amazon CloudSearch?

The latest version of Amazon CloudSearch features advanced index compression and supports larger indexes on each instance type. This makes the new version of Amazon CloudSearch more efficient than the previous version and can result in significant cost savings.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Amazon Elastic Transcoder FAQ

General

Q: What is Amazon Elastic Transcoder?

Amazon Elastic Transcoder is a highly scalable, easy to use and cost effective way for developers and businesses to convert (or “transcode”) video and audio files from their source format into versions that will playback on devices like smartphones, tablets and PCs.

Q: What can I do with Amazon Elastic Transcoder?

You can use Amazon Elastic Transcoder to convert video and audio files into supported output formats optimized for playback on desktops, mobile devices, tablets, and televisions. In addition to supporting a wide range of input and output formats, resolutions, bitrates, and frame rates, Amazon Elastic Transcoder also offers features for automatic video bit rate optimization, generation of thumbnails, overlay of visual watermarks, caption support, DRM packaging, progressive downloads, encryption and more. For more details, please visit the Product Details page.

Q: Why should I use Amazon Elastic Transcoder?

Amazon Elastic Transcoder manages all the complexity of running media transcoding in the AWS cloud. Amazon Elastic Transcoder enables you to focus on your content, such as the devices you want to support and the quality levels you want to provide, rather than managing the infrastructure and software needed for conversion. Amazon Elastic Transcoder scales to handle the largest encoding jobs. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources that you use. We offer a free tier that enables you to explore the service and transcode up to up to 20 minutes of SD video or 10 minutes of HD video a month free of charge. To see terms and additional information on the free tier program, please visit the AWS Free Usage Tier page.

Q: How do I get started with Amazon Elastic Transcoder?

You can sign up for Amazon Elastic Transcoder through theAWS Management Console. You can then use the console to create a pipeline, set up an IAM role, and create your first transcoding job. To help you test Amazon Elastic Transcoder, the first 20 minutes of SD content (or 10 minutes of HD content) transcoded each month is provided free of charge. Once you exceed the number of minutes in this free usage tier, you will be charged at the prevailing rates. We do not watermark the output content or otherwise limit the functionality of the service, so you can use it and truly get a feel for its capabilities. To see terms and additional information on the free tier program, please visit the AWS Free Usage Tier page. If you do not have an AWS account, you can create one by clicking the Sign Up button at the top of this page.

Q: How do I use Amazon Elastic Transcoder?

To use Amazon Elastic Transcoder you need to have at least one media file in an Amazon S3 bucket. The easiest way to use Amazon Elastic Transcoder is to try it through the console. Create a transcoding pipeline that connects the input Amazon S3 bucket to the output Amazon S3 bucket. Create a transcoding job that will transcode your media file, choose a transcoding preset (a template), and submit the job. Your transcoded file will appear in your output bucket once it has been processed.

Q: What tools and libraries work with Amazon Elastic Transcoder?

Amazon Elastic Transcoder uses a JSON API, and we provide SDKs for Python, Node.js, Java, .NET, PHP, and Ruby. The new AWS Command Line Interface also supports Amazon Elastic Transcoder. You can see a full list of our SDKs here.

Q: Can I use the AWS Management Console with Amazon Elastic Transcoder?

Yes. Amazon Elastic Transcoder has a console that is accessed through the AWS Management Console. You can use our console to create pipelines, jobs, and presets as well as manage and view existing pipelines and jobs.

Q: How do I get my media files into Amazon S3?

There are many ways to get content into Amazon S3, from the simple web-based uploader in the AWS Management Console to programmatic approaches through APIs. For very large files, you may wish to use AWS Import/Export, AWS Direct Connect, or file-acceleration solutions available in the AWS Marketplace. For more information please refer to the Amazon S3 documentation and the AWS Digital Media website.

Q: How do I retrieve my media files from Amazon S3?

You can retrieve files from Amazon S3 programmatically, using the AWS Management Console or a third party tool. You can also mark Amazon S3 objects as public and download them directly from Amazon S3.

Q: Can I use a Content Distribution Network (CDN) to distribute my media files?

Yes. You can easily use CDNs to distribute your content; for example, you can use Amazon CloudFront to distribute your content to end-users with low latency, high data transfer speeds, and no commitments. You can use an output bucket that contains your transcoded content in Amazon S3 as the origin server for Amazon CloudFront. For more information, please visit the detail page for Amazon CloudFront.

Q: How long does it take to transcode a job?

Jobs start processing in the order in which they are received in a pipeline. Once a job is ready to be transcoded, many variables affect the speed of transcoding, for example, the input file size, resolution, and bitrate. For example, if you were to submit a 10 minute video using the iPhone 4 preset, it would take approximately 5 minutes. If a large number of jobs are received they are backlogged (queued). Please note that the transcoding speed may be different between regions.

Q: When will my job be ready?

You can use Amazon SNS notifications to be informed of job status changes. For example, you can be notified when your job starts to transcode and when it has finished transcoding. For more information on Amazon SNS notifications, please see the detail page on Amazon SNS.

Q: How many jobs are processed at once?

Pipelines operate independently from one another. Each pipeline processes jobs in parallel up to a default limit set for that pipeline. Within a job, each individual output also progresses in parallel. For more information on limits and capacity, visit the limits section in the Elastic Transcoder Developer Guide. You can request higher limits by opening a support case.

Q: How many jobs can I submit?

Currently, we allow a maximum of 100,000 jobs per pipeline. Once you exceed this limit, you will receive a 429 Rate Limit Exception. If you require this limit to be raised, please contact us here.

Q: Can I create multiple outputs per job?

Each transcoding job relates to a single input file and can create one or more output files. For example, you may wish to create audio only, low- and high-resolution renditions of the same input file and could do so as part of a single transcoding job. The number of outputs per job is limited. For more information on Amazon Elastic Transcoder limits, please refer to the documentation.

Multiple outputs are charged individually: each output is charged as a separate transcode.

Q: How do I generate clips?

You can create a clip from your source media in your transcoding job. You specify a start time and a duration (both specified as HH:mm:ss.SSS or sssss.SSS.) To cut off the start of a file, you would just specify a start time. You can generate different length clips (or transcode the entire file) for each different output in your transcoding job. You will be charged based on the output duration of your transcode, so if you have a five-minute input file and you create a one-minute output from it, you will only be charged for one minute of transcoding. Please remember that fractional minutes are rounded up, so if you create a clip that is one minute and thirty seconds in duration, you will be charged for two minutes of transcoding.

Q: What is a transcoding pipeline, what can I use it for, and how many can I have?

A pipeline is a queue-like structure that manages your transcoding jobs. A pipeline can process multiple jobs simultaneously, and generally starts to process jobs in the order in which you added them to the pipeline. Jobs often finish in a different order based on job specifications. It is up to you how you wish to use pipelines. Some examples include submitting jobs to different pipelines based on the priority or the duration of a transcode, or using different pipelines for your development, test and production environments. The number of pipelines per AWS account is limited. For more information on Amazon Elastic Transcoder limits, please refer to the documentation.

Q: What are transcoding presets?

A preset is a template that contains the settings that you want Amazon Elastic Transcoder to apply during the transcoding process, for example, the codec and the resolution that you want in the transcoded file. When you create a job, you specify which preset you want to use. We provide presets that create media files that play on any device and presets that target specific devices. For maximum compatibility, choose a “breadth preset” that creates output that plays on a wide range of devices. For optimum quality and file size, choose an “optimized preset” that creates output for a specific device or class of devices.

Q: What do I do if none of your transcoding presets work for me?

You can create your own custom presets based on an existing preset. Once you create your own custom preset, it is available across your AWS account for the Amazon Elastic Transcoder service within a specific region. For more information on presets, please refer to the Amazon Elastic Transcoder Developer Guide. The number of pipelines per AWS account is limited. For more information on Amazon Elastic Transcoder limits, please refer to the documentation.

Q: Why do I need to assign a role to a transcoding pipeline?

Amazon Elastic Transcoder uses AWS Identity and Access Management (IAM) roles to enable you to securely control access to your media assets. The IAM role sets a policy that defines what permissions you have for accessing Amazon S3 resources. You can assign different roles to different pipelines, and an IAM administrator can create specific roles for use with Amazon Elastic Transcoder. More information about IAM can be found here.

Q: How can I configure roles to be more restrictive?

You can use the AWS Management Console to edit and create new IAM roles. IAM roles that are created by Amazon Elastic Transcoder are visible in the AWS Management Console and can also be edited.

Q: How do I use notifications?

Amazon Elastic Transcoder uses Amazon SNS to notify you of specific events. You can choose to be notified about jobs that start to process, jobs that complete, warnings, and errors. Each event type is assigned to an SNS topic, and you can use the same topic or different topics for each event. The Amazon Elastic Transcoder console will create an SNS topic for you or you can specify an existing one.

Q: Why should I use notifications?

Notifications are a much more efficient way to check transcoding status than polling the API. Notifications provide a way to be notified on specific events that occur in the system. For example, you can be notified on a completed event. This is useful if you want to know when a job has finished transcoding and this is far more efficient than calling the 'List Jobs By Status' or 'Read Job' API at regular intervals.

Q: Why does my job keep failing?

The most common reason for jobs to fail is that the input file is corrupted in some way. If you receive an error about the format not being supported, we are unable to decode your source file and we’d love for you to tell us more about on our Discussion Forum. We need the following information to assist with diagnosis: AWS Account ID, Region and Job ID. For a list of error codes, please refer to the documentation.

Q: How can I generate more than one thumbnail per job?

You can specify a thumbnail creation interval in seconds to create one thumbnail every n seconds. To create thumbnails in more than one size, you need to create different jobs.

Q: Can I reserve a transcoder for my exclusive use?

Amazon Elastic Transcoder provides a shared transcoding service and does not enable a transcoder to be reserved or allocated to an individual customer.

Q: Do I need to pay license fees?

We have licensed relevant intellectual property from the applicable patent pools for transcoding content. Like any other transcoder, customers are responsible for evaluating and, if necessary, securing licenses for distribution of content in various formats.

Q: Do you support live encoding?

Amazon Elastic Transcoder is a file-based transcoding service and does not support live transcoding.

Q: Are there limits to the service?

The number of transcoding pipelines, transcoding presets and outputs per job have limits. Most of these limits can be adjusted on a customer-by-customer basis. For the current limits, please refer to the documentation.

Q: How do I increase service limits?

If you require an increase in the service limits, please contact ushere and provide all the information requested on the form. We will then contact you to discuss your requirements.

Q: Where is Amazon Elastic Transcoder available?

Amazon Elastic Transcoder is available in the following AWS regions: US East (N Virginia), US West (Oregon), US West (N California), EU (Ireland), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Sydney), and Asia Pacific (Mumbai).

The service operates standalone in each region, so jobs created in one region may not be transferred to another region.

You can create a transcoding pipeline in one region that would specify Amazon S3 buckets in another region. However, if you choose to do this, you should be aware that you will incur Amazon S3 transfer costs when content is read from or written out to an Amazon S3 bucket in a region other than the one where the transcoding work is taking place.

Q: Can I pass metadata when creating a job?

You have the option to attach up to 10 custom metadata key-value pairs to your Elastic Transcoder jobs. This metadata will be included in the job notifications and when reading the job via the API or console. You provide this information in the “UserMetadata” field on the Job object.

Format Support

Q: What input formats do you support?

We support popular web, consumer and professional media formats. Examples include 3GP, AAC, AVI, FLV, MP4 and MPEG-2. If there is a format that you’ve found does not work, please let us know through our forum.

Q: Where can I find a comprehensive list of support formats?

We add new input formats on an ongoing basis, so such a list would age quickly. Please take advantage of our free tier and console to try a format not mentioned above and if you run into problems, please let us know!

Q: When creating MP4 files, do you support "fast start"?

We locate the MOOV atom for an MP4 at the start of the file so that your player can start playback immediately without waiting for the entire file to finish downloading.

Q: Do you support Apple ProRes or digital cinematography formats?

We do not support reading Apple ProRes files or raw camera formats like ARRI and RED at this time.

Q: What video formats can I transcode into?

We support the following video codecs: H.264, VP9, VP8, MPEG-2, and animated GIF. File formats supported include MPEG-2 TS container (for HLS), fmp4 (for Smooth Streaming and MPEG-DASH), MP4, WebM, FLV, MPG, and MXF (XDCAM-compatible). For information on file formats that are supported by specific codecs, please visit the Product Details page.

Q: What audio formats can I transcode into?

We support the following audio codecs: AAC, MP3, MP2, PCM, FLAC, and Vorbis. Audio-only file formats supported include MP3, MP4, FLAC, OGA, OGG, and WAV. For information on file formats that are supported by specific codecs, please visit the Product Details page.

Q: How is album art supported for audio files?

Album art is supported in MP4 files containing AAC audio, in MP3 files, and in FLAC files. Album art is not supported for OGA, OGG, WAV, WebM or MPEG-2 TS outputs. You can specify whether album art from the source file is passed through to the output, removed, or whether new album art should replace it or be appended to it.

Q: How do I create an audio file from a video file?

To strip out video and create an output that only contains the audio track, run a transcoding job with your input file and use one of the system transcoding presets that contains Audio in its name. Alternatively, you can create your own audio only custom transcoding preset. The output file will only contain the audio portion of the input file.

Q: Do you support surround sound formats?

The audio portion of the transcoded output from Amazon Elastic Transcoder is two-channel AAC, MP3 or Vorbis.

Q: Do you support audio channel remapping?

If the source file contains multi-channel audio, the output will contain the first two channels, which are frequently left and right audio tracks. For the MXF container, we support multiple modes of packaging the audio into the file, including optional insertion of motor only shots (MOS).

Q: Can I generate XDCAM-compatible video?

Yes, the easiest way to generate XDCAM-compatible outputs is to specify one of the XDCAM system presets when creating a transcoding job. You can also create a custom preset by choosing the MXF container with MPEG-2 video and PCM audio.

Q: Do you support closed captions?

Yes, you can add, remove, or preserve captions as you transcode your video from one format to another.

Supported input formats:
Embedded: CEA-608, CEA-708 (MPEG-2 only) and mov-text
Sidecar captions: DFXP, EBU-TT, SCC, SMPT, SRT, TTML, WebVTT

Supported output formats:
Embedded captions: mov-text (MP4), and CEA-708 (MP4 and MPEG-TS)
Sidecar captions: DFXP, EBU-TT, SCC, SMPT, SRT, TTML, and WebVTT

CEA-708 captions are embedded in the H.264 SEI user data of the stream.

Q: Can you support multiple caption tracks?

Yes, you can add one track per language.

Q: How do I create content for HLS output?

There are two steps:

1. Create a transcoding job containing outputs for each variation using one of our supplied system presets or your own, based on the MPEG-2 TS container and H.264 and AAC codecs. The lowest rate stream should be an audio only stream.

2. Specify that the transcoding job create a playlist that references the outputs. You should order your bit rates from lowest to highest, with the audio only stream last, since this order will be maintained in the generated playlist file. Once your transcoding job has completed, the output bucket will contain a proper arrangement of your master and individual M3U8 playlists, and MPEG-2 TS media stream fragments.

Note: When selecting the HLSv4 option, your outputs should be matched to audio-only and video-only presets. For system presets, these can be identified by words "Audio" or "Video" as part of their name. For example, “System preset: HLS Video – 600k,” would match with the HLSv4 option whereas "System preset: HLS – 600k,” would be used with the HLSv3 option.

Q: How do I create content for Smooth Streaming?

There are two steps:

1. Create a transcoding job containing outputs for each variation using one of our supplied system presets or your own, based on the fragmented MP4 container and H.264 and AAC codecs.
2. Specify that the transcoding job create a playlist that references the outputs. Once your transcoding job has completed, the output bucket specified by the transcoding pipeline will contain your manifest ISM file, client ISMC file, and fragmented MP4 media files.

Q: How do I create content for MPEG-DASH streaming?

There are two steps:

1. Create a transcoding job containing the video-only outputs (with the desired resolutions and bitrates) and the audio-only output using either the system presets or your own customized presets, based on the fragmented MP4 container with H.264 video and AAC audio.
2. Create an MPEG-DASH playlist for the transcoding job by selecting MPEG-DASH as the Playlist Format. Specify the outputs that this playlist will reference. Once your transcoding job has completed, the output bucket specified by the transcoding pipeline will contain your manifest MPD file, and the fragmented MP4 media files.

Q: Should I use the HLSv3 or the HLSv4 option?

HLS version 3 has been supported natively on iOS 2+ devices since July 2008 and on Android 4.0+ since Oct. 2011. HLS version 4 has been supported natively on iOS 5+ devices since Oct. 2011 and on Android 4.4+ since Sept. 2013.

If you able to reach your target devices with HLS version 4, you will be able to generate playlists that use byte range requests, late-binding audio, and I-frame only playback. Playlists with byte range requests are able to use just one file per bit rate, eliminating the need to manage thousands of small segment files. Late-binding audio allows the audio to be streamed separately from the video, eliminating redundant audio storage. I-frame only playback enables trick-play modes used to enhance fast-forward, rewind, and seeking through the video.

Q: Can I stream HLS directly from S3?

Yes, you can play your HLS renditions directly from S3 by pointing the player to the M3U8 playlist. We recommend you use a CDN such as Amazon CloudFront, which provides a better end user experience with improved scalability and performance. See Configuring On-Demand Apple HTTP Live Streaming (HLS).

Q: Do I need a streaming server to deliver my Smooth Streaming content?

Usually playing back Smooth Streaming requires an IIS origin server, and you cannot stream directly from S3. However, if you distribute your content with CloudFront you can simply configure a CloudFront Smooth Streaming distribution, eliminating the need for a streaming server. See Configuring On-Demand Smooth Streaming.

Q: Why is the codec parameter that I want to change not exposed by the API?

In designing Amazon Elastic Transcoder, we wanted to create a service that was simple to use. Therefore, we expose the most frequently used codec parameters. If there is a parameter that you require, please let us know by letting us know through our forum.

Q: What settings do I use to preserve the dimensions of my video?

Use the following settings in your custom preset:
MaxWidth: auto; MaxHeight: auto; SizingPolicy: ShrinkToFit; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: How do I scale my output to a specified width and set the height to preserve the aspect ratio of the source content?

Use the following settings in your custom preset:
MaxWidth: [Desired Width]; MaxHeight: auto; SizingPolicy: Fit; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: How do I limit the height or width of a video without stretching the output to fit my set limit while preserving the input aspect ratio?

Use the following settings in your custom preset:
MaxWidth: [Desired Width Limit]; MaxHeight: [Desired Height Limit]; SizingPolicy: ShrinkToFit; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: What settings should I use to create a preset that causes the output video to fill the screen without distortion, if necessary cropping some of the edges ("center cut")?

Use the following settings in your custom preset:
MaxWidth: [Desired Width]; MaxHeight: [Desired Height]; SizingPolicy: Fill; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: What settings should I use to create a preset that causes the output video to fill the screen without cropping any image area, if necessary distorting the image ("squeeze" or "stretch")?

Use the following settings in your custom preset:
MaxWidth: [Desired Width]; MaxHeight: [Desired Height]; SizingPolicy: Stretch; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: How do I make my watermark scale with my video?

In the watermark settings of your transcoding preset, set the HorizontalAlign, VerticalAlign, and Target parameters as desired. Then set the HorizontalOffset and VerticalOffset with relative parameters. For example, to place the watermark 10% away from the edges, set both values to 10%.

Q: How do I avoid distorting my watermark?

If you do not want your watermark to be distorted when the video output is resized, set the SizingPolicy to ShrinkToFit while setting MaxWidth and MaxHeight to 100%. With these settings, Elastic Transcoder will never up-sample, expand, or distort your watermark.

Q: What are the settings for placing my watermark over the active video region rather than over the matte?

To place your watermark so that it is always over the active video content, use relative size for the MaxWidth and MaxHeight settings, and set the Target to be Content. For example, to fix the watermark size to 10% of the active output video size, set both MaxWidth and MaxHeight to 10%.

Q: How do I use multiple watermarks?

Presets specify placement settings for up to four watermarks. Each setting has an associated watermark ID. You can create a job with up to four watermarks by specifying an array of watermarks in the job creation call. Each element of the array specifies the Id of the watermark setting to use, and the watermark image file.

Q: Can I generate NTSC or PAL outputs?

Yes, you can generate both NTSC and PAL compliant outputs. The easiest way to generate NTSC and PAL compliant outputs is to specify the NTSC or PAL system preset when creating a transcoding job. Via the console, this is done by the preset drop down for each output of your transcoding job.

Pricing

Q: How much does Amazon Elastic Transcoder cost to use?

Pricing for Amazon Elastic Transcoder is described here. Our pricing does not require any commitment or minimum volume of jobs. We also offer a free tier that enables you to explore the service and transcode up to up to 20 minutes of audio-only output, 20 minutes of SD video output and 10 minutes of HD video output a month free of charge. To see terms and additional information on the free tier program, please visit the AWS Free Usage Tier page.

Q: How are jobs charged?

Transcoding jobs are charged according to the duration of the content. For example, media that lasts 60 minutes costs twice as much as media that lasts 30 minutes. High definition (HD) content costs twice as much as standard definition (SD). Audio-only output is priced lower than standard definition (SD) output. The minimum charge for a job is one minute. We do not charge for thumbnail generation, for API calls, or for Amazon S3 transfer within the same region. For more information, please refer to the Amazon Elastic Transcoder pricing page.

Q: How are fractional minutes charged?

Fractional minutes are rounded up. For example, if your output duration is less than a minute, you are charged for one minute. If your output duration is 1 minutes and 10 seconds, you are charged for 2 minutes.

Q: Do you charge for failed jobs?

Our policy is to forgive customers for failed jobs unless the number of failed jobs becomes excessive.

Q: Is it cheaper to use multiple outputs per job than to use separate jobs?

When you use multiple outputs per job, transcoding costs remain the same as if you had submitted multiple jobs for each output. However, the processing time will be quicker for larger jobs since the source file is only being transferred from your S3 bucket to Amazon Elastic Transcoder once.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Security

Q: Are my media assets secure?

You are in complete control of your media assets because they are stored in your own Amazon S3 buckets. You use IAM roles to grant us access to your specific Amazon S3 bucket.

Q: Can I set S3 permissions and storage options?

Amazon Elastic Transcoder enables you to specify which users, groups, and canonical IDs you want to grant access to your transcoded files, thumbnails and playlists, as well as the type of access that you want them to have. You can also specify whether to store transcoded content using Standard or Reduced Redundancy Storage. Please refer to Amazon Elastic Transcoder documentation for further information.

Q: Can I use encrypted input media files or encrypt my output files?

Yes. You can use encrypted mezzanine files as input to Amazon Elastic Transcoder, or protect your transcoded files by letting the service encrypt the output. Supported options range from fully managed integration with Amazon S3's Server-Side Encryption, to keys that you manage on your own and protect using AWS Key Management Service (KMS). Furthermore, encryption support is not limited to your video files. You can protect thumbnails, captions, and even watermarks.

Q: Do you support DRM?

Yes, we support packaging for Microsoft PlayReady DRM. Our Smooth Streaming packaging is compatible with the Microsoft PIFF 1.1, and our HLSv3 packaging is compatible with the Discretix 3.0.1 specification for Microsoft PlayReady.

Q: What are the best practices for securing my media?

Please refer to our AWS Security Best Practices whitepaper.

Q: Can I get a history of all Amazon Elastic Transcoder API calls made on my account for security, operational or compliance auditing?

Yes. To start receiving a history of all Elastic Transcoder API calls made on your account, you simply turn on AWS CloudTrail in CloudTrail's AWS Management Console. For more information, visit the AWS CloudTrail home page.

Q: Do I need to setup AWS KMS before using the Elastic Transcoder encryption and DRM packaging features?

Yes. You must first create a master AWS KMS key and add the role used by Elastic Transcoder as an authorized user of that key. Elastic Transcoder uses your KMS master key to protect the data encryption keys that it exchanges with you.

Q: Can I save the keys used to encrypt my HLS streams to S3?

Yes. If you elect to store your keys in S3, Elastic Transcoder will write your keys to the same folders as your playlist files, and your keys will be protected using Server-Side Encryption with Amazon S3-Managed Encryption Keys (SSE-S3).

Q: Can I rotate the keys used for HLS with AES-128 encryption?

Key rotation is not supported. All renditions and file segments share the same key.

Service Level Agreement (SLA)

Q: Does Amazon Elastic Transcoder offer a Service Level Agreement (SLA)?

Amazon Elastic Transcoder does not offer an SLA at this time.

Amazon SES FAQ

General

Q: What is Amazon Simple Email Service (Amazon SES)?

Amazon Simple Email Service (Amazon SES) is a highly scalable and cost-effective email service for businesses 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 for this type of email communication. In addition, the service integrates with other AWS services, making it easy to send emails from applications being hosted on AWS. With Amazon SES there is no long-term commitment, minimum spend or negotiation required — businesses can utilize a free usage tier, and beyond that pay only low fees for the number of emails sent or received plus data transfer fees.

Q: Who should use Amazon SES?

Any business or developer that needs a reliable, scalable, and inexpensive way to send or receive email — without having to build their own solution or license, install, and operate third-party software.

Sending Email

General

Q: What kinds of email can I send via Amazon SES?

You should send email that recipients expect and appreciate, and that comply with applicable laws and regulations, and the AWS Customer Agreement (including the AWS Acceptable Use Policy). Amazon SES can reliably deliver merchandising, subscription, transactional, and notification email messages.

Q: How does Amazon SES help ensure reliable email delivery?

For high email deliverability, Amazon SES uses content filtering technologies to scan a business's outgoing email messages to help ensure that the content meets ISP standards. To help businesses further improve the quality of email communications with their customers, Amazon SES provides a built-in feedback loop that includes bounce, complaint, and delivery notifications.

Q: What prevents Amazon SES users from sending spam?

Amazon SES uses in-house content filtering technologies to scan email content for spam and malware. In exceptional cases, accounts identified as sending spam or other low-quality email may be suspended, or AWS may take such other action as it deems appropriate. When malware is detected, Amazon SES prevents these emails from being sent.

Q: Do I need to sign up for Amazon EC2 or any other AWS services to use Amazon SES?

Amazon SES users do not need to sign up for any other AWS services. Any application with Internet access can use Amazon SES to deliver email, whether that application runs in your own data center, within Amazon EC2, or as a client software solution.

Q: How is Amazon SES different from Amazon SNS?

Amazon SES is for applications that need to send arbitrary communications via email. Amazon SES supports custom email header fields, and many MIME types.

By contrast, Amazon Simple Notification Service (Amazon SNS) is for messaging-oriented applications, with multiple subscribers requesting and receiving "push" notifications of time-critical messages via a choice of transport protocols, including HTTP, Amazon SQS, and email. The body of an Amazon SNS notification is limited to 8192 characters of UTF-8 strings, and is not intended to support multimedia content.

Getting Started Sending Email

Q: How can I get started using Amazon SES?

To use Amazon SES, you simply:

1. **Sign up**: After signing up for AWS, you can access the Amazon SES sandbox – an environment specifically designed for developers to test and evaluate the service.

2. **Verify domains or email addresses:** Before you can send an email using Amazon SES, you need to verify that you own the domain or address from which you will send email. To start the verification process, visit the Amazon SES console.

3. **Send a test email:** You can use the Amazon SES console, SMTP interface, or API to send a test email to an email address or domain that you verified.

4. **Apply to increase your sending limits:** When you are ready to use Amazon SES to send production email, you can apply to increase your sending limits and move your account out of the sandbox environment. It only takes a few minutes to apply, and you will typically receive a response within 24 hours.

5. **Send production email:** You can use either SMTP or the Amazon SES API to queue email messages for delivery.

6. **Get feedback:** Amazon SES provides useful statistics about your sending activities. With a simple API query or Amazon SES console visit, you can quickly obtain vital statistics such as volume sent, bounces and complaints.

For more information about how toset up email with Amazon SES, see theAmazon SES Developer Guide.

Q: What should I do after I'm finished testing and evaluating Amazon SES?

Once you are ready to use Amazon SES to send email, you can request an Amazon SES sending limit increase in Support Center. If granted, this increase will move your account out of the sandbox environment so that you can begin sending email to your customers. You will no longer need to verify recipient email addresses or recipient domains, and you will be able to send much larger quantities of email.

To request a sending limit increase, please complete ourbrief request form in Support Center. We generally respond to these requests within 24 hours.

Q: Do I need to set up reverse DNS records in order to use Amazon SES?

Amazon SES users do not need to do this. Amazon Web Services manages the IP addresses used by Amazon SES, and provides reverse DNS records for these addresses.

Q: I am sending email using my own mail servers hosted on Amazon EC2. Do I have to start using Amazon SES instead?

Amazon SES does not affect any Amazon EC2-based solution that you may currently have. You can continue to use your existing solution, or use Amazon SES, or do both at the same time.

Email-Sending Features and Functionality

Q: Does Amazon SES provide an SMTP endpoint?

Yes. Amazon SES provides a full-featured SMTP interface for seamless integration with applications that can send email via SMTP. You can connect directly to this SMTP interface from your applications, or configure your existing email server to use this interface as an SMTP relay.

To connect to the Amazon SES SMTP interface, you must create SMTP credentials. To create your credentials, go to the Amazon SES console and click the SMTP link.

Q: How can I use the Amazon SES SMTP interface?

To use the Amazon SES SMTP interface, all you need are your SMTP username and password, the SMTP endpoint name, and the port number. Using this information, you can connect to the Amazon SES SMTP interface in the same manner as any other SMTP relay.

For example, you can integrate your existing packaged software so that it sends email through Amazon SES. You can add email sending capability to your applications, using a programming language that supports SMTP. You can integrate Amazon SES sending with popular mail transfer agents (MTAs) such as Sendmail, Postfix, and Exim. You can even connect to the SMTP interface from the command line, and send SMTP commands directly.

For more information about the SMTP interface, go to theAmazon SES Developer Guide.

Q: Can I use Amazon SES to send bulk email?

Yes. Simply call the *SendEmail* or *SendRawEmail* APIs repeatedly for each email you would like to send. Software running on Amazon EC2, Amazon Elastic MapReduce, or your own servers can compose and deliver bulk emails via Amazon SES in whatever way best suits your business. If you already have your own bulk mailing software, it's easy to update it to deliver through Amazon SES — either by modifying the software to directly call Amazon SES, or reconfiguring it to deliver email through an Amazon SES SMTP relay as described above.

Q: Can Amazon SES send emails with attachments?

Yes. Amazon SES supports many popular content formats, including documents, images, audio, and video.

You can send email with attachments by using an email client that supports SMTP. When you configure such a client to send outgoing email through Amazon SES, the client constructs the appropriate MIME parts and email headers before sending the message. All of this happens automatically in your client with no additional user intervention.

You can also send email with attachments programmatically. To include an attachment in your email, construct a new multipart email message. In this message, include a MIME part that contains an appropriate *Content-Type* header, along with the MIME-encoded content. Next, use the *Content-Disposition* header to specify whether the content is to be displayed inline or treated as an attachment.

Once you have constructed your message, you can send it using the*SendRawEmail* API; you can also use the AWS Software Development Kits (SDKs) or a third-party library such as *boto* for Python.

To learn more about attachment pricing for Amazon SES, please seeAmazon SES Pricing.

Q: How do I control the character encoding of my emails with Amazon SES?

The SMTP protocol requires that all data be sent in 7-bit ASCII format. If you want to use a different character encoding with the Amazon SES SMTP interface, you will need to apply your desired encoding to your subject and body, and then convert them to a valid 7-bit ASCII message before sending it to the SMTP endpoint.

The *SendEmail* API accepts UTF-8 subject and body inputs, transcodes them into whatever format you specify via an optional encoding parameter, and automatically converts the resulting content into 7-bit ASCII with appropriate encoded-word syntax and content-transfer-encoding headers before transmission. The *SendRawEmail* API requires you to apply your desired encoding to your subject and body and then convert them to a valid 7-bit ASCII message before submitting each request.

Q: What happens if I try to send a malformed email message or send an email that is disallowed for any other reason?

If Amazon SES determines that it is unable to deliver your message it will return an error specifying that delivery failed and providing the reason. In rare cases, Amazon SES may not detect the problem with your email until after accepting your request. In such cases, your email will be returned to you as a bounce with a corresponding error code and reason.

Q: Does Amazon SES support Sender Policy Framework (SPF)?

Yes, and you may or may not need to publish an SPF record, depending on your use case. If you do not need to comply with Domain-based Message Authentication, Reporting and Conformance (DMARC) using SPF, you do not need to publish an SPF record to pass SPF authentication because by default, Amazon SES sends your emails from a MAIL FROM domain that Amazon SES owns. If you want to comply with DMARC using SPF, you must set up Amazon SES to use your own MAIL FROM domainand publish an SPF record.

Q: Does Amazon SES support Domain Keys Identified Mail (DKIM)?

Yes. Amazon SES will DKIM-sign outgoing messages on your behalf if you haveEasy DKIM configured and enabled. If you wish, you can also DKIM-sign your email yourself. To ensure maximum deliverability, there are a few DKIM headers that you should not sign. For more information, see Manual DKIM Signing in Amazon SES.

Q: Can emails from Amazon SES comply with DMARC?

Yes. With Amazon SES, your emails can comply withDMARC through SPF, DKIM, or both.

Q: Does Amazon SES send email over an encrypted connection using Transport Layer Security (TLS)?

Yes. If the receiving mail server advertises the STARTTLS extension, Amazon SES will attempt to upgrade the connection to a TLS connection. If that fails, Amazon SES will fall back to plain text.

Q: What TLS version does Amazon SES use to send email?

Amazon SES only supports TLS v1.

Q: Can I test Amazon SES responses without sending email to real recipients?

Yes. The Amazon SES mailbox simulator provides an easy way to test your sending rate and generic email responses, including bounces and complaints, without sending to actual recipients. Emails to the mailbox simulator do not affect your bounce and complaint metrics, and do not count against your sending quota.

For more information on the Amazon SES mailbox simulator, go to theAmazon SES Developer Guide.

Email-Sending Performance and Reliability

Q: How long will it take for emails sent via Amazon SES to arrive?

Generally, Amazon SES attempts to deliver emails to the Internet within a few seconds of each request. However, due to a number of factors and the inherent uncertainties of the Internet, we cannot predict with certainty when your email will arrive nor the exact route the message will take to get to its destination. For example, an ISP might be unable to deliver the email to the recipient because of a temporary condition such as "mailbox full." In these cases, Amazon attempts to retry the message for a length of time. If the error is permanent, such as "mailbox does not exist," Amazon SES does not retry the delivery attempt and you will receive a hard bounce notification. You can set up delivery notifications to alert you when Amazon SES successfully delivers one of your emails to a recipient's mail server. For troubleshooting delivery issues, see the Amazon SES Developer Guide.

Q: Does Amazon SES guarantee receipt of my emails?

Amazon SES closely monitors ISP guidelines worldwide to help ensure that legitimate, good quality email will be delivered reliably to recipient inboxes. However, neither Amazon SES nor any other email-sending service can guarantee that emails will be received. ISPs can drop or lose email messages, recipients can accidentally provide the wrong email address, and if recipients do not wish to receive your email messages, ISPs may choose to reject or silently drop them.

Sending Emails Programmatically

Q: Do the AWS Software Development Kits contain support for Amazon SES?

Yes. You can use the AWS Software Development Kits (SDKs) for Android, iOS, Java, .NET, Node.js, Python, PHP, and Ruby to access the Amazon SES API. These SDKs make it easy to email-enable your applications, allowing them to send email with a simple API call.

Q: How do I make requests to Amazon SES?

Amazon SES accepts Query requests over HTTPS. These requests use verbs such as GET or POST, and a parameter named Action to indicate the action being performed. For security reasons, Amazon SES does not support HTTP requests; you must use HTTPS instead.

Q: Can I use Amazon SES for email-to-text delivery?

Yes. For example, if you know the email address associated with a mobile phone, you can use Amazon SES to send an email message to an SMS gateway, and the message will be delivered to the phone.

Q: Can I use Amazon SES to send email from my existing applications?

Yes. The Amazon SES Developer Guide provides instructions for configuring common mail transfer agents (MTAs) to use Amazon SES as an email transport. By following these instructions, you can create a private SMTP relay for use with any existing SMTP client software. This includes any software that you write, or any third-party software that supports SMTP, such as content management and database management systems.

Email-Sending Notifications

Q: How does Amazon SES send bounce, complaint, and delivery notifications to me?

Amazon SES forwards bounce and complaint notifications to you by email or sends them to an Amazon SNS topic, depending on your configuration. Delivery notifications, which are triggered when Amazon SES successfully delivers one of your emails to a recipient’s mail server, are sent to you only through Amazon SNS.

Q. What actions should I take if I receive a bounce or a complaint?

You will need to analyze each bounce and complaint email or Amazon SNS JSON object that you receive to determine the cause. Bounces are usually caused by attempting to send to a nonexistent recipient; complaints arise when the recipient indicates that they do not want to receive your message. In either case, we recommend that you stop sending to these email addresses.

Q. Is there an additional cost to use Amazon SNS to receive bounce, complaint, and delivery notifications?

You will incur normal Amazon SNS expenses if you use it for bounce, complaint, and/or delivery notifications. Please see Amazon SNS Pricing for information about their free tier and full details about their pricing.

Q. When can I expect to be notified of bounces, complaints, and deliveries?

After an ISP sends a bounce or complaint to Amazon SES, we will usually pass it to you within a few seconds via Amazon SNS or email. However, we may not receive the bounce or complaint notification from the recipient's ISP for a period of time ranging from seconds to weeks or longer, depending on how quickly the ISP notifies us. Delivery notifications are published as soon as Amazon SES delivers an email to a recipient's mail server. In most cases, email sent through Amazon SES is delivered within seconds, but occasionally it might take longer.

Q: Will I be affected by any bounces or complaints that are caused by other Amazon SES users?

Even if other Amazon SES users cause bounces or complaints, your ability to send email should remain unchanged.

There is one exception. Amazon SES maintains a suppression list of recipient email addresses that have recently caused a hard bounce for *any* Amazon SES customer. If you try to send an email through Amazon SES to an address that is on the suppression list, the call to Amazon SES succeeds, but Amazon SES treats the email as a hard bounce instead of attempting to send it. Like any hard bounce, suppression list bounces count towards your sending quota and your bounce rate. An email address can remain on the suppression list for up to 14 days. For details, go to the Amazon SES Developer Guide.

Q: What if I'm sure that a recipient address on the Amazon SES suppression list is valid?

You can submit a suppression list removal request using the Amazon SES console.

Email-Sending Limits and Restrictions

Q: Are there limits as to whom I can send emails to?

As described in the AWS Customer Agreement (including the AWS Acceptable Use Policy), each user is responsible for remaining compliant with applicable laws and regulations. Further, each user is responsible for sending only email that recipients want and expect to receive. AWS may suspend any accounts identified as sending spam or other unwanted low-quality email, or take other action as AWS deems appropriate.

Q: Can I send emails from any source email address?

Yes. You can specify any "From" address in the email messages that you send using Amazon SES, but to prevent phishing, you must demonstrate your ownership and control of each email address or domain that you send from. Otherwise, your email will not be accepted for delivery. You can verify ownership and control of email addresses and domains by using either the Amazon SES console or the Amazon SES API. For details on address verification and domain verification, see the Amazon SES Developer Guide.

You can verify a total of up to 1000 email addresses and domains, in any combination.

Q: Is there a limit on the size of emails Amazon SES will deliver?

Amazon SES will accept email messages up to 10 MB in size. This includes any attachments that are part of the message.

Q: Is there a limit on the number of recipients per email message?

Amazon SES lets you specify a maximum of 50 recipients for every message you send. In other words, the combined number of *To:*, *CC:*, and *BCC:* recipients must not exceed 50. If you need to send an email message to more than 50 recipients, then you need to send multiple messages, each addressed to 50 recipients or fewer.

Q: Are there any limits on how many emails I can send?

Every Amazon SES sender has a unique set of sending limits, which are calculated by Amazon SES on an ongoing basis:

Sending quota — the maximum number of emails you can send in a 24-hour period.

Maximum send rate — the maximum number of emails that Amazon SES can accept from your account per second.

Note: The rate at which Amazon SES accepts your messages might be less than the maximum send rate.

New Amazon SES users start in the Amazon SES sandbox, which is a test environment that has a sending quota of 200 emails per 24-hour period, at a maximum rate of 1 email per second. To request an increase in these limits, you can submit an SES Sending Limits Increase case at any time.

Sending limits are based on recipients rather than on messages. You can check your sending limits at any time by using the Amazon SES console.

Note that if your email is detected to be of poor or questionable quality (e.g., high complaint rates, high bounce rates, spam, or abusive content), Amazon SES might temporarily or permanently reduce your permitted send volume, or take other action as AWS deems appropriate.

Q: Why are these sending limits in place?

Using these limits to steadily "ramp up" your sending activity helps you improve your deliverability. This approach helps Amazon SES adapt to your particular sending needs. As you continue to send high-quality email, Amazon SES adjusts to your particular usage patterns, and gradually increases your sending limits. You can also submit an SES Sending Limits Increase case at any time if you anticipate needing to raise your sending limits.

Q: How can I monitor the email I send using Amazon SES?

Amazon SES provides three main ways to monitor your bounces, complaints, deliveries, sent emails, and rejected emails. The first method is to use the Amazon SES console, Amazon SES API, or Amazon CloudWatch to access basic email sending metrics across your entire AWS account. The second method is to set up Amazon SES to send you detailed feedback notifications through email or through Amazon SNS. The third method is to use Amazon SES event publishing. With event publishing, you categorize your emails and collect event data for each category of emails separately using either Amazon CloudWatch or Amazon Kinesis Firehose. You can set up Amazon Kinesis Firehose to send the event records to Amazon Redshift, Amazon S3, or Amazon Elasticsearch Service. If you use Amazon Elasticsearch Service, you can visualize your event data using Kibana. For more information about monitoring methods, see the Amazon SES Developer Guide.

Receiving Email

General

Q: How do I get set up to receive mail using Amazon SES?

You must first verify your domain with Amazon SES to prove that you own it by using the procedure described in the documentation. This process is identical to the domain verification process Amazon SES uses for sending mail. If you are already using your domain to send mail with Amazon SES, you do not need to verify it again. Once you have successfully verified your domain, the next step is to publish a DNS mail exchanger (MX) record for your domain that points to the regional Amazon SES endpoint that you want to use to receive email. Publishing the MX record is not required to receive mail through Amazon SES, but you must do so if you want your incoming mail to be automatically routed to Amazon SES, rather than route it yourself. The final step is to create a receipt rule using Amazon SES's console or API. A receipt rule tells Amazon SES what you'd like done with the mail received on your behalf. A basic rule would simply write all of your mail to an Amazon S3 bucket you own.

Q: What happens when Amazon SES receives my mail?

When Amazon SES receives a message, it references your active receipt rule set to determine whether or not you have any rules that match any of the incoming message's recipients. If there aren't any matches, or if the mail was sent from an IP address on your IP address block list, Amazon rejects the mail synchronously in the SMTP conversation. Otherwise, Amazon SES accepts the mail. After Amazon SES accepts the mail, it evaluates your active receipt rule set, and all of the receipt rules that apply to the message are applied in the order that they are defined.

Amazon SES's next steps are determined by the actions you defined in your receipt rules. You can set up your receipt rule to have Amazon SES deliver your messages to an Amazon S3 bucket, call your custom code via an AWS Lambda function, or publish notifications to Amazon SNS. You can also configure Amazon SES to drop or bounce messages you do not want to receive.

Q: How do I access my mail in Amazon S3?

When you set up a receipt rule to specify that Amazon SES should write your messages to an Amazon S3 bucket, you have the option of setting up Amazon SNS notifications as well. The notifications, which contain general information about the message and the action taken on it, will include the unique message ID of the message. Use this message ID to retrieve the corresponding message from Amazon S3.

Q: How can I process emails I receive?

There are two ways to process mail that you receive. You can either write an application that listens for Amazon SNS notifications from Amazon SES, retrieves the mail from Amazon S3, and processes it. Alternatively, you can write a custom AWS Lambda function. The AWS Lambda event contains all of the metadata about the message that was received, but does not include the actual message content. If you need access to the message content from within the AWS Lambda, then you need to first write the message to Amazon S3 using an Amazon S3 action before your AWS Lambda action is evaluated. AWS Lambda actions can be executed synchronously or asynchronously, depending on whether or not the AWS Lambda function needs to return a result that influences how other actions are executed. We recommend that you use asynchronous unless synchronous is absolutely necessary for your use case.

Q: Can multiple different AWS accounts receive mail on the same domain?

Yes. It is valid for more than one AWS account to receive mail for the same domain. This means that for each email that arrives on the shared domain, a copy of the message is processed by each account's receipt rule set independently.

Security

Q: How can I encrypt my emails?

Amazon SES integrates with AWS Key Management Service (KMS) to optionally encrypt the mail that it writes to your Amazon S3 bucket. You can either use the default Amazon SES KMS master key in your account for encryption, which does not require additional setup, or you can set up a new master KMS key that grants the Amazon SES service principal permission to generate data keys. Amazon SES uses client-side encryption to encrypt your mail prior to writing it to Amazon S3. This means that it is necessary for you to decrypt the content on your side after retrieving the mail from Amazon S3. The AWS Java SDK and AWS Ruby SDK provide a client that is able to handle the decryption for you.

Q: What can I do about incoming email that was sent over an unencrypted connection?

Within the receipt rules you set up, you can specify for Amazon SES to reject mail from connections that don't use Transport Layer Security (TLS).

Unwanted Mail

Q: What spam protections are in place?

Amazon SES has a number of different spam protection measures in place. It uses block lists to prevent mail from known spammers from even entering the system. Every email Amazon SES accepts undergoes virus and spam scanning. Amazon SES makes the verdicts available to you, enabling you to ultimately decide whether or not you trust the message. In addition to the spam and virus verdicts, Amazon SES provides the DKIM and SPF check results.

Q: Am I billed for spam?

You are not billed for any mail that's rejected during the SMTP conversation, so it's to your advantage to reject as much unwanted mail as possible at that time. Currently, Amazon SES exposes two different mechanisms that enable you to control whether or not mail is accepted during the SMTP conversation. The first way is to use custom IP address allow lists and block lists. By adding the IP addresses that the spam is coming from to your IP address block list, you cut off the flow, which prevents you from being charged for the unwanted mail. The second mechanism is to set up receipt rules. Amazon SES only accepts messages if there is at least one receipt rule that matches a recipient of the message. Additionally, Amazon SES maintains its own IP address block list (subject to change) and will block mail from the IP addresses in that list without your intervention. If you want to enable delivery from one of those addresses, you can add it to your allow list.

Email-Receiving Limits and Restrictions

Q: Is there any size limit to the messages that I can receive through Amazon SES?

If you choose for your messages to be stored in an Amazon S3 bucket, the maximum message size (including headers) is 30 MB. If you choose to receive your messages through Amazon

SNS notifications, the maximum message size (including headers) is 150 KB.

Q: Is there a limited throughput at which I can receive messages through Amazon SES?

No.

Q: Does Amazon SES put any restrictions on AWS Lambda functions in addition to the restrictions imposed by AWS Lambda?

Yes. There is a 30-second timeout on *RequestResponse* invocations

AWS Integration

Q: How does Amazon SES integrate with Amazon WorkMail?

Amazon WorkMail uses Amazon SES to send and receive mail. When you set up Amazon WorkMail, Amazon WorkMail creates two items within your Amazon SES configuration settings: 1) a sending authorization policy that allows Amazon WorkMail to send mail through your domain, and 2) a receipt rule with a WorkMail action that delivers your domain's incoming mail to Amazon WorkMail. Do not remove these items.

Privacy

Q: Who, if anyone, has access to my email content?

We take our privacy and data protection responsibilities very seriously. Amazon SES uses in-house anti-spam/anti-virus technologies to filter email messages containing poor-quality content and prevent them from being sent. We will only access email content under very limited circumstances, such as system troubleshooting, or investigating fraudulent or abusive activity. Furthermore, other Amazon SES customers do not have access to your email content.

Billing

Q: How much does Amazon SES cost?

Pay only for what you use. There is no minimum fee. In addition, Amazon EC2 users can get started with Amazon SES for free. You pay only low charges for messages sent, attachments, and data transfer. Please refer to Amazon SES Pricing for more information on pricing, data transfer costs, and free usage.

Q: How will I be charged and billed for my use of Amazon SES?

There are no set-up fees to begin using the service. At the end of the month, you will be charged for that month's usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site. To view your charges for the current billing period, log into your Amazon Web Services account, click "My Account/Console," and then click "Account Activity."

Q: When does billing of my Amazon SES use begin and end?

Your Amazon SES billing cycle begins on the first day of each month and ends on the last day of each month. Your monthly charges will be totaled at the end of each month.

Q: How do I get started with the Amazon SES free tier?

To benefit from Amazon SES free tier pricing, you need to call Amazon SES from within Amazon EC2 or AWS Elastic Beanstalk. If you do this, then your AWS bill will reflect your free tier usage.

Q: Can I take advantage of Amazon SES free tier pricing if I'm in the Amazon SES sandbox?

Yes. However, to take full advantage of the Amazon SES free tier, you shouldrequest higher sending limits for Amazon SES.

Q: Where can I track my Amazon SES usage?

You can track your usage on your AWS Account Activity page.

Amazon SNS FAQ

General

Q: What is 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 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 for developers. Amazon SNS follows the “publish-subscribe” (pub-sub) messaging paradigm, with notifications being delivered to clients using a “push” mechanism that eliminates the need to periodically check or “poll” for new information and updates. With simple APIs requiring minimal up-front development effort, no maintenance or management overhead and pay-as-you-go pricing, Amazon SNS gives developers an easy mechanism to incorporate a powerful notification system with their applications.

Q: What are some example uses for Amazon SNS notifications?

The Amazon SNS service can support a wide variety of needs including monitoring applications, workflow systems, time-sensitive information updates, mobile applications, and any other application that generates or consumes notifications. For example, Amazon SNS can be used in workflow systems to relay events among distributed computer applications, move data between data stores or update records in business systems. Event updates and notifications concerning validation, approval, inventory changes and shipment status are immediately delivered to relevant system components as well as end-users. Another example use for Amazon SNS is to relay time-critical events to mobile applications and devices. Since Amazon SNS is both highly reliable and scalable, it provides significant advantages to developers who build applications that rely on real-time events.

Q: What are the benefits of using Amazon SNS?

Amazon SNS offers several benefits making it a versatile option for building and integrating loosely-coupled, distributed applications:

- Instantaneous, push-based delivery (no polling)
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols
- Inexpensive, pay-as-you-go model with no up-front costs
- Web-based AWS Management Console offers the simplicity of a point-and-click interface

Q: How does Amazon SNS work?

It is very easy to get started with Amazon SNS. Developers must first create a “topic” which is an “access point” – identifying a specific subject or event type – for publishing messages and allowing clients to subscribe for notifications. Once a topic is created, the topic owner can set policies for it such as limiting who can publish messages or subscribe to notifications, or specifying which notification protocols will be supported (i.e. HTTP/HTTPS, email, SMS). Subscribers are clients interested in receiving notifications from topics of interest; they can subscribe to a topic or be subscribed by the topic owner. Subscribers specify the protocol and end-point (URL, email address, etc.) for notifications to be delivered. When publishers have information or updates to notify their subscribers about, they can publish a message to the topic – which immediately triggers Amazon SNS to deliver the message to all applicable subscribers.

Q: How is Amazon SNS different from Amazon SQS?

Amazon Simple Queue Service (SQS) and Amazon SNS are both messaging services within AWS, which provide different benefits for developers. Amazon SNS allows applications to send time-critical messages to multiple subscribers through a “push” mechanism, eliminating the need to periodically check or “poll” for updates. Amazon SQS is a message queue service used by distributed applications to exchange messages through a polling model, and can be used to decouple sending and receiving components. Amazon SQS provides flexibility for distributed components of applications to send and receive messages without requiring each component to be concurrently available.

Q: How can I get started using Amazon SNS?

To sign up for Amazon SNS, click the “Sign up for Amazon SNS” button on the Amazon SNS detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon SNS sign-up process. After signing up, please refer to the Amazon SNS documentation and Getting Started Guide to begin using Amazon SNS. Using the AWS Management Console, you can easily create topics, add subscribers, send notifications, and edit topic policies – all from your browser.

Q: Is Amazon SNS supported in the AWS Management Console?

Amazon SNS is supported in the AWS Management Console which provides a point-and-click, web-based interface to access and manage Amazon SNS. Using the AWS Management Console, you can create topics, add subscribers, and send notifications – all from your browser. In addition, the AWS Management Console makes it easy to publish messages over your protocol of choice (HTTP, email, SQS protocol, etc.) and edit topic policies to control publisher and subscriber access. The AWS Management Console is provided free of charge at: http://aws.amazon.com/console

Q: What are the Amazon SNS service access points in each region?

The US East (Northern Virginia) end-point is: http://sns.us-east-1.amazonaws.com

The US West (Oregon) end-point is: http://sns.us-west-2.amazonaws.com

The US West (Northern California) end-point is: http://sns.us-west-1.amazonaws.com

The EU(Ireland) end-point is: http://sns.eu-west-1.amazonaws.com

The EU(Frankfurt) end-point is: http://sns.eu-central-1.amazonaws.com

The Asia Pacific (Singapore) end-point is: http://sns.ap-southeast-1.amazonaws.com

The Asia Pacific (Tokyo) end-point is: http://sns.ap-northeast-1.amazonaws.com

The Asia Pacific (Sydney) end-point is: http://sns.ap-southeast-2.amazonaws.com

The South America (Sao Paulo) end-point is: http://sns.sa-east-1.amazonaws.com

Q: Can I get a history of SNS API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. SNS supports AWS CloudTrail, a web service that records AWS API calls for your account and delivers log files to you. With CloudTrail, you can obtain a history of such information as the identity of the API caller, the time of the API call, the source IP address of the API caller, the request parameters, and the response elements returned by SNS.

SNS currently supports CloudTrail auditing for authenticated calls only. CloudTrail Audit logs for unauthenticated ConfirmSubscription and Unsubscribe calls are not available at this time. For more information, see the CloudTrail section of the SNS Developer Guide.

To receive a history of SNS API calls made on your account, simply turn on AWS CloudTrail in the AWS Management Console. To learn more about AWS CloudTrail,click here.

Billing

Q: How much does Amazon SNS cost?

With Amazon SNS, there is no minimum fee and you pay only for what you use. Users pay $0.50 per 1 million Amazon SNS Requests, $0.06 per 100,000 Notification deliveries over HTTP, $0.75 per 100 Notification deliveries over SMS and $2.00 per 100,000 Notification deliveries over Email.

Amazon SNS also includes a Free Tier, where users can get started with Amazon SNS for free. Each month, Amazon SNS customers pay no charges for the first 1 million Amazon SNS Requests, no charges for the first 100,000 Notifications over HTTP, no charges for the first 100 Notifications over SMS and no charges for the first 1,000 Notifications over Email.

Please refer to the Amazon SNS Details page for additional details on pricing and data transfer costs.

Q: How will I be charged and billed for my use of Amazon SNS?

There are no set-up fees to begin using the service. At the end of the month, your credit card will automatically be charged for that month’s usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site by logging into your Amazon Web Services account and clicking “Account Activity” under “Your Web Services Account”.

Q: When does billing of my Amazon SNS use begin and end?

Your Amazon SNS billing cycle begins on the first day of each month and ends on the last day of each month. Your monthly charges will be totaled at the end of each month.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.

Features and Functionality

Q: What is the format of an Amazon SNS topic?

Topic names are limited to 256 characters. Alphanumeric characters plus hyphens (-) and underscores (_) are allowed. Topic names must be unique within an AWS account. After you delete a topic, you can reuse the topic name. When a topic is created, Amazon SNS will assign a unique ARN (Amazon Resource Name) to the topic, which will include the service name (SNS), region, AWS ID of the user and the topic name. The ARN will be returned as part of the API call to create the topic. Whenever a publisher or subscriber needs to perform any action on the topic, they should reference the unique topic ARN.

The following is the ARN for a topic named “mytopic” created by a user with the AWS account ID “123456789012” and hosted in the US East region:

arn:aws:sns:us-east-1:1234567890123456:mytopic Note: Users should NOT attempt to build the topic ARN from its separate components – they should always use the name returned from the API call to create the topic.

Q: What are the available operations for Amazon SNS and who can perform these operations?

Amazon SNS provides a set of simple APIs to enable event notifications for topic owners, subscribers and publishers.

Owner operations:

- CreateTopic – Create a new topic.
- DeleteTopic – Delete a previously created topic.
- ListTopics – List of topics owned by a particular user (AWS ID).
- ListSubscriptionsByTopic – List of subscriptions for a particular topic

- SetTopicAttributes – Set/modify topic attributes, including setting and modifying publisher/subscriber permissions, transports supported, etc.
- GetTopicAttributes – Get/view existing attributes of a topic
- AddPermission – Grant access to selected users for the specified actions
- RemovePermission – Remove permissions for selected users for the specified actions

Subscriber operations:

- Subscribe – Register a new subscription on a particular topic, which will generate a confirmation message from Amazon SNS
- ConfirmSubscription – Respond to a subscription confirmation message, confirming the subscription request to receive notifications from the subscribed topic
- UnSubscribe – Cancel a previously registered subscription
- ListSubscriptions – List subscriptions owned by a particular user (AWS ID)

Publisher operations:

- Publish: Publish a new message to the topic.

Q: Why are there two different APIs to list subscriptions?

The two APIs to list subscriptions perform different functions and return different results:

- The ListSubscriptionsByTopic API allows a topic owner to see the list of all subscribers actively registered to a topic.
- The ListSubscriptions API allows a user to get a list of all their active subscriptions (to one or more topics).

Q: What are the different delivery formats/transports for receiving notifications?

In order for customers to have broad flexibility of delivery mechanisms, Amazon SNS supports notifications over multiple transport protocols. Customers can select one the following transports as part of the subscription requests:

- “HTTP”, “HTTPS” – Subscribers specify a URL as part of the subscription registration; notifications will be delivered through an HTTP POST to the specified URL.
- ”Email”, “Email-JSON” – Messages are sent to registered addresses as email. Email-JSON sends notifications as a JSON object, while Email sends text-based email.

- “SQS” – Users can specify an SQS queue as the endpoint; Amazon SNS will enqueue a notification message to the specified queue (which subscribers can then process using SQS APIs such as ReceiveMessage, DeleteMessage, etc.)
- “SMS” – Messages are sent to registered phone numbers as SMS text messages.

Q: Can topic owners control the transports that are allowed on topics they create/own?

Topic owners can configure specific transports on their topics by setting the appropriate permissions through access control policies.

Q: How does an owner set Access Control policies?

Please refer to the Amazon SNS Getting Started Guide for an overview of setting access control policies.

Q: Can a single topic support subscriptions over multiple protocols/transports?

Subscribers to an Amazon SNS topic can receive notifications on any transport supported by the topic. A topic can support subscriptions and notification deliveries over multiple transports.

Q: Can Amazon SNS be used with other AWS services?

Amazon SNS can be used with other AWS services such as Amazon SQS, Amazon EC2 and Amazon S3. Here is an example of how an order processing workflow system uses Amazon SNS with Amazon EC2, SQS, and SimpleDB. In this workflow system, messages are sent between application components whenever a transaction occurs or an order advances through the order processing pipeline. When a customer initially places an order, the transaction is first recorded in Amazon SimpleDB and an application running on Amazon EC2 forwards the order request to a payment processor which debits the customer’s credit card or bank account. Once approved, an order confirmation message is published to an Amazon SNS topic. In this case, the topic has various subscribers over Email/HTTP – merchant, customer and supply chain partners – and notifications sent by Amazon SNS for that topic can instantly update all of them that payment processing was successful. Notifications can also be used to orchestrate an order processing system running on EC2, where notifications sent over HTTP can trigger real-time processing in related components such as an inventory system or a shipping service. By integrating Amazon SNS with Amazon SQS, all notifications delivered are also persisted in an Amazon SQS queue where they are processed by an auditing application at a future time.

Q: Is Amazon SNS available in all regions where AWS services are available?

Amazon SNS is available in the US East (Northern Virginia), US West (Oregon), US West (Northern California), EU (Ireland), EU (Frankfurt), Asia Pacific (Singapore), Asia Pacific

(Tokyo), Asia Pacific (Sydney) and South America (Sao Paulo) regions.

Q: How soon can customers recreate topics with previously used topic names?

Topic names should typically be available for reuse approximately 30-60 seconds after the previous topic with the same name has been deleted. The exact time will depend on the number of subscriptions which were active on the topic – topics with a few subscribers will be available instantly for reuse, topics with larger subscriber lists may take longer.

Transports

Q: How would a user subscribe for notifications to be delivered over email?

To receive email notifications for a particular topic, a subscriber should specify “Email” or “Email-JSON” as the protocol and provide a valid email address as the end-point. This can be done using the AWS Management Console or by calling the Amazon SNS API directly. Amazon SNS will then send an email with a confirmation link to the specified email address, and require the user monitoring the email address to explicitly opt-in for receiving email notifications from that particular topic. Once the user confirms the subscription by clicking the provided link, all messages published to the topic will be delivered to that email address. The AWS Management Console is available at: http://aws.amazon.com/console

Q: Why does Amazon SNS provide two different transports to receive notifications over email?

The two email transports are provided for two distinct types of customers/end-users. “Email-JSON” sends notifications as a JSON object, and is meant for applications to programmatically process emails. The ”Email” transport is meant for end-users/consumers and notifications are regular, text-based messages which are easily readable.

Q: Can a user change the Subject and Display name for notifications sent over Email/Email-JSON?

Amazon SNS allows users to specify the Subject field for emails as a parameter passed in to the Publish API call and can be different for every message published. The Display name for topics can be set using the SetTopicAttributes API – this name applies to all emails sent from this topic.

Q: Do subscribers need to specifically configure their email settings to receive notifications from Amazon SNS?

In most cases, users should be able to receive subscription confirmations and notifications from Amazon SNS without doing anything specific. However, there could be cases where the email provider’s default settings or other user-specific configurations mistakenly redirect the emails to the junk/spam folder. To ensure that users see confirmation messages and notifications sent from Amazon SNS, users can add “no-reply@sns.amazonaws.com” to their contact lists and check their junk/spam folders for messages from Amazon SNS.

Q: In the case of passing in an SQS queue as an end-point, will users need to create the queue prior to subscribing? What permissions will the queue require?

Using the SQS console, users should create the SQS queue prior to subscribing it to a Topic. Select this queue on the console, and from the ‘Queue Actions’ in the menu bar, select ‘Subscribe Queue to SNS Topic’ from the drop-down list. In the subscribe dialog box, select the topic from the ‘Choose a Topic’ drop-down list, and click the ‘Subscribe’ button. For complete step-by-step instructions, please refer to the Amazon SNS documentation.

Q: How would a developer setup an Amazon SQS queue to receive Amazon SNS notifications?

To have Amazon SNS deliver notifications to an SQS queue, a developer should subscribe to a topic specifying “SQS” as the transport and a valid SQS queue as the end-point. In order to allow the SQS queue to receive notifications from Amazon SNS, the SQS queue owner must subscribe the SQS queue to the Topic for Amazon SNS to successfully deliver messages to the queue.

If the user owns both the Amazon SNS topic being subscribed to and the SQS queue receiving the notifications, nothing further is required. Any message published to the topic will automatically be delivered to the specified SQS queue. If the user owning the SQS queue is not the owner of the topic, Amazon SNS will require an explicit confirmation to the subscription request.

Please refer to the Amazon SNS documentation for further details on subscribing an SQS queue to a topic and setting access control policies for SQS queues.

Q: How can I fanout identical messages to multiple SQS queues?

Create an SNS topic first using SNS. Then create and subscribe multiple SQS queues to the SNS topic. Now whenever a message is sent to the SNS topic, the message will be fanned out to the SQS queues, i.e. SNS will deliver the message to all the SQS queues that are subscribed to the topic.

Q: What is the format of structured notification messages sent by Amazon SNS?

The notification message sent by Amazon SNS for deliveries over HTTP, HTTPS, Email-JSON and SQS transport protocols will consist of a simple JSON object, which will include the following information:

- MessageId: A Universally Unique Identifier, unique for each notification published.
- Timestamp: The time (in GMT) at which the notification was published.
- TopicArn: The topic to which this message was published
- Type: The type of the delivery message, set to “Notification” for notification deliveries.
- UnsubscribeURL: A link to unsubscribe the end-point from this topic, and prevent receiving any further notifications.
- Message: The payload (body) of the message, as received from the publisher.
- Subject: The Subject field – if one was included as an optional parameter to the publish API call along with the message.
- Signature: Base64-encoded “SHA1withRSA” signature of the Message, MessageId, Subject (if present), Type, Timestamp, and Topic values.
- SignatureVersion: Version of the Amazon SNS signature used.

Notification messages sent over the “Email” transport only contain the payload (message body) as received from the publisher.

Q: How would a user subscribe for notifications to be delivered over SMS?

Please refer to the 'SMS Related Question' section below.

Security

Q: How can users secure the messages sent to my topics?

All API calls made to Amazon SNS are validated for the user’s AWS Id and the signature. In addition, we recommend that users secure their data over the wire by connecting to our secure SSL end-points.

Q: Who can create a topic?

Topics can only be created by users with valid AWS IDs who have signed up for Amazon SNS. The easiest way to create a topic is to use the AWS Management Console. It can also be created through the CreateTopic API. The AWS Management Console is available at: http://aws.amazon.com/console

Q: Can multiple users publish to a single topic?

A topic owner can set explicit permissions to allow more than one user (with a valid AWS ID) to publish to a topic. By default, only topic owners have permissions to publish to a topic.

Q: How can the owner grant/revoke publish or subscribe permissions on a topic?

The AddPermission and RemovePermission APIs provide a simple interface for developers to add and remove permissions for a topic. However, for conditional access and more advanced use cases, users should use access control policies to manage permissions. The easiest way to manage permissions is to use the AWS Management Console. The AWS Management Console is available at: http://aws.amazon.com/console

Q: How does a topic owner give access to subscribers?Do subscribers have to have valid AWS IDs?

Amazon SNS makes it easy for users with and without AWS IDs to receive notifications. The owner of the topic can grant/restrict access to subscribers by setting appropriate permissions for the topic using Access Control policies. Users can receive notifications from Amazon SNS in two ways:

- Users with AWS IDs: Subscribers with valid AWS IDs (please refer to this link for details on obtaining AWS IDs) can subscribe to any topic directly – as long as the topic owner has granted them permissions to do so. The AWS IDs will be validated as part of the subscription registration.
- Other users: Topic owners can subscribe and register end-points on behalf of users without AWS IDs.

In both cases, the owner of the subscription endpoint needs to explicitly opt-in and confirm the subscription by replying to confirmation message sent by Amazon SNS.

Q: How will Amazon SNS authenticate API calls?

All API calls made to Amazon SNS will validate authenticity by requiring that requests be signed with the secret key of the AWS ID account and verifying the signature included in the requests.

Q: How does Amazon SNS validate a subscription request to ensure that notifications will not be sent to users as spam?

As part of the subscription registration, Amazon SNS will ensure that notifications are only sent to valid, registered subscribers/end-points. To prevent spam and ensure that a subscriber end-point is really interested in receiving notifications from a particular topic, Amazon SNS requires an explicit opt-in from subscribers using a 2-part handshake:

i. When a user first calls the Subscribe API and subscribes an end-point, Amazon SNS will send a confirmation message to the specified end-point.

ii. On receiving the confirmation message at the end-point, the subscriber should confirm the subscription request by sending a valid response. Only then will Amazon SNS consider the subscription request to be valid. If there is no response to the challenge, Amazon SNS will not send any notifications to that end-point. The exact mechanism of confirming the subscription varies by the transport protocol selected:

- For HTTP/HTTPS notifications, Amazon SNS will first POST the confirmation message (containing a token) to the specified URL. The application monitoring the URL will have to call the ConfirmSubscription API with the token included token.

- For Email and Email-JSON notifications, Amazon SNS will send an email to the specified address containing an embedded link. The user will need to click on the embedded link to confirm the subscription request.

- For SQS notifications, Amazon SNS will enqueue a challenge message containing a token to the specified queue. The application monitoring the queue will have to call the ConfirmSubscription API with the token.

Note: The explicit “opt-in” steps described above are not required for the specific case where you subscribe your Amazon SQS queue to your Amazon SNS topic – and both are “owned” by the same AWS account.

Q: How long will subscription requests remain pending, while waiting to be confirmed?

Token included in the confirmation message sent to end-points on a subscription request are valid for 3 days.

Q: Who can change permissions on a topic?

Only the owner of the topic can change permissions for that topic.

Q: How can users verify that notification messages are sent from Amazon SNS?

To ensure the authenticity of the notifications, Amazon SNS will sign all notification deliveries using a cryptographically secure, asymmetric mechanism (private-public key pair based on certificates). Amazon SNS will publish its certificate to a well-known location (e.g. http://sns.us-east-1.amazonaws.com/SimpleNotificationService.pem for the US East region) and sign messages with the private key of that certificate. Developers/applications can obtain the certificate and validate the signature in the notifications with the certificate’s public key, to ensure that the notification was indeed sent out by Amazon SNS. For further details on certificate locations, please refer to the Amazon SNS details page.

Q: Do publishers have to sign messages as well?

Amazon SNS requires publishers with AWS IDs to validate their messages by signing messages with their secret AWS key; the signature is then validated by Amazon SNS.

Q: Can a publisher/subscriber use SSL to secure messages?

Yes, both publishers and subscribers can use SSL to help secure the channel to send and receive messages. Publishers can connect to Amazon SNS over HTTPS and publish messages over the SSL channel. Subscribers should register an SSL-enabled end-point as part of the subscription registration, and notifications will be delivered over a SSL channel to that end-point.

Q: What permissions does a subscriber need to allow Amazon SNS to send notifications to a registered endpoint?

The owner of the end-point receiving the notifications has to grant permissions for Amazon SNS to send messages to that end-point.

Q: How can subscriptions be unsubscribed?

Subscribers can be unsubscribed either by the topic owner, the subscription owner or others – depending on the mechanism used for confirming the subscription request.

- A subscription that was confirmed with the AuthenticateOnUnsubscribe flag set to True in the call to the ConfirmSubscription API call can only be unsubscribed by a topic owner or the subscription owner.

- If the subscription was confirmed anonymously without the AuthenticateOnUnsubscribe flag set to True, then it can be anonymously unsubscribed.

In all cases except when unsubscribed by the subscription owner, a final cancellation message will be sent to the end-point, allowing the endpoint owner to easily re-subscribe to the topic (if the Unsubscribe request was unintended or in error). For further details on the ConfirmSubscription API, please refer to the Amazon SNS documentation.

Reliability

Q: How reliable is my data once published to Amazon SNS?

Amazon SNS stores all topic and message information within Amazon’s proven network infrastructure and datacenters. All messages are stored redundantly on multiple servers and in multiple data centers, which means that no single computer or network failure renders Amazon SNS inaccessible.

Q: Will a notification contain more than one message?

No, all notification messages will contain a single published message.

Q: How many times will a subscriber receive each message?

Although most of the time each message will be delivered to your application exactly once, the distributed nature of Amazon SNS and transient network conditions could result in occasional, duplicate messages at the subscriber end. Developers should design their applications such that processing a message more than once does not create any errors or inconsistencies.

Q: Will messages be delivered to me in the exact order they were published?

The Amazon SNS service will attempt to deliver messages from the publisher in the order they were published into the topic. However, network issues could potentially result in out-of-order messages at the subscriber end.

Q: Can a message be deleted after being published?

No, once a message has been successfully published to a topic, it cannot be recalled.

Q: Will Amazon SNS guarantee that messages will be delivered to the subscribed end-point?

When a message is published to a topic, Amazon SNS will attempt to deliver notifications to all subscribers registered for that topic. Due to potential Internet issues or Email delivery restrictions, sometimes the notification may not successfully reach an HTTP or Email end-point. In the case of HTTP, an SNS Delivery Policy can be used to control the retry pattern (linear, geometric, exponential backoff), maximum and minimum retry delays, and other parameters. If it is critical that all published messages be successfully processed, developers should have notifications delivered to an SQS queue (in addition to notifications over other transports).

Worldwide SMS

Q: What features are part of the new Worldwide SMS capability?

You can use Amazon SNS to deliver SMS (text) messages to 200+ countries, and you do not require recipients to explicitly opt in as before. You must obtain prior permission from recipients to send SMS messages to their phone numbers, where required by local law and regulations. Additionally, you can now mark your SMS messages as Transactional to optimize for reliable delivery, or you can mark it as Promotional to optimize for cost savings. Furthermore, you can set account and message-level spend limits to avoid inadvertent overruns.

Q: When should I mark an SMS message as Transactional?

SMS messages that are of high priority to your business should be marked as Transactional. This ensures that messages such as those that contain one-time passwords (OTP) or PINs get delivered over routes with the highest delivery reliability. These routes tend to be more expensive than Promotional messaging routes in countries other than the US. You should never mark marketing messages as Transactional, because this violates the local regulatory policies in certain countries, and your account may be marked for abuse and suspended.

Q: When should I mark an SMS message as Promotional?

SMS messages that carry marketing messaging should be marked Promotional. Amazon SNS ensures that such messages are sent over routes that have a reasonable delivery reliability but are substantially cheaper than the most reliable routes. This also allows Amazon SNS to handle and deliver your messages in compliance with on local laws and regulation

Q: What are account-level and message-level spend limits and how do they work?

Spend limits can be specified for an AWS account and for individual messages, and the limits apply only to the cost of sending SMS messages. The default spend limit per account (if not specified) is 50.00 USD per month. If you need a larger spend limit, please fill out a limit increase request here. Amazon SNS sends SMS messages that you publish while the total cost incurred for your SMS traffic is below your spend limit for that calendar month. Once the spend limit is exceeded, Amazon SNS stops delivering messages until you either increase the spend limit or a new calendar month begins. Similarly, you can also specify a spend limit for an individual message, and Amazon SNS will send the message only if the cost is below the limit. Amazon SNS will not send your SMS messages if the account-level spend limit is exceeded, regardless of whether the message-level spend limit is exceeded.

Q: Is two-way SMS supported?

Amazon SNS does not currently support two-way SMS capabilities, except for opt out where required by local regulations.

Q: Do I need to subscribe phone numbers to an SNS Topic before sending an SMS message to it?

You no longer need to subscribe a phone number to an Amazon SNS topic before you publish messages to it. Now, you can directly publish messages to a phone number using the Amazon SNS console or the Publish request in the Amazon SNS API.

Q: Does AWS offer long codes or short codes for purchase?

AWS does not currently offer long codes or short codes for purchase.

Q: Will SMS notifications come from a specific number of short codes or long codes?

Amazon SNS uses a pool of long codes or short codes to send SMS notifications out. So while there is a possibility that SMS notifications come from multiple numbers, Amazon SNS ensures that the messages sent from an AWS account to a specific phone number always come from the same long code or short code. This is called "Sticky Sender ID".

Q: Which countries does Amazon SNS support for Worldwide SMS?

Amazon SNS supports more than 200 countries, and we keep growing our reach. Please refer to the SMS Supported Country List for a comprehensive list of supported calling countries.

Q: Which AWS regions support Worldwide SMS?

1) US-East-1 (Virginia), 2) US-West-2 (Oregon), 3) EU-West-1 (Dublin), 4) Asia Pacific (Tokyo), 5) Asia Pacific (Singapore), and 6) Asia Pacific (Sydney).

Q: Do the AWS phone numbers change?

Yes. Amazon SNS uses a pool of long codes or short codes to send SMS notifications. So while there is a possibility that SMS notifications come from multiple numbers, Amazon SNS ensures that the messages sent from an AWS account to a specific phone number, always come from the same long code or short code. This is called "Sticky Sender ID".

Q: Why do some devices on the same carrier receive messages from different phone numbers?

Amazon SNS uses a pool of long codes or short codes to send SMS notifications. So while there is a possibility that SMS notifications come from multiple numbers, Amazon SNS ensures that the messages sent from an AWS account to a specific phone number always come from the same long code or short code. This is called "Sticky Sender ID".

Q: What is the phone number format for sending messages to other countries?

AWS strongly encourages E.164 number formatting for all phone numbers both in the ‘to’ and ‘from’ (when applicable) fields. Please refer to the SMS Supported Country List for a comprehensive list of supported countries.

Q: Does Amazon SNS determine if a phone number is a mobile, landline, or VoIP number?

No. Currently, Amazon SNS does not detect whether a phone number is mobile, landline, or VoIP.

Q: Is time-based or scheduled delivery supported for SMS messages?

No. Amazon SNS does not currently support time-based or scheduled delivery.

Q: How do I track the delivery status of my SMS messages?

By enabling the Delivery Status feature in Amazon SNS, you can get information on the following for each message: MessageID, Time Sent, Destination Phone Number, Disposition, Disposition Reason (if applicable), Price, and Dwell Time.

Q: Do you support MMS ?

No. Currently Amazon SNS does not support MMS messages.

Q: What is the cost of receiving SMS messages from Amazon SNS?

Costs for receiving SMS messages depend on the Data and Messaging of the recipient's wireless / mobile carrier plans.

Q: How do recipients opt out from receiving SMS messages from AWS?

To opt out, recipients reply back with the message "STOP" to the same short or long code that originated the message. Once opted out, recipients will stop receiving any SMS message that is sent from your AWS account.

Q: How do I know if a recipient device has ‘opted out’ of Global SMS?

The SNS console displays the list of opted out numbers for your account. Additionally, the Amazon SNS API provides the ListPhoneNumbersOptedOut request for listing opted out phone numbers.

Q: If a user opts out, will that number be unsubscribed automatically from the SNS Topic?

No. Opt-outs do not unsubscribe a number from an Amazon SNS topic, but rather disable the subscription. This means if you opt-in a phone number you do not need to re-subscribe the phone number to the topic.

Q: How do I confirm the end user received the SMS message?

You can use our Delivery Status feature to get information on the final disposition of your SMS message. For more information on the feature and how to use it, please refer to our documentation.

Q: Does Amazon SNS provide delivery receipts for SMS messages?

Our Delivery Status feature provides information based on delivery receipts received from the destination carrier. For more information on the Delivery Status feature and how to use it, please refer to our documentation.

Q: Does SMS support delivery to VoIP services like Google Voice or Hangouts?

Yes. Amazon SNS does support delivery to VoIP services that can receive SMS messages.

SMS Pricing

Q: Where can I find the current SMS pricing per country?

Our pricing is based on destination country and carrier, and it is providedhere.

Q: Why does pricing for SMS keeps changing for the same destination country and carrier?

Pricing in the SMS industry is not static and the cost of sending to different countries and carriers in those countries tend to vary over time. Amazon SNS has adopted a transparent approach and has exposed those price variations to customers, so that you get the maximum cost benefits.

Q: Will I be charged for failed deliveries or messages rejected by carriers?

You may be charged for failed deliveries if the destination carrier reports back that you attempted to send messages to an invalid phone number. Phone numbers can be invalid for a number of reasons, such as the phone number doesn't exist, the phone holder's account does not have sufficient credits, or if the destination number is a landline number.

Q: Is there a 'free tier' allowance for SMS messages?

There is a monthly free tier allowance for SMS messages. The first 100 SMS messages sent to US phone numbers each month are free. Additional SMS messages to the US or any messages sent to non-US phone numbers are charged based on current pricing provided here.

Limits and Restrictions

Q: Are there limits to the number of topics or number of subscribers per topic?

By default, SNS offers 10 million subscriptions per topic, and 100,000 topics per account. To request a higher limit, please contact us at at http://aws.amazon.com/support

Q: How much and what kind of data can go in a message?

Amazon SNS messages can contain up to 256 KB of text data, including XML, JSON and unformatted text. The following Unicode characters are accepted:

> #x9 | #xA | #xD | [#x20 to #xD7FF] | [#xE000 to #xFFFD] | [#x10000 to #x10FFFF]
>
> (according to http://www.w3.org/TR/REC-xml/#charsets).

Note: For SMS subscriptions, only 140 characters will be included in the payload of the message delivered. This will include the DisplayName of the topic and as many characters of the published message as can be accomodated.

Note: Each 64KB chunk of published data is billed as 1 request. For example, a single API call with a 256KB payload will be billed as four requests.

Q: Are there TCP ports that should be used for cross-region communication between SNS and EC2?

Yes, cross-region communication between SNS and EC2 on ports other than 80/443/4080/8443 is not guaranteed to work and should be avoided.

Raw Message Delivery

Q: What is raw message delivery?

You can now opt-in to get your messages delivered in raw form, i.e. exactly as you published them. By default, messages are delivered encoded in JSON that provides metadata about the message and topic. Raw message delivery can be enabled by setting the “RawMessageDelivery” property on the subscriptions. This property can be set by using the AWS Management Console, or by using the API SetSubscriptionAttributes.

Q: What is the default behavior if the raw message delivery property on the subscription is not set?

By default, if this property is not set, messages will be delivered in JSON format, which is the current behavior. This ensures existing applications will continue to operate as expected.

Q: Which types of endpoints support raw message delivery?

New raw message delivery support is added to endpoints of type SQS Queue and HTTP(S). Deliveries to Email and SMS endpoints will behave the same independent of the “RawMessageDelivery” property.

Q: How will raw messages be delivered to HTTP endpoints?

When raw-formatted messages are delivered to HTTP/s endpoints, the message body will be included in the body of the HTTP POST.

Mobile Push Notifications

Q: What is SNS Mobile Push?

SNS Mobile Push lets you use Simple Notification Service (SNS) to deliver push notifications to Apple, Google, Fire OS, and Windows devices, as well as Android devices in China with Baidu Cloud Push. With push notifications, an installed mobile application can notify its users immediately by popping a notification about an event, without opening the application. For example, if you install a sports app and enable push notifications, the app can send you the latest score of your favorite team even if the app isn’t running. The notification appears on your device, and when you acknowledge it, the app launches to display more information. Users’ experiences are similar to receiving an SMS, but with enhanced functionality and at a fraction of the cost.

Q: How do I get started sending push notifications?

Push notifications can only be sent to devices that have your app installed, and whose users have opted in to receive them. SNS Mobile Push does not require explicit opt-in for sending push notifications, but iOS, Android and Kindle Fire operating systems do require it. In order to send push notifications with SNS, you must also register your app and each installed device with SNS. For more information, see Using Amazon SNS Mobile Push Notifications.

Q: Which push notifications platforms are supported?

Currently, the following push notifications platforms are supported:

- Amazon Device Messaging (ADM)
- Apple Push Notification Service (APNS)
- Google Cloud Messaging (GCM)
- Windows Push Notification Service (WNS) for Windows 8+ and Windows Phone 8.1+
- Microsoft Push Notification Service (MPNS) for Windows Phone 7+
- Baidu Cloud Push for Android devices in China

Q: How many push notifications can I send with the SNS Free Tier?

The SNS free tier includes 1 million publishes, plus 1 million mobile push deliveries. So you can send 1 million free push notifications every month. Notifications to all mobile push endpoints are all counted together toward your 1 million free mobile push deliveries.

Q: Does enabling push notifications require any special confirmations with SNS Mobile Push?

No, they do not. End-users opt-in to receive push notifications when they first run an app, whether or not SNS delivers the push notifications.

Q: Do I have to modify my client app to use SNS Mobile Push?

SNS does not require you to modify your client app. Baidu Cloud Push requires Baidu-specific components to be added to your client code in order to work properly, whether or not you choose to use SNS.

Q: How do SNS topics work with Mobile Push?

SNS topics can have subscribers from any supported push notifications platform, as well as any other endpoint type such as SMS or email. When you publish a notification to a topic, SNS will send identical copies of that message to each endpoint subscribed to the topic. If you use platform-specific payloads to define the exact payload sent to each push platform, the publish will fail if it exceeds the maximum payload size imposed by the relevant push notifications platform.

Q: What payload size is supported for various target platforms?

SNS will support maximum payload size that is supported by the underlying native platform. Customers can use a JSON object to send platform specific messages. See Using SNS Mobile Push API for additional details.

Q: How do platform-specific payloads work?

When you publish to a topic and want to have customized messages sent to endpoints for the different push notification platforms then you need to select “Use different message body for different protocols” option on the Publish dialog box and then update the messages. You can use platform-specific payloads to specify the exact API string that is relayed to each push notifications service. For example, you can use platform-specific payloads to manipulate the badge count of your iOS application via APNS. For more information, see Using Amazon SNS Mobile Push Notifications.

Q: Can one token subscribe to multiple topics?

Yes. Each token can be subscribed to an unlimited number of SNS topics.

Q: What is direct addressing? How does it work?

Direct addressing allows you to deliver notifications directly to a single endpoint, rather than sending identical messages to all subscribers of a topic. This is useful if you want to deliver precisely targeted messages to each recipient. When you register device tokens with SNS, SNS creates an endpoint that corresponds to the token. You can publish to the token endpoint just as you would publish to a topic. You can direct publish either the text of your notification, or a platform-specific payload that takes advantage of platform-specific features such as updating the badge count of your app. Direct addressing is currently only available for push notifications endpoints.

Q: Does SNS support direct addressing for SMS or Email?

At this time, direct addressing is only supported for mobile push endpoints (APNS, GCM, ADM, WNS, MPNS, Baidu) and SMS. Email messaging requires the use of topics.

Q: How does SNS Mobile Push handle token feedback from notification services?

Push notification services such as APNS and GCM provide feedback on tokens which may have expired or may have been replaced by new tokens. If either APNS or GCM reports that a particular token has either expired or is invalid, SNS automatically "disables" the application endpoint associated with the token, and notifies you of this change via an event.GCM specifically, at times not only indicates that a token is invalid, but also provides the new token associated with the application endpoint in its response to SNS. When this happens, SNS automatically updates the associated endpoint with the new token value, leaving the endpoint enabled, and then notifies you of this change via an event.

Q: Can I migrate existing apps to SNS Mobile Push?

Yes. You can perform a bulk upload of existing device tokens to Amazon SNS, either via the console interface or API. You would also register your app with SNS by uploading your credentials for the relevant push notifications services, and configure your proxy or app to register future new tokens with SNS.

Q: Can I monitor my push notifications through Amazon CloudWatch?

Yes. SNS publishes Cloudwatch metrics for number of messages published, number of successful notifications, number of failed notifications and size of data published. Metrics are available on per application basis. You can access Cloudwatch metrics via AWS Management Console or CloudWatch APIs.

Q: What types of Windows Push Notifications does Amazon SNS support?

SNS supports all types of push notifications types offered by Microsoft WNS and MPNS, including toast, tile, badge and raw notifications. Use the TYPE message attribute to specify which notification type you wish to use. When you use default payloads to send the same message to all mobile platforms, SNS will select toast notifications by default for Windows platforms. It is required to specify a notification type for Windows platforms when you use platform-specific payloads.

Q: Does SNS support Windows raw push notifications?

Yes. You must encode the notification payload as text to send raw notifications via SNS.

Q: What is Baidu Cloud Push?

Baidu Cloud Push is a third-party alternative push notifications relay service for Android devices. You can use Baidu Cloud Push to reach Android customers in China, no matter what Android app store those customers choose to use for downloading your app. For more information about Baidu Cloud Push, visit: http://developer.baidu.com/cloud/push.

Q: Can I publish Baidu notifications from all public AWS regions?

Yes, SNS supports Baidu push notifications from all public AWS regions.

Q: Can I use Baidu notifications to any Android app store?

Yes, Baidu push notifications work for apps installed via any Android app store.

Q: What are message attributes?

Message attributes allow you to provide structured metadata items (such as timestamps, geospatial data, signatures, and identifiers) about the message. Message attributes are optional and separate from, but sent along with, the message body. This information can be used by the receiver of the message to help decide how to handle the message without having to first process the message body.

You can use SNS message attributes in conjunction with SQS and mobile push endpoints. To learn more about message attributes, please see the SNS Getting Started Guide.

Q: What message attributes are supported in SNS?

SNS supports different message attributes for each endpoint type, depending on what the endpoint types each support themselves.

- **For SQS endpoints**, you can specify up to 10 name-type-value triples per message. Types supported include: String, Binary and Number (including integers, floating point, and doubles).
- **For mobile push endpoints**, you can take advantage of specific message attributes that each mobile platform supports (such as notification type).

Q: What is Time to Live (TTL)?

Some messages that you can send with SNS are relevant or valuable only for a limited period of time. Amazon SNS now allows you to set a TTL (Time to Live) value for each message. When the TTL expires for a given message that was not delivered and read by an end user, the message is deleted. TTL is specified in seconds and is relative to the time Publish call is made.

Q: How do I specify a TTL for my messages?

You can specify a TTL using the console or via API. TTL can be specified at publish time for a message, using the message attribute below. There is a different attribute for each platform. An attribute specified for a platform is applicable only for notification deliveries to that platform.

Q: What is the default TTL?

SNS uses a default Time to Live (TTL) of 4 weeks for all mobile platforms.

Q: Do TTL message attributes override TTLs specified in a message payload?

Yes. Google GCM and Amazon ADM allow you to specify a TTL within the message payload. If you specify TTL within the message payload and also within a message attribute, SNS will follow the message attribute.

Q: What happens if I specify TTL=0?

Some platforms treat TTL = 0 as a special case and attempt to deliver the message immediately, else let it expire. If you specify TTL = 0, SNS will relay your message to the appropriate service with TTL = 0 in order to take advantage of this special case.

Q: What SNS endpoints support TTL?

You can use TTL with the following mobile push endpoints: APNS, APNS_Sandbox, GCM, ADM, Baidu, and WNS. Microsoft MPNS does not currently support TTL. TTL is also not supported for SQS, HTTP, email or SMS endpoints.

Q: What does the Delivery Status feature of Amazon SNS do?

The Delivery Status feature lets you collect information on success rates, failure rates and dwell times of your push notifications for the supported mobile notification platforms. The currently supported platforms include Apple (APNS), Google (GCM), Windows (WNS and MPNS), Amazon (ADM), and Baidu. The status information is captured in the Amazon CloudWatch log groups created by Amazon SNS on your behalf. Additionally, you can create actionable metrics in Amazon CloudWatch and trigger alarms based on the patterns you are interested in.

Q: Is the Delivery Status feature in Amazon SNS available only for mobile push notifications? Do you plan to support this feature for other endpoint types?

Currently the Delivery Status feature is available for mobile push notifications and SMS. We will evaluate extending this to other endpoint types based on feedback from customers.

Q: How do I activate the Delivery Status feature?

You can activate the Delivery Status feature from the Amazon SNS console. From your Application, choose the Delivery Status option in the Application Actions drop-down menu. For details, please read our documentation.

Q: Can I activate the Delivery Status feature from the Amazon SNS APIs?

Yes, you can activate this feature from Amazon SNS APIs by adding the relevant application-level attributes. Our documentation goes over the application-level attributes that you need to add and the specific API calls that need to be made to enable this feature.

Q: How much does the Delivery Status feature cost?

There is currently no additional Amazon SNS charge for using the Delivery Status feature. However, depending upon your usage, you may incur charges for using CloudWatch since this feature creates Amazon CloudWatch log groups. Read our pricing page for more information about CloudWatch pricing and free tier.

Q: Why can you only choose a sampling percentage for successful delivery attempts and not sample failed delivery attempts?

Based on feedback we received from customers, we found that most developers are interested in knowing all the delivery attempt failures for their applications – and prefer to only store sample successful deliveries rather than logging all of them.

Q: How can I set alarms based on failure metrics or dwell time metrics?

After activating the Delivery Status feature, you need to define a Log Metrics Filter in Amazon CloudWatch Logs for the log group that gets created by Amazon SNS on your behalf. This metrics filter can be defined to extract information that you are interested in, such as failure rate and dwell time. Once a Metric Filter is defined, you can create it and assign it to a Metric. This metric can then be used to set alarms or send notifications based on thresholds you define. For more information, take a look at our documentation or blog.

SNS Support for AWS Lambda

Q: What does support for AWS Lambda endpoints in Amazon SNS mean?

You can now invoke your AWS Lambda functions by publishing messages to Amazon SNS topics that have AWS Lambda functions subscribed to them. Because Amazon SNS supports message fan-out, publishing a single message can invoke different AWS Lambda functions or invoke Lambda functions in addition to delivering notifications to supported Amazon SNS destinations such as mobile push, HTTP endpoints, SQS, email and SMS (US only).

Q: What is AWS Lambda?

AWS Lambda is a compute service that runs your code in response to events and automatically manages the compute resources for you, making it easy to build applications that respond quickly to new information. More information on AWS Lambda and how to create AWS Lambda functions can be found here.

Q: What can I do with AWS Lambda functions and Amazon SNS?

By subscribing AWS Lambda functions to Amazon SNS topics, you can perform custom message handling. You can invoke an AWS Lambda function to provide custom message delivery handling by first publishing a message to an AWS Lambda function, have your Lambda function modify a message (e.g. localize language) and then filter and route those messages to other topics and endpoints. Apps and services that already send Amazon SNS notifications, such as Amazon CloudWatch, can now immediately take advantage of AWS Lambda without having to provision or manage infrastructure for custom message handling. You can also use delivery to an AWS Lambda function as a way to publish to other AWS services such as Amazon Kinesis or Amazon S3. You can subscribe an AWS Lambda function to the Amazon SNS topic, and then have the Lambda function in turn write to another service.

Q: How do I activate AWS Lambda endpoint support in Amazon SNS?

You need to first create an AWS Lambda function via your AWS account and theAWS Lambda console, and then subscribe that AWS Lambda function to a topic using theAmazon SNS console or the Amazon SNS APIs. Once that is complete, any messages that you publish to the Amazon SNS topics which have Lambda functions subscribed to them will be delivered to the appropriate Lambda functions in addition to any other destinations subscribed to that topic.

Q: What does delivery of a message from Amazon SNS to an AWS Lambda function do?

A message delivery from Amazon SNS to an AWS Lambda function creates an instance of the AWS Lambda function and invokes it with your message as an input. For more information on message formats, please refer to the Amazon SNS documentation and the AWS Lambda documentation.

Q: How much does this feature cost?

Publishing a message with Amazon SNS costs $0.50 per million requests. Aside from charges incurred in using AWS services, there are no additional fees for delivering a message to an AWS Lambda function. Amazon SNS has a Free Tier of 1 million requests per month. For more information, please refer to Amazon SNS pricing. AWS Lambda function costs are based on the number of requests for your functions and the time your code executes. The AWS Lambda Free-Tier includes 1M requests per month and 400,000 GB-seconds of compute time per month. For more information, please refer to AWS Lambda pricing.

Q: Can I subscribe AWS Lambda functions created by someone else to Amazon SNS topics that I own?

We currently do not allow an AWS account owner to subscribe an AWS Lambda function that belongs to another account. You can subscribe your own AWS Lambda functions to your own Amazon SNS topics or subscribe your AWS Lambda functions to an Amazon SNS topic that was created by another account so long as the topic policy for that SNS topic allows it.

Q: Is there a limit to the number of AWS Lambda functions that I can subscribe to an Amazon SNS topic?

Amazon SNS treats AWS Lambda functions like any other destination. By default, SNS offers 10 million subscriptions per topic. To request a higher limit, please contact us.

Q: What data can I pass to my AWS Lambda function?

When an AWS Lambda function is invoked as a result of an Amazon SNS message delivery, the AWS Lambda function receives data such as the Message ID, the topic ARN, the message payload and message attributes via an SNS Event. For more information on the event structure passed to the AWS Lambda function please read our blog.

Q: Can I track delivery status for message delivery attempts to AWS Lambda functions?

To track the success or failure status of message deliveries, you need to activate the Delivery Status feature of Amazon SNS. For more information about how to activate this feature please read our blog.

Q: What regions is AWS Lambda available in?

AWS Lambda is currently available in US East (N. Virginia), US West (Oregon) and EU (Ireland).

Q: Do my AWS Lambda functions need to be in the same region as my Amazon SNS usage?

You can subscribe your AWS Lambda functions to an Amazon SNS topic in any region.

Q: Are there any data transfer costs for invoking AWS Lambda functions?

Data transfer costs are applicable to message deliveries to AWS Lambda functions. Please refer to our pricing for more information.

Q: Are there any limits to the concurrency of AWS Lambda functions?

AWS Lambda currently supports 100 concurrent requests per AWS account. If your Amazon SNS message deliveries to AWS Lambda contribute to crossing these concurrency limits, your Amazon SNS message deliveries will be throttled. If AWS Lambda throttles an Amazon SNS message, Amazon SNS will retry the delivery attempts. For more information about AWS Lambda concurrency limits, please refer to AWS Lambda documentation.

Q: Can Amazon SNS use the same AWS Lambda functions that I use with other services (e.g. Amazon S3)?

You can use the same AWS Lambda functions that you use with other services as long as the same function can parse the event formats from Amazon SNS in addition to the event format of the other services. For the SNS event format please read our blog.

VoIP iOS and Mac OS Notifications

Q: What are VoIP Push Notifications for iOS?

In iOS 8 and later, voice-over-IP (VoIP) apps can register for VoIP remote notifications such that iOS can launch or wake the app, as appropriate, when an incoming VoIP call arrives for the user. The procedure to register for VoIP notifications is similar to registering for regular push notifications on iOS. For more information, please refer to our documentation.

Q: Can I use VoIP Push Notifications and other Push Notifications in the same iOS app?
Yes, you can have an iOS application that is registered to receive both types of push notifications. However, you will need to obtain the VoIP push notification certificate from Apple in addition to the regular push notification certificate and create a new Platform Application in Amazon SNS and choose Apple VoIP Push as the platform type. For more information, please refer to our documentation.

Q: What are Mac OS push notifications?
You can now send push notifications to Mac OS desktops that run Mac OS X Lion (10.7) or later using Amazon SNS. For more information, please refer to our documentation.

Amazon SQS FAQTechnical FAQ

Overview

Q: What is Amazon SQS?

Amazon Simple Queue Service (Amazon SQS) is a web service that gives you access to message queues that store messages waiting to be processed. With Amazon SQS, you can quickly build message queuing applications that can run on any computer.

Amazon SQS offers a reliable, highly-scalable, hosted queue for storing messages in transit between computers. With Amazon SQS, you can move data between diverse, distributed application components without losing messages and without requiring each component to be always available.

Amazon SQS can help you build a distributed application with decoupled components, working closely with the Amazon Elastic Compute Cloud (Amazon EC2) and the other AWS infrastructure web services.

Q: What can I do with Amazon SQS?

Because Amazon SQS is highly-scalable and you pay only for what you use, you can start small and grow your application alongside your business needs, with no performance or reliability compromises. Amazon SQS lets you stop worrying about how your messages are stored and managed and helps you focus on building robust, sophisticated message-based applications.

Here are just a few ideas:

- Integrate Amazon SQS with other AWS services to make applications more flexible and reliable.

- Use Amazon SQS to create work queues with each message as a task to be completed by a process. Let one (or many) computers read tasks from the message queue and process them.
- Build a microservice architecture and use message queues to connect your microservices.
- Keep notifications of significant business events in an Amazon SQS message queue. Each event can have a corresponding message in a message queue, and applications that need to be aware of the event can read and process the messages.

Q: How can I get started using Amazon SQS?

You can get started with Amazon SQS in a few steps:

1. Register for an AWS account.
2. After signing up, visit the AWS Management Console, select **SQS**, and on the next page select **Create New Queue.**
3. In the **Create New Queue** dialog box, enter a **Queue Name** (for example, MyQueue), and click **Create Queue.**
4. Select a queue, and from the **Queue Actions** drop-down list select **Send a Message.**
5. In the **Send a Message to MyQueue** dialog box, on the **Message Body** tab, enter the text of your message and click **Send Message.**

The message is sent and a confirmation with the sent message attributes is displayed. Click **Close** to finish.

For more information, see the *Amazon SQS Getting Started Guide*, the *Amazon SQS Developer Guide*, and sample code in the Resource Center.

Q: What are the benefits of Amazon SQS over homegrown or packaged message queuing systems?

Using Amazon SQS provides several advantages over building your own software for managing message queues or using commercial or open-source message queuing systems that require significant up-front time for development and configuration.

These alternatives require ongoing hardware maintenance and system administration resources. The complexity of configuring and managing these systems is compounded by the need for redundant storage of messages that ensures messages are not lost if hardware fails.

In contrast, Amazon SQS requires no administrative overhead and little configuration. Moreover, Amazon SQS works on a massive scale, processing billions of messages per day. You can scale the amount of traffic you send to Amazon SQS up or down without any configuration.

Amazon SQS also provides extremely high message durability, giving you and your stakeholders added confidence.

Q: When should I use Amazon Simple Workflow Service (Amazon SWF) instead of Amazon SQS?

You can use either Amazon SWF or Amazon SQS to develop distributed, decoupled applications:

- Amazon SWF is a service designed for orchestrating highly-scalable applications; it also provides auditability.
- Amazon SQS provides a reliable, highly-scalable, hosted queue for sending and receiving messages.

While you can use Amazon SQS to build basic workflows to coordinate your distributed application, you can get this facility out-of-the-box with Amazon SWF, alongside other application-level capabilities.

We recommend trying both Amazon SQS and Amazon SWF to determine which solution best fits your needs.

Q: Does Amazon use Amazon SQS for its own applications?

Yes. Developers at Amazon use Amazon SQS for a variety of applications that process large numbers of messages every day. Key business processes in both Amazon.com and Amazon Web Services use Amazon SQS.

Billing

Q: What can I do with the Amazon SQS Free Tier?

The Amazon SQS Free Tier provides you with 1 million requests per month at no charge.

Many small-scale applications are able to operate entirely within the limits of the Free Tier. However, data transfer charges might still apply. For more information, see Amazon SQS Pricing.

The Free Tier is a monthly offer. Free usage does not accumulate across months.

Q: How much does Amazon SQS cost?

You pay only for what you use, and there is no minimum fee.

In most regions, the cost of Amazon SQS is $0.50 for every 1 million requests, plus data transfer charges for data transferred out of Amazon SQS (unless data is transferred to Amazon EC2 instances or to AWS Lambda functions within the same region). For more information, see Amazon SQS Pricing.

Q: Will I be charged for all Amazon SQS requests?

Yes. All Amazon SQS requests are chargeable, and they are billed at the same rate.

Q: Do Amazon SQS batch operations cost more than other requests?

No. Batch operations (that include SendMessageBatch, DeleteMessageBatch, and ChangeMessageVisibilityBatch) all cost the same as other Amazon SQS requests. By grouping messages into batches, you can reduce your Amazon SQS costs.

Q: How will I be charged and billed for my use of Amazon SQS?

There are no initial fees to begin using Amazon SQS. At the end of the month, your credit card will be automatically charged for the month's usage.

You can view your charges for the current billing period at any time on the AWS website:

1. Log into your AWS account.
2. Under **Your Web Services Account**, select **Account Activity**.

Q: Do your prices include taxes?

Except as noted otherwise, our prices do not include any applicable taxes and duties such as VAT or applicable sales tax.

For customers with a Japanese billing address, the use of AWS in any region is subject to Japanese Consumption Tax. For more information, see the Amazon Web Services Consumption Tax FAQ.

Features, Functionality, and Interfaces

Q: Can I use Amazon SQS with other AWS services?

Yes. You can make your applications more flexible and scalable by using Amazon SQS with compute services such as Amazon EC2, Amazon EC2 Container Service (Amazon ECS), and AWS Lambda, as well as with storage and database services such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

One common use case is a distributed, decoupled application whose multiple components and modules need to communicate with each other, but can’t do the same amount of work simultaneously. In this case, Amazon SQS message queues carry messages to be processed by the application running on Amazon EC2 instances.

The Amazon EC2 instances can read the message queue, process the job, and then post the results as messages to another Amazon SQS message queue (for example, for further processing by another application). Because Amazon EC2 allows applications to scale up and down dynamically, application developers can vary the number of compute instances based on the amount of messages in the Amazon SQS queues using Auto Scaling, to ensure that jobs are executed in a timely manner.

Q: Can you give me an example use case for Amazon SQS?

Here is how a video transcoding website uses Amazon EC2, Amazon SQS, Amazon S3, and Amazon DynamoDB together:

1. End users submit videos to be transcoded to the website.
2. The videos are stored in Amazon S3, and a request message is placed in an incoming Amazon SQS queue with a pointer to the video and to the target video format within the message.
3. The transcoding engine that runs on a set of Amazon EC2 instances reads the request message from the incoming queue, retrieves the video from Amazon S3 using the pointer, and transcodes the video into the target format.
4. The converted video is put back into Amazon S3 and another response message is placed in another outgoing Amazon SQS queue with a pointer to the converted video.
5. At the same time, metadata about the video (format, date created, length, and so on) is indexed into Amazon DynamoDB for querying.

During this workflow, a dedicated Auto Scaling instance can constantly monitor the incoming queue. Based on the number of messages in the incoming queue, the Auto Scaling instance dynamically adjusts the number of transcoding Amazon EC2 instances to meet the response time requirements of the website's customers.

Q: How do I interact with Amazon SQS?

You can access Amazon SQS using the AWS Management Console, which has a visual, web-based interface for setting up and managing Amazon SQS.

Amazon SQS also provides a web services API and it is integrated with the AWS SDKs, allowing you to work in the programming language of your choice.

Q: What are the available operations for message queues?

For information on message queue operations, see Amazon SQS Product Details.

Q: Who can perform operations on a message queue?

Only an AWS account owner (or an AWS account that the account owner has delegated rights to) can perform operations on an Amazon SQS message queue.

Q: Can I use Java Message Service (JMS) with Amazon SQS?

Yes. You can take advantage of the scale, low cost, and high availability of Amazon SQS without the worry and high overhead of running your own JMS cluster.

Amazon provides the Amazon SQS Java Messaging Library that implements the JMS 1.1 specification and uses Amazon SQS as the JMS provider. For more information, see Using JMS with Amazon SQS.

Q: How are messages identified in the system?

All messages have a global unique ID that Amazon SQS returns when the message is delivered to the message queue. The ID isn't required to perform any further actions on the message, but it is useful for tracking the receipt of a particular message in the message queue.

When you receive a message from the message queue, the response includes a receipt handle that you must provide when deleting the message.

Q: How are unsuccessfully-processed messages handled?

In Amazon SQS, you can use the API or the console to configure*dead letter queues*, which are queues that you configure to receive messages from other source queues.

Any queue can become a dead letter queue that will receive messages after a maximum number of processing attempts cannot be completed. You can use dead letter queues to isolate messages that can't be processed for later analysis.

For more information, see Using Amazon SQS Dead Letter Queues.

Q: Does Amazon SQS provide first-in-first-out (FIFO) access to messages?

Amazon SQS provides a loose-FIFO capability that attempts to preserve the order of messages. However, we have designed Amazon SQS to be massively scalable using a distributed architecture. Thus, we can't guarantee that you will always receive messages in the exact order you sent them (FIFO).

If your system requires the order of messages to be preserved, place sequencing information in each message so that messages can be ordered when they are received.

Q: Does Amazon SQS provide at-least-once delivery of messages?

Yes. Amazon SQS guarantees that each message is delivered at least once. Amazon SQS stores copies of your messages on multiple servers for redundancy and high availability. On rare occasions, one of the servers that stores a copy of a message might be unavailable when you receive or delete the message.

If this occurs, the copy of the message will not be deleted on that unavailable server, and you might get a copy of that message again when you receive messages (at-least-once delivery).

You must design your applications to be *idempotent* (that is, they must not be affected adversely when processing the same message more than once).

Q: What is a visibility timeout?

The visibility timeout is a period of time during which Amazon SQS prevents other consuming components from receiving and processing a message. For more information, see Visibility Timeout.

Q: How does Amazon SQS allow multiple readers to access the same message queue without losing messages or processing them multiple times?

Every Amazon SQS queue has a configurable visibility timeout. A message is not visible to any other reader for a designated amount of time when it is read from a message queue. As long as the amount of time it takes to process the message is less than the visibility timeout, every message is processed and deleted.

If the component processing of the message fails or becomes unavailable, the message again becomes visible to any component reading the message queue once the visibility timeout ends. This allows multiple components to read messages from the same message queue, each one working to process different messages.

Q: What is the maximum limit for message visibility?

The maximum visibility timeout for an Amazon SQS message is 12 hours.

Q: Does Amazon SQS support message metadata?

Yes. An Amazon SQS message can contain up to 10 metadata attributes. You can use message attributes to separate the body of a message from the metadata that describes it. This helps process and store information with greater speed and intelligence because your applications do not have to inspect an entire message before understanding how to process it.

Amazon SQS message attributes take the form of name-type-value triples. The supported types include string, binary, and number (including integer, floating-point, and double). For more information, see Using Amazon SQS Message Attributes.

Q: How can I determine the time-in-queue value?

To determine the time-in-queue value, you can request the SentTimestamp attribute when receiving a message. Subtracting that value from the current time results in the time-in-queue value.

Q: What is the typical latency for Amazon SQS?

Typical latencies for SendMessage, ReceiveMessage, and DeleteMessage API requests are in the tens or low hundreds of milliseconds.

Q: For anonymous access, what is the value of the SenderId attribute for a message?

When the AWS account ID is not available (for example, when an anonymous user sends a message), Amazon SQS provides the IP address.

Q: What is Amazon SQS long polling?

Amazon SQS long polling is a way to retrieve messages from your Amazon SQS queues. While the regular short polling returns immediately, even if the message queue being polled is empty, long polling doesn’t return a response until a message arrives in the message queue, or the long poll times out.

Long polling makes it inexpensive to retrieve messages from your Amazon SQS queue as soon as the messages are available. Using long polling might reduce the cost of using SQS, because you can reduce the number of empty receives. For more information, see Amazon SQS Long Polling.

Q: Is there an additional charge for using Amazon SQS long polling?

No. Long-polling ReceiveMessage calls are billed exactly the same as short-polling ReceiveMessage calls.

Q: When should I use Amazon SQS long polling, and when should I use Amazon SQS short polling?

In almost all cases, Amazon SQS long polling is preferable to short polling. Long-polling requests let your queue consumers receive messages as soon as they arrive in your queue while reducing the number of empty ReceiveMessageResponse instances returned.

Amazon SQS long polling results in higher performance at reduced cost in the majority of use cases. However, if your application expects an immediate response from a ReceiveMessage call, you might not be able to take advantage of long polling without some application modifications.

For example, if your application uses a single thread to poll multiple queues, switching from short polling to long polling will probably not work, because the single thread will wait for the long-poll timeout on any empty queues, delaying the processing of any queues that might contain messages.

In such an application, it is a good practice to use a single thread to process only one queue, allowing the application to take advantage of the benefits that Amazon SQS long polling provides.

Q: What value should I use for my long-poll timeout?

In general, you should use maximum 20 seconds for a long-poll timeout. Because higher long-poll timeout values reduce the number of empty ReceiveMessageResponse instances returned, try to set your long-poll timeout as high as possible.

If the 20-second maximum doesn't work for your application (see the example in the previous question), set a shorter long-poll timeout, as low as 1 second.

All AWS SDKs work with 20-second long polls by default. If you don't use an AWS SDK to access Amazon SQS, or if you configured your AWS SDK to specifically have a shorter timeout, you might need to modify your Amazon SQS client to allow longer requests or to use a shorter long-poll timeout.

Q: What is the AmazonSQSBufferedAsyncClient for Java?

The AmazonSQSBufferedAsyncClient for Java provides an implementation of the AmazonSQSAsyncClient interface and adds several important features:

- Automatic batching of multiple SendMessage, DeleteMessage, or ChangeMessageVisibility requests without any required changes to the application

- Prefetching of messages into a local buffer that allows your application to immediately process messages from Amazon SQS without waiting for the messages to be retrieved

Working together, automatic batching and prefetching increase the throughput and reduce the latency of your application while reducing your costs by making fewer Amazon SQS requests. For more information, see Client-Side Buffering and Request Batching.

Q: Where can I download the AmazonSQSBufferedAsyncClient for Java?

You can download the AmazonSQSBufferedAsyncClient as part of theAWS SDK for Java.

Q: Do I have to rewrite my application to use the AmazonSQSBufferedAsyncClient for Java?

No. The AmazonSQSBufferedAsyncClient for Java is implemented as a drop-in replacement for the existing AmazonSQSAsyncClient.

If you update your application to use the latest AWS SDK and change your client to use the AmazonSQSBufferedAsyncClient for Java instead of the AmazonSQSAsyncClient, your application will receive the added benefits of automatic batching and prefetching.

Q: How can I subscribe Amazon SQS message queues to receive notifications from Amazon Simple Notification Service (Amazon SNS) topics?

1. In the Amazon SQS console, select an Amazon SQS message queue.
2. Under **Queue Actions**, select **Subscribe Queue to SNS Topic** from the drop-down list.
3. In the dialog box, select the topic from the**Choose a Topic** drop-down list, and click **Subscribe**.

For more information, see Subscribing a Queue to an Amazon SNS Topic.

Q: How can I fan-out identical messages to multiple Amazon SQS queues?

1. Use Amazon SNS to create a topic.
2. Create and subscribe multiple Amazon SQS message queues to the Amazon SNS topic.
3. Whenever a message is sent to the Amazon SNS topic, it is fanned out to the Amazon SQS message queues.

Amazon SNS delivers the message to all Amazon SQS message queues subscribed to the topic.

Q: Can I delete all messages in a message queue without deleting the message queue itself?

Yes. You can delete all messages in an Amazon SQS message queue using the PurgeQueue action.

When you purge a message queue, all the messages previously sent to the message queue are deleted. Because your message queue and its attributes remain, there is no need to reconfigure the message queue; you can continue using it.

To delete only specific messages, use the DeleteMessage or DeleteMessageBatch actions.

Security and Reliability

Q: How reliable is the storage of my data in Amazon SQS?

Amazon SQS stores all message queues and messages within a single, highly-available AWS region with multiple redundant Availability Zones (AZs), so that no single computer, network, or AZ failure can make messages inaccessible. For more information, see Regions and Availability Zones.

Q: How can I secure the messages in my message queues?

Authentication mechanisms ensure that messages stored in Amazon SQS message queues are secured against unauthorized access. You can control who can send messages to a message queue and who can receive messages from a message queue. For additional security, you can build your application to encrypt messages before they are placed in a message queue.

Amazon SQS has its own resource-based permissions system that uses policies written in the same language as AWS Identity and Access Management (IAM) policies: for example, you can use variables, just like in IAM policies. For more information, see Amazon SQS Policy Examples.

Amazon SQS supports the HTTP over SSL (HTTPS) and Transport Layer Security (TLS) protocols. Most clients can automatically negotiate to use newer versions of TLS without any code or configuration change. Amazon SQS supports versions 1.0, 1.1, and 1.2 of the Transport Layer Security (TLS) protocol in all regions.

Q: Why are there separate ReceiveMessage and DeleteMessage operations?

When Amazon SQS returns a message to you, the message stays in the message queue whether or not you actually receive the message. You are responsible for deleting the message and the deletion request acknowledges that you’re done processing the message.

If you don’t delete the message, Amazon SQS will deliver it again on when it receives another receive request. For more information, see Visibility Timeout.

Q: Can a deleted message be received again?

Yes. Under rare circumstances, you might receive a previously-deleted message a second time. This can happen in the rare situation when a DeleteMessage operation doesn’t delete all copies of a message because one of the servers in the distributed Amazon SQS system isn’t available at the time of deletion. This message copy can be delivered again.

To avoid this behavior, design your application to be *idempotent* (that is, no errors or inconsistencies occur if you receive a deleted message a second time).

Q: What happens if I issue a DeleteMessage request on a previously-deleted message?

When you issue a DeleteMessage request on a previously-deleted message, Amazon SQS returns a *success* response.

Compliance

Q: Is Amazon SQS PCI DSS certified?

Yes. Amazon SQS is PCI DSS Level 1 certified. For more information, seePCI Compliance.

Q: Is Amazon SQS HIPAA compliant?

No. Amazon SQS is not yet eligible for HIPAA compliance.

However, you can use the Extended Client Library to send Amazon SQS message payloads through Amazon S3 (Amazon S3 is an HIPAA-eligible service).You can achieve HIPAA compliance in this manner, because no personally identifiable information (PII) is transferred via Amazon SQS.

For more information, see Using the Amazon SQS Extended Client Library for Java.

Limits and Restrictions

Q: How long can I keep my messages in Amazon SQS message queues?

Longer message retention provides greater flexibility to allow for longer intervals between message production and consumption.

You can configure the Amazon SQS message retention period to a value from 1 minute to 14 days. The default is 4 days. Once the message retention limit is reached, your messages are automatically deleted.

Q: How do I configure Amazon SQS to support longer message retention?

To configure the message retention period, set the MessageRetentionPeriod attribute using the console or using the Distributiveness method. Use this attribute to specify the number of seconds a message will be retained in Amazon SQS.

You can use the MessageRetentionPeriod attribute to set the message retention period from 60 seconds (1 minute) to 1,209,600 seconds (14 days). For more information on working with this message attribute, see the *Amazon SQS API Reference*.

Q: How do I configure the maximum message size for Amazon SQS?

To configure the maximum message size, use the console or the SetQueueAttributes method to set the MaximumMessageSize attribute. This attribute specifies the limit on bytes that an Amazon SQS message can contain. Set this limit to a value between 1,024 bytes (1 KB), and 262,144 bytes (256 KB).

For more information, see Using Amazon SQS Message Attributes.

To send messages larger than 256 KB, use the Amazon SQS Extended Client Library for Java. This library lets you send an Amazon SQS message that contains a reference to a message payload in Amazon S3 that can be as large as 2 GB.

Q: What kind of data can I include in a message?

Amazon SQS messages can contain up to 256 KB of text data, including XML, JSON and unformatted text. The following Unicode characters are accepted:

#x9 | #xA | #xD | [#x20 to #xD7FF] | [#xE000 to #xFFFD] | [#x10000 to #x10FFFF]

For more information, see the XML 1.0 Specification.

Q: How large can Amazon SQS message queues be?

A single Amazon SQS message queue can contain an unlimited number of messages. However, there is a 120,000 limit for the number of inflight messages per queue. Messages are inflight after they have been received from the queue by a consuming component, but have not yet been deleted from the queue.

Q: How many message queues can I create?

You can create any number of message queues.

Q: Is there a size limit on the name of Amazon SQS message queues?

Queue names are limited to 80 characters.

Q: Are there restrictions on the names of Amazon SQS message queues?

You can use alphanumeric characters, hyphens (-), and underscores (_).

Q: Can I reuse a message queue name?

A message queue's name must be unique within an AWS account and region. You can reuse a message queue's name after you delete the message queue.

Q: What happens if there is no activity against a message queue for an extended period of time?

We reserve the right to delete a message queue if none of the following requests are issued against the message queue for more than 30 consecutive days:

- SendMessage
- ReceiveMessage
- DeleteMessage
- GetQueueAttributes
- SetQueueAttributes

Queues that act as dead letter queues are not deleted as long as any of their source queues still exist.

Queue Sharing

Q: How do I share a message queue?

You can associate an access policy statement (and specify the permissions granted) with the message queue to be shared. Amazon SQS provides APIs for creating and managing access policy statements:

- AddPermission
- RemovePermission
- SetQueueAttributes
- GetQueueAttributes

For more information, see the*Amazon SQS API Reference*.

Q: Who pays for shared queue access?

The message queue owner pays for shared message queue access.

Q: How do I identify another AWS user I want to share a message queue with?

The Amazon SQS API uses the AWS account number to identify AWS users.

Q: What do I need to provide to an AWS user I want to share a message queue with?

To share a message queue with an AWS user, provide the full URL from the message queue you want to share. The CreateQueue and ListQueues operations return this URL in their responses.

Q: Does Amazon SQS support anonymous access?

Yes. You can configure an access policy that allows anonymous users to access a message queue.

Q: When should I use the permissions API?

The permissions API provides an interface for sharing access to a message queue to developers. However, this API cannot allow conditional access or more advanced use cases.

Q: When should I use the SetQueueAttributes operation with JSON objects?

The SetQueueAttributes operation supports the full access policy language. For example, you can use the policy language to restrict access to a message queue by IP address and time of day. For more information, see Amazon SQS Policy Examples.

Service Access and Regions

Q: What regions is Amazon SQS available in?

For region availability, see the AWS Global Infrastructure Region Table.

Q: Can I share messages between queues in different regions?

No. Each Amazon SQS message queue is independent within each region.

Q: Is there a pricing difference between regions?

Amazon SQS pricing is the same for all regions, except Asia Pacific (Tokyo) and AWS GovCloud (US). For more information, see Amazon SQS Pricing.

Q: What is the pricing structure between various regions?

You can transfer data between Amazon SQS and Amazon EC2 or AWS Lambda free of charge within a single region.

When you transfer data between Amazon SQS and Amazon EC2 or AWS Lambda in different regions, you will be charged the normal data transfer rate. For more information, see Amazon SQS Pricing.

Amazon SWF FAQ

General

Q: What is Amazon SWF?
Amazon Simple Workflow Service (SWF) is a web service that makes it easy to coordinate work across distributed application components. Amazon SWF enables applications for a range of use cases, including media processing, web application back-ends, business process workflows, and analytics pipelines, to be designed as a coordination of tasks. Tasks represent invocations of various processing steps in an application which can be performed by executable code, web service calls, human actions, and scripts.

The coordination of tasks involves managing execution dependencies, scheduling, and concurrency in accordance with the logical flow of the application. With Amazon SWF, developers get full control over implementing processing steps and coordinating the tasks that drive them, without worrying about underlying complexities such as tracking their progress and keeping their state. Amazon SWF also provides the AWS Flow Framework to help developers use asynchronous programming in the development of their applications. By using Amazon SWF, developers benefit from ease of programming and have the ability to improve their applications' resource usage, latencies, and throughputs.

Q: What are the benefits of designing my application as a coordination of tasks?How does Amazon SWF help me with this?
In Amazon SWF, tasks represent invocations of logical steps in applications. Tasks are processed by workers which are programs that interact with Amazon SWF to get tasks, process them, and return their results. A worker implements an application processing step. You can build workers in different programming languages and even reuse existing components to quickly create the worker. For example, you can use cloud services, enterprise applications, legacy systems, and even simple scripts to implement workers. By independently controlling the number of workers for processing each type of task, you can control the throughput of your application efficiently.

To coordinate the application execution across workers, you write a program called the decider in your choice of programming language. The separation of processing steps and their coordination makes it possible to manage your application in a controlled manner and give you the flexibility to deploy, run, scale and update them independently. You can choose to deploy workers and deciders either in the cloud (e.g. Amazon EC2 or Lambda) or on machines behind corporate firewalls. Because of the decoupling of workers and deciders, your business logic can be dynamic and you application can be quickly updated to accommodate new requirements. For example, you can remove, skip, or retry tasks and create new application flows simply by changing the decider.

By implementing workers and deciders, you focus on your differentiated application logic as it pertains to performing the actual processing steps and coordinating them. Amazon SWF handles the underlying details such as storing tasks until they can be assigned, monitoring assigned tasks, and providing consistent information on their completion. Amazon SWF also provides ongoing visibility at the level of each task through APIs and a console.

Q: What can I do with Amazon SWF?
Amazon SWF can be used to address many challenges that arise while building applications with distributed components. For example, you can use Amazon SWF and the accompanying AWS Flow Framework for:

- Writing your applications as asynchronous programs using simple programming constructs that abstract details such as initiating tasks to run remotely and tracking the program’s runtime state.
- Maintaining your application’s execution state (e.g. which steps have completed, which ones are running, etc.). You do not have to use databases, custom systems, or ad hoc solutions to keep execution state.
- Communicating and managing the flow of work between your application components. With Amazon SWF, you do not need to design a messaging protocol or worry about lost and duplicated tasks.
- Centralizing the coordination of steps in your application. Your coordination logic does not have to be scattered across different components, but can be encapsulated in a single program.
- Integrating a range of programs and components, including legacy systems and 3rd party cloud services, into your applications. By allowing your application flexibility in where and in what combination the application components are deployed, Amazon SWF helps you gradually migrate application components from private data centers to public cloud infrastructure without disrupting the application availability or performance.
- Automating workflows that include long-running human tasks (e.g. approvals, reviews, investigations, etc.) Amazon SWF reliably tracks the status of processing steps that run up to several days or months.
- Building an application layer on top of Amazon SWF to support domain specific languages for your end users. Since Amazon SWF gives you full flexibility in choosing your programming language, you can conveniently build interpreters for specialized languages (e.g. XPDL) and customized user-interfaces including modeling tools.
- Getting detailed audit trails and visibility into all running instances of your applications. You can also incorporate visibility capabilities provided by Amazon SWF into your own user

interfaces using the APIs provided by Amazon SWF.

Customers have used Amazon SWF to build applications for video encoding, social commerce, infrastructure provisioning, MapReduce pipelines, business process management, and several other use cases. For more details on use cases, please see What are some use cases that can be solved with SWF?. To see how customers are using Amazon SWF today, please read our case studies.

Q: What are the benefits of Amazon SWF vs. homegrown solutions and existing workflow products?
When building solutions to coordinate tasks in a distributed environment, developers have to account for several variables. Tasks that drive processing steps can be long-running and may fail, timeout, or require restarts. They often complete with varying throughputs and latencies. Tracking and visualizing tasks in all these cases is not only challenging, but is also undifferentiated work. As applications and tasks scale up, developers face difficult distributed systems' problems. For example, they must ensure that a task is assigned only once and that its outcome is tracked reliably through unexpected failures and outages. By using Amazon SWF, developers can focus on their differentiated application logic, i.e. how to process tasks and how to coordinate them.

Existing workflow products often force developers to learn specialized languages, host expensive databases, and give up control over task execution. The specialized languages make it difficult to express complex applications and are not flexible enough for effecting changes quickly. Amazon SWF, on the other hand, is a cloud-based service, allows common programming languages to be used, and lets developers control where tasks are processed. By adopting a loosely coupled model for distributed applications, Amazon SWF enables changes to be made in an agile manner.

Q: What are workers and deciders?
In Amazon SWF, an application is implemented by building workers and a decider which communicate directly with the service. Workers are programs that interact with Amazon SWF to get tasks, process received tasks, and return the results. The decider is a program that controls the coordination of tasks, i.e. their ordering, concurrency, and scheduling according to the application logic. The workers and the decider can run on cloud infrastructure, such as Amazon EC2, or on machines behind firewalls. Amazon SWF brokers the interactions between workers and the decider. It allows the decider to get consistent views into the progress of tasks and to initiate new tasks in an ongoing manner. At the same time, Amazon SWF stores tasks, assigns them to workers when they are ready, and monitors their progress. It ensures that a task is assigned only once and is never duplicated. Since Amazon SWF maintains the application's state durably, workers and deciders don't have to keep track of execution state. They can run independently, and scale quickly. Please see Functionality section of the Amazon SWF detail page to learn more about the steps in building applications with Amazon SWF.

You can have several concurrent runs of a workflow on Amazon SWF. Each run is referred to as a workflow execution or an execution. Executions are identified with unique names. You use the Amazon SWF Management Console (or the visibility APIs) to view your executions as a whole and to drill down on a given execution to see task-level details.

Q: What programming conveniences does Amazon SWF provide to write applications?
Like other AWS services, Amazon SWF provides a core SDK for the web service APIs. Additionally, Amazon SWF offers an SDK called the AWS Flow Framework that enables you to develop Amazon SWF-based applications quickly and easily. AWS Flow Framework abstracts the details of task-level coordination with familiar programming constructs. While running your program, the framework makes calls to Amazon SWF, tracks your program's execution state using the execution history kept by Amazon SWF, and invokes the relevant portions of your code at the right times. By offering an intuitive programming framework to access Amazon SWF, AWS Flow Framework enables developers to write entire applications as asynchronous interactions structured in a workflow. For more details, please see What is the AWS Flow Framework?

Q: How is Amazon SWF different from Amazon SQS?
Amazon SWF provides an infrastructure that is designed for coordinating tasks when building highly scalable and auditable applications. Amazon Simple Queue Service (SQS), on the other hand, provides a reliable, highly scalable, hosted queue for storing messages. While you may use Amazon SQS to build the messaging support needed to implement your distributed application, you get this facility out-of-the-box with Amazon SWF together with other application-level capabilities. The following are the key differences between Amazon SWF and Amazon SQS:

- Amazon SWF presents a task-oriented API, whereas Amazon SQS offers a message-oriented API.
- Amazon SWF ensures that a task is assigned only once and is never duplicated. With Amazon SQS, you need to handle duplicated messages and may also need to ensure that a message is processed only once.
- Amazon SWF keeps track of all the tasks and events in an application. With Amazon SQS, you need to implement your own application-level tracking, especially if your application uses multiple queues.
- The Amazon SWF Console and visibility APIs provide an application-centric view that lets you search for executions, drill down into an execution's details, administer executions, etc. With Amazon SQS, you have to implement this type of functionality.
- Amazon SWF offers several features to facilitate application development, such as uniqueness for executions, passing data between tasks, signaling, flexibility in distributing tasks, etc. With Amazon SQS, you implement all such application-level functionality yourself.

- In addition to a core SDK to call the service APIs, Amazon SWF provides the AWS Flow Framework with which you can easily write distributed applications using programming constructs to structure asynchronous interactions.

Q: What are some use cases that can be solved with Amazon SWF?
Amazon SWF has been applied to use cases in media processing, business process automation, data analytics, migration to the cloud, and batch processing. Some examples are:

Use case #1: Video encoding using Amazon S3 and Amazon EC2. In this use case, large videos are uploaded to Amazon S3 in chunks. The upload of chunks has to be monitored. After a chunk is uploaded, it is encoded by downloading it to an Amazon EC2 instance. The encoded chunk is stored to another Amazon S3 location. After all of the chunks have been encoded in this manner, they are combined into a complete encoded file which is stored back in its entirety to Amazon S3. Failures could occur during this process due to one or more chunks encountering encoding errors. Such failures need to be detected and handled.

With Amazon SWF: The entire application is built as a workflow where each video file is handled as one workflow execution. The tasks that are processed by different workers are: upload a chunk to Amazon S3, download a chunk from Amazon S3 to an Amazon EC2 instance and encode it, store a chunk back to Amazon S3, combine multiple chunks into a single file, and upload a complete file to Amazon S3. The decider initiates concurrent tasks to exploit the parallelism in the use case. It initiates a task to encode an uploaded chunk without waiting for other chunks to be uploaded. If a task for a chunk fails, the decider re-runs it for that chunk only. The application state kept by Amazon SWF helps the decider control the workflow. For example, the decider uses it to detect when all chunks have been encoded and to extract their Amazon S3 locations so that they can be combined. The execution's progress is continuously tracked in the Amazon SWF Management Console. If there are failures, the specific tasks that failed are identified and used to pinpoint the failed chunks.

Use case #2: Processing large product catalogs using Amazon Mechanical Turk. While validating data in large catalogs, the products in the catalog are processed in batches. Different batches can be processed concurrently. For each batch, the product data is extracted from servers in the datacenter and transformed into CSV (Comma Separated Values) files required by Amazon Mechanical Turk's Requester User Interface (RUI). The CSV is uploaded to populate and run the HITs (Human Intelligence Tasks). When HITs complete, the resulting CSV file is reverse transformed to get the data back into the original format. The results are then assessed and Amazon Mechanical Turk workers are paid for acceptable results. Failures are weeded out and reprocessed, while the acceptable HIT results are used to update the catalog. As batches are processed, the system needs to track the quality of the Amazon Mechanical Turk workers and adjust the payments accordingly. Failed HITs are re-batched and sent through the pipeline again.

With Amazon SWF: The use case above is implemented as a set of workflows. A BatchProcess workflow handles the processing for a single batch. It has workers that extract the data, transform it and send it through Amazon Mechanical Turk. The BatchProcess workflow outputs the acceptable HITs and the failed ones. This is used as the input for three other workflows: MTurkManager, UpdateCatalogWorkflow, and RerunProducts. The MTurkManager workflow makes payments for acceptable HITs, responds to the human workers who produced failed HITs, and updates its own database for tracking results quality. The UpdateCatalogWorkflow updates the master catalog based on acceptable HITs. The RerunProducts workflow waits until there is a large enough batch of products with failed HITs. It then creates a batch and sends it back to the BatchProcess workflow. The entire end-to-end catalog processing is performed by a CleanupCatalog workflow that initiates child executions of the above workflows. Having a system of well-defined workflows enables this use case to be architected, audited, and run systematically for catalogs with several million products.

Use case #3: Migrating components from the datacenter to the cloud. Business critical operations are hosted in a private datacenter but need to be moved entirely to the cloud without causing disruptions.

With Amazon SWF: Amazon SWF-based applications can combine workers that wrap components running in the datacenter with workers that run in the cloud. To transition a datacenter worker seamlessly, new workers of the same type are first deployed in the cloud. The workers in the datacenter continue to run as usual, along with the new cloud-based workers. The cloud-based workers are tested and validated by routing a portion of the load through them. During this testing, the application is not disrupted because the workers in the datacenter continue to run. After successful testing, the workers in the datacenter are gradually stopped and those in the cloud are scaled up, so that the workers are eventually run entirely in the cloud. This process can be repeated for all other workers in the datacenter so that the application moves entirely to the cloud. If for some business reason, certain processing steps must continue to be performed in the private data center, those workers can continue to run in the private data center and still participate in the application.

See our case studies for more exciting applications and systems that developers and enterprises are building with Amazon SWF.

Q: Does Amazon use Amazon SWF for its own applications?
Yes. Developers within Amazon use Amazon SWF for a wide variety of projects and run millions of workflow executions every day. Their use cases include key business processes behind the Amazon.com and AWS web sites, implementations for several AWS web services and their APIs, MapReduce analytics for operational decision making, and management of user-facing content such as web pages, videos and Kindle books.

Getting Started

Q: How can I get started with Amazon SWF?

To sign up for Amazon SWF, go to the Amazon SWF detail page and click the “Sign Up Now” button. If you do not have an Amazon Web Service account, you will be prompted to create one. After signing up, you can run a sample walkthrough in the AWS Management Console which takes you through the steps of running a simple image conversion application with Amazon SWF. You can also download the AWS Flow Framework samples to learn about the various features of the service. To start using Amazon SWF in your applications, please refer to the Amazon SWF documentation.

Q: Are there sample workflows that I can use to try out Amazon SWF?
Yes. When you get started with Amazon SWF, you can try the sample walkthrough in the AWS Management Console which takes you through registering a domain and types, deploying workers and deciders and starting workflow executions. You can download the code for the workers and deciders used in this walkthrough, run them on your infrastructure and even modify them to build your own applications. You can also download the AWS Flow Framework samples, which illustrate the use of Amazon SWF for various use cases such as distributed data processing, Cron jobs and application stack deployment. By looking at the included source code, you can learn more about the features of Amazon SWF and how to use the AWS Flow Framework to build your distributed applications.

Q: What are the different ways to access SWF?
You can access SWF in any of the following ways:

- AWS SDK for Java, Ruby, .NET, and PHP
- AWS Flow Framework for Java (Included in the AWS SDK for Java)
- Amazon SWF web service APIs
- AWS Management Console

Functionality

Q: What is registration?
Registration is a one-time step that you perform for each different types of workflows and activities. You can register either programmatically or through the Amazon SWF Management Console. During registration, you provide unique type-ids for each activity and workflow type. You also provide default information that is used while running a workflow, such as timeout values and task distribution parameters.

Q: What are domains?
In SWF, you define logical containers called domains for your application resources. Domains can only be created at the level of your AWS account and may not be nested. A domain can

have any user-defined name. Each application resource, such as a workflow type, an activity type, or an execution, belongs to exactly one domain. During registration, you specify the domain under which a workflow or activity type should be registered. When you start an execution, it is automatically created in the same domain as its workflow type. The uniqueness of resource identifiers (e.g. type-ids, execution ID) is scoped to a domain, i.e. you may reuse identifiers across different domains.

Q: How can I manage my application resources across different environments and groupings?
You can use domains to organize your application resources so that they are easier to manage and do not inadvertently affect each other. For example, you can create different domains for your development, test, and production environments, and create the appropriate resources in each of them. Although you may register the same workflow type in each of these domains, it will be treated as a separate resource in each domain. You can change its settings in the development domain or administer executions in the test domain, without affecting the corresponding resources in the production domain.

Q: How does a decider coordinate a workflow in Amazon SWF?
The decider can be viewed as a special type of worker. Like workers, it can be written in any language and asks Amazon SWF for tasks. However, it handles special tasks called decision tasks. Amazon SWF issues decision tasks whenever a workflow execution has transitions such as an activity task completing or timing out. A decision task contains information on the inputs, outputs, and current state of previously initiated activity tasks. Your decider uses this data to decide the next steps, including any new activity tasks, and returns those to Amazon SWF. Amazon SWF in turn enacts these decisions, initiating new activity tasks where appropriate and monitoring them. By responding to decision tasks in an ongoing manner, the decider controls the order, timing, and concurrency of activity tasks and consequently the execution of processing steps in the application. SWF issues the first decision task when an execution starts. From there on, Amazon SWF enacts the decisions made by your decider to drive your execution. The execution continues until your decider makes a decision to complete it.

To help the decider in making decisions, SWF maintains an ongoing record on the details of all tasks in an execution. This record is called the history and is unique to each execution. A new history is initiated when an execution begins. At that time, the history contains initial information such as the execution’s input data. Later, as workers process activity tasks, Amazon SWF updates the history with their input and output data, and their latest state. When a decider gets a decision task, it can inspect the execution’s history. Amazon SWF ensures that the history accurately reflects the execution state at the time the decision task is issued. Thus, the decider can use the history to determine what has occurred in the execution and decide the appropriate next steps.

Q: How do I ensure that a worker or decider only gets tasks that it understands?

You use task lists to determine how tasks are assigned. Task lists are Amazon SWF resources into which initiated tasks are added and from which tasks are requested. Task lists are identified by user-defined names. A task list may have tasks of different type-ids, but they must all be either activity tasks or decision tasks. During registration, you specify a default task list for each activity and workflow type. Amazon SWF also lets you create task lists at run time. You create a task list simply by naming it and starting to use it. You use task lists as follows:

- While initiating an activity task, a decider can add it into a specific task list or request Amazon SWF to add it into the default task list for its activity type.
- While starting an execution, you can request Amazon SWF to add all of its decision tasks to a specific task list or to the default task list for the workflow type.
- While requesting tasks, deciders and workers specify which task list they want to receive tasks from. If a task is available in the list, SWF sends it in the response and also includes its type-id.

Based on the above, you control which task list a task gets added into and who asks for tasks from each list. Thus, you can ensure that workers and deciders only get the tasks that they understand.

Q: What is the AWS Flow Framework?How does it help me with coordinating my workflow?
AWS Flow Framework is a programming framework that enables you to develop Amazon SWF-based applications quickly and easily. It abstracts the details of task-level coordination and asynchronous interaction with simple programming constructs. Coordinating workflows in Amazon SWF involves initiating remote actions that take variable times to complete (e.g. activity tasks) and implementing the dependencies between them correctly.

AWS Flow Framework makes it convenient to express both facets of coordination through familiar programming concepts. For example, initiating an activity task is as simple as making a call to a method. AWS Flow Framework automatically translates the call into a decision to initiate the activity task and lets Amazon SWF assign the task to a worker, monitor it, and report back on its completion. The framework makes the outcome of the task, including its output data, available to you in the code as the return values from the method call. To express the dependency on a task, you simply use the return values in your code, as you would for typical method calls. The framework's runtime will automatically wait for the task to complete and continue your execution only when the results are available. Behind the scenes, the framework's runtime receives worker and decision tasks from Amazon SWF, invokes the relevant methods in your program at the right times, and formulates decisions to send back to Amazon SWF. By offering access to Amazon SWF through an intuitive programming framework, the AWS Flow Framework makes it possible to easily incorporate asynchronous and event driven programming in the development of your applications.

Q: How do workers and deciders communicate with Amazon SWF?Isn't a poll protocol resource-intensive?
Typically poll based protocols require developers to find an optimal polling frequency. If developers poll too often, it is possible that many of the polls will be returned with empty results. This leads to a situation where much of the application and network resources are spent on polling without any meaningful outcome to drive the execution forward. If developers don't poll often enough, then messages may be held for longer increasing application latencies.

To overcome the inefficiencies inherent in polling, Amazon SWF provides long-polling. Long-polling significantly reduces the number of polls that return without any tasks. When workers and deciders poll Amazon SWF for tasks, the connection is retained for a minute if no task is available. If a task does become available during that period, it is returned in response to the long-poll request. By retaining the connection for a period of time, additional polls that would also return empty during that period are avoided. With long-polling, your applications benefit with the security and flow control advantages of polling without sacrificing the latency and efficiency benefits offered by push-based web services.

Q: Can I use an existing web service as a worker?
Workers use standard HTTP GET requests to get tasks from Amazon SWF and to return the results. To use an existing web service as a worker, you can write a wrapper that gets tasks from Amazon SWF, invokes your web service's APIs as appropriate, and returns the results back to Amazon SWF. In the wrapper, you translate input data provided in a task into the parameters for your web service's API. Similarly, you also translate the output data from the web service APIs into results for the task and return those to Amazon SWF.

Q: Does Amazon SWF restrict me to use specific programming languages?
No, you can use any programming language to write a worker or a decider, as long as you can communicate with Amazon SWF using web service APIs. The AWS SDK is currently available in Java, .NET, PHP and Ruby. The AWS SDK for Java includes the AWS Flow Framework.

Q: I want to ensure that there is only one execution for each activation of my business process (e.g. a transaction, a submission, or an assignment). How do I accomplish this?
When you start new workflow executions you provide an ID for that workflow execution. This enables you to associate an execution with a business entity or action (e.g. customer ID, filename, serial number). Amazon SWF ensures that an execution's ID is unique while it runs. During this time, an attempt to start another execution with the same ID will fail. This makes it convenient for you to satisfy business needs where no more than one execution can be running for a given business action, such as a transaction, submission or assignment. Consider a workflow that registers a new user on a website. When a user clicks the submit button, the user's unique email address can be used to name the execution. If the execution already exists, the call to start the execution will fail. No additional code is needed to prevent conflicts as a result of the user clicking the button more than one when the registration is in progress.

Once the workflow execution is complete (either successfully or not), you can start another workflow execution with the same ID. This causes a new run of the workflow execution with the same execution ID but a different run ID. The run ID is generated by Amazon SWF and multiple executions that have the same workflow execution ID can be differentiated by the run ID. By allowing you to reuse workflow execution IDs in such a manner, Amazon SWF allows you to address use cases such as retries. For example, in the above user registration example, assume that the workflow execution failed when creating a database record for the user. You can start the workflow execution again with the same execution ID (user's email address) and do not have to create a new ID for retrying the registration.

Q: How does Amazon SWF help with scaling my applications?
Amazon SWF lets you scale your applications by giving you full control over the number of workers that you run for each activity type and the number of instances that you run for a decider. By increasing the number of workers or decider instances, you increase the compute resources allocated for the corresponding processing steps and, thereby, the throughput for those steps. To auto-scale, you can use run-time data that Amazon SWF provides through its APIs. For example, Amazon SWF provides the number of tasks in a task list. Since an increase in this number implies that the workers are not keeping up with the load, you can spin up new workers automatically whenever the backlog of tasks crosses a threshold.

Q: I run a large number of mission critical application executions. How can I monitor and scale them?
In addition to a Management Console, Amazon SWF provides a comprehensive set of visibility APIs. You can use these to get run-time information to monitor all your executions and to auto-scale your executions depending on load. You can get detailed data on each workflow type, such as the count of open and closed executions in a specified time range. Using the visibility APIs, you can also build your own custom monitoring applications.

Q: I have numerous executions running at any time, but a handful of them often fail or stall. How can I detect and troubleshoot these problematic executions?
Amazon SWF lets you search for executions through its Management Console and visibility APIs. You can search by various criteria, including the time intervals during which executions started or completed, current state (i.e. open or closed), and standard failure modes (e.g. timed out, terminated). To group workflow executions together, you can use upto 5 tags to associate custom text with workflow executions when you start them. In the AWS Management Console, you can use tags when searching workflow executions.

To find executions that may be stalled, you can start with a time-based search to hone in on executions that are running longer than expected. Next, you can inspect them to see task level details and determine if certain tasks have been running too long or have failed, or whether the decider has simply not initiated tasks. This can help you pinpoint the problem at a task-level.

Q: I have an activity type that can be used in multiple applications. Can I share it across

these applications?
Yes. Multiple applications can share a given activity type provided the applications and the activity are all registered within the same domain. To implement this, you can have different deciders initiate tasks for the activity type and add it to the task list that the workers for that activity poll on. The workers of that activity type will then get activity tasks from all the different applications. If you want to tell which application an activity task came from or if you want to deploy different sets of workers for different applications, you can use multiple task lists. Refer to How do I ensure that a worker or decider only gets tasks that it understands?

Security and Reliability

Q: Can I use AWS Identity and Access Management (IAM) to manage access to Amazon SWF?
Yes. You can grant IAM users permission to access Amazon SWF. IAM users can only access the SWF domains and APIs that you specify.

Q: Can I run my workers behind a firewall?
Yes. Workers use standard HTTP GET requests to ask Amazon SWF for tasks and to return the computed results. Since workers always initiate requests to Amazon SWF, you do not have to configure your firewall to allow inbound requests.

Q: Isn’t it a security risk to expose my business logic as workers and deciders?
Workers use standard HTTP GET requests to ask Amazon SWF for tasks and to return the computed results. Thus, you do not have to expose any endpoint for your workers. Furthermore, Amazon SWF only gives tasks to workers when the decider initiates those tasks. Since you write the decider, you have full control over when and how tasks are initiated, including the input data that gets sent with them to the workers.

Q: How does Amazon SWF help in coordinating tasks reliably in my application?
Amazon SWF provides useful guarantees around task assignment. It ensures that a task is never duplicated and is assigned only once. Thus, even though you may have multiple workers for a particular activity type (or a number of instances of a decider), Amazon SWF will give a specific task to only one worker (or one decider instance). Additionally, Amazon SWF keeps at most one decision task outstanding at a time for a workflow execution. Thus, you can run multiple decider instances without worrying about two instances operating on the same execution simultaneously. These facilities enable you to coordinate your workflow without worrying about duplicate, lost, or conflicting tasks.

Limits

Q: How many workflow types, activity types, and domains can I register with Amazon SWF?
You can have a maximum of 10,000 workflow and activity types (in total) that are either

registered or deprecated in each domain. You can have a maximum of 100 Amazon SWF domains (including registered and deprecated domains) in your AWS account. If you think you will exceed the above limits, please use this form to contact the Amazon SWF team to discuss your scenario and request higher limits.

Q: Are there limits on the number of workflow executions that I can run simultaneously?
At any given time, you can have a maximum of 100,000 open executions in a domain. There is no other limit on the cumulative number of executions that you run or on the number of executions retained by Amazon SWF. If you think you will exceed the above limits, please use this form to contact the Amazon SWF team to discuss your scenario and request higher limits.

Q: How long can workflow executions run?
Each workflow execution can run for a maximum of 1 year. Each workflow execution history can grow up to 25,000 events. If your use case requires you to go beyond these limits, you can use features Amazon SWF provides to continue executions and structure your applications using child workflow executions.

Q: What happens if my workflow execution is idle for an extended period of time?
Amazon SWF does not take any special action if a workflow execution is idle for an extended period of time. Idle executions are subject to the timeouts that you configure. For example, if you have set the maximum duration for an execution to be 1 day, then an idle execution will be timed out if it exceeds the 1 day limit. Idle executions are also subject to the Amazon SWF limit on how long an execution can run (1 year).

Q: How long can a worker take to process a task?
Amazon SWF does not impose a specific limit on how long a worker can take to process a task. It enforces the timeout that you specify for the maximum duration for the activity task. Note that since Amazon SWF limits an execution to run for a maximum of 1 year, a worker cannot take longer than that to process a task.

Q: How long can Amazon SWF keep a task before a worker asks for it?
Amazon SWF does not impose a specific limit on how long a task is kept before a worker polls for it. However, when registering the activity type, you can set a default timeout for how long Amazon SWF will hold on to activity tasks of that type. You can also specify this timeout or override the default timeout through your decider code when you schedule an activity task. Since Amazon SWF limits the time that a workflow execution can run to a maximum of 1 year, if a timeout is not specified, the task will not be kept longer than 1 year.

Q: Can I schedule several activity tasks by issuing one decision?
Yes, you can schedule up to 100 activity tasks in one decision and also issue several decisions one after the other.

Q: How many worker tasks, signals, and markers can I have in a workflow execution and across executions?

There is no limit on the total number of activity tasks, signals, and timers used during a workflow execution. However at this time, you can only have a maximum of 1,000 open activity tasks per workflow execution. This includes activity tasks that have been initiated and activity tasks that are being processed by workers. Similarly there can be up to 1,000 open timers per workflow execution and up to 1,000 open child executions per workflow execution.

Q: How much data can I transfer within a workflow execution?
There is no limit on the total amount of data that is transferred during a workflow execution. However, Amazon SWF APIs impose specific maximum limits on parameters that are used to pass data within an execution. For example, the input data that is passed into a activity task and the input data that is sent with a signal can each be a maximum of 32,000 characters.

Q: Does Amazon SWF retain completed executions? If so, for how long?

Amazon SWF retains the history of a completed execution for any number of days that you specify, up to a maximum of 90 days (i.e. approximately 3 months). During retention, you can access the history and search for the execution programmatically or through the console.

Q: When are API calls throttled?
Beyond infrequent spikes, You may be throttled if you make a very large number of API calls in a very short period of time. If you find that you are frequently throttled or your application encounters frequent spikes, please use this form to contact the Amazon SWF team to discuss your usage scenario and request different throttle settings for your account.

Access and Availability

Q: Which regions is Amazon SWF available in?
Amazon SWF (SWF) is available in each of the following regions: US East (Northern Virginia), US West (Oregon), US West (Northern California), EU (Ireland), EU (Frankfurt), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Sydney), South America (Sao Paulo), and AWS GovCloud (US).

Q: Is Amazon SWF available across availability zones?
Yes, Amazon SWF manages your workflow execution history and other details of your workflows across 3 availability zones so that your applications can continue to rely on Amazon SWF even if there are failures in one availability zone.

Q: What are the Amazon SWF service access points?
Please visit the AWS General Reference documentation for more information on access endpoints.

Billing

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax. Learn more.