JON BONSO

AWS CERTIFIED CLOUDOPS ENGINEER ASSOCIATE

VERSION 3.0

Tutorials Dojo Study Guide

TABLE OF CONTENTS

INTRODUCTION

Today, we live in a world of fast-paced innovation and the invention of new technologies, where competitors race to develop the next disruptive product in the market. Companies with on-premises resources are quickly shifting to the cloud, such as AWS, for the many advantages that it brings. Furthermore, AWS has been the leading cloud service provider for the past few years and is continually releasing brand-new offerings. Millions of users and businesses have already adopted the AWS platform for their operations, but not all can capitalize on the benefits that AWS brings to its customers. It takes well-trained individuals to operate on the AWS cloud platform effectively.

AWS is built and managed by highly experienced engineers who offer their expertise to deliver the best products and solutions. That is why you can almost always find a function or service in AWS that would fulfill whatever need or requirement you have. A lot of the heavy lifting is offloaded from you as the customer so that you can dedicate your efforts and resources to your business operations. Another significant benefit of the AWS cloud is that it is extremely cost-effective and offers a way to expedite the launch of your products and services compared with the traditional methods. Resources can be quickly provisioned for a very low price and can be decommissioned quickly once you don't need them anymore. The cloud is an essential infrastructure piece in most companies, as you can quickly spin up or tear down test environments with just a push of a button. It can simplify deployment processes that are usually difficult and expensive to do in traditional data center setups.

The AWS Certified CloudOps Engineer Associate (SOA-C03) is a well-recognized certificate in the IT industry and is a major booster for career opportunities and salary increases. Having this certificate under your belt means that you indeed have the relevant knowledge and skills in deployment, management, and operations on AWS. Once you have gained more experience with AWS, you can also aim for higher-level certifications, such as the AWS Certified DevOps Engineer Professional certificate. The Professional and Specialty level certification exams in AWS are quite difficult and require extensive hands-on experience in order to ensure a pass. So if you are planning to pursue a career in Cloud DevOps, passing the AWS Certified CloudOps Engineer Associate is a great way to start the journey.

Note: We took extra care to come up with these concise articles and cheat sheets; however, this is meant to be just a supplementary resource when preparing for the exam. We highly recommend doing these **hands-on lab sessions**, **video course**, and **practice exams** to expand your knowledge further and improve your test-taking skills.

AWS CERTIFIED CLOUDOPS ENGINEER ASSOCIATE EXAM OVERVIEW

In 2013, Amazon Web Services (AWS) began the Global Certification Program with the primary purpose of validating the technical skills and knowledge for building secure and reliable cloud-based applications using the AWS platform. By successfully passing the AWS exam, individuals can prove their AWS expertise to their current and future employers. The AWS Certified Solutions Architect - Associate exam was the first AWS certification that was launched, followed by two other role-based certifications: Systems Operations (SysOps) Administrator and Developer Associate later that year.

As of September 2025, the AWS Certified SysOps Administrator - Associate exam has been renamed and updated. The new version is AWS Certified CloudOps Engineer - Associate (SOA-C03), reflecting evolving industry terms and the broader scope of modern cloud operations. This new title applies exclusively to individuals who pass the updated SOA-C03 exam. Those who previously earned the AWS Certified SysOps Administrator - Associate will retain that original designation; the change will not be applied retroactively.

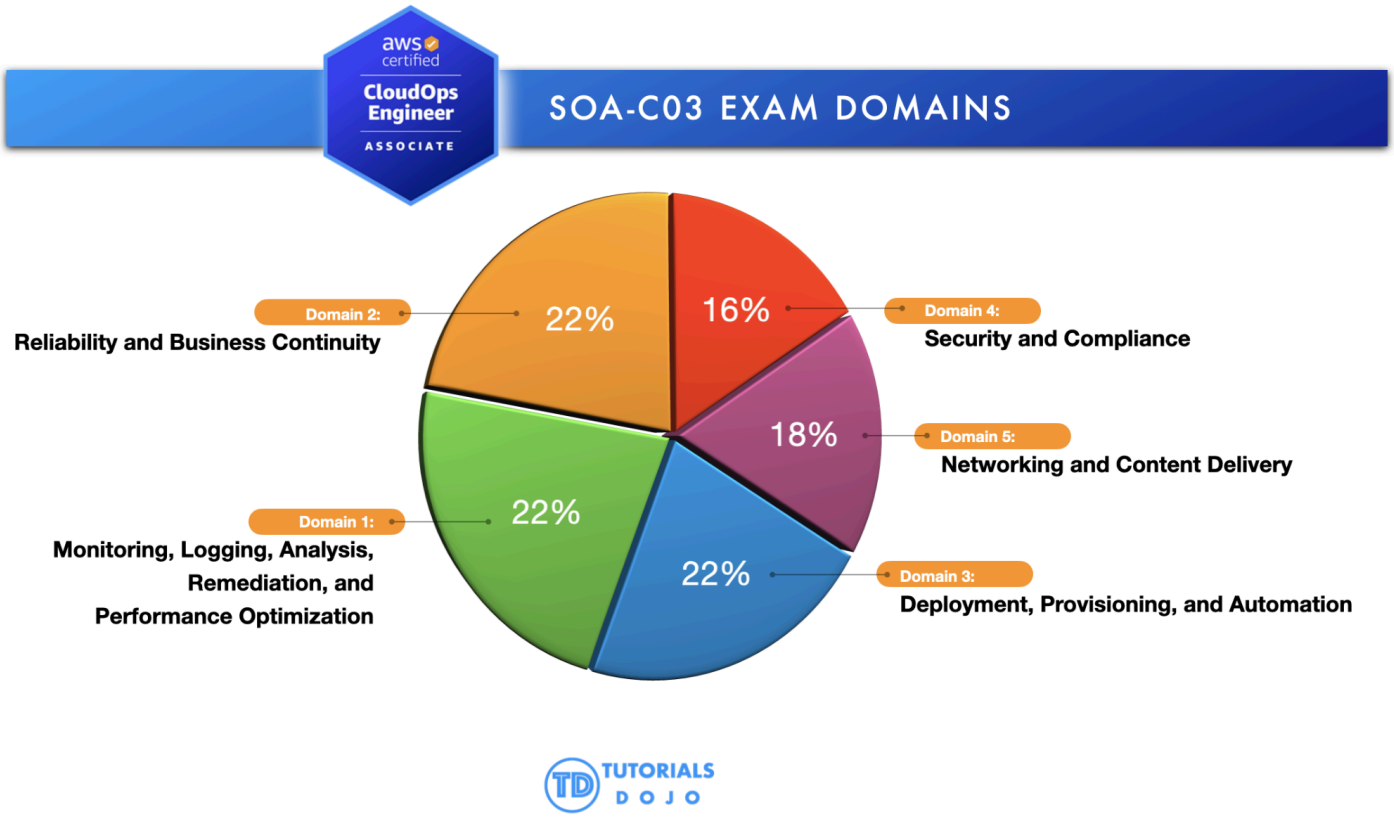
Exam Details

The AWS Certified CloudOps Engineer - Associate examination is intended for CloudOps engineers with at least one year of experience in deployment, management, troubleshooting, networking, and security on AWS. The number of questions varies, depending on the exam.

Exam Code:	SOA-C03
No. of Questions:	65
Score Range:	100/1000
Passing Score:	720/1000
Time Limit:	130 minutes
Format:	Multiple choice/multiple response questions
Delivery Method:	Testing center or online proctored exam

Exam Domains

The AWS Certified CloudOps Engineer - Associate exam has five (5) different domains, each with corresponding weight and topic coverage. The domains are: Monitoring, Logging, Analysis, Remediation, and Performance Optimization (22%), Reliability and Business Continuity (22%), Deployment, Provisioning, and Automation (22%), Security and Compliance (16%), and Networking and Content Delivery (18%).



Domain 1: Monitoring, Logging, Analysis, Remediation, and Performance Optimization (22%)

1.1 Implement metrics, alarms, and filters by using AWS monitoring and logging services
1.2 Identify and remediate issues by using monitoring and availability metrics
1.3 Implement performance optimization strategies for compute, storage, and database resources

Domain 2: Reliability and Business Continuity (22%)

2.1 Implement scalability and elasticity
2.2 Implement highly available and resilient environments
2.3 Implement backup and restore strategies

Domain 3: Deployment, Provisioning, and Automation (22%)

3.1 Provision and maintain cloud resources
3.2 Automate the management of existing resources

Domain 4: Security and Compliance (16%)

- 4.1 Implement and manage security and compliance tools and policies
- 4.2 Implement strategies to protect data and infrastructure

Domain 5: Networking and Content Delivery (18%)

- 5.1 Implement and optimize networking features and connectivity
- 5.2 Configure domains, DNS services, and content delivery
- 5.3 Troubleshoot network connectivity issues

Exam Scoring

You can get a score from 100 to 1,000 with a minimum passing score of **720** when you take the AWS Certified CloudOps Engineer Associate exam. AWS uses a scaled scoring model to associate scores across multiple exam types that may have different levels of difficulty. Your complete score report will be sent to you by email about 3 to 5 business days after your exam.

For individuals who unfortunately did not pass their exams, they must wait for 14 days before they are allowed to retake the exam. There is no hard limit on the number of attempts you can retake an exam. Once you pass, you'll receive various benefits, such as a 50% discount coupon, which you can use for your next AWS exam.

Once you receive your score report via email, the result should also be saved in your AWS Certification account already. The score report contains a table of your performance on each domain, and it will indicate whether you have met the level of competency required for these domains.

Take note that you do not need to achieve competency in all domains for you to pass the exam. At the end of the report, there will be a score performance table that highlights your strengths and weaknesses, which will help you determine the areas you need to improve on.

Exam Benefits

If you have successfully passed any AWS exam, you will be eligible for the following benefits:

- **Exam Discount** - You'll get a 50% discount voucher that you can apply for your recertification or any other exam you plan to pursue. To access your discount voucher code, go to the "Benefits" section of your AWS Certification account and apply the voucher when you register for your next exam.

- **Certification Digital Badges** - You can showcase your achievements to your colleagues and employers with digital badges on your email signatures, LinkedIn profile, or social media accounts. You can also show your Digital Badge to gain exclusive access to Certification Lounges at AWS re:Invent, regional Appreciation Receptions, and select AWS Summit events. To view your badges, simply go to the "Digital Badges" section of your AWS Certification Account.

- **Event Recognition** - You can receive invitations to regional Appreciation Receptions and use your digital badge for exclusive access to AWS Certification Lounges at AWS re:Invent and select AWS Summit events.

You can visit the official AWS Certification FAQ page to view the frequently asked questions about getting AWS Certified and other information about the AWS Certification: https://aws.amazon.com/certification/faqs/.

AWS CERTIFIED CLOUDOPS ENGINEER ASSOCIATE EXAM - STUDY GUIDE AND TIPS

If you are a Systems Administrator or a DevOps Engineer, then this certification will test your knowledge on various knowledge areas in the AWS Cloud platform. Your experience will come in handy in passing the exam, but this should be complemented by actual experience in working with your cloud workloads. Doing several hands-on labs is quite helpful, too, as it will give you that much-needed experience and insight into how the different AWS services work.

The AWS Certified CloudOps Engineer Associate exam will verify your ability to perform the following:

- Deploy, manage, and operate workloads on AWS
- Support and maintain AWS workloads according to the AWS Well-Architected Framework
- Perform operations by using the AWS Management Console and the AWS CLI
- Implement security controls to meet compliance requirements
- Monitor, log, and troubleshoot systems
- Apply networking concepts (for example, DNS, TCP/IP, firewalls)
- Implement architectural requirements (for example, high availability, performance, capacity)
- Perform business continuity and disaster recovery procedures
- Identify, classify, and remediate incidents

You should also be knowledgeable about the AWS Well-Architected Framework, AWS Deployment Options. Having prior knowledge of fundamental networking and security will also be very valuable. This guide aims to provide you with a straightforward guide when reviewing for this exam.

Study Materials

There are many study resources that you can use to prepare for the AWS Certified CloudOps Engineer - Associate (SOA-C03) exam; however, your source of truth should always be the **official exam guide**. You should also know the latest exam code of the CloudOps exam, which is SOA-C03, so you can ensure that the study materials you are using is not for the previous exam version (e.g. SOA-C02). Use the official exam guide to know the relevant exam topics of the test and verify if the resources you are using adequately cover the required knowledge areas.

- AWS Certified SysOps Administrator - Associate video course
- AWS Certified CloudOps Engineer - Associate SOA-C03 Practice Exams
- AWS Cheat Sheets
- AWS Hands-On Labs

Recommended Whitepapers

The whitepapers listed below are arranged in such a way that you will learn the concepts first, before proceeding to application and best practices. If you need a refresh on your AWS fundamentals, go check out our guide on the AWS Certified Cloud Practitioner Exam before proceeding below.

1. Amazon Virtual Private Cloud Connectivity Options - Study how you can connect different VPCs together, your VPCs to your on-premises network, and vice versa.
2. How AWS Pricing Works - Study the fundamental drivers of cost in AWS, the pricing models of commonly used services in compute, storage, and database, and how to optimize your costs.
3. AWS Well-Architected Framework - This whitepaper is one of the most important papers that you should study for the SOA-C03 exam. It discusses the different pillars that make up a well-architected cloud environment.
4. Overview of Deployment Options on AWS - This is an optional whitepaper that you can read to be aware of your deployment options in AWS. There is a chance that this might come up in the exam.
5. AWS Disaster Recovery Plans - As a CloudOps Engineer, you should be familiar with your DR options when outages occur. Having knowledge of DR will determine how fast you can recover your infrastructure.

AWS Services to Focus On

AWS offers extensive documentation and well-written FAQs for all of their services. These two will be your primary source of information when studying. Furthermore, as an AWS CloudOps Engineer, you need to be well-versed in a number of AWS products and services since you will almost always be using them in your work. I recommend checking out Tutorials Dojo's AWS Cheat Sheets which provides a summarized but highly informative set of notes and tips for your review of these services.

Core AWS Services to focus on:

1. EC2 - As the most fundamental computing service offered by AWS, you should know about EC2 inside out.
2. Elastic Load Balancer - Load balancing is very important for a highly available system. Study the different types of ELBs and the features each of them supports.
3. Auto Scaling - Study what services in AWS can be auto-scaled, what triggers scaling, and how auto-scaling increases/decreases the number of instances.
4. Elastic Block Store - As the primary storage solution of EC2, study the types of EBS volumes available. Also study how to secure, backup, and restore EBS volumes.
5. S3 / Glacier - Study what the S3 storage types are and the differences between them. Also review the capabilities of S3, such as hosting a static website, securing access to objects using policies, lifecycle policies, etc. Learn as much about S3 as you can.

6. VPC - Study every service that is used to create a VPC (subnets, route tables, internet gateways, nat gateways, VPN gateways, etc). Also, review the differences between network access control lists and security groups and during which situations they are applied.
7. Route 53 - Study the different types of records in Route 53 and the various routing policies. Know what hosted zones and domains are.
8. RDS - Know how each RDS database differs from one another, and how they are different from Aurora. Determine what makes Aurora unique, and when it should be preferred from other databases (in terms of function, speed, cost, etc). Learn about parameter groups, option groups, and subnet groups.
9. DynamoDB - Consider how DynamoDB compares to RDS, Elasticache, and Redshift. This service is also commonly used for serverless applications along with Lambda.
10. ElastiCache - Familiarize yourself with ElastiCache Redis and its functions. Determine the areas/services where you can place a caching mechanism to improve data throughput, such as managing the session state of an ELB, optimizing RDS instances, etc.
11. SQS - Gather info on why SQS is helpful in decoupling systems. Study how messages in the queues are being managed (standard queues, FIFO queues, dead letter queues). Know the differences between SQS, SNS, SES, and Amazon MQ.
12. SNS - Study the function of SNS and what services can be integrated with it. Also, be familiar with the supported recipients of SNS notifications.
13. IAM - Services such as IAM Users, Groups, Policies, and Roles are the most important to learn. Study how IAM integrates with other services and how it secures your application through different policies. Also, read on the best practices when using IAM.
14. CloudWatch - Study how monitoring is done in AWS and what types of metrics are sent to CloudWatch. Also read upon CloudWatch Logs, CloudWatch Alarms, and the custom metrics made available with CloudWatch Agent.
15. CloudTrail - Familiarize yourself with how CloudTrail works, and what kinds of logs it stores as compared to CloudWatch Logs.
16. Config - Be familiar with the situations where AWS Config is useful.
17. CloudFormation - Study how CloudFormation is used to automate infrastructure deployment. Learn the basic makeup of a CloudFormation template, stack, and stack set.
18. KMS – Familiarize how KMS integrates with other services in storing encryption keys.
19. Secrets Manager – Understand how Secrets Manager stores secrets and how you can use them with other AWS services.
20. Parameter Store – Know when to use Parameter store and how compute services like EC2, ECS, and Lambda utilize it.
21. DataSync – Familiarize yourself with which AWS services can be used to migrate data from an on-premises data center.

Some additional services we recommend to review:

1. Trusted Advisor
2. Systems Manager

3. CodeDeploy
4. CodePipeline
5. CloudFront
6. Cost and Billing Management Console
7. Direct Connect
8. Amazon FSx for Windows File Server and Amazon FSx for Lustre
9. AWS Backup
10. EC2 Image Builder
11. S3 Transfer Acceleration
12. AWS Global Accelerator
13. RDS Proxy
14. IAM Access Analyzer

Common Exam Scenarios

Scenario	Solution
Monitoring, Logging, Analysis, Remediation, and Performance Optimization	
You need to set up an alert that notifies the IT manager about EC2 instances service limits.	Use Amazon Eventbridge to detect and react to changes in the status of Trusted Advisor checks
You need to track the deletion and rotation of KMS keys.	Use AWS CloudTrail to log AWS KMS API calls
You need to investigate if the traffic is reaching the EC2 instance.	Use VPC flow logs
You need to ensure that the SSH protocol is always disabled on private servers.	Use AWS Config Rules
You need to retrieve the instance metadata of an EC2 instance.	`http://169.254.169.254/latest/`
You have to monitor the CPU usage of a single process in your EC2 instance.	Use the CloudWatch Agent procstat plugin to monitor system utilization.
You need to generate a report on the replication and encryption status of all of the objects stored in the S3 bucket.	Use S3 Inventory
Metric to use to alarm when all instances behind an ALB becomes unhealthy	`AWS/ApplicationELB HealthyHostCount <= 0`

Monitor restricted CIDR changes on a security group and remove them automatically.	Use AWS Config to evaluate the security group and AWS Systems Manager Automation document to remove the unwanted CIDR range.
Monitor CreateUser API call via email	Utilize Amazon EventBridge, declare CloudTrail as a source, and CreateUser as an event pattern. Create an SNS topic and set it as an event target on Amazon EventBridge.
You have to automate the process of patching managed instances with security-related updates.	Use AWS Systems Manager Patch Manager
You need to analyze the data hosted in Amazon S3 using standard SQL.	Use Amazon Athena
Improving the site speed of a static S3 web hosting with customers around the globe	Create a CloudFront web distribution and set Amazon S3 as the origin.
You need to implement a solution to enforce the tagging of all instances that will be launched in the VPC.	Use AWS Service Catalog TagOption library
You need to get billing alerts once it reaches a certain limit.	Enable billing alerts in Account Preferences of the AWS Console.
Resize an Amazon Elasticache for Redis cluster.	Use online resizing for Amazon Elasticache Redis cluster
No sharing of Reserved Instance (RI) discounts between AWS accounts in the Organization	Disable RI discount sharing via management account and provision instances using individual AWS accounts.
Reliability and Business Continuity	
When the incoming message traffic increases, the EC2 instances fall behind and it takes too long to process the messages.	Create an Auto Scaling group that can scale out based on the number of messages in the queue.
You need to log the client's IP address, latencies, request paths, and server responses that go through your Application Load Balancer.	Enable access logging in ALB and store the logs on an S3 bucket.
You need to determine which cipher is used for the SSL connection in your ELB.	Enable Server Order Preference
You need to monitor the total number of requests or connections in your load balancer.	Monitor the `SurgeQueueLength` metric

You need to ensure that the backups of an Amazon Redshift cluster are always available.	Configure the Amazon Redshift cluster to automatically copy snapshots of a cluster to another region.
Highly available File Server that supports SMB and manages file permissions using Windows Access Control List (A.	Multi-AZ Amazon FSx for Windows File Server
Slow load time when uploading objects to S3	S3 Transfer Acceleration
PercentIOLimit metric hits 100% on EFS	Create a new Max I/O performance mode EFS file system and migrate data to the new file system using AWS DataSync.
Must ensure data integrity when performing EBS backups	Build a Lambda function that uses CreateImage API to generate AMI of the EC2 instance and include a reboot parameter. Create an Amazon EventBridge rule to execute the Lambda function daily.
An online retail application hosted on EC2 must remain available during peak sales. In case of an Availability Zone failure, the system must be restored and accessible within 30 minutes (Recovery Time Objective (RTO) = 30 minutes).	Deploy in Multi-AZ with Auto Scaling and ELB for rapid failover.
A healthcare system stores patient records in Amazon RDS. During a regional outage, the business requires that recovery ensures no more than 15 minutes of data loss (Recovery Point Objective (RPO) = 15 minutes).	Enable cross-region replication and point-in-time recovery.
A company requires objects uploaded to a destination S3 bucket in another AWS Region to replicate within a predictable time frame for compliance requirements.	Enable Amazon S3 Replication Time Control (S3 RTC) for the cross-Region replication configuration.
A company needs newly uploaded objects in an Amazon S3 bucket to automatically replicate to another AWS Region immediately for disaster recovery purposes.	Configure Amazon S3 Live Replication using Cross-Region Replication (CRR) between the source and destination S3 buckets.
Deployment, Provisioning, and Automation	

You must remotely execute shell scripts and securely manage the configuration of EC2 instances.	Use Systems Manager Run Command
You need to identify the configuration changes in the CloudFormation resources.	Use drift detection
Requires a CloudFormation template that can be reused for multiple environments. If the template has been updated, all the stack that is referencing it will automatically use the updated configuration.	Use Nested Stacks
You need to automate the process of updating the CloudFormation templates to map to the latest AMI IDs.	Use CloudFormation with Systems Manager Parameter Store
The eviction count in Amazon ElastiCache for Memcached has exceeded its threshold.	Scale the cluster by increasing the number of nodes.
You need to provide each department a new AWS account with governance guardrails and a defined baseline in place.	Set up AWS Control Tower
An S3 bucket must be configured to move the objects older than 60 days to the Infrequent Access storage class.	Set up a lifecycle policy
You need to monitor all the COPY and UNLOAD traffic in the Redshift cluster.	Enable Enhanced VPC routing on the Redshift cluster.
TLS certificate should be renewed automatically	Request a public certificate via AWS Certificate Manager (ACM)
Get cost expenses of each AWS user account	Enable the createdBy tag in the Billing and Management console
Provisioning instances on ASG takes time because of software dependencies installed via the UserData script.	EC2 Image Builder
Get cost expenses of each AWS user account	Enable the createdBy tag in the Billing and Management console
Security and Compliance	

You have to rotate an existing KMS key with imported key material every 6 months	Create a new KMS key with imported key material and update the key ID to point to the new KMS key
A company needs to restrict access to the data in an S3 bucket.	Use S3 ACL and bucket policy
Mitigate malicious attacks such as SQL injection and DDoS attacks from unknown origins.	Use AWS WAF and Shield
You need to define an IAM policy to enable the user to pass a role to an AWS service.	Define iam:PassRole in the IAM policy
You need to create a solution that allows multiple EC2 instances in a private subnet to use AWS KMS and the traffic must not pass through the public Internet.	Configure a VPC endpoint
You need to encrypt all the objects at rest in your S3 bucket	Use SS3-S3, AWS KMS keys (SSE-KMS) or SSE-C
Enable authentication to AWS services using Active Directory Federation Services.	Amazon Cognito user pool
Create a bucket policy to only allow AWS accounts in the organization to access an S3 bucket.	Set principal to (*) and create a condition for PrincipalOrgId
Read, update, delete messages from SQS queues from an instance.	Create a policy with sqs:SendMessage, sqs:ReceiveMessage, sqs:DeleteMessage, and attach the policy to a new role that can perform API calls to AWS. Associate the new role to the instance.
RDS credentials should not be hardcoded on Lambda functions	Use Secrets Manager to store credentials
Networking and Content Delivery	
You need to allow the EC2 instances in your VPC that support IPv6 to connect to the Internet but block any incoming connection.	Set up an egress-only Internet gateway
You have to establish a dedicated connection between their on-premises network and their Amazon VPC.	Set up a Direct Connect connection

You need to increase the cache hit ratio for a CloudFront web distribution.	Add a `Cache-Control max-age` and increase the TTL by specifying the longest value for max-age
You need to ensure that users are consistently directed to the AWS region nearest to them.	Set up a Route 53 Geoproximity routing policy
A company plans to implement a hybrid cloud architecture. You need to allow your resources on AWS the connectivity to external networks.	Assign an Internet Gateway to the VPC Create a Virtual Private Gateway
Users being served desktop version on mobile phones	Add a `User-Agent` header to the list of origin custom header on CloudFront.
DNS record at the apex domain	ALIAS record

Validate Your Knowledge

Once you have finished your review and you are more than confident of your knowledge, test yourself with some practice exams available online. AWS offers a practice exam that you can try out at their aws.training portal. **Tutorials Dojo** also offers a top-notch set of **AWS Certified CloudOps Engineer Associate practice tests**. Each test contains unique questions that will surely help verify if you have missed out on anything important that might appear on your exam. You can pair our practice exams with this study guide eBook to further help in your exam preparations.

Sample Practice Test Questions:

Question 1

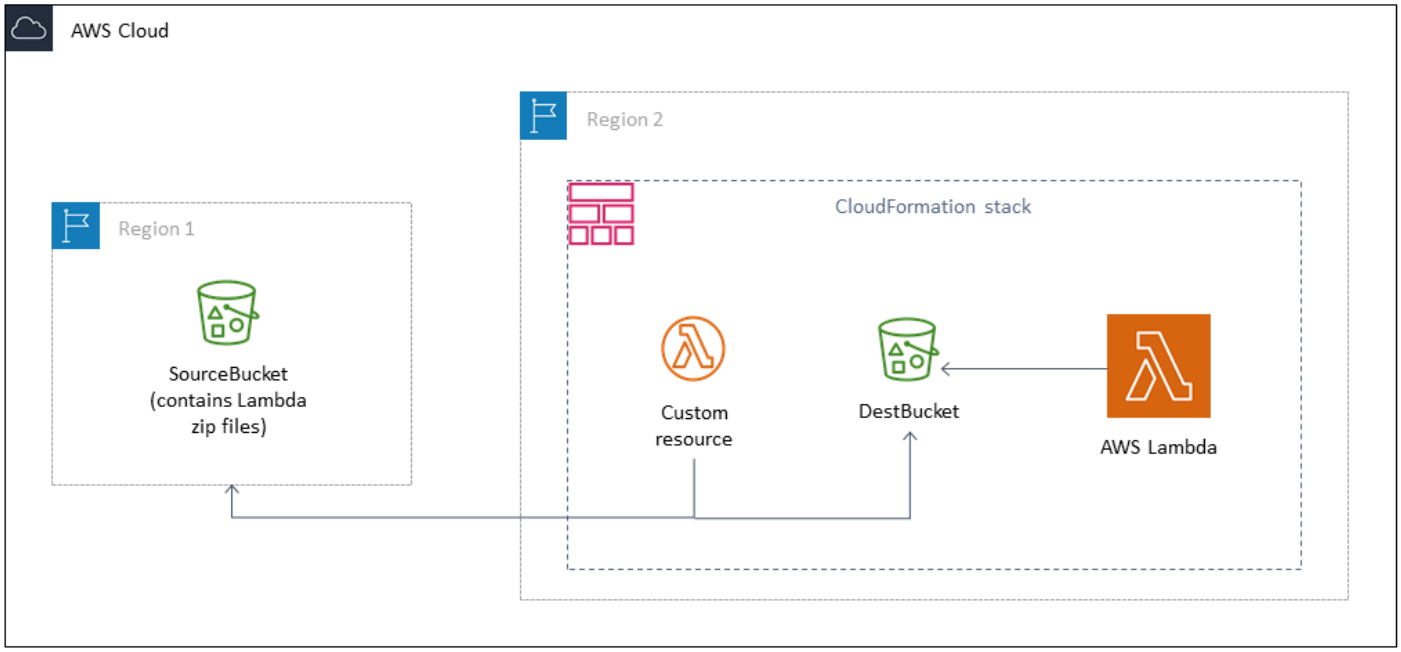
A company is heavily using AWS CloudFormation templates to automate the deployment of their cloud resources. The CloudOps Engineer needs to write a template that will automatically copy objects from an existing S3 bucket into the new one.

Which of the following is the most suitable configuration for this scenario?

1. Set up an AWS Lambda function and configure it to perform the copy operation. Integrate the Lambda function to the CloudFormation template as a custom resource.
2. Configure the CloudFormation template to modify the existing S3 bucket to allow cross-origin requests.
3. Configure the CloudFormation template to set up an AWS Step Functions state machine to orchestrate the copy process from the existing S3 bucket to the new one.
4. Configure the CloudFormation template to enable cross-region replication on the existing S3 bucket and select the new S3 bucket as the destination.

Correct Answer: 1

AWS CloudFormation gives you an easy way to model a collection of related AWS and third-party resources, provision them quickly and consistently, and manage them throughout their lifecycles, by treating infrastructure as code. A CloudFormation template describes your desired resources and their dependencies so you can launch and configure them together as a stack. You can use a template to create, update, and delete an entire stack as a single unit, as often as you need to, instead of managing resources individually. You can manage and provision stacks across multiple AWS accounts and AWS Regions.



In an AWS CloudFormation template, you can specify a Lambda function as the target of a custom resource. Use custom resources to process parameters, retrieve configuration values, or call other AWS services during stack lifecycle events. When you associate a Lambda function with a custom resource, the function is invoked whenever the custom resource is created, updated, or deleted. AWS CloudFormation calls a Lambda API to invoke the function and to pass all the request data (such as the request type and resource properties) to the function. The power and customizability of Lambda functions in combination with AWS CloudFormation

enable a wide range of scenarios, such as dynamically looking up AMI IDs during stack creation, or implementing and using utility functions, such as string reversal functions.

The requirement for this scenario is to copy all the objects from an existing S3 bucket to a new S3 bucket created by the CloudFormation template. To accomplish this requirement, you need to create a custom Lambda function that can copy the objects from the source bucket to the new S3 bucket. You can also define the options you want Amazon S3 to apply during replication, such as server-side encryption, replica ownership, and transitioning replicas to another storage class.

Hence, the correct answer is: **Set up an AWS Lambda function and configure it to perform the copy operation. Integrate the Lambda function to the Cloudformation template as a custom resource.**

The option that says: **Configure the Cloudformation template to enable cross-region replication on the existing S3 bucket and select the new S3 bucket as the destination** is incorrect because this option won't be able to copy the existing objects to the new S3 bucket. For this configuration, you need to invoke Lambda first to copy the objects in the S3 bucket.

The option that says: **Configure the CloudFormation template to set up an AWS Step Functions state machine to orchestrate the copy process from the existing S3 bucket to the new one** is incorrect because AWS Step Function is primarily designed for orchestrating complex workflows. Using it for a simple file copy operation adds unnecessary complexity and overhead.

The option that says: **Configure the CloudFormation template to modify the existing S3 bucket to allow cross-origin requests** is incorrect because the scenario did not state anything about allowing cross-origin access to your Amazon S3 resources. Also, this option does not have the capability to copy all the objects from an existing S3 bucket to a new S3 bucket.

References:
https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/template-custom-resources-lambda.html
https://aws.amazon.com/blogs/infrastructure-and-automation/deploying-aws-lambda-functions-using-aws-cloudformation-the-portable-way/
https://aws.amazon.com/blogs/devops/use-aws-cloudformation-to-automate-the-creation-of-an-s3-bucket-with-cross-region-replication-enabled/

Check out this AWS CloudFormation Cheat Sheet:
https://tutorialsdojo.com/aws-cloudformation/

Question 2

An eCommerce company has a suite of microservices-based retail applications on a Kubernetes cluster using Amazon Elastic Kubernetes Service (Amazon EKS) in AWS.

The application suite has been running for a few months when the DevOps team notices a surge of application traffic whenever there's a scheduled promotion event. To avoid loss of revenue, the CloudOps Engineer must ensure an uninterrupted service and preempt any potential degradation of the service in the production environment.

Which combination of actions can provide the MOST operationally efficient solution that can meet the above requirement? (Select TWO.)

1. Enable the built-in Kubernetes Horizontal Pod Autoscaler option in Amazon EKS.

2. Set up a Kubernetes Metrics Server first in the Amazon EKS cluster and enable the Kubernetes Vertical Pod Autoscaler.

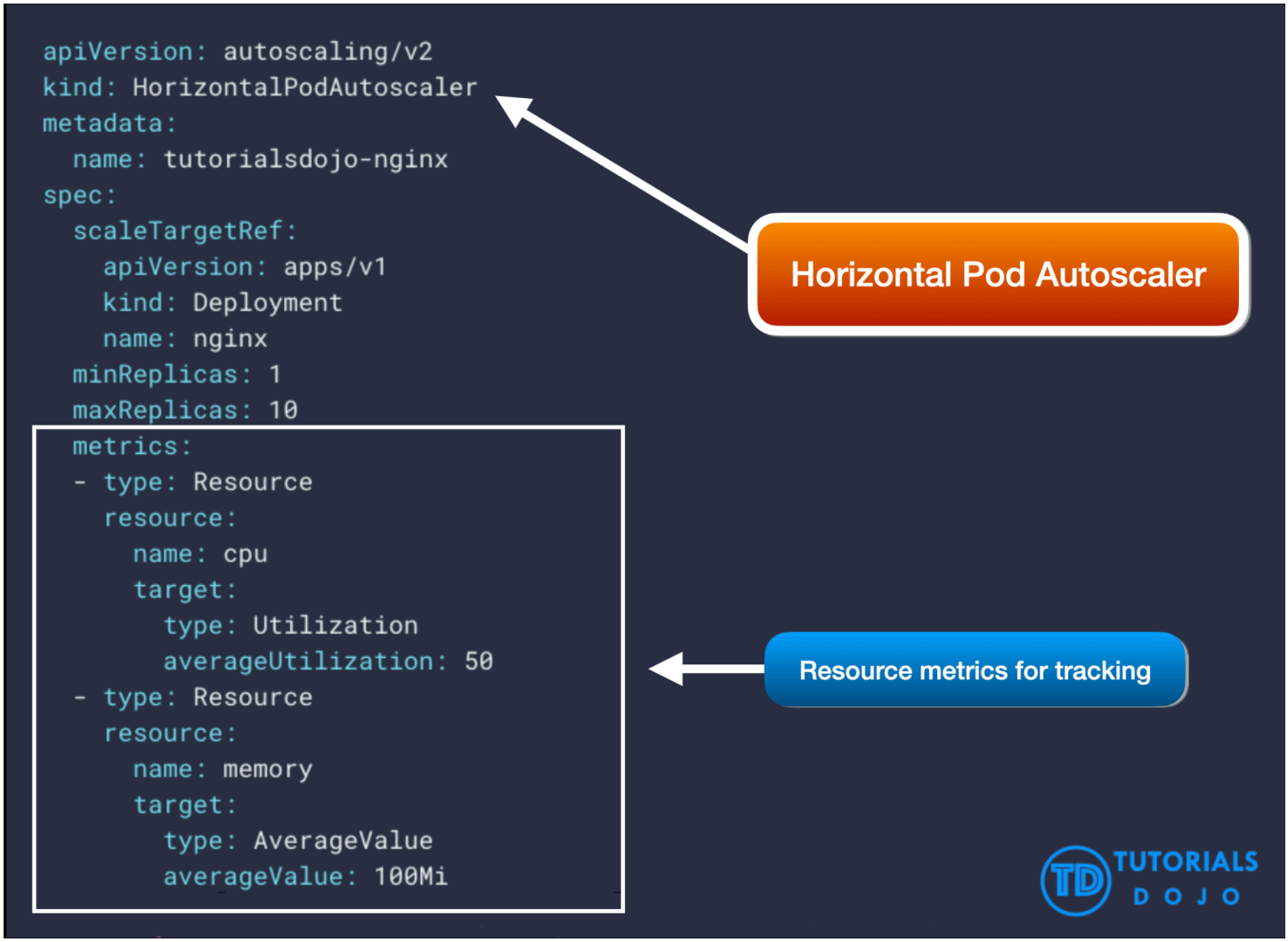
3. Integrate both Amazon CloudWatch Contributor Insights and Amazon CloudWatch Application Insights in the EKS cluster.

4. Install the Kubernetes Metrics Server in the Amazon EKS cluster and use the Kubernetes Horizontal Pod Autoscaler.

5. Configure the kubectl client to communicate with the Amazon EKS cluster.

Correct Answers: 4 & 5

Take note that the Kubernetes Horizontal Pod Autoscaler feature automatically scales the number of Pods in a deployment, replication controller, or replica set based on that resource's CPU utilization. The Horizontal Pod Autoscaler can help your applications scale out automatically to meet increased demand or scale in when resources are not needed, thus freeing up your nodes for other applications. When you set a target CPU utilization percentage, the Horizontal Pod Autoscaler scales your application in or out to try to meet that target.



Hence, the correct answers are:

- Install the Kubernetes Metrics Server in the Amazon EKS cluster and use the Kubernetes Horizontal Pod Autoscaler.

- Configure the `kubectl` client to communicate with the Amazon EKS cluster.

The option that says: **Set up a Kubernetes Metrics Server first in the Amazon EKS cluster and enable the Kubernetes Vertical Pod Autoscaler** is incorrect. While setting up a Kubernetes Metrics Server is right, the use of a Kubernetes Vertical Pod Autoscaler (VPA) is not required at all. VPA is useful for optimizing resource utilization within individual pods as it involves more fine-tuning and requires a deeper understanding of the application's resource requirements. In contrast, Horizontal Pod Autoscaler (HPA) is more geared toward

handling changes in demand by adjusting the number of pod replicas, making it a simpler and more automated solution for handling increased application traffic.

The option that says: **Enable the built-in Kubernetes Horizontal Pod Autoscaler option in Amazon EKS** is incorrect because there are no built-in Horizontal Pod Autoscaler in Amazon EKS. You have to manually install the Kubernetes Metrics Server and the Horizontal Pod Autoscaler programs first in your Amazon EKS cluster.

The option that says: **Integrate both Amazon CloudWatch Contributor Insights and Amazon CloudWatch Application Insights in the EKS cluster** is incorrect. Even though CloudWatch Contributor Insights and Application Insights are great for monitoring, this combination is not capable of scaling the Amazon EKS cluster. You need to use a Kubernetes Horizontal Pod Autoscaler for this case.

References:

https://docs.aws.amazon.com/eks/latest/userguide/horizontal-pod-autoscaler.html

https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/#how-does-a-horizontalpodautoscaler-work

https://docs.aws.amazon.com/eks/latest/userguide/eks-workloads.html

Check out this Amazon Elastic Kubernetes Service Cheat Sheet:

https://tutorialsdojo.com/amazon-elastic-kubernetes-service-eks/

Click **here** for more **AWS Certified CloudOps Engineer - Associate practice exam questions**.

It is best to get some rest before the day of your exam, and review any notes that you have written down. If you have done well in the **practice tests**, go over the questions where you made a mistake and understand why so. If you are not feeling so confident after trying the practice tests, you can just reschedule your exam and take your time preparing. The exam will not be easy to pass, but it'll be worth it when you do.

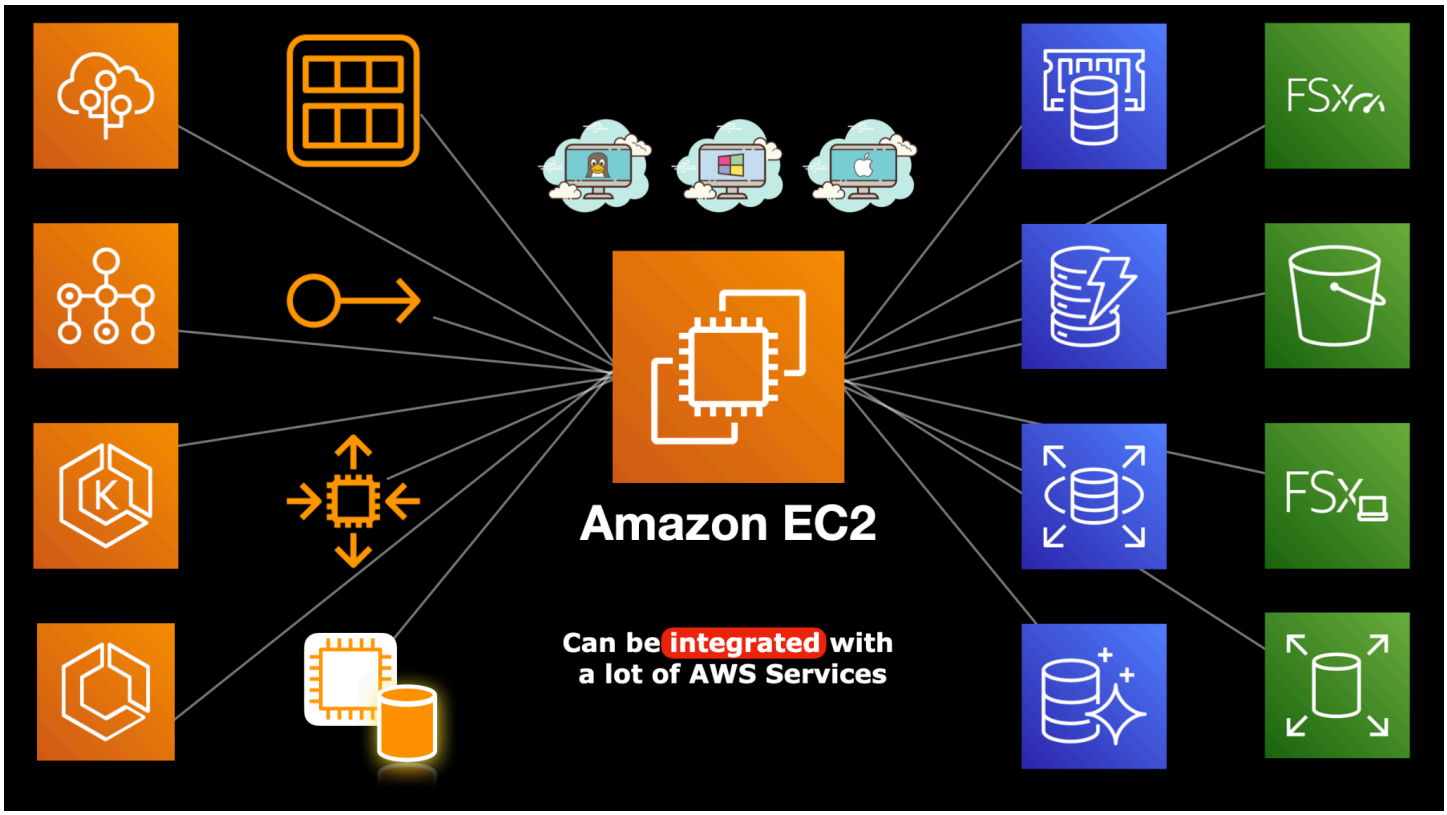
AWS Deep Dives

Amazon EC2

Amazon Elastic Compute Cloud (EC2) is a computing service that runs virtual servers in the cloud. It allows you to launch Linux or Windows virtual machines to host your applications and manage them remotely – wherever you are in the globe.

You and AWS have a shared responsibility in managing your Amazon EC2 virtual machines. AWS manages the data centers, physical facilities, the hardware components, the host operating system, and the virtualization layer that powers the entire Amazon EC2 service. On the other hand, you are responsible for your guest operating system, applying OS patches, setting up security access controls, and managing your data.

Amazon EC2 can be integrated into other AWS services to accomplish a certain task or to meet your specifications. It can be used to do a variety of functions – from running applications, hosting a self-managed database, processing batch jobs and so much more.



An Amazon EC2 virtual machine is somewhat similar to your desktop or laptop that you may be using right now. It also has a CPU, a Random Access Memory, a Network Interface, an IP address, and even a system image backup. You can also attach a Solid State Drive, or a Hard Disk Drive (HDD) to your EC2 instance for more storage; You can even connect it to a shared network file system, to allow multiple computers to access the same files.

Just like your computer, you can also integrate a lot of other AWS services with Amazon EC2. You can attach various storage, networking, and security services to an Amazon EC2 instance. There are many options available to purchase your EC2 instance, that can help you lower down your operating costs. Some AWS Services are even using Amazon EC2 as its underlying compute component. These services orchestrate or control a group of EC2 instances to perform a specific function, such as scaling or batch processing. It is also used on AWS-managed databases, containers, serverless computing engines, microservices, and many more! This is why Amazon EC2 is considered as the basic building block in AWS – it is used in almost every service!

For storage, you can use different AWS Storage services with your Amazon EC2 instance to store and process data. You can attach an Instance store for your temporary data or an Amazon EBS volume for persistent storage.

You can also mount a file system to your EC2 instances. You can connect it to Amazon EFS or Amazon FSx. For your static media files or object data, you can store them in Amazon S3 and then retrieve them back to your EC2 instance via an API or through an HTTP and FTP client.

For networking, you launch your EC2 instance on either a public or a private subnet in a Virtual Private Cloud or VPC. You can associate an Elastic IP address to your instance for it to have a static IPv4 address. An elastic network interface can also be used as a virtual network card for your EC2 instance. If you have a group of interdependent instances, you can organize them into a placement group. This placement group can be a cluster, a spread, or a partition type that enables you to minimize correlated failures, lower network latency, and achieve high throughput.

AWS also offers enhanced networking features to provide high-performance networking capabilities by using an Elastic Network Adapter or an Intel 82599 Virtual Function (VF) interface. If you have a High-Performance Computing workload or machine learning applications, you can attach an Elastic Fabric Adapter to your instance to provide a higher network throughput than your regular TCP transport.

For scaling, you can use Amazon EC2 Auto Scaling to automatically add more EC2 instances to process the increasing number of traffic in your application. Auto Scaling can also terminate the underutilized instances if the demand decreases – this can cut down your server expenses in half, or even more!

For system image backup, you can take a snapshot of your EC2 instance by creating an Amazon Machine Image, or AMI.

The AMI is just like a disk image of your Mac, Linux, or Windows computer that contains custom data and system configurations that you have set. It enables you to launch a pre-configured Amazon EC2 instance that can be used for auto-scaling, migration, and backups. If your EC2 instance crashes, you can easily restore your data using an AMI. It is also helpful if you want to move your server to another Available Zone, another Region, or even another AWS account. You can also launch one or more EC2 instances using a single AMI.

There are more AWS services and features that you can integrate with Amazon EC2. We will cover these services in the succeeding chapters of this eBook.

Components of an EC2 Instance

You must know the components of an EC2 instance since this is one of the core AWS services that you'll be encountering the most in the exam.

1) When creating an EC2 instance, you always start off by choosing a **base AMI or Amazon Machine Image**. An AMI contains the OS, settings, and other applications that you will use in your server. AWS has many pre-built AMIs for you to choose from, and there are also custom AMIs created by other users which are sold on the AWS Marketplace for you to use. If you have created your own AMI before, it will also be available for you to select. AMIs cannot be modified after launch.
2) After you have chosen your AMI, you select the **instance type and size** of your EC2 instance. The type and size will determine the physical properties of your instance, such as CPU, RAM, network speed, and more. There are many instance types and sizes to choose from and the selection will depend on your workload for the instance. You can freely modify your instance type even after you've launched your instance, which is commonly known as “right-sizing”.
3) Once you have chosen your AMI and your hardware, you can now configure your instance settings.
    a) If you are working on the console, the first thing you'll indicate is the **number of instances** you'd like to launch with these specifications you made.
    b) You specify whether you'd like to launch **spot instances** or use another instance billing type (on-demand or reserved).
    c) You configure which **VPC and subnet** the instance should be launched in, and whether it should receive a **public IP address** or not.
    d) You choose whether to include the instance in a **placement group** or not.
    e) You indicate if the instance will be joined to one of your **domains/directories**.
    f) Next is the **IAM role** that you'd like to provide to your EC2 instance. The IAM role will provide the instance with permission to interact with other AWS resources indicated in its permission policy.
    g) **Shutdown behavior** lets you specify if the instance should only be stopped or should be terminated once the instance goes into a stopped state. If the instance supports **hibernation**, you can also enable the hibernation feature.
    h) You can enable the **termination protection** feature to protect your instance from accidental termination.

i) If you have **EFS file systems** that you'd like to immediately mount to your EC2 instance, you can specify them during launch.
j) Lastly, you can specify if you have commands you'd like your EC2 instance to execute once it has launched. These commands are written in the **user data** section and submitted to the system.

4) After you have configured your instance settings, you now need to add **storage** to your EC2 instance. A volume is automatically created for you since this volume will contain the OS and other applications of your AMI. You can add more storage as needed and specify the type and size of EBS storage you'd like to allocate. Other settings include specifying which EBS volumes are to be included for termination when the EC2 instance is terminated, as well as encryption.
5) When you have allocated the necessary storage for your instances, next is adding **tags** for easier identification and classification.
6) After adding in the tags, you now create or add **security groups** to your EC2 instance, which will serve as firewalls to your servers. Security groups will moderate the inbound and outbound traffic permissions of your EC2 instance. You can also add, remove, and modify your security group settings later on.
7) Lastly, access to the EC2 instance will need to be secured using one of your **key pairs**. Make sure that you have a copy of this key pair so that you'll be able to connect to your instance when it is launched. There is no way to reassociate another key pair once you've launched the instance. You can also proceed without selecting a key pair, but then you would have no way of directly accessing your instance unless you have enabled some other login method in the AMI or via Systems Manager.
8) Once you are happy with your instance, proceed with the launch. Wait for your EC2 instance to finish preparing itself, and you should be able to connect to it if there aren't any issues.

References:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html
https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/

Types of EC2 Instances

1. **General Purpose** — Provides a balance of computing, memory, and networking resources and can be used for a variety of diverse workloads. Instances under the T-family have burstable performance capabilities to provide higher CPU performance when the CPU is under high load in exchange for CPU credits. Once the credits run out, your instance will not be able to burst anymore. More credits can be earned at a certain rate per hour, depending on the instance size.
2. **Compute Optimized** — Ideal for compute-bound applications that benefit from high-performance processors. Instances belonging to this family are well suited for batch processing workloads, media transcoding, high-performance web servers, high-performance computing, scientific modeling, dedicated gaming servers and ad server engines, machine learning inference, and other compute-intensive applications.
3. **Memory Optimized** — Designed to deliver fast performance for workloads that process large data sets in memory.

4. **Accelerated Computing** — Uses hardware accelerators or co-processors to perform functions such as floating point number calculations, graphics processing, or data pattern matching more efficiently than on CPUs.
5. **Storage Optimized** — Designed for workloads that require high, sequential read and write access to very large data sets on local storage. They are optimized to deliver tens of thousands of low-latency, random I/O operations per second (IOPS) to applications.
6. **Nitro-based** — The Nitro System provides bare metal capabilities that eliminate virtualization overhead and support workloads that require full access to host hardware. When you mount EBS Provisioned IOPS volumes on Nitro-based instances, you can provision from 100 IOPS up to 64,000 IOPS per volume compared to just up to 32,000 on other instances.

References:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-types.html
https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/

Storage with Highest IOPS for EC2 Instance

When talking about storage and IOPS in EC2 instances, the first thing that pops into the minds of people is Amazon EBS Provisioned IOPS. Amazon EBS Provisioned IOPS volumes are the highest performing EBS volumes designed for your critical, I/O intensive applications. These volumes are ideal for both IOPS-intensive and throughput-intensive workloads that require extremely low latency. And since they are EBS volumes, your data will also persist even after shutdowns or reboots. You can create snapshots of these volumes and copy them over to your other instances, and much more.

But what if you require both high IOPS and low latency performance, and the data doesn't necessarily have to be stored on the volume? If you have this requirement, then the instance store volumes on specific instance types might be preferable to EBS-provisioned IOPS volumes. EBS volumes are attached to EC2 instances virtually, so there is still some latency in there. Instance store volumes are physically attached to the EC2 instances themselves, so your instances are able to access the data much faster. Instance store volumes can come in HDD, SSD, or NVME SSD, depending on the instance type you choose. Available storage space will depend on the instance type as well.

Reference:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/InstanceStorage.html

Instance Purchasing Options

AWS offers multiple options for you to purchase compute capacity that will best suit your needs. Aside from pricing on different instance types and instance sizes, you can also specify how you'd like to pay for the compute capacity. With EC2 instances, you have the following purchase options:

1) **On-Demand Instances** – You pay by the hour or the second depending on which instances you run for each running instance. If your instances are in a stopped state, then you do not incur instance charges. No long-term commitments.

2) **Savings Plans** – Receive discounts on your EC2 costs by committing to a consistent amount of usage, in USD per hour, for a term of 1 or 3 years. You can achieve higher discount rates by paying a portion of the total bill upfront, or paying full upfront. There are two types of Savings Plans available:
    a) **Compute Savings Plans** provide the most flexibility since it automatically applies your discount regardless of instance family, size, AZ, region, OS or tenancy, and also applies to Fargate and Lambda usage.
    b) **EC2 Instance Savings Plans** provide the lowest prices but you are committed to usage of individual instance families in a region only. The plan reduces your cost on the selected instance family in that region regardless of AZ, size, OS, or tenancy. You can freely modify your instance sizes within the instance family in that region without losing your discount.

3) **Reserved Instances (RI)** – Similar to Saving Plans but less flexible since you are making a commitment to a consistent instance configuration, including instance type and Region, for a term of 1 or 3 years. You can also pay partial upfront or full upfront for higher discount rates. A Reserved Instance has four instance attributes that determine its price:
    a) Instance type
    b) Region
    c) Tenancy - shared (default) or single-tenant (dedicated) hardware.
    d) Platform or OS

    Reserved Instances are automatically applied to running On-Demand Instances provided that the specifications match. A benefit of Reserved Instances is that you can sell unused Standard Reserved Instances in the AWS Marketplace. There are also different types of RIs for you to choose from:
    a) Standard RIs - Provide the most significant discount rates and are best suited for steady-state usage.
    b) Convertible RIs - Provide a discount and the capability to change the attributes of the RI as long as the resulting RI is of equal or greater value.
    c) Scheduled RIs - These are available to launch within the time windows you reserve. This option allows you to match your capacity reservation to a predictable recurring schedule that only requires a fraction of a day, a week, or a month.

	Standard RI	Convertible RI
Applies to usage across all Availability Zones in an AWS region	Yes	Yes
Can be shared between multiple accounts within a consolidated billing family.	Yes	Yes
Change Availability Zone, instance size (for Linux OS), networking type	Yes	Yes
Change instance families, operating system, tenancy, and payment option	No	Yes
Benefit from Price Reductions	No	Yes
Can be bought/sold in Marketplace	Yes	No

4) **Spot Instances** – Unused EC2 instances that are available for a cheap price, which can reduce your costs significantly. The hourly price for a Spot Instance is called a Spot price. The Spot price of each instance type in each Availability Zone is set by Amazon EC2, and is adjusted gradually based on the long-term supply of and demand for Spot Instances. Your Spot Instance runs whenever capacity is available and the maximum price per hour that you've placed for your request exceeds the Spot price. When the Spot price goes higher than your specified price, your Spot Instance will be stopped or terminated after a two minute warning. Use Spot Instances only when your workloads can be interrupted
5) **Dedicated Hosts** – You pay for a physical host that is fully dedicated to running your instances, and bring your existing per-socket, per-core, or per-VM software licenses to reduce costs. Support for multiple instance sizes on the same Dedicated Host is available for the following instance families: c5, m5, r5, c5n, r5n, and m5n. Dedicated Hosts also offers options for upfront payment for higher discounts.
6) **Dedicated Instances** – Pay by the hour for instances that run on single-tenant hardware. Dedicated Instances that belong to different AWS accounts are physically isolated at a hardware level. Only your compute nodes run in single-tenant hardware; EBS volumes do not.

	Dedicated Hosts	Dedicated Instances
Billing	Per-host billing	Per-instance billing
Visibility of sockets, cores, and host ID	Provides visibility on the number of sockets and physical cores	No visibility
Host and instance affinity	Allows you to consistently deploy	Not supported

	your instances to the same physical server over time	
Targeted instance placement	Provides additional visibility and control over how instances are placed on a physical server	Not supported
Automatic instance recovery	Supported	Supported
Bring Your Own License (BYOL)	Supported	Not supported
Instances must run within a VPC	Yes	Yes
Can be combined with other billing options	On-demand Dedicated Hosts, Reserved Dedicated Hosts, Savings Plans	On-demand Instances, Reserved Dedicated Instances, Dedicated Spot Instances

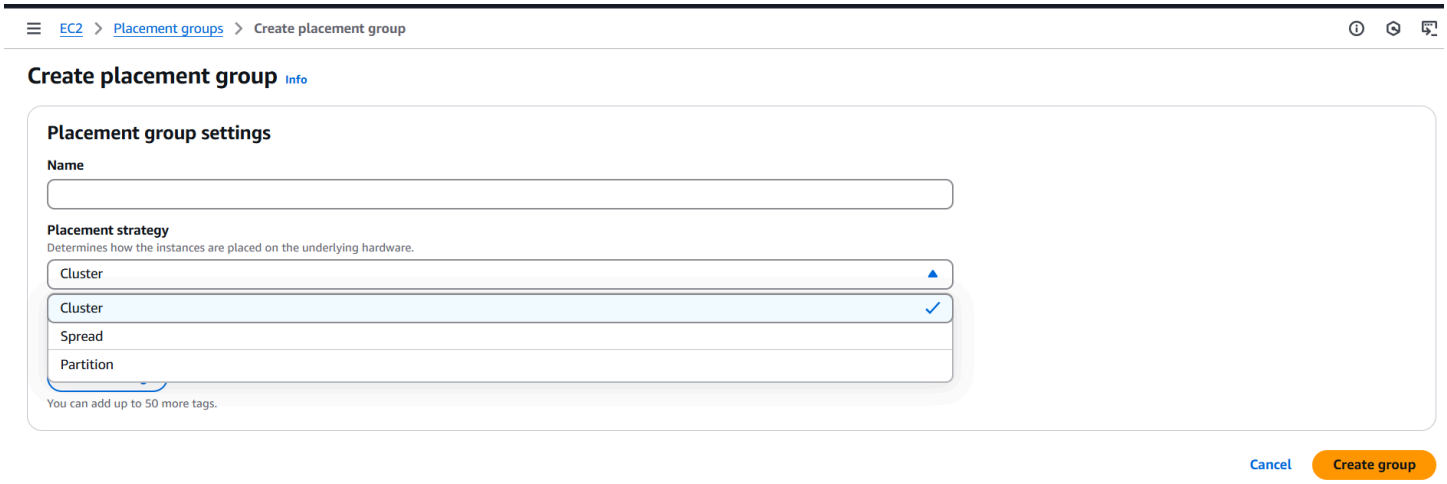
7) **Capacity Reservations** – Allows you to reserve capacity for your EC2 instances in a specific Availability Zone for any duration. No commitment required.

References:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-purchasing-options.html
https://aws.amazon.com/ec2/pricing/
https://tutorialsdojo.com/amazon-elastic-compute-cloud-amazon-ec2/

EC2 Placement Groups

Placement Groups is a logical grouping of your interdependent instances in AWS. This logical grouping affects how your instances are placed on the underlying hardware. Having the instances in a placement group has particular benefits in terms of network latency, throughput, and minimizing correlated hardware failure. By default, AWS automatically spreads out your instances across underlying hardware to reduce this correlated hardware failure.

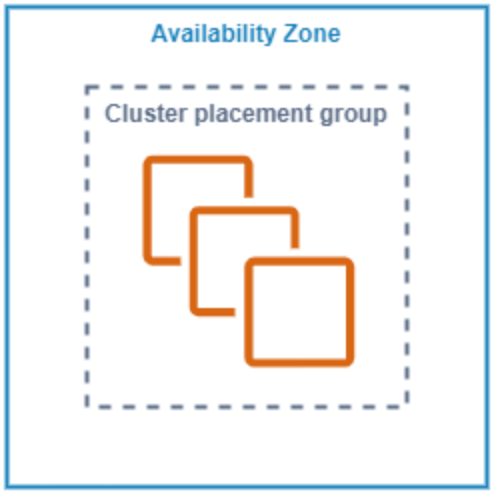
AWS offers different placement strategies which can suit the placement requirements of your application hosted in Amazon EC2.



Cluster Placement Group

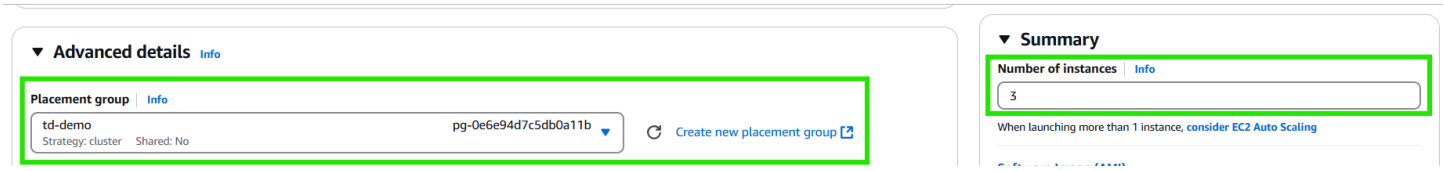
A cluster placement group is a logical group of instances within a single Availability Zone and instances from peered VPC in the same region. Through VPC Peering, you can still add instances from different Availability Zones to your cluster placement group.

Instances on the same cluster group have a low network latency and high throughput. A cluster placement group is beneficial to applications with a high volume of network traffic between their instances. To further maximize these network performance benefits, you can choose instance types with enhanced networking for your cluster placement group.



AWS recommends launching the instances for the cluster placement group through a single launch request. They also recommend using the same instance type for all the instances in the placement group to minimize

the chance of getting an insufficient capacity error. This error comes out when there is not enough hardware capacity to launch an instance. For example, when adding more instances to an existing placement group or adding instances with a different instance type. The capacity error can also be encountered when you stop and then start an instance again in a placement group.



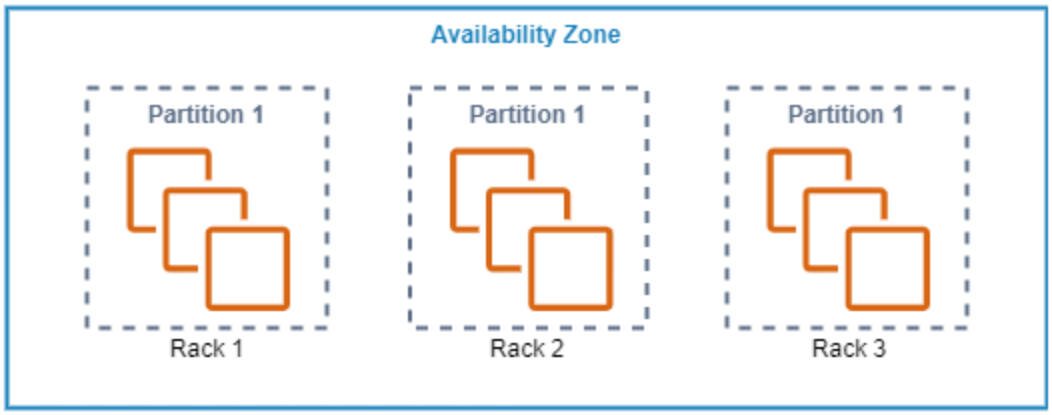
However, if you need to launch an instance to an existing placement group with running instances and encounter an insufficient capacity error, try to stop and start all of the running instances inside the placement group, then relaunch the instance. Doing so may force the instances to boot into new hardware capable of accommodating all the instance requests for the placement group.

Cluster placement groups are commonly used for High-Performance Computing (HPC) applications, like genomics, computational chemistry, financial risk modeling, machine learning, deep learning, etc.

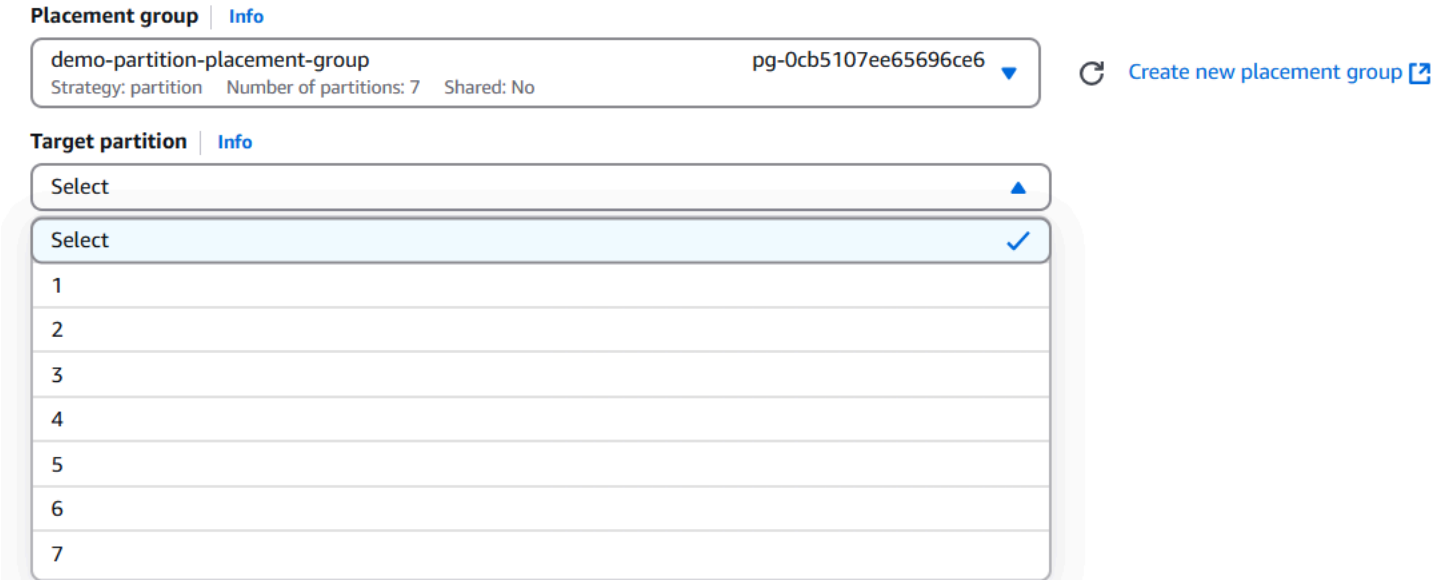
Partition Placement Group

A partition placement group spreads all instances into logical segments called partitions. Each partition has a dedicated rack with its network and power source. This placement strategy ensures that all partitions are isolated from each other, reducing the risk of correlated hardware failures.

Also, partition placement groups can have partitions from different Availability Zones in the same Region with a limit of seven partitions per AZ. The account limit determines the maximum number of instances. However, a maximum of two partitions is allowed for the partition placement group with Dedicated Instance.



When launching instances to the partition placement group, you can specify the specific partition.

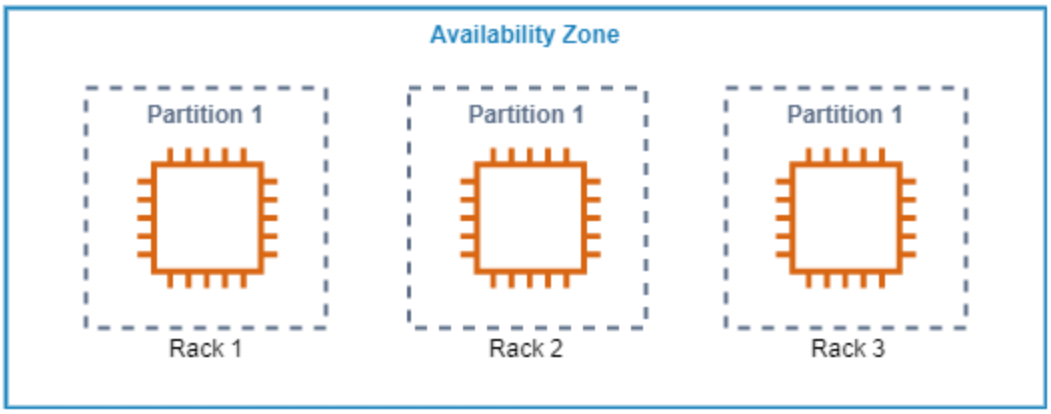


To achieve high availability for the application, we often go to multi-AZ deployment, but some applications are dependent on internode latency, thus making it unavailable for multi-AZ deployment. With a partition placement group, you can deploy this kind of application in a single Availability Zone but with improved performance and less chance for correlated hardware faults.

Applications like HDFS, HBase, and Cassandra are benefiting from this kind of placement strategy. Because they are topology-aware applications, they can use the topology information to make intelligent data storage decisions.

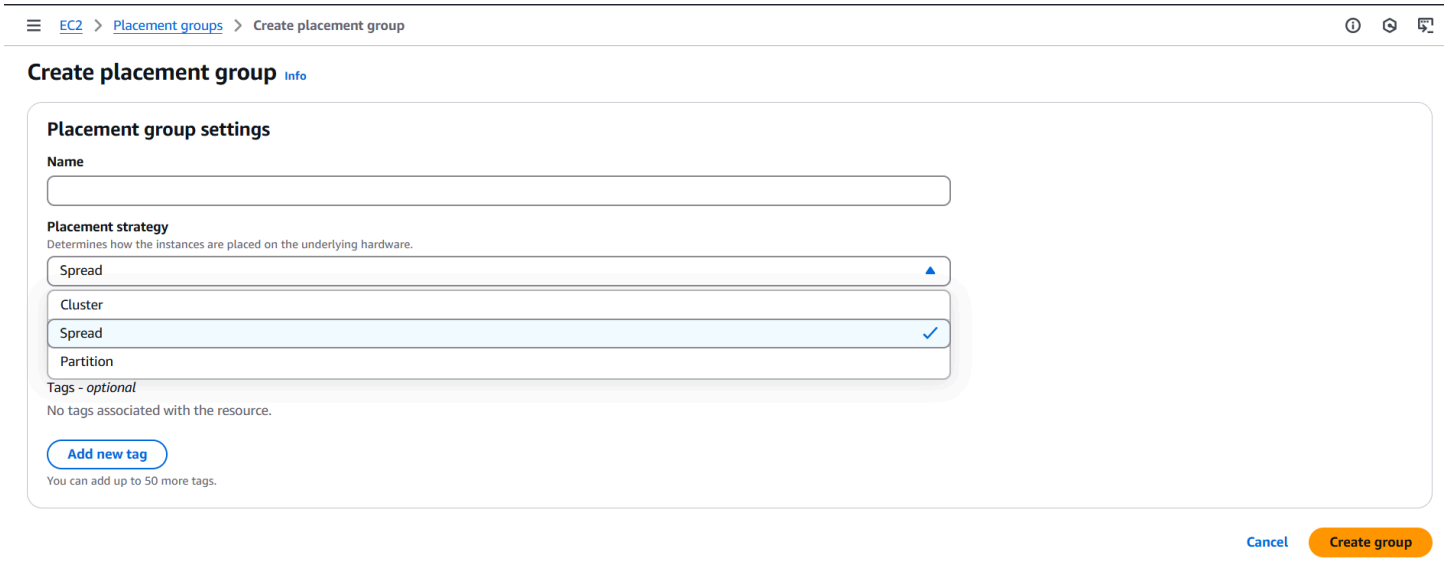
Spread Placement Group

A Spread placement group is a placement strategy that strictly hosts instances separately on a distinct rack that has an individual network and power source. Since all instances are hosted on distinct racks, you can freely have multiple instance types or add instances over time on your spread placement group.



Since instances on the spread placement group are isolated from each other, the chance of having hardware faults is reduced when compared to instances sharing the same rack.

Like partition placement groups, spread placement groups can also span on different Availability Zones with a maximum of seven running instances per AZ.



For the Partition and Spread placement group, there are times when a unique hardware is unavailable to accommodate all instance requests. When this happens, try to request again later as more hardware becomes available over time.

Reference:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html

EC2 Image Builder

EC2 Image Builder is an AWS service that automates the process of creating, managing, and deploying machine and docker images both for your AWS environment and on-premises. You can keep your images updated through the image builder, automate image customization, validate image integrity and functionality through testing, and deploy images in different AWS regions. The image builder is pretty straightforward; it lets you create an Image Pipeline, configure an Image recipe, define the infrastructure, and set the image distribution.

Image Pipelines

To automate the creation of images, AWS allows you to create a pipeline where you can configure the necessary components of your custom images. The image creation will run based on the defined build schedule and frequency or can be manually run.

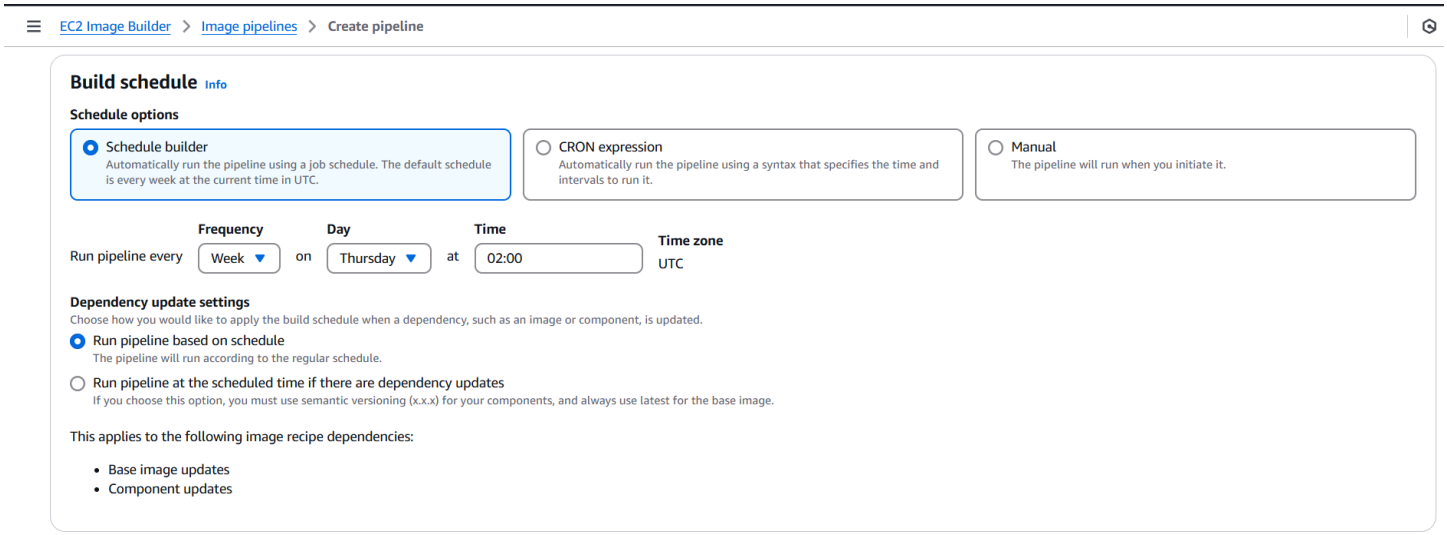
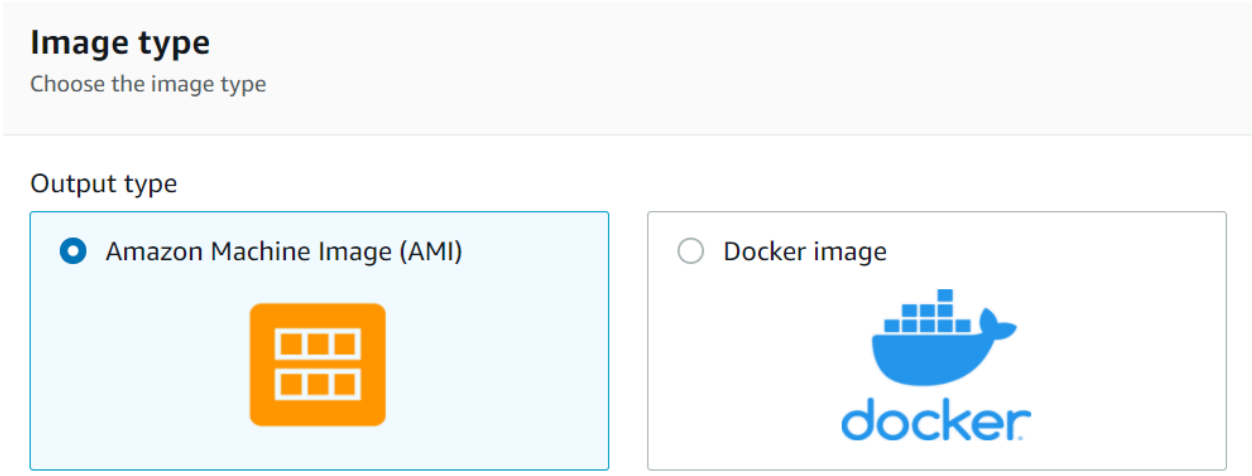


Image Recipes Configuration

Image recipes are where you define the customization and testing of your images. Image recipes are reusable and have version control. It consists of the following components.

Source Images

The source image will be the baseline of your custom image. Image builder supports the customization for Amazon Machine Image (AMI) and Docker image. For AMI, this can be AWS-managed images or a custom AMI. Likewise, for Docker images, it can be AWS-managed images, an ECR image, or a public image from Docker Hub.



You can select from different operating systems and versions; availability depends on the image type.

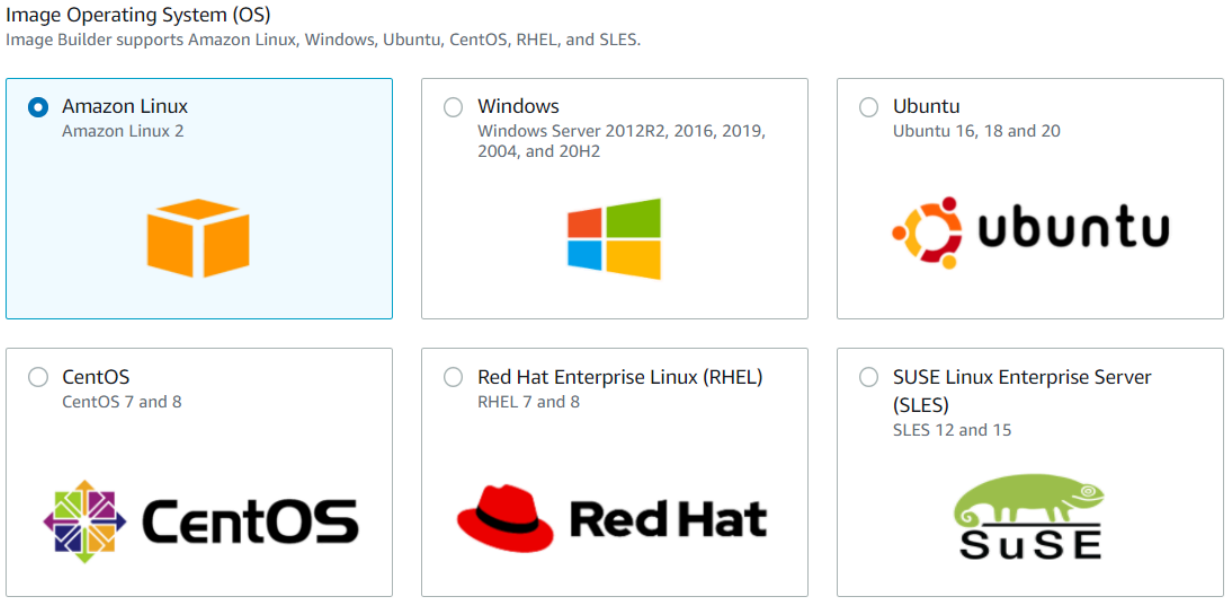
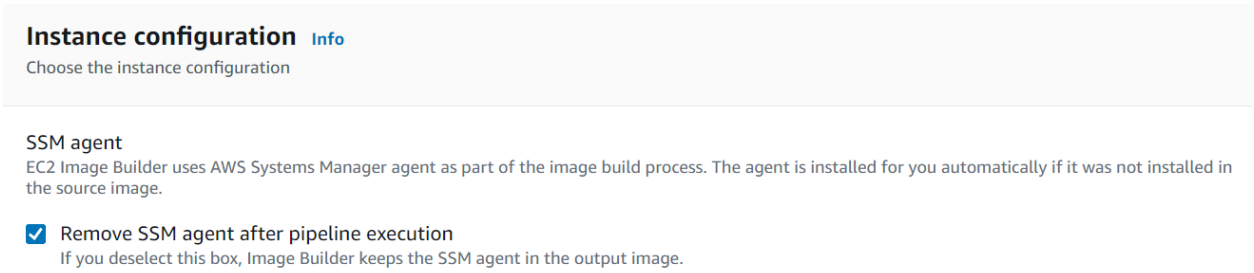
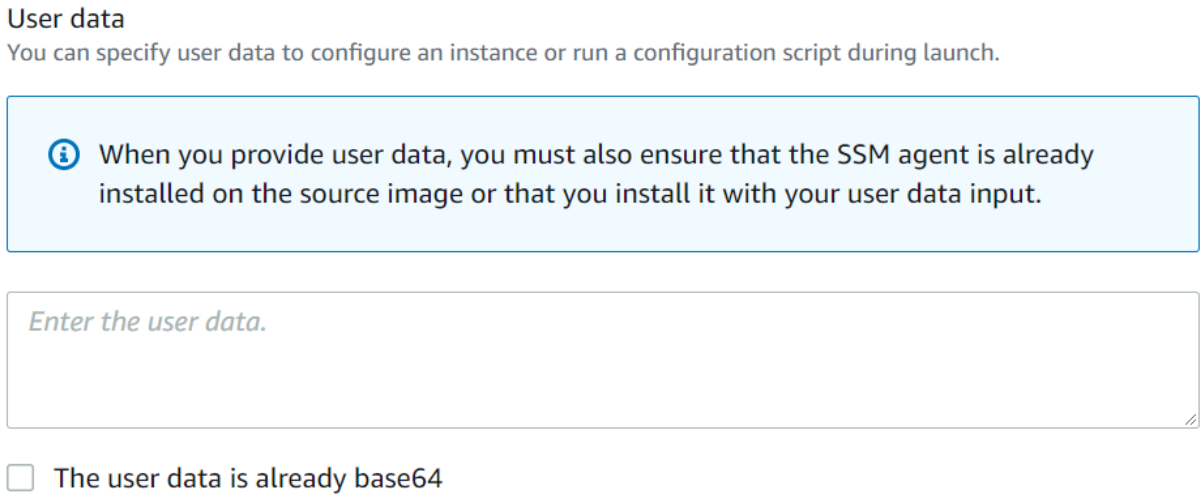


Image builder installs SSM Agent during the build process, but you can remove the agent after the pipeline execution.

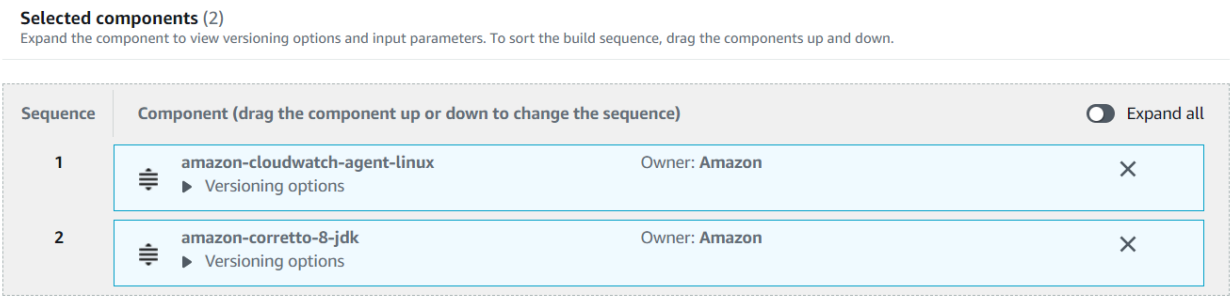


Should it be necessary to run a command on the instance launch, you can set it on the User Data. Note that defining User data requires your source image to have the SSM Agent pre-installed or that you include SSM Agent installation on the User Data.

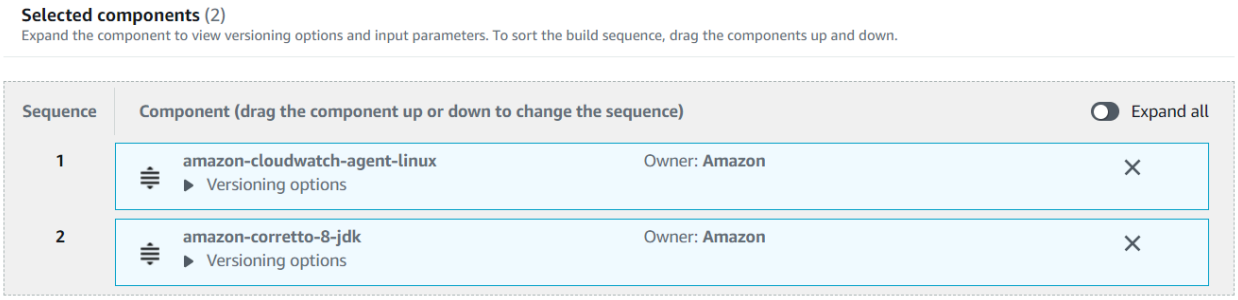


Build and Test Components

A build component installs software packages to your source image. You can select from Amazon-managed build components, share build components to your AWS account, or create a new one. See the example Amazon-managed build components below.

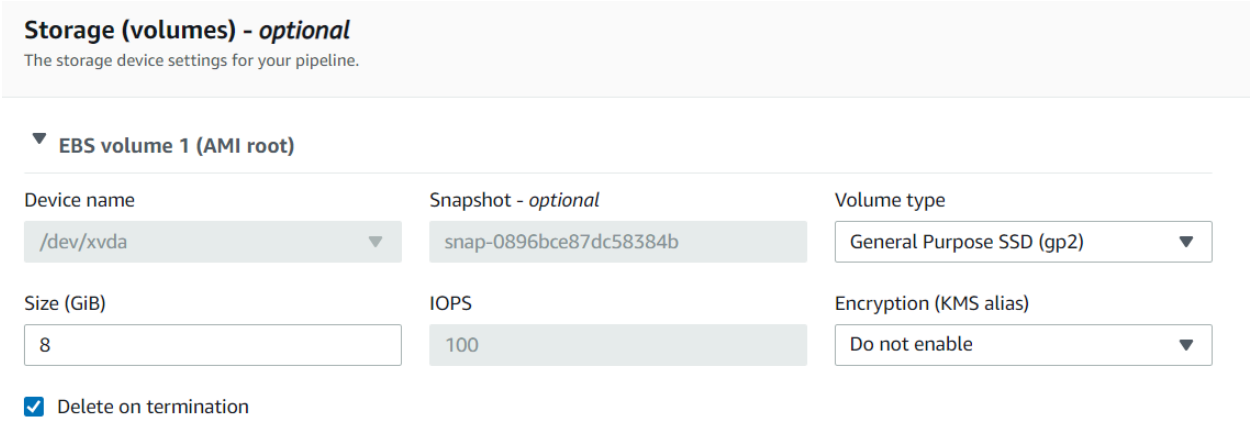


Test Components are optional, but it's a better option to configure this to validate the integrity and functionality of the output image. You can also use Amazon-managed, shared, or create a new test component. See the example Amazon-manage test components below.



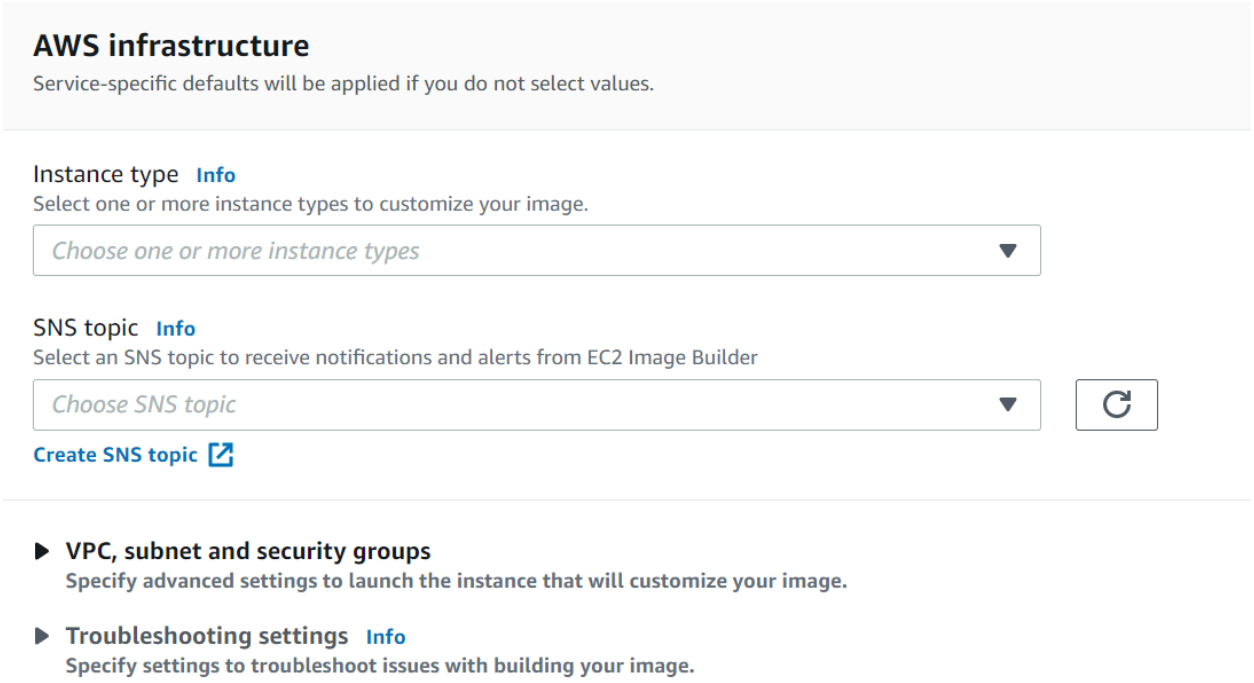
Storage

Storage configuration is optional. You can configure this during the instance launch.



Infrastructure Configuration

The Infrastructure configuration is an optional configuration on the image pipeline. You can configure the Instance Type, VPC settings, IAM role, and Tags for the output image. A notification can also be published using SNS.



Besides the default IAM policies that the image builder uses, the configured IAM role should also have the necessary permissions to execute all the build and test components defined on the image recipe.

Default IAM Policies for Image Builder:

- *EC2InstanceProfileForImageBuilder*
- *EC2InstanceProfileForImageBuilderECRContainerBuilds*
- *AmazonSSMManagedInstanceCore*

Distribution Settings

You can configure the image deployment on the Distribution settings. You can choose multiple AWS Regions as image destinations. For Amazon Machine Images, you can configure the output image name, AMI sharing, and the license and launch template configuration. For the docker images, you need to specify the Regions and ECR repository name.

Reference:
https://docs.aws.amazon.com/imagebuilder/latest/userguide/what-is-image-builder.html

Amazon EC2Rescue

While AWS takes care of the underlying infrastructure for EC2, customers are responsible for configuring, maintaining, and troubleshooting their instances.

EC2Rescue for Windows Server

EC2Rescue for Windows Server is a downloadable tool for Windows Server instances to help you diagnose and troubleshoot issues. You can also use EC2Rescue to detect potential problems in your current instances.

Diagnose and Rescue an Offline Instance

EC2Rescue scans and diagnoses the Amazon EBS root volumes of the problematic instances. To do this, EC2Rescue requires a host instance where it will be installed. The EBS root volume should be detached from the problematic instance and attached to the EC2Rescue instance host.

Reminders:

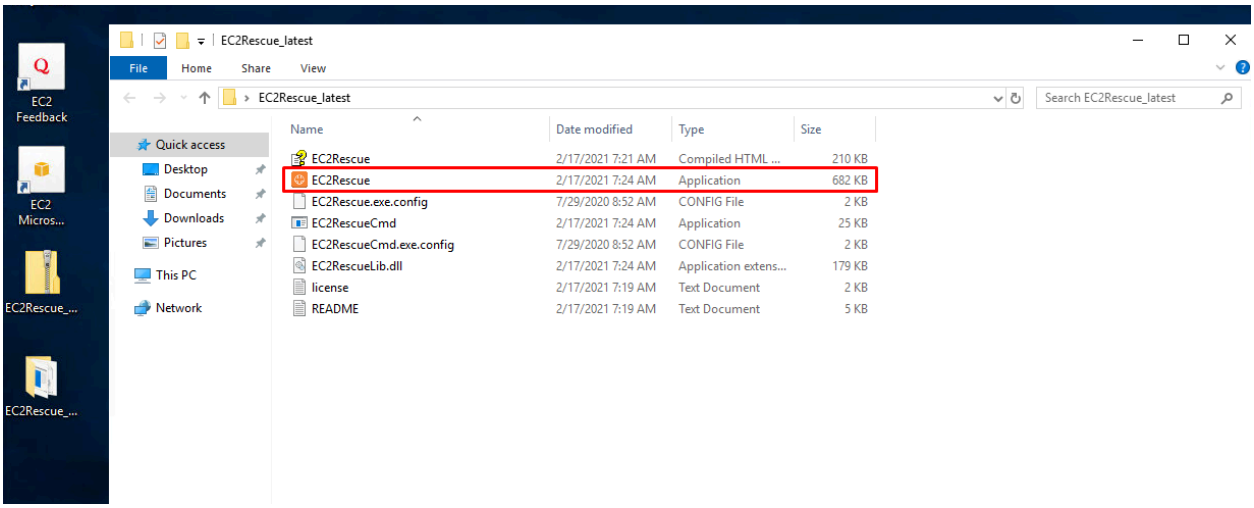
- The EC2Rescue tool only runs on Windows Server 2008 R2 or later and requires .NET Framework 3.5 SPI or later.
- The EC2Rescue instance host should also be accessible using an RDP connection.
- The instance where EC2Rescue is installed and the instance to be diagnosed should reside on the same Availability Zone.

The following instructions will guide you on how to check an instance using EC2Rescue.

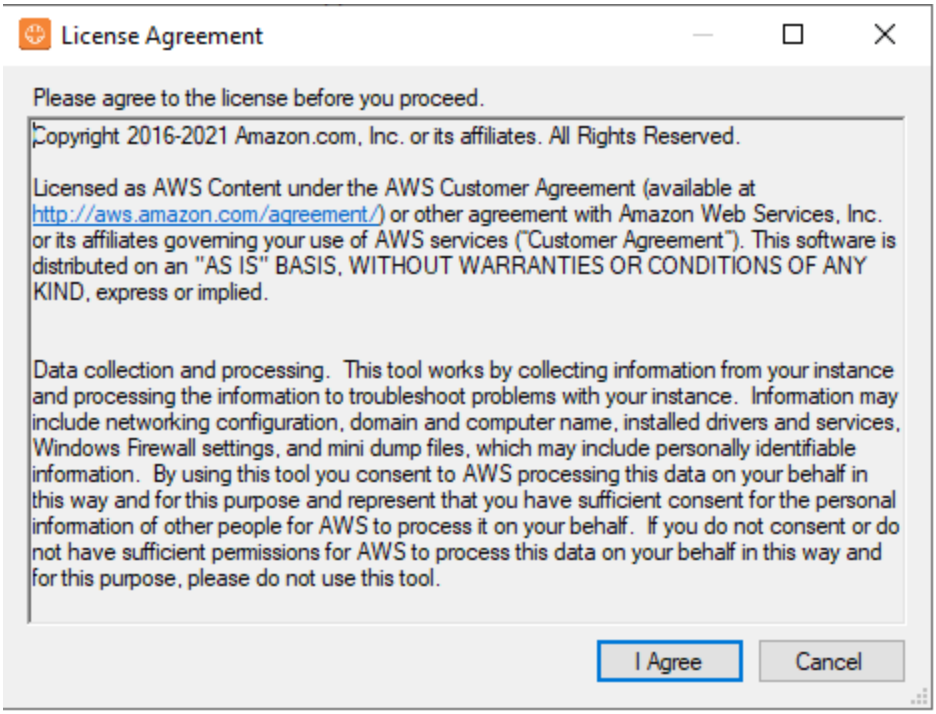
1. Connect to the EC2Rescue host and download the tool here using a browser or using the PowerShell command below.

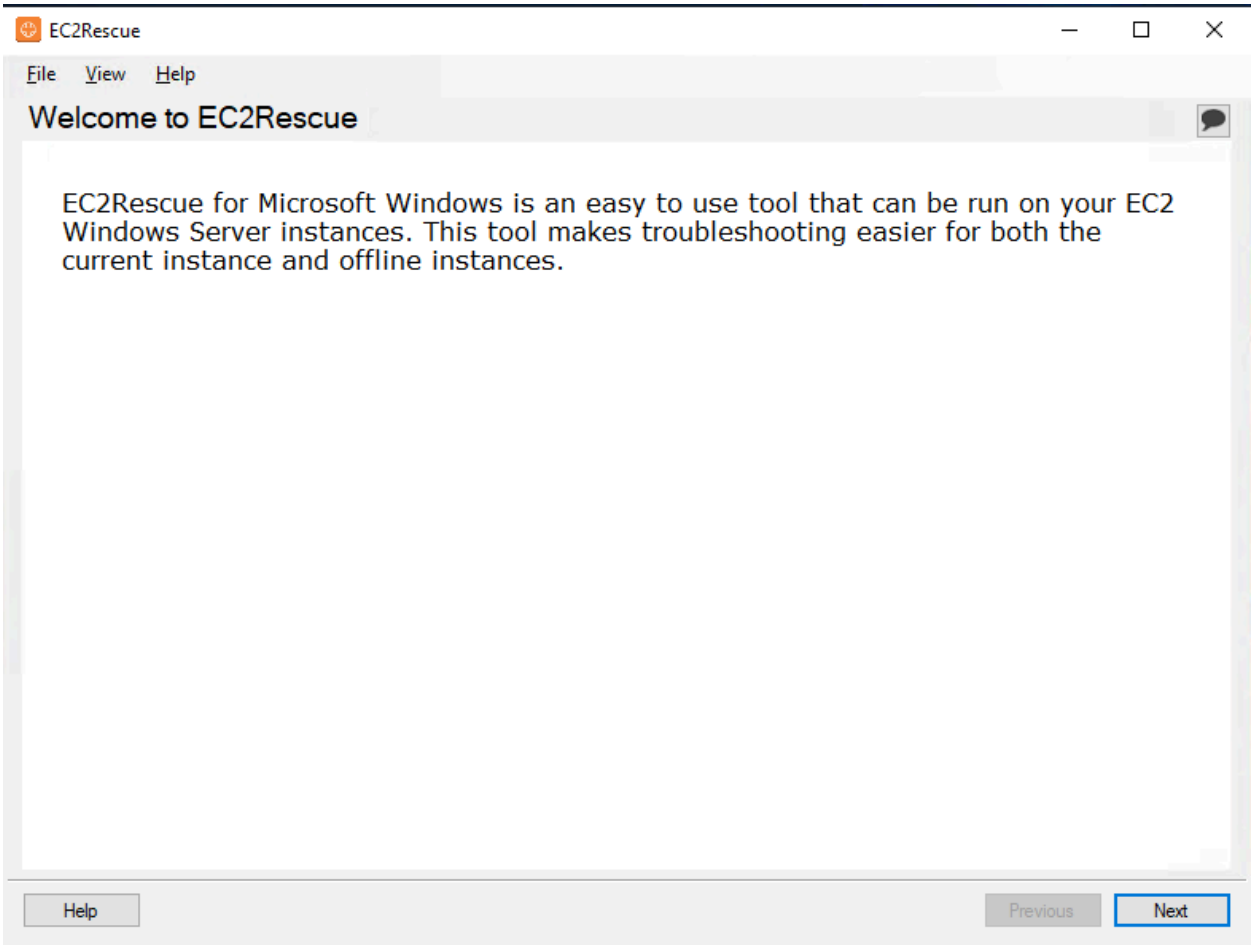
```
Invoke-WebRequest https://s3.amazonaws.com/ec2rescue/windows/EC2Rescue_latest.zip -OutFile
$env:USERPROFILE\Desktop\EC2Rescue_latest.zip
```

2. Extract the downloaded zip file. Once extracted, run the EC2Rescue application.

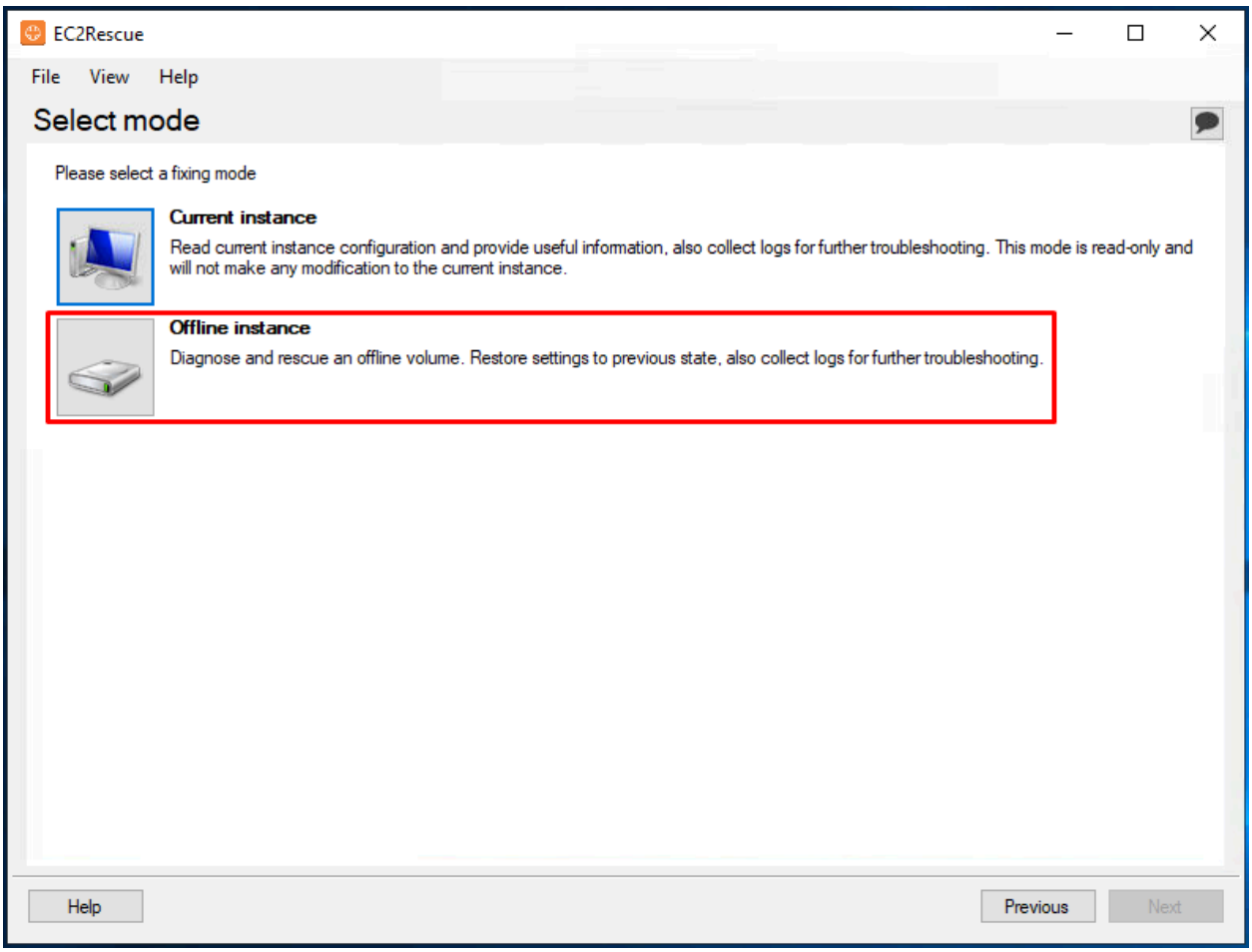


3. Click ***I agree*** on the license agreement and click ***Next*** on the Welcome screen.

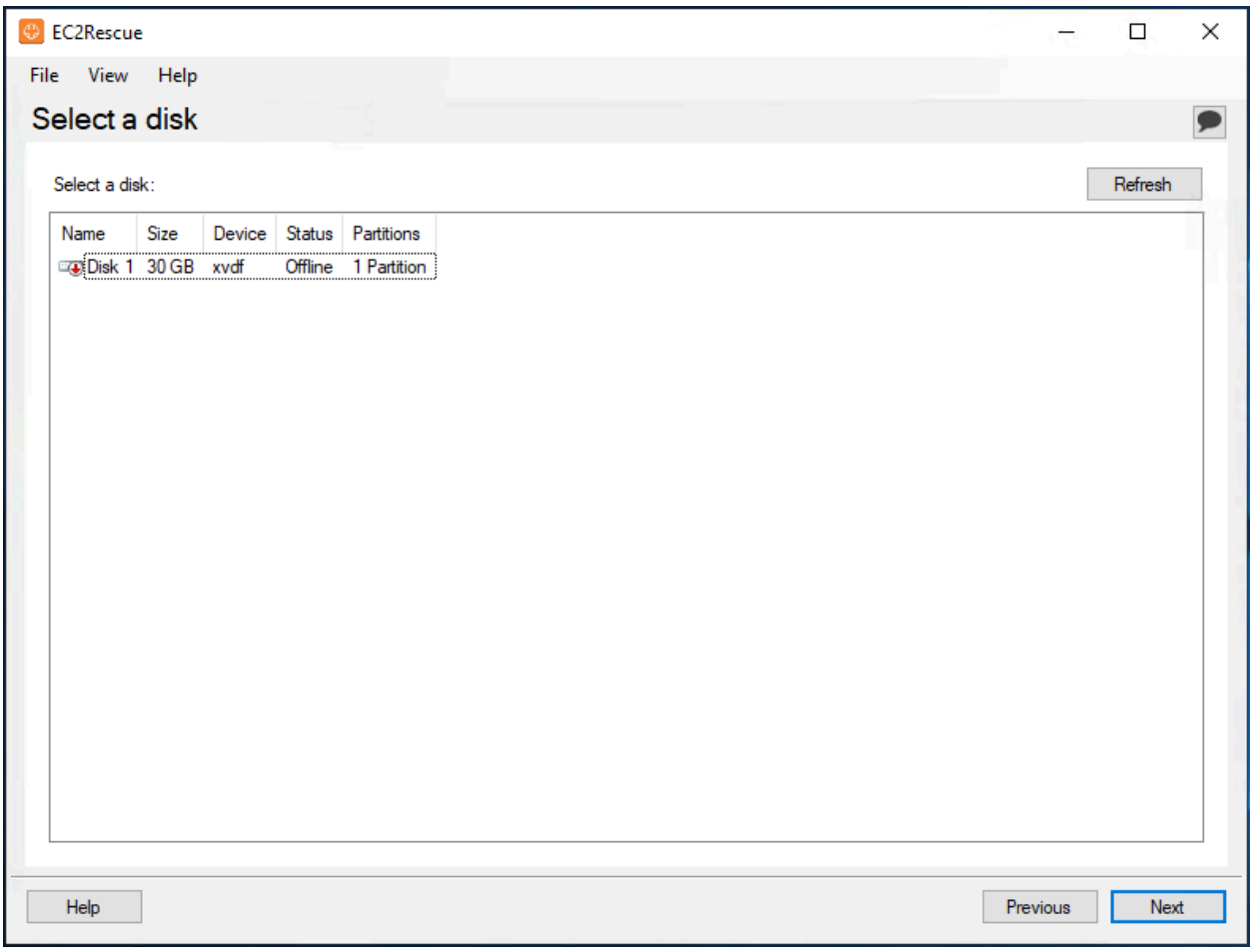




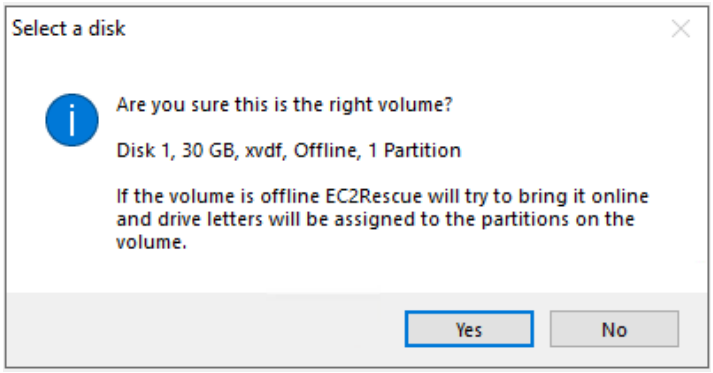
4. Select ***Offline instance*** and click ***Next***.

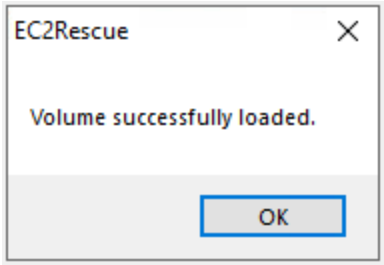


5. Select the ***EBS volume*** of the problematic instance and click ***Next***. If you are checking multiple root volumes, note the device name when attaching the volumes to the EC2Rescue host.

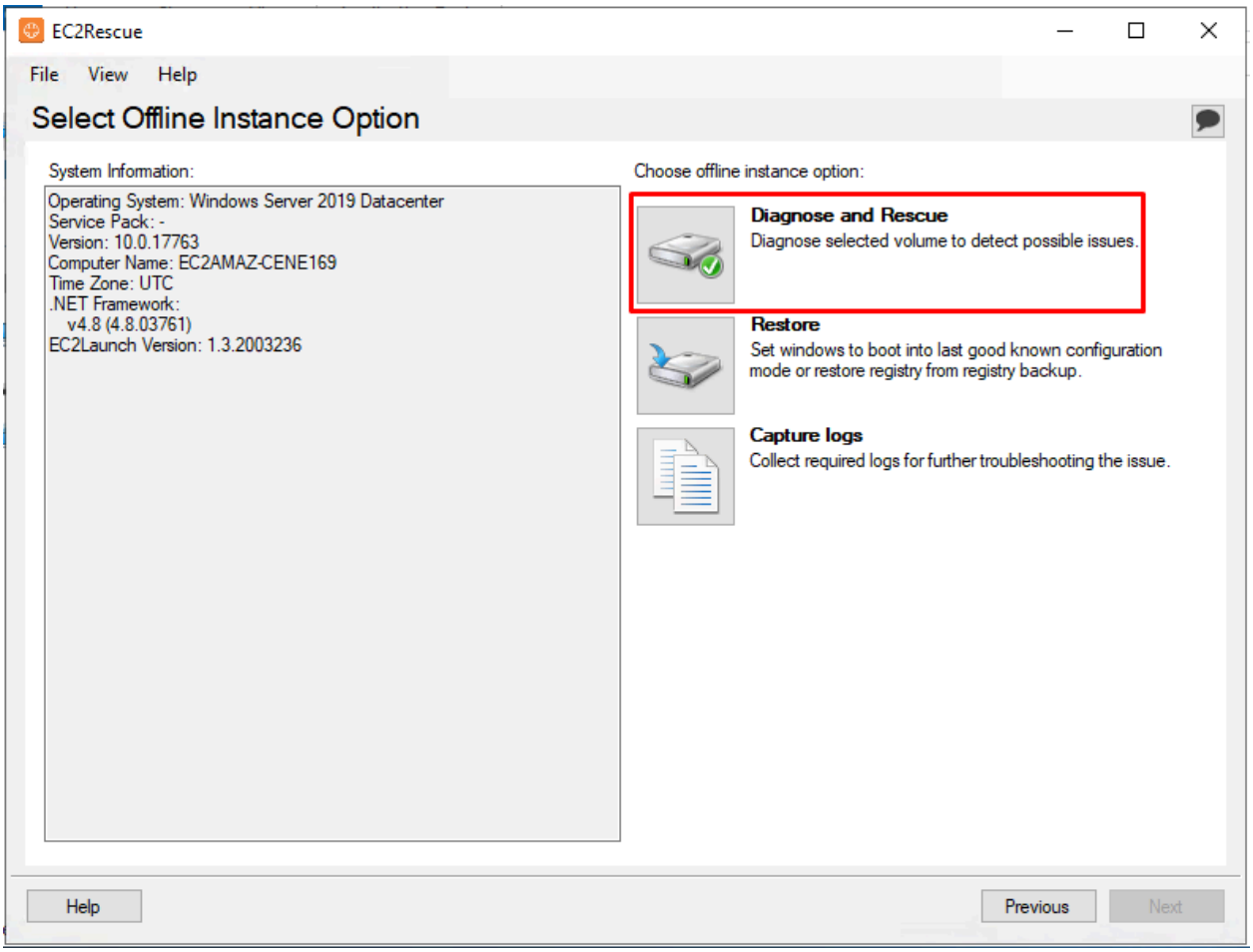


6. Click ***Yes*** to confirm. A popup window will show once the EBS volume is successfully loaded.

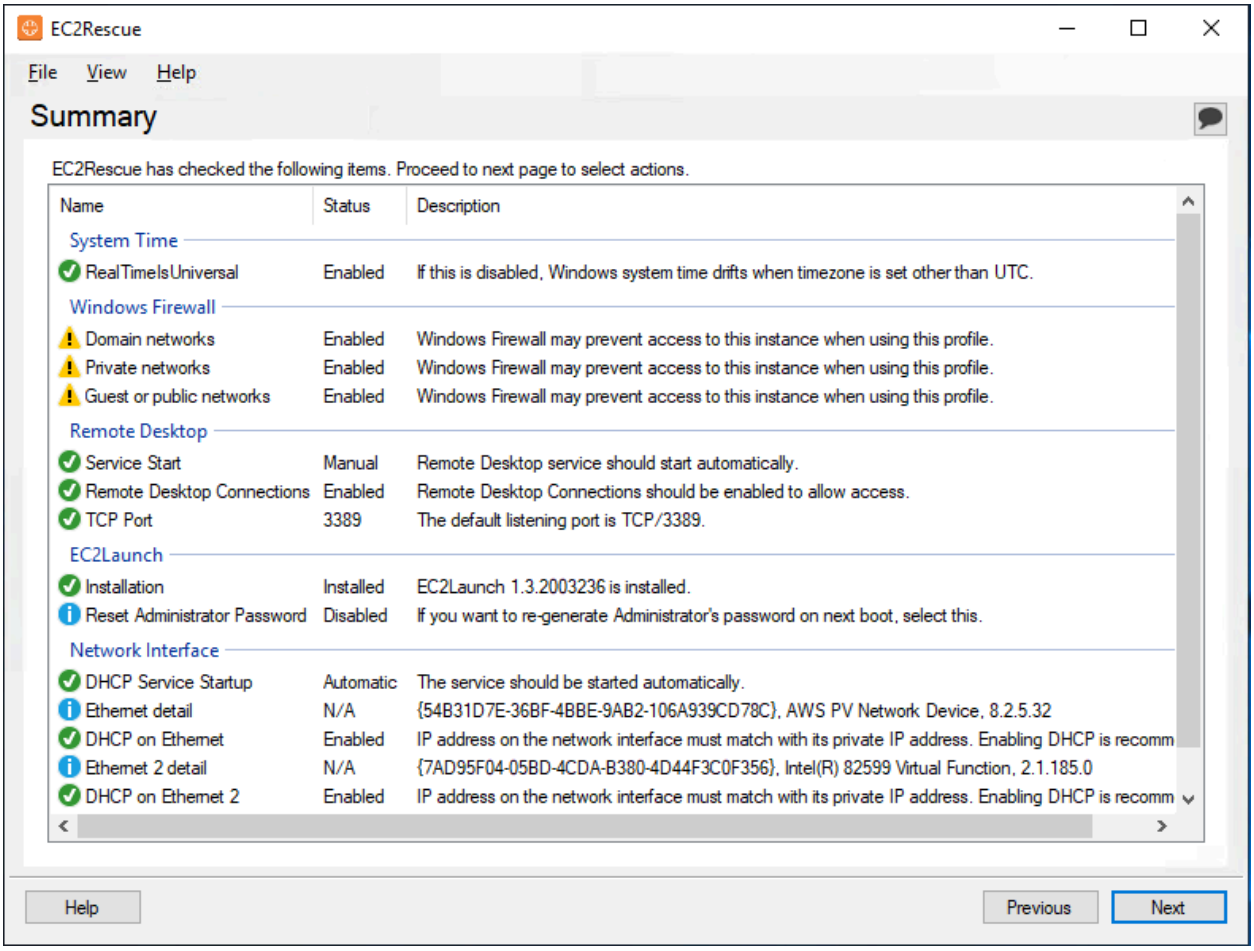




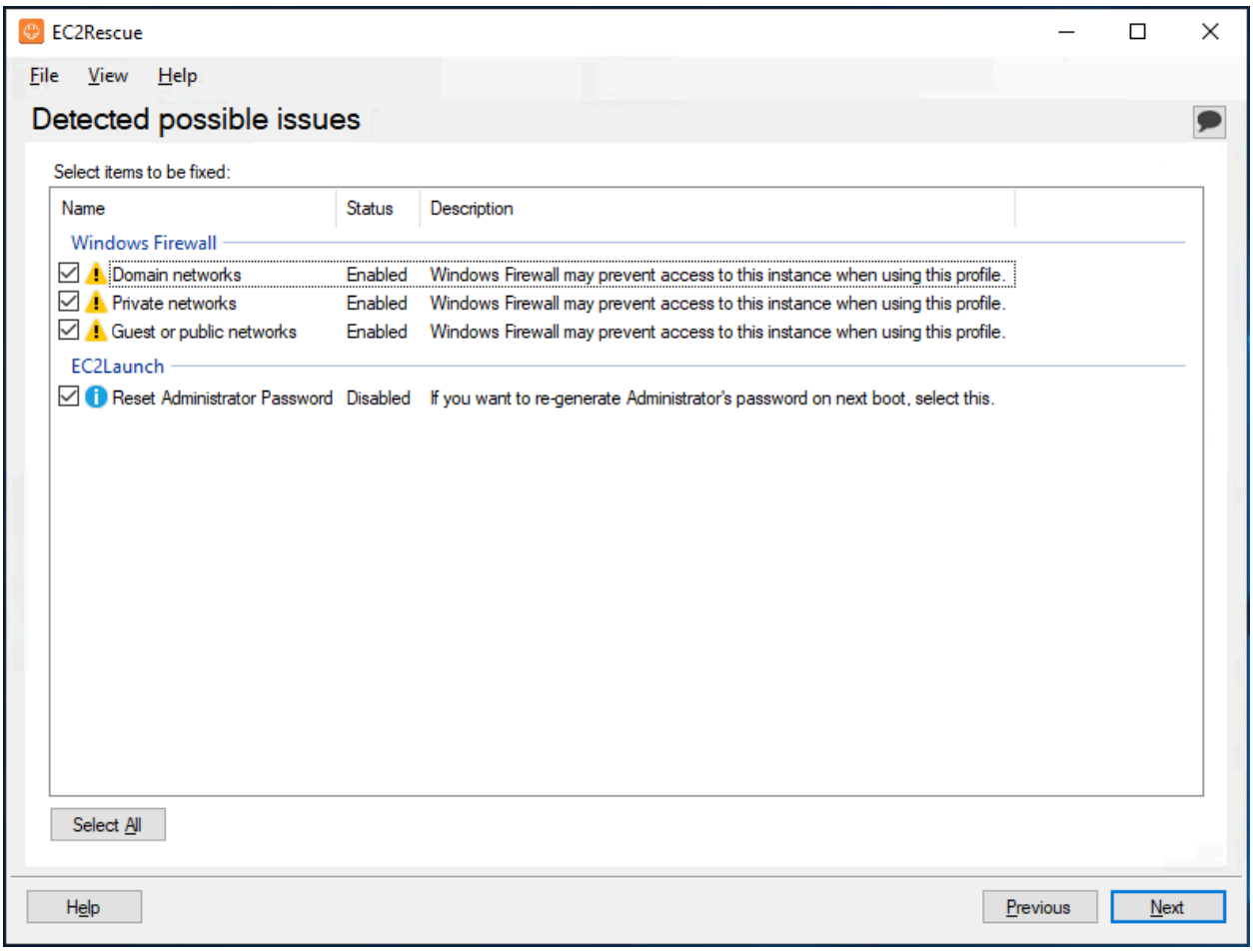
7. Once the volume is loaded, the EC2Rescue tool will display system information of the instance. You will also see different offline instance options. In this case, select ***Diagnose and Rescue*** to proceed.



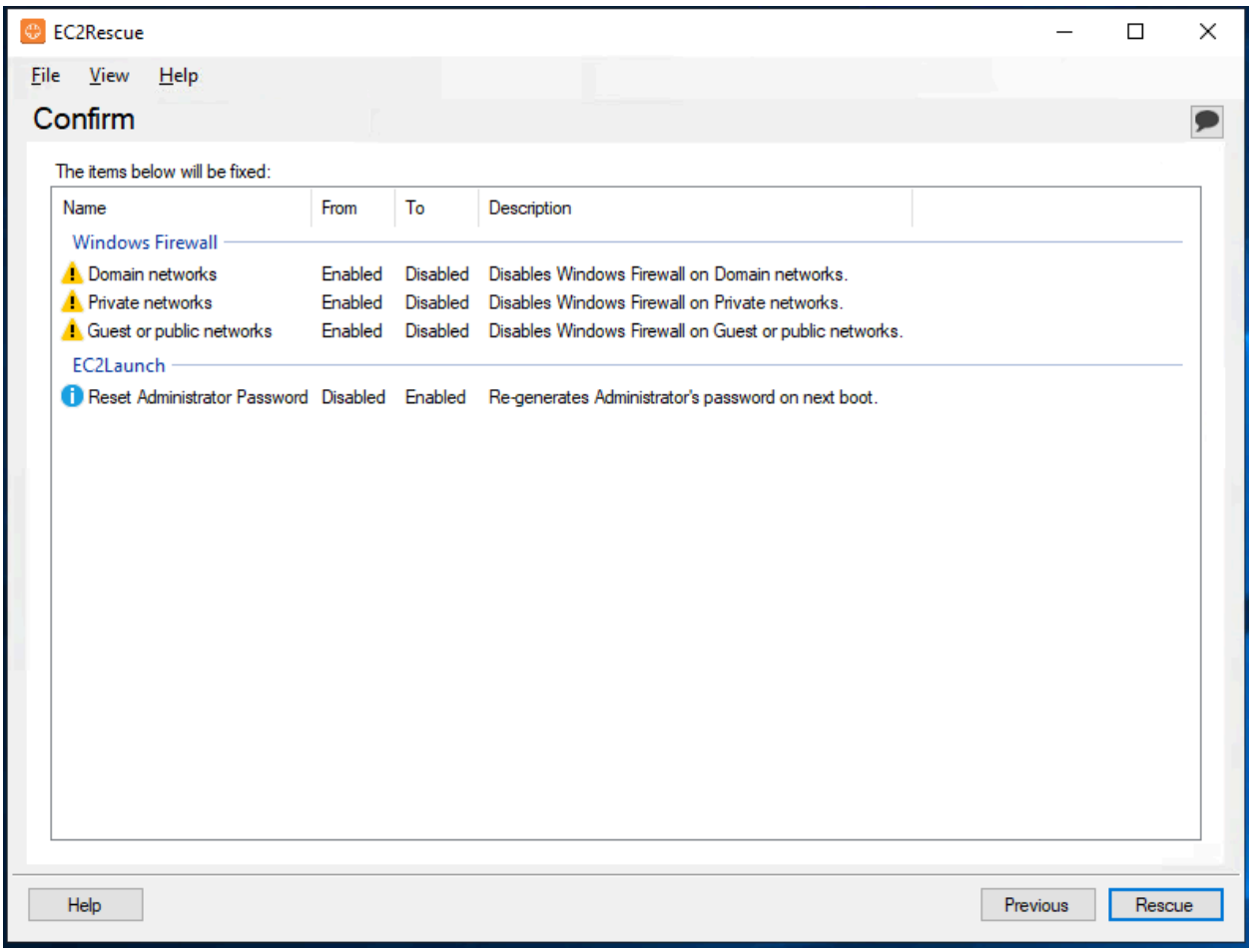
8. The EC2Rescue tool will now start scanning and diagnosing the volume. Once the diagnostic is done, it will summarize the necessary configurations, including their status and description. Click ***Next*** to proceed.



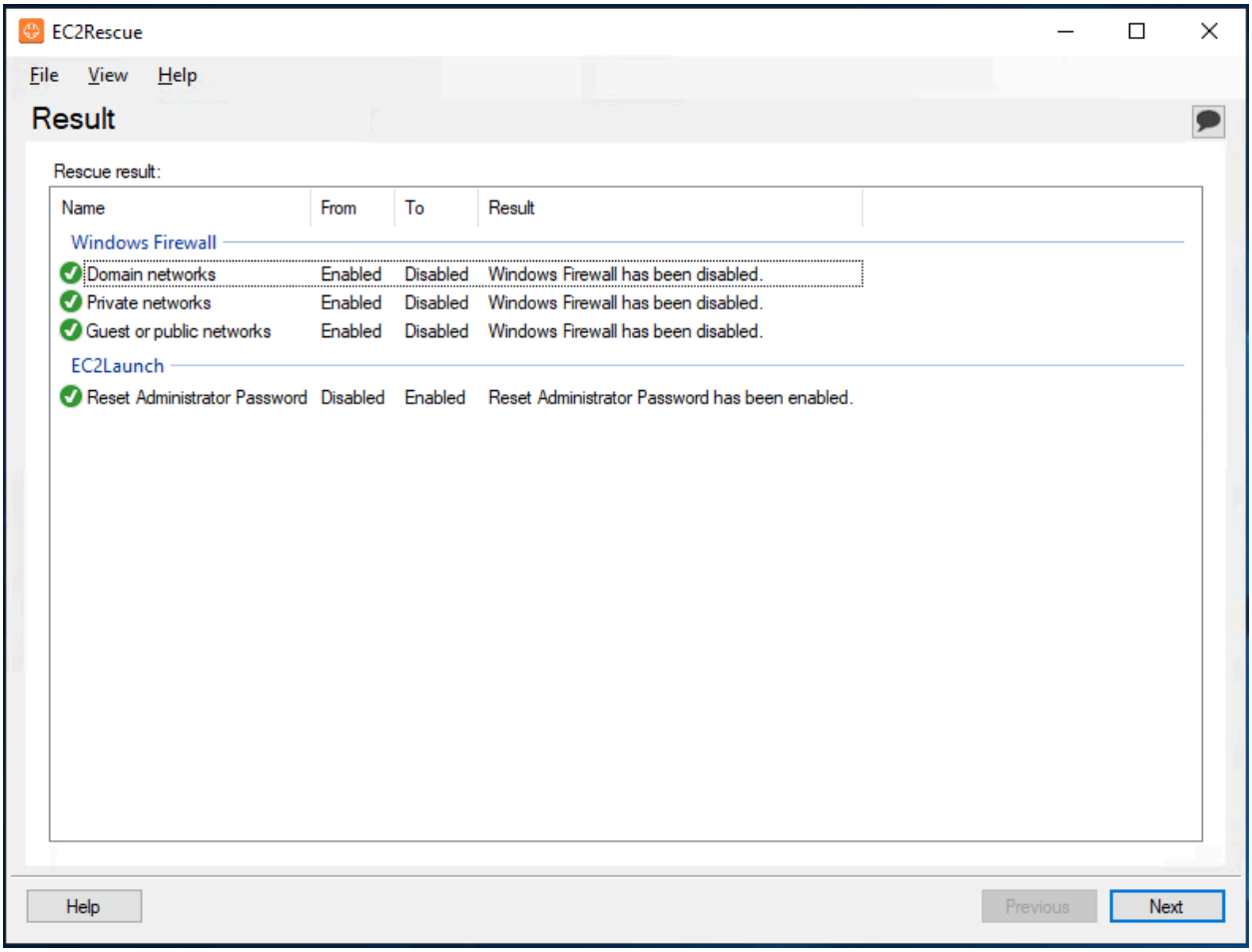
9. EC2Rescue will give you a list of potential issues of the instance. From this, you can select the fixes you find necessary for your instance. Click ***Next*** to proceed.



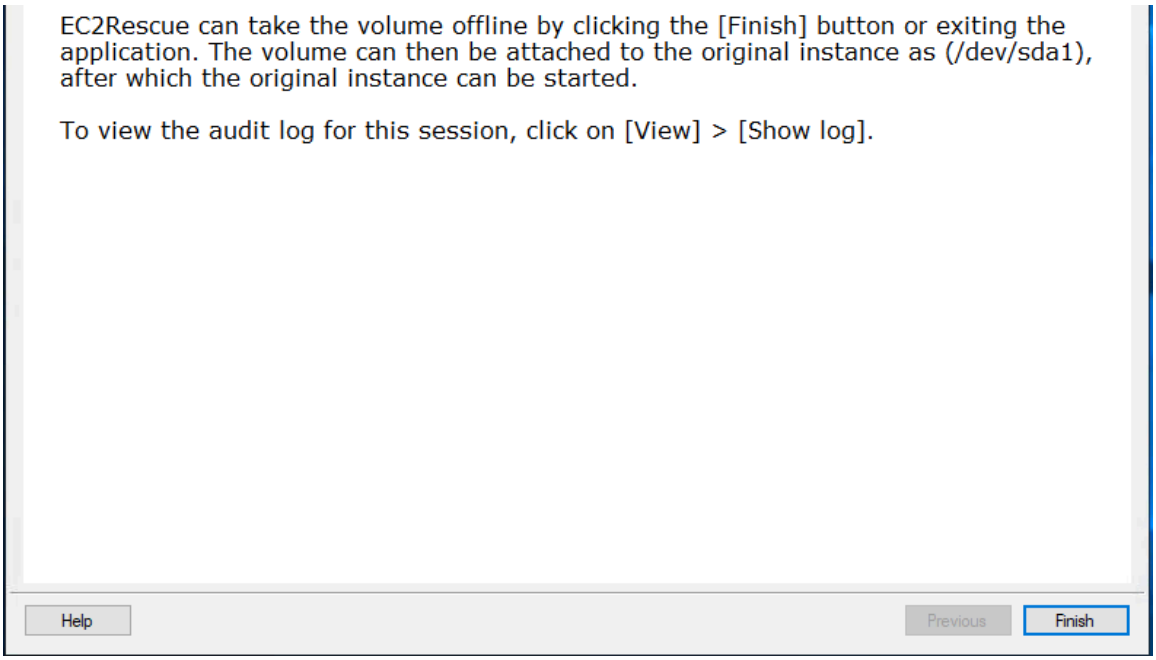
10. Once confirmed, click ***Rescue***.



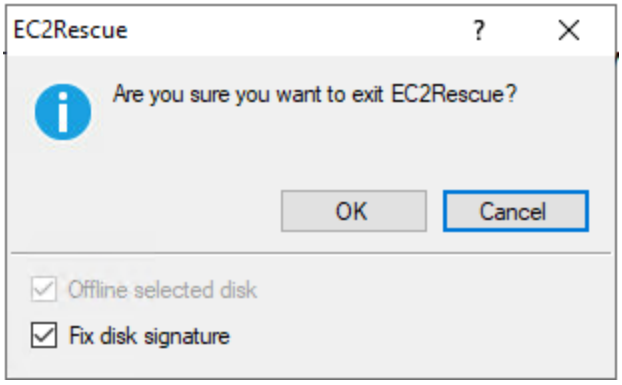
11. Click ***Next*** to continue applying the changes.



12. Click ***Finish***.



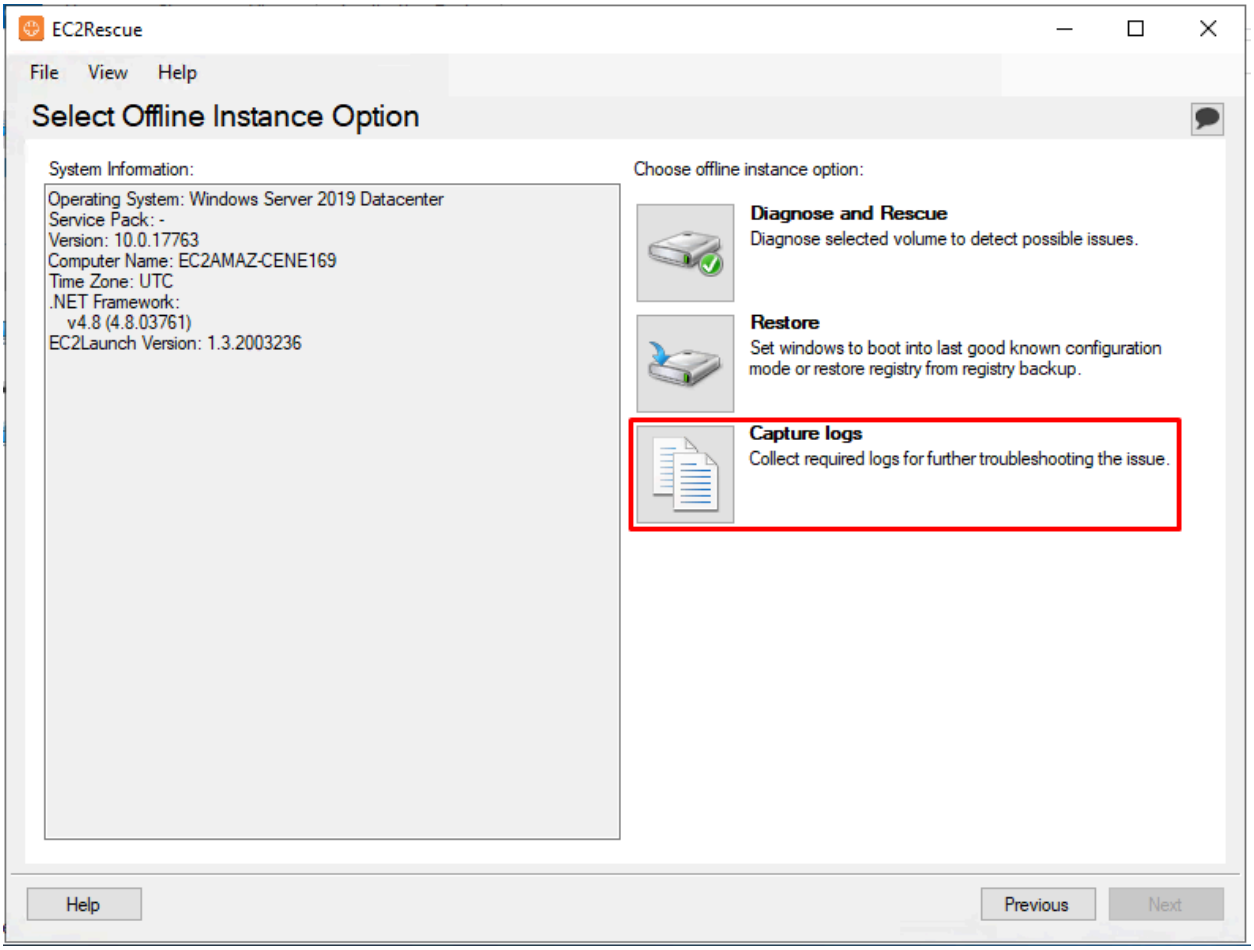
13. When exiting the EC2Rescue tool, check the ***Fix disk signature*** to resolve the boot issue caused by disk signature mismatch. The offline selected disk is also checked when the Fix disk signature is checked.



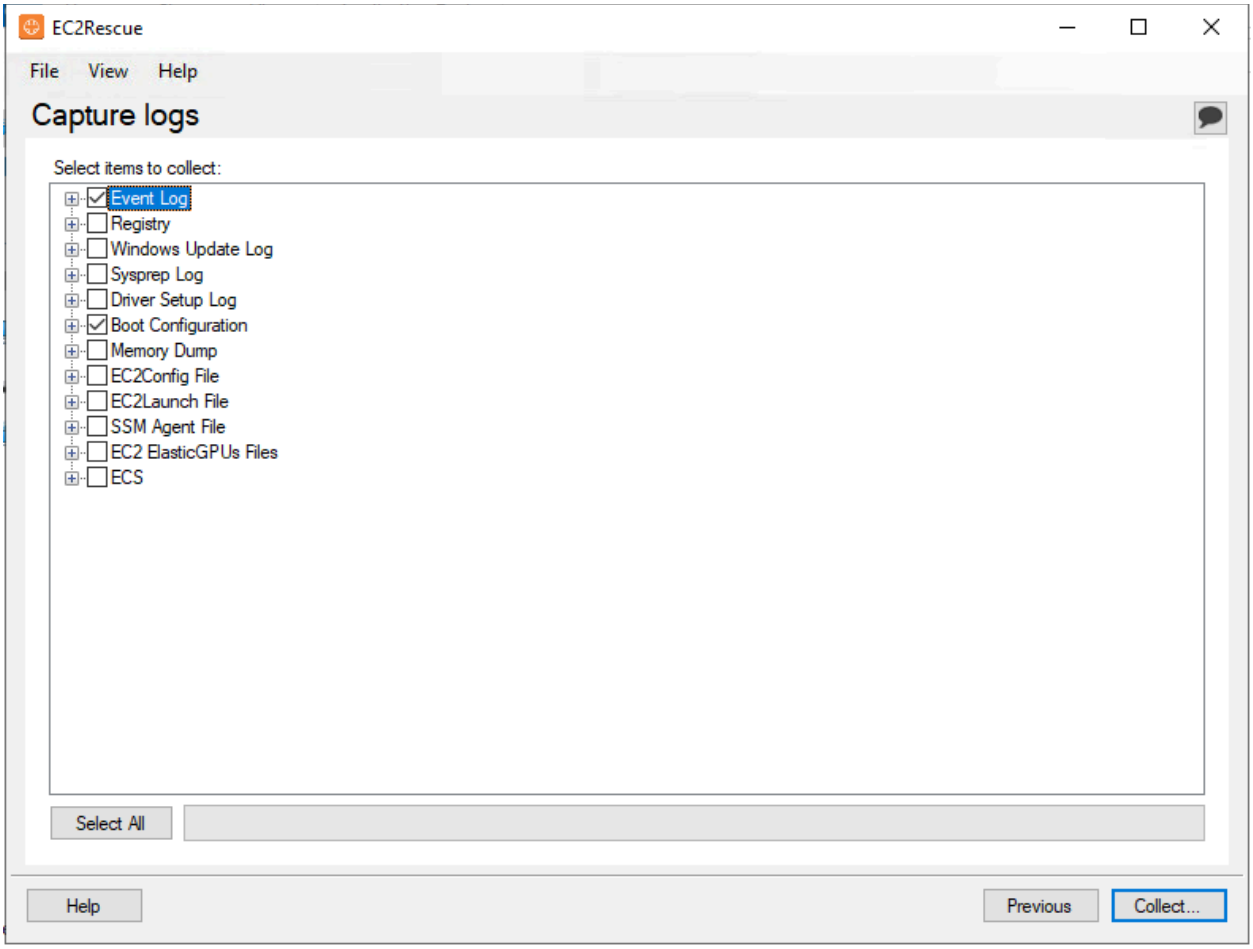
Collecting Logs from an Offline Instance

To collect logs from the instance, select Capture Logs from the offline instance option. The following instructions will guide you on how to collect logs from an offline instance.

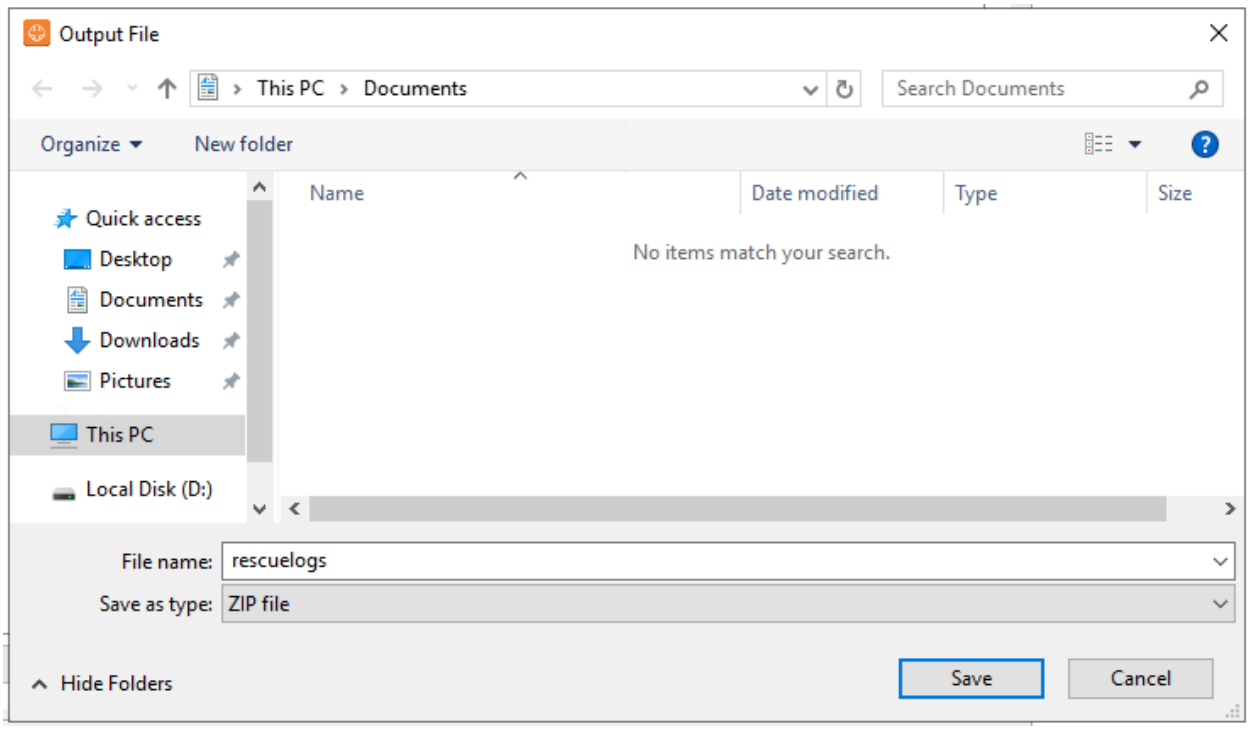
1. Select ***Capture logs*** from the offline instance option.



2. Select the logs you want to collect. Click ***Collect*** to proceed.



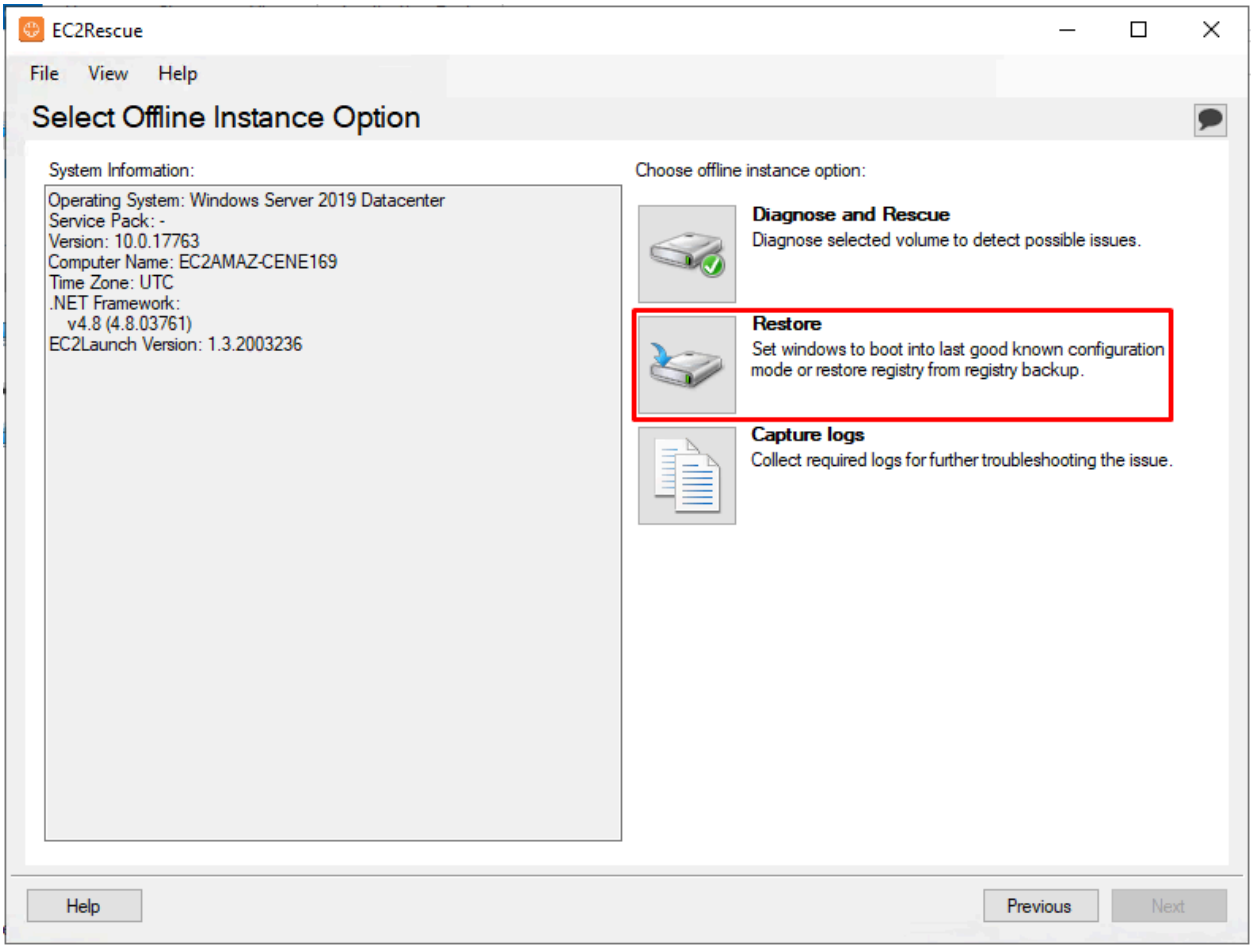
3. Save the logs as a zip file.



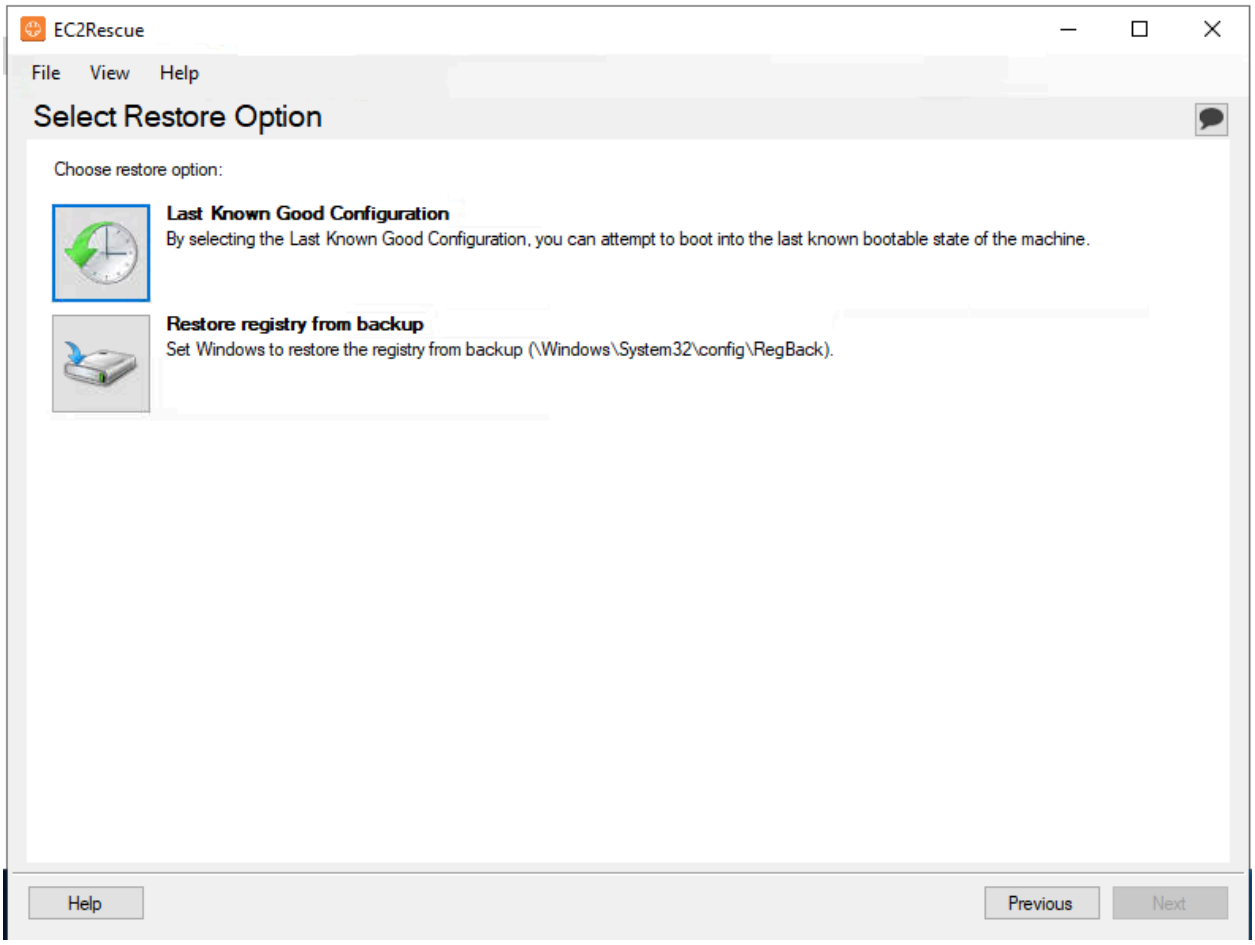
Restore Options for an Offline Instance

Using the EC2Rescue tool, you can also restore an instance. EC2Rescue provides two restore options: ***Last Known Good Configuration*** and ***Restore registry from backup***.

You restore an offline instance by selecting Restore on the offline instance options.

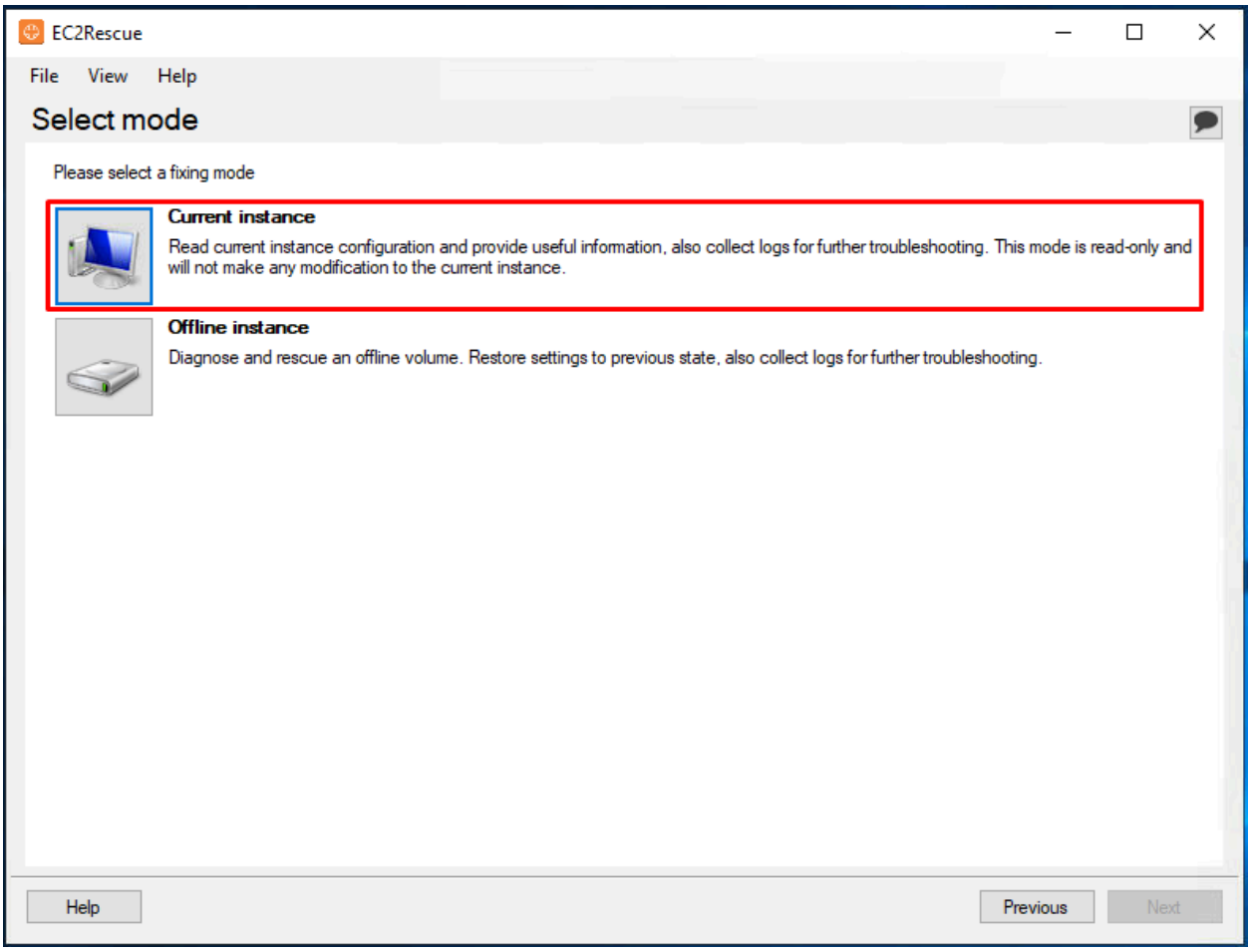


Then select a Restore option to begin the restoration.

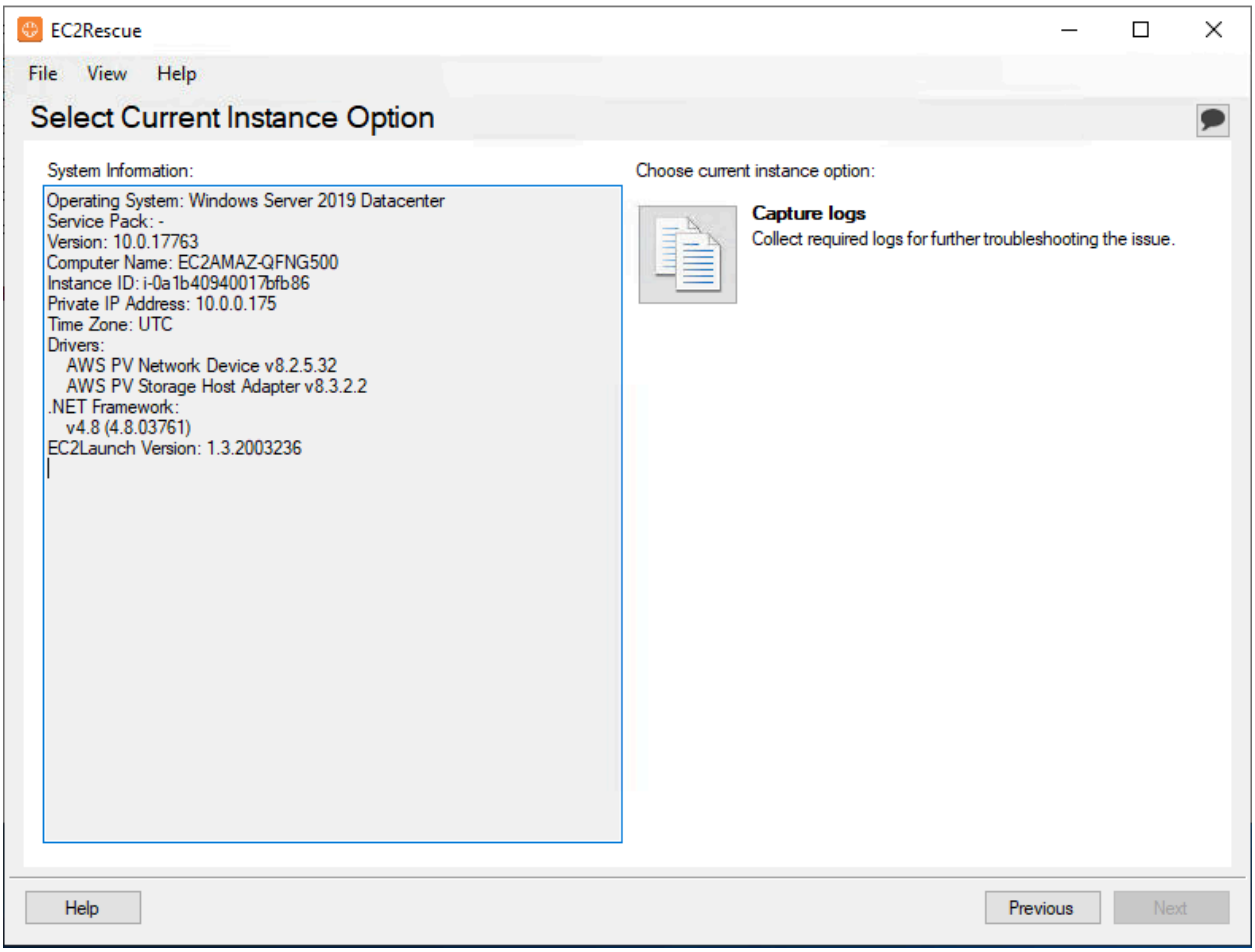


Checking the Current Instance

EC2Rescue also allows you to check the configuration and logs of its host instance. To check the current instance, select ***Current Instance*** mode.

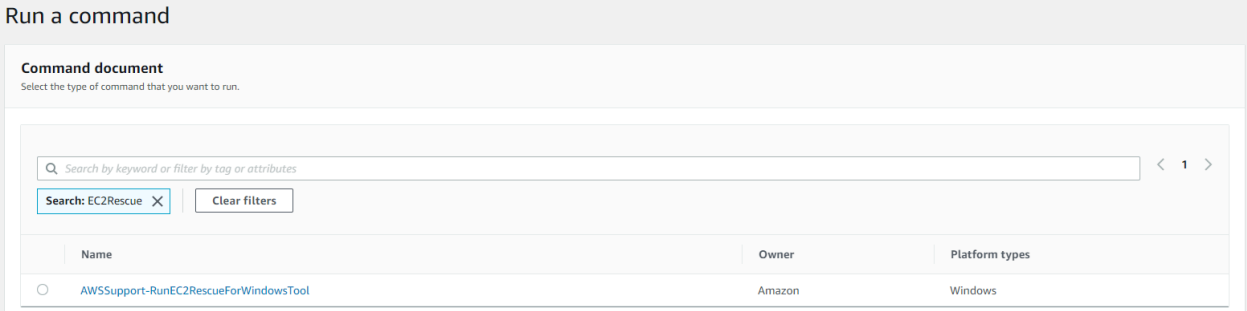


The EC2Rescue tool only allows you to capture logs of the current instance, unlike the offline instance mode where you can run a diagnostic and restore.



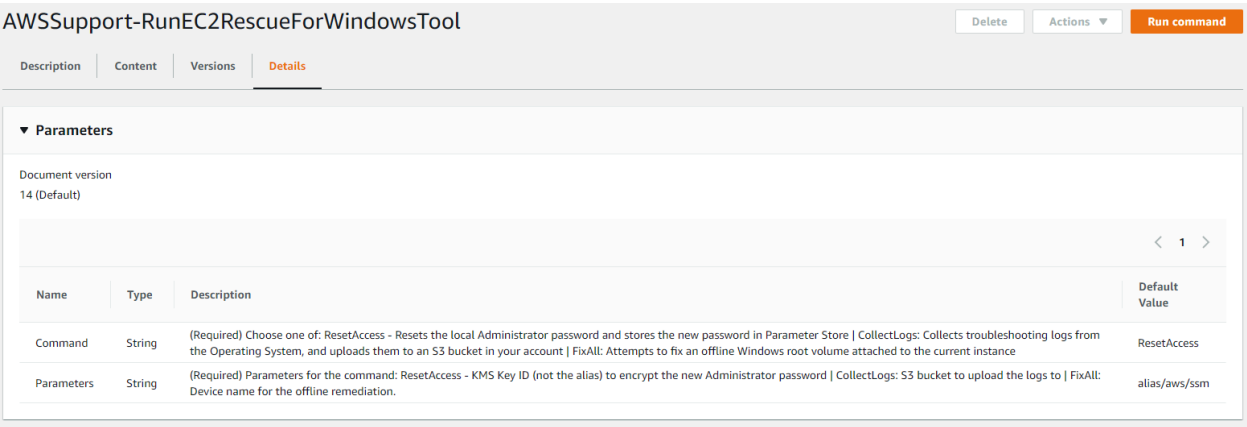
EC2Rescue for Windows on Systems Manager

You can also use EC2Rescue using the Run command in AWS Systems Manager. The command document for EC2Rescue for Windows Server is named *AWSSupport-RunEC2RescueForWindowsTool*. When you run this command, it will download and verify EC2Rescue for Windows Server and install a PowerShell module that allows you to run EC2RescueCmd. You can find this command document by simply typing EC2Rescue on the command search.



Given that your windows instance is added as Managed Instances on the Systems Manager, you will be able to run and do the following commands:

1. ***ResetAccess*** - resets the password of the local windows administrator and creates a new password in the Parameter Store. Requires KMS Key ID to encrypt new administrator password.
2. ***CollectLogs*** - collects and uploads valuable logs from the operating system to an S3 bucket.
3. ***FixAll*** - fix an offline root volume attached to the instance.



EC2Rescue for Linux

You can diagnose and troubleshoot EC2 instances running in Linux using EC2Rescue for Linux. This tool has over 100 modules on its library ready to use in checking Linux instances. EC2Rescue for Linux supports the following Linux distributions.

- Amazon Linux 2
- Amazon Linux 2016.09+
- SUSE Linux Enterprise Server 12+
- RHEL 7+
- Ubuntu 16.04+

EC2Rescue for Linux also requires Python 2.7.9+ or 3.2+ installed.

Installing EC2Rescue for Linux

To use EC2Rescue for Linux, you need to download and install the tool in a working Linux machine which will serve as its host. You can download the EC2Rescue tool using the command below.

```
curl -O https://s3.amazonaws.com/ec2rescuelinux/ec2rl.tgz
```

You also need to download the sha56 hash file to verify the integrity.

```
curl -O https://s3.amazonaws.com/ec2rescuelinux/ec2rl.tgz.sha256
```

Verify the integrity of the tool tar archive.

```
sha256sum -c ec2rl.tgz.sha256
```

Extract the archive.

```
tar -xvf ec2rl.tgz
```

Run EC2Rescue help to view help and to verify the installation.

```
cd ec2rl-<version_number>
./ec2rl help
```

Diagnose Issues Using EC2Rescue for Linux

The EC2Rescue for Linux contains modules that you can use to diagnose and get valuable information about an instance. You can run either all the modules or any specific module you want. Some modules require root access so you may want to use an account with root access or run the command with *sudo*.

Run all modules.

```
./ec2rl run
```

Run a specific module.

```
./ec2rl run --only-modules=module_name --arguments
```

If you are unsure what module to use, you can list and find information about the module using the following command.

List all modules.

```
./ec2rl list

#get help for a module
./ec2rl help module_name
```

EC2Rescue generates a log file that you can view in the */var/tmp/ec2rl* directory after running a module. You can either upload the log file to an S3 bucket or AWS Support.

Upload log file to S3 bucket.

```
./ec2rl upload --upload-directory=/var/tmp/ec2rl/2017-05-11T15_39_21.893145 --presigned-url="s3presignedurl"
```

Upload log file to AWS Support

```
./ec2rl upload --upload-directory=/var/tmp/ec2rl/2017-05-11T15_39_21.893145
--support-url="URLfromAWSsupport"
```

Creating Instance Backup Using EC2Rescue for Linux

With the EC2Rescue tool, you can create a backup for the Linux instance. Backup can be done as AMI or volume snapshot.

Backup using AMI

```
./ec2rl run --backup=ami
```

Backup all volume.

```
./ec2rl run --backup=allvolumes
```

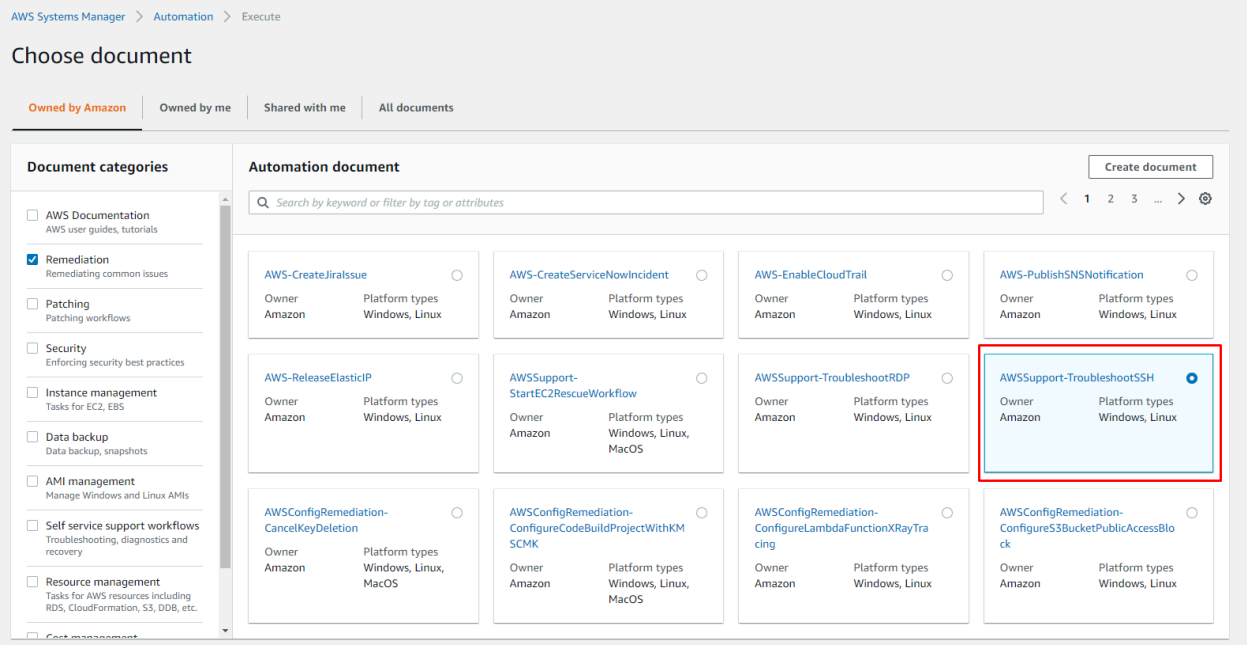
Backup specific volume.

```
./ec2rl run --backup=volumeID
```

EC2Rescue for Linux on Systems Manager

Systems Manager also has automation documents that check Linux instances called *AWSSupport-TroubleshootSSH*. When you execute this automation, it will install EC2Rescue for Linux to check and troubleshoot the instance.

You can find this document under the Remediation category from the Automation in Systems Manager.



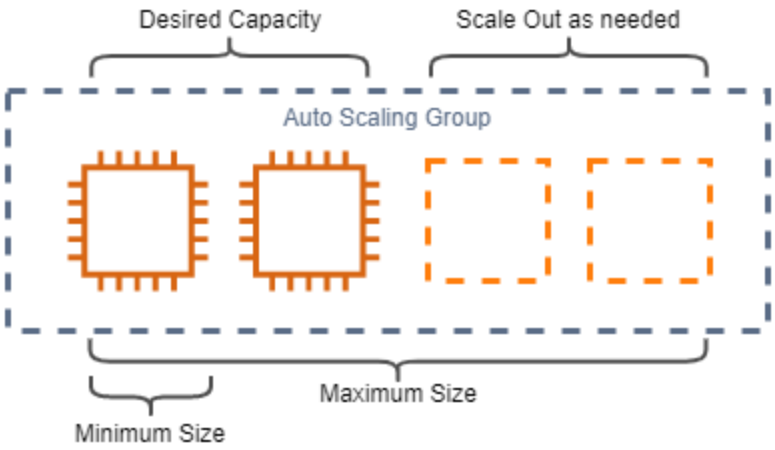
References:

https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/Windows-Server-EC2Rescue.html
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Linux-Server-EC2Rescue.html

AWS Auto Scaling

Auto Scaling Group

Auto Scaling allows AWS resources to scale up and down quickly to handle your application load. On EC2, you can create an Auto Scaling Group to provision a fleet of instances. You can specify its minimum, maximum, and desired number of instances according to your requirements.



EC2 Auto Scaling offers different scaling options that include manual, dynamic, and scheduled scaling. As the word implies, manual scaling allows you to manually add and remove instances from the Auto Scaling Group. Dynamic scaling enables you to set a specific condition to trigger autoscaling. Scheduled scaling scales your Auto Scaling Group at a scheduled time. During scale-out, additional instances spawn while instances are terminated during scale-in.

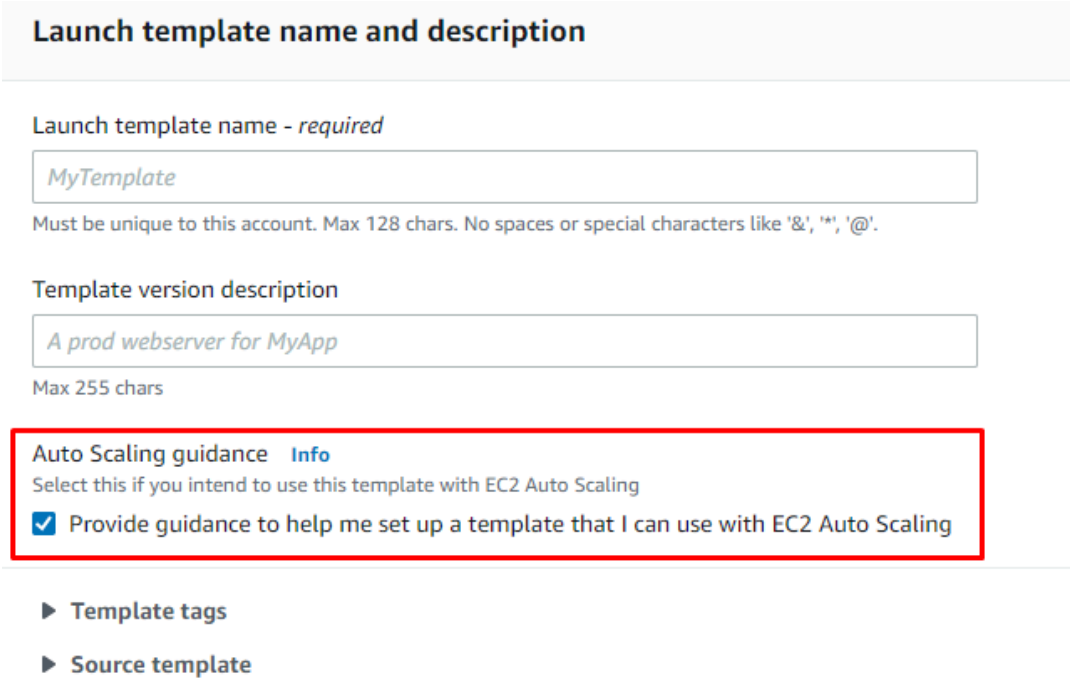
Auto Scaling Templates

To create an Auto Scaling Group, you need to configure Launch Templates. These templates contain the necessary instance parameters for the Auto Scaling Group, like Amazon machine image (AMI) and Instance Type.

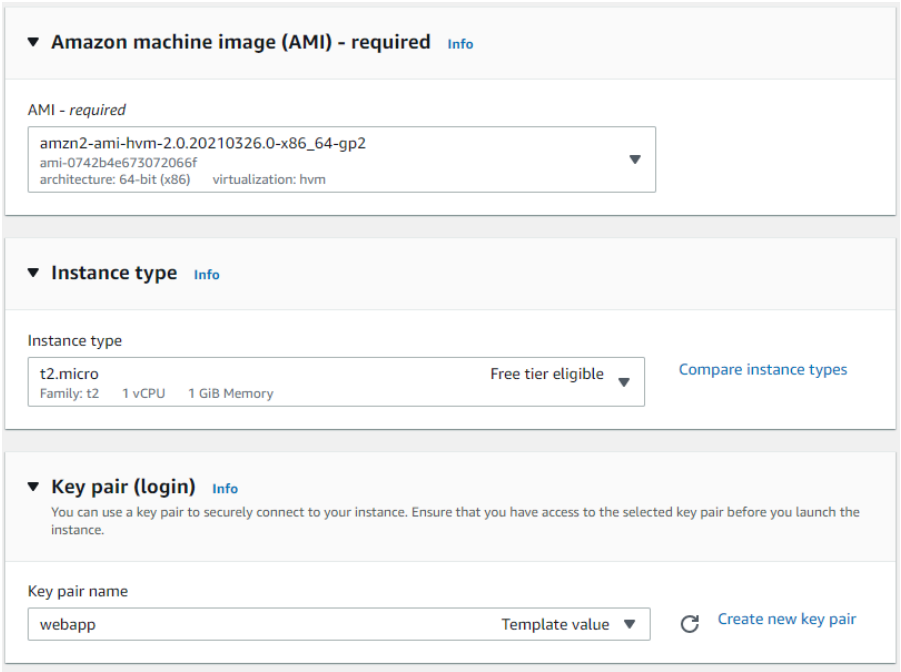
Launch Templates

Launch Templates are reusable templates that contain parameters needed to provision EC2 instances. When creating Launch Templates, defining only the common parameters like template name, AMI, and networking platform is recommended.

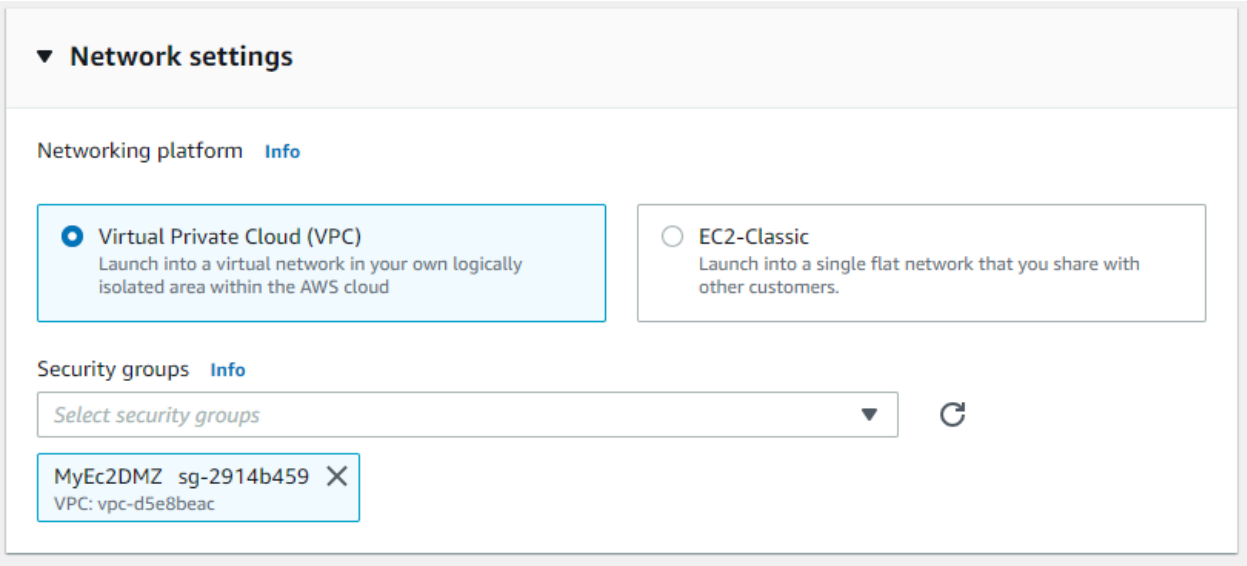
You may want to enable Auto Scaling guidance to guide you on the template creation. If enabled, the Amazon machine image (AMI) parameter is required.



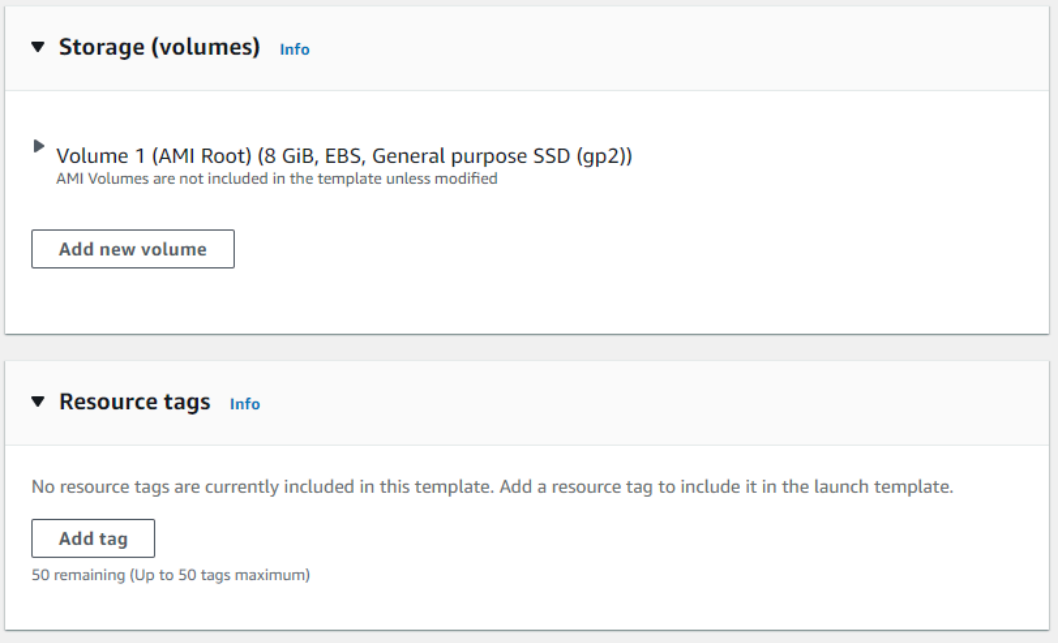
You can already define the Instance Type, but this can be changed or overridden during the Auto Scaling Group configuration. The key pair is the security credentials that will be used when connecting to the instance.



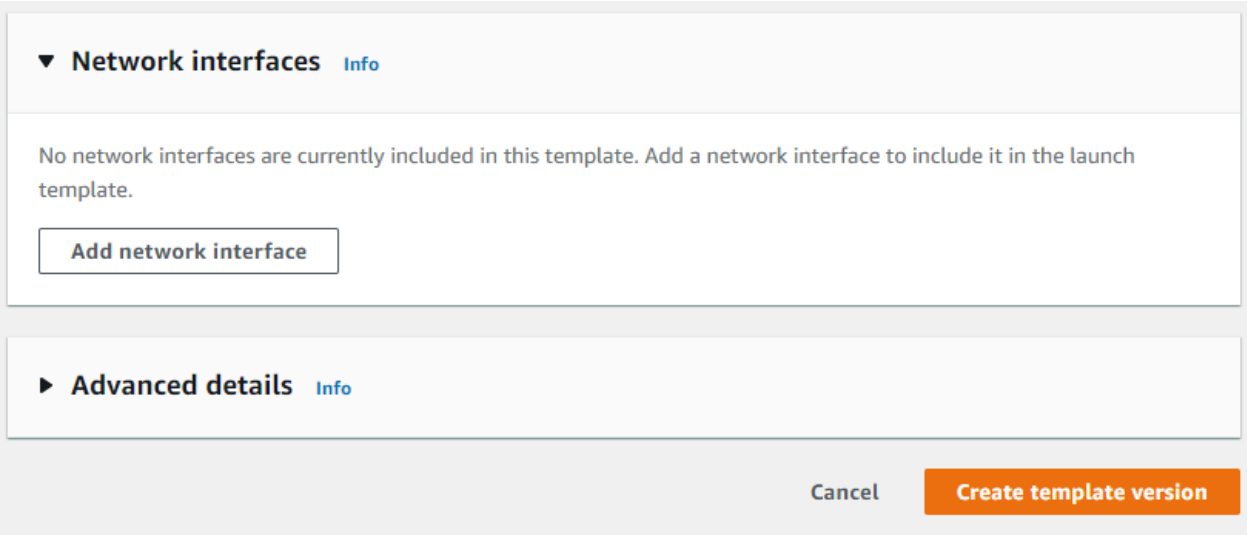
You can select multiple security groups as well. Note that when no security group is assigned, instances will use the default security group. You can only leave the security group empty if you intend to configure this using the network interface.



Volume is automatically created and will depend on the selected AMI; you can optionally create additional volume and add tags.

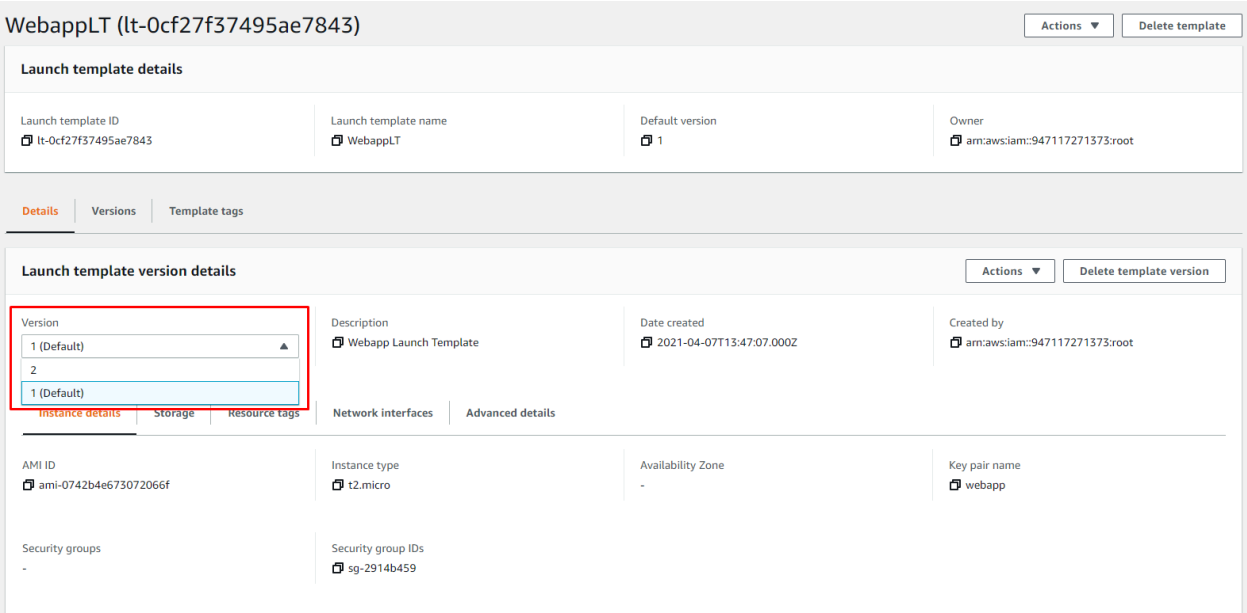


You can skip adding a network interface if you want to keep the default network interface. If you don't want to use auto-assignment of IP addresses to your instances and assign specific IP addresses, you may want to change the default network interface. See further instructions here.



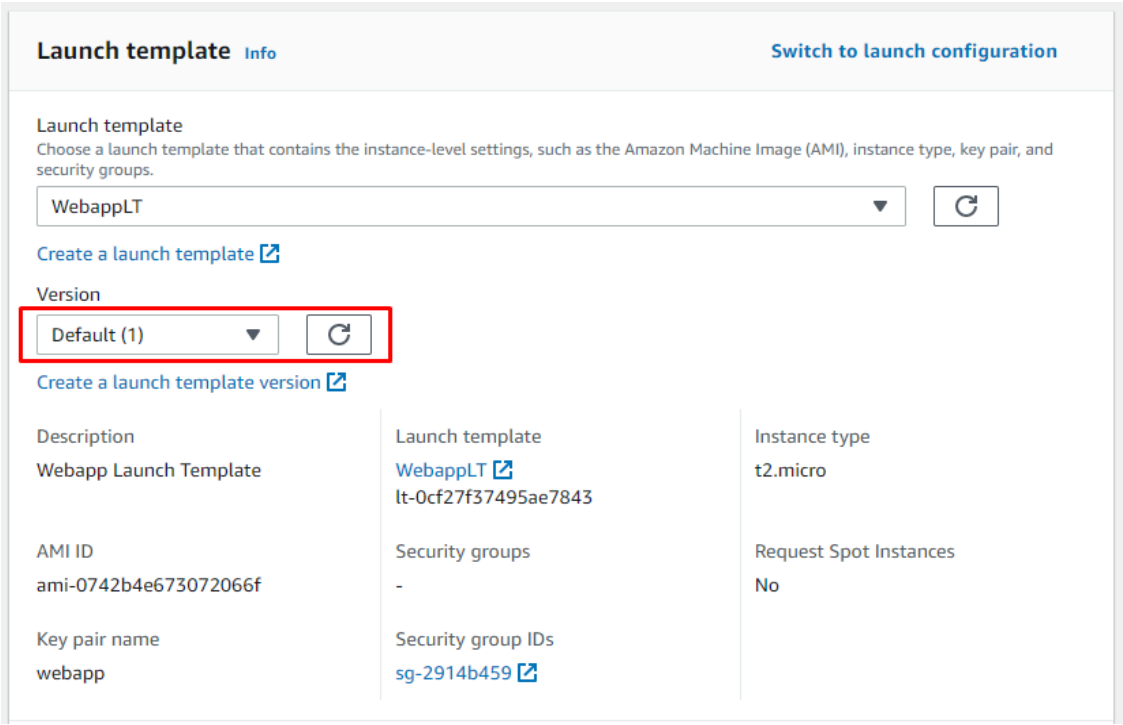
Don't forget to review the parameters before you proceed with the creation of the template.

You can keep multiple template versions when modifying an existing launch template. You can also specify the default version.

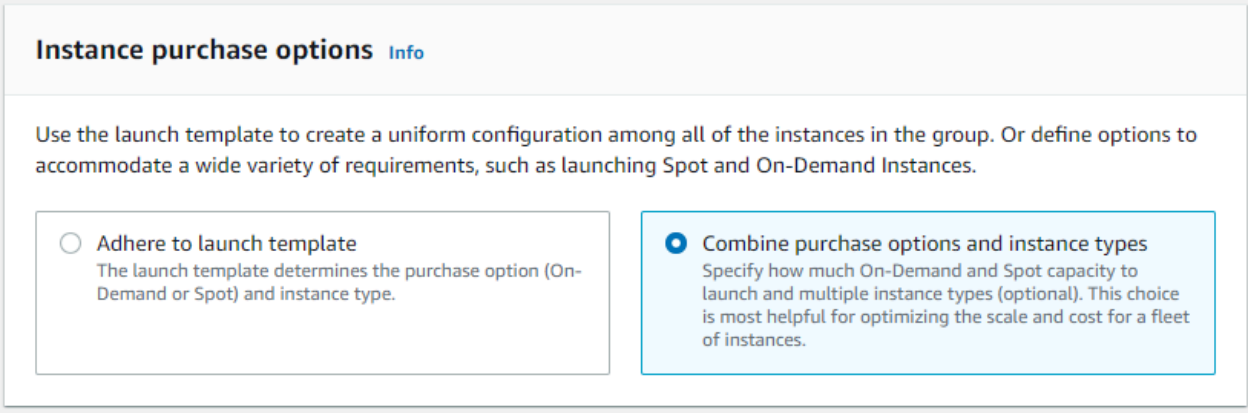


Auto Scaling Group Configuration

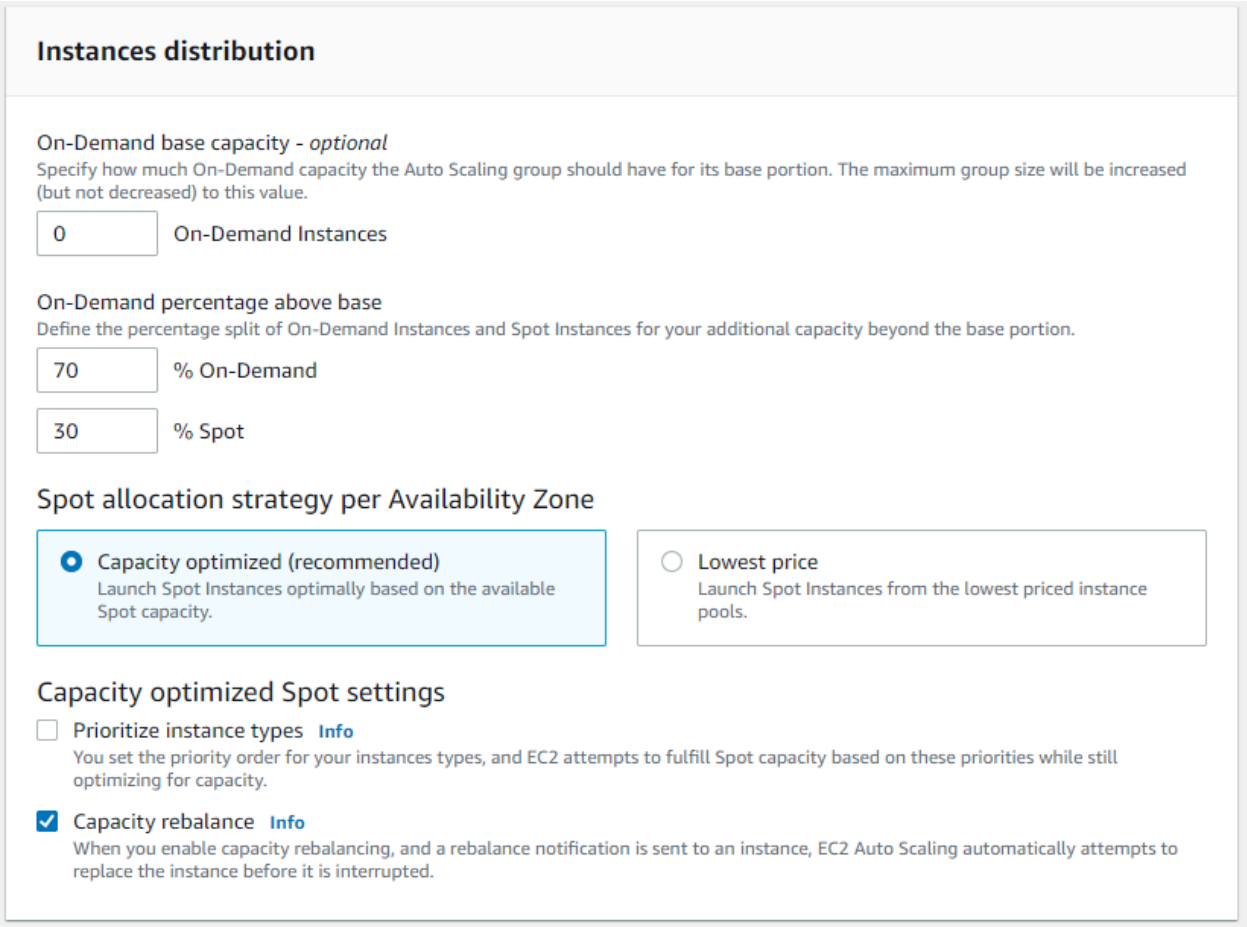
As we mentioned above, you can create an Auto Scaling Group using Launch Templates. Instances in Auto Scaling Group can be On-Demand, Spot instances, or both. You can specify the template version when you use Launch Template.



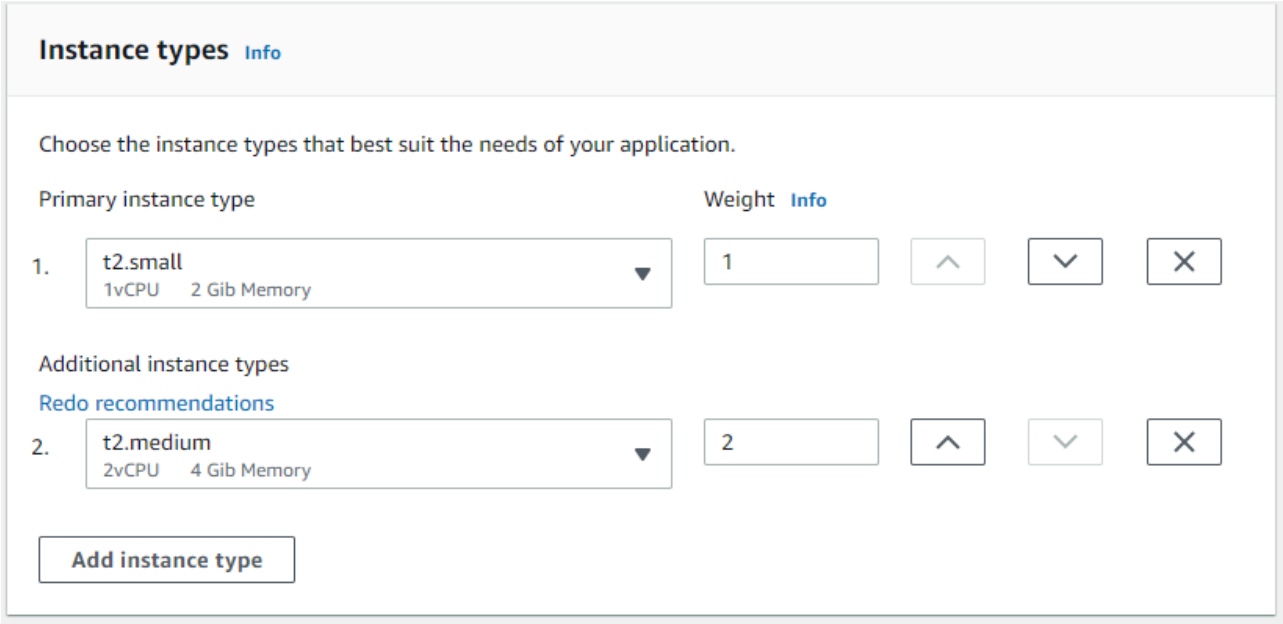
Launch templates allow you to combine purchase options and instance types.



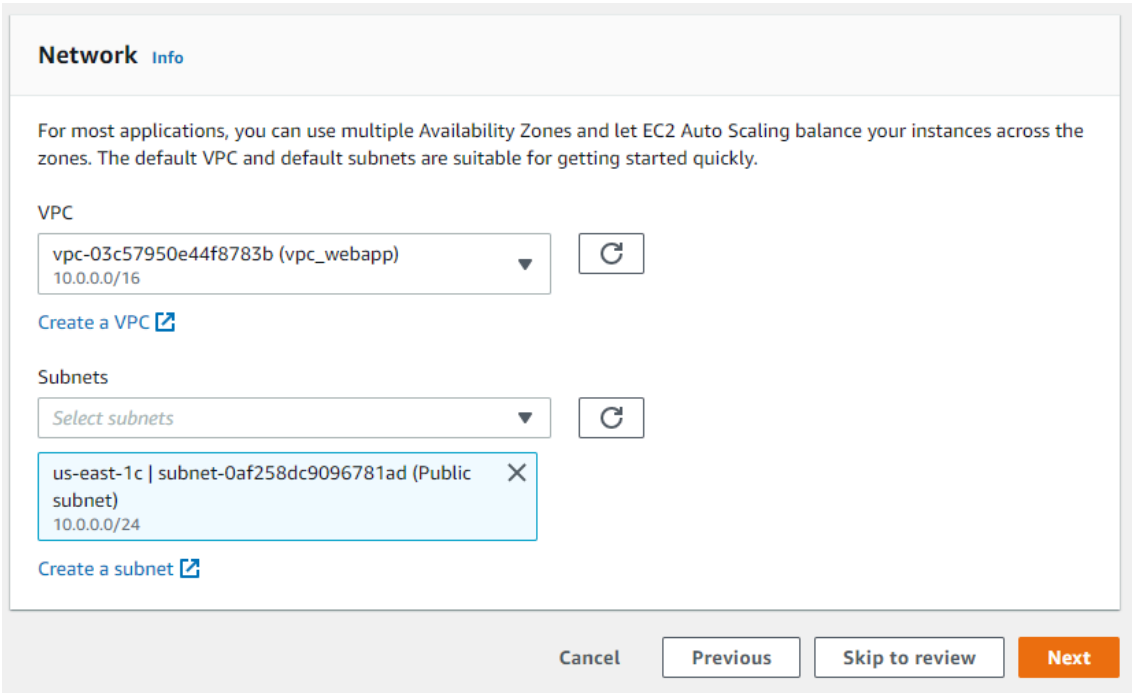
You can maximize the scaling and cost savings by setting the percentage split for the On-Demand and Spot instances. You can also set the Spot allocation strategy and configure Capacity optimized Spot settings.



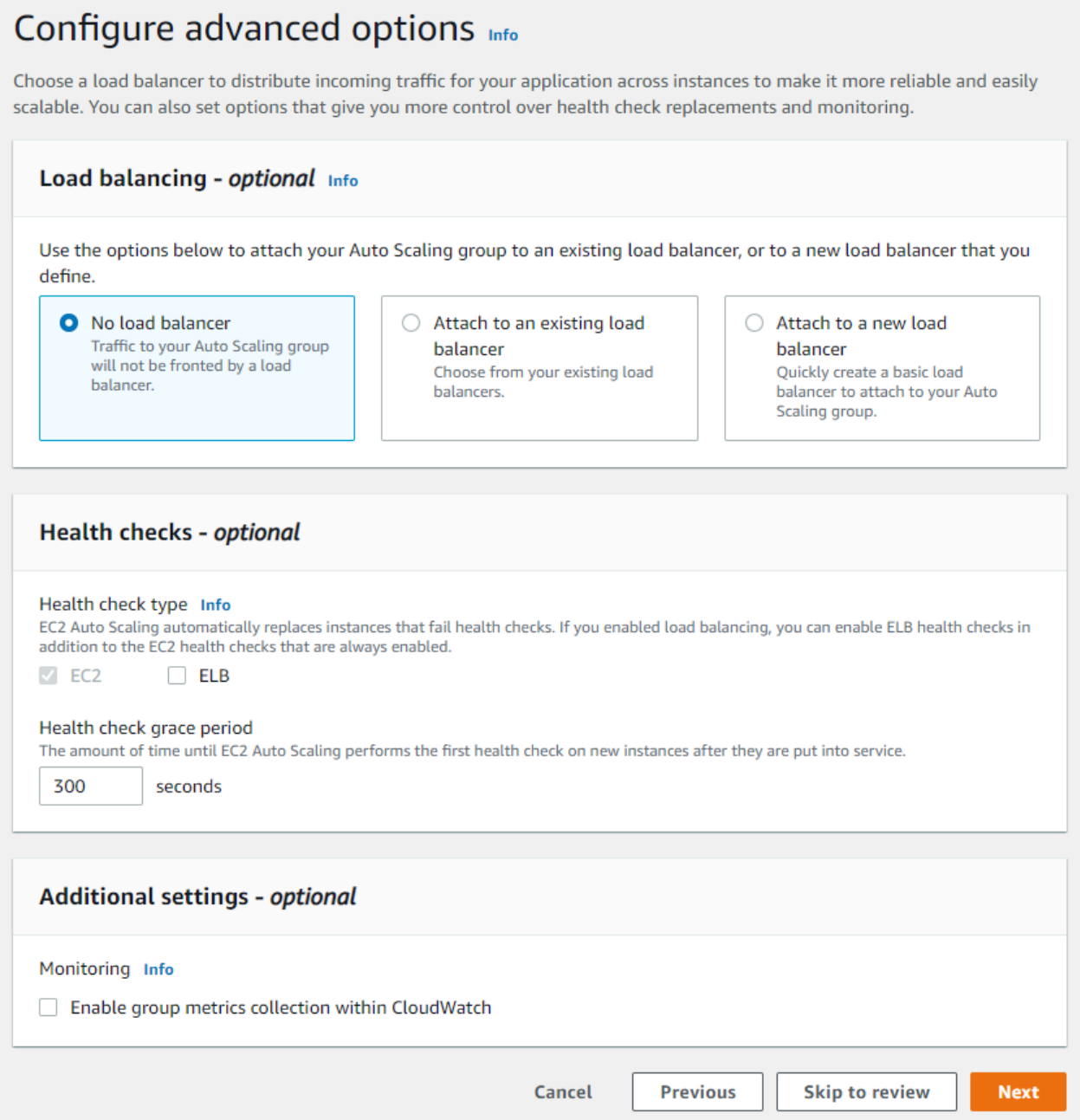
You can still select the instance type for your instances. If you do, the instance type on the launch template is overridden. By default, instances are equally weighted. To use instance weighting for your instances, use units like vCPU or memory to determine the weight. The weight determines the number of units that the instance type represents towards the desired capacity.



Set the VPC and Subnet for the Auto Scaling Group. Make sure that the security group set on the Launch Template resides within the VPC that you selected.



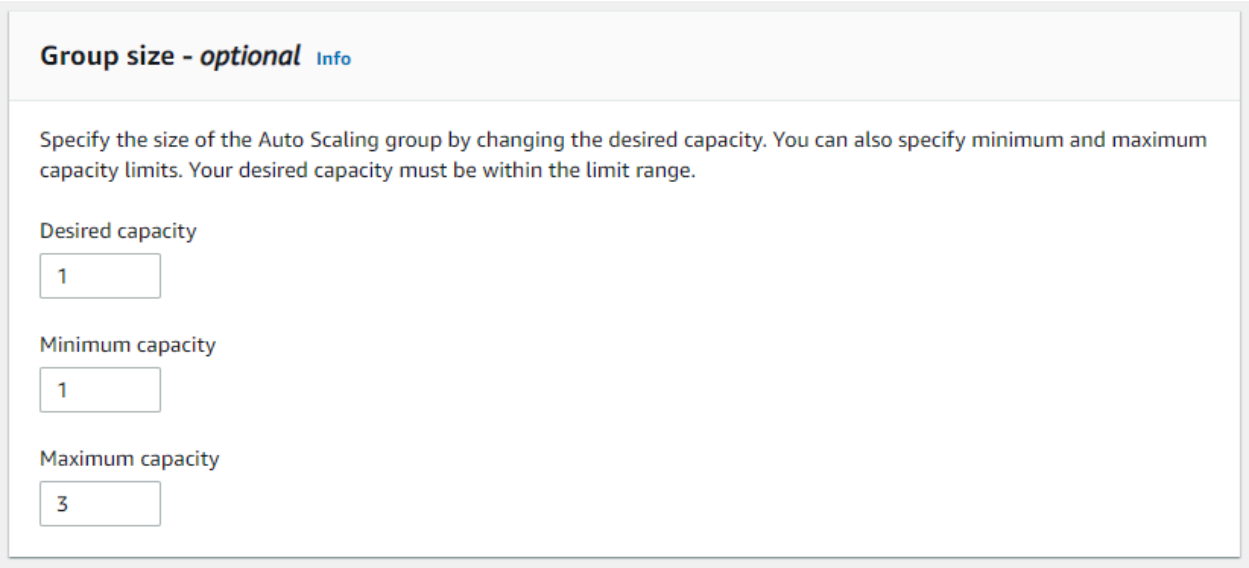
You have an option to add a Load Balancer on top of your Auto Scaling Group. By default, EC2 health checks are enabled, but you can add an ELB health check as well.



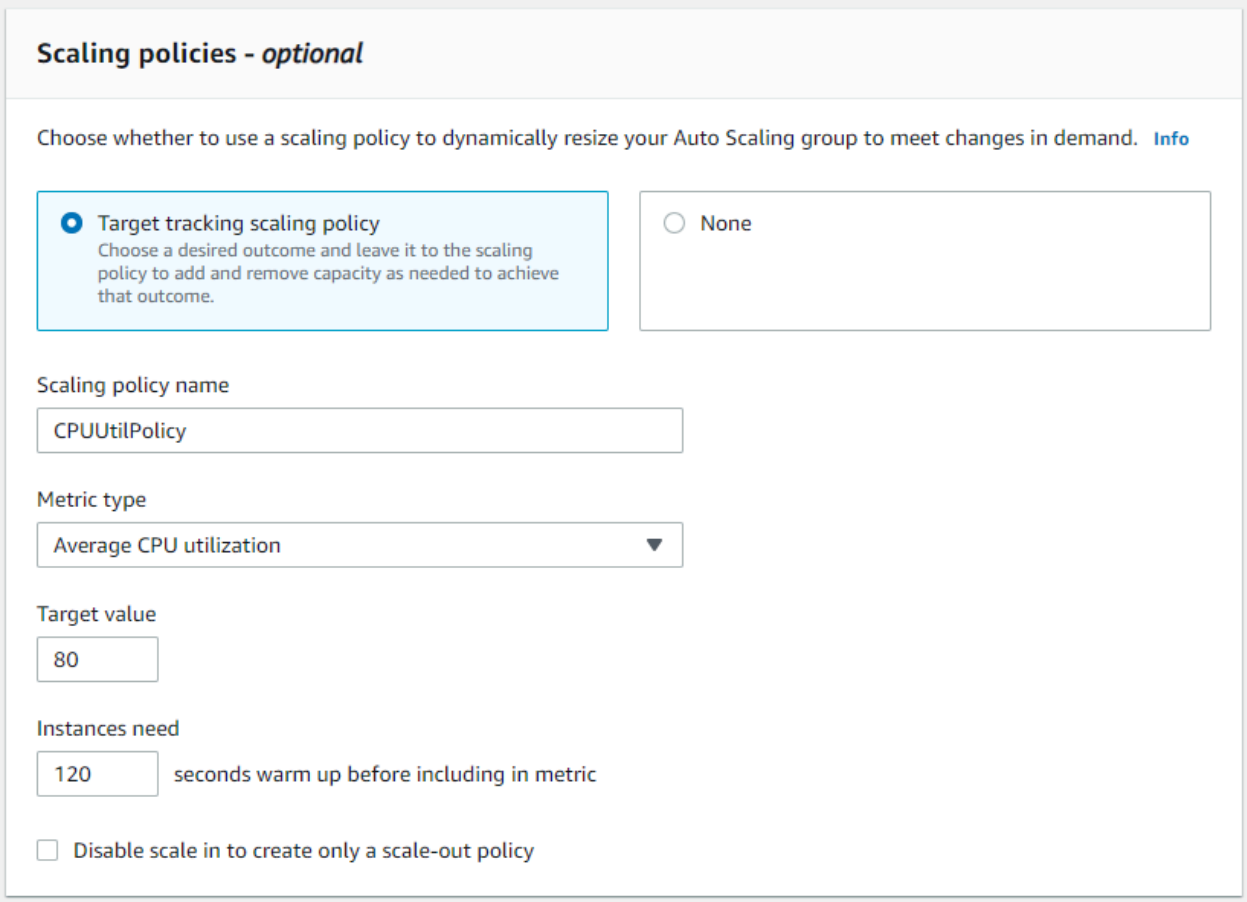
It is recommended to enable group metrics collection within CloudWatch to monitor your Auto Scaling Group. We will further discuss monitoring in the latter part of this eBook.

You can set the desired, minimum, and maximum number of instances for your Auto Scaling Group. When the Auto Scaling Group is created, it will spawn a number of instances equal to its desired capacity. If there are no

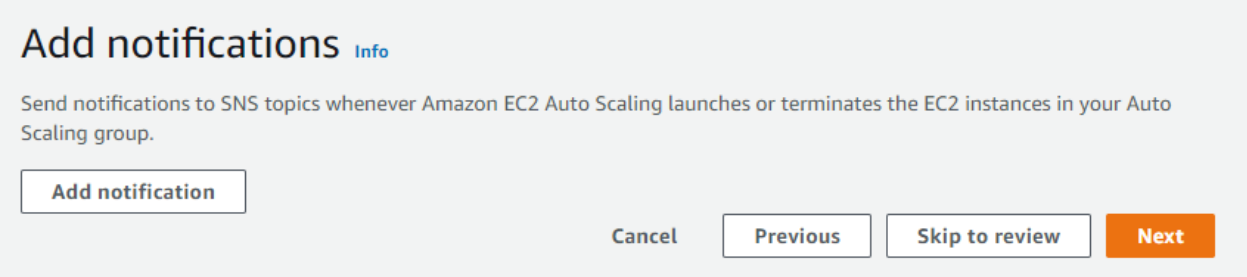
scaling policies defined, the desired capacity will be maintained and will go through periodic health checks. Unhealthy instances will be terminated and will be replaced by new instances.



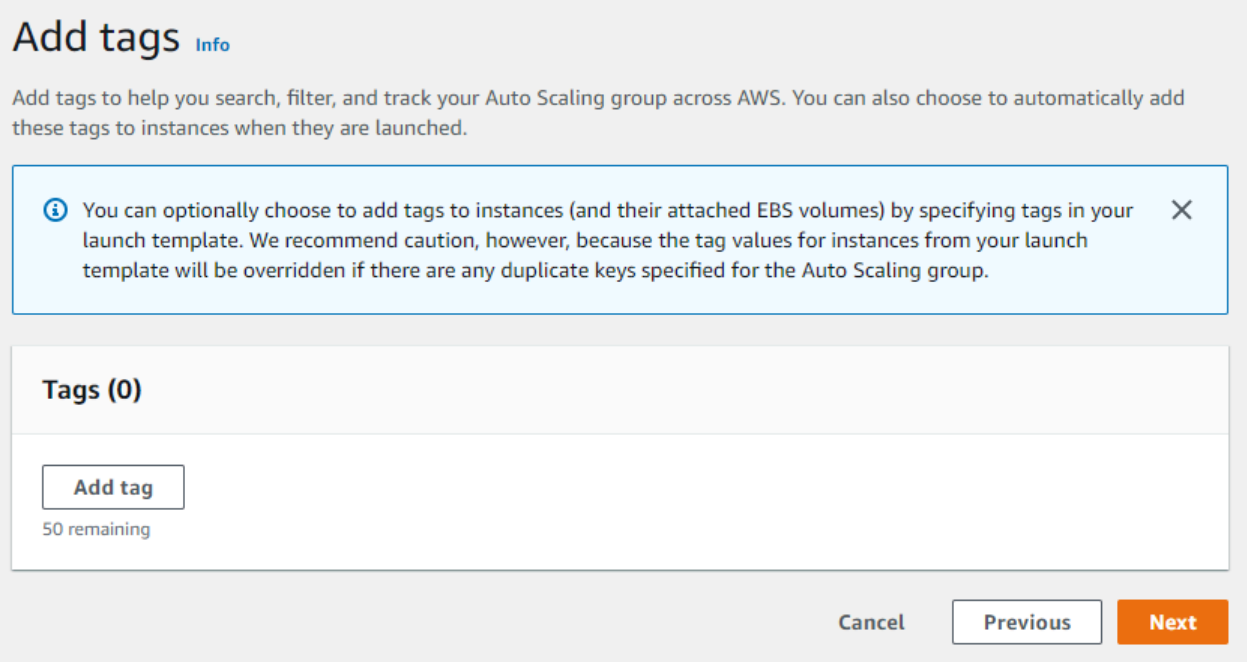
Scaling policies will determine if an Auto Scaling Group will scale out or scale in. You can use different instance metrics to define the scaling policy. For example, if an instance reaches an average CPU utilization of 80%, a new instance will be spawned. And once it drops below 80%, the instance will be terminated. A cooldown will be applied during a scaling process to ensure that the previous scaling has been completely done before starting another scale process. You may want to add a warmup time for your instance to exclude their metric since instances tend to have a high utilization when they are launched.



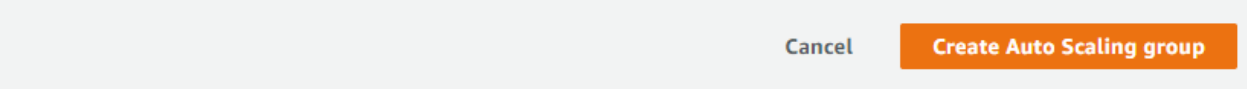
Lastly, you can also set up an SNS integration if you want to get notified of the launch and termination of instances.



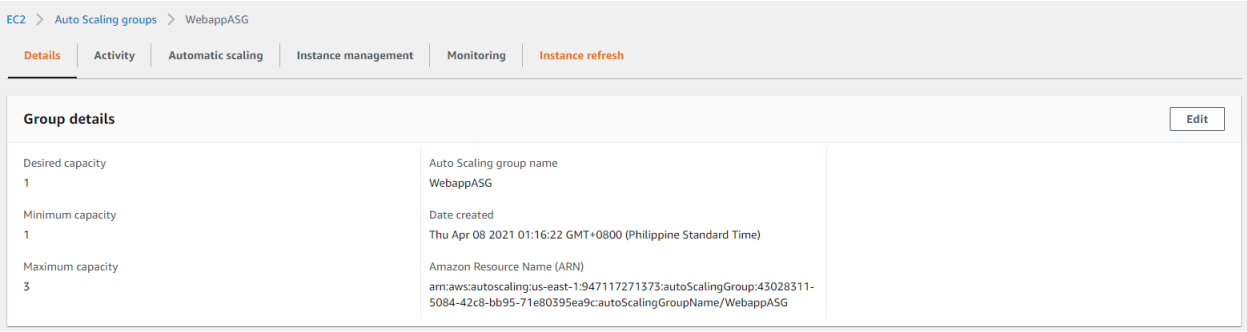
You can optionally add tags as well. Tags will be applied to all of the instances inside your Auto Scaling Group.



After reviewing all of the group configurations, you can now proceed with creating the Auto Scaling Group.



Now you can view vital information like scaling policy, instance management, and monitoring inside your Auto Scaling Group.



If you expect a significant traffic change towards your Auto Scaling Group, you can set a Scheduled Action under the Automatic scaling tab.

Reference:
https://docs.aws.amazon.com/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html

Kubernetes Vertical Pod Autoscaler

The Kubernetes Vertical Pod Autoscaler (VPA) dynamically adjusts the amount of computational resources allocated to Pods in a Kubernetes cluster. Pods are allocated optimal CPU and memory based on actual resource usage and demand, ensuring efficient task performance. The VPA continuously monitors the resource consumption of Pods and compares it with the requested resources. If it detects a discrepancy, it automatically rescales the Pod's CPU and memory requests to match the observed usage patterns. This dynamic adjustment capability enhances the efficiency of resource utilization in the cluster, leading to improved application performance and cost-effectiveness. However, it's important to note that the VPA should be used judiciously as it can lead to frequent Pod restarts due to resizing, which might not be suitable for all applications. Therefore, understanding the nature of your workloads and their resource requirements is crucial when implementing the VPA in your Kubernetes environment.

AWS Global Accelerator

AWS Global Accelerator optimizes and accelerates data transfer for your applications over the AWS global network. It achieves this by directing user traffic to the nearest edge location, which minimizes latency and improves the overall user experience. In addition, it provides consistent performance by intelligently routing user requests to the healthiest endpoints. It is also protected by AWS Shield, which safeguards your applications from DDoS attacks, ensuring their safety and security.

With AWS Global Accelerator, you can be confident that your applications will perform at their best, regardless of your users' location.

AWS Compute Optimizer

AWS Compute Optimizer is a free tool that generates recommendations on cost-savings and performance improvement according to your workloads. It checks the current configuration of your AWS resources, analyzes the historical data from CloudWatch metrics to identify your usage patterns, gives a calculated projection, and provides recommendations. AWS Compute Optimizer leverages the capabilities of AI/ML-based analytics to provide optimal recommendations.

Avoid commonly encountered provisioning issues with AWS Compute Optimizer

- **Over-provisioned Resources** = Unnecessary infrastructure cost
- **Under-provisioned Resources** = Service/Performance degradation

AWS Compute Optimizer supports the following services.

- Amazon Elastic Compute Cloud (Amazon EC2) instances

- Amazon EC2 Auto Scaling groups
- Amazon Elastic Block Store (Amazon EBS) volumes
- AWS Lambda functions

Prerequisites

Only AWS resources with a minimum of **30 consecutive hours of CloudWatch metric** data should receive recommendations from AWS Compute Optimizer, except for Lambda functions which don't need a metric requirement. Once a resource meets the CloudWatch metrics requirement, AWS Compute Optimizer will begin analyzing data which may take up to 12 hours.

Supported AWS Resource	Requirement
Amazon EC2	• Supported Instance types: C1, C3, C4, C5, C5a, C5ad, C5d, C5n, C6a, C6g, C6gn, C7g, D2, D3en, H1, Hpc6a, I2, I3, I3en, Im4gn, Is4gen, M1, M3, M4, M5, M5a, M5ad, M5d, M5dn, M5n, M5zn, M6a, M6g, M6gd, M6i, R3, R4, R5, R5a, R5ad, R5b, R5d, R5dn, R5n, R6g, R6gd, R6i, T1, T2, T3, T3a, T4g, X1, X1e, X2gd, z1d
Auto Scaling Group	• All supported EC2 Instance Types • Single instance type within ASG (no mixed instance types). • The same values for desired, minimum, and maximum capacity. • No scaling policy is attached. • No overrides (changes on Launch Template).
EBS Volume	• EBS Volumes must be attached to an EC2 instance. • Supported EBS Volume Types: gp2, gp3, io1, io2 Block Express
Lambda Function	• Lambda functions with memory ≤ 1,792 MB • Lambda functions with 50 times invocations in the last 14 days

AWS Compute Optimizer Dashboard

AWS Compute Optimizer has a dashboard that helps users visualize the data like Savings opportunities and Under/Over-provisioned percentages.

Dashboard Info

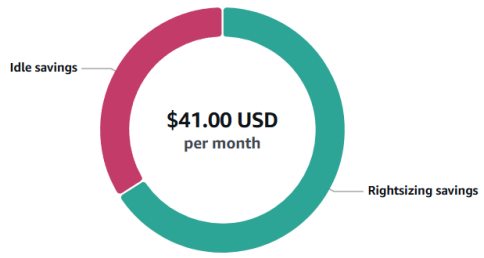
This dashboard provides your savings overview, performance enhancements, and optimization recommendations. These findings are refreshed daily.

View:

Account: [redacted] | Selected regions (1)

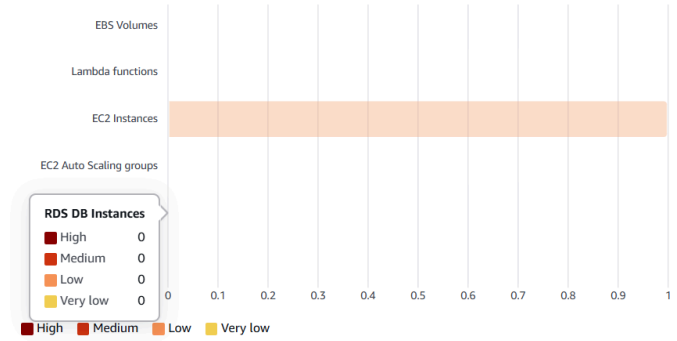
Estimated savings opportunities

Savings type



Percent	Estimated monthly savings	Estimated monthly idle savings
61.36%	$41.00 USD	$13.84 USD

Estimated performance improvement opportunities



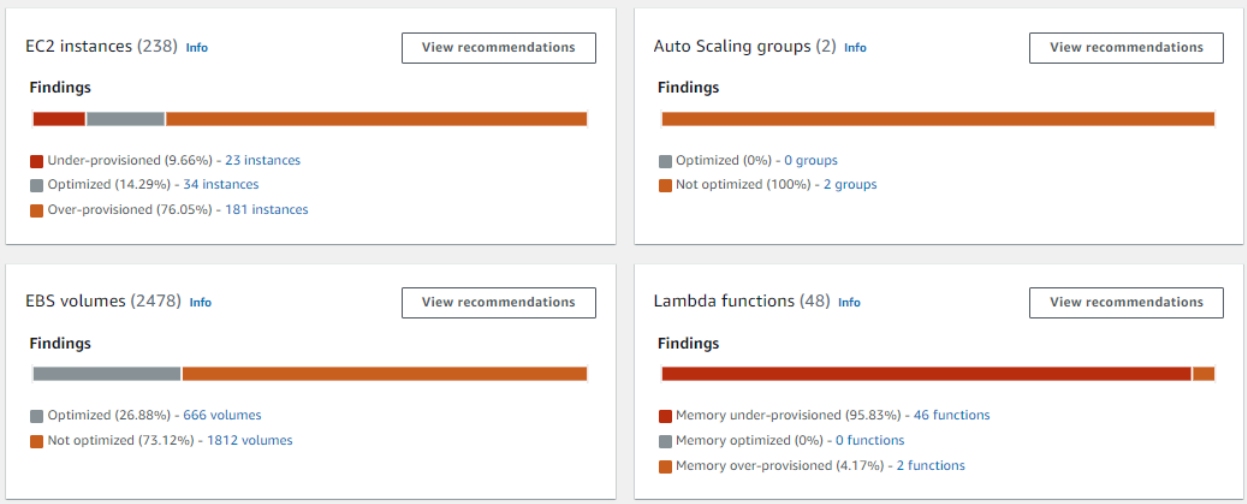
Under-provisioned (percent)	Under-provisioned (count)
25%	1/4

Optimization options for your resources Info

Savings opportunity	Resource type	Optimized	Not optimized	Idle	Idle savings	Rightsizing savings	License savings
$27.16	EC2 Instances	0	2	-	$0.00	$27.16	$0.00
$13.84	EBS Volumes	2	0	7	$13.84	$0.00	-
$0.00	EC2 Auto Scaling groups	0	0	-	$0.00	$0.00	-
$0.00	Lambda functions	0	0	-	-	$0.00	-
$0.00	RDS DB Instances	0	0	-	$0.00	$0.00	-
$0.00	RDS DB Storage	0	0	-	-	$0.00	-
$0.00	ECS services on Fargate	0	0	-	$0.00	$0.00	-

Note: Savings info provides total max savings by resource type. Max savings opportunity may not equal to the sum of idle, rightsize and license savings.

Sample findings Summary per AWS Resources



Amazon Compute Optimizer will provide vital information to help you strategize accordingly from a performance and cost standpoint. This information includes Estimated monthly savings, Savings opportunity (%), Recommended instance type, Price difference (%), and many others.

Recommendations for EC2 Instances

AWS Compute Optimizer analyzes the metrics collected on the CloudWatch Metrics. For Amazon EC2 instances, these metrics are the following:

- CPUutilization
- Memory utilization
- NetworkIn/NetworkOut
- NetworkPacketsIn/NetworkPacketsOut
- DiskReadOps/DiskWriteOps
- DiskReadBytes/DiskWriteBytes
- VolumeReadBytes/VolumeWriteBytes
- VolumeReadOps/VolumeWriteOps

An Amazon EC2 Instance can be classified as:

- Over-provisioned - underutilized infrastructure, unnecessary costs
- Under-provisioned - over utilized infrastructure, performance degradation
- Optimized - instance is correctly configured, and optimized base on the workloads

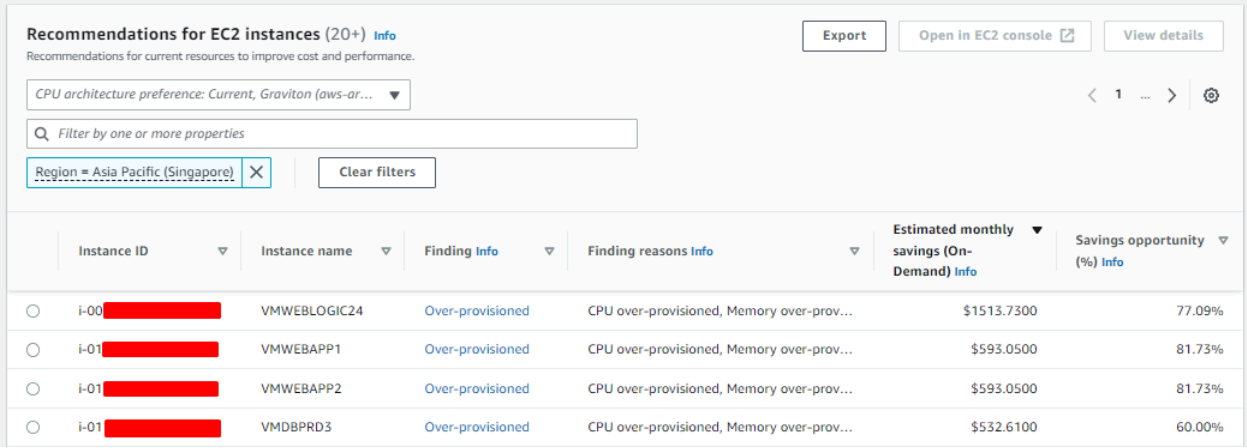
The finding reasons determine the overall finding classification for the current instance. Below are some examples of finding reasons.

- CPU over-provisioned
- Memory over-provisioned
- EBS IOPS over-provisioned
- EBS throughput over-provisioned
- Network bandwidth over-provisioned

AWS Compute Optimizer may give the following recommendations for EC2 instances.

- Recommended instance type
- Recommended On-Demand price
- Recommended RI coverage (%)
- Recommended RI utilization (%)

Sample EC2 recommendations from taken AWS Console



Recommendations for Auto Scaling Group

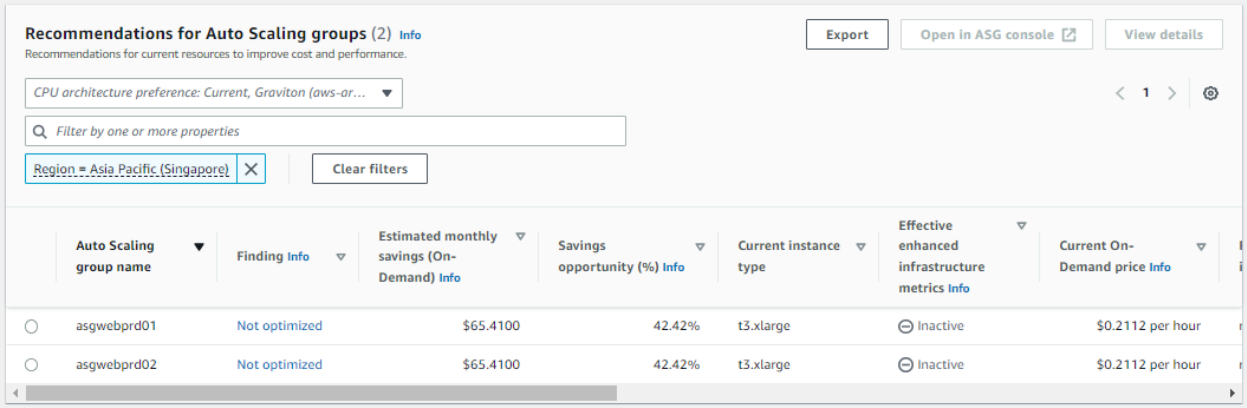
An Auto Scaling group can be classified as Not optimized, or Optimized.

- Optimized - correctly configured and optimized based on the workloads
- Not Optimized - at least one specification/metric is Over-provisioned, or Under-provisioned

AWS Compute Optimizer may give the following recommendations for Auto Scaling Group.

- Recommended instance type
- Recommended On-Demand price
- Recommended RI coverage (%)
- Recommended RI utilization (%)

Sample Auto Scaling Group recommendations taken from AWS Console



Recommendations for EBS Volume Instances

The following EBS Volume metrics are collected.

- VolumeReadBytes
- VolumeWriteBytes
- VolumeReadOps
- VolumeWriteOps

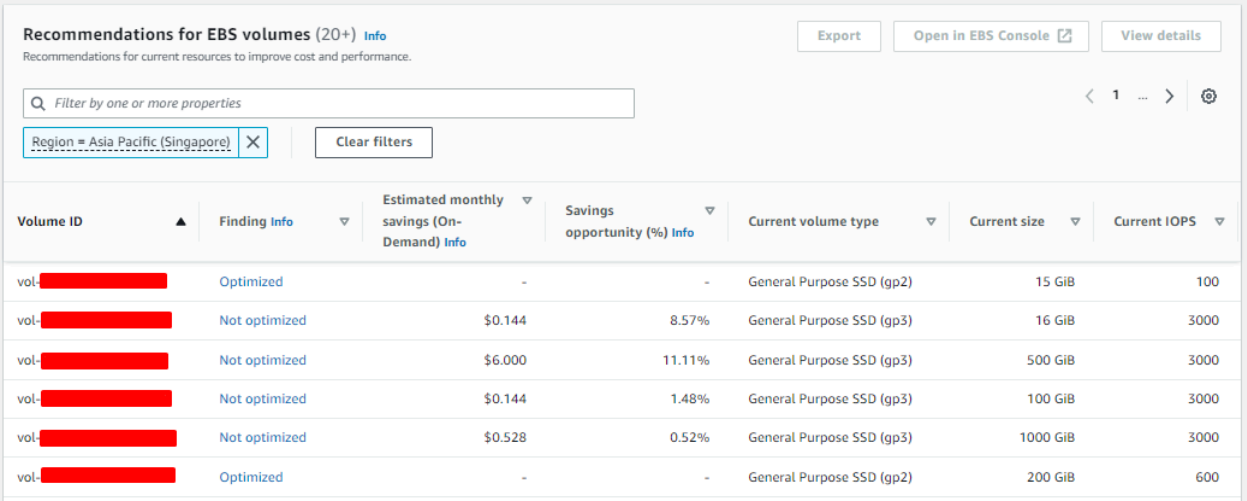
An EBS Volume can be classified as Not optimized, or Optimized.

- Optimized - correctly configured and optimized based on the workloads
- Not Optimized - at least one specification/metric is Over-provisioned, or Under-provisioned

AWS Compute Optimizer may give the following recommendations for EBS Volumes.

- Recommended volume type
- Recommended size
- Recommended IOPS
- Recommended throughput
- Recommended monthly price

Sample EBS Volume recommendations taken from AWS Console



Recommendations for EBS volumes (20+) Info
Recommendations for current resources to improve cost and performance.

Export | Open in EBS Console | View details

Filter by one or more properties

Region = Asia Pacific (Singapore) | Clear filters

Volume ID	Finding Info	Estimated monthly savings (On-Demand) Info	Savings opportunity (%) Info	Current volume type	Current size	Current IOPS
vol-[redacted]	Optimized	-	-	General Purpose SSD (gp2)	15 GiB	100
vol-[redacted]	Not optimized	$0.144	8.57%	General Purpose SSD (gp3)	16 GiB	3000
vol-[redacted]	Not optimized	$6.000	11.11%	General Purpose SSD (gp3)	500 GiB	3000
vol-[redacted]	Not optimized	$0.144	1.48%	General Purpose SSD (gp3)	100 GiB	3000
vol-[redacted]	Not optimized	$0.528	0.52%	General Purpose SSD (gp3)	1000 GiB	3000
vol-[redacted]	Optimized	-	-	General Purpose SSD (gp2)	200 GiB	600

Lambda function metrics being used:

- Invocations
- Duration
- Errors
- Throttles

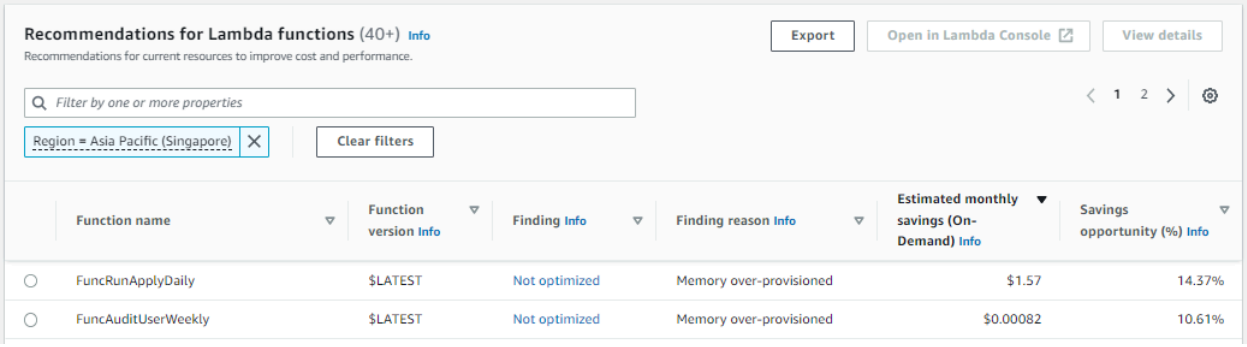
Lambda function can be classified as Not optimized, or Optimized.

- Optimized - correctly configured and optimized based on the workloads
- Not Optimized - at least one specification/metric is Over-provisioned, or Under-provisioned

AWS Compute Optimizer may give the following recommendations for Lambda Volumes.

- Recommended configured memory
- Recommended cost (low)
- Recommended cost (high)

Sample Lambda recommendations taken from AWS Console



Recommendations for Lambda functions (40+) Info
Recommendations for current resources to improve cost and performance.

Export | Open in Lambda Console | View details

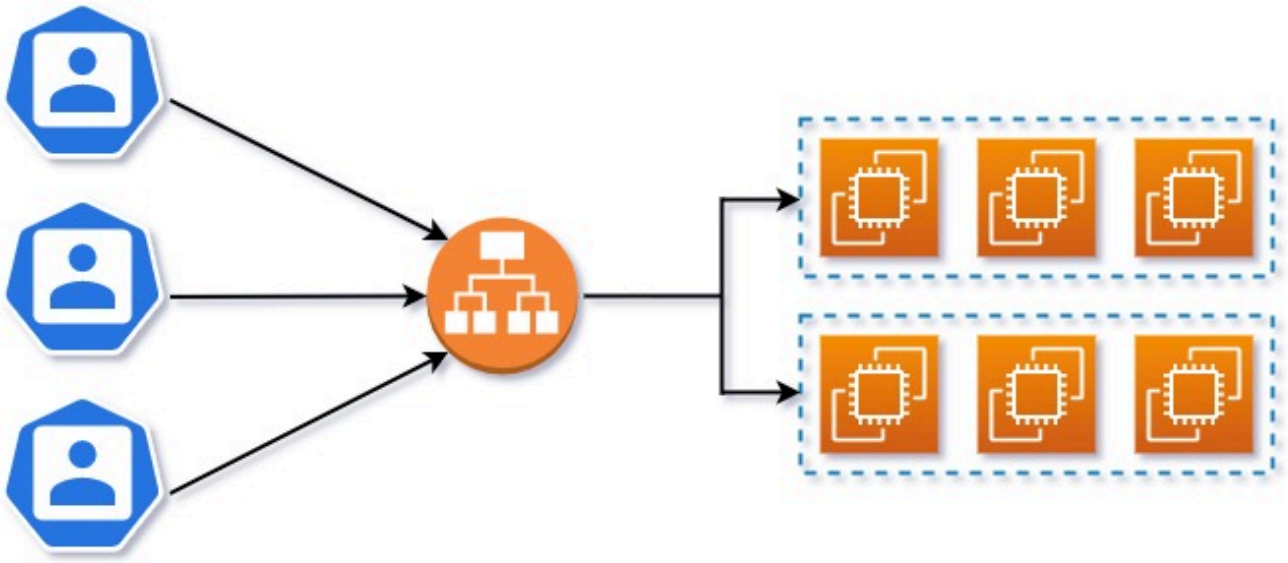
Filter by one or more properties

Region = Asia Pacific (Singapore) | Clear filters

Function name	Function version Info	Finding Info	Finding reason Info	Estimated monthly savings (On-Demand) Info	Savings opportunity (%) Info
FuncRunApplyDaily	$LATEST	Not optimized	Memory over-provisioned	$1.57	14.37%
FuncAuditUserWeekly	$LATEST	Not optimized	Memory over-provisioned	$0.00082	10.61%

Elastic Load Balancing

Elastic Load Balancing (ELB) is a traffic distribution service that offers high availability and elasticity for your application. ELB is a fully managed service. Its load balancer scales as traffic changes allowing it to handle millions of requests per second. A load balancer works by accepting traffic and routing it across multiple targets in one or more Availability Zones (AZ). A load balancer node is created on each enabled Availability Zone. The load balancer automatically handles the distribution of the traffic.



Load Balancer Types

ELB offers different types of load balancers. The type of load balancer depends on the traffic that you want to distribute. Another thing to check when setting up an ELB is whether the load balancer would be internet-facing or internal.

ELB supports the following types of load balancers for your workloads:

- Application Load Balancer - Layer 7 load balancing/HTTP/HTTPS traffic
- Network Load Balancer - Layer 4 load balancing/TCP/UDP traffic
- Gateway Load Balancer - Layer 3 Gateway + Layer 4 Load Balancing

ELB Features and Components

Elastic Load Balancing is placed on top of your application for traffic distribution. A load balancer is designed to increase the availability and security of your application. Aside from the load balancer type, you need to define other load balancer components and features to maximize its capabilities.

Load Balancer Scheme

You can set a load balancer to either internal or internet-facing. An internet-facing load balancer has a Public IP address assigned on its node, and the requests are routed over the internet from the client to a target. On the other hand, the nodes of an internal load balancer have a Private IP address, and requests are routed privately from client to target.

IP Addresses Type

For Application and Network Load Balancer, you can specify the IP Address type that it will support. You can choose between IPv4 or dualstack (both IPv4 and IPv6).

Listener

The listener is a process that identifies the request through ports and protocols. By defining the listener rule, a load balancer will know the traffic it will accept. Here's an Application Load Balancer listener as an example.

Target Group

The requests are routed to one or more targets through the port and protocol defined on the listener through the target groups. ELB supports various types of targets such as:

- EC2 Instances
- Containers
- Lambda Functions
- IP Addresses
- Virtual Appliances

Security Groups

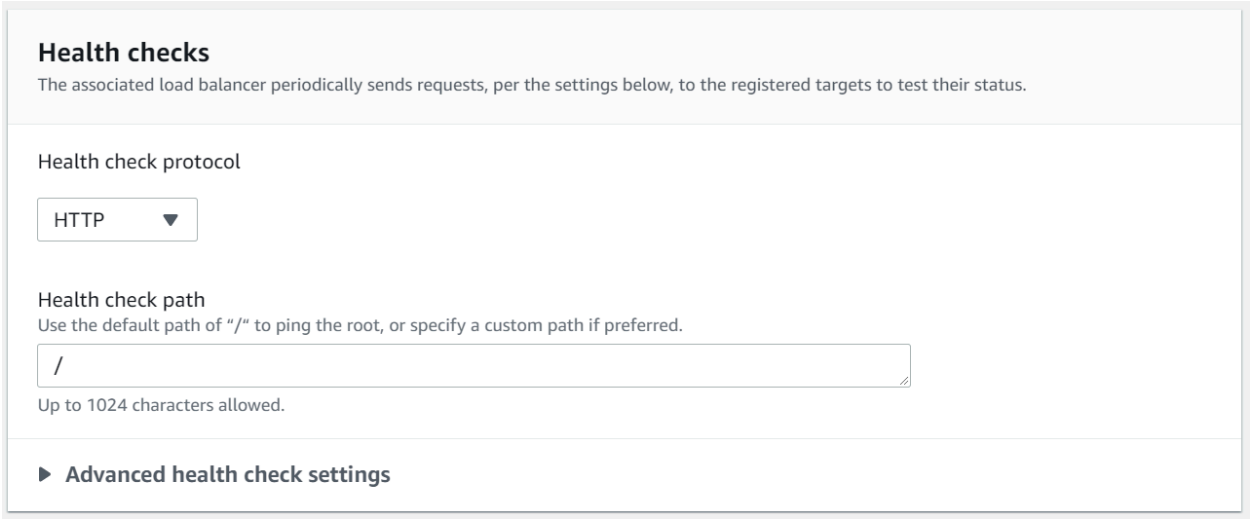
A security group is an additional layer of security for load balancers. Only the traffic allowed by the security group will reach the load balancer.

Availability Zones

Availability Zones are locations within an AWS Region. You should always enable the AZ where your targets are located. Take note also that you can only specify one subnet per Availability Zone.

Health Checks

Health checks are configured inside a target group. It ensures that the requests are routed to healthy targets by monitoring your registered targets' HTTP and HTTPS endpoints. Health checks are also configurable; you can change the value of the threshold, interval, timeout, and success code.



Sticky Sessions

The sticky session is an ELB feature that guarantees that traffic from the same client is routed only to the same target.

Cross-zone Load Balancing

Cross-zone Load Balancing is an ELB feature that enables the load balancer node to distribute all traffic across all Availability Zones with healthy targets. When disabled, the load balancer node distributes the traffic only to its AZ.

Connection Draining

Connection draining or deregistration delay is a feature that allows a request to be completed before unregistering a target.

Load Balancer Monitoring

CloudWatch integrates with Elastic Load Balancing to monitor and log its metrics. API calls to ELB are also logged on CloudTrail.

Delete Protection

Deletion protection prevents the accidental deletion of a load balancer.

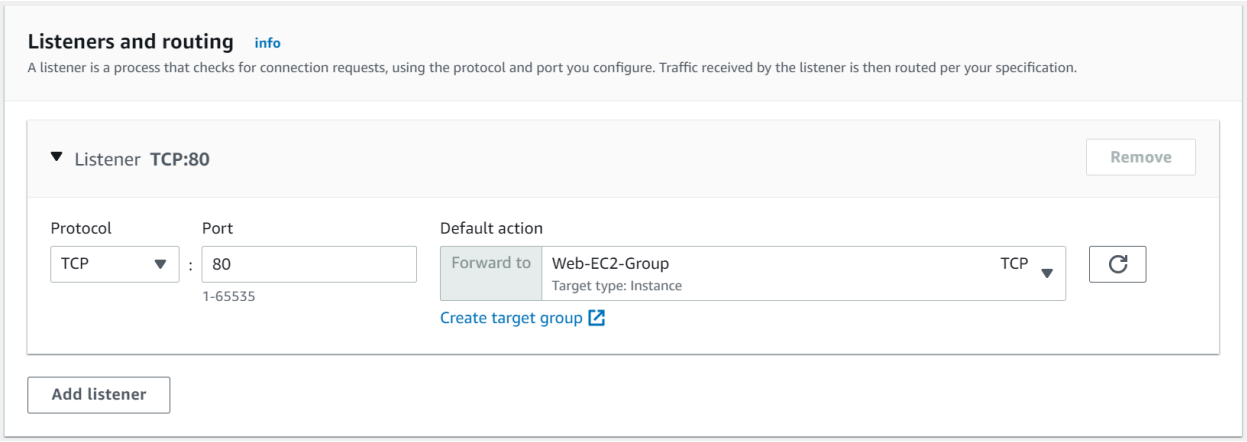
Choosing the Right Load Balancer

Application Load Balancer (ALB)

For routing HTTP and HTTPS requests, an Application Load Balancer is a good choice. Since this load balancer works on Layer 7, it can distribute your web application's HTTP and HTTPS requests. An ALB can work as an internet-facing and internal load balancer, and work for IPv4 and dualstack IP addresses. It requires a minimum of two different subnets on different Availability Zones. Note that a server certificate is required if you have a listener configured for HTTPS protocol. You can either use a certificate from AWS Certificate Manager or IAM.

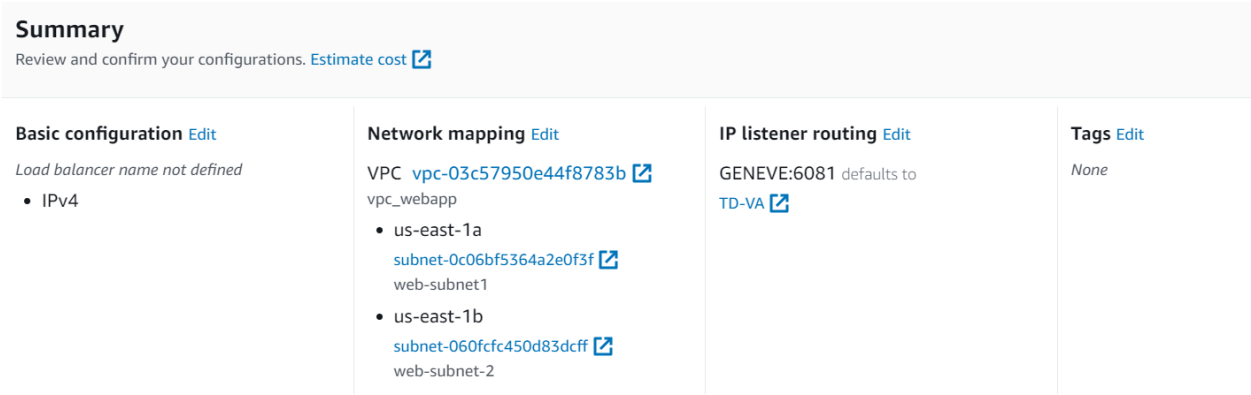
Network Load Balancer (NLB)

A Network Load Balancer is ideal for distributing TCP and UDP traffic from the client to your application. Like an Application Load Balancer, the Network Load Balancer can also be either internal or internet-facing. It also supports both IPv4 and dualstack IP addresses. You can set up a Network Load Balancer using only one subnet, but two subnets are ideal for improving your application's fault tolerance. Make sure as well that the target group has at least one valid target to accept the requests. If multiple AZs are specified, you can enable Cross-Zone Load Balancing to distribute the traffic in all targets equally. The direction of the network traffic will depend on the rules defined by the listener and routing.



Gateway Load Balancer (GWLB)

A Gateway Load Balancer eases the management of multiple virtual appliances. A GWLB can distribute traffic across multiple virtual appliances that support Generic Network Virtualization Encapsulation (GENEVE). You can enable GWLB on a single subnet, but the target group should have at least one target that accepts the GENEVE protocol. Take note that Gateway Load Balancer only supports IPv4 addresses; it doesn't support IPv6 and dualstack.



References:
https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/what-is-load-balancing.html
https://docs.aws.amazon.com/elasticloadbalancing/latest/userguide/how-elastic-load-balancing-works.html

S3 Presigned URL

All buckets and objects are stored privately by default in S3. This is a much more secure design than having your objects publicly accessible by default. But if you have a storage requirement where you must share an object or bucket with your customers to upload an object quickly, you can use the S3 presigned URL to generate a URL for your customer. The URL generated can be used by a customer to view or upload an object. The action and the duration of the access can also be defined when generating a URL.

Sharing S3 objects using Presigned URL

There are multiple ways of creating a presigned URL to share an object, such as:

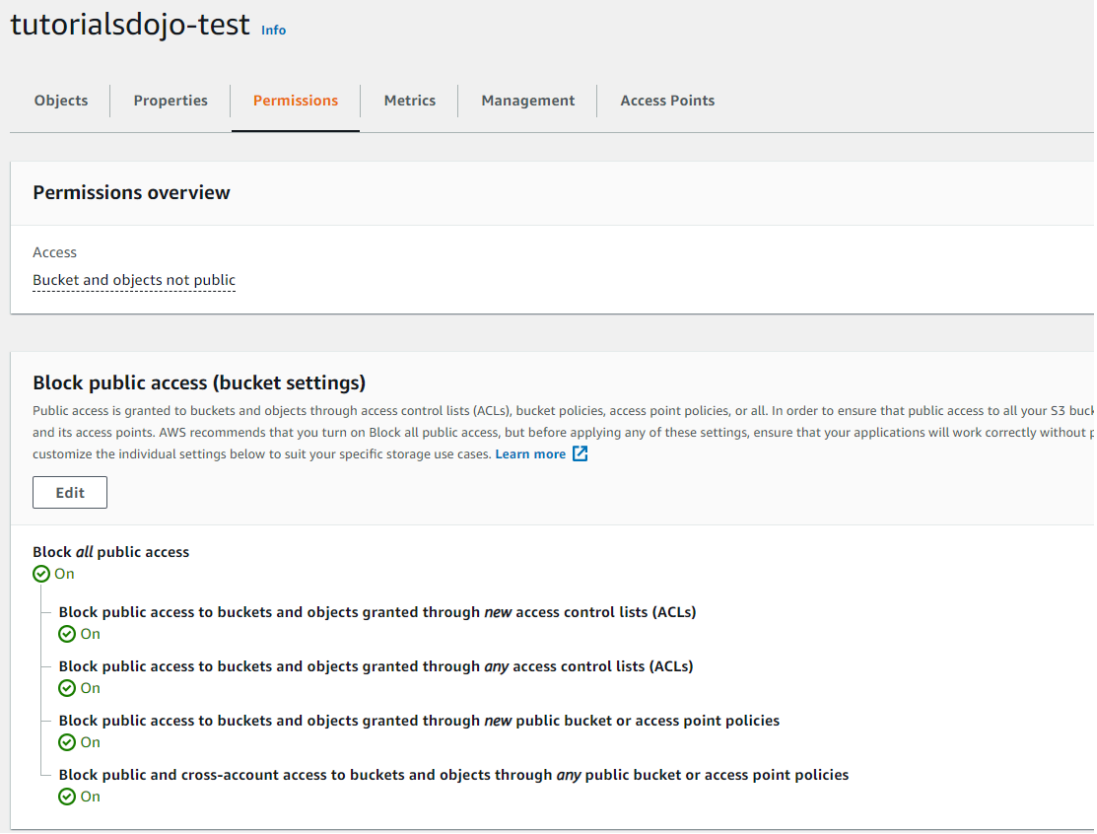
- AWS CLI
- AWS SDK for Java
- AWS Explorer for Visual Studio
- REST API
- .NET
- Ruby

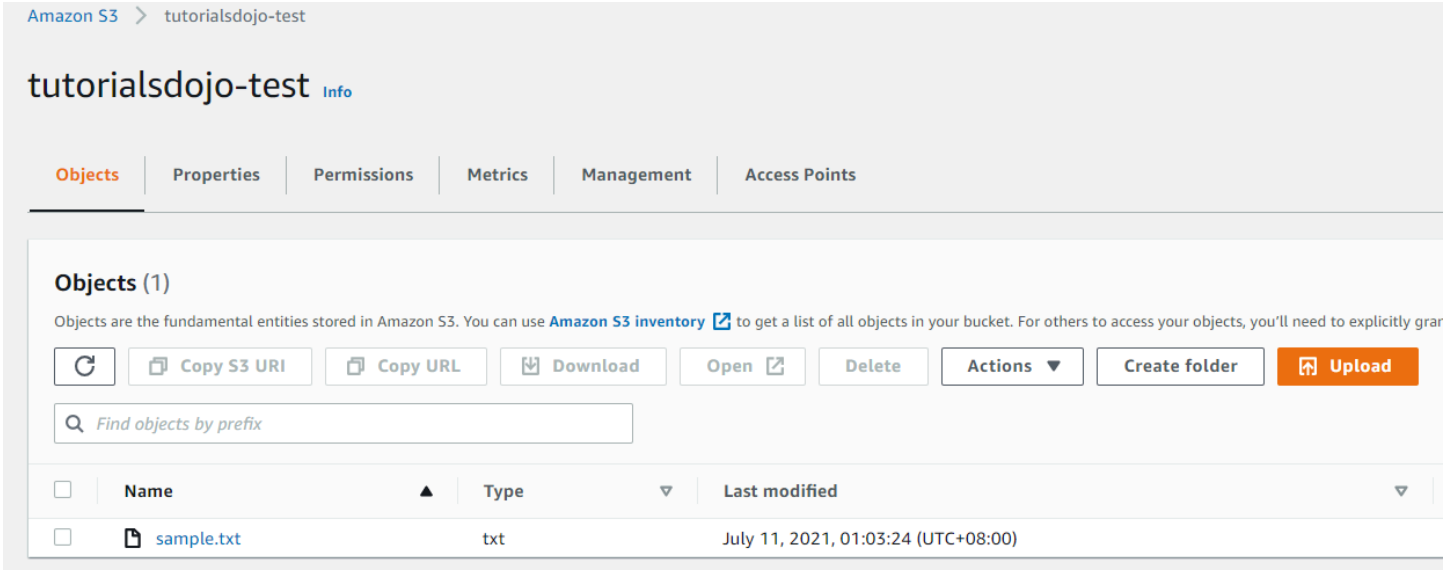
- PHP
- Node.js
- Python
- Go

You can use the following security credentials when generating a presigned URL. Note that the user who generates a presigned URL should have the necessary permission to execute the action of the generated URL.

Security Credentials	Validity
IAM instance profile	6 hours max validity
AWS Security Token Service	36 hours max validity for permanent credentials
IAM user (access key and secret access key)	Seven days max validity when using AWS Signature Version 4

In this example, we use AWS Command Line Interface to generate a URL and share an object in S3.





To generate a presigned URL, follow the command below.

```
aws s3 presign <s3objectURL>
```

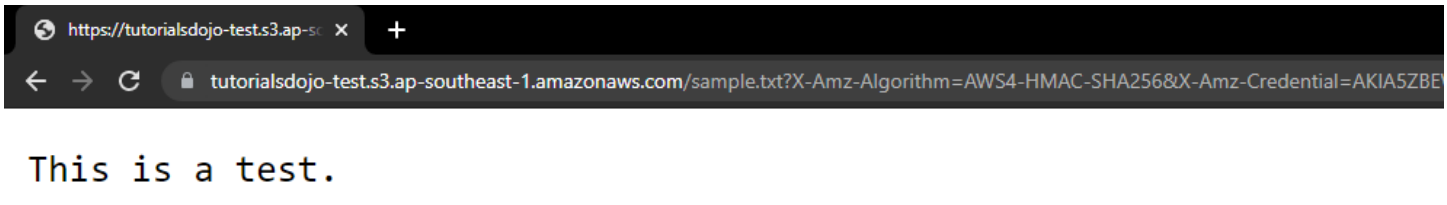
Example:

```
aws s3 presign s3://tutorialsdojo-test/sample.txt
```

Sample Output

```
https://tutorialsdojo-test.s3.ap-southeast-1.amazonaws.com/sample.txt?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIA5ZBEWVFGZ6VYNZLZ%2F20210710%2Fap-southeast-1%2Fs3%2Faws4_request&X-Amz-Date=20210710T171224Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=409f641d1b06f2122b17fb26f9ef56a84b475714b4eb7fec29151746d963df5a
```

Use the generated URL to grant temporary access on the object.



Optionally, you can define an access duration for the URL by adding the `--expires-in` parameter. If you didn't specify the duration, it would get a default value of 3600 seconds.

```
aws s3 presign <s3objectURL> --expires-in <value>
```

Uploading S3 Objects Using Presigned URL

Presigned URL also enables you to upload an object to a specific bucket. Users who upload an object using the presigned URL need no credentials or permissions.

To generate a presigned URL for uploading objects, you can use AWS SDK. You need to provide valid security credentials and specify the following parameters.

- Bucket Name
- Object Key
- HTTP method (PUT method)
- Expiration

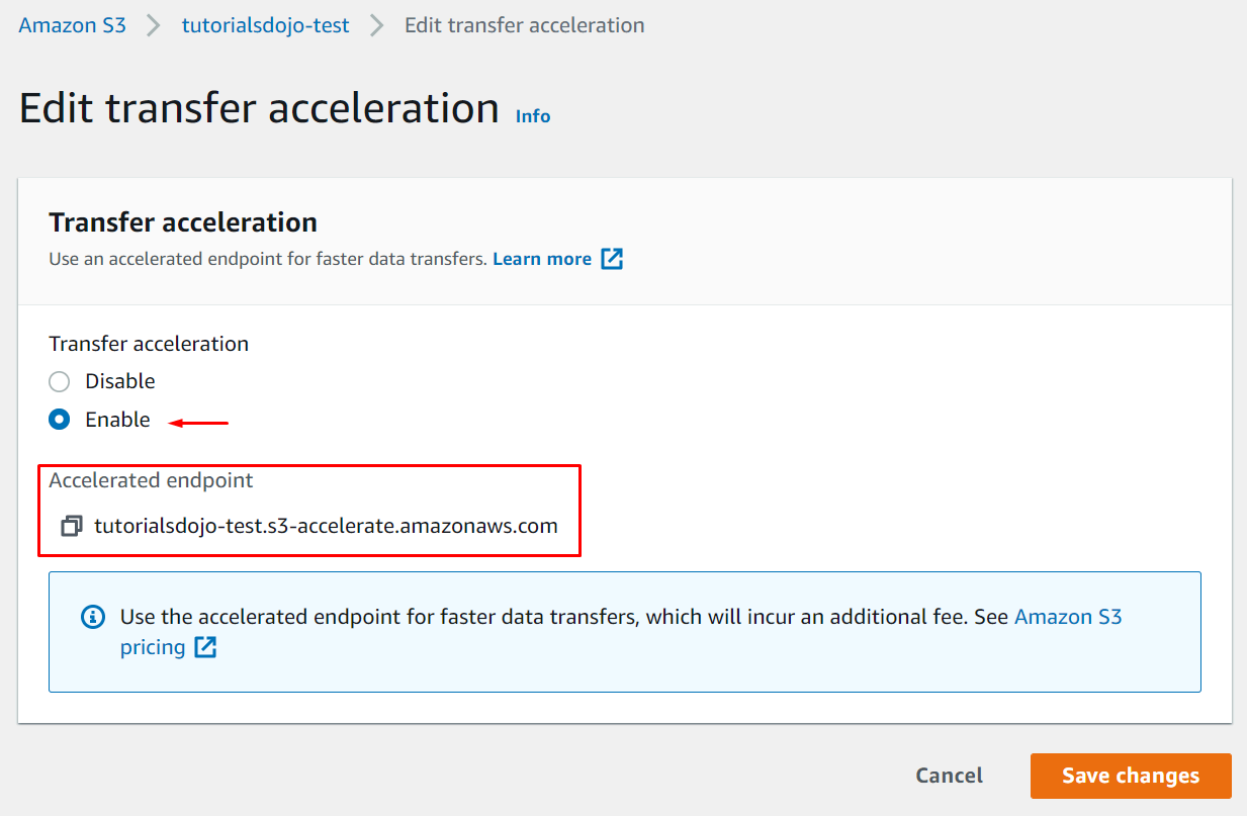
In this example, we use AWS SDK for Python. The function *generate_presigned_url* will generate the presigned URL with the defined parameters. You can set the validity by defining the ExpiresIn value in seconds.

```
import boto3
   url = boto3.client('s3').generate_presigned_url(
   ClientMethod='put_object',
   Params={'Bucket': 'S3BucketName', 'Key': 'ObjectKey'},
   ExpiresIn=3600)
```

Reference:
https://docs.aws.amazon.com/AmazonS3/latest/userguide/using-presigned-url.html

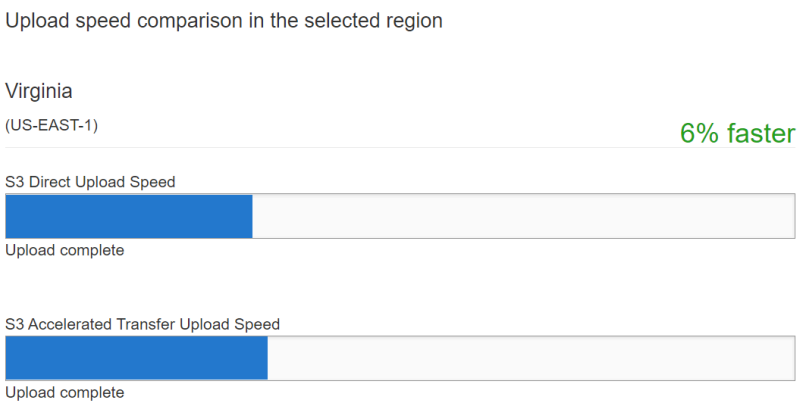
S3 Transfer Acceleration

When catering to a global audience, the distance between you and your client is undoubtedly an issue. This is also applicable for the S3 service, especially when uploading large objects across the regions. It will surely take a long time to upload hundreds of gigabytes of things to an S3 bucket located on the other side of the globe. To address this issue, you can enable the S3 Transfer Acceleration feature on S3 buckets.



S3 Transfer Acceleration integrates with Amazon CloudFront and uses its AWS Edge location around the globe to create an accelerated endpoint for a much faster data transfer. In this way, your S3 bucket is much closer to your users. Data will be uploaded on an AWS Edge Location and routed to your S3 bucket through the AWS internal network.

AWS offers a Speed Comparison tool to compare the performance of accelerated transfer and standard S3 data transfer in different AWS regions.



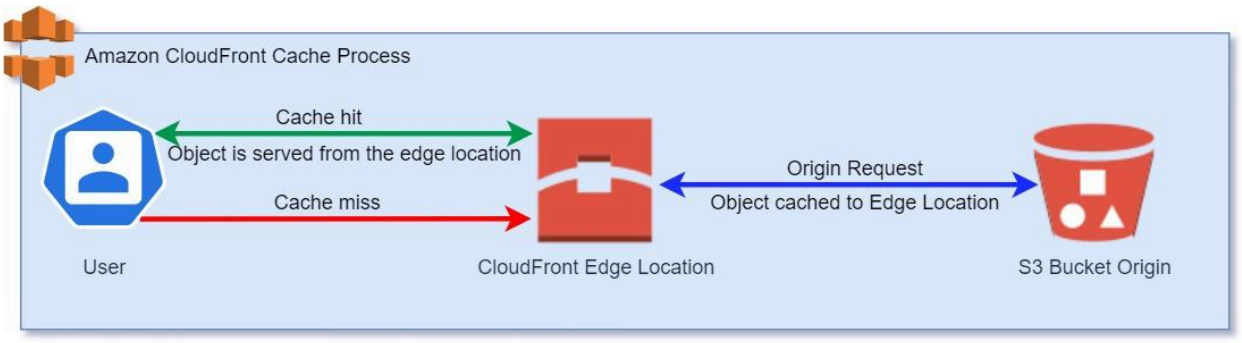
Reference:
https://docs.aws.amazon.com/AmazonS3/latest/userguide/transfer-acceleration.html

Amazon CloudFront

CloudFront enables customers to deliver content throughout the globe by using CloudFront edge location. Content is cached to different edge locations from an origin server, allowing viewer requests to be directed to the nearest edge location instead of the origin server.

Caching Process

Every object being cached has a unique identifier called **Cache Key** associated with them. This key may contain values like domain name, URL, HTTP header, query strings, and cookies. When a viewer sends a request to an edge location, CloudFront determines if the request is a cache hit or miss by comparing the values of the request with the values in the cache key.



A cache hit occurs when the request and cache key's value matches, thus confirming that the requested object is cached on the edge location. On the other hand, the request is forwarded to the origin server (origin request) to retrieve the requested object during a cache miss event. The **Origin Request** contains values like URL, request body, and headers. The origin server then sends the object to the user through the edge location and caching simultaneously.

The cache key directly affects the cache hit ratio of an object; the more values defined on the cache key, the lower the number of hits that can affect your application or website performance.

CloudFront Policies

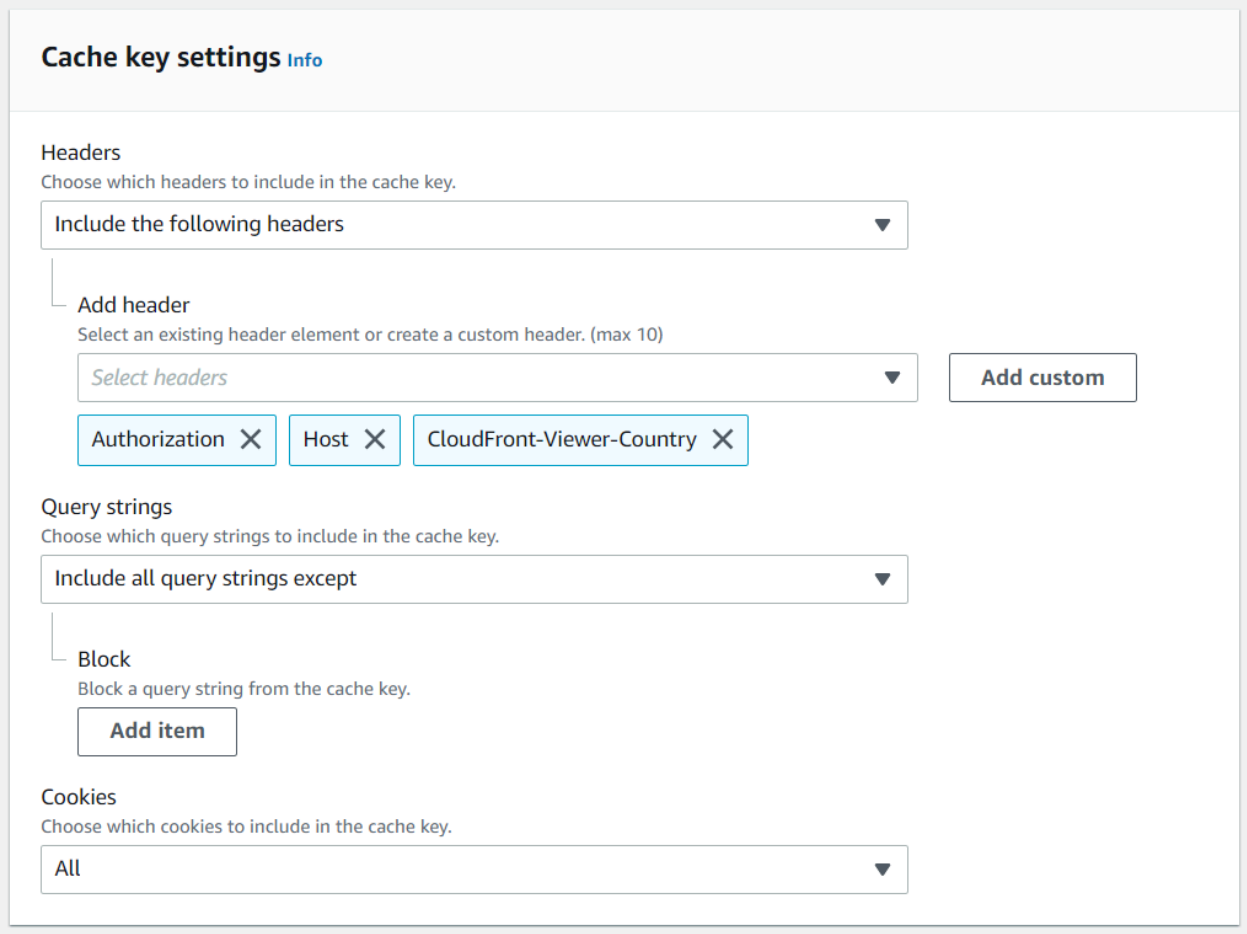
CloudFront Policies help you to granularly define what values you want to include in Cache Key and Origin Request. First, it is essential to determine what values your application/website requires from the viewer's request. You may also want to choose the values being forwarded to the origin server if you plan to collect these data.

Cache Policy

Amazon CloudFront provides a preconfigured **AWS-managed cache policy** that you can use for your distribution. Below is a comparison of the AWS-managed cache policy.

	Cache Key Settings			Compression Support	
Cache Policy	**Header**	**Cookies**	**Query Strings**	**Gzip**	**Brotli**
Amplify	Authorization CloudFront-Viewer-Country Host	All	All	Yes	Yes
CachingDisabled	None	None	None	No	No
CachingOptimized	None	None	None	Yes	Yes
CachingOptimizedForUncompressedObjects	None	None	None	No	No
Elemental-MediaPackage	origin	None	aws.manifestfilter start end	Yes	No

If the AWS-managed cache policy doesn't fit your requirements, you can always create a **custom cache policy**. A custom cache policy gives you the flexibility to specify header, and block or accept a particular query string and cookies.



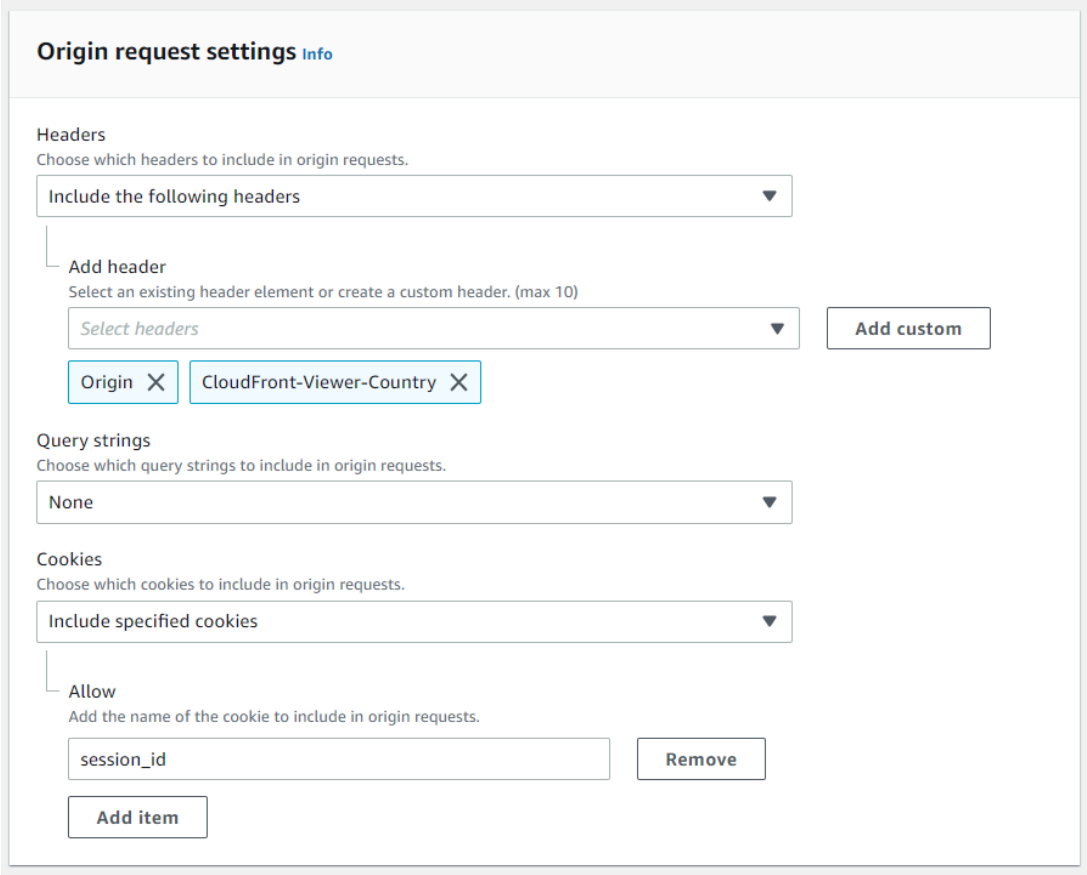
Origin Request Policy

Origin Request Policy defines the information included in the origin requests. AWS also provides a preconfigured **AWS-managed origin request** policy for origin requests. Below is a comparison of the AWS-managed origin request policy.

	Cache Key Settings		
Origin Request Policy	**Header**	**Cookies**	**Query Strings**
AllViewer	All	All	All
CORS-CustomOrigin	origin	None	None
CORS-S3Origin	origin access-control-request-headers access-control-request-method	None	None

Elemental-MediaTailor-PersonalizedManifests	origin access-control-request-headers x-forwarded-for access-control-request-method user-agent	None	All
UserAgentRefererHeaders	referer user-agent	None	None

If the AWS-managed origin request policy doesn't fit your requirements, you can always create a custom cache policy. A custom origin request policy gives you the flexibility to specify header, query strings, and cookies.

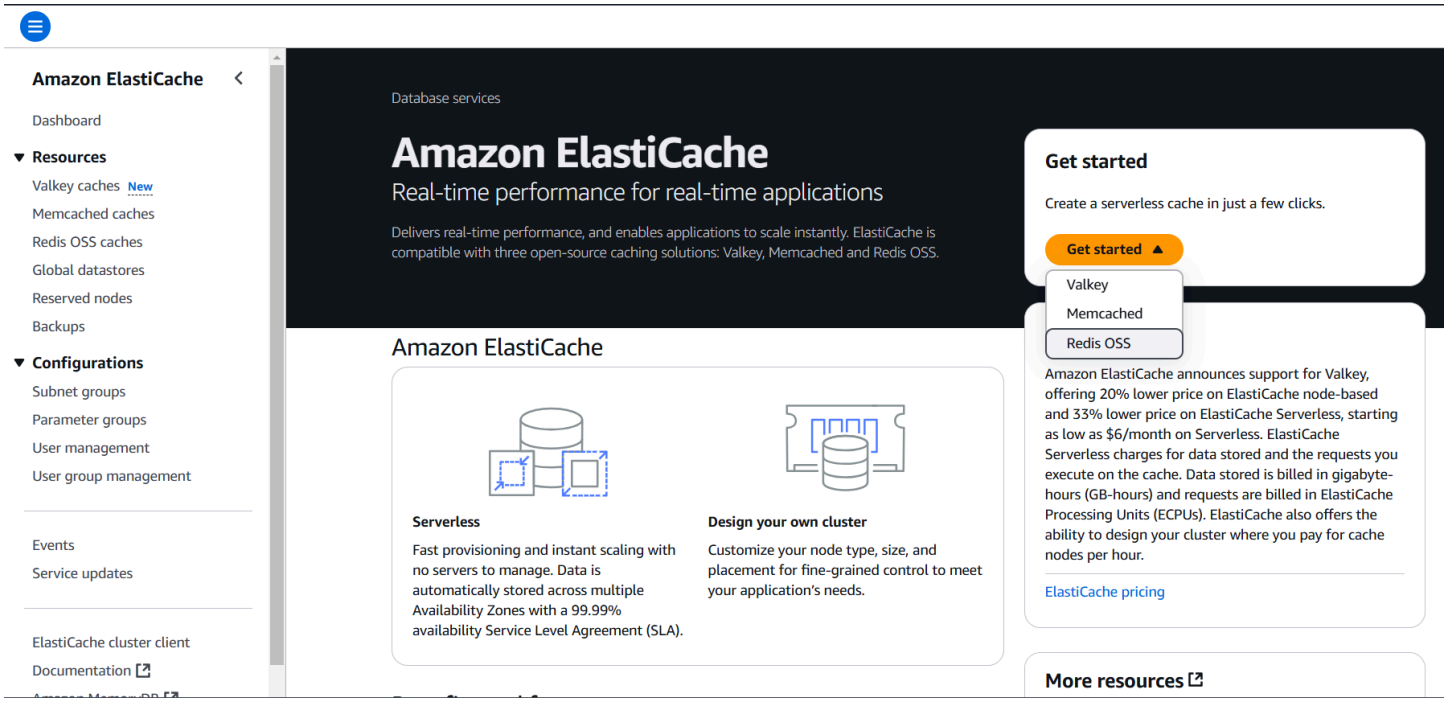


Reference:
https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/HowCloudFrontWorks.html
https://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/working-with-policies.html

Amazon ElastiCache

ElastiCache Memcached and Redis Engine

AWS offers a fully managed service for in-memory caching for your applications. ElastiCache generally supports both Memcached and Redis engines. Both are identical, but you certainly shouldn't easily choose a specific engine over the other since both of them have specific use cases. Here are the things you might want to consider when choosing an ElastiCache engine.



Clusters

To start using ElastiCache, you create a cluster and specify the engine to be used. A cluster contains one or more cache nodes with their memory and compute resources according to the node type. Each node in a cluster runs a cache engine. For the Redis engine, you can choose to enable cluster mode during cluster creation. Redis (Cluster Mode enabled) offers a limited option for cluster modification, but Memcached and Redis (Cluster Mode disabled) support adding and removing nodes from the cluster.

Sharding

One key difference between Memcached and Redis is the sharding feature. Memcached doesn't support the use of shards, while Redis does. Shard is a hierarchical arrangement of multiple nodes that allows you to have primary and replica nodes. Redis (Cluster Mode enabled) may contain up to 500 shards, while Redis (Cluster

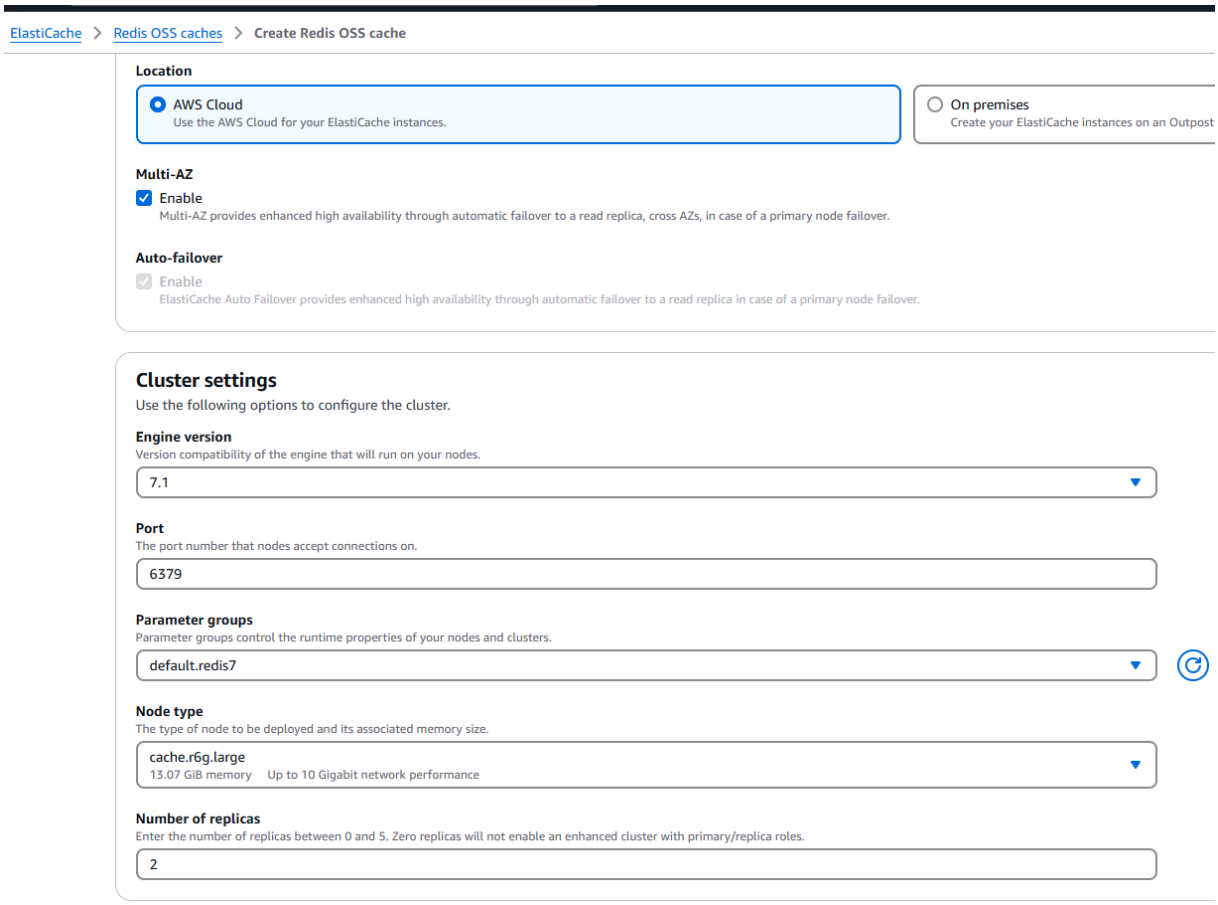
Mode disabled) may only have a single shard on a cluster. Both Redis and Memcached can split their data across all nodes. Still, sharding allows for much efficient processing, which is beneficial when working with a large scale in-memory caching.

Multithreading

Memcache supports multithreading allowing it to process through multiple cores and making it easy to scale up its computing resources. On the other hand, Redis runs on a single thread, which allows it to scale horizontally on a cluster.

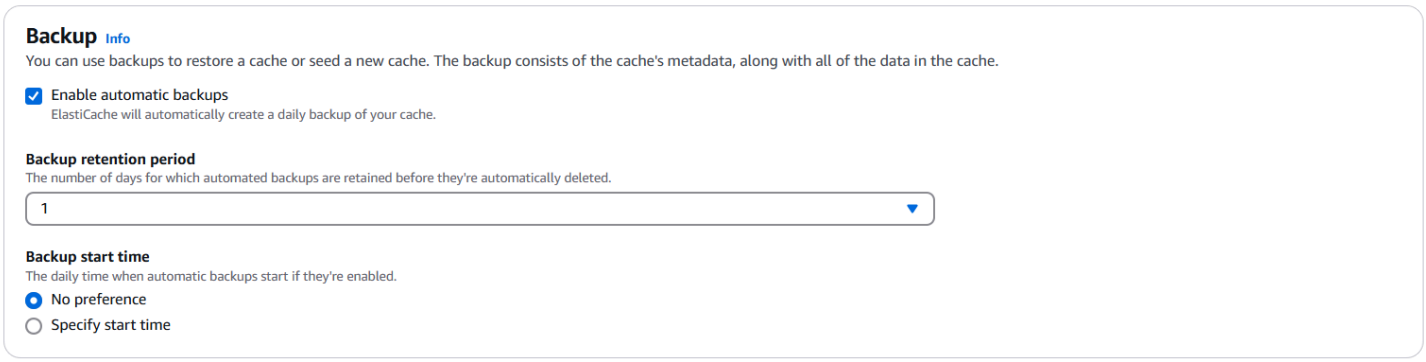
High Availability

Suppose high availability is a crucial requirement for your caching workloads. In that case, Redis is the engine to choose from since Memcache doesn't support replication and high availability. Redis supports replication through sharding. For Redis (Cluster Mode disabled), all nodes can contain all the cluster's data in a single shard, while for Redis (Cluster Mode enabled), data are partitioned across all shards. This feature also enables Redis for automatic failover. Below is an example of a node, replica, and Multi-AZ configuration on ElastiCache Redis.



Backup and Restore

An ElastiCache cluster running on the Redis engine can use S3 to export a backup. You can set up a scheduled automatic backup or trigger a manual backup for your Redis cluster. You can also create a Redis cluster using a backup from an S3 bucket. On the other hand, Memcached doesn't support backup and restore. Here's an example backup configuration of the ElastiCache cluster using the Redis engine.



Key Points:

The workloads you'll be processing must be considered when choosing the right engine for ElastiCache. If you'll be working with simple data types like objects and need the capability to scale depending on the number of workloads, it is recommended to use Memcached. If your data includes complex data types (string, sets, sorted sets, lists, hashes) and requires high availability and failover capabilities, Redis is the right choice.

References:

https://docs.aws.amazon.com/AmazonElastiCache/latest/red-ug/SelectEngine.html
https://docs.aws.amazon.com/AmazonElastiCache/latest/red-ug/WhatIs.html
https://docs.aws.amazon.com/AmazonElastiCache/latest/mem-ug/WhatIs.html

Virtual Private Cloud

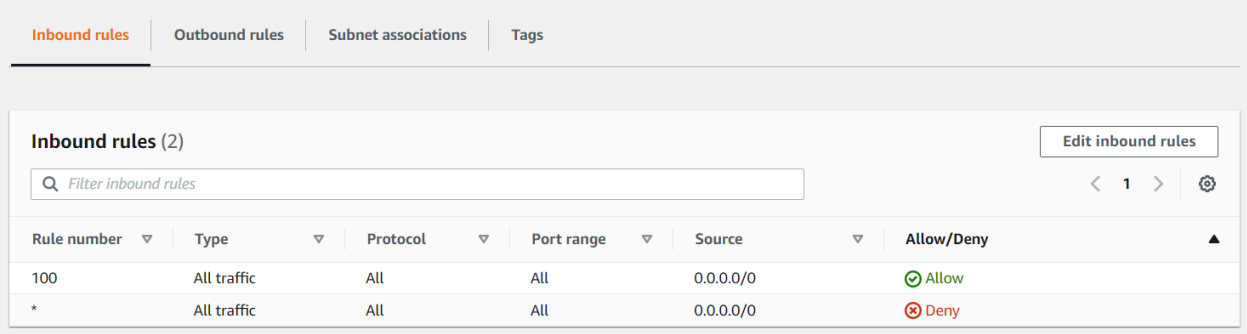
Network Access Control List (NACL)

AWS recommends applying security on any level of your architecture. For VPC Subnets, it's the **Network Access Control List (NACL)**. The NACL may allow or deny network traffic according to the rules you applied. Unlike a security group, the NACLs are stateless. Accepted inbound traffic is not automatically permitted on outbound; you need to define it explicitly. Changes made to an NACL are also applied immediately. You can't associate multiple NACLs to a subnet simultaneously, but you can associate an NACL to various subnets. You create a rule by specifying the following details.

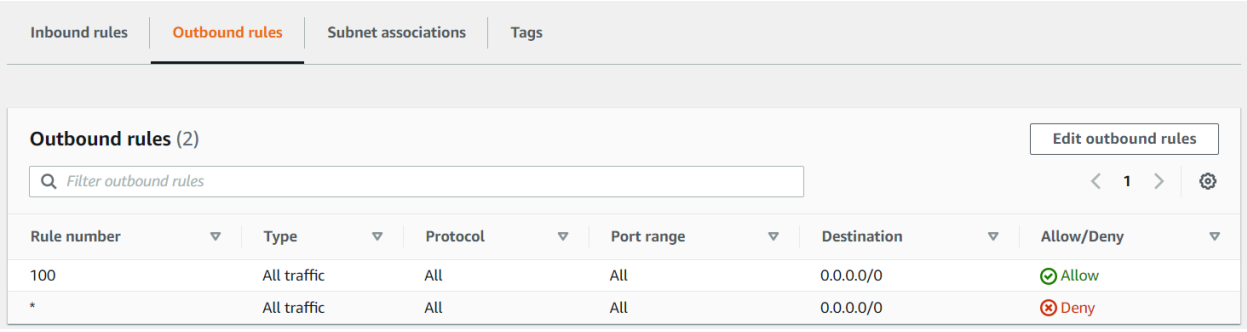
- Rule number.
- Type
- Protocol
- Port range
- Source (Inbound rules only)
- Destination (Outbound rules only)
- Action (Allow/Deny)

Each subnet requires an NACL associated with it. AWS provides a default network ACL that accepts both inbound and outbound traffic.

Default NACL Inbound Rules



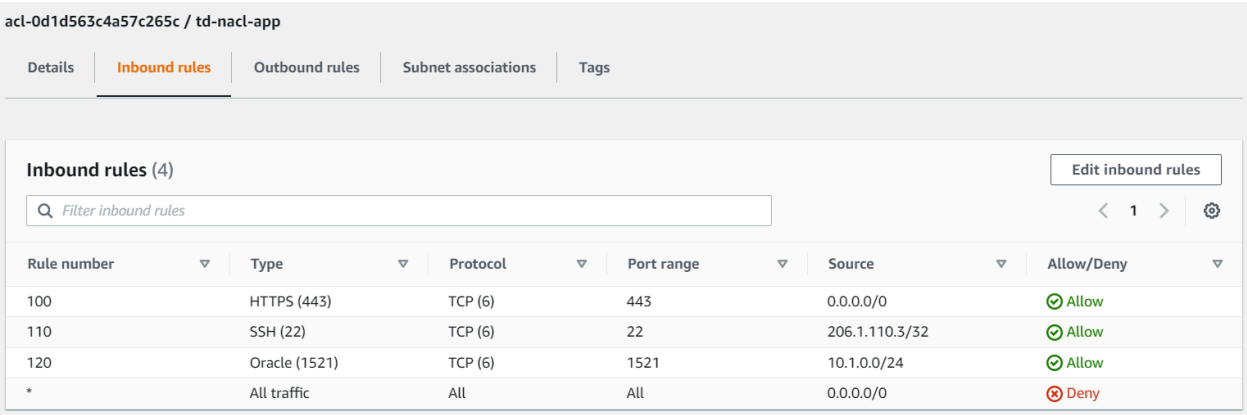
Default NACL Outbound Rules

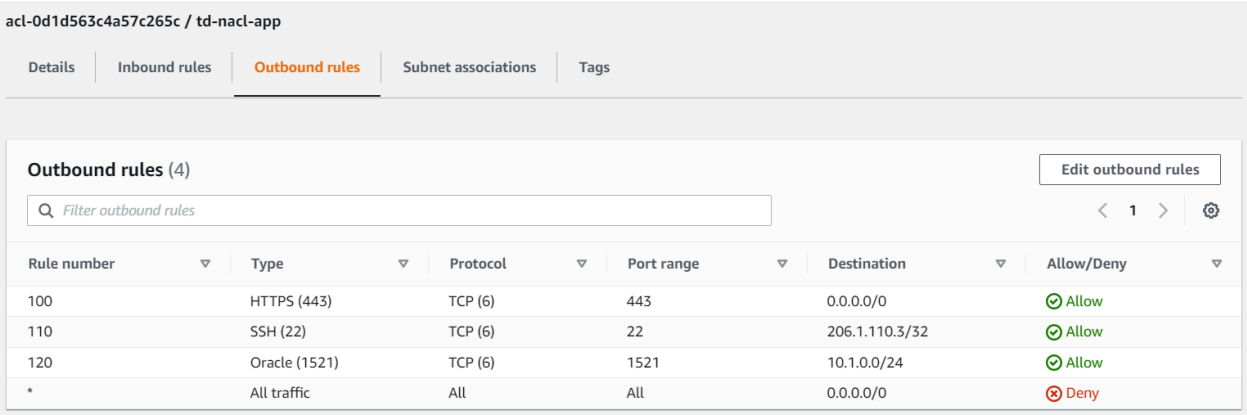


The default network ACL works fine if you don't have strict security requirements. Still, it's a good practice to leverage this feature to another layer of security in your cloud architecture.

You can create your own fine-grained rules on a custom network ACL and associate them with your subnets. Always consider and evaluate your application's traffic and traffic from your users to avoid any conflicts on the rule. By default, a newly created custom NACL denies all inbound traffic.

Below is an example of a custom NACL with fine-grained rules for inbound and outbound. The custom NACL allows HTTPS from IPv4 addresses, SSH from a specific IP address, and Oracle database connection from a particular CIDR block. The rule with the lowest rule number is evaluated first; in this case, the HTTPS rule is applied first, then followed by the rule for SSH and Oracle database traffic. The rule with an asterisk (*) on rule number guarantees that all network traffic that is not explicitly defined is denied. Note that a rule is immediately applied for specific traffic and doesn't consider conflicts with other rules with a higher rule number.



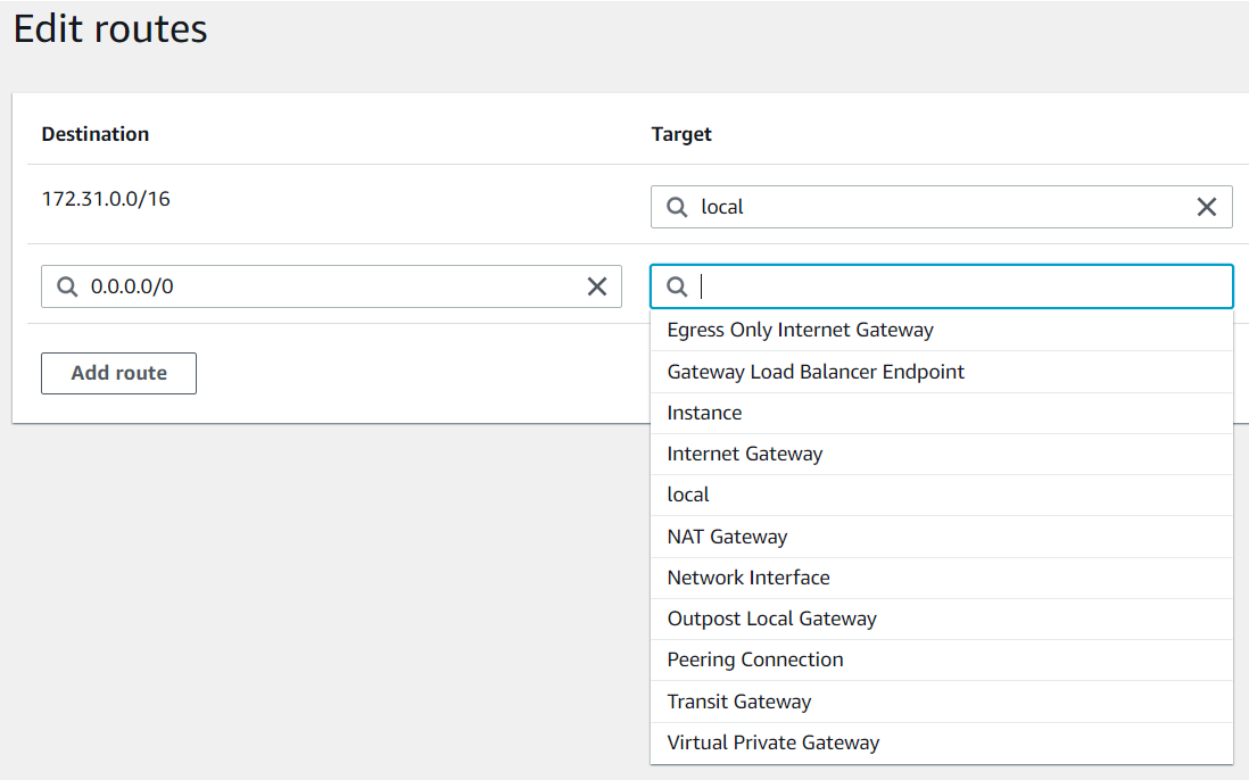


Route Tables

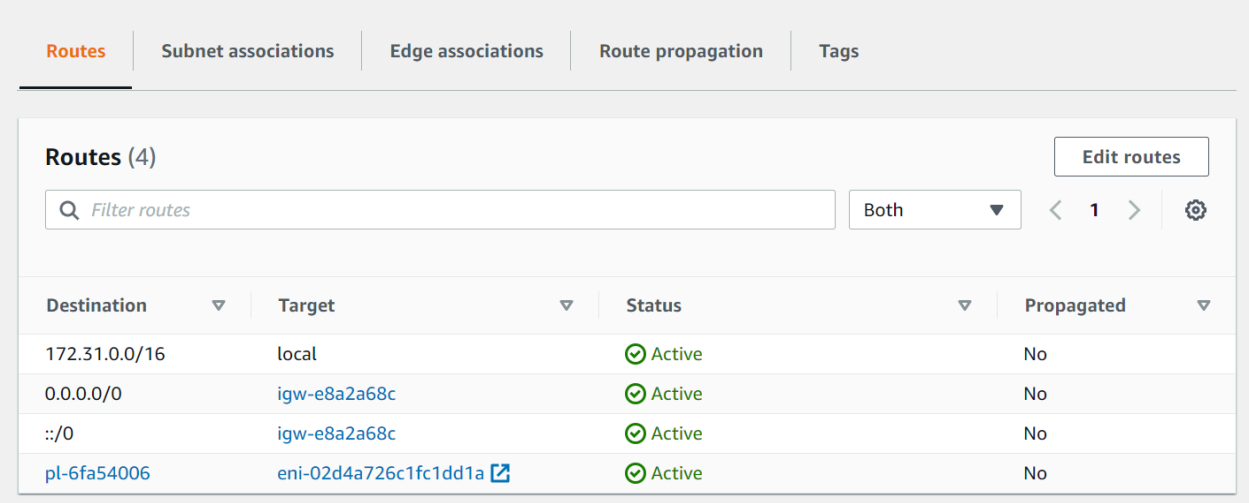
The Route tables control the traffic routing of your subnets and gateways. The route table evaluates the destination and target defined in the route to direct traffic. All subnets in a VPC require a route table. When a VPC is created, the main route table, which handles the traffic routing within that VPC, is also created. The main route table is implicitly associated with a VPC until you explicitly associate a new custom route table with it. Remember that you can associate a route table with multiple subnets. However, a subnet can be associated with only one route table at a time.

On the route table, you define a destination and target. A destination determines where the traffic from an IP address range or a prefix goes. Traffic from the destination is routed to whatever is defined on the route target.

- **Egress Only Internet Gateway** - allows only outbound IPv6 traffic to the public internet. Prohibits IPv6 connection from the internet
- **Gateway Load Balancer Endpoint** - a VPC endpoint that the Gateway Load Balancer uses for distributing traffic between the service provider and service consumer VPC
- **Instance** - EC2 instance within the same instance
- **Internet Gateway** - allows inbound and outbound communication between VPC and public internet
- **Local** - route for the local VPC
- **NAT Gateway** - resides on a public subnet. Allows internet connection for private subnets
- **Network Interface** - virtual network card attached to an instance
- **Outpost Local Gateway** - allows communication between VPC and on-premises network
- **Peering Connection** - the connection between two VPCs
- **Transit Gateway** - interconnects multiple VPC and on-premises networks using a network transit hub
- **Virtual Private Gateway** - virtual router used for VPN tunnel



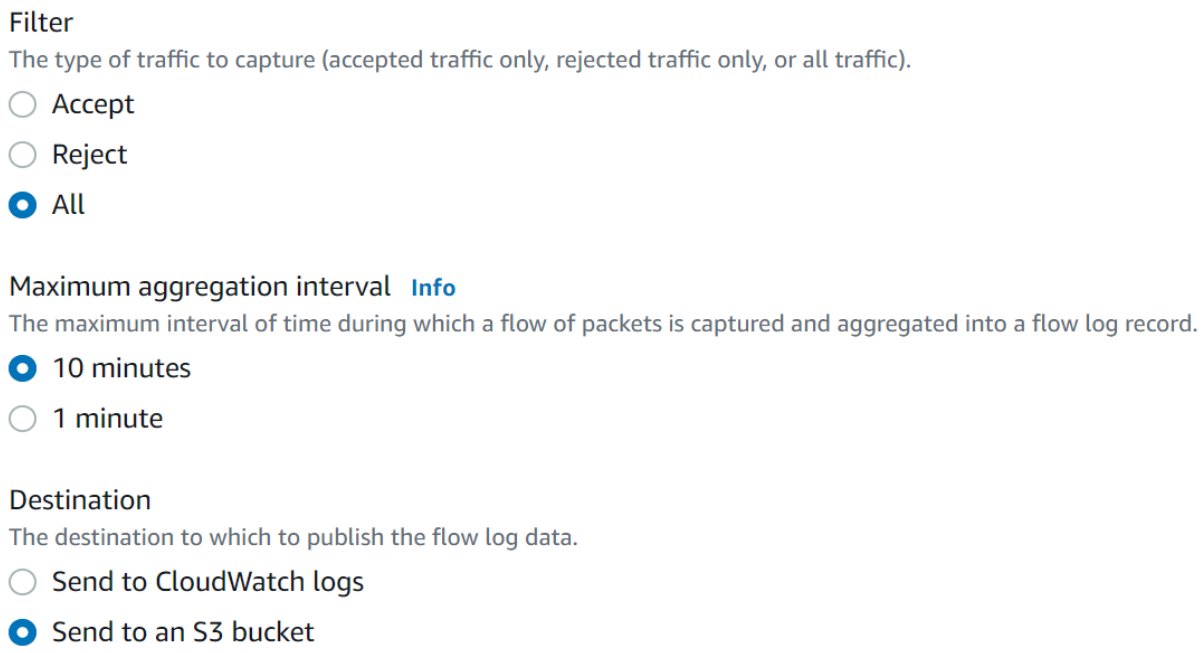
Here is an example of a custom route table. Creating a route table automatically creates a local route to enable traffic routing within the VPC. You also see a route for all IPv4 and IPv6 addresses routed to an Internet Gateway. You can route traffic from a prefix (Amazon S3) to a specific network interface within a VPC.



VPC Flow logs

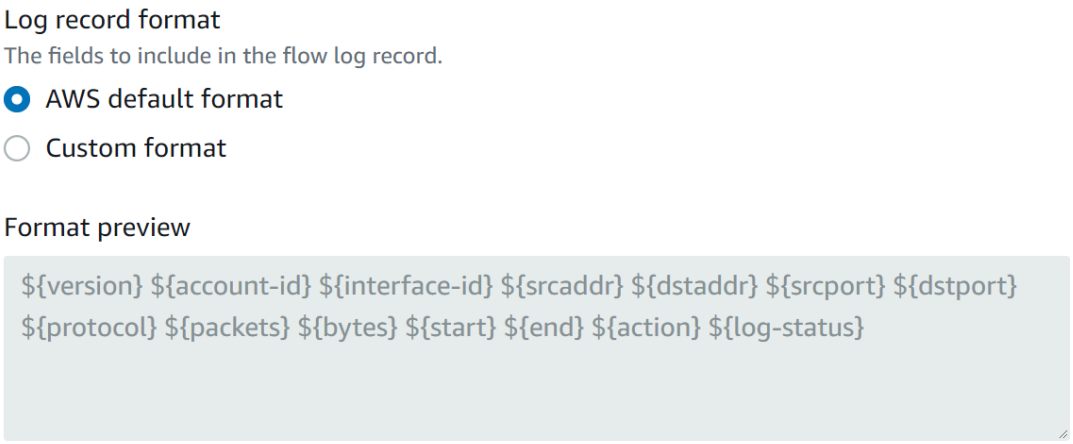
Using VPC Flow logs, you can track IP traffic to and from your network interfaces in a VPC. This feature helps you to monitor traffic or troubleshoot a network-related operational issue.

You create a Flow Log on the Network Interface on the EC2 console. You can set the Flow log to capture accepted and rejected traffic only, or both. You also set the logging interval to either 10 minutes or 1 minute.

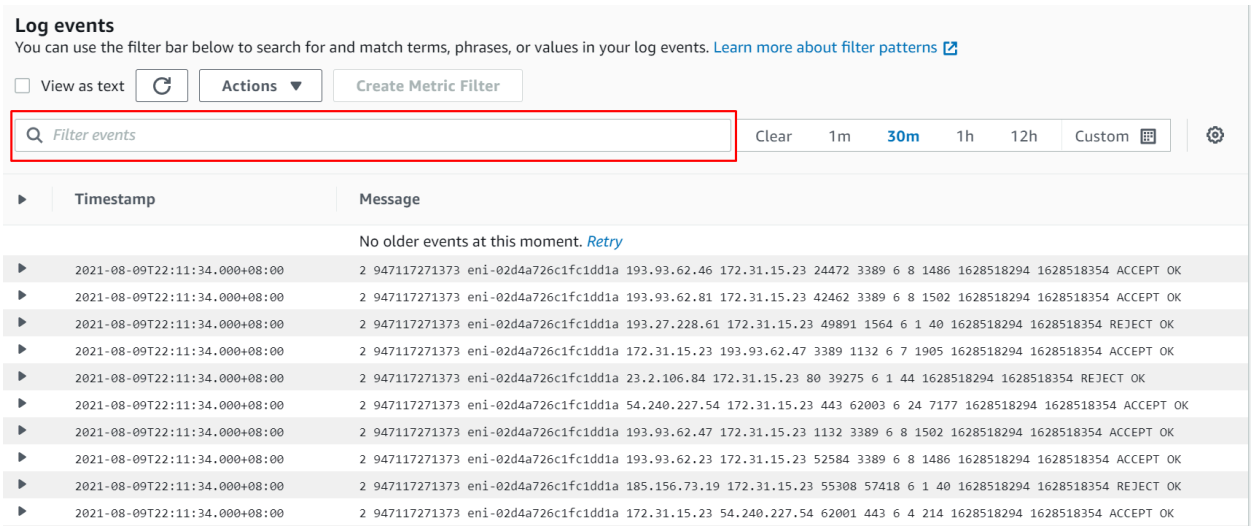


When you create a Flow log, you have an option to publish the data into an Amazon CloudWatch logs or Amazon S3 bucket. When you choose to publish the data to Amazon CloudWatch logs, you need to specify the log and an IAM role; otherwise, set the ARN for the Amazon S3 bucket.

You can also choose the log format. You can use the AWS default format or choose only specific log attributes by selecting a custom format.



Here is an example of a flow log published to an Amazon CloudWatch Logs log group. The log filter is helpful for things like viewing traffic from a specific address, checking traffic on particular ports, and others.



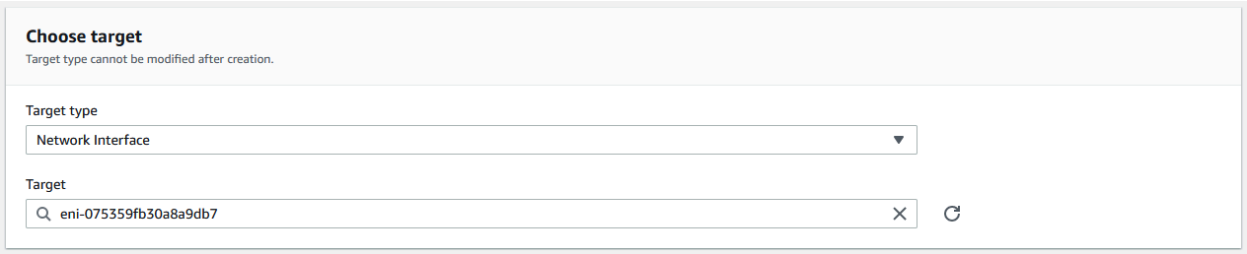
Traffic Mirroring

Traffic Mirroring is a VPC feature that allows you to stream both inbound and outbound traffic from a network interface to multiple targets. With traffic mirroring, you can analyze your network traffic in real time by sending the mirrored traffic to many monitoring/security devices. The device can be on the same VPC or a different VPC connected via VPC Peering, AWS Transit Gateway, or the Gateway Load Balancer endpoint. Since the network packets are captured from the network interface level, running an agent on your instance is unnecessary. With a simple configuration, you can quickly detect threats and identify vulnerabilities.

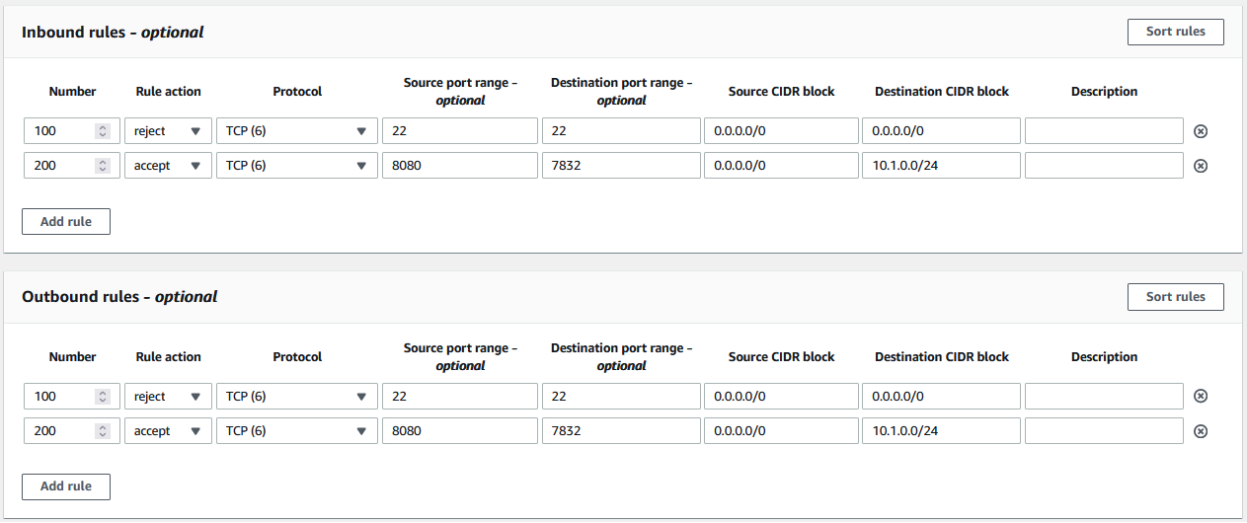
To use Traffic Mirroring, you need to configure the following.

Mirror targets - the destination of mirrored traffic. A target can be any of the following:

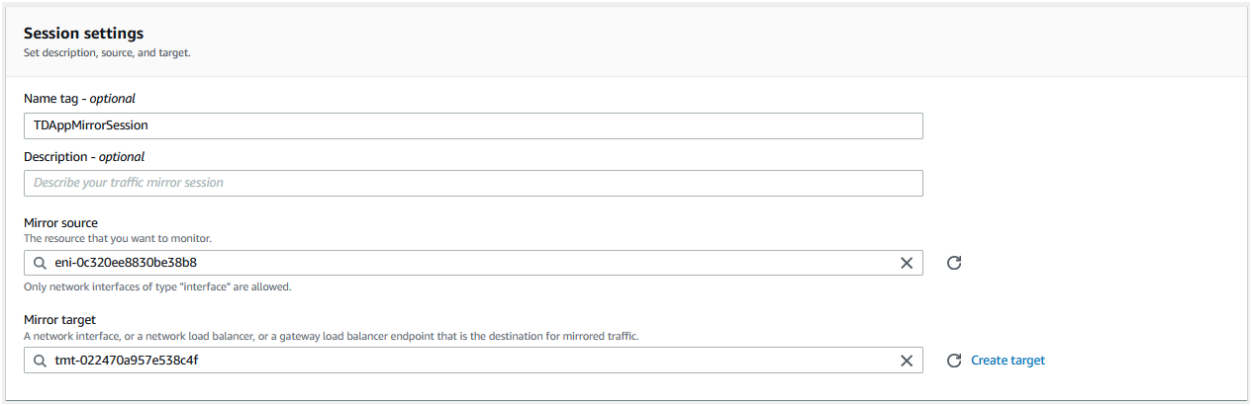
- Network interface
- Network Load Balancer with UDP listener
- Gateway Load Balancer with UDP listener



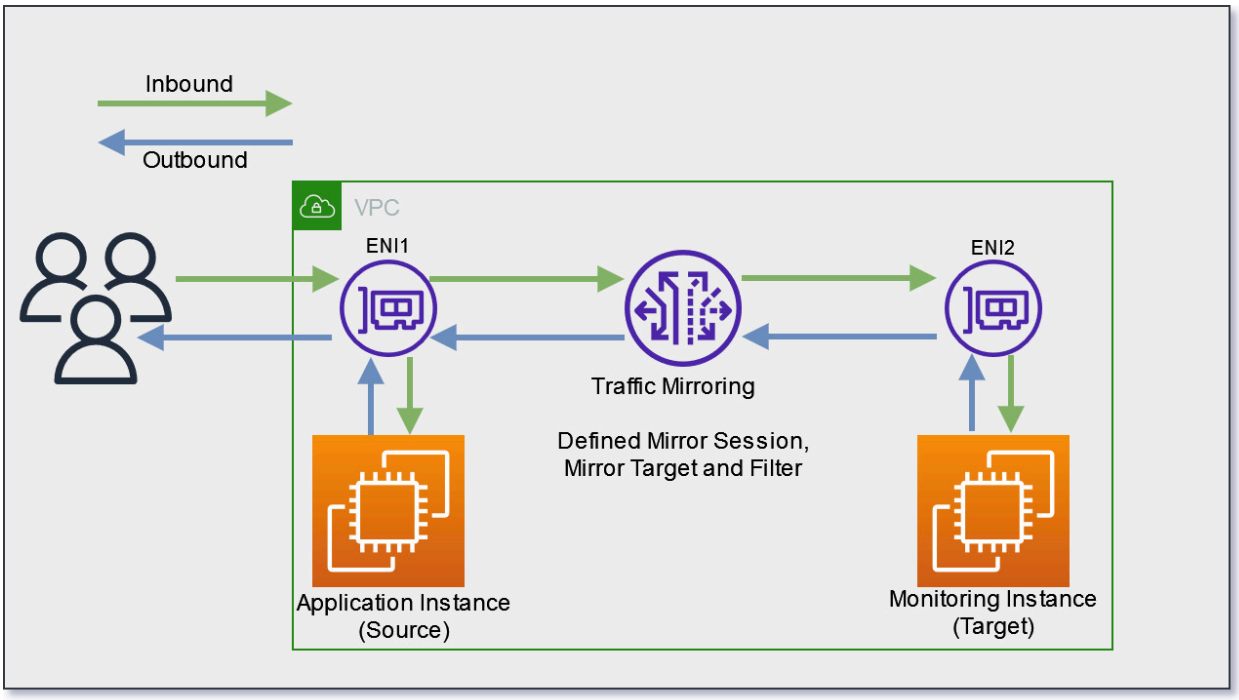
Mirror filter - defines the specific inbound and outbound traffic that is mirrored. No traffic is accepted by default; you need to accept or reject traffic explicitly.



Mirror sessions - set up the traffic mirroring by specifying the Mirror source, target, and Filter.



In the example below, users access an application hosted in an EC2 instance. On the same VPC, a monitoring instance is spawned (a SaaS or an instance with any network/packet analyzer capabilities). According to the defined Filter, inbound and outbound traffic from the users going to the application instance is mirrored in the monitoring instance. From there, the monitoring instance can see and analyze the same traffic the application has.



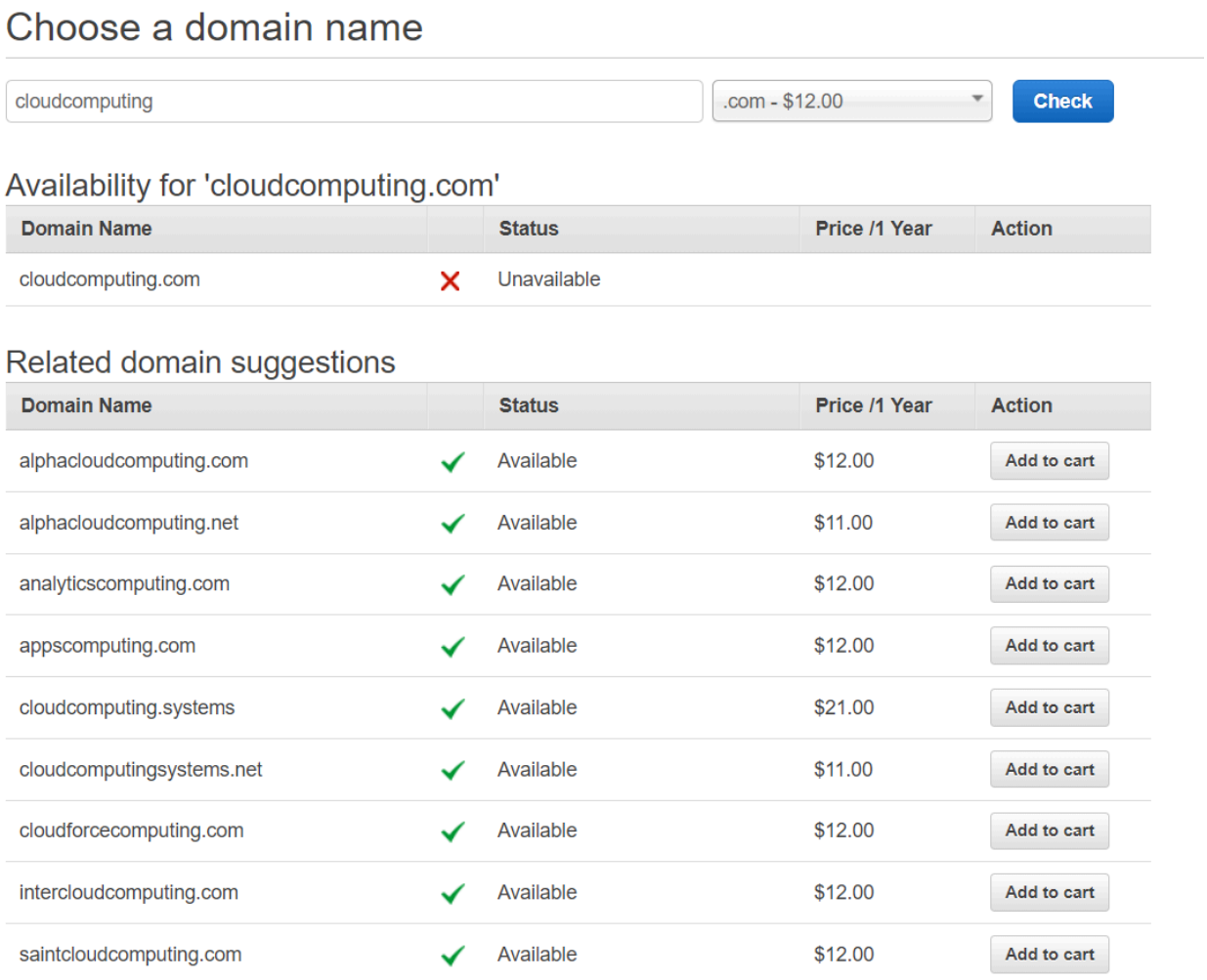
References:
https://docs.aws.amazon.com/vpc/latest/userguide/vpc-network-acls.html
https://docs.aws.amazon.com/vpc/latest/mirroring/what-is-traffic-mirroring.html

Amazon Route 53

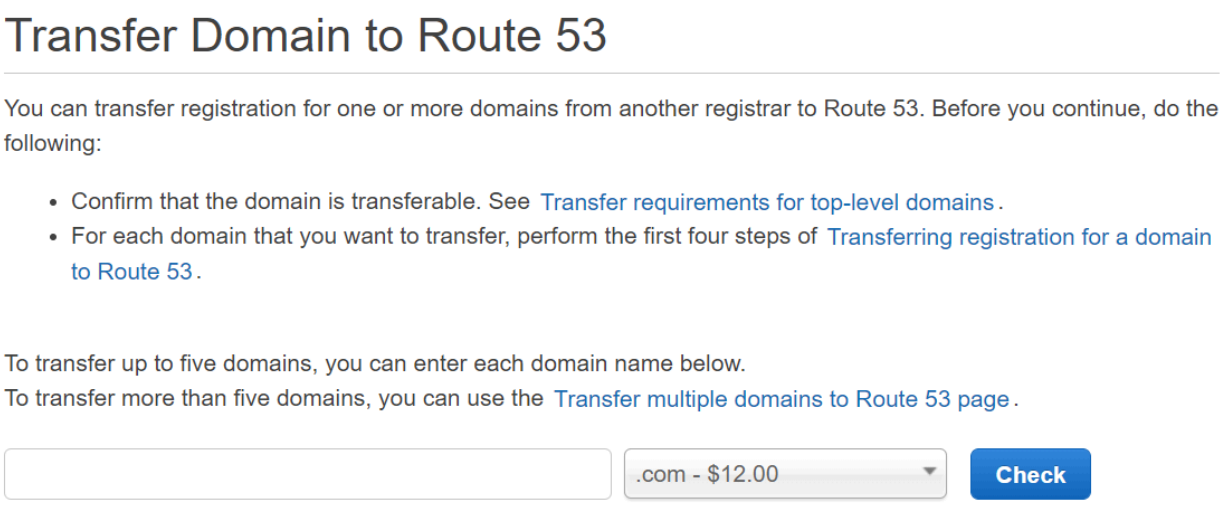
Amazon Route 53 is a scalable Domain Name Service (DNS) for your resources inside and outside an AWS environment. Route 53 gives you the capabilities to register a domain, use different routing policies for DNS routing, and monitor resources through health checks.

Domain Registration

Domain registration on Route 53 is straightforward. You search by providing a domain name and extension (.com, .net, etc.). Route 53 will tell you the availability of the domain and gives you domain suggestions. Domains are priced per year, and the prices also vary per domain extension.



If you have an existing domain on another DNS provider, you can also transfer its registration under Route 53.



Route 53 Service Integrations

Route 53 works with multiple AWS resources and even resources outside AWS. For AWS resources, you can route traffic to the following:

- **Amazon EC2** - route traffic to website and application hosted on EC2 instances using IP address.
- **Amazon VPC** - route traffic to a VPC endpoint using its IPv4 address
- **Elastic Load Balancing** - route traffic to Load Balancer's IPv4 address
- **Amazon API Gateway** - route traffic to an API endpoint using IPv4 address
- **Amazon CloudFront** - route traffic to CloudFront distribution using IPv4/IPv6 address
- **AWS Elastic Beanstalk** - route traffic to Elastic Beanstalk environment using IPv4 address
- **Amazon Lightsail** - route traffic to a Lightsail instance using IPv4 address
- **Amazon WorkMail** - route traffic to WorkMail using MX, TXT, and CNAME record
- **Amazon RDS** - route traffic to DB instance using its domain name
- **Amazon S3** - route traffic to a static website hosted on an S3 bucket using S3 website endpoint

Hosted Zones

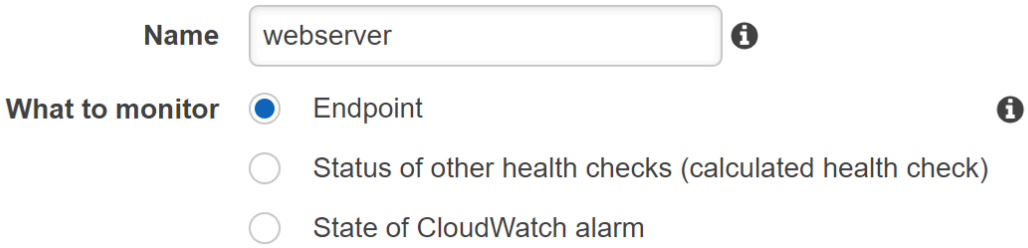
For easy routing management, hosted zones are configured to hold routing information. You can create a hosted zone by specifying your domain and the zone type. A hosted zone can be either of the two:

- **Public Hosted Zones** - traffic are routed on the internet
- **Private Hosted Zones** - traffic are routed within Amazon VPC

Route 53 Health Checks

Route 53 can monitor the health status of multiple resources for more efficient traffic distribution. You can set up endpoint monitoring for your resources like web servers by sending HTTP/HTTPS or TCP requests. Aside from monitoring endpoints, Route 53 can also monitor other health checks and CloudWatch Alarm. Monitoring the other health checks can be helpful if you are monitoring the number of resources with the same function. You still have health checks (child) on the individual resources, but instead of receiving multiple alerts, you only receive alerts from the parent health check that monitors the other health checks.

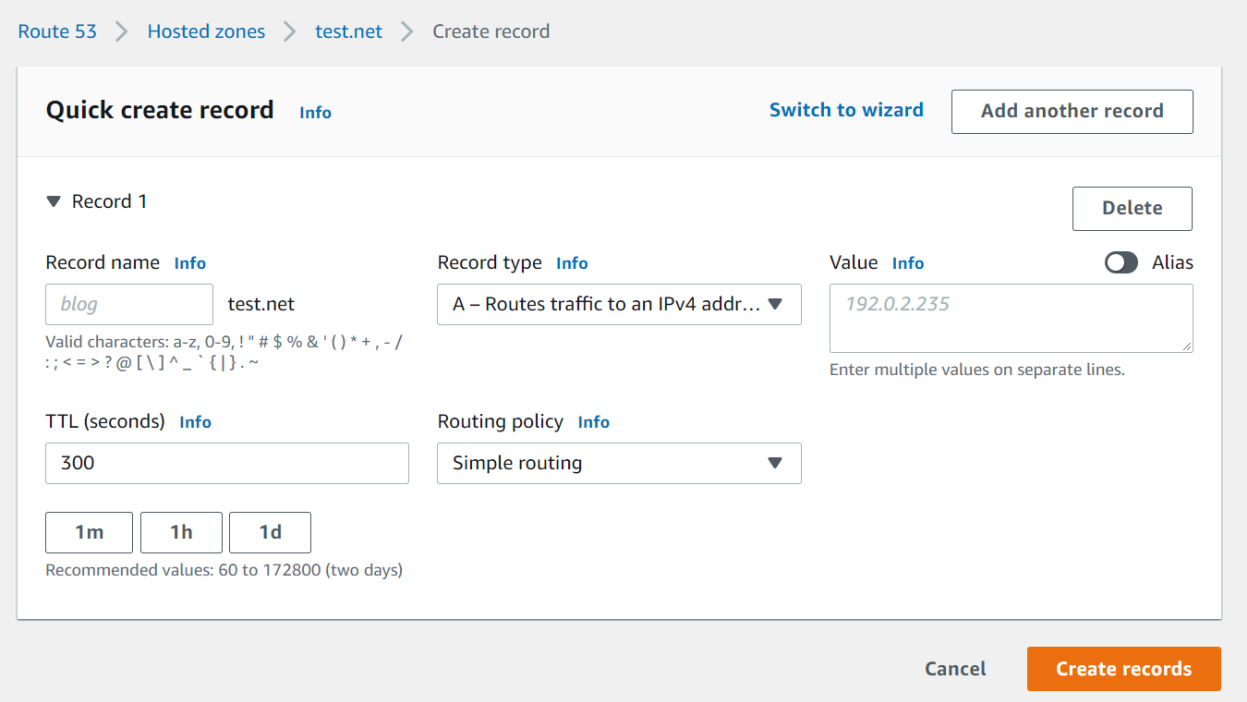
For CloudWatch Alarms, Route 53 health check returns a healthy status for *OK* state and unhealthy status for *ALARM* state. For *INSUFFICIENT* alarm state, you can configure a health check to either return a healthy, unhealthy, or the last known status.



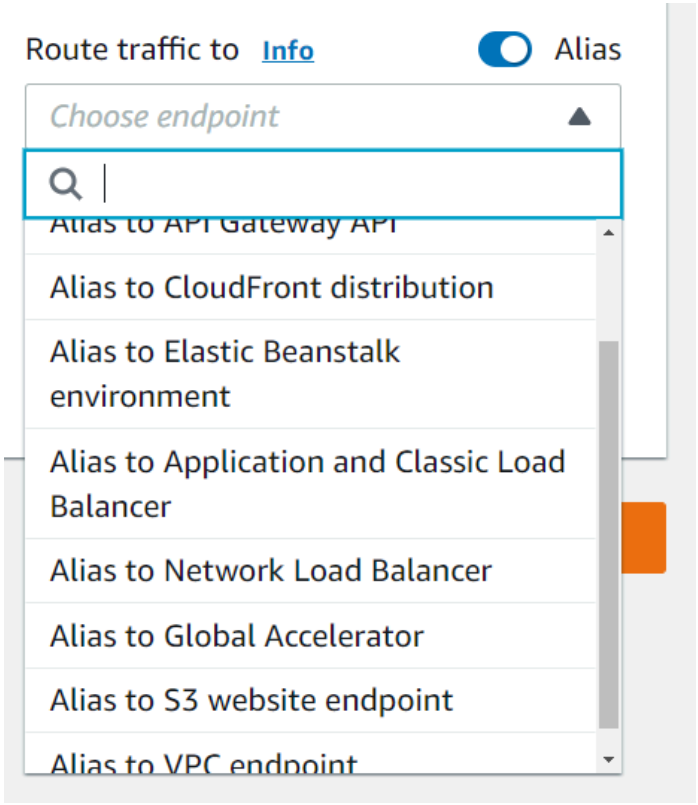
Route 53 health check can also monitor resources listed on a DNS record set for DNS failover. Route 53 can monitor resources specified in a record and do the DNS queries only to the healthy resources.

Route 53 Records

When you create a hosted zone, you can create records where you specify record types of your resources as well as the routing policy.



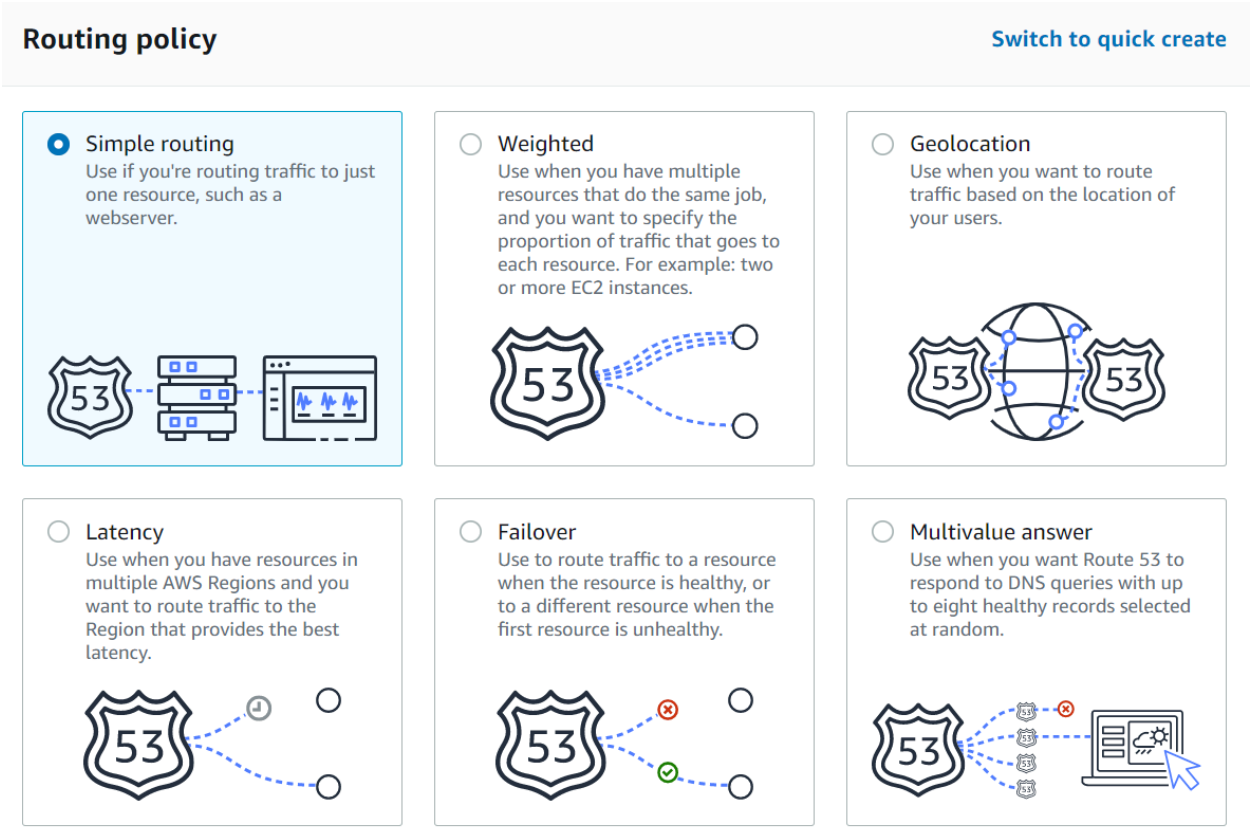
When routing traffic to other AWS resources, Route 53 supports the use of Aliases.



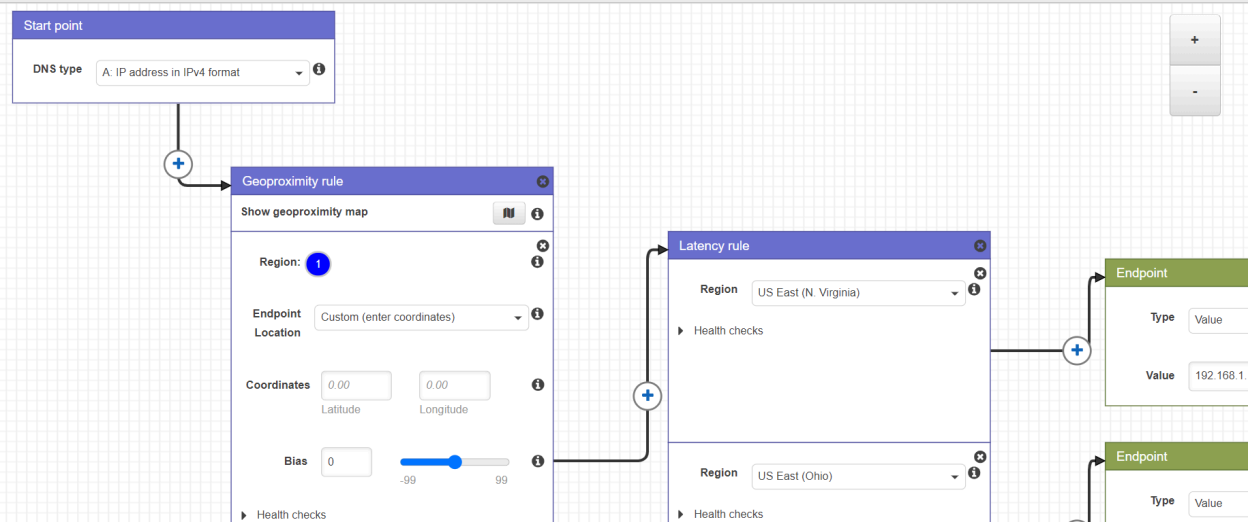
Routing Policy

How Route 53 distributes traffic to your resources will depend on the routing policy you specified. Route 53 offers the following routing policies.

- **Simple Routing Policy** - routes traffic to a single resource
- **Failover Routing Policy** - lets you define a primary and secondary resource. Traffic is automatically transferred to secondary resources when the primary resource is unhealthy
- **Geolocation Routing Policy** - route user traffic to specific resources based on their location
- **Latency Routing Policy** - route traffic to your resource's AWS region with the lowest latency
- **Multivalue Answer Routing Policy** - supports multiple value return for DNS queries
- **Weighted Routing Policy** - route traffic by specifying the weight to various resources.
- **Geoproximity Routing Policy** - determines the traffic routing according to the location of the user and resources. It can be used on Traffic Flow only.



For complex DNS routing of many resources, you can use Route 53 Traffic Flow. It is a visual editor to help you create complex DNS routing like using multiple routing policies for different resources.



DNS Record Types

When adding a resource to a record, you specify the record type and its value. Route 53 supports the following record types.

Record Type	Value/Used for
A	IPv4 address
AAAA	IPv6 address
CAA	Verification of certificates used by a domain/subdomain
CNAME	Map a domain to another domain/subdomain
DS	Establishing the chain of trust for DNS Security Extensions (DNSSEC)
MX	mail server
NAPTR	Dynamic Delegation Discovery System (DDDS) applications
NS	Provides Name Servers for a hosted zone
PTR	Map IP address to Domain
SOA	Provide DNS information about a domain
SPF	Verification of email sender (Not recommended)
SRV	Accessing services by defining priority, weight, and port

TXT	For application-specific values and verification of email sender

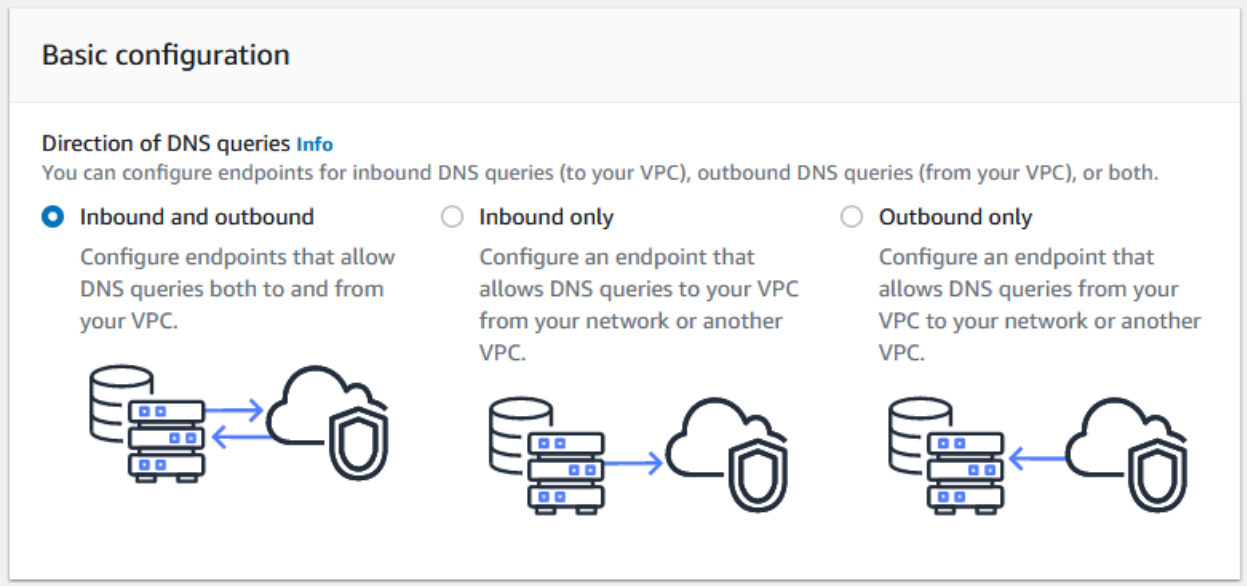
Route 53 Resolver

Route 53 Resolver is a DNS resolver that automatically answers DNS queries within a VPC. By default, it resolves queries for

- EC2 instances domain names
- Records in Private Hosted Zones
- Records in Public Name Servers

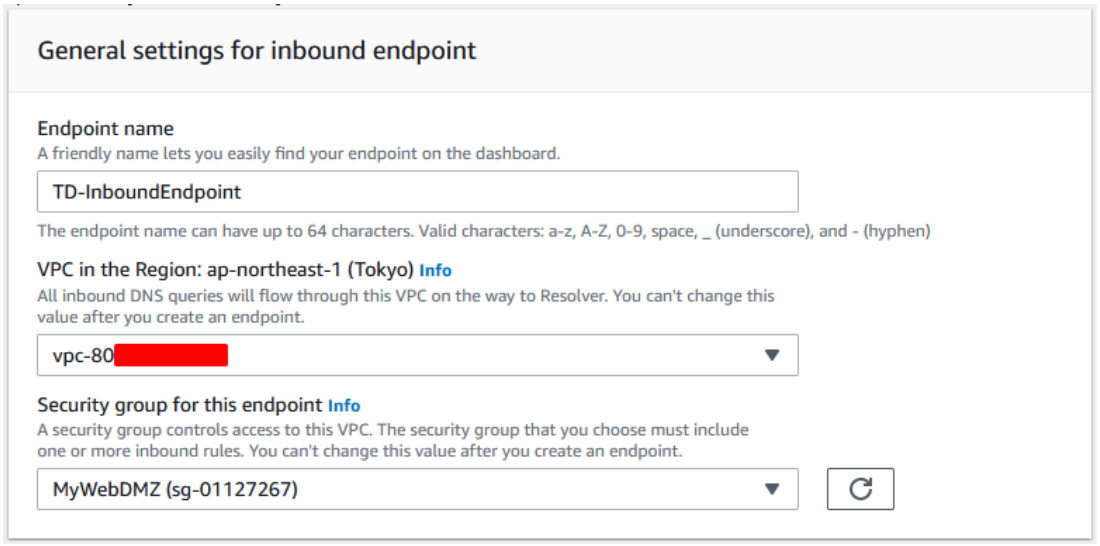
Resolver Endpoints

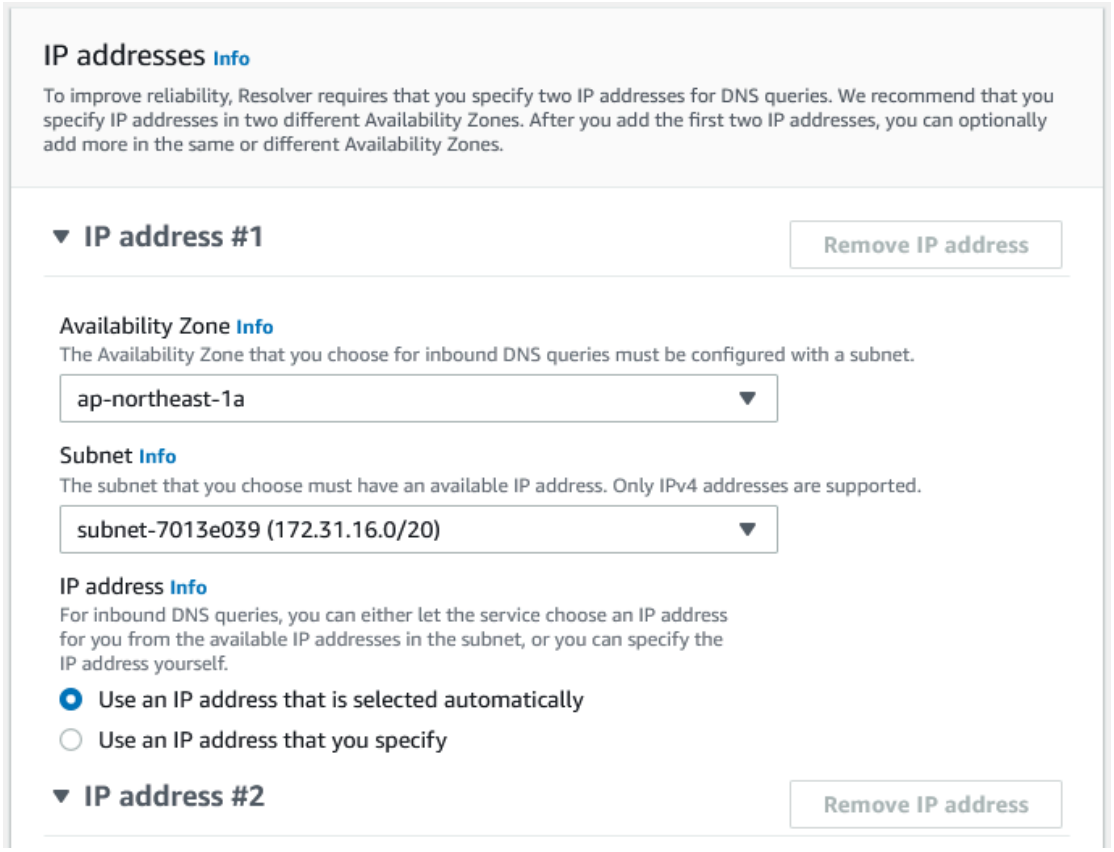
If there is a need to set up DNS resolution between a VPC and an on-premise network, you will need to create Resolver Endpoints. You can create an Inbound Endpoint (going to Route 53 resolver), Outbound Endpoint (going to another DNS resolver), or both.



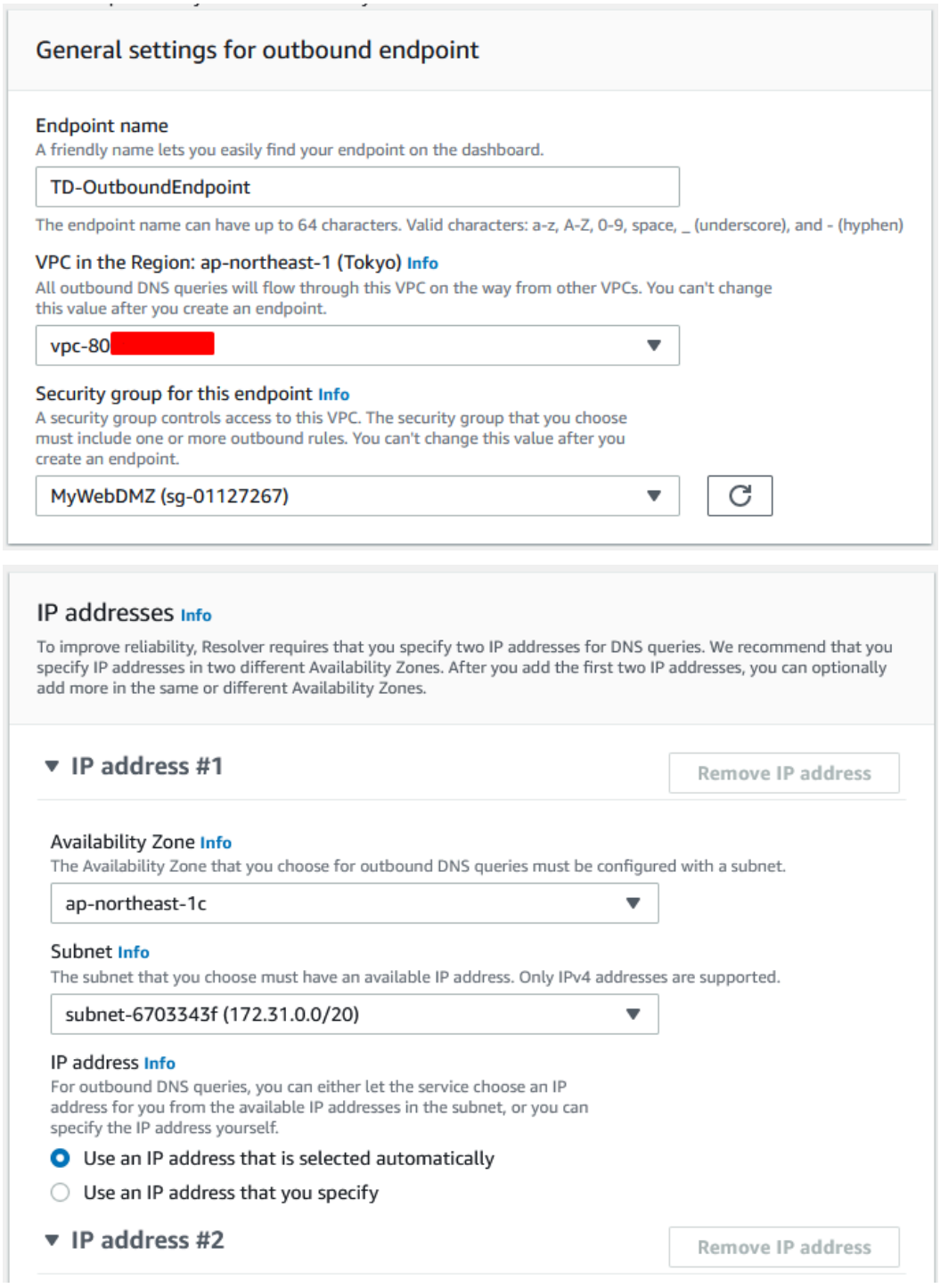
An on-premise network must be connected to your VPC via AWS Direct Connect/AWS VPN, or if with a different VPC, it must be connected using VPC Peering.

Inbound Endpoints - DNS Resolvers from on-premise network/peered VPC forward DNS queries to your Route 53 Resolver.





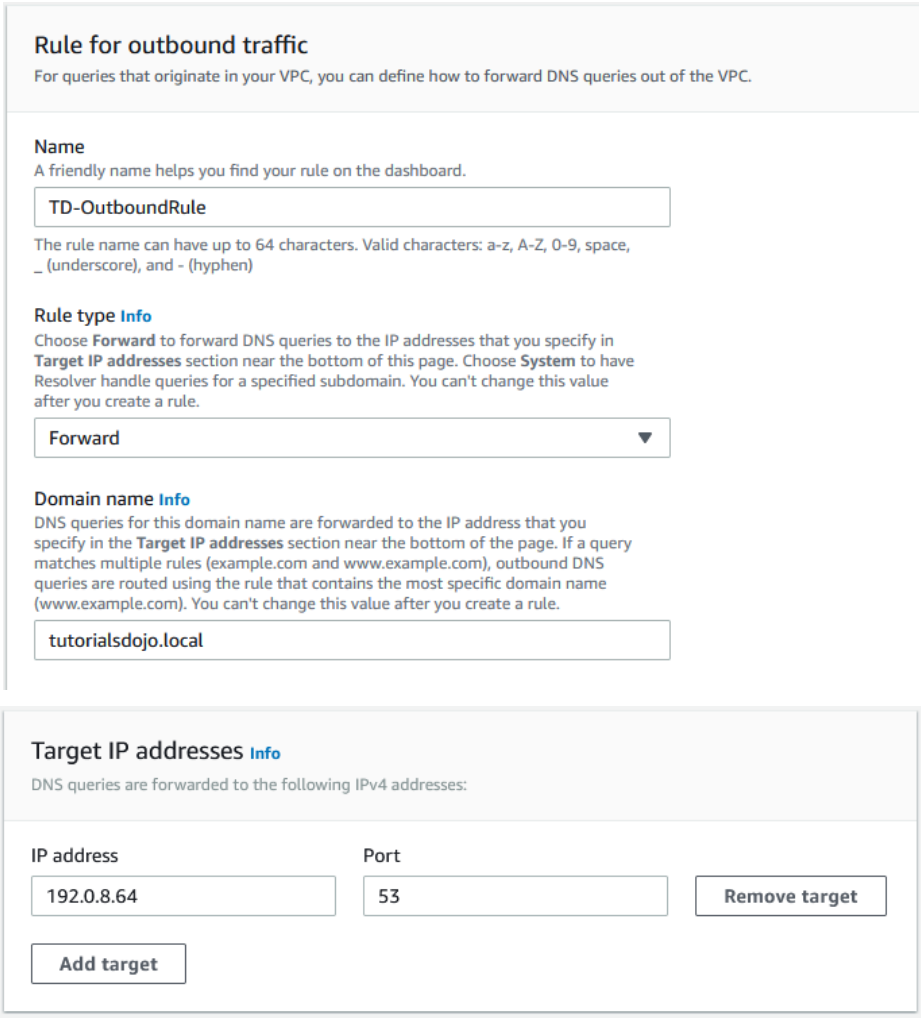
Outbound Endpoints - Route 53 Resolver uses Rules to conditionally forward DNS queries to DNS Resolvers from on-premise network/peered VPC.



Resolver Rules

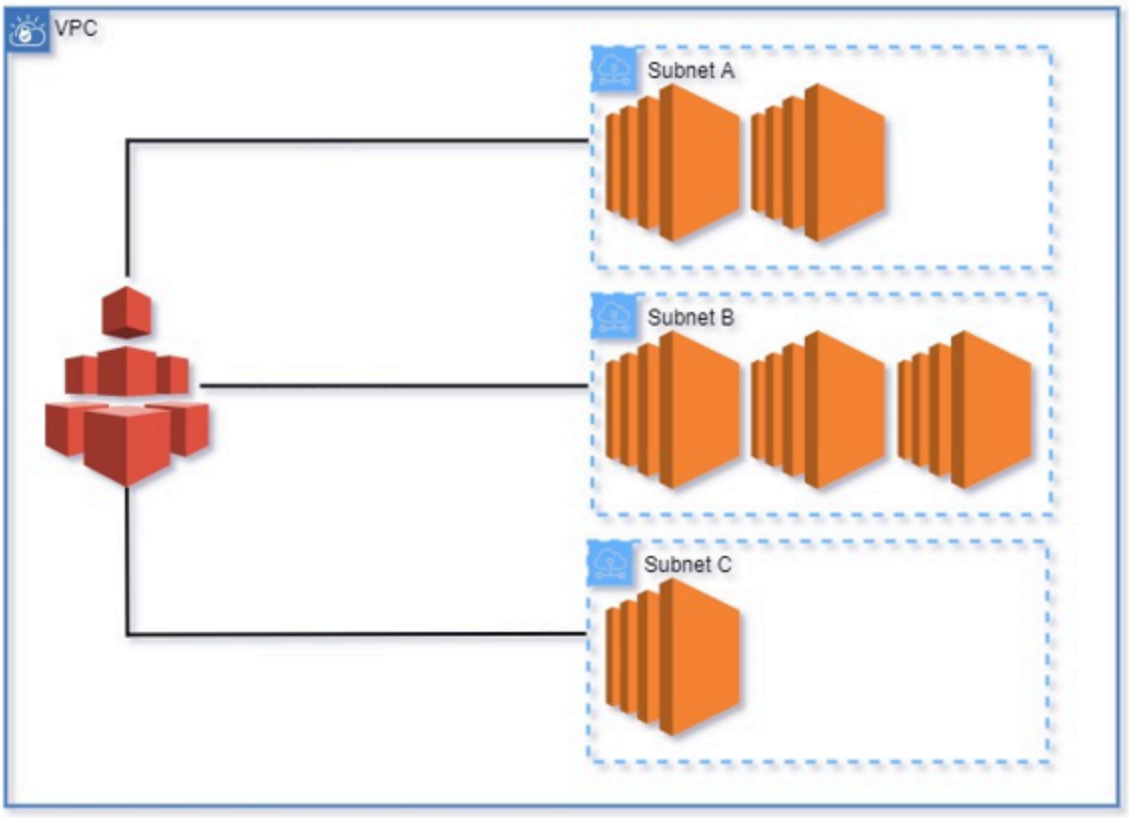
Rules are associated with an Outbound Endpoint to define how the DNS queries from your VPC will be forwarded.

- Conditional forwarding rules - forward DNS queries for specified domain names to DNS resolvers on your network. Queries to the specified domain name are forwarded to the defined Target IP Addresses
- System rules - handles the DNS queries for the specified domain names instead of the other DNS resolvers outside your VPC
- Recursive rule (Internet Resolver) - created automatically by the Resolver. Other domain names that are not defined explicitly on the rules go to this recursive resolver



Amazon Elastic File System (EFS)

Elastic File System (EFS) is a fully managed storage service from AWS that enables you to provision a simple and scalable file system that can be used with AWS Cloud services and on-premises resources. EFS can accommodate multiple connections all at the same time. For example, you can create an EFS on your VPC with various instances and mount it on the instances.



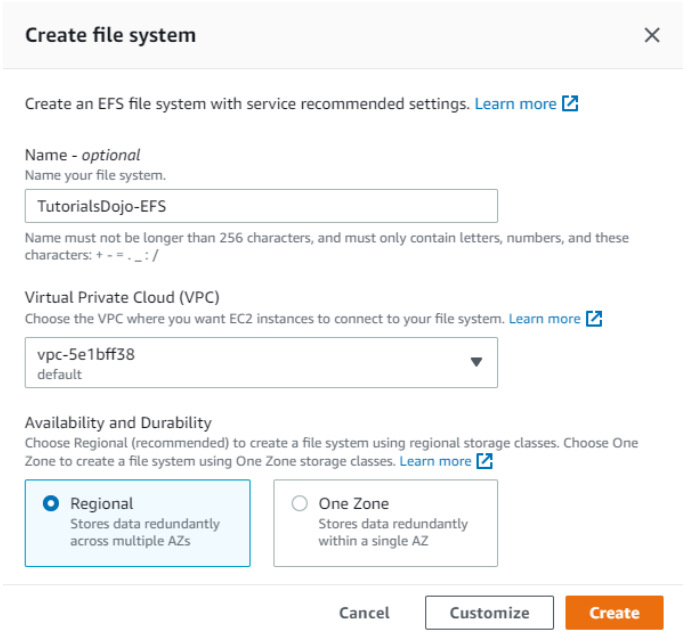
EFS Storage Classes

Like Amazon S3, Amazon EFS also has different storage classes to offer for different file system requirements. The EFS Classes differ on availability and durability.

- **EFS Standard** - frequently accessed data / multi-AZ storage
- **EFS Standard–Infrequent Access (IA)** - long term storage for infrequently accessed data / multi-AZ storage
- **EFS One Zone** - frequently accessed data / stored on single AZ
- **EFS One Zone-IA** - long term storage for infrequently accessed data / stored on single AZ

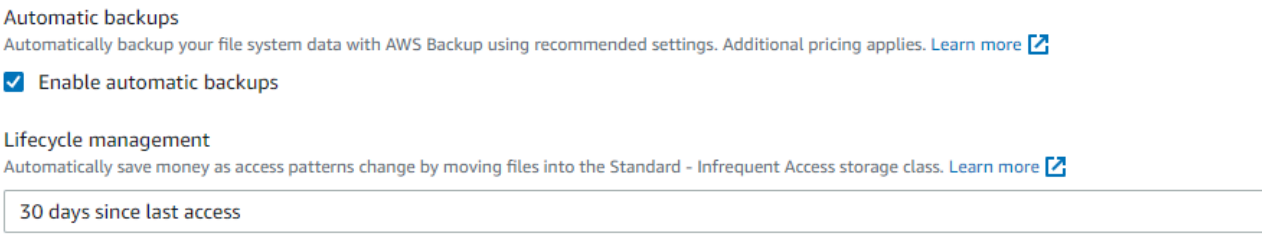
Creating a File System

To create an EFS, you need to specify a VPC and storage classes.

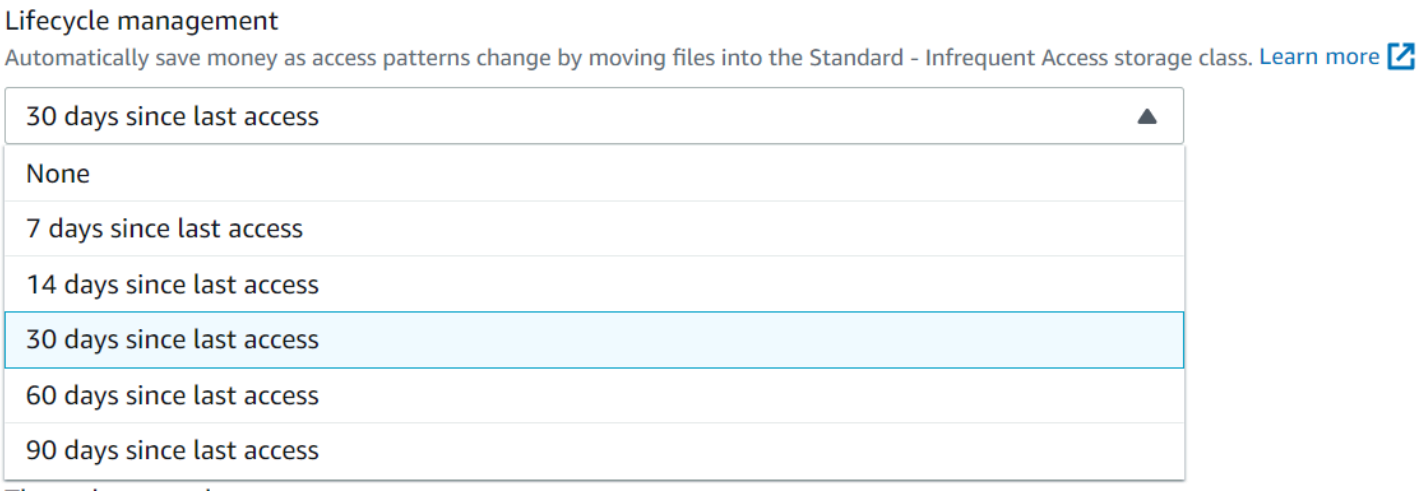


More settings can be configured when you click on the Customize button.

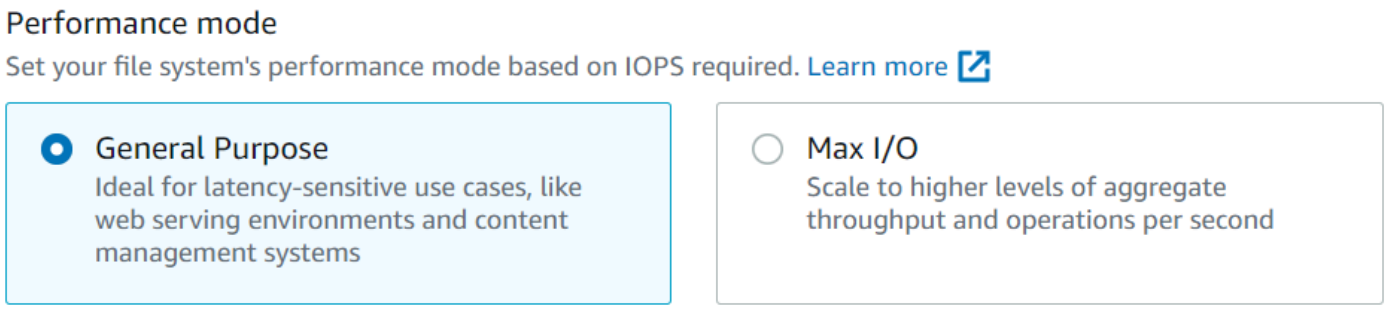
Enabling automatic backups allows you to backup your file system using AWS Backup.



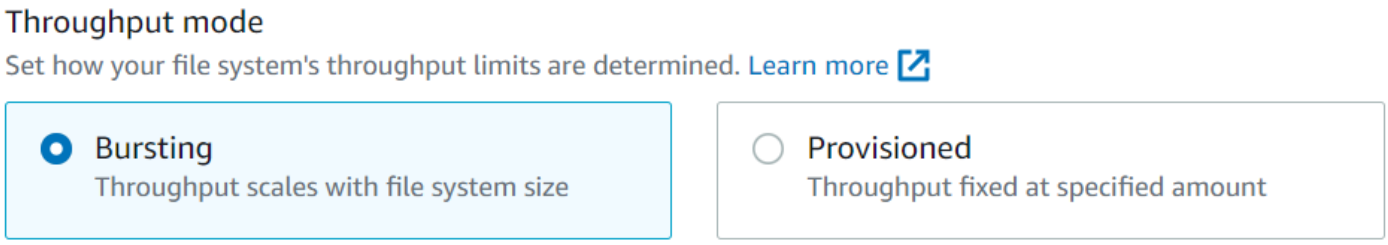
Like S3, EFS also has Lifecycle management that moves files to IA storage classes (Standard-IA/One Zone-IA) to save on costs effectively. The default policy set for lifecycle management is 30 days, but this can be changed.



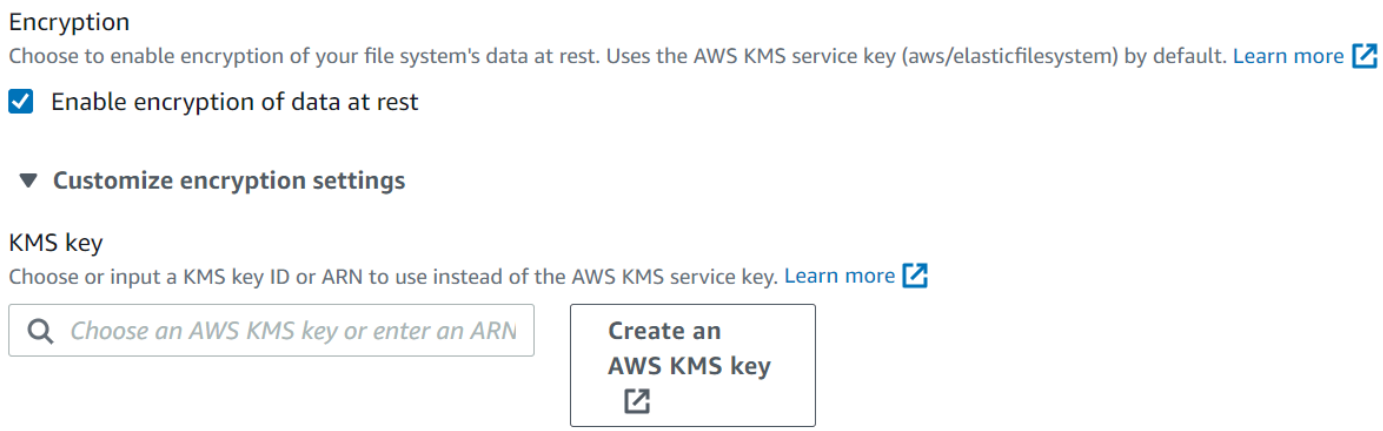
You can also set the performance mode of your file system to either **General Purpose** or **Max I/O** with no additional cost. You can select General Purpose mode for latency-sensitive use cases and Max I/O if you want a throughput performance that scales up. Max I/O performance mode is not available for file systems with the One Zone storage class. Performance mode can't also be changed once the file system is created.



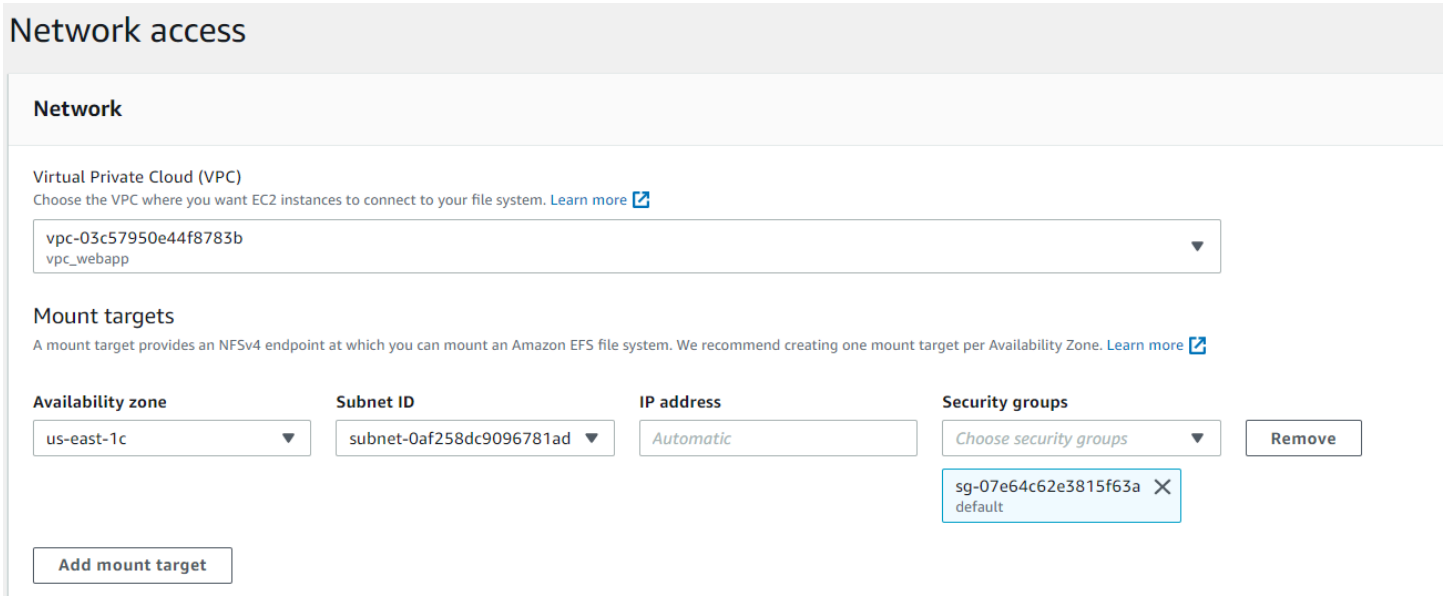
EFS also has two throughput modes: bursting for throughput scaling and provisioned for fixed throughput.



EFS uses AWS KMS service to encrypt the data at rest on the file system. You can also use a custom KMS key for the encryption.

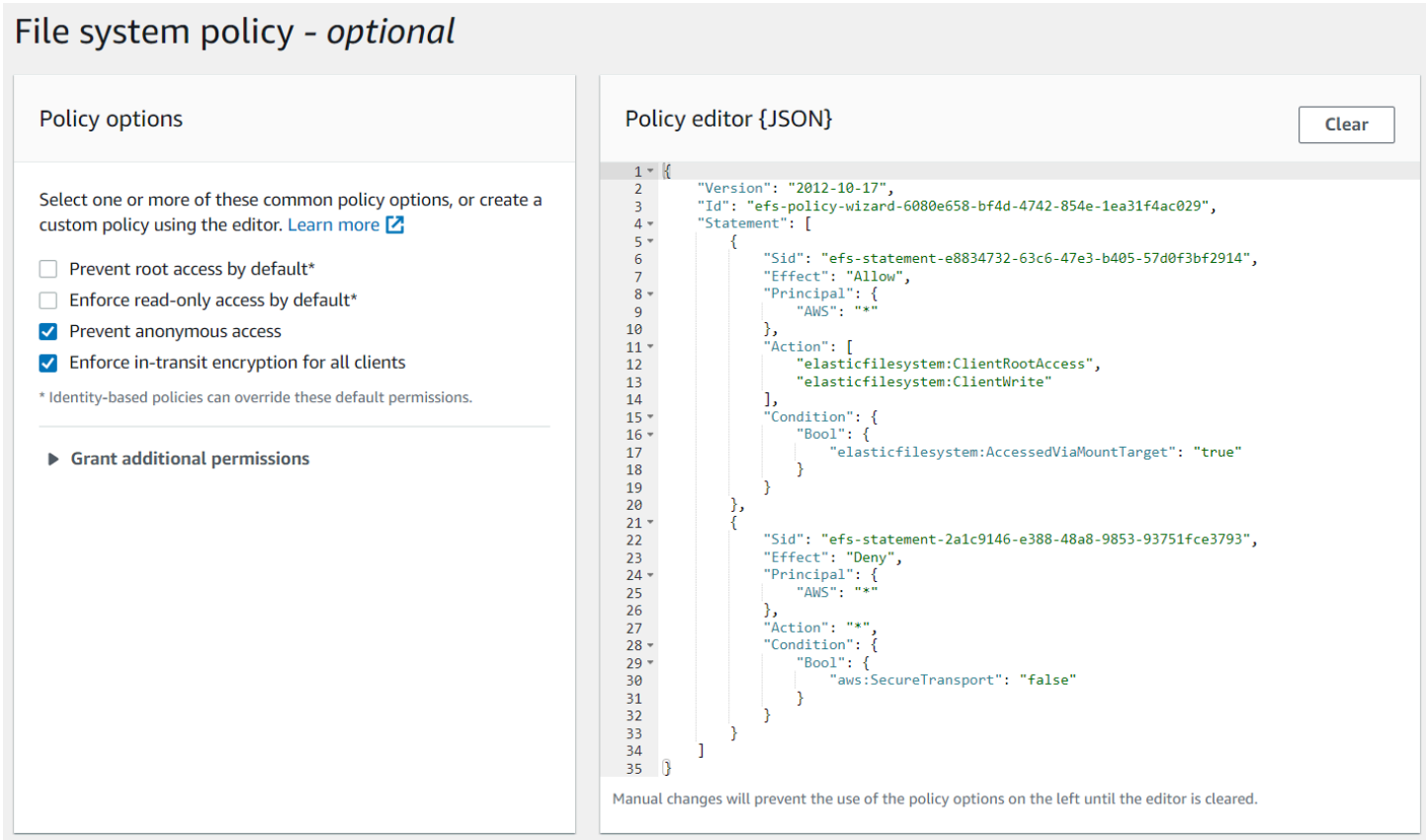


Depending on the availability you selected, your file system will have different mount targets per availability zone.

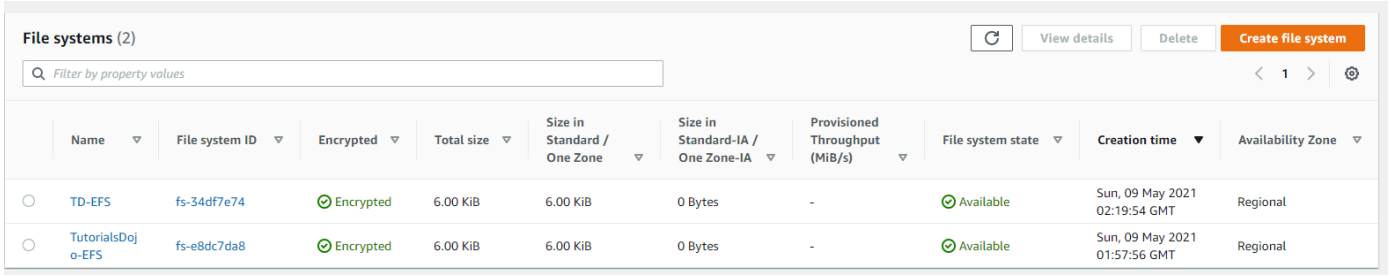


For additional control, EFS also has a file system policy as an optional feature. The following are the current policies available which can also be customized using the policy editor.

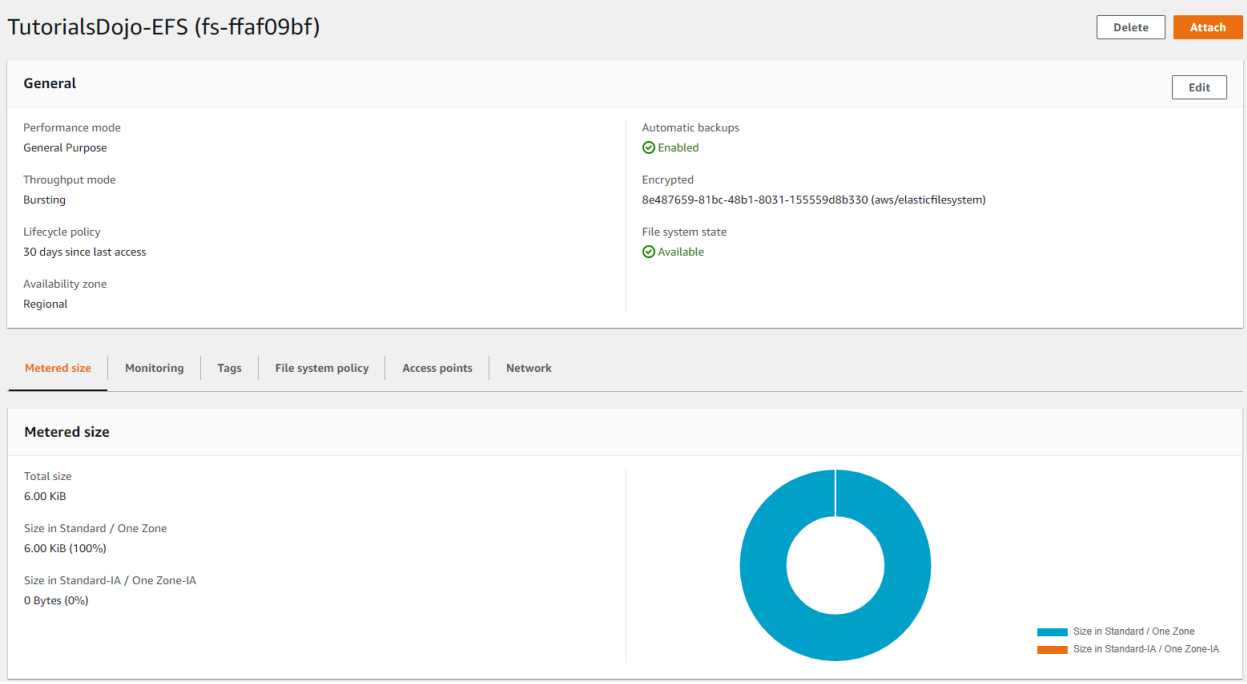
- Prevent root access by default
- Enforce read-only access by default
- Prevent anonymous access
- Enforce in-transit encryption for all clients



Once the file system is created, you can view a summary of its details with its management and monitoring options.

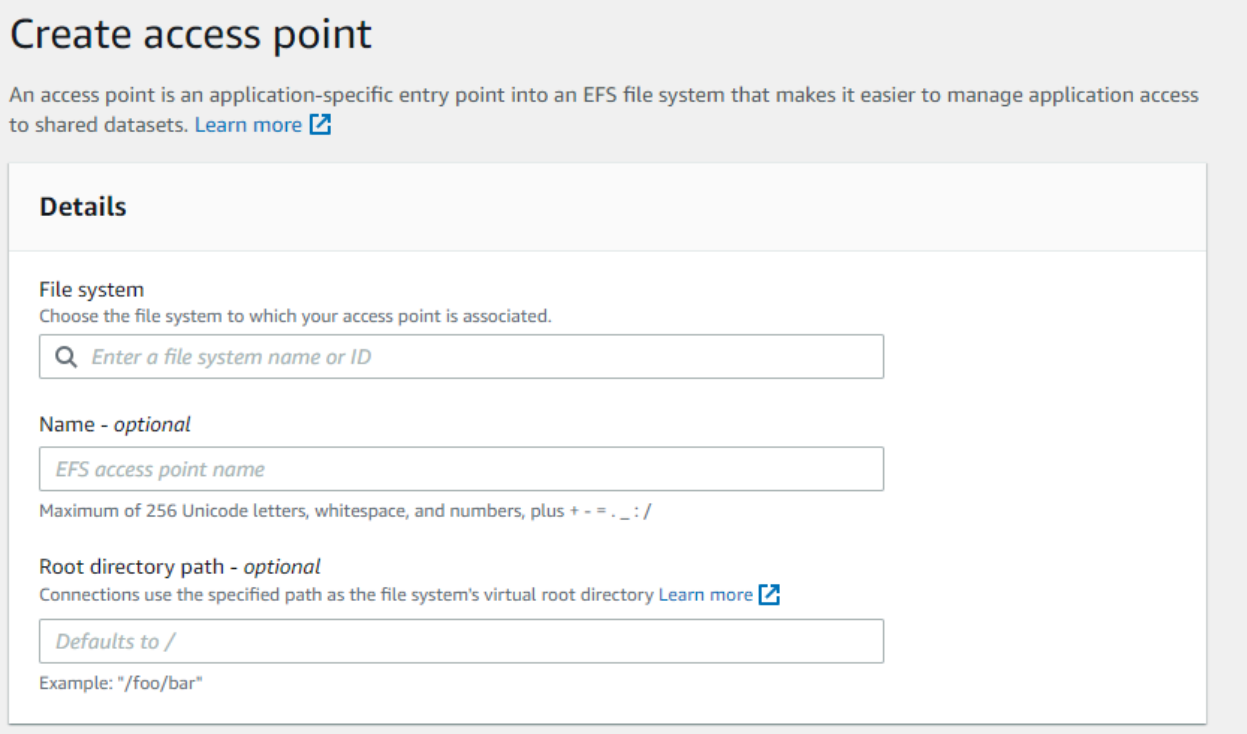


EFS automatically scales as you add or remove files. You only pay for the storage used by the file system.

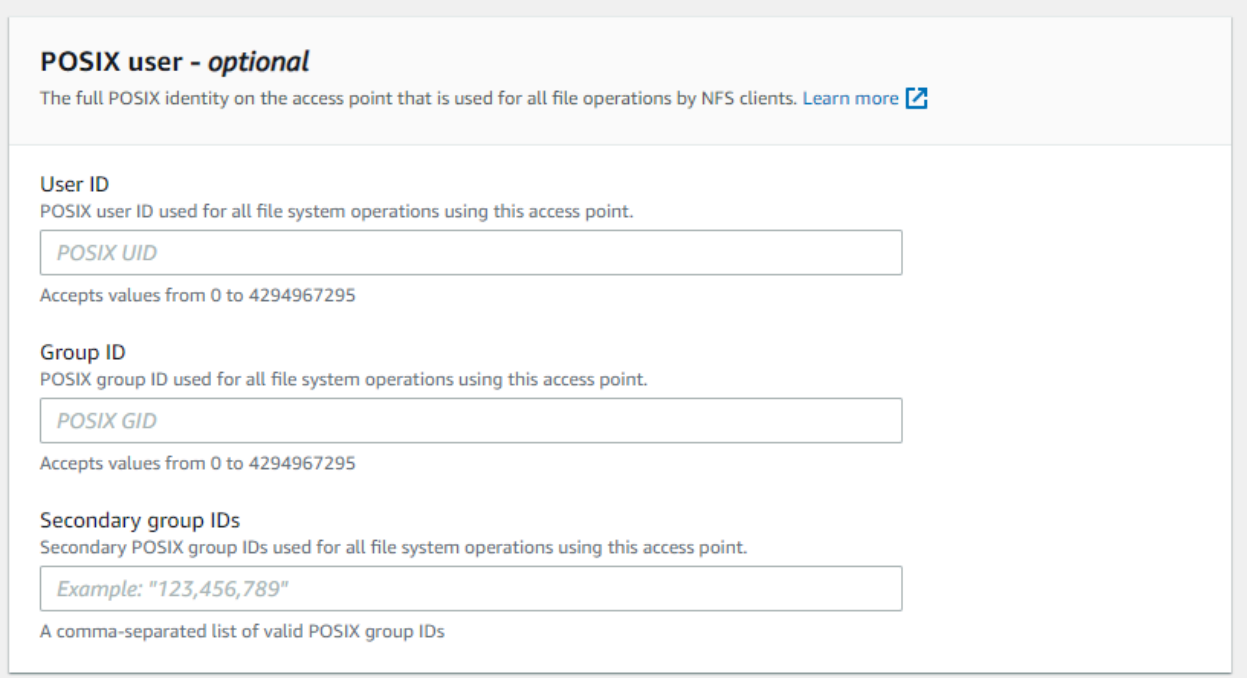


File System Access Point

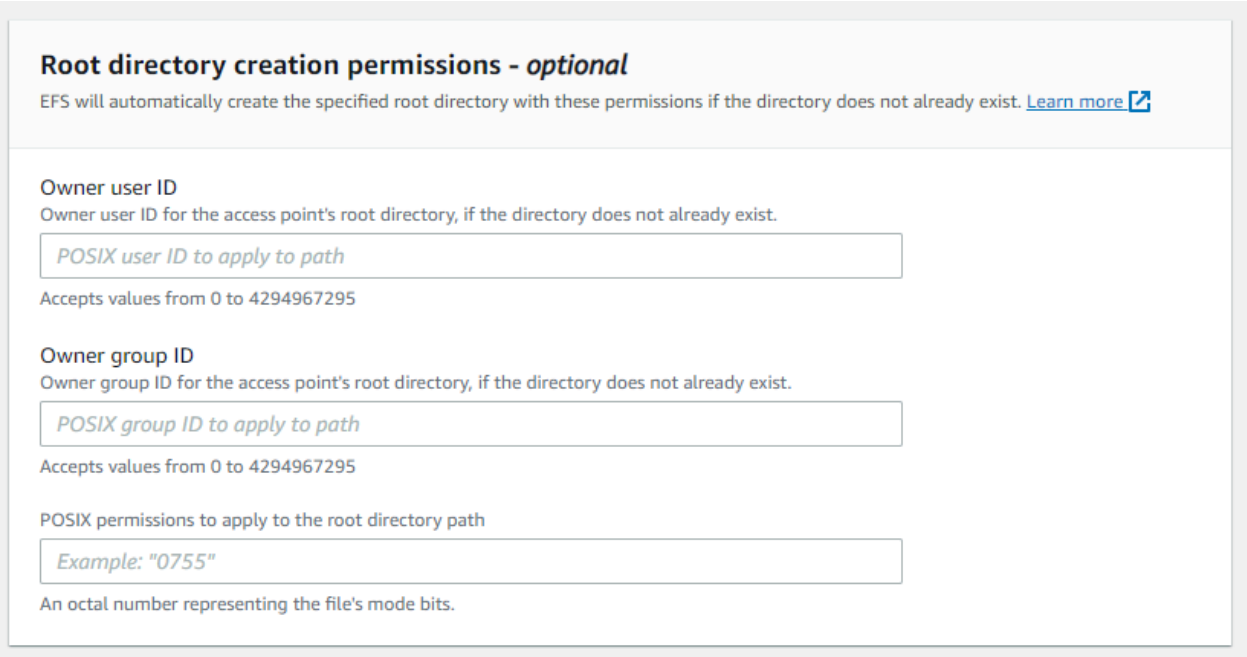
Access points are entry points for your applications to connect with your file system. Access points help you efficiently manage your application access. When creating an access point, you need to provide the File System Name/ID, Access Point Name *(Optional),* and Root directory path *(Optional).*



You also have an option to enable operating system (POSIX) identity for your access point by providing the User ID, Group ID, and Secondary group IDs.

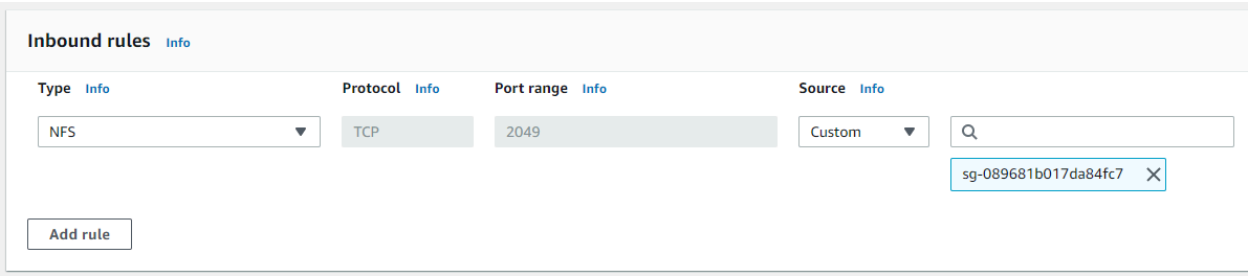


Furthermore, you can set permission for your file system root directory by providing the User ID, Group ID, and Permission (defined in octal number).



File System Security Group

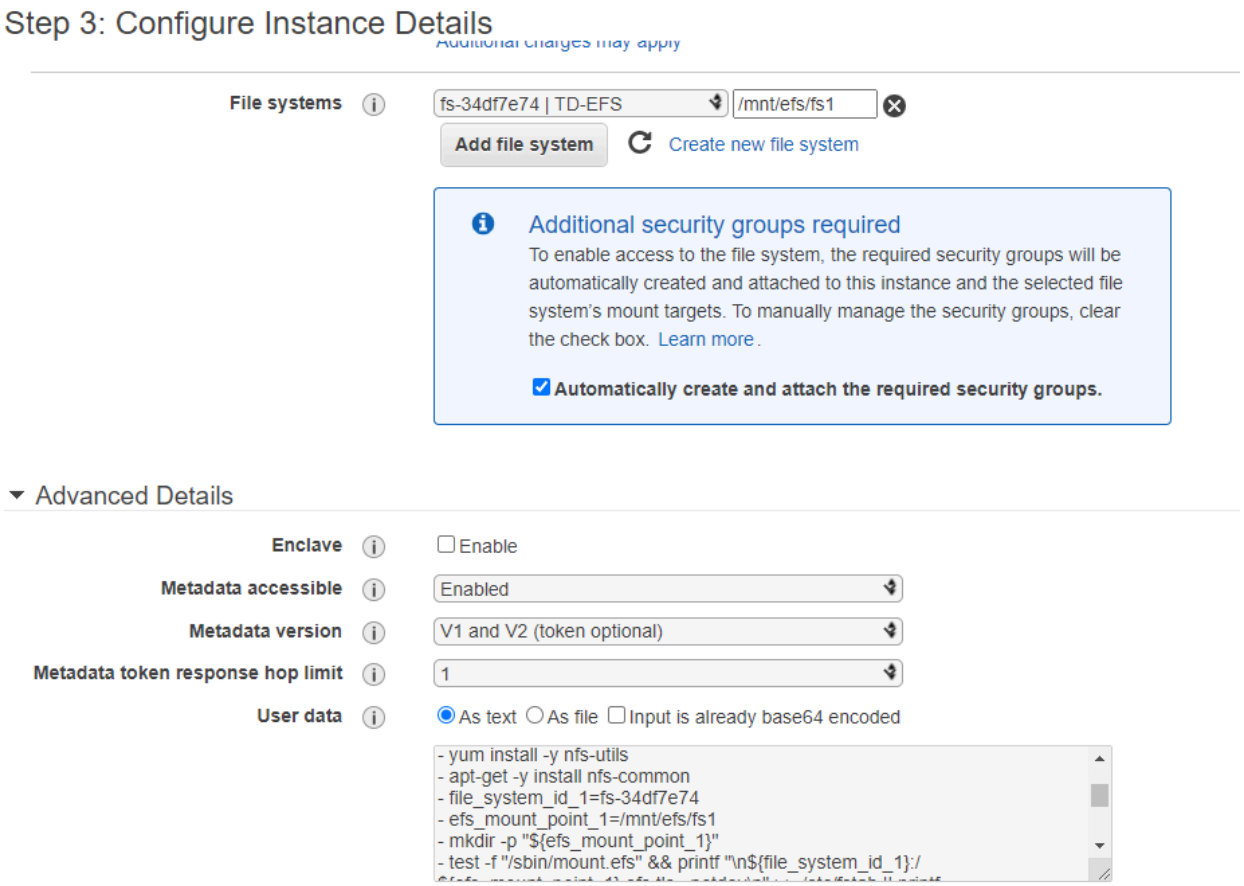
In addition to the security group for your Linux instance to allow incoming ssh traffic on port 22, you also need to create another firewall rule that allows NFS traffic to your instance. The source will depend on your requirement.



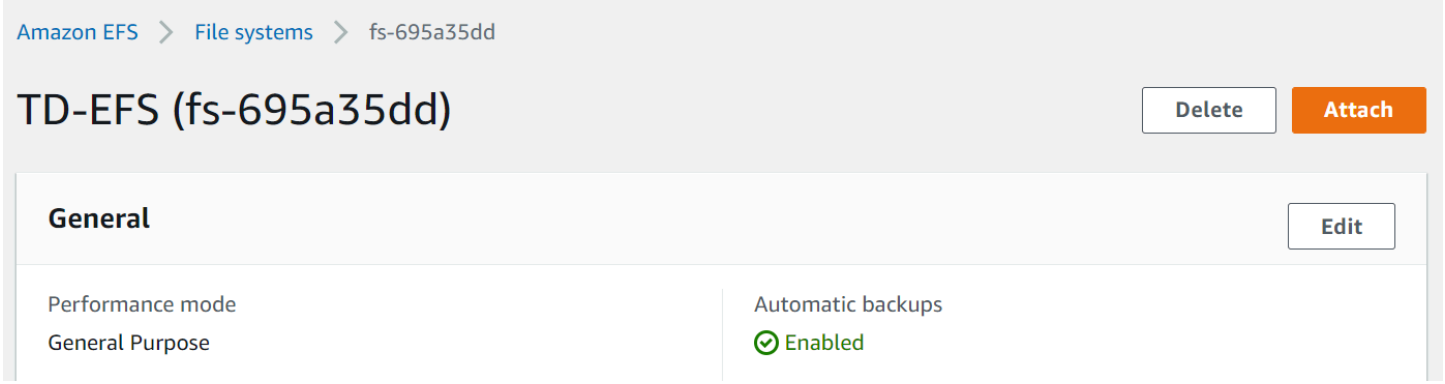
Mounting a File System

When launching an EC2 instance, you can select and mount an existing file system that you created. Notice that a script for the file system is automatically generated on the user data. If you haven't made a security

group for your file system, you have an option to allow AWS to create the required security group on your behalf. Please note that EFS only supports Linux instances.



Additionally, you can mount your file system on an existing Linux instance via DNS or IP address. EFS provides both commands when using the EFS mount helper and the NFS client. You can view these commands by selecting the file system then clicking the Attach button.



Mount via DNS

Mount your Amazon EFS file system on a Linux instance. Learn more

(•) Mount via DNS ( ) Mount via IP

Using the EFS mount helper:

```
sudo mount -t efs -o tls fs-695a35dd:/ efs
```

Using the NFS client:

```
sudo mount -t nfs4 -o nfsvers=4.1,rsize=1048576,wsize=1048576,hard,timeo=600,retrans=2,noresvport fs-695a35dd.efs.us-east-1.amazonaws.com:/ efs
```

Mount via IP

Mount your Amazon EFS file system on a Linux instance. Learn more

( ) Mount via DNS (•) Mount via IP

Availability zone

us-east-1a ▼

Using the NFS client:

```
sudo mount -t nfs4 -o nfsvers=4.1,rsize=1048576,wsize=1048576,hard,timeo=600,retrans=2,noresvport 172.31.26.238:/ efs
```

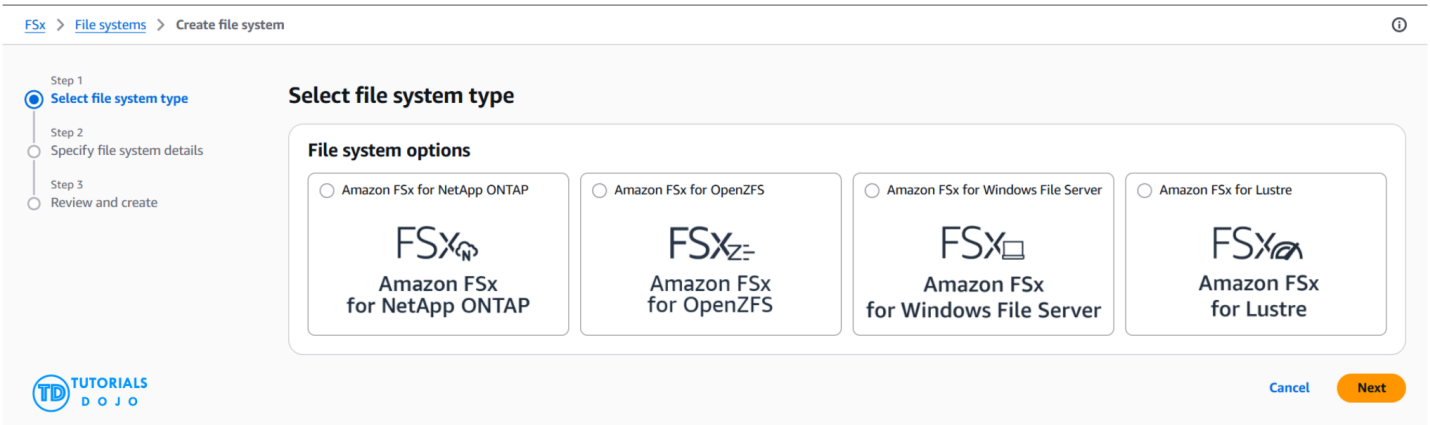
Reference:
https://docs.aws.amazon.com/efs/latest/ug/whatisefs.html

Amazon FSx

Amazon FSx is a simple and fully managed file system service. It offers a fast and reliable file system performance while maintaining durability and scalability for its data. Underlying hardware and software are AWS managed and are continuously monitored. Amazon FSx also replicates the data on multiple devices within an Availability Zone and supports different file systems optimized for specific workloads..

Amazon FSx supports the following file system types:

- Amazon FSx for Windows File Server
- Amazon FSx for Lustre
- Amazon FSx for NetApp ONTAP
- Amazon FSx for OpenZFS



Amazon FSx for Windows File Server

Amazon FSx provides a file system to Windows Operating System accessible through Server Message Block (SMB) Protocol. Like a native Windows file system, Amazon FSx for Windows also has management features like Active Directory integration and user quotas. The file system is also accessible outside AWS Environment through AWS Direct Connect or AWS VPN. Furthermore, it is also accessible by Linux and macOS devices.

File System Details

When setting up an Amazon FSx for Windows Server, the following details are provided.

Deployment Type

Amazon FSx for Windows Server is available on the following deployment options.

- **Single-AZ** - creates the Windows File Server on a single Availability Zone only.
- **Multi-AZ** - creates the Windows File Server across multiple Availability Zones.

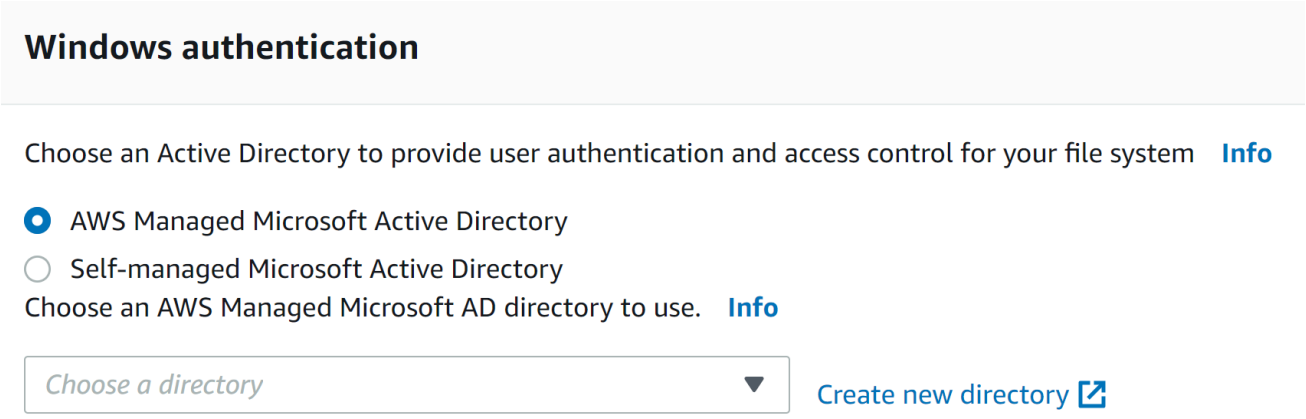
Storage Capacity and Storage Type
The minimum and maximum storage capacity vary depending on the storage type. To obtain much higher storage and throughput, you create multiple file systems and combine them using Microsoft Distributed File System (DFS). The following storage type is supported.

- **Solid State Drives (SSD)** - high performance and low latency workloads.
- **Hard Disk Drives (HDD)** - general workloads, user file sharing

Throughput Capacity
The data transfer rate can be configured on throughput capacity. You can select the AWS recommended throughput capacity (based on storage capacity) or specify the throughput capacity yourself. Throughput bursts for periods of time are also supported.

Virtual Private Cloud
The VPC is where your file system resides. The file system is also accessible from other VPC through peering or Transit Gateway and is accessible outside the AWS environment through AWS Direct Connect or AWS VPN. For Multi-AZ deployment, two subnet/availability zones are required for preferred and standby file systems. Amazon FSx creates an Elastic Network Interface for the instances to connect to the file system.

Windows Authentication
Amazon FSx for Windows Server integrates with Microsoft Active Directory for authentication and folder/file access control. You can select an AWS Managed Microsoft Active Directory or a Self-managed AD if you have it running in an Amazon EC2 instance.



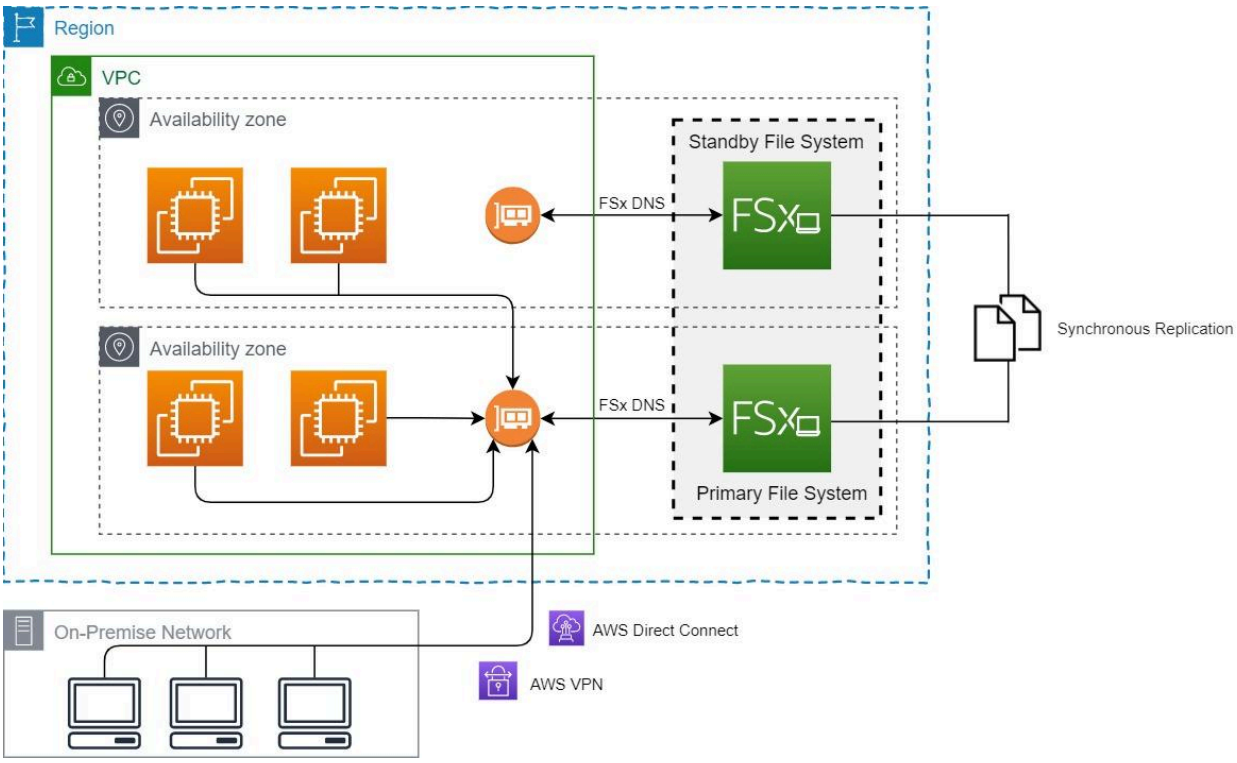
Encryption
Amazon FSx integrates with AWS KMS for encryption. You can use the default *aws/fsx* key or provide another key ARN.

Optional Details

- **Auditing** - allow audit logging for files and folders. Log destination can be CloudWatch logs or Kinesis Data Firehouse
- **Access** - define additional DNS names for the file system
- **Backup and Maintenance** - configure the backup schedule and maintenance window
- **Tags** - Metadata/Label for AWS resources

Working with Amazon FSx for Windows File Server

The diagram below shows an Amazon FSx for Windows File Server with Multi-AZ deployment. The file system is created on both preferred and standby availability zones. Any changes on the primary file system are synchronously replicated to the standby file system. The instances access the primary file system through a network interface. In cases when the primary file system becomes unavailable, the Traffic is automatically migrated to the standby file system. The file system is also accessible from an on-premises network using AWS Direct Connect or AWS VPN.



Amazon FSx for Lustre

Lustre is an open-source, parallel file system used for High-Performance Computing (HPC). Through Amazon FSx, you can create a fast, scalable, and cost-effective file system capable of Lustre workloads. Amazon FSx for Lustre offers capabilities like sub-millisecond latencies, hundreds of gigabytes per second of throughput, and millions of IOPS. Amazon FSx for Lustre is POSIX-compliant and can be used for Linux applications. To mount Amazon FSx for Lustre, Linux instances should have Lustre Client installed.

File System Details

When setting up an Amazon FSx for Lustre, the following details are provided.

Deployment and Storage Type

Persistent deployment options have automatic data replication and file server failover capabilities, which is highly recommended for long-term data processing. On the other hand, the scratch deployment option is ideal for temporary, non-critical, and short-term processing. Storage also comes on SSD and HDD storage. The available deployment options and storage types are as follows.

- **Persistent, SSD** - for long term latency-sensitive workloads that require high IOPS/throughput
- **Persistent, HDD** - for long-term workloads that require high throughput but are not latency-sensitive. Has an optional SSD cache feature.
- **Scratch, SSD** - for short-term workloads and non-critical data

Storage Capacity and Throughput Capacity

The data transfer rate can be configured on throughput capacity. The total throughput capacity depends on the storage capacity. The total throughput increases as the storage capacity increases.

- 50 MB/s/TiB (up to 1.3 GB/s/TiB burst)
- 100 MB/s/TiB (up to 1.3 GB/s/TiB burst)
- 200 MB/s/TiB (up to 1.3 GB/s/TiB burst)

*Throughput capacity = Storage capacity (TiB) * Per unit storage throughput (MB/s)*

Data Compression Type

Amazon FSx for Lustre uses the LZ4 algorithm for data compression. All new files are compressed before being written on the file system and are decompressed when being read. Because of the tremendous LZ4 algorithm, the compression and decompression process has zero to minimal effects on the file system performance.

Virtual Private Cloud

The VPC is where your file system resides. The file system is also accessible from other VPC through peering or Transit Gateway and is accessible outside the AWS environment through AWS Direct Connect or AWS VPN. Amazon FSx for Lustre doesn't come with Multi-AZ deployment, but the data is replicated within the Availability

Zone for persistent deployment. Amazon FSx creates an Elastic Network Interface for the instances to connect to the file system.

Encryption

Amazon FSx integrates with AWS for encryption. You can use the default *aws/fsx* key or provide another key ARN.

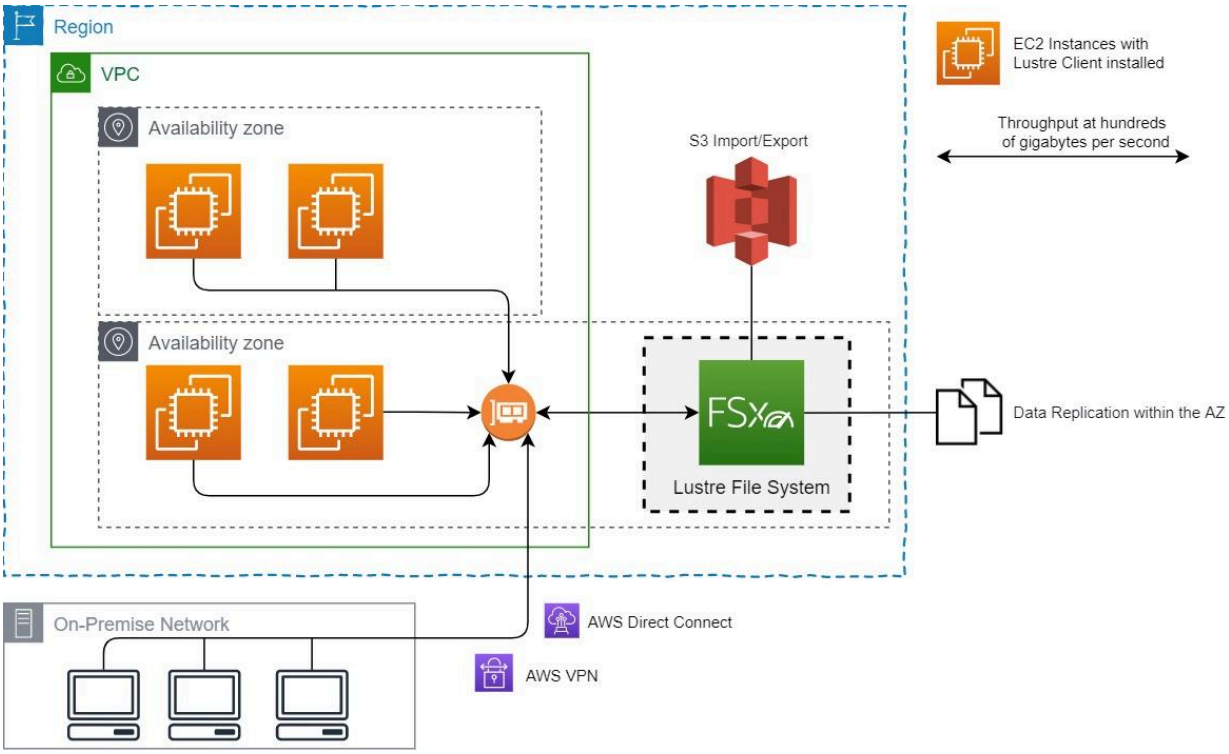
Optional Details

- **Data Repository Import/Export** - enable import data from S3/export data to S3
- **Backup and Maintenance** - configure the backup schedule and maintenance window
- **Tags** - Metadata/Label for AWS resources

Working with Amazon FSx for Lustre

The diagram below shows an Amazon FSx for Lustre in persistent deployment. The file system is being accessed by multiple instances across two availability zones through a network interface. Data is transferred from and to the file system at a throughput rate of hundreds of gigabytes per second. It also shows S3 integration for exporting/importing files in an S3 bucket.

The file system is also accessible from an on-premises network using AWS Direct Connect or AWS VPN.



Amazon FSx for NetApp ONTAP

Amazon FSx for NetApp ONTAP provides a fully managed shared storage service built on NetApp ONTAP technology. It supports multiprotocol access and advanced enterprise storage capabilities such as snapshots, cloning, deduplication, compression, and replication.

Amazon FSx for NetApp ONTAP supports Network File System (NFS), Server Message Block (SMB), Internet Small Computer Systems Interface (iSCSI), and Non-Volatile Memory Express over TCP (NVMe), making it suitable for Linux, Windows, and high-performance block storage workloads.

File System Details

When setting up an Amazon FSx for NetApp ONTAP file system, the following details are provided.

Deployment and Storage Type

Amazon FSx for NetApp ONTAP supports the following deployment options.

- **Single-AZ** - deploys the ONTAP file system within a single Availability Zone.
- **Multi-AZ** - deploys highly available ONTAP file servers across multiple Availability Zones with automatic failover support.

Storage Capacity and Throughput Capacity

Amazon FSx for NetApp ONTAP uses SSD storage for primary storage workloads. Storage can also automatically tier infrequently accessed data to capacity pool storage for cost optimization.

The storage layers are:

- **SSD Storage** - high-performance primary storage for active workloads.
- **Capacity Pool Storage** - low-cost elastic storage tier for cold data.

Throughput capacity defines the maximum network throughput available to the file system. Higher throughput configurations are recommended for workloads with intensive read/write operations. The throughput capacity is independent of storage capacity and can be scaled as needed.

Storage Virtual Machines (SVM)

Amazon FSx for NetApp ONTAP uses Storage Virtual Machines (SVMs) to isolate workloads and manage storage resources. Each SVM can support different protocols and authentication configurations.

Volume Configuration

Volumes are logical storage containers created inside an SVM. Each volume can be configured independently with:

- Security styles
- Storage efficiency settings
- Snapshot policies
- Tiering policies

Data Efficiency Features

Amazon FSx for NetApp ONTAP supports advanced storage optimization capabilities such as:

- Deduplication
- Compression
- Compaction
- Thin provisioning

These features help reduce storage consumption and cost.

Snapshots and Cloning

Amazon FSx for NetApp ONTAP supports near-instant point-in-time snapshots and FlexClone technology for rapid cloning of datasets without requiring full copies.

Protocol Support

Amazon FSx for NetApp ONTAP supports the following protocols.

- Network File System (NFS)
- Server Message Block (SMB)
- iSCSI

This allows Linux and Windows applications to access the same shared storage.

Virtual Private Cloud

The VPC is where your file system resides. The file system is also accessible from other VPCs through peering or Transit Gateway and is accessible outside the AWS environment through AWS Direct Connect or AWS VPN. Amazon FSx creates Elastic Network Interfaces for client connectivity.

Encryption

Amazon FSx integrates with AWS for encryption. You can use the default *aws/fsx* key or provide another key ARN.

Optional Details

- Backup and Maintenance - configure backup schedules and maintenance windows
- Data Tiering - automatically tier cold data to capacity pool storage
- Snapshots - configure automatic snapshot schedules
- Replication - enable SnapMirror replication
- Tags - Metadata/Label for AWS resources

Amazon FSx for OpenZFS

Amazon FSx for OpenZFS provides fully managed shared file storage powered by the OpenZFS file system. It is optimized for Linux-based workloads requiring low-latency performance and high throughput.

Amazon FSx for OpenZFS supports NFS protocols and offers advanced OpenZFS capabilities such as snapshots, cloning, compression, and data integrity validation.

File System Details

When setting up an Amazon FSx for OpenZFS file system, the following details are provided.

Deployment and Storage Type

Amazon FSx for OpenZFS supports the following deployment options.

- **Multi-AZ (HA)** - deploys highly available file servers across multiple Availability Zones with automatic failover capabilities.
- **Single-AZ (HA)** - deploys a primary and standby file server within a single Availability Zone. This deployment type provides high availability through automatic failover and failback within the same AZ.
- **Single-AZ (non-HA)** - deploys a single file server within one Availability Zone without standby failover capability. Amazon FSx automatically detects and replaces failed infrastructure components, but recovery events and maintenance operations may result in temporary downtime.

Storage Capacity and Throughput Capacity

Amazon FSx for OpenZFS uses SSD-based storage for low-latency workloads and intelligent caching.

The storage components include:

- **SSD storage** – high-performance storage for active datasets
- **NVMe cache devices** – accelerate read and write operations through intelligent caching

Throughput capacity controls the amount of network bandwidth available to the file system. Higher throughput configurations improve performance for concurrent workloads and large data transfers.

File Systems and Volumes

Amazon FSx for OpenZFS organizes storage into file systems and datasets. Each dataset can have independent settings such as:

- Compression
- Quotas
- Record size
- Snapshot policies

Data Compression

Amazon FSx for OpenZFS supports transparent inline compression to reduce storage consumption while maintaining performance.

Snapshots and Cloning

Amazon FSx for OpenZFS supports lightweight snapshots and near-instant writable clones for rapid testing and recovery operations.

Data Integrity

OpenZFS continuously validates data integrity using checksums and automatically repairs corrupted data blocks whenever possible.

Protocol Support

Amazon FSx for OpenZFS supports the following protocols.

- **Network File System (NFS) v3**
- **Network File System (NFS) v4**

Virtual Private Cloud

The VPC is where your file system resides. The file system is also accessible from other VPCs through peering or Transit Gateway and is accessible outside the AWS environment through AWS Direct Connect or AWS VPN. Amazon FSx creates Elastic Network Interfaces for client connectivity.

Encryption

Amazon FSx integrates with AWS for encryption. You can use the default *aws/fsx* key or provide another key ARN.

Optional Details

- Backup and Maintenance - configure backup schedules and maintenance windows
- Snapshots - configure automatic snapshot schedules
- Tags - Metadata/Label for AWS resources

References:
https://docs.aws.amazon.com/fsx/latest/WindowsGuide/what-is.html
https://docs.aws.amazon.com/fsx/latest/LustreGuide/what-is.html
https://docs.aws.amazon.com/fsx/latest/ONTAPGuide/what-is-fsx-ontap.html

AWS DataSync

Moving data from on-premises to cloud for data is a bit of a challenge. It could be for a Disaster Recovery (DR) requirement, data archiving, data migration, or any business requirement. AWS DataSync simplifies the process of moving data from on-premises to AWS Storage services and between AWS storage services. AWS DataSync, as the name implies, syncs the data between the data source and its destination. It is an online data transfer service that supports different file-sharing protocols like Network File System (NFS) and Server Message Block Protocol (SMB).

Supported AWS Storage Service

- Amazon S3
- Amazon EFS
- Amazon FSx for Windows File Server

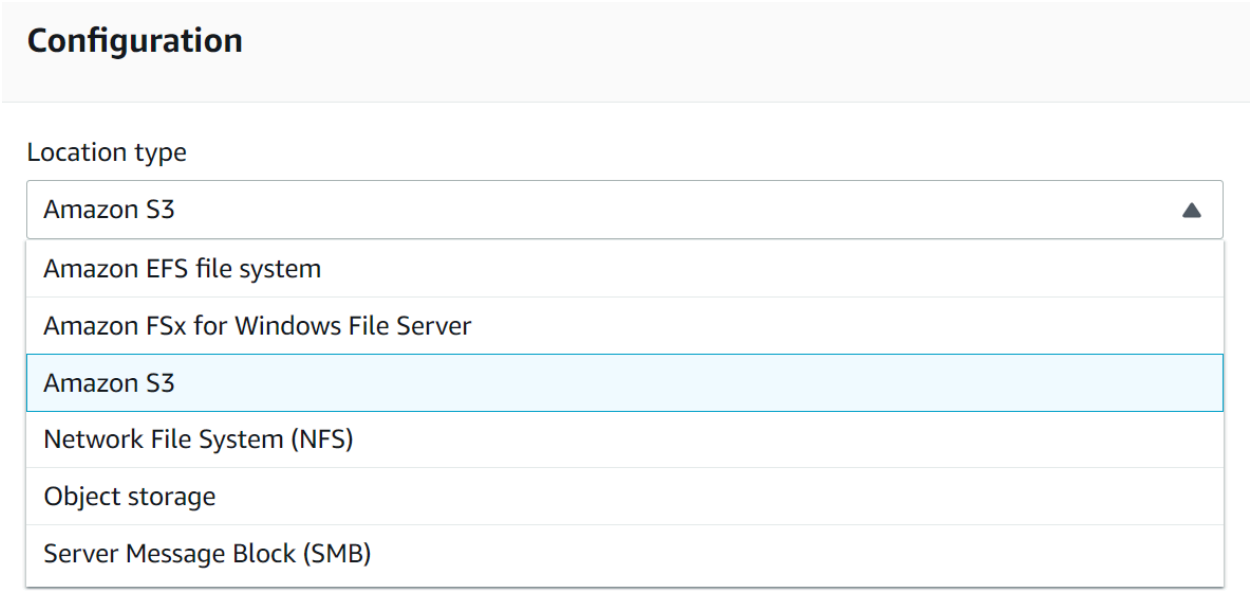
Working with DataSync

A **DataSync Agent** running on the on-premises environment is required when transferring data to AWS storage service. The agent is not required when transferring data between AWS storage services. A DataSync Agent is a virtual machine that is deployed on the following hypervisors and in Amazon EC2.

- VMware ESxi
- Microsoft Hyper-V
- Kernel-based Virtual Machine (KVM)
- AWS EC2

A **Service endpoint** and **Activation key** is also required. A service endpoint defines where your agent will connect while the activation key associates the agent with your AWS account.

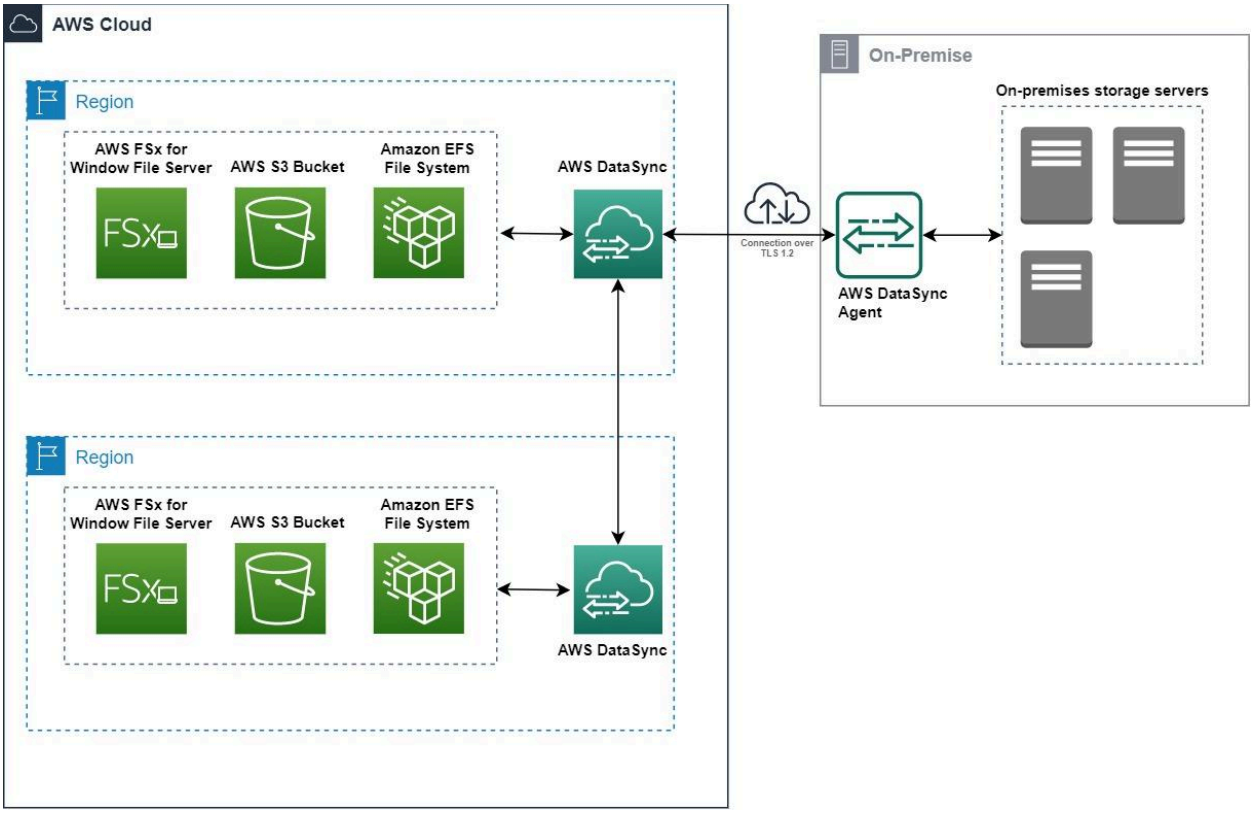
A **Task** is where you configure the source and destination by specifying the location type and its region. You also select the agent here if transferring data from on-premises. You also need to define an IAM role for both the source and destination; you can have AWS auto generate the required IAM role.



Moreover, the following settings are configurable for a more refined data transfer task.

- **Data Verification** - AWS DataSync does a data integrity check during the transfer process.
- **Bandwidth limit** - sets the maximum bandwidth (MiB/s) for the task
- **Transfer Mode** - choose whether to copy all the data or only the data that has been changed. Set the behavior for the destination when a deleted file from the source and an existing file is detected.
- **Filter** - specify a folder or file you want to exclude
- **Schedule** - defines the time and frequency of the task
- **Logging** - enable logging of information to a CloudWatch log group

The diagram below shows a typical setup when transferring data from an on-premises storage device to any AWS storage service. The DataSync agent deployed on the on-premises environment enables the transfer process as the data travels to the AWS environment over the Internet using Transport Layer Security (TLS) 1.2 protocol. The diagram also shows the data transfer process between AWS storage (e.g., S3 to EFS, FSx to S3).


AWS Cloud
Region
AWS FSx for Window File Server
AWS S3 Bucket
Amazon EFS File System
AWS DataSync
Region
AWS FSx for Window File Server
AWS S3 Bucket
Amazon EFS File System
AWS DataSync
Connection over TLS 1.2
On-Premise
On-premises storage servers
AWS DataSync Agent

AWS Backup

AWS Backup is a fully managed backup service that enables you to manage and automate the backup of your AWS resources in one place easily. You can do this by creating a backup plan, assigning an AWS Service to that backup plan, and having it monitored. The AWS Backup service allows you to view all your backup jobs on a single dashboard with details like backup and restoration job status for centralized monitoring.

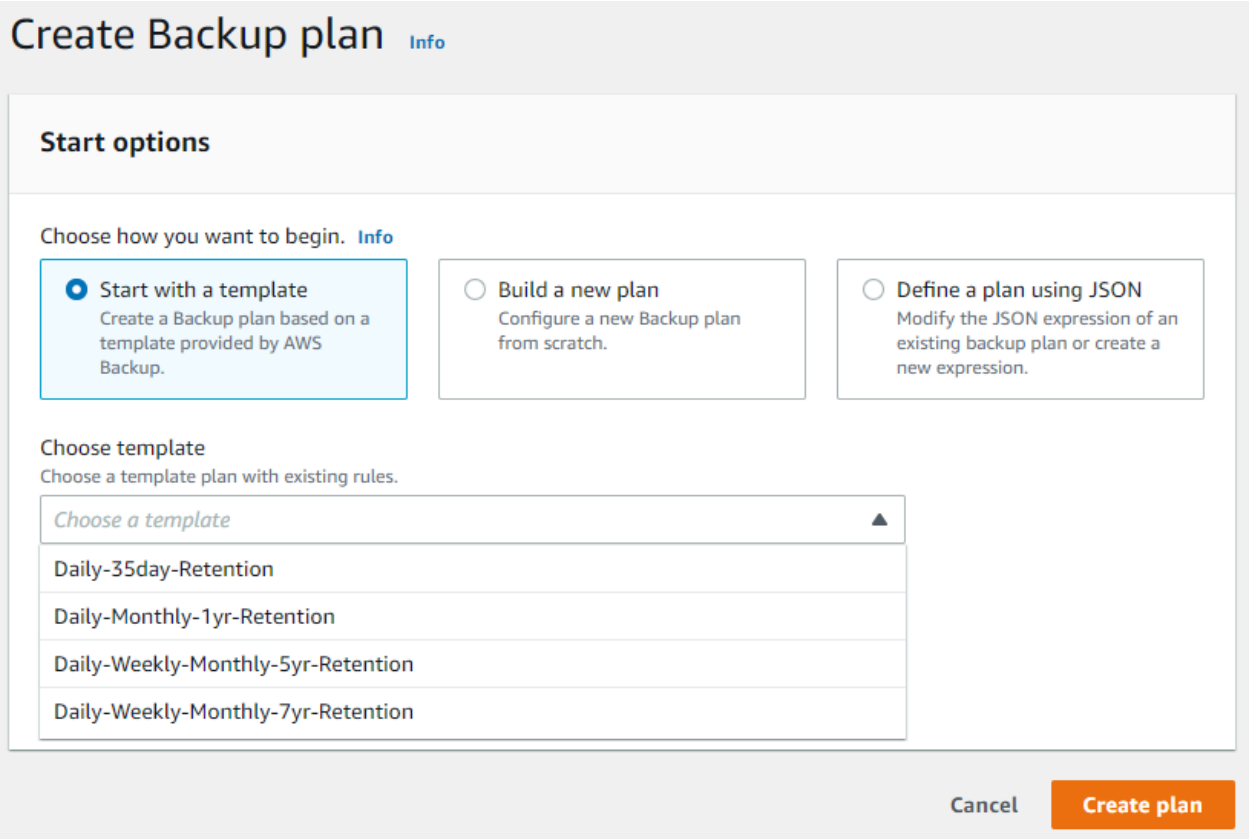
Supported AWS Resources

- Aurora
- DynamoDB
- EBS
- EC2
- EFS
- FSx
- RDS
- Storage Gateway

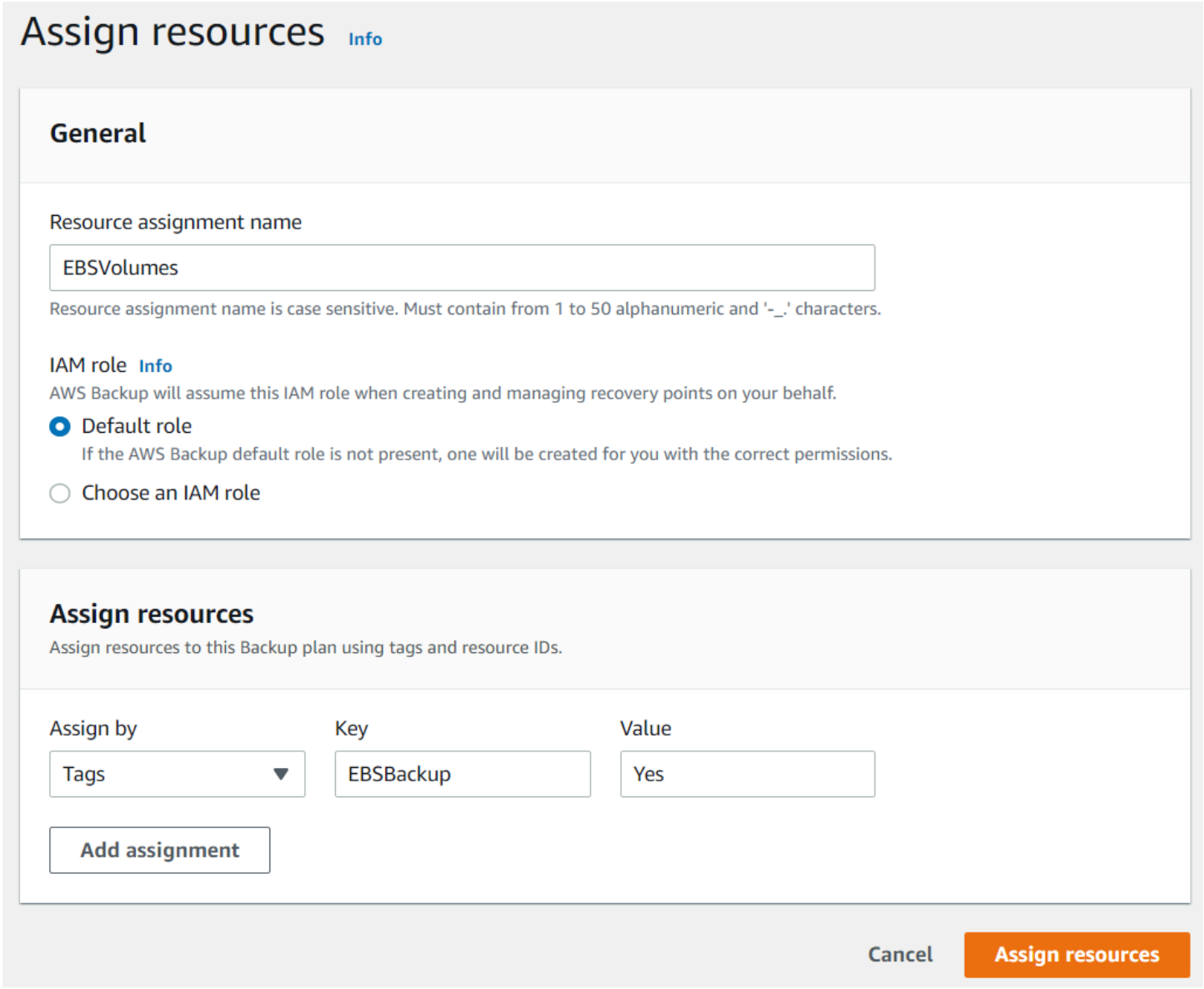
You can also back up your Windows VSS-supported applications on Amazon EC2 using AWS Backup.

Backup Plan

This is where you define your backup name and backup rules. The backup rules indicate the schedule, frequency, and lifecycle rules of your backup. You can create a backup plan by either using a template, creating a new plan, or using JSON. Once a backup plan is created, you can then assign AWS resources to it. That AWS resources will be backed up according to its backup plan.

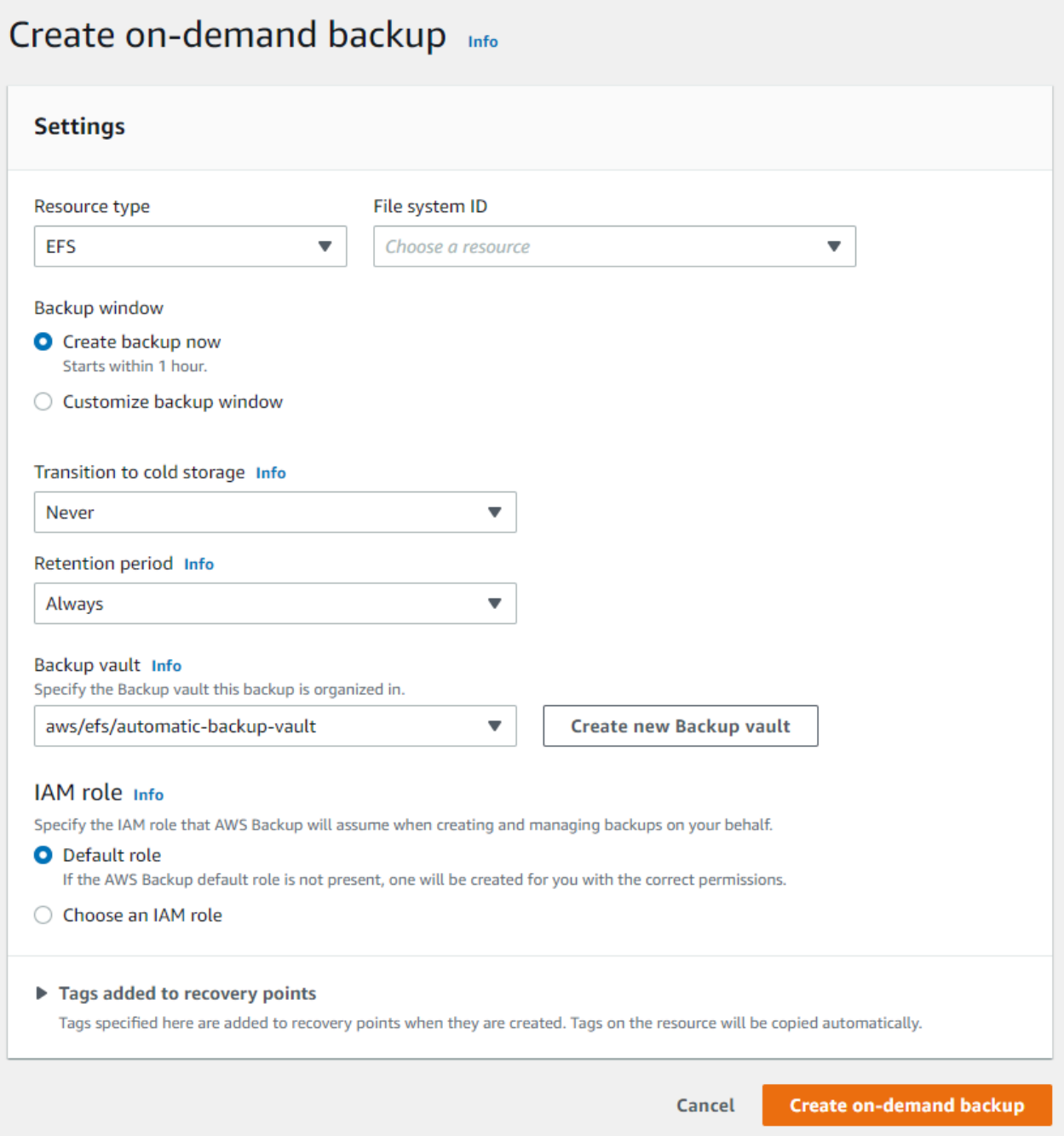


AWS Backup uses tagging to identify the resources that you want to include in your backup plan. Here in this example, all resources with the tag *EBSBackup:Yes* are included on this backup plan.



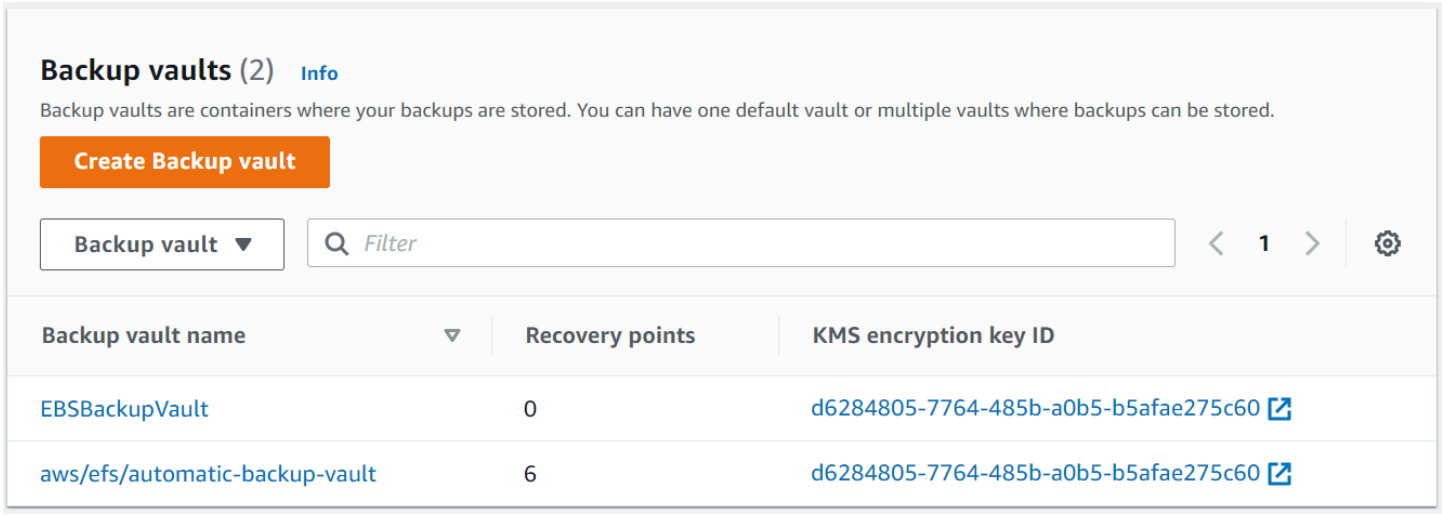
On-demand Backup

If you want to create a backup of your AWS resource immediately, you can do so through On-demand Backup. You just need to select the AWS resource to backup and define its lifecycle and retention rules.



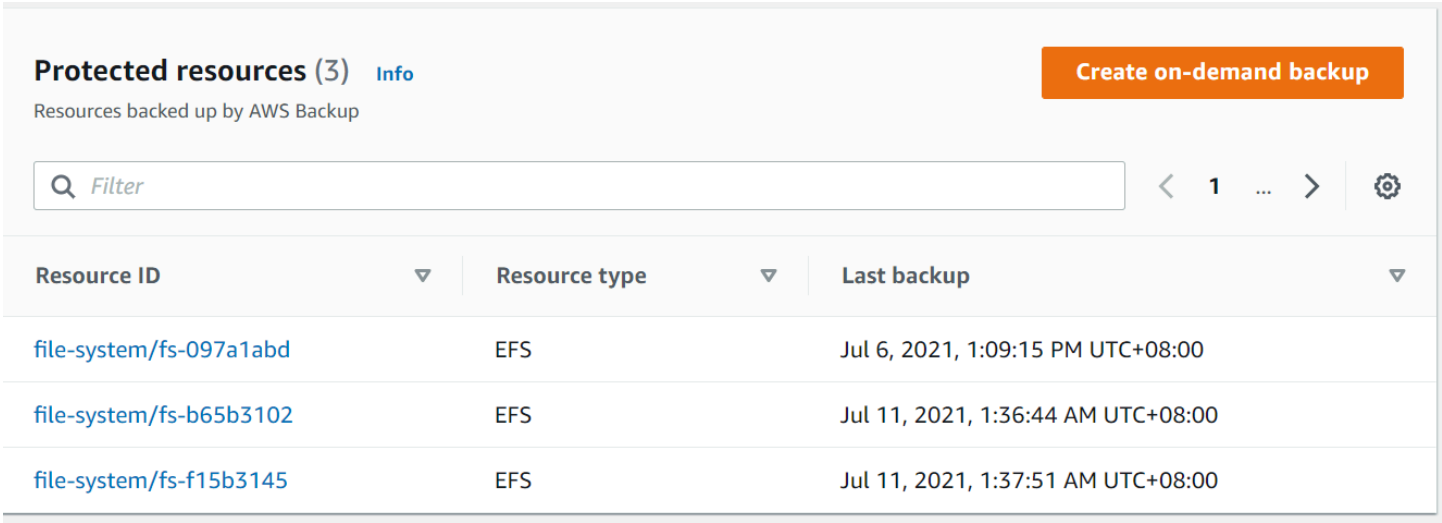
Backup Vault

Backup Vault is where your backups are stored. It uses KMS for encryption to further secure the backups.



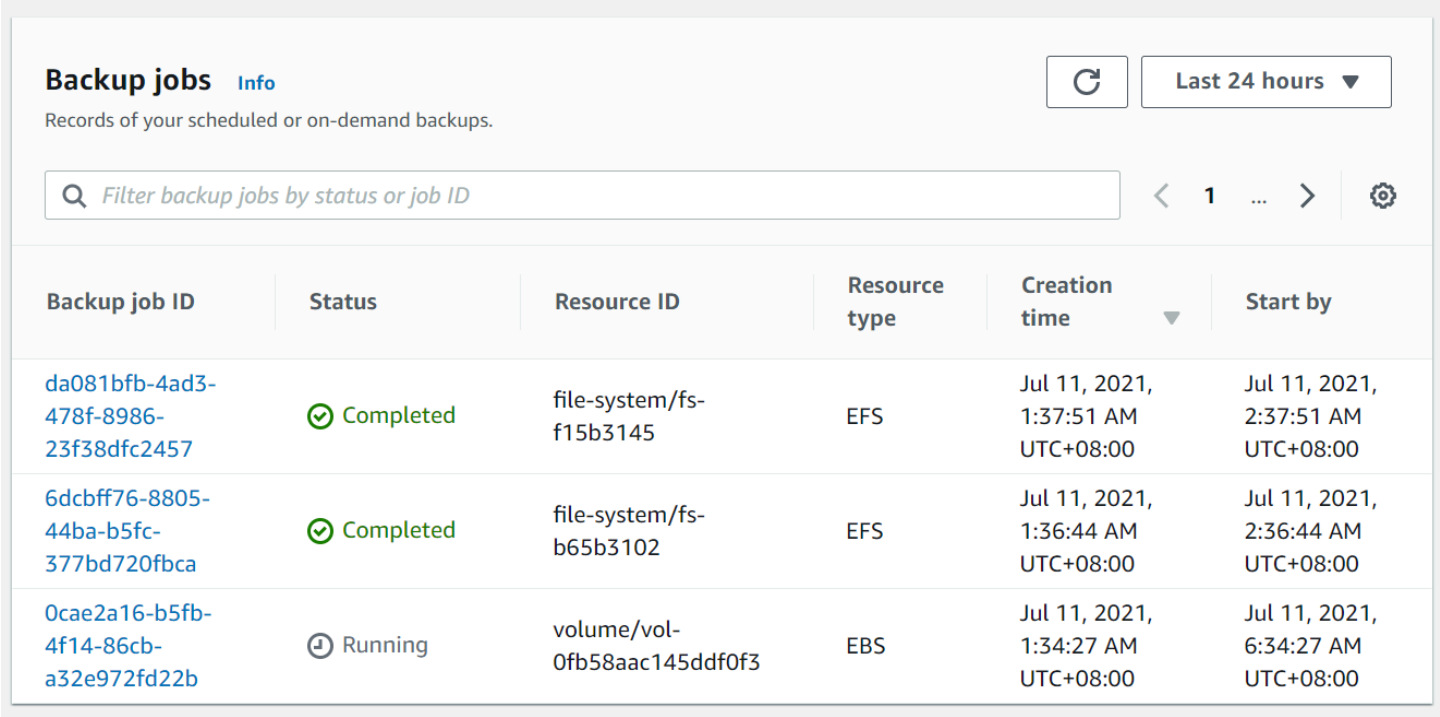
Protected Resources

This is a list of AWS resources that are being backed up.



Backup Jobs

Jobs hold the records of your scheduled and on-demand backups as well as recovery jobs.



Backup jobs Info

Records of your scheduled or on-demand backups.

Last 24 hours

Filter backup jobs by status or job ID

1 ...

Backup job ID	Status	Resource ID	Resource type	Creation time	Start by
da081bfb-4ad3-478f-8986-23f38dfc2457	Completed	file-system/fs-f15b3145	EFS	Jul 11, 2021, 1:37:51 AM UTC+08:00	Jul 11, 2021, 2:37:51 AM UTC+08:00
6dcbff76-8805-44ba-b5fc-377bd720fbca	Completed	file-system/fs-b65b3102	EFS	Jul 11, 2021, 1:36:44 AM UTC+08:00	Jul 11, 2021, 2:36:44 AM UTC+08:00
0cae2a16-b5fb-4f14-86cb-a32e972fd22b	Running	volume/vol-0fb58aac145ddf0f3	EBS	Jul 11, 2021, 1:34:27 AM UTC+08:00	Jul 11, 2021, 6:34:27 AM UTC+08:00

Cross-account Management

By integrating AWS Backup to AWS Organizations, you will be able to do a cross-account backup and monitoring. It also allows you to create and use backup policies across different AWS accounts.

Reference:
https://docs.aws.amazon.com/aws-backup/latest/devguide/whatisbackup.html

Amazon Relational Database Service (RDS)

AWS introduced Amazon Relational Database Service (RDS) to cater to the most challenging and demanding requirements of setting up a database on the cloud. RDS eases the process of creating and maintaining a relational database on the cloud with notable advantages in availability, flexibility, performance, and cost when compared to databases in a traditional environment.

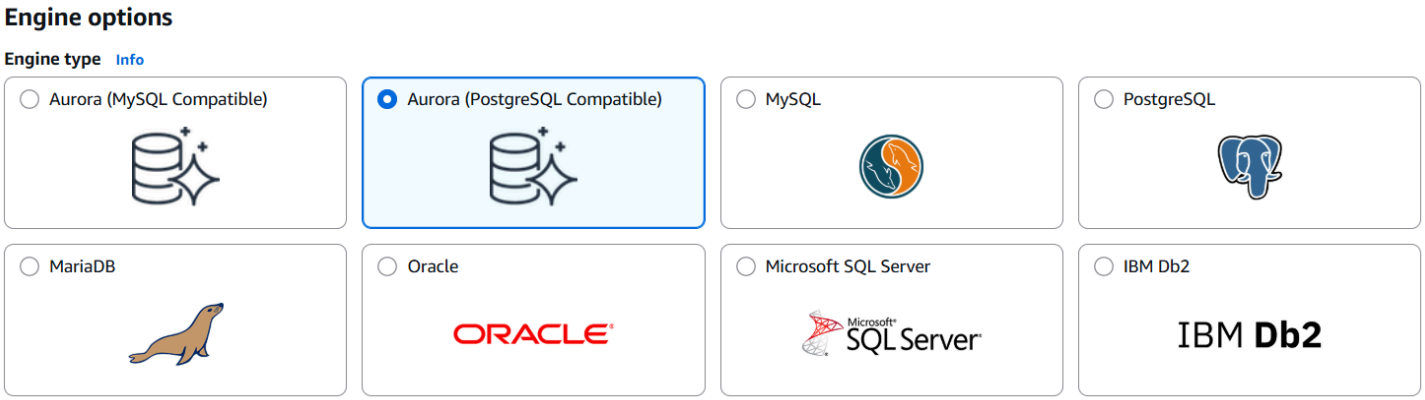
Amazon RDS Features and Components

Amazon Relational Database Service (RDS) is a managed relational database service. Having a managed relational database service means that maintaining and managing the physical infrastructure for the databases is offloaded from the customer and is being taken care of by AWS allowing customers to focus on the database itself. This includes fault management, software patches, backups, and recoveries. To further understand RDS, let's break down its components and features.

Amazon RDS Database Engines

Amazon RDS supports the following relational database engines.

- **Aurora (PostgreSQL Compatible)** - Amazon Aurora's PostgreSQL-compatible edition that combines PostgreSQL compatibility with a cloud-native distributed storage architecture. It provides higher performance, scalability, and availability compared to standard PostgreSQL deployments.
- **Aurora (MySQL Compatible)** - Amazon Aurora's MySQL-compatible edition that combines MySQL compatibility with a distributed and fault-tolerant storage system. It delivers significantly improved performance, automatic scaling, and high availability compared to standard MySQL deployments.
- **PostgreSQL** - an open-source object-relational database known for being reliable and stable. This database engine supports both SQL (relational) and JSON (non-relational) queries.
- **MySQL** - is an easy-to-use, reliable, and very responsive database, making it the most popular relational database engine.
- **MariaDB** - also a well-known high-performing open-source relational database created by the same people who developed MySQL. MariaDB is compatible with MySQL databases.
- **Oracle** - is a commercial multi-model database management system. Built a superior reputation in terms of performance, flexibility, availability, and security over the years. Widely used for transactional database workloads as well as data warehousing. RDS offers both *Bring-Your-Own-License (BYOL)* and *License Included* licensing models when setting up Oracle Database on RDS.
- **Microsoft SQL Server** - a relational database management system offering from Microsoft. Widely used for transactional workloads, analytics, and business intelligence.
- **IBM Db2** - an enterprise-grade relational database management system optimized for transactional processing, analytics, and mission-critical workloads. Amazon RDS for Db2 automates administrative tasks such as backups, patching, monitoring, and high availability while supporting existing Db2 applications and tools.



Choosing Suitable RDS DB Instance Classes

A DB instance is a primary component of RDS. It is a secured, isolated environment for your database. Launching a database engine on RDS requires you to define a DB instance identifier unique within the Region. Every RDS API call and command is associated with a DB instance using its identifier.

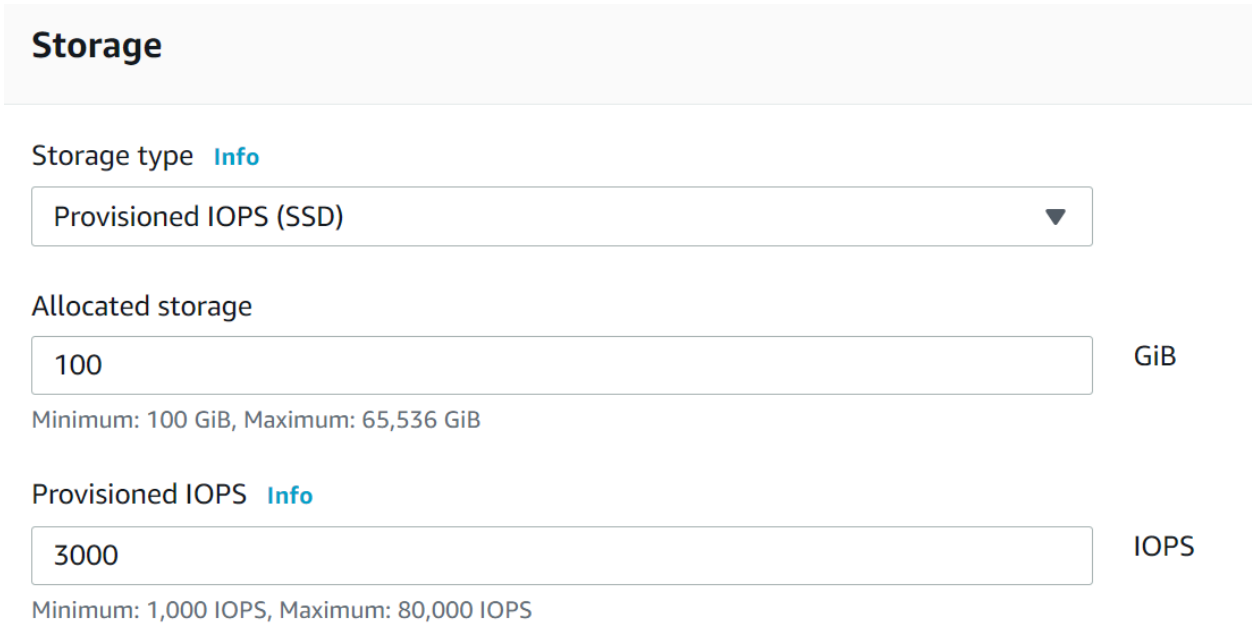
RDS offers different DBS Instance classes to match the different processing power and memory requirements. The availability of the instance class options varies per database engine.

- **Standard** - General-purpose instance with balance performance.
- **Memory-Optimized** - Designed for memory-intensive database workloads.
- **Burstable Performance** - has a baseline performance level that can burst to higher performance; ideal for unpredictable database workloads.

Choosing the Right RDS DB Instance Storages

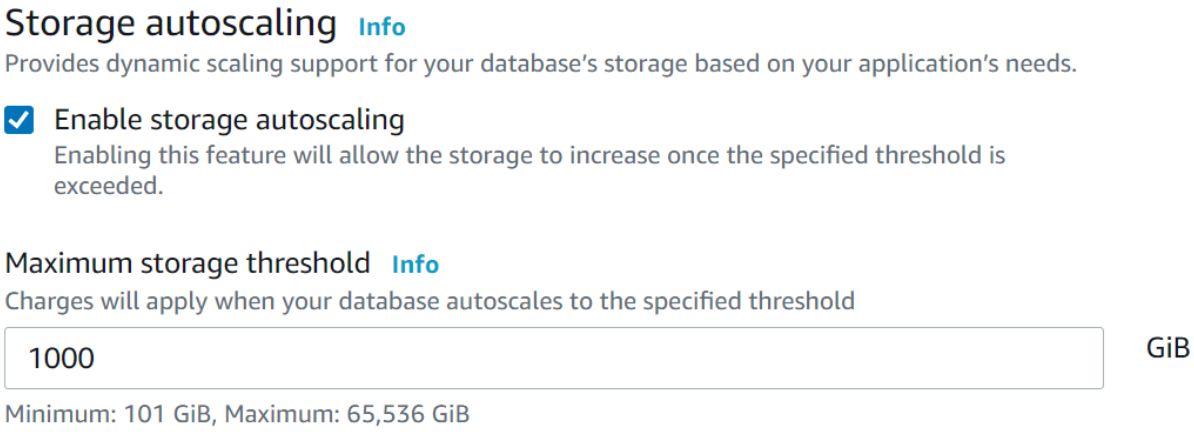
Amazon RDS utilizes Amazon EBS volumes for its storage. The different storage types have distinct performance and cost; choosing the right storage type will highly depend on the database storage requirement.

- **General Purpose SSD (gp2) Storage** - cost-effective storage with burstable performance that suits most database workloads. Storage performance is relational to volume size i.e. the larger the volume, the better the performance.
- **Provisioned IOPS SSD (io1) Storage** - designed for I/O-intensive workloads that require low latency and consistent I/O throughput. Ideal for OLTP workloads with consistent performance.
- **Magnetic (standard) Storage** - ideal for small workloads and infrequently used data. Magnetic storage uses hard disk drives (HDD). It doesn't offer much performance when compared to gp2 and io1 storage.



Right-sizing database storage is also a crucial part, especially for databases with unpredictable workloads. In these cases, it is ideal to use RDS storage autoscaling. On this feature, RDS will monitor and trigger autoscaling according to the following factors.

- Free space is less than 10% of the total storage.
- The low-storage condition lasts at least 5 minutes.
- At least 6 hours have passed since the last storage modification.



Choosing a Region and Availability Zone for RDS Instance

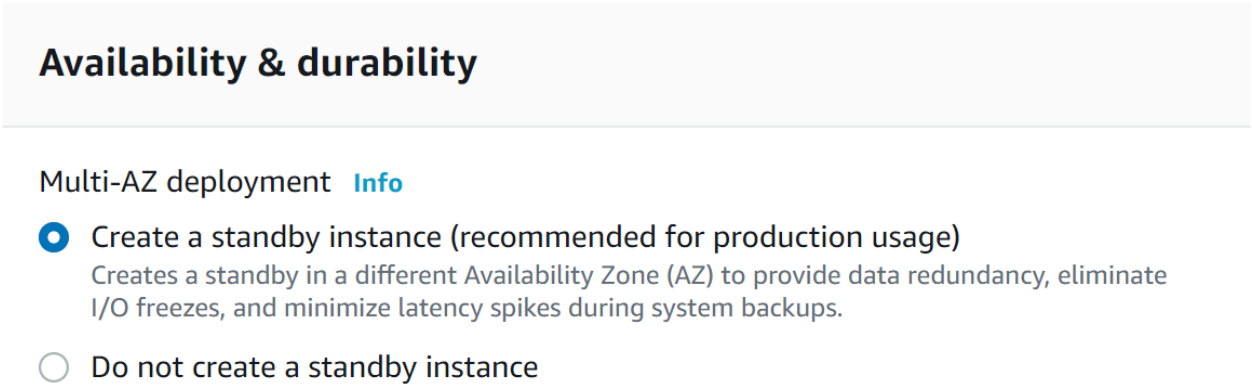
When choosing the Region for your DB instance, it is vital to consider the region of the application connecting to the database and its end users. AWS Regions are isolated and completely independent from other Regions. Likewise, an Availability Zone (AZ) is isolated from different Availability Zones within a Region.

You can also extend an AWS Region to your end-users using Local Zones. A local zone can host services like compute services, databases, and storage. A local zone is set up by creating a VPC Subnet in your Region and assigning it to a Local Zone.

Increasing Database Availability Using Multi-AZ Deployment

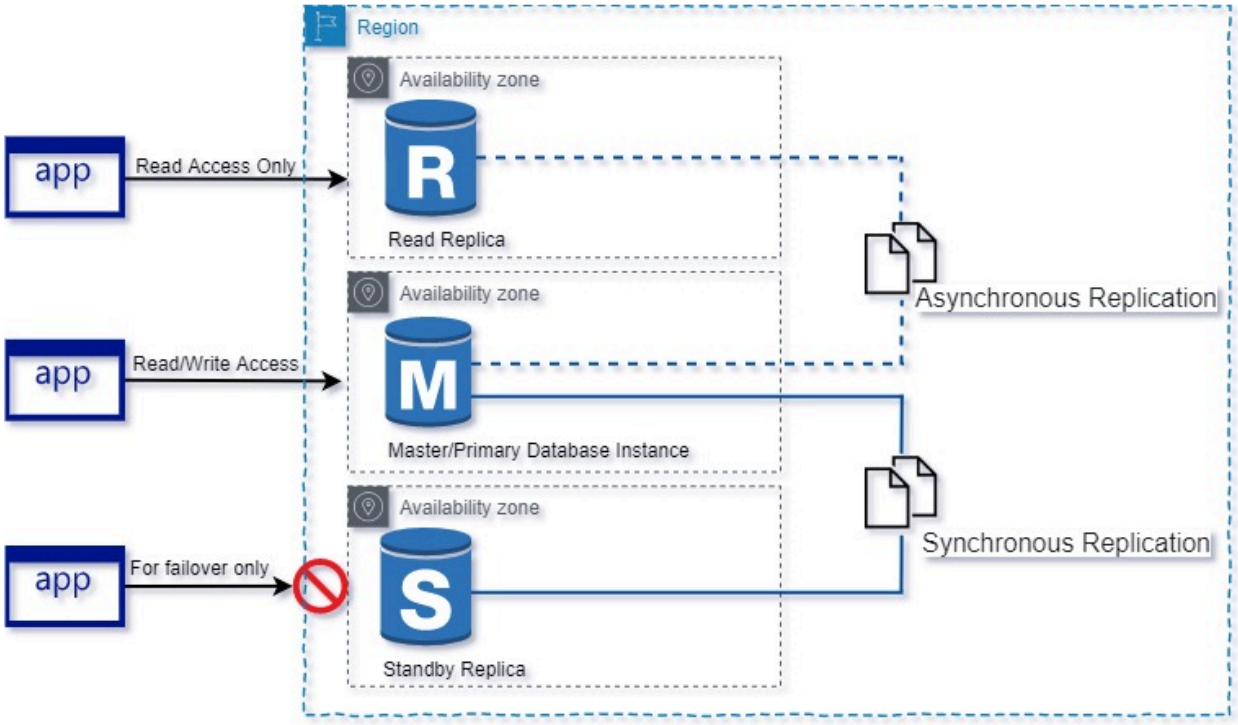
A database holds a collection of essential data. Always consider a Multi-AZ deployment, especially for the production databases, as it offers high availability and failover features. Though the odds are minimal, there is always a chance that an Availability Zone will go down. The worst case is an offline production database because of a faulty Availability Zone. Database engines like MySQL, MariaDB, Oracle, and PostgreSQL DB instances leverage Amazon's failover technology for Multi-AZ deployment. At the same time, Microsoft SQL Server uses SQL Server Database Mirroring (DBM) or Always On Availability Groups (AGs) for high availability. Just note that you can't use Local Zones for Multi-AZ deployment.

You can enable Multi-AZ deployment during the database creation or when editing the DB instance for existing databases. When the Multi-AZ deployment is enabled, RDS creates a standby replica in an Availability Zone that is different from the primary database. RDS automatically handles the synchronous replication between the primary database and replicas and the failover process during unexpected disruption on the primary database.



Amazon Aurora has durability and high availability features native to it. Amazon Aurora always keeps a copy of the data in a DB cluster that spans multiple Availability Zones within a Region. Additionally, Amazon Aurora has replication capabilities that are set up to either Single-master or Multi-master. A Single-master replication can

create up to 15 Aurora Replicas or Read Replicas. If a primary database becomes problematic, a replica will be promoted as the primary database.



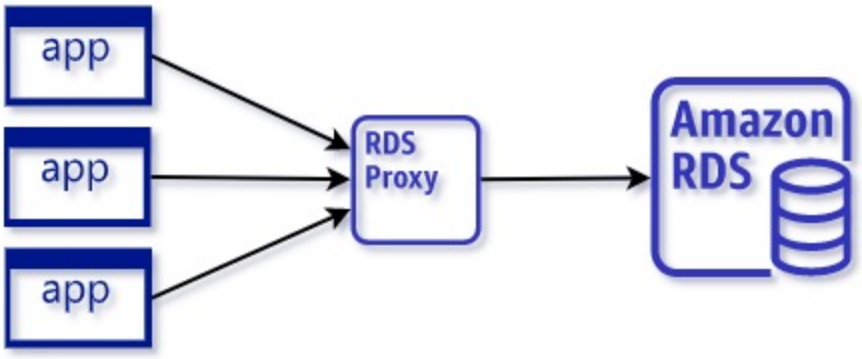
Improving Database Performance using Read Replica and DB Clusters

A Read Replica improves the availability and increases the performance of your database at the same time. You can do this by routing most of the read queries of your application to a read replica instead of the primary database. Changes made on the primary database are asynchronously replicated to its read replica. The database engines MySQL, MariaDB, Oracle, PostgreSQL, Microsoft SQL Server have built-in replication features that RDS uses to create read replicas.

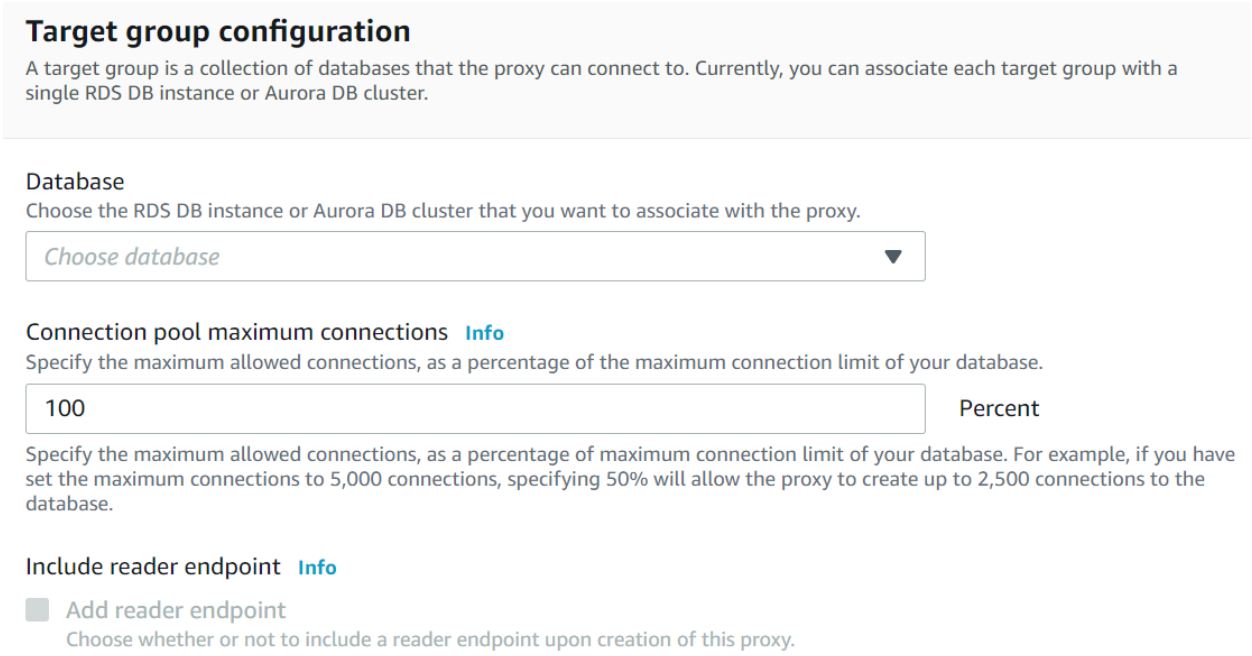
A database cluster is a collection of database servers that shares storage to improve performance and availability. Similarly, an Amazon Aurora DB Cluster consists of one or more database instances with a cluster volume that spreads to multiple Availability Zones. This cluster volume's size increases as needed. RDS replicates the Primary DB Instance to create multiple Aurora Replica. These replicas can accept read queries to minimize the queries on the primary database, significantly improving the database performance and improving availability, as discussed earlier.

Adding an RDS Proxy

If you expect your application to have a high volume of connections to your database or have an application with unpredictable workloads, you might want to use a proxy. Some applications have a high rate of opening and closing a database connection or otherwise keep a database connection on idle status, leading to high CPU and memory utilization. You can over-provision your database to accommodate all the workloads, or better yet, add an RDS proxy.

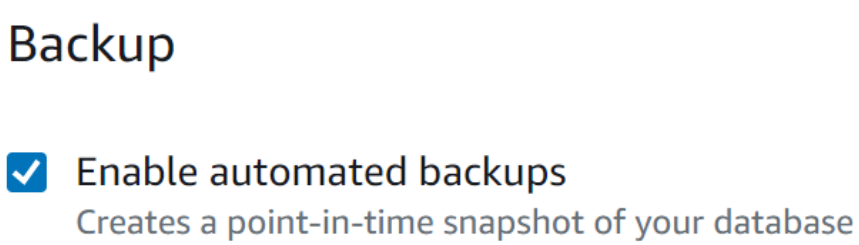


RDS proxy operates between your application and database to manage all the connections between these two. You can set a certain percentage of connections for your database according to its number of connection limits; thus, you can expect that your database will consistently process a good percentage of connections avoiding a sudden burst on the number of connections. You can also set a connection timeout for idle connections. RDS proxy enables an application to reuse the pooled connections to minimize the need to establish a new connection. RDS proxy is MySQL and PostgreSQL-compatible and can also require a connection over TLS.

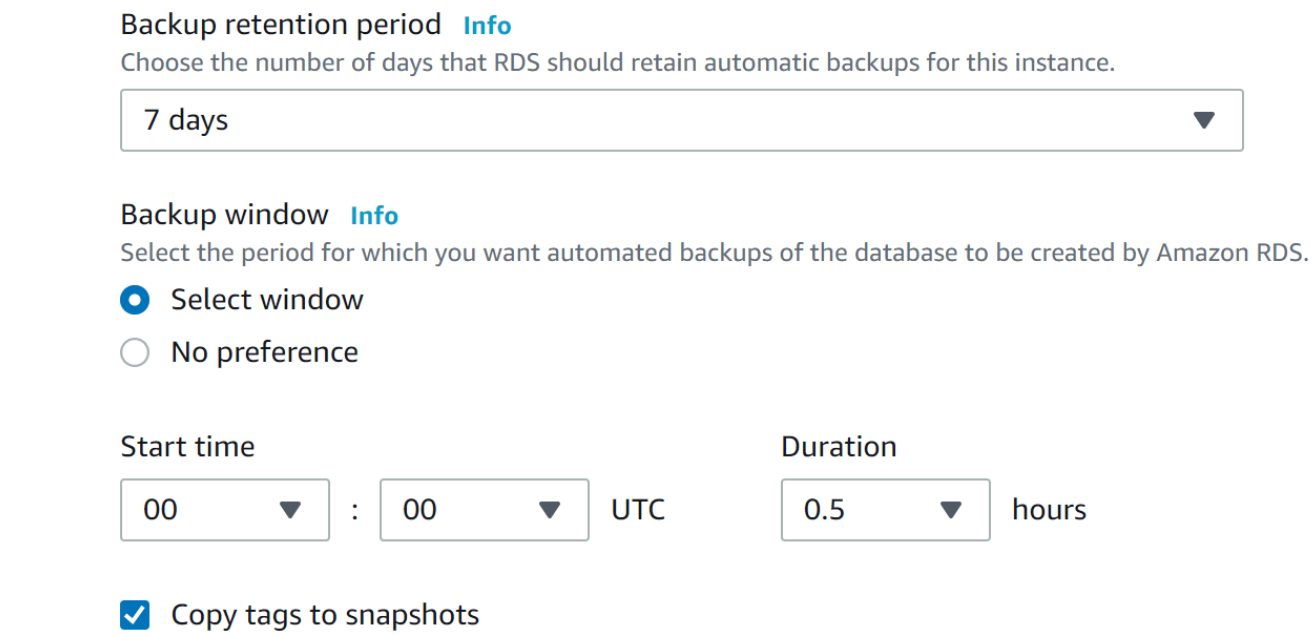


Working with RDS Backup

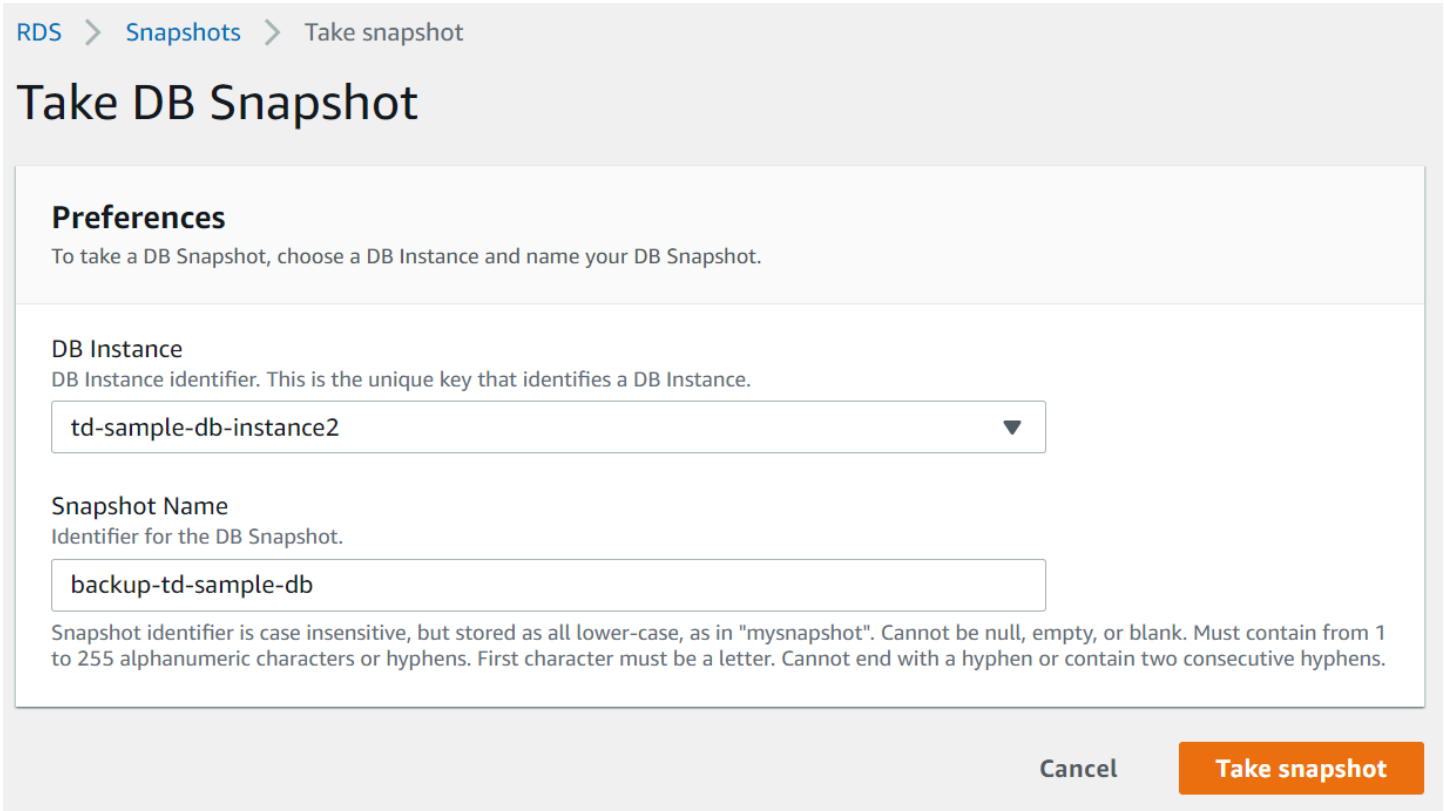
Amazon RDS takes a snapshot of the storage volume used by the DB instance when you enable the automated backup feature. The automated backup allows you to make a Point-in-Time Recovery (PITR) when the necessity occurs.



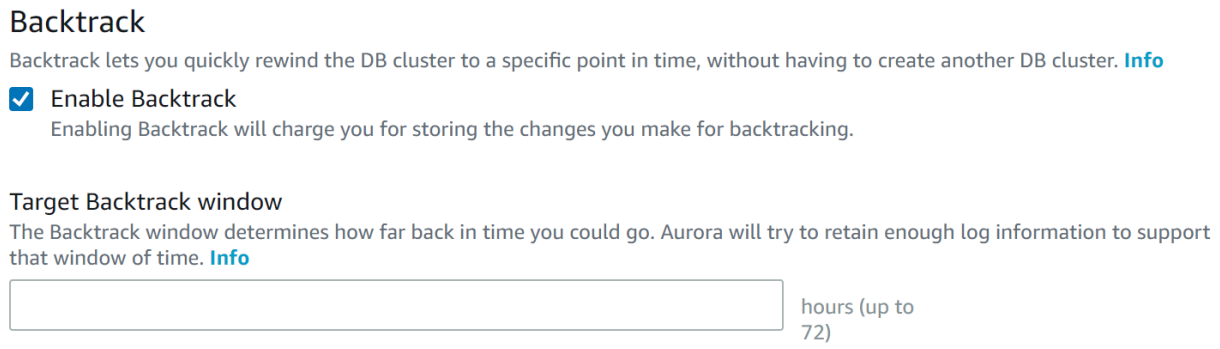
RDS will create a snapshot based on the specified backup window. The retention period can also be set from 0 to 35 days.



If needed, you can manually take a snapshot at any given time.

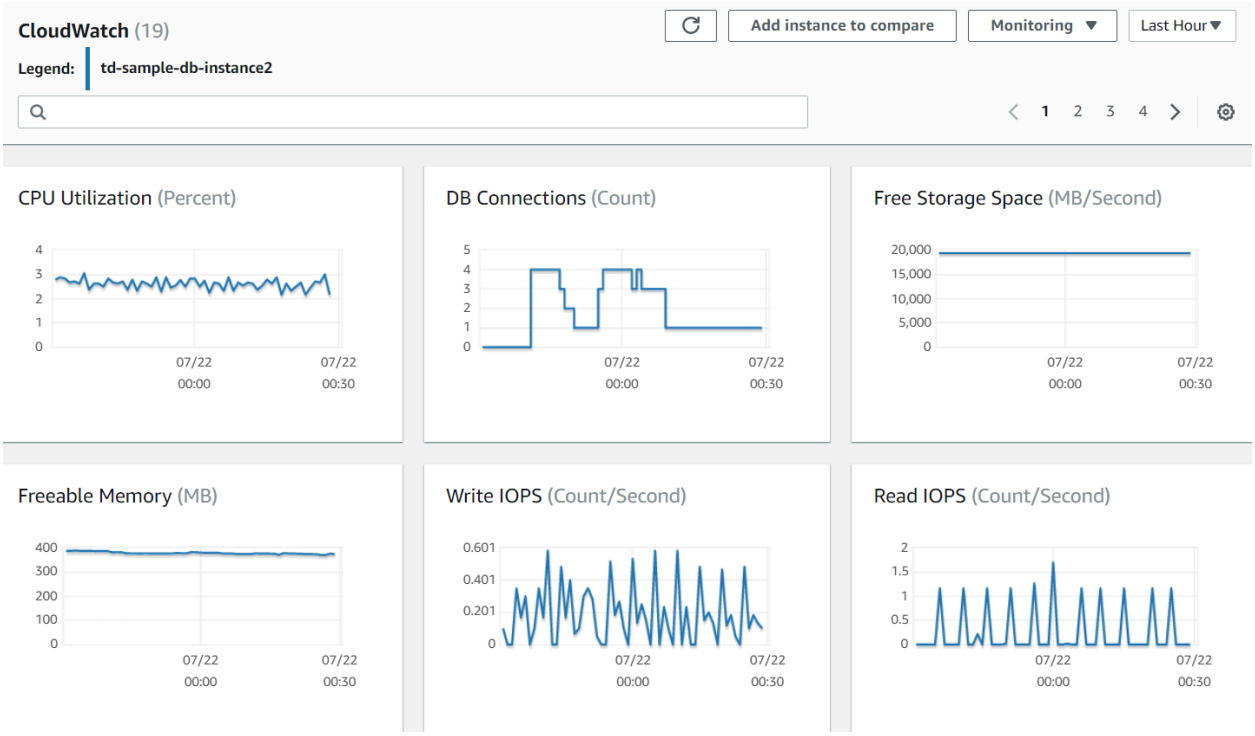


Amazon Aurora offers a Backtrack option to “rewind” a DB cluster. Backtrack is a way to undo data changes in minutes compared to restoring a DB cluster using a snapshot that could take hours to complete. Backtrack can rewind a DB cluster to its previous state for as long as 72 hours.

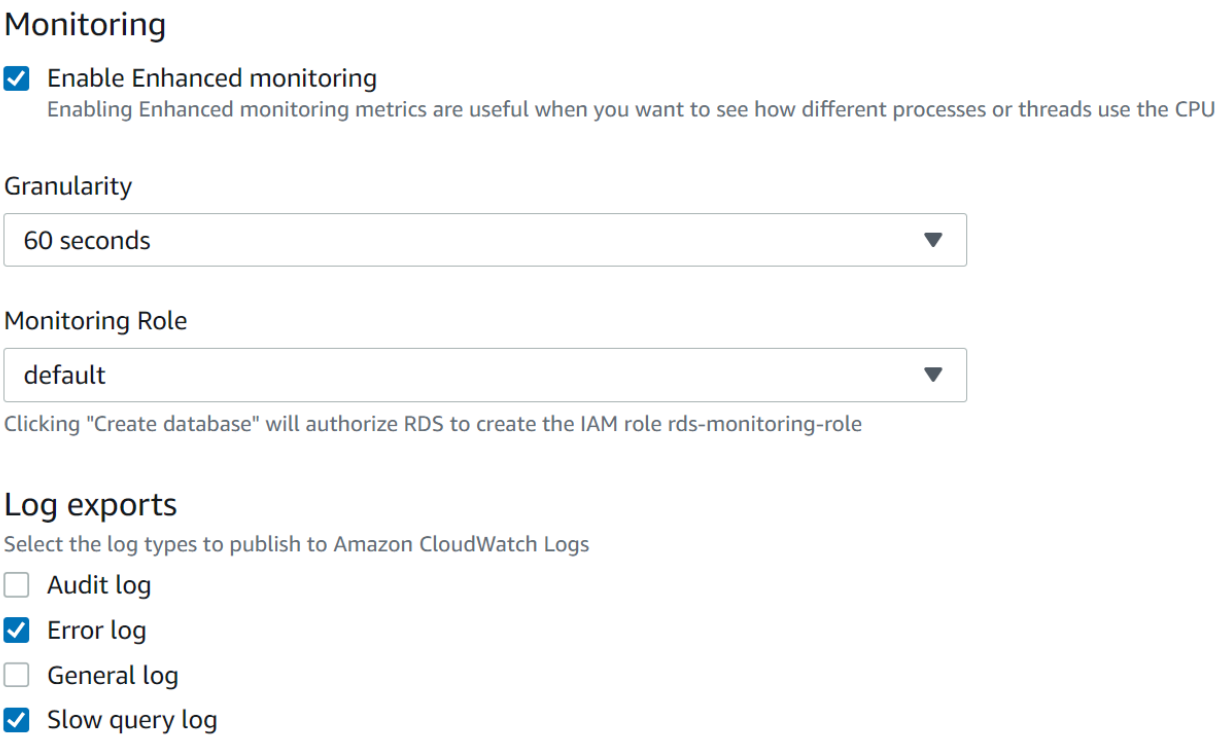


Monitoring a Database Instance

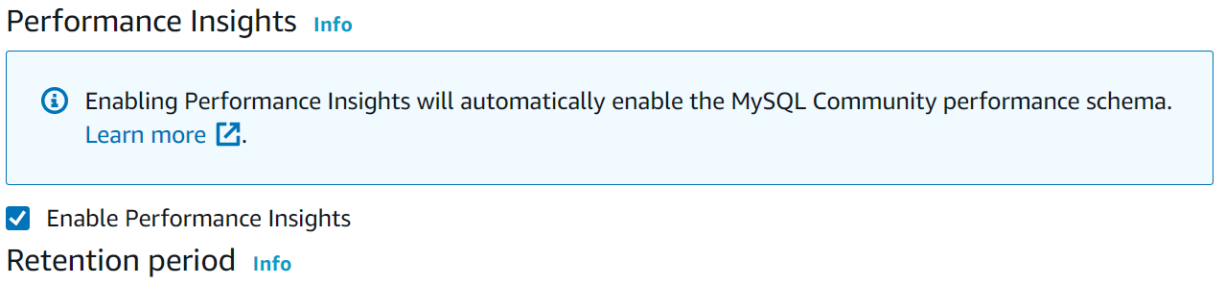
Amazon RDS integrates with different AWS services to provide efficient monitoring of your databases. Metrics of the active database are sent to Amazon CloudWatch every minute. You can then create Cloudwatch Alarms using these metrics. The default monitoring that CloudWatch provides seems to be enough for testing of non-critical databases.



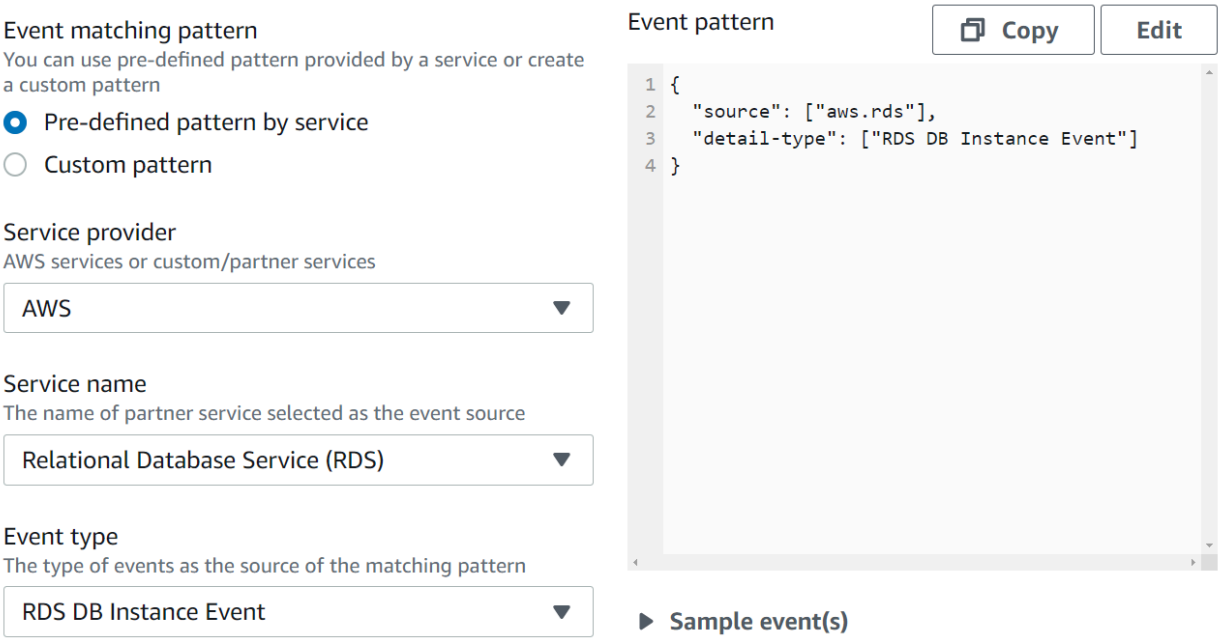
For production and databases with critical workloads, it is best to enhance monitoring options for a thorough and granular performance monitoring. Moreover, database log files can be stored and accessed using Amazon CloudWatch Logs. Moreover, AWSCloudTrail logs all the activities on RDS by recording all API calls. These features are helpful when troubleshooting or investigating particular database issues.



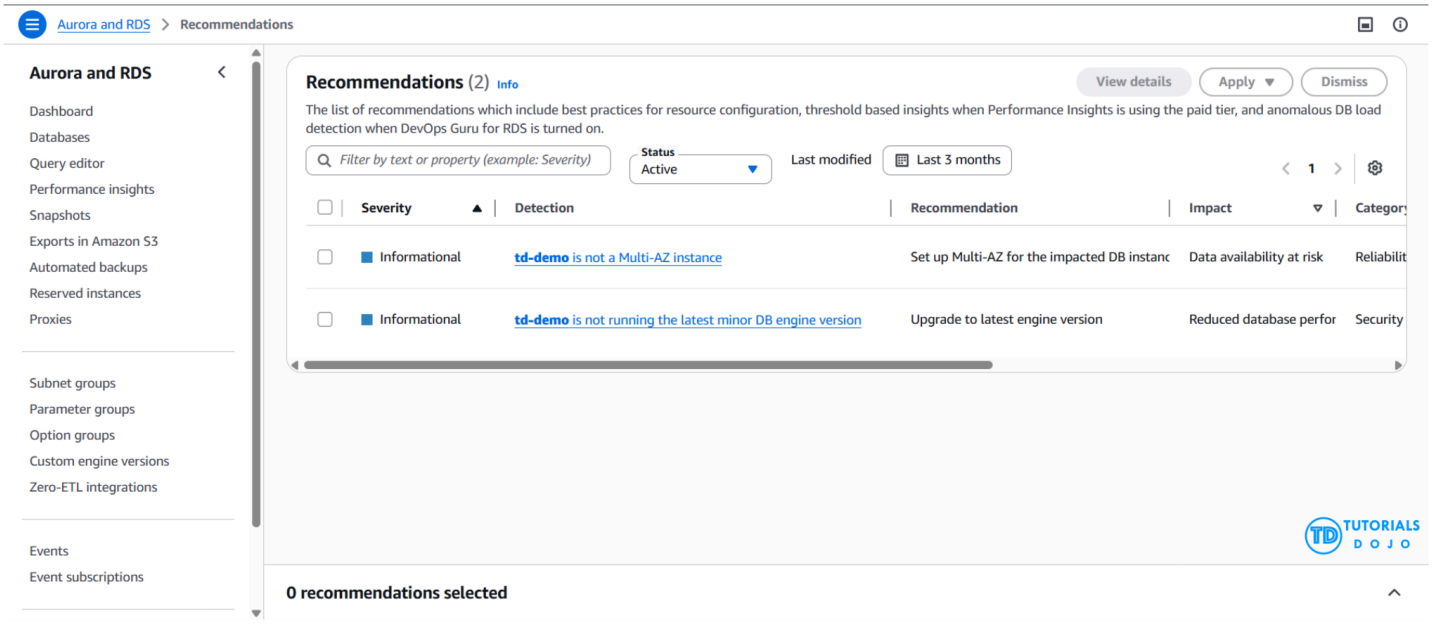
Performance Insights provides a dashboard to visualize and analyze your database's performance. This feature allows you to view essential insights like database sessions, queries, database load, and top charts for users, connections, queries, and others.



You can also create a rule on Amazon EventBridge to monitor and log events of your database using an AWS Lambda function.

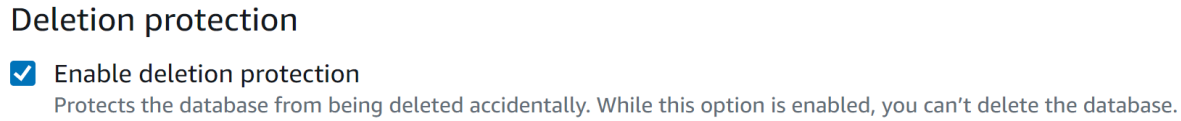


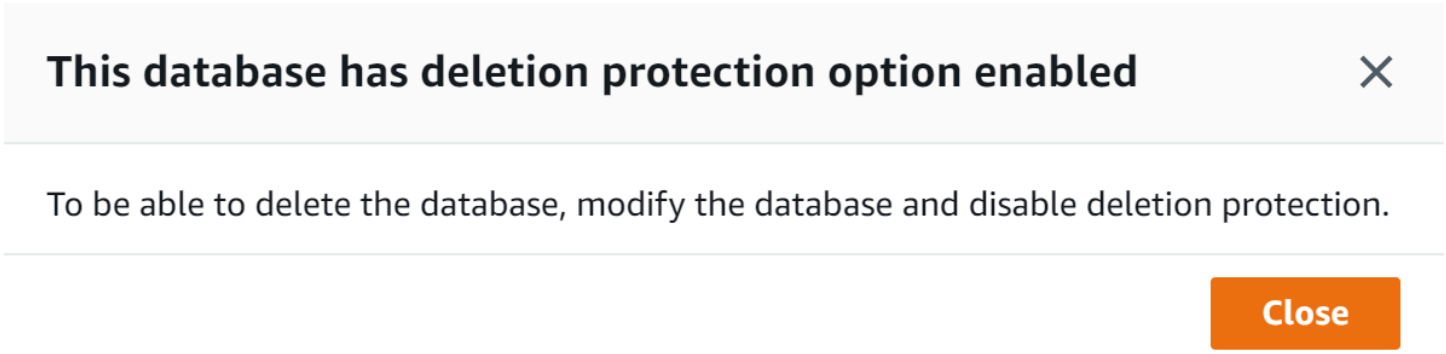
Performance Insights proactive recommendations identify potential performance issues before they impact your database, offering suggested actions to prevent future problems. They work by monitoring metrics, setting dynamic thresholds, and generating recommendations when anomalies or problematic patterns are detected.



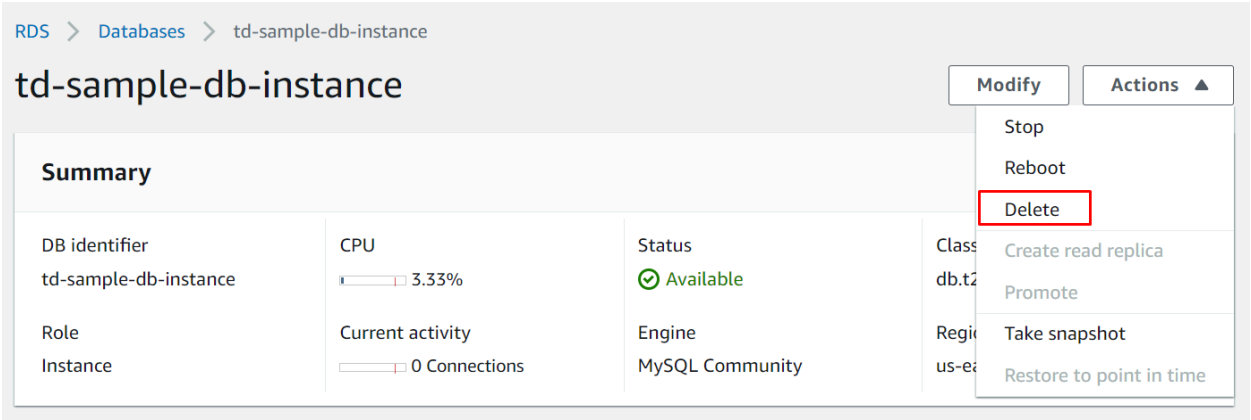
Deleting a Database Instance

Like other AWS services, Amazon RDS also has a deletion protection feature to avoid accidental deletion of the database.

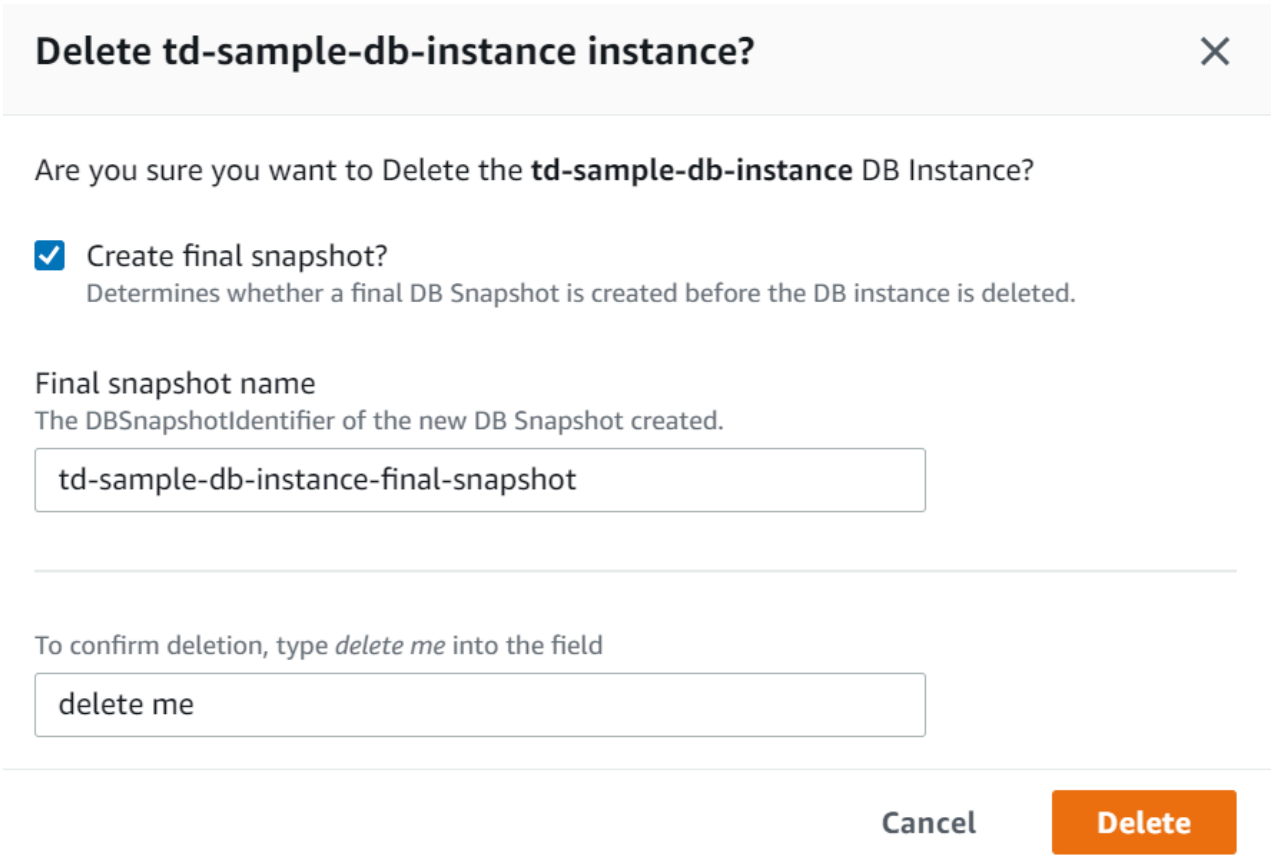




You can trigger a database deletion by selecting a database and clicking *Delete* from the Actions.



You will have an option to take a snapshot of your database before deleting a database instance in case you want to keep a backup. Confirmation should also be done by typing "*delete me*."



Reference:
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html

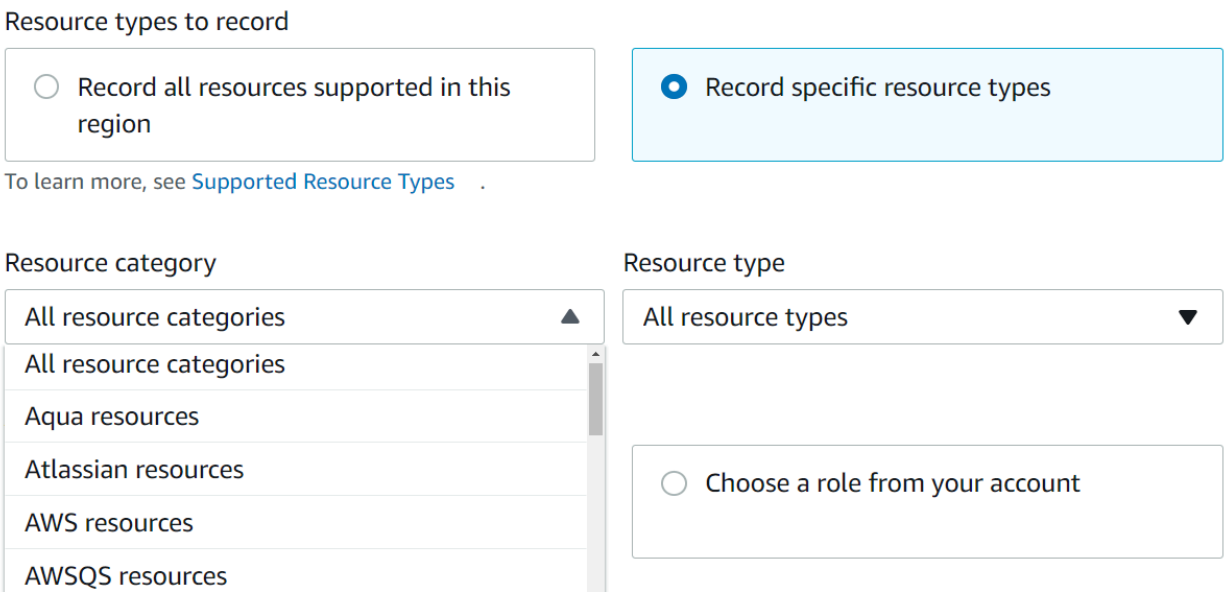
Amazon OpenSearch Service

Amazon OpenSearch Service is designed to simplify the deployment, operation, and scaling of OpenSearch (an open-source search and analytics suite derived from Elasticsearch 7.10). It provides robust capabilities for full-text search, log analytics, real-time application monitoring, and security analytics. This service works smoothly with other AWS services and provides features that ensure it's always available, can grow with your needs, and has security measures like encryption and access control. It's an ideal solution for businesses requiring near real-time data and insight access.

AWS Config

AWS Config is a service that keeps track of the configuration of your AWS resources. This service allows you to continuously monitor and assess configurations beneficial for audit and compliance requirements. AWS Config keeps the records of resource configuration in an S3 bucket. Additionally, aside from AWS resources, AWS

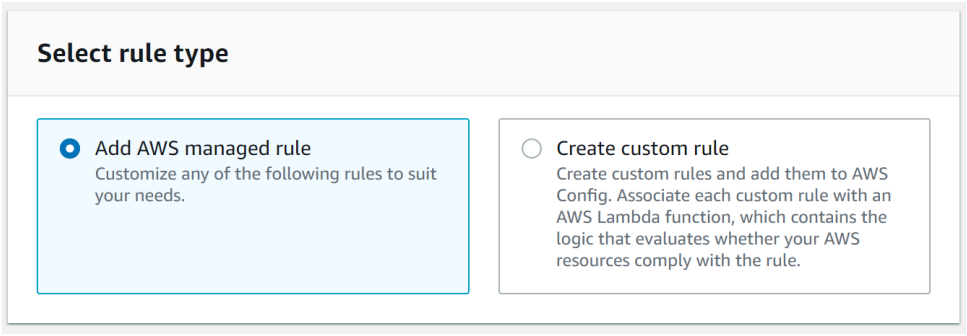
config also supports a wide array of third-party resources to monitor. Resources are tracked according to their record type and category. You also have the option to discover and watch all resources in an entire region.



AWS Config stores the configuration details of its monitored resources to an **Amazon S3** bucket. Each record type has its configuration history. You can also trigger an **Amazon SNS** to notify you of any changes in the configurations.

AWS Config Continuous Configuration Monitoring

Resources are monitored and evaluated according to a set of desired configurations called **Rules**. You can use a predefined but customizable AWS managed rule or create your own Custom Rule using the AWS Lambda function.

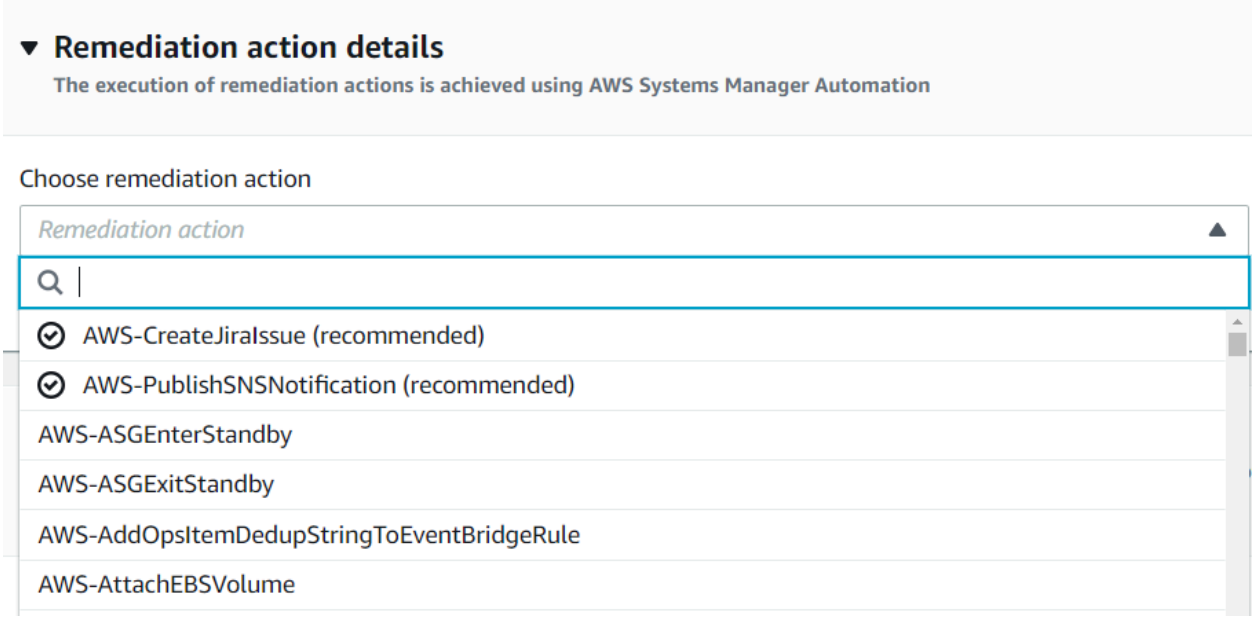


Resources evaluation happens when a configuration change occurs, and during the schedule you set. AWS Config marks the resources as *COMPLIANT* or *NON_COMPLIANT* according to its config rules. See below example of an AWS-managed rule for EC2.

○	desired-instance-type	EC2	Checks whether your EC2 instances are of the specified instance types.
○	ebs-optimized-instance	EC2	Checks whether EBS optimization is enabled for your EC2 instances that can be EBS-optimized.

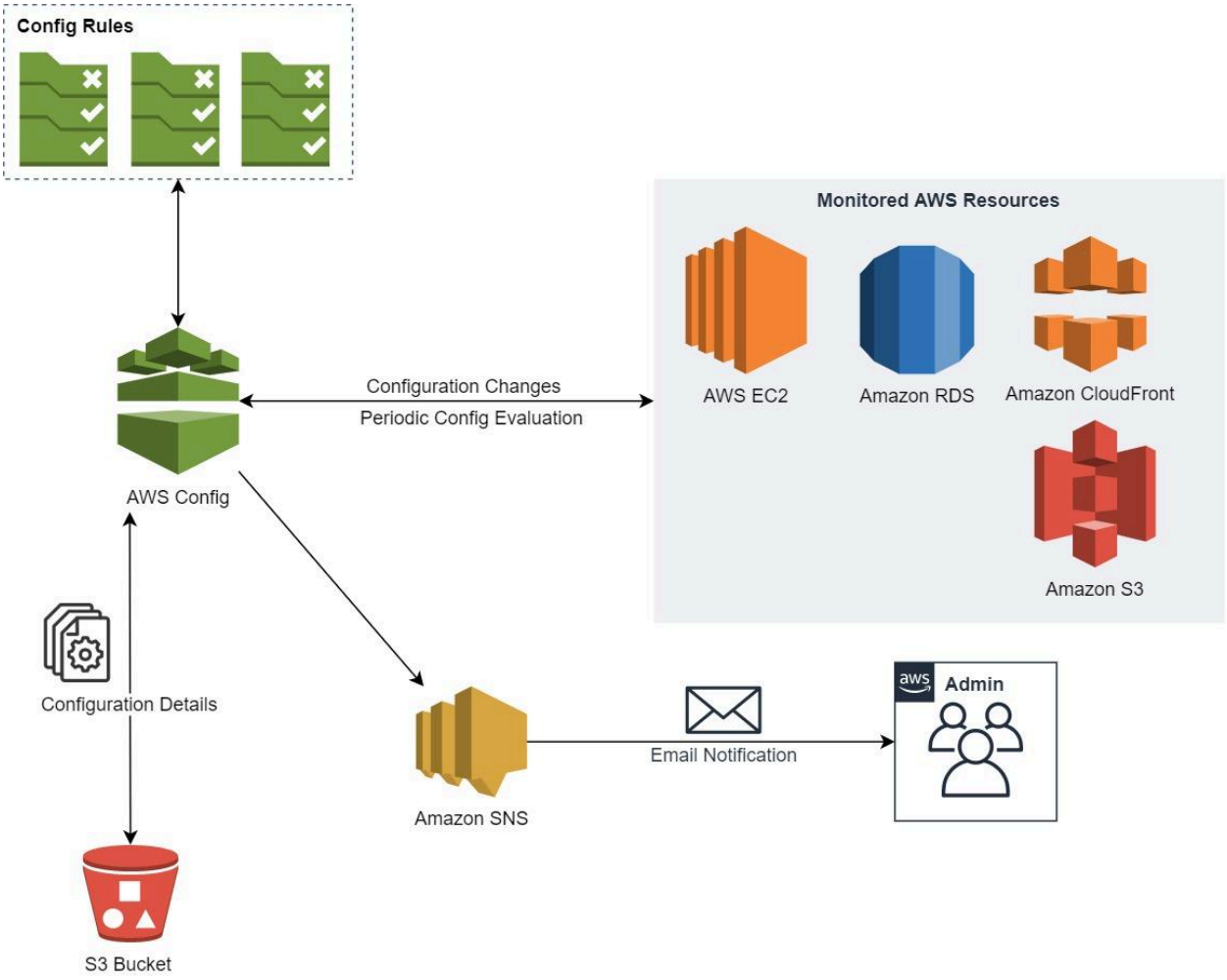
.

For reported non-compliant resources, a remediation action can be applied either automatically or manually. AWS also recommends remediation for the non-compliant resource.



To further broaden the scope of your configuration monitoring, you can set up an **AWS Config Aggregator**. Data aggregation will allow you to collect all configuration data across multiple AWS accounts and Regions and consolidate it all in one place.

The diagram below shows a high-level architecture of AWS Config. Multiple AWS resources are being evaluated using the rules periodically or in the event of configuration changes. It also shows the configuration details being stored in an Amazon S3 bucket and email notifications sent via Amazon SNS.



Reference:
https://docs.aws.amazon.com/config/latest/developerguide/WhatIsConfig.html

Deploying Resources with CloudFormation

AWS CloudFormation lets you model and create resources for your environment using programming language, leveraging the concept of Infrastructure as Code (IaC). You don't need to make all of your resources one by one manually; these resources are all defined on the AWS CloudFormation template. In this way, an AWS environment can be reliably and quickly reproduced since everything is declared on a template. An example would be creating a new Test environment similar to your current setup or creating a Disaster Recovery environment in another region.

With AWS CloudFormation, you can upload your template, use a sample template, or create a template using Designer. These templates can be in JSON or YAML format. Resources defined from these templates are

treated and managed as a single unit called stacks. A template may contain the following sections, with Resources being only the required section.

- **Format Version** - AWS CloudFormation template version
- **Description** - Template description
- **Metadata** - Additional details of the template
- **Parameters** - values used by the template during stack creation
- **Rules** - used to validate the parameters being used by the template
- **Mappings** - matches a key to its corresponding value
- **Conditions** - specifies the condition required for creating resources
- **Transform** - specifies the version of the AWS Serverless Application Model
- **Resources** - identifies the stack resources and their properties
- **Outputs** - defines the output value

StackSets

AWS CloudFormation is used mainly for automating deployments of different applications. For creating resources for multiple regions and multi-accounts deployment, you should consider using StackSets. With StackSets, you can create resources for the Target Accounts using an Administrator Account. The administrator account centrally manages the templates, and you only specify the target accounts to where you want to create your resources.

Nested Stacks

As your infrastructure grows, there will be some cases where you need to declare the same resources to multiple CloudFormation templates. In these cases, it is better to use nested stacks. You can create separate templates for the most used resources and reference them on other templates. This way, you'll avoid copying and pasting the same configuration on your templates, and this also simplifies stack updates.

Below is an example CloudFormation Template main stack in YAML format.

```
AWSTemplateFormatVersion: 2010-09-09
Resources:
  MyStack:
    Type: AWS::CloudFormation::Stack
    Properties:
      TemplateURL: https://aws-account.testbucket.amazonaws.com/s3bucketcreate.yaml
      TimeoutInMinutes: 60
Outputs:
  StackRef:
    Value: !Ref MyStack
  OutputFromNestedStack:
    Value: !GetAtt MyStack.Outputs.BucketName
```

The example CloudFormation template below is referenced on **AWS::CloudFormation::Stack** using its S3 object URL, allowing the ***GetAtt*** function to pull out the output values from the referenced template.

s3bucketcreate.yaml (uploaded on S3)

```
AWSTemplateFormatVersion: 2010-09-09
Resources:
  SampleBucket:
    Type: AWS::S3::Bucket
Outputs:
    BucketName:
        Value: !Ref 'SampleBucket'
        Description: This is a sample bucket
```

Deleting a Stack

Deleting a stack on CloudFormation also removes all the provisioned resources in it. In some cases, you want some resources to be retained even after deleting its stack. The good thing is that you can do this by defining its *DeletionPolicy*.

To keep the resources when deleting a stack, you need to define its *DeletionPolicy* with *Retain* value on the template. You can set Snapshot as its value for the resources that support snapshot (like RDS databases). With *DeletionPolicy: Snapshot,* a snapshot is created before a resource is deleted. In this way, you will have a backup of the resource deleted from the stack.

Retain

Adding *DeletionPolicy: Retain* on the template will retain the provisioned resources even after deleting its stack.

```
AWSTemplateFormatVersion: 2010-09-09
Resources:
  SampleBucket:
    Type: AWS::S3::Bucket
    DeletionPolicy: Retain
Outputs:
    BucketName:
        Value: !Ref 'SampleBucket'
        Description: This is a sample bucket
```

Snapshot

DeletionPolicy: *Snapshot* can be added on resources that support snapshots like the following:

```
AWS::EC2::Volume

AWS::ElastiCache::CacheCluster

AWS::ElastiCache::ReplicationGroup

AWS::Neptune::DBCluster

AWS::RDS::DBCluster

AWS::RDS::DBInstance

AWS::Redshift::Cluster
```

In this example, we have a Volume created along with an EC2 instance. Because Snapshot is defined as its DeletionPolicy, we expect this to create a snapshot when the stack is deleted.

```
AWSTemplateFormatVersion: 2010-09-09
Resources:
  Ec2Instance:
    Type: AWS::EC2::Instance
    Properties:
      ImageId: ami-0615132a0f36d24f4
  TestVolume:
    Type: AWS::EC2::Volume
    DeletionPolicy: Snapshot
    Properties:
      Size: 10
      AvailabilityZone: !GetAtt Ec2Instance.AvailabilityZone
```

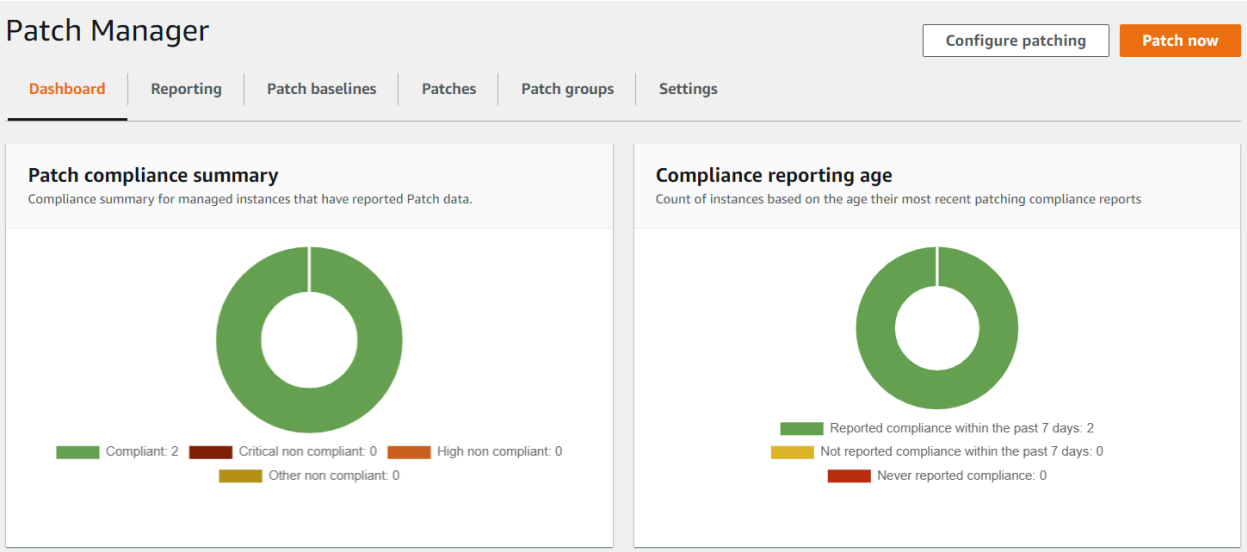
References:
https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/template-guide.html
https://tutorialsdojo.com/aws-cloudformation-stacksets-and-nested-stacks/
https://tutorialsdojo.com/aws-cloudformation-deletion-policy/

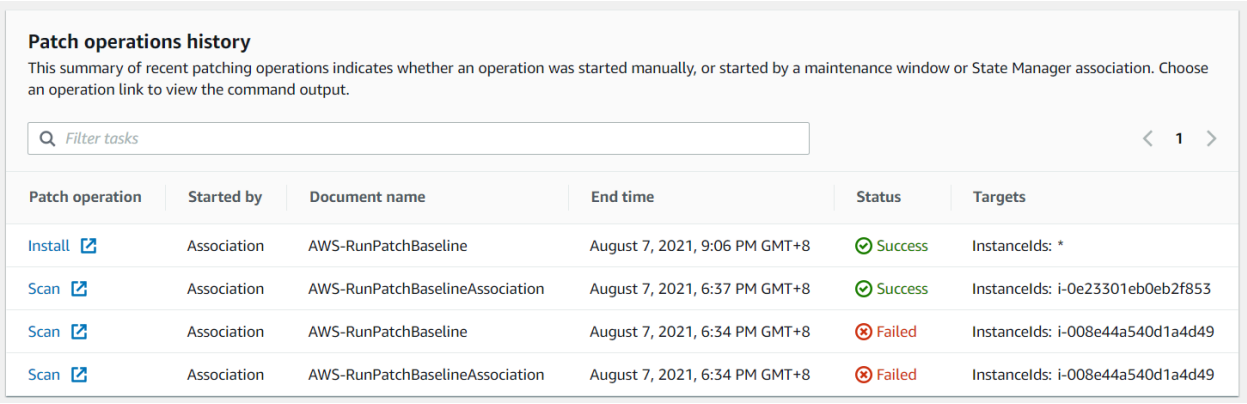
AWS Systems Manager Patch and Change Manager

AWS Systems Manager Patch Manager

Patch Manager is an AWS Systems Manager feature that helps you identify and install the necessary application and operating system updates on your managed instances. Managed instances can be EC2 instances or machines from an on-premises environment with SSM Agent installed.

Patch Manager provides a dashboard that gives essential information like compliance status and patch history.



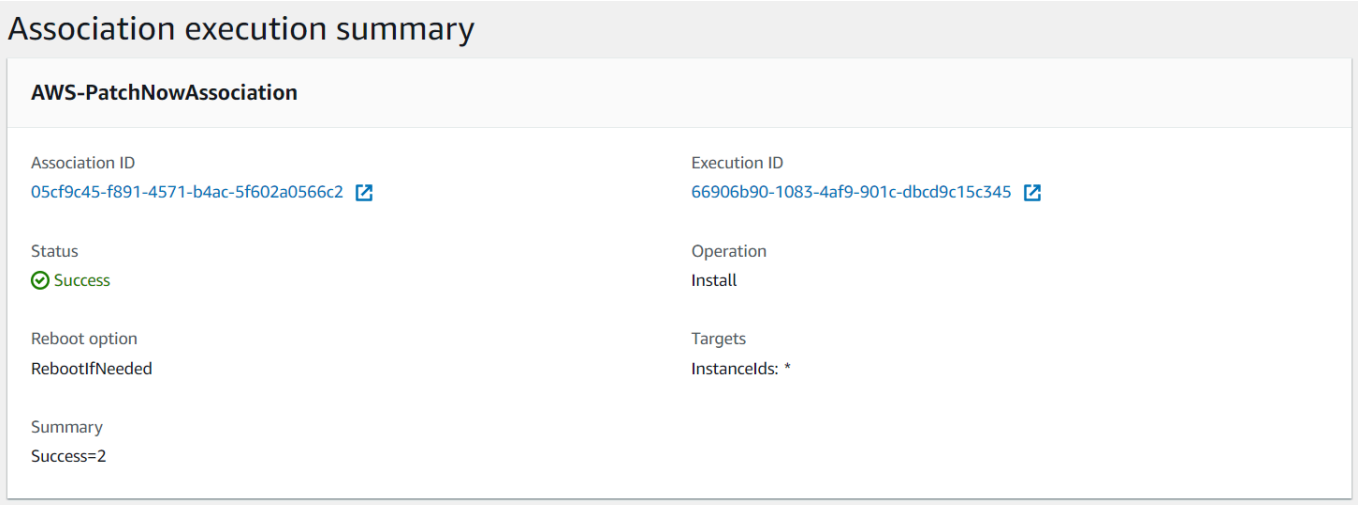


To review the list of your managed instances, you can check **Systems Manager Fleet Manage**r. The Fleet Manager consolidates all managed instances in a single view allowing you to do administration tasks like RDP/SSH, viewing file system, and user/group management.

To further understand how a Patch Manager works, here are the concepts that you need to know:

- **Patch Baseline** defines the patches that are approved for installation on your managed instances. AWS provides a list of Patch Baselines per operating system, but you can configure a custom Patch Baseline.
- When defining instances to patch, you can select instances manually, select instances using tags, or select a Patch Group. A **Patch Group** allows you to group identical instances for a particular patch baseline. You can create a patch by tagging your managed instances with the *key: Patch Group*.
- **Patch Operation** lets you choose between *Scan* and *Scan and Install* operation for your selected instances. Scan operation only identifies and generates a patched list while Install operation downloads and installs all approved patches on the instances.
- **Patch Schedule** defines the maintenance window for the patching. You can also trigger a patch instantly.

The Patch Manager provides a summary after a patch is executed.

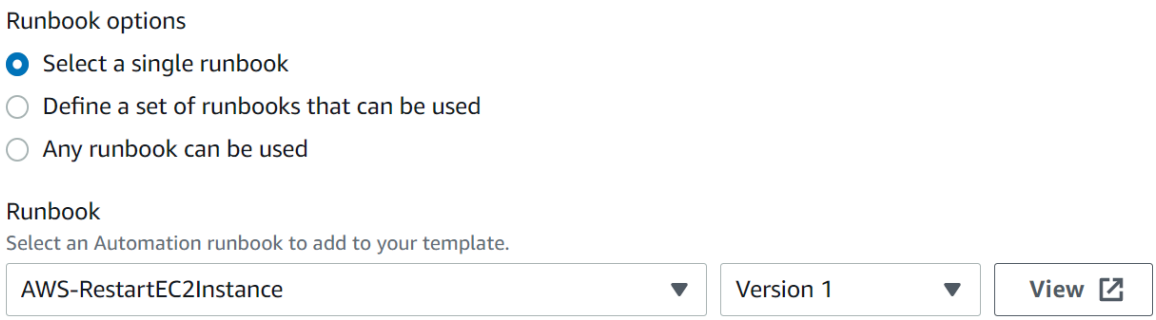


AWS Systems Manager Change Manager

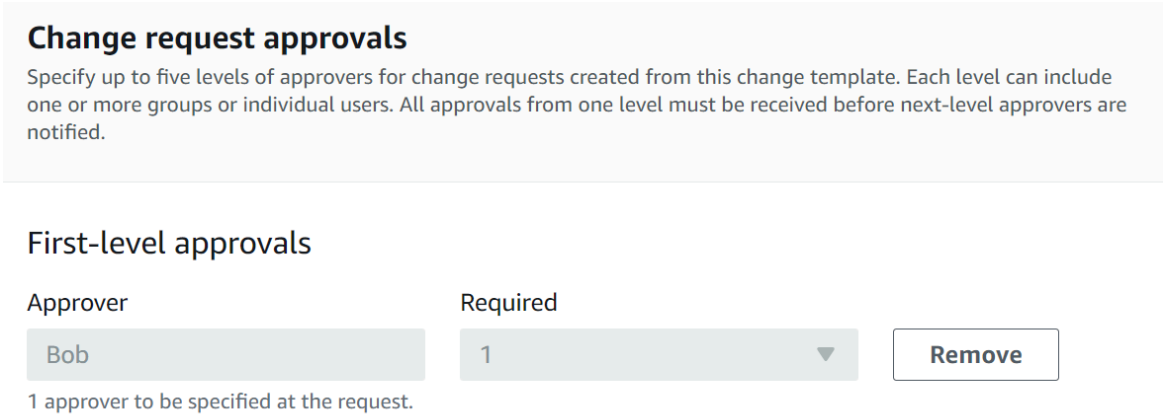
Change Manager is a change management capability of AWS Systems Manager. Change management is a practice done by many organizations to control changes in a given system. In an AWS environment, this capability simplifies submitting, reviewing, approving, and implementing operational changes to application and infrastructure configurations. The Change Manager can manage all changes across AWS Regions and AWS Accounts.

To further understand how a Change Manager works, here are the concepts you need to know:

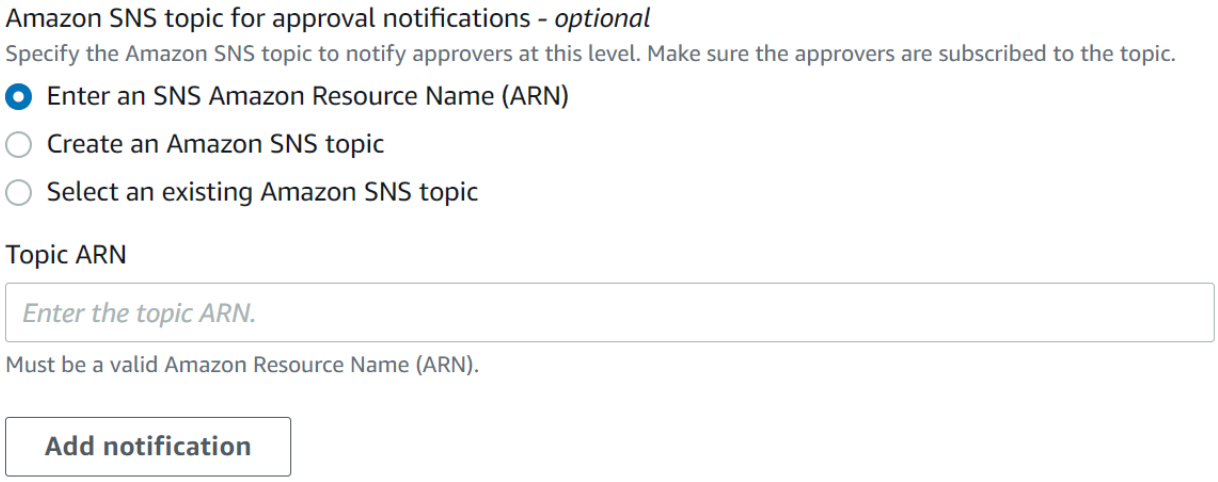
- **Change Templates** contain information like level of approvals, runbook, and notification configuration. You can create a Standard and Emergency change template.
- **Change Request** is a request created from a change template. You can choose the request operation to start at a scheduled time or once the request has been approved.
- A **Runbook** contains the actions that will be performed for a particular request. Below is an example of a runbook for restarting an EC2 instance.



- **Automation Runbook** is a feature of AWS Systems Manager that automates tasks on AWS resources. It can handle complex workflows, manage errors, and can be customized based on operational needs.
- **Change Request Approval** defines the needed approval for a particular request. The number of approval depends on the levels defined on the template.



- You can add an **Approval Notification** using the SNS topic when configuring a template.

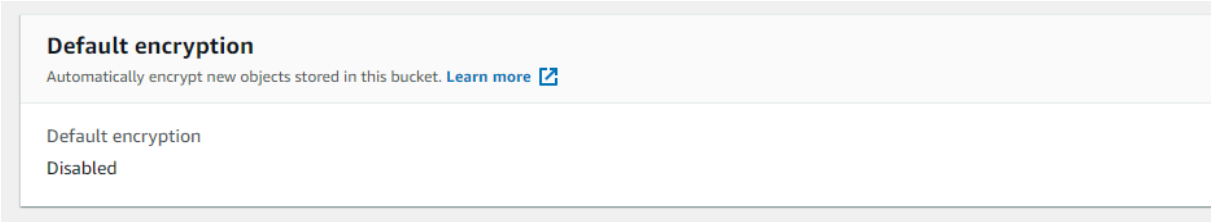


References:
https://docs.aws.amazon.com/systems-manager/latest/userguide/systems-manager-patch.html
https://docs.aws.amazon.com/systems-manager/latest/userguide/systems-manager-actions-and-change.html

Encryption on AWS Storage Services

S3 Encryption

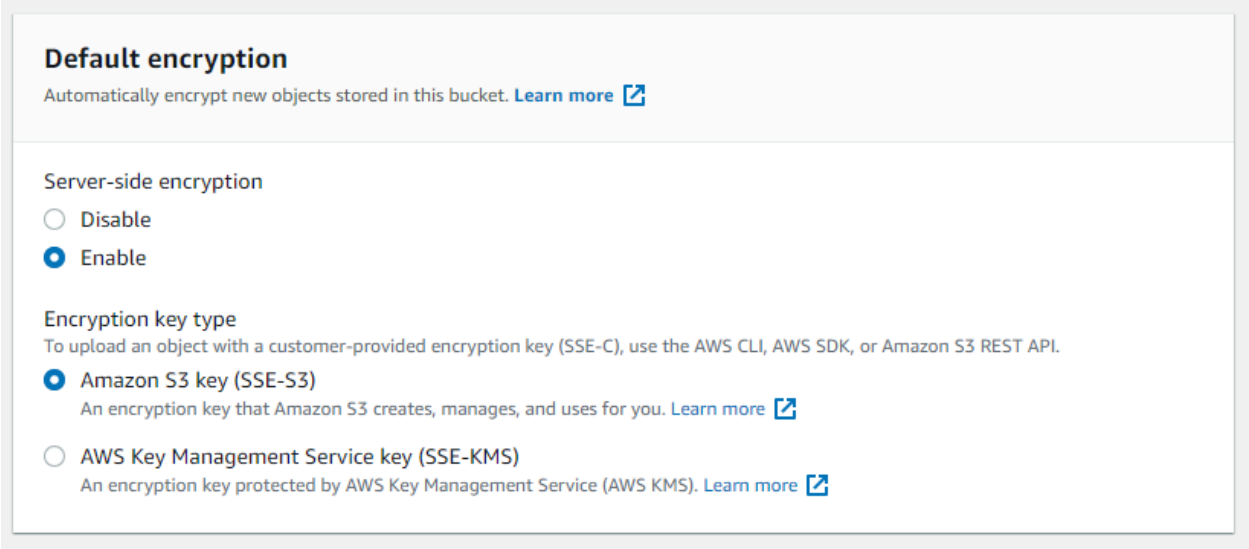
Encryption is another security feature that S3 has as data protection. Users can opt to use Server-Side Encryption or Client-Side encryption, and both have their use cases depending on your storage requirements. S3 default encryption uses SSE and can be set on your bucket properties tab.



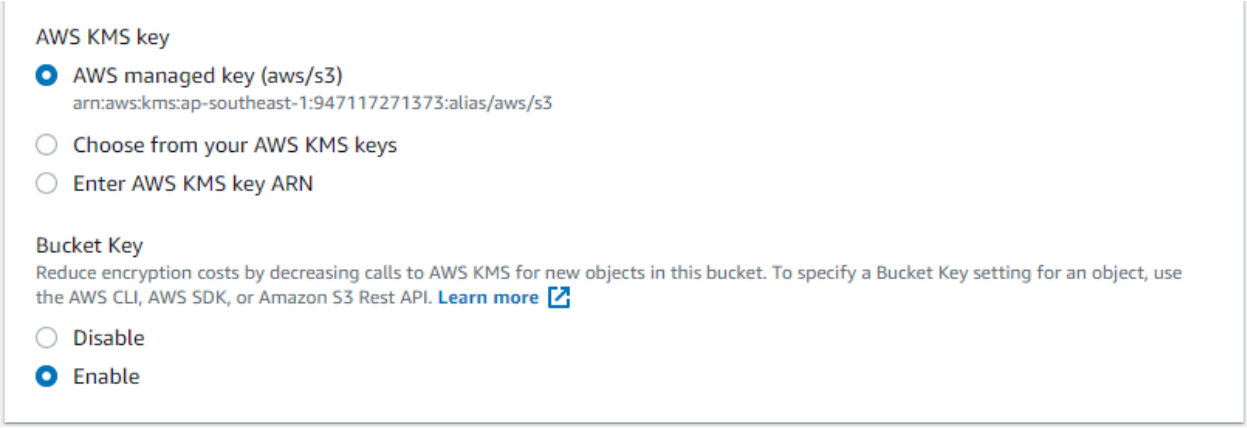
Server-Side Encryption

When using Server-Side Encryption, the new objects uploaded in a bucket are encrypted as it is written on a disk and decrypted when being accessed. There are three different options to use for your encryption keys.

- **SSE-S3** - Encryption keys are managed by the S3 service. AES-256 encryption is used.



- **AWS KMS keys (SSE-KMS)** - Leverage AWS KMS capabilities. Users can use Customer Managed Key stored on AWS KMS as encryption keys. You can create your own keys or use the default encryption for S3. This comes with additional charges since you are using another AWS service. Enabling the Bucket Key feature will reduce the number of calls to AWS KMS to save cost.



- **SSE-C** - Customer-provided keys are used by S3 to encrypt objects.

Client-Side Encryption

While SSE is a great option to encrypt objects in S3, users may also choose to encrypt the objects before uploading to S3. You can use your own keys for encryption or use a Customer managed key from AWS KMS.

Encrypting Existing S3 Objects

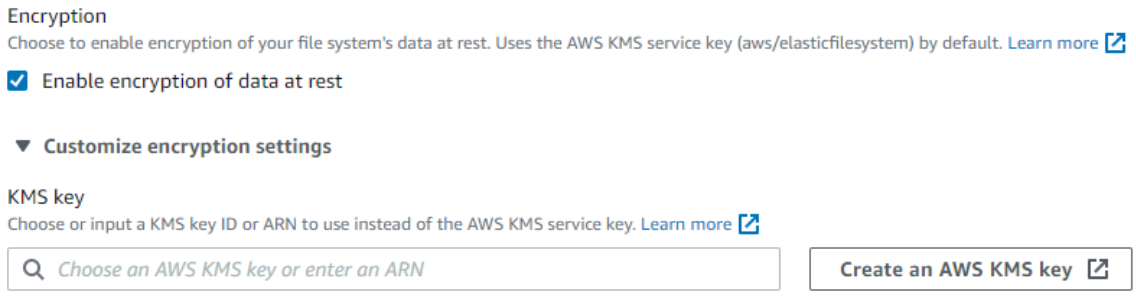
Only new objects uploaded on an S3 bucket are being encrypted when encryption is enabled. To encrypt the existing objects, you can use Amazon S3 Batch operations. This will allow you to identify and copy unencrypted objects, encrypt it, then write it on the same bucket.

EFS Encryption

Like other storage services, EFS also has encryption available for its data. EFS offers encryption both for Data at rest and Data in transit.

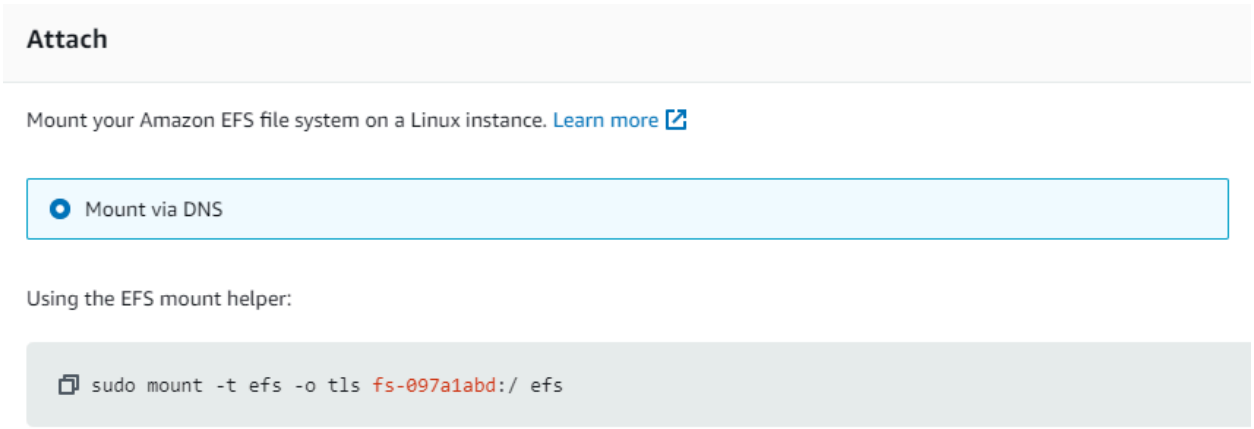
Data at Rest Encryption

You can enable encryption on a filesystem when creating it. Note that once a filesystem is created, you can't modify its encryption settings. EFS also uses KMS service to do the encryption; by default it uses *aws/elasticfilesystemkey* but you can also create your own keys.



Data In Transit Encryption

Using TLS when mounting your filesystem secures your data in transit. You can use the command provided on the EFS mount helper.

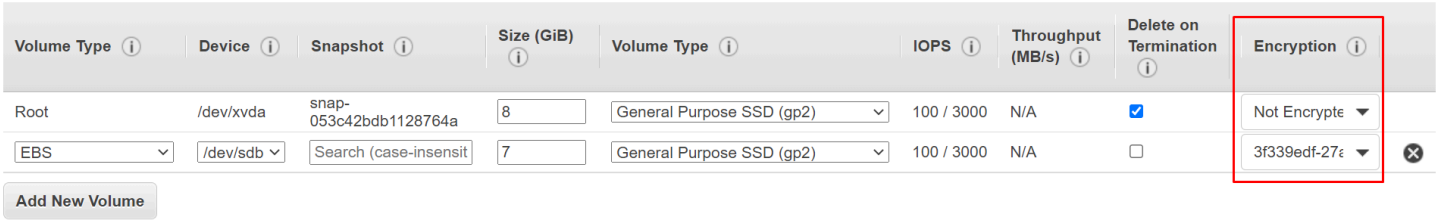


EBS Encryption

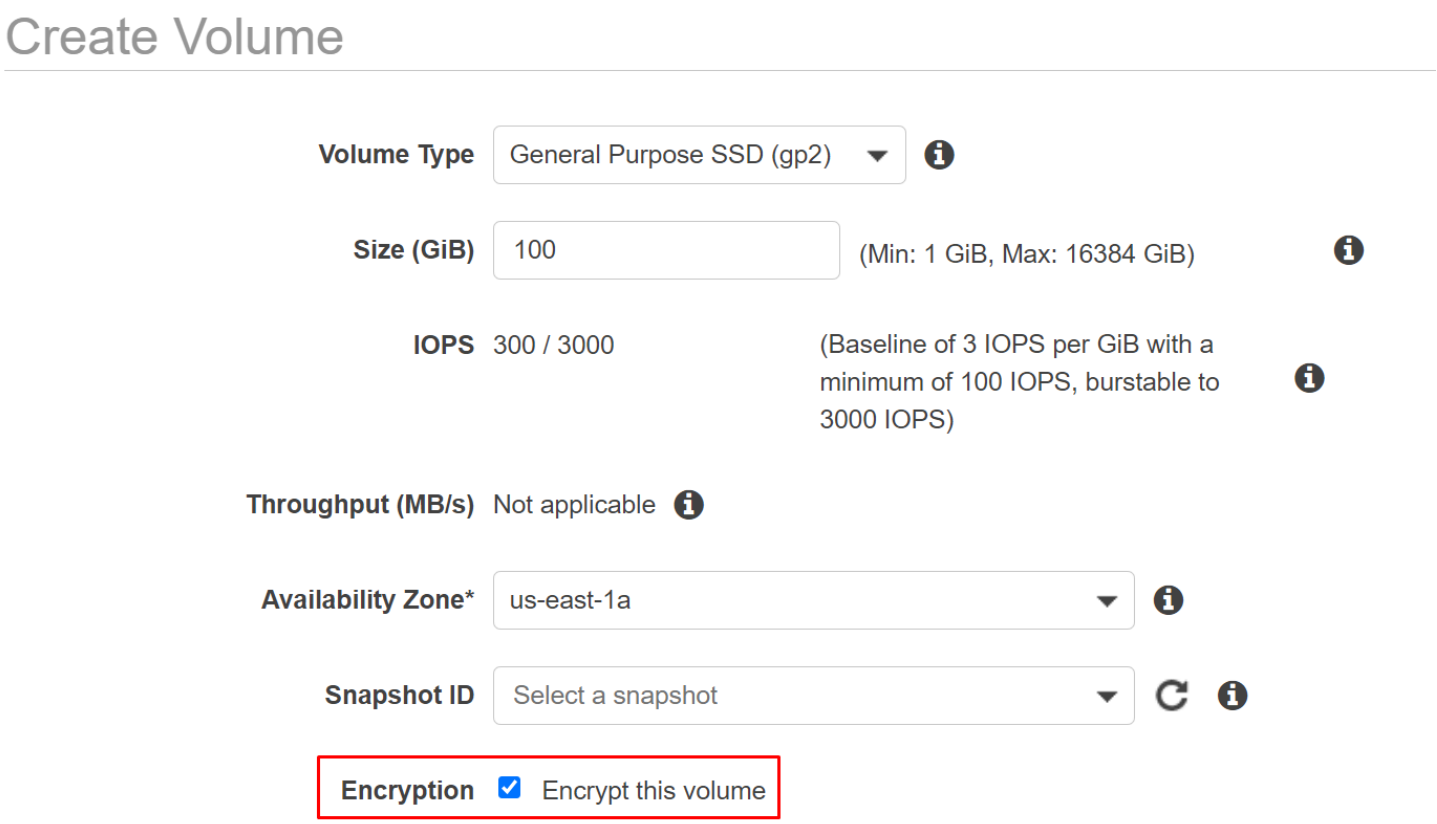
Elastic Block Storage serves as block storage volumes for EC2 instances. EBS has encryption for both data at rest and data in transit. Like other AWS services, EBS utilizes AWS KMS to handle the encryption.

Creating Encrypted EBS Volume

You can enable the volume encryption when launching an AWS instance. Both root and data volume can be encrypted.



Likewise, you can also enable encryption during volume creation.

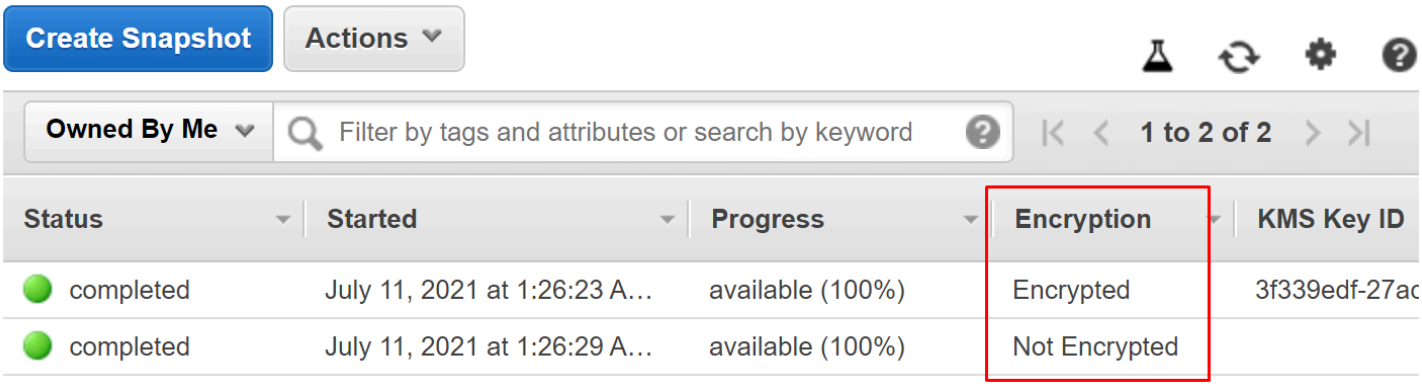


Note that once a volume is created, you can only modify its volume type and size. You can't modify its encryption.

Snapshots

Snapshots are backups of your EBS volumes. The same keys are used in encrypting a volume and its snapshot. When you take a snapshot from an encrypted volume, the snapshot will automatically be encrypted. Same goes with the snapshots of unencrypted volume - the snapshot will also be unencrypted.

Another thing to note. You can create an encrypted volume from an unencrypted snapshot but you can't create an unencrypted volume from an encrypted snapshot.



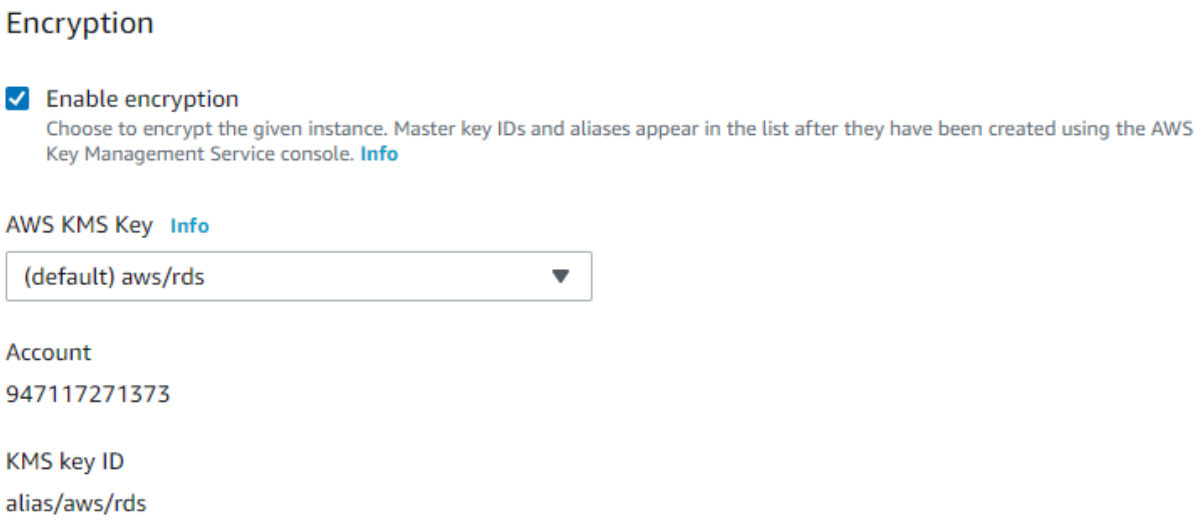
RDS Encryption

Encryption for storage is a must especially for databases. On AWS Relational Database Services, encryption is available both for data at rest and data in transit for its database instances. It uses AES-256 encryption to secure the data on the RDS database instances.

Encrypting RDS Database Instance with AWS KMS

RDS also leverages AWS KMS for the encryption. During database creation, you have an option to enable the encryption and use your preferred key. For an encrypted database instance, all of its data will be encrypted including read replicas, snapshots, and automated backups.

Encryption using Transparent Data Encryption (TDE) is also supported for Oracle and SQL database instances. Take note that using TDE simultaneously with encryption at rest will lead to some performance issues.



Note that encryption is not available for the following database instance classes:

- General Purpose (db.m1.small, db.m1.medium, db.m1.large, db.m1.xlarge)
- Memory Optimized (db.m2.xlarge, db.m2.2xlarge, db.m2.4xlarge)
- Burst Capable (db.t2.micro)

Securing Database Connection on RDS

For data in transit encryption requirements, you can SSL/TLS for the connection between your application and your database instance. A server certificate is used to validate a connection to your database instance; these certificates are also rotated for additional security. The implementation of this method varies depending on the database being used.

References:
https://docs.aws.amazon.com/AmazonS3/latest/userguide/bucket-encryption.html
https://docs.aws.amazon.com/efs/latest/ug/encryption.html
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSEncryption.html
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.Encryption.html

Security on AWS

AWS KMS Key Rotation

AWS Key Management Service (AWS KMS) is a managed service that lets you create and manage KMS Keys. A *KMS Key* is a logical representation of a data key used to encrypt your data, making it a valuable and sensitive asset to your security. It makes sense, and it is a best practice to do a rotation on these keys.

AWS Managed KMS Key has an automatic rotation feature enabled by default. This feature is optional for Customer Managed Key. However, automatic key rotation is not supported by the following:

- **Asymmetric KMS key** where public and private key pairs are used to encrypt/decrypt or sign/verify operations
- **KMS keys in custom key stores** where the key material is stored on the AWS CloudHSM cluster
- **KMS Keys that have imported key material, which is supported only for symmetric KMS key in AWS KMS key stores**
- **Hash-Based Message Authentication Code (HMAC) KMS keys is used for generating and verifying HMAC tags.**

You may not have an automatic key rotation available on these KMS key types, but you can still do a rotation manually.

AWS Owned KMS key Rotation
AWS Owned KMS key is, as the name implies, AWS-owned and managed and is usable by multiple AWS Accounts. The key rotation for AWS-owned KMS key varies depending on the AWS service that uses the key.

AWS Managed KMS key Rotation
AWS Managed KMS Keys are the ones that are created on your behalf. You can't manage the key rotation for AWS Managed Keys. The automatic key rotation is handled by AWS KMS and is automatically rotated every three years.

AWS Customer Managed KMS key Rotation
Unlike AWS Managed KMS key, you have full control over Customer Managed key, including key rotation. By default, Customer Managed key can be rotated every year.

KMS Key Automatic Key Rotation

TYPE OF KMS KEY	Can view key metadata	Can manage key	Used only for your AWS account	Automatic Rotation
Customer Managed KMS key	Yes	Yes	Yes	Optional. Every 365 days.
AWS Managed KMS key	Yes	No	Yes	Required. Every 1095 days.
AWS Owned KMS key	No	No	No	Varies

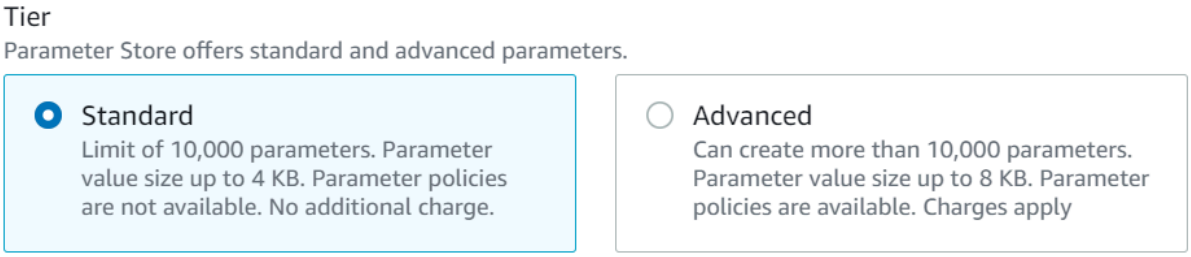
Secrets Manager vs Parameter Store

The best practices are always ideal to follow when storing essential and sensitive application information in the cloud. Information like database credentials, application keys, configuration, or any other security information that your applications consume are business-critical, hence should be secured at any given time. AWS recommends securely storing this sensitive information and only retrieving it when needed to avoid embedding credentials on application source code. Moreover, audit and key rotation are also advisable. AWS offers two relevant services for these requirements, each of them with particular use cases.

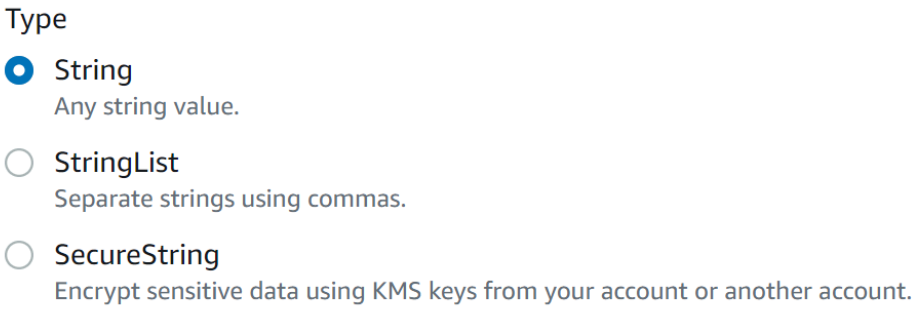
Secrets Manager – securely store, rotate, and monitor secrets. This service leverages AWS KMS for encryption-at-rest, which is enabled by default. Likewise, by default, Secrets Manager only allows programmatic retrieval of secrets over TLS and Perfect Forward Secrecy (PFS). Secrets can be credentials of databases like RDS, DocumentDB, Redshift, or other secrets like API keys.

Parameter Store – a feature of AWS Systems Manager that manages different configuration data like passwords, strings, and parameter values. Information on Parameter Store can be stored on plain text or as encrypted data using AWS KMS.

Parameter Stores offer two tiers:



It also supports the following parameter type:



Here is a side-by-side comparison of Secrets Manager and Parameter Store.

	Parameter Store	**Secrets Manager**
Key-Value Size	4KB/8KB	10KB
Encryption	Yes	Yes
Reference in CloudFormation	Yes	Yes
Built-in password generation	No	Yes
Key Rotation	No	Yes
Cross-account Access	No	Yes
Cost	Free for standard parameters	Yes
Use Cases	Hierarchical storage for passwords, strings, and parameter values. Single store for configuration and secrets	Business-critical secrets like database passwords. Dedicated secrets store with lifecycle management with rotation

		capabilities.

IAM Access Analyzer

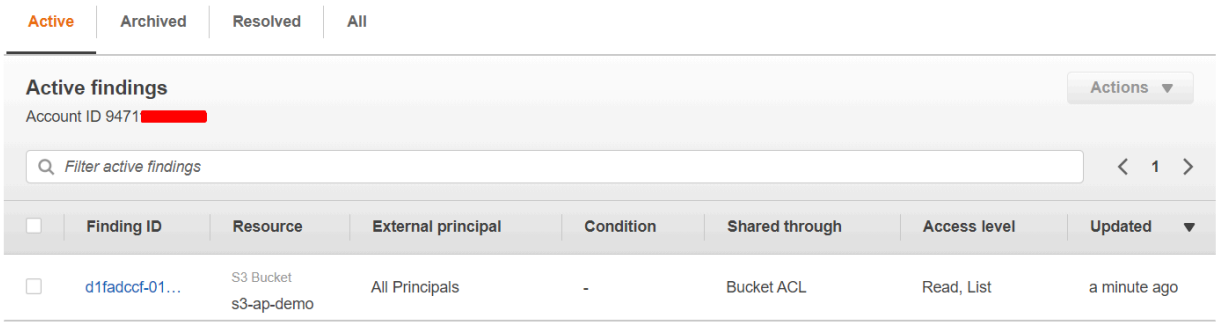
The Access Analyzer is a feature of AWS IAM that evaluates the policies associated with AWS resources. It identifies the resources that are shared with an external entity. Moreover, you can preview your policies before proceeding with creation. IAM Access Analyzer will check your policy grammar and validate against the best practices.This feature is a simple yet effective way to point out unintended access, over-provisioned privileges, and other security-related risks. It helps you maintain the principle of least privilege on your resources.

Supported AWS Services

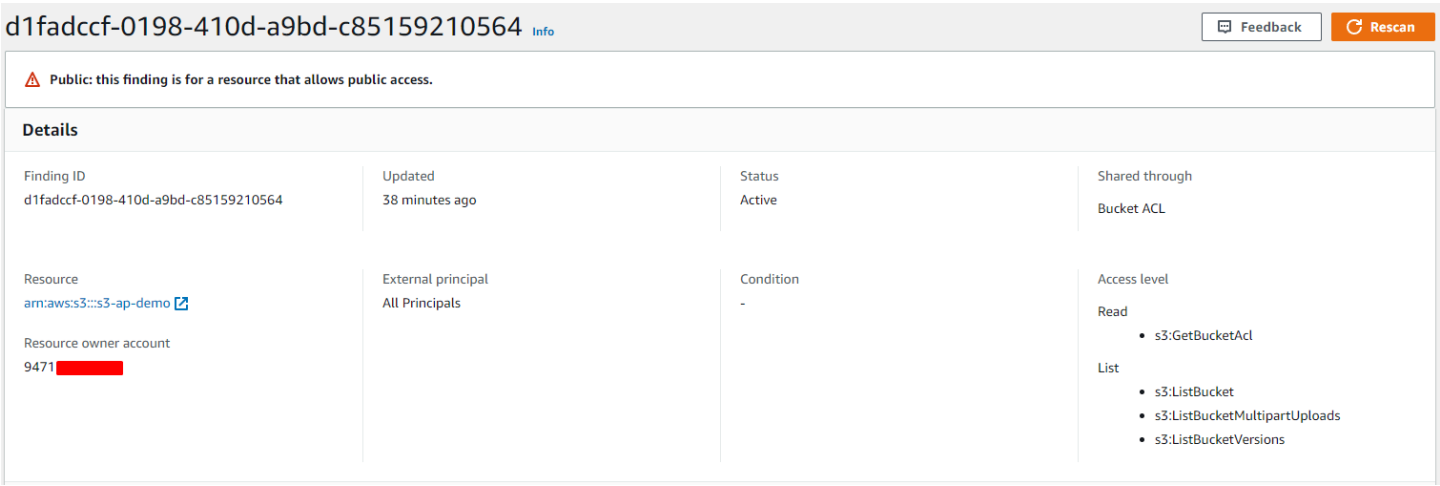
- **Amazon S3** - evaluates the bucket policies, access control list, and access point
- **AWS IAM** - evaluates the trust policies of an IAM role
- **AWS Lambda** - evaluates the policies associated with the functions
- **Amazon SQS** - evaluates the policies associated with a queue
- **AWS KMS** - evaluates the key policies and grants of keys
- **AWS Secrets Manager** - evaluates the policies associated with a secret

The following statements summarize the process of Access Analyzer.

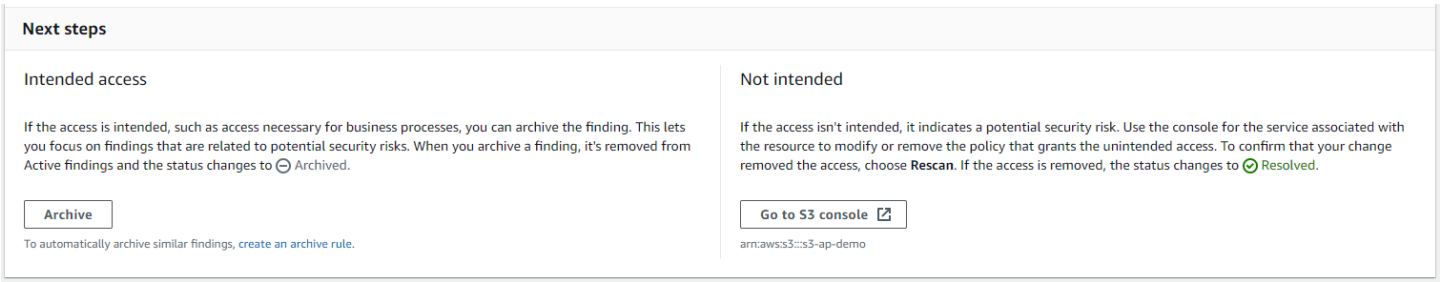
1. Define the scope or zone of trust. The zone of trust will serve as a basis of the Access Analyzer; scope can be either within an AWS Account or Organization. Any entities outside the scope are treated as an external entity; this can be an AWS Account, IAM Account/Role, public entities.
2. Scan all resources policy within the trust zone and determine the resources shared outside the defined trust zone/scope.
3. Create findings to evaluate if access to a particular resource is intended or not.



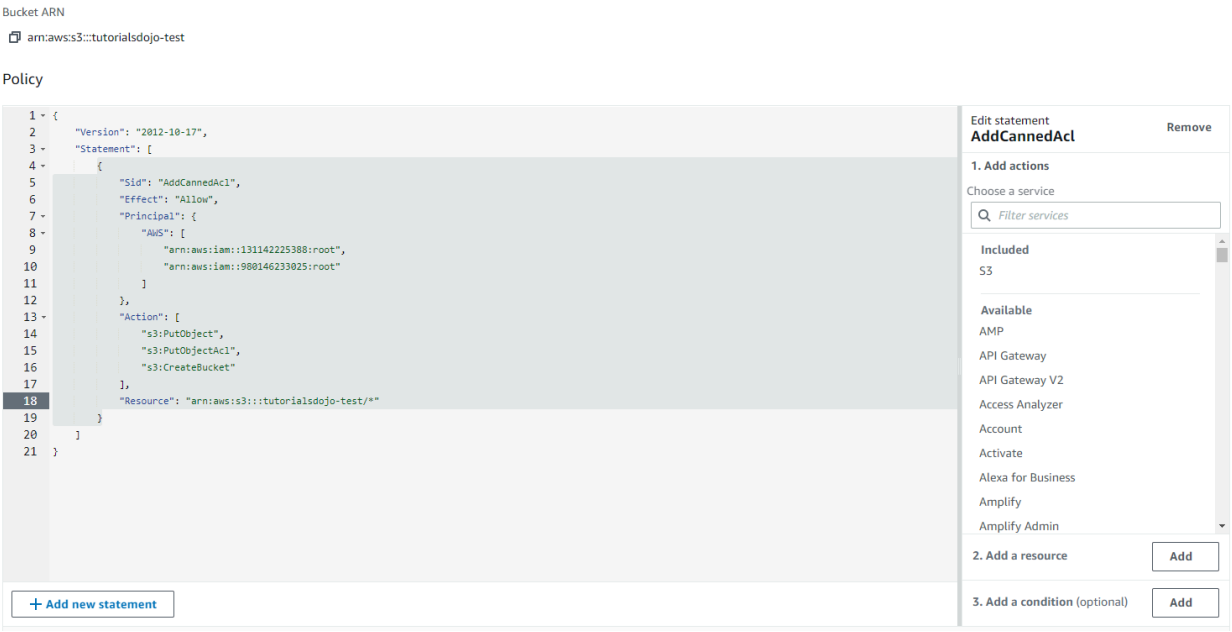
Here is an example of an active finding on an S3 bucket that is publicly exposed. It shows essential details like the ARN, the policy, and the access level.



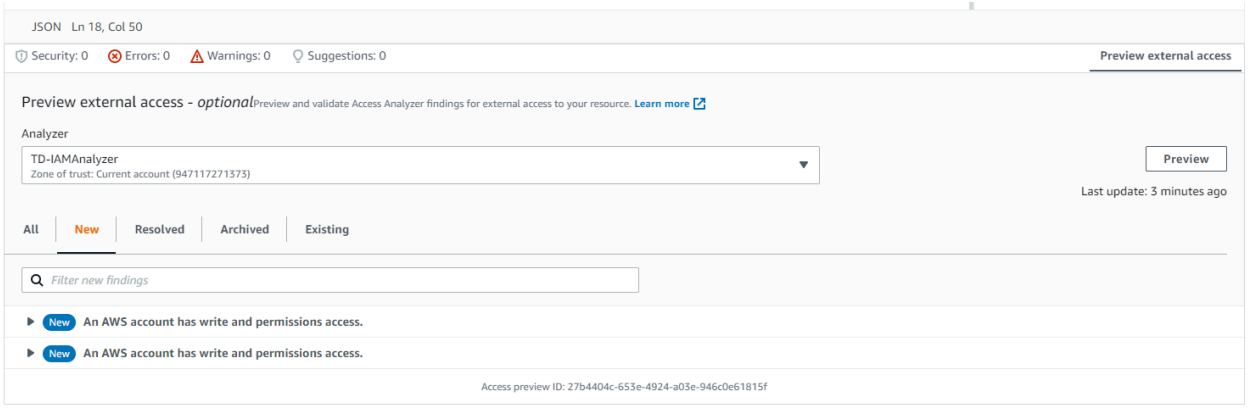
IAM Access Analyzer allows you to archive (for intended access) the findings or apply the necessary fix (not intended access) to the resource.



Another popular use of the IAM Access Analyzer is when creating a bucket policy. IAM Access Analyzer will interactively check the syntax/grammar of the policy being created.

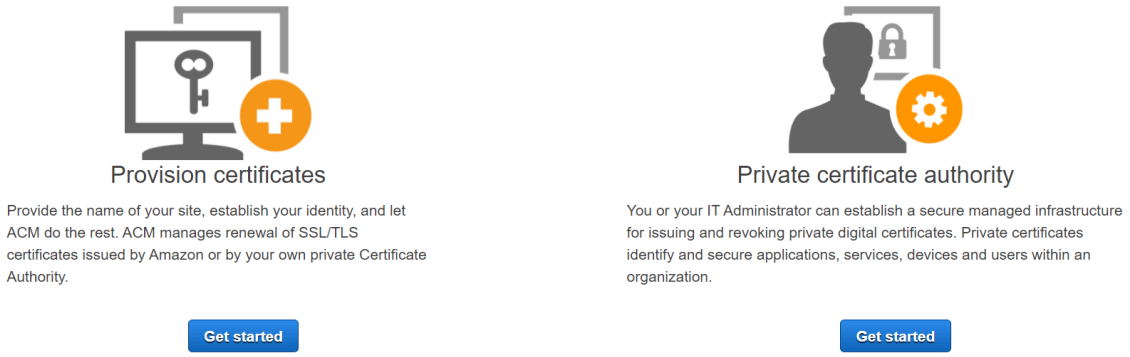


Using the Preview feature will trigger the IAM Access Analyzer to scan your policy and provide recommendations according to security best practices. See sample findings below.



AWS Certificate Manager

The AWS Certificate Manager (ACM) helps you easily provision and manage keys and certificates for your application or websites. With ACM, you can create both public and private SSL/TLS X.509 certificates.

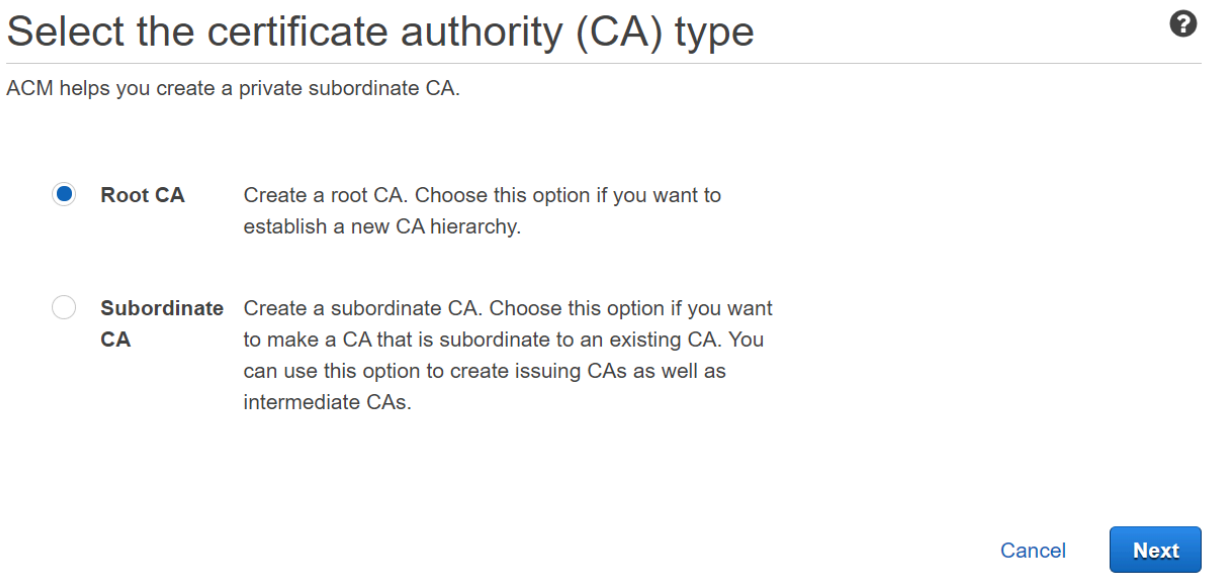


AWS Certificate Manager (ACM) allows you to secure your public websites and applications over TLS. You can request new certificates with just a few clicks. You can even request certificates for multiple domains and wildcard domains. Before ACM issues the requested certificates, the request is validated through DNS or Email validation. ACM also handles the auto-renewal of certificates. If you want to use existing certificates, you can do so by importing them to ACM.

The ACM Certificates are free to use and can be integrated into the following AWS services.

- Elastic Load Balancing
- Amazon CloudFront
- AWS Elastic Beanstalk
- AWS App Runner
- Amazon API Gateway
- AWS Nitro Enclaves
- AWS CloudFormation

ACM Private CA allows you to create your certificate authority (CA) hierarchy in AWS. Certificates generated from ACE Private CA can only be privately used and accessible via ACM Private CA API or the AWS CLI. The ACM Private CA isn't free to use; you pay a monthly fee for each CA and certificate that you create and issue.



Amazon Detective

Amazon Detective is designed to simplify and expedite the investigation of potential security issues. The service automatically collects log data from AWS resources and employs machine learning, statistical analysis, and graph theory to construct a comprehensive dataset. This dataset, presented through interactive visualizations, enables security teams to conduct efficient and detailed investigations. Amazon Detective's prebuilt data aggregations, summaries, and context provide a unified view of user and resource interactions. This helps to quickly identify and understand the nature and extent of potential security issues. Its capabilities include triaging security findings, investigating incidents with interactive visualizations, and tracking threats. This service is a powerful tool for any organization looking to enhance its security posture in the cloud.

Amazon GuardDuty

Amazon GuardDuty is designed to safeguard your AWS accounts and workloads. It continuously monitors your environment, examining various data sources like AWS CloudTrail management events, VPC flow logs, and DNS logs. The service uses machine learning and threat intelligence feeds to spot unusual or potentially harmful activities. This could include identifying compromised EC2 instances, detecting abnormal login patterns, or spotting communication with suspicious IP addresses. By providing detailed security findings, GuardDuty empowers you to take action and enhance the security of your AWS environment.

AWS Firewall Manager

AWS Firewall Manager is a service that simplifies your security tasks by enabling you to manage and organize firewall rules from a single, central location within your AWS Organization. This means you can easily set up and maintain security protocols across all your accounts and applications. As your organization grows and you build more VPCs and accounts, you can quickly propagate your security and compliance policies across these

new resources. This ensures all your security rules are enforced consistently, even as new resources are added. AWS Firewall Manager simplifies your security management and reduces the risk of misconfigurations by providing a single place to set up and maintain your rules. It integrates with AWS WAF, AWS Shield Advanced, and Amazon VPC security groups, enabling you to use these services more effectively to protect your AWS environment.

AWS Directory Service

AWS Directory Service enables the integration of AWS resources with an existing Microsoft Active Directory located on-premises or creating a new, standalone directory in the AWS Cloud. It provides four directory types:

1. AWS Directory Service for Microsoft Active Directory (Available in Standard Edition or Enterprise Edition): It is capable of supporting workloads aware of Active Directory, such as Remote Desktop Licensing Manager, Microsoft SharePoint, and Microsoft SQL Server Always On in the AWS Cloud. It is also suitable for use with AWS applications and services like Amazon WorkSpaces and Amazon Quick, or if there is a need for LDAP support for applications based on Linux.
2. AD Connector: This is suitable if your requirement is solely to enable users from your on-premises setup to authenticate into AWS applications and services using their Active Directory credentials. Additionally, AD Connector can be utilized to associate Amazon EC2 instances with your pre-existing Active Directory domain.
3. Simple AD: This is suitable if you're looking for a directory that is cost-effective and operates on a smaller scale, offering fundamental compatibility with Active Directory. It supports applications compatible with Samba 4 and provides LDAP compatibility for applications that are LDAP-aware.
4. Amazon Cognito: This is ideal for large-scale SaaS applications that require a directory capable of managing and authenticating users, and it is compatible with social media identities.

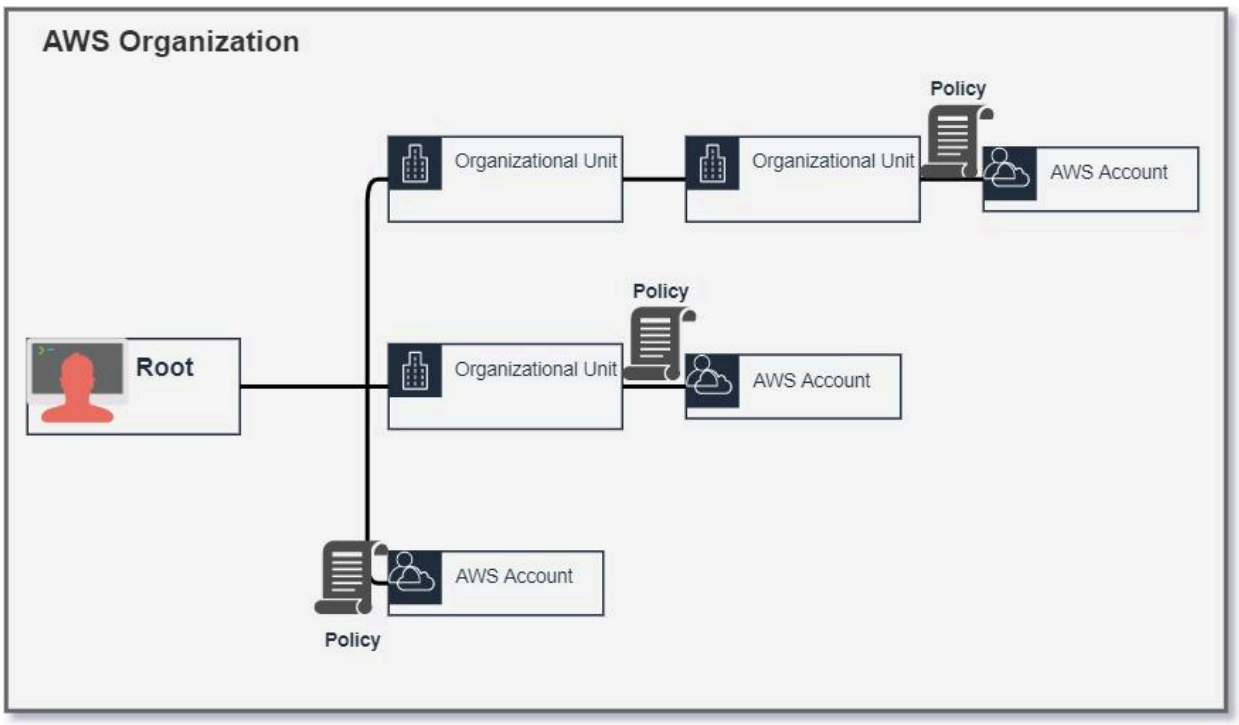
By using AWS Directory Service, you can decrease the time required to establish and operate a directory in the AWS Cloud, integrate your AWS resources with your existing infrastructure seamlessly, and simplify your hybrid cloud deployments. It also offers multiple features, such as access control, auditing, and scaling, to accommodate high-performance workloads. AWS Directory Service is a crucial component for enterprises that are transitioning their workloads to the AWS Cloud as it enables them to leverage single sign-on to AWS applications and services.

References:
https://docs.aws.amazon.com/kms/latest/developerguide/concepts.html
https://docs.aws.amazon.com/secretsmanager/latest/userguide/intro.html
https://docs.aws.amazon.com/systems-manager/latest/userguide/systems-manager-parameter-store.html
https://docs.aws.amazon.com/IAM/latest/UserGuide/what-is-access-analyzer.html
https://docs.aws.amazon.com/acm/latest/userguide/acm-overview.html
https://tutorialsdojo.com/aws-secrets-manager-vs-systems-manager-parameter-store/
https://docs.aws.amazon.com/detective/latest/adminguide/what-is-detective.html
https://docs.aws.amazon.com/guardduty/latest/ug/what-is-guardduty.html
https://aws.amazon.com/firewall-manager/
https://docs.aws.amazon.com/directoryservice/latest/admin-guide/what_is.html
https://aws.amazon.com/blogs/security/introducing-aws-directory-service-for-microsoft-active-directory-standard-edition/

AWS Billing and Governance

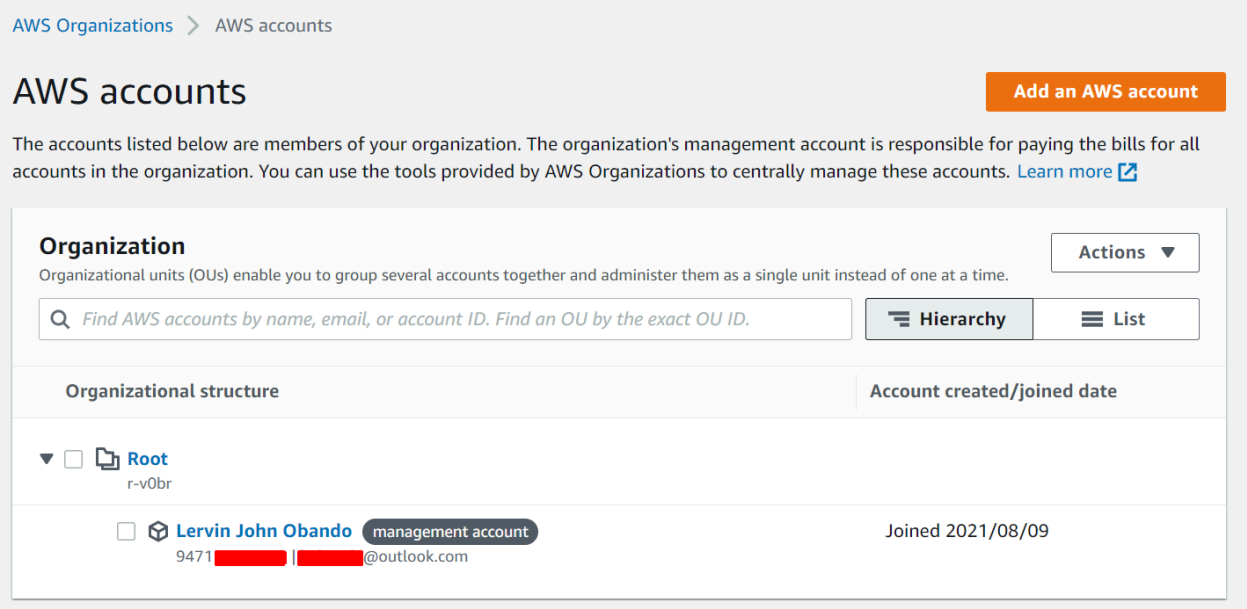
AWS Organizations

AWS Organizations' purpose is to ease the management of multiple AWS Accounts by consolidating them into a group called Organization.

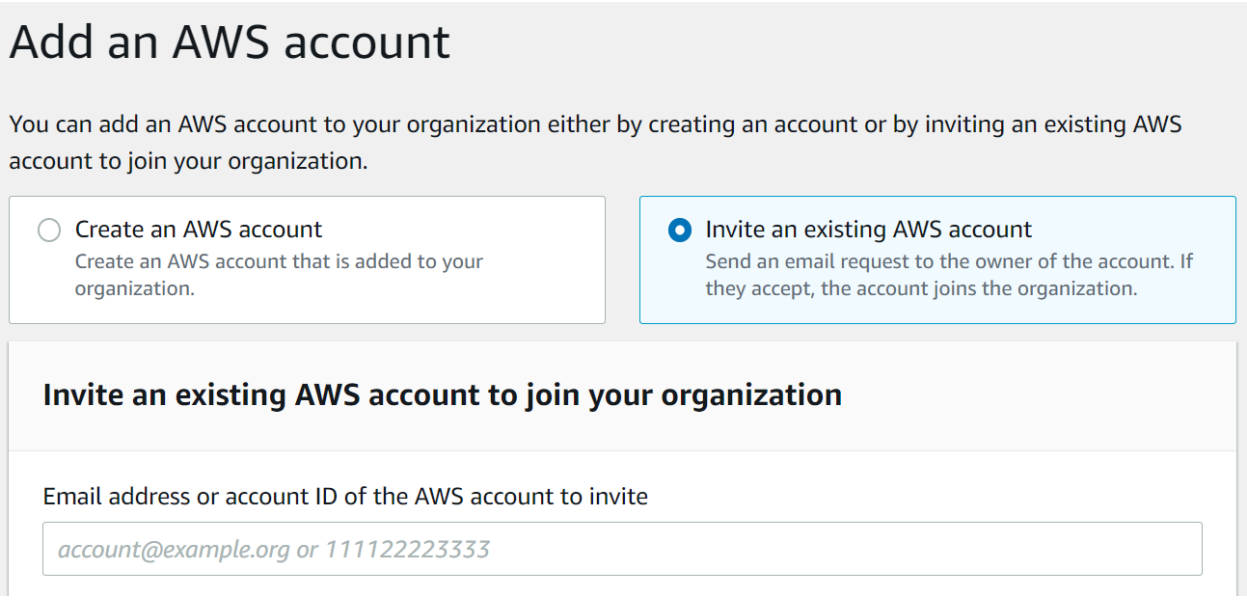


Customers mainly benefit from AWS Organizations in terms of billing and account management. Instead of paying bills from each AWS Account, customers can consolidate and pay all bills on their AWS Master Account. Consolidated billing allows the customers to centrally manage their cloud expenditure while saving costs through different pricing discounts.

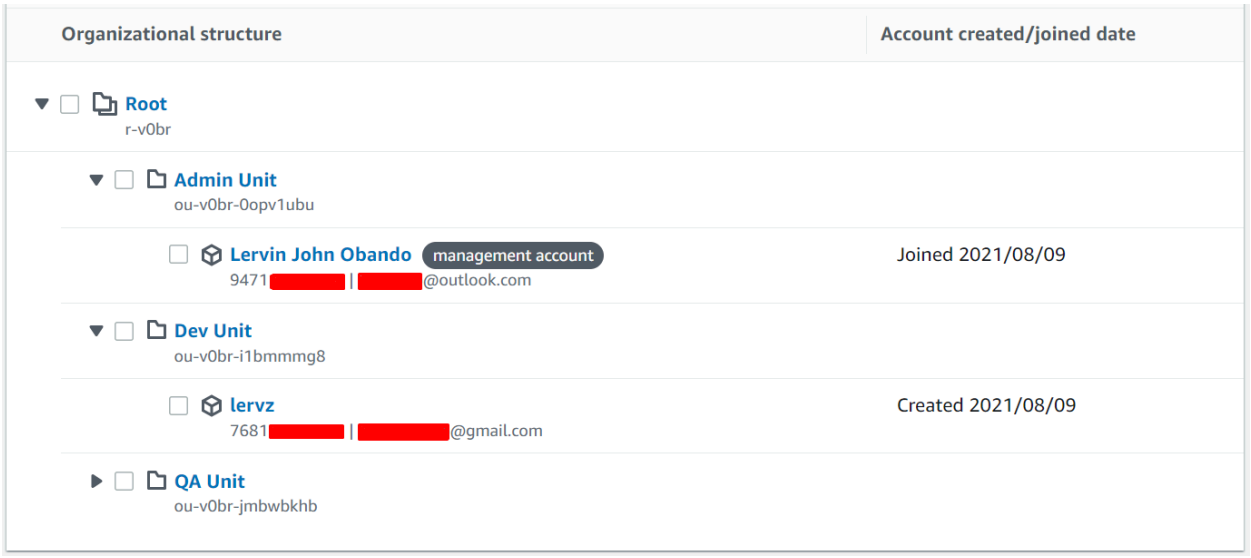
You can centrally manage all of your AWS Account from your Master Account. When you enable AWS Organizations on your account, it automatically becomes the Master Account; an email verification is also sent to your AWS Account email.



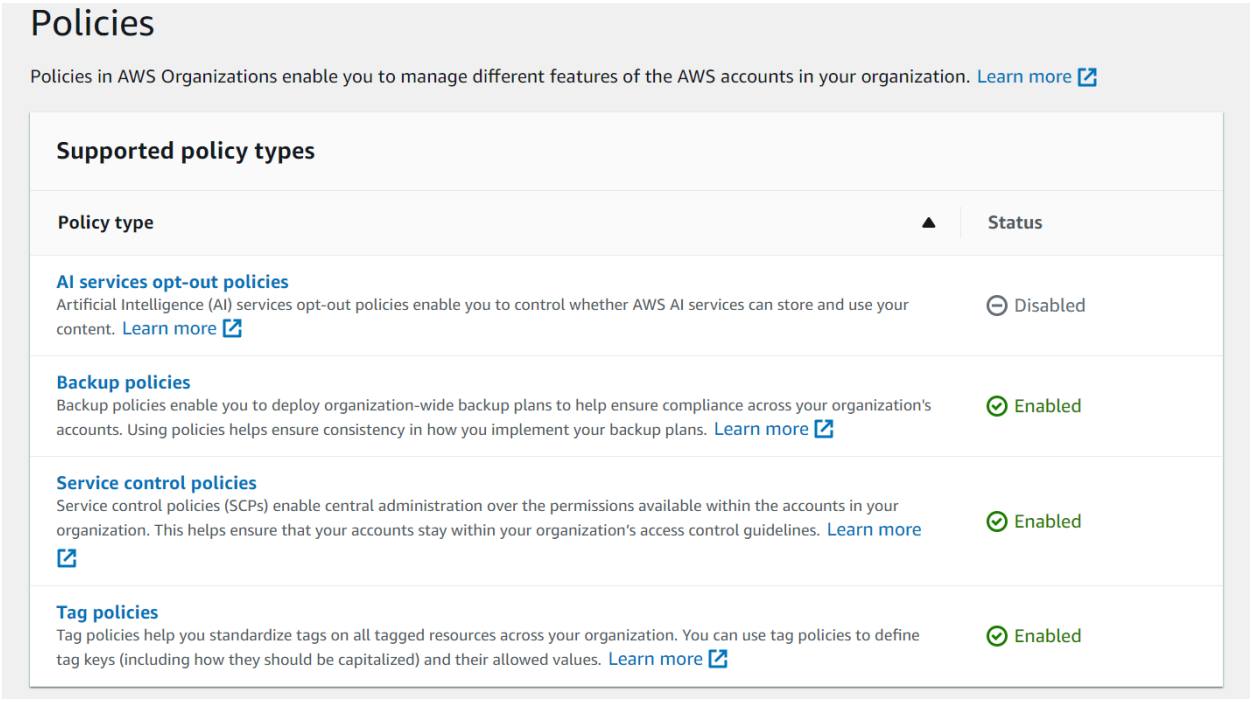
You can then create an AWS account or invite an existing AWS Account for your Organization; the added AWS accounts are considered member accounts. You can invite an existing AWS Account by specifying the AWS Account ID or the email address.



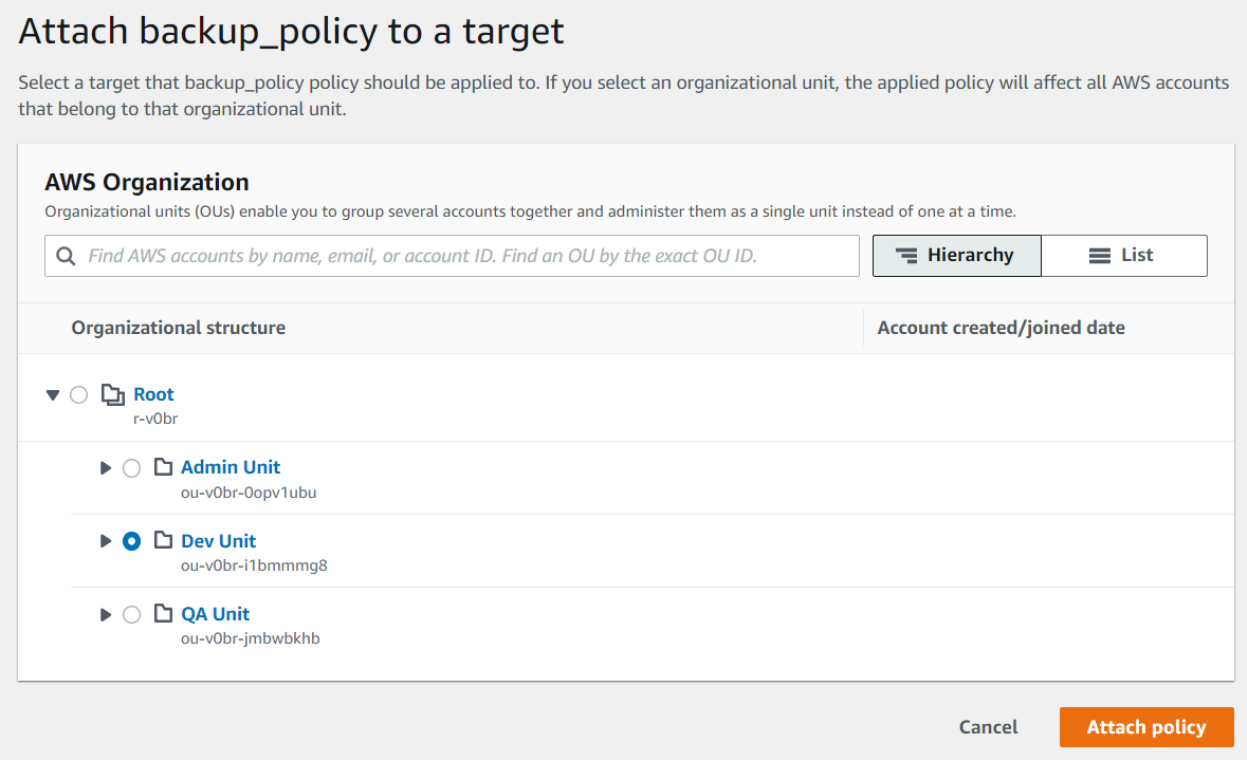
You can also have a hierarchical grouping on your AWS Organizations. You can do this by creating Organizational Units (OU), a logical group, or container for your accounts. OUs may either have an AWS Account or another OU into it. You can also freely move an AWS Account at any level in your Organization.



You can attach policies at any level on your Organizational hierarchy as another management capability. Organization policies are disabled by default, but you can explicitly enable them. Below are the supported policy types for the AWS Organization.



When you attach a policy to one of the hierarchy nodes, all Organizational entities (root, Organizational unit (OU), or account) beneath it will inherit that policy. Here's an example of a Backup Policy being attached to an Organizational Unit (OU).

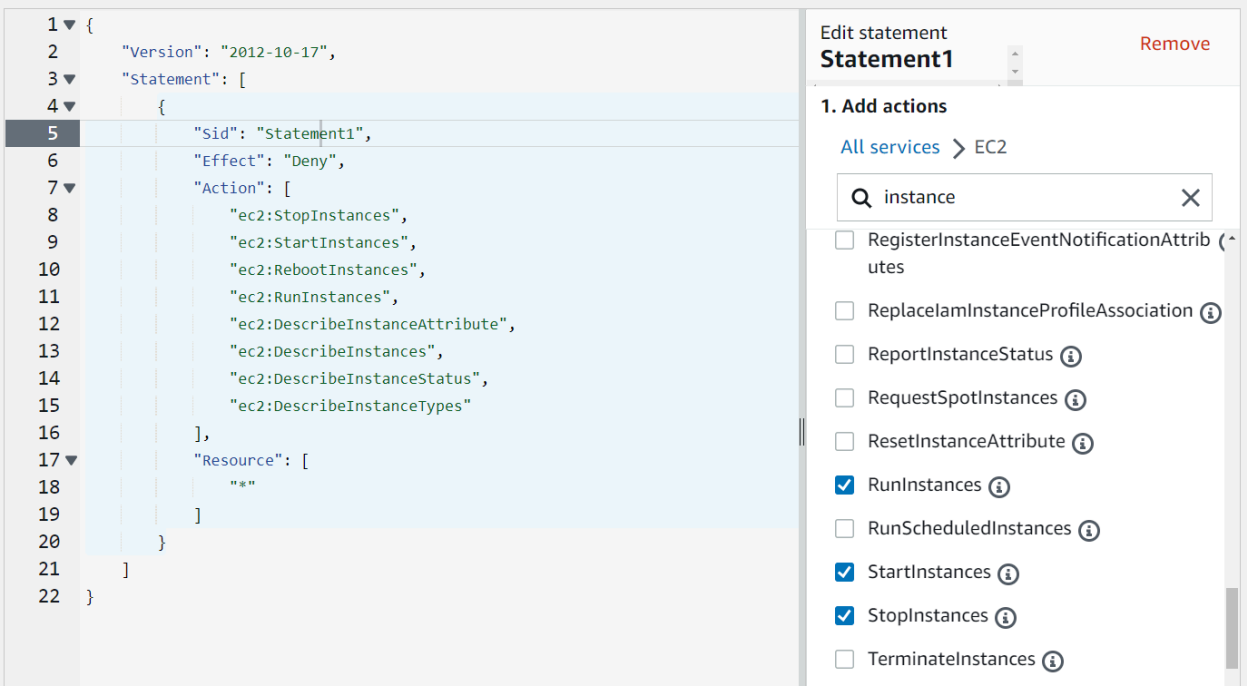


Service Control Policies (SCP)

Service control policies (SCPs) are a type of organization policy that you can use to manage permissions in your organization. SCPs are similar to IAM permissions policies except that they don't grant any permissions. Instead, SCPs specify the maximum permissions for an AWS Organizations entity. Attaching an SCP to your organization root or an OU limits the permissions of all the entities beneath it; this includes the root user of any member account.

Since SCPs don't grant any permissions, you still need to attach identity-based or resource-based policies to IAM users or roles or the resources in your organization's accounts to grant permissions.

When creating a service control policy, you make a statement that identifies the maximum permission of an entity to a particular AWS Service. Below is an example of an SCP that limits an account's action to simple EC2 instance management.



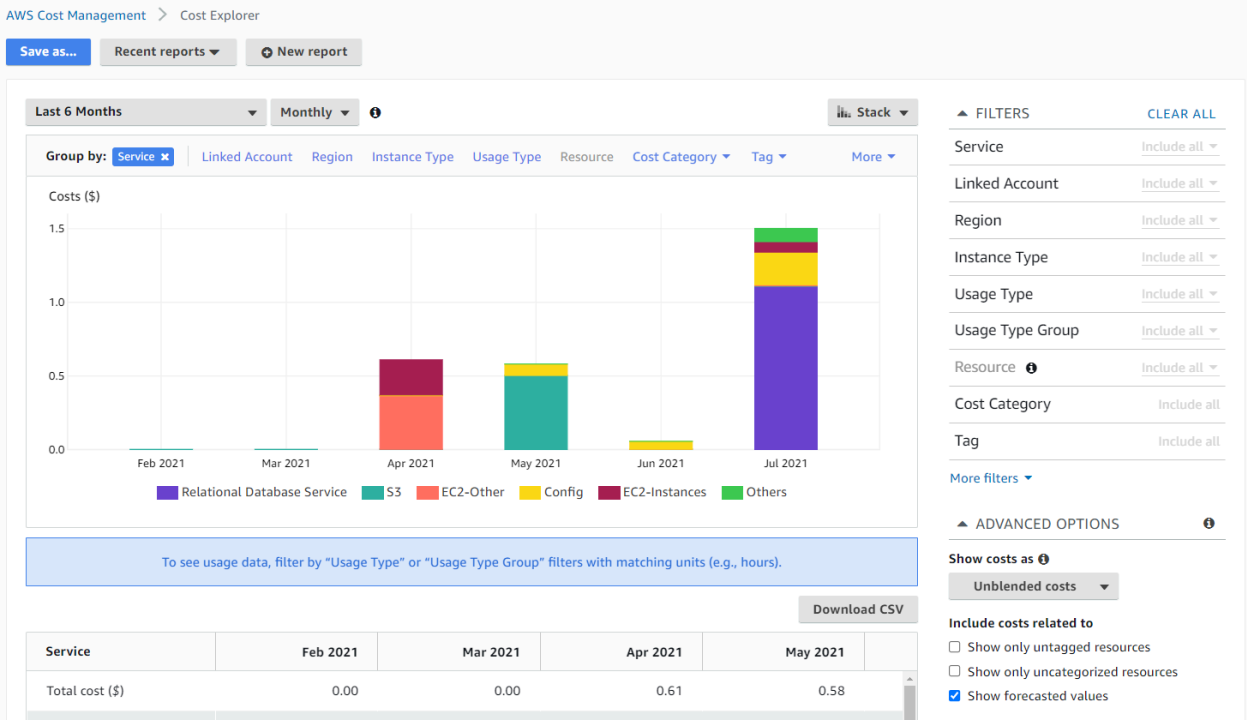
Cost Explorer

The Cost Explorer allows customers to view and analyze their consumption on AWS. This service can help you identify usage patterns like most used services and unusual high service usage, which are crucial for analyzing and maintaining a reasonable cloud expenditure. You can create reports from the Cost Explorer. Below are the types of reports that are supported.

- Cost and usage (recommended)
- Savings Plans reports (Savings Plans utilization, Savings Plans coverage)
- Reservation reports (Reservation utilization, Reservation coverage)

The Cost Explorer presents the costs data into a dashboard graphical report, making it easy to visualize your AWS usage. The dashboard is pre-configured but is customizable if you want to modify the view; it also provides a data filter to help you navigate your data. Data can also be visualized according to groups like Service, Regions, and Allocation Tags.

The Cost Explorer can present up to 12 months' worth of data and give a billing forecast according to your usage and billing patterns for the past months. Furthermore, it can also give a Reserve Instance purchase recommendation. Reports are downloadable as a CSV file.

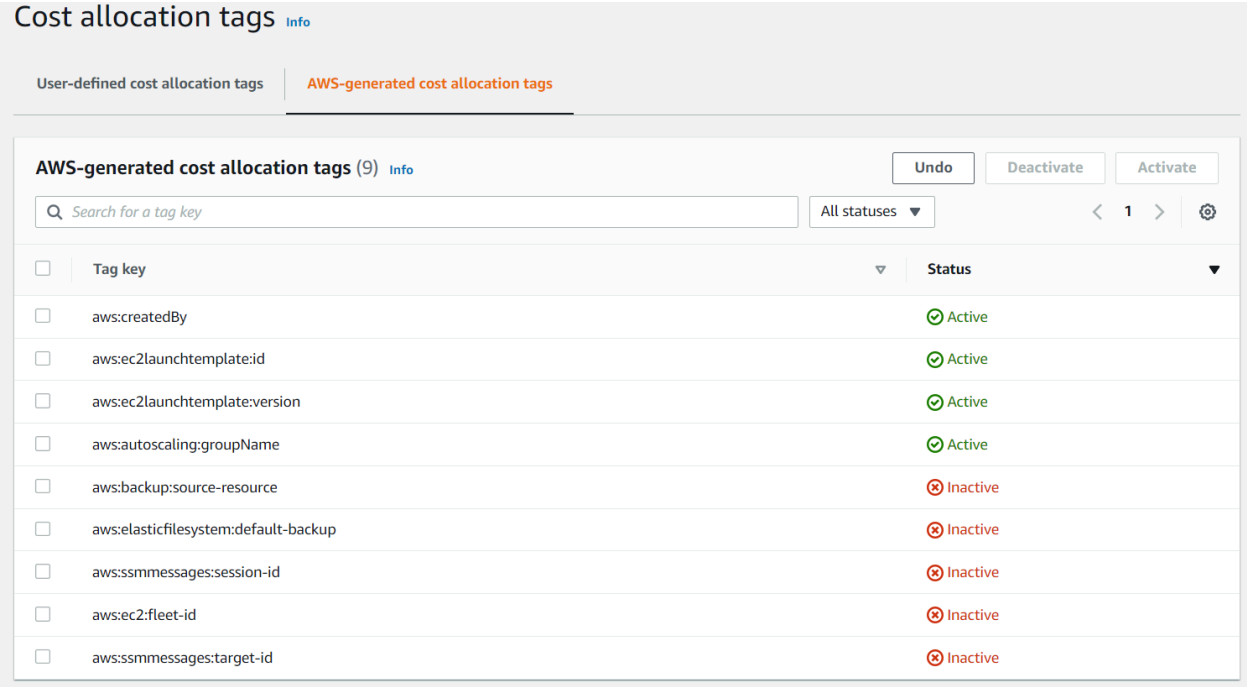


Cost Allocation Tags

The Cost Allocation Tags is another effective way to organize and track your costs in your AWS cloud environment. You can enable Cost Allocation Tags on your AWS Billing console. AWS associates your resources with cost allocation tags to organize your cost allocation report.

There are two types of Cost Allocation Tags that you can simultaneously use.

- **User-defined cost allocation tags** - tags you created for your resources
- **AWS-generated cost allocation tags** - AWS-created and managed cost allocations tags



AWS Cost and Usage Report

AWS Cost and Usage Report is a powerful tool that gives users detailed insights into their usage of AWS resources and services. It offers detailed information about the expenses associated with various AWS services and the patterns of their use. This allows companies to gain a clearer understanding of their expenditures, manage their financial plans more efficiently, and fine-tune the distribution of resources. Utilizing the AWS Cost and Usage Report, organizations can formulate cost-effective strategies, improve financial transparency, and make decisions based on data regarding their AWS investments. It's an essential tool for maintaining control over cloud expenditure and ensuring the efficient use of AWS services.

The AWS Cost and Usage Report (CUR) is created by gathering detailed billing and usage data from all your AWS services. To set up the CUR, you need to enable it in the AWS Billing and Cost Management console and specify an Amazon S3 bucket where the report will be stored. The CUR provides updates based on your chosen configuration: daily, hourly, or monthly. Each update creates a new file in the specified S3 bucket, capturing the latest usage and cost data in either CSV or Parquet format. The report includes detailed information such as product usage, operation type, cost allocation tags, and account-level details, which are crucial for thorough financial analysis and cost management. The stored data can be queried using Amazon Athena, visualized with Amazon Quick.

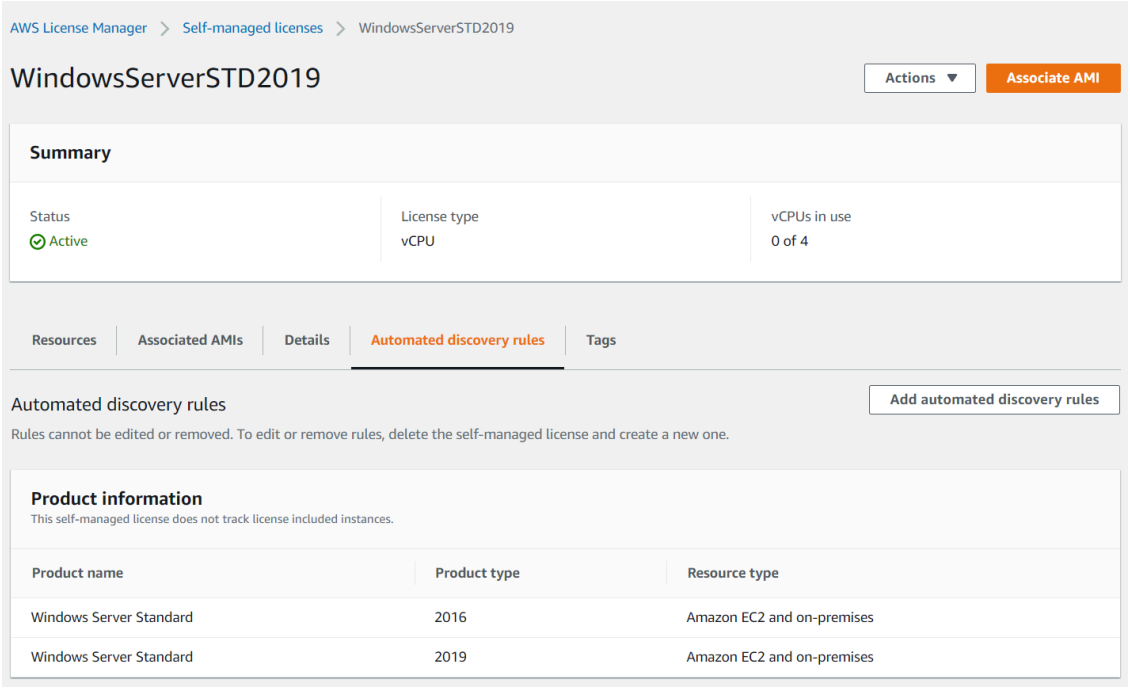
AWS License Manager

When you have multiple software licenses, tracking and managing them can be challenging, primarily if numerous vendors have issued them. With AWS License Manager, you can have your licenses centralized into one single plane, giving you a better view and control. Its scope can be extended across all AWS Organization Accounts, enabling cross-account license inventory.

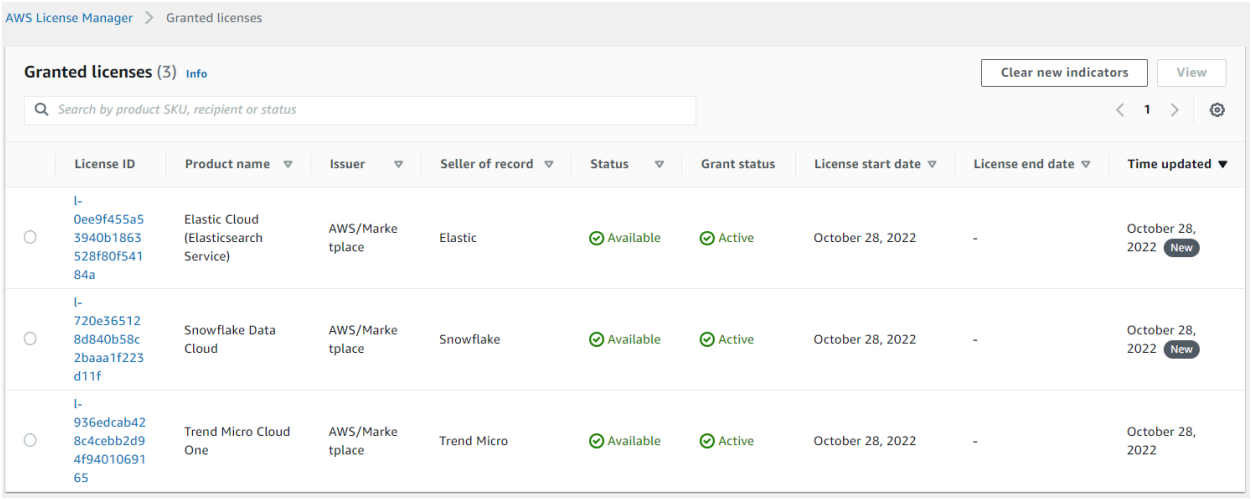
Here are the following notable features of the AWS License Manager.

Manage License from multiple sources

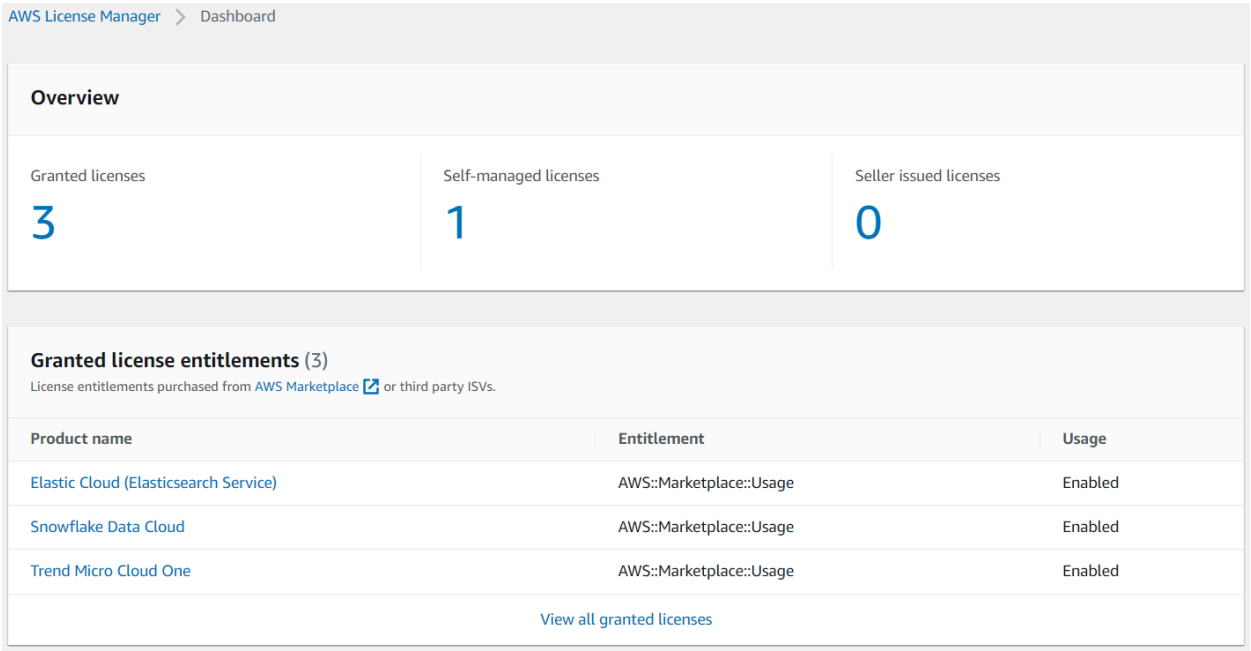
Self-managed licenses - create/manage your software license (BYOL). When creating a self-managed license, you define the software license type (vCPUs, Cores, Sockets, Instances). Optionally, you can create Automated discovery rules to automatically discover and track the license of the software you defined from your instances. Once created, you can associate it with an AMI.



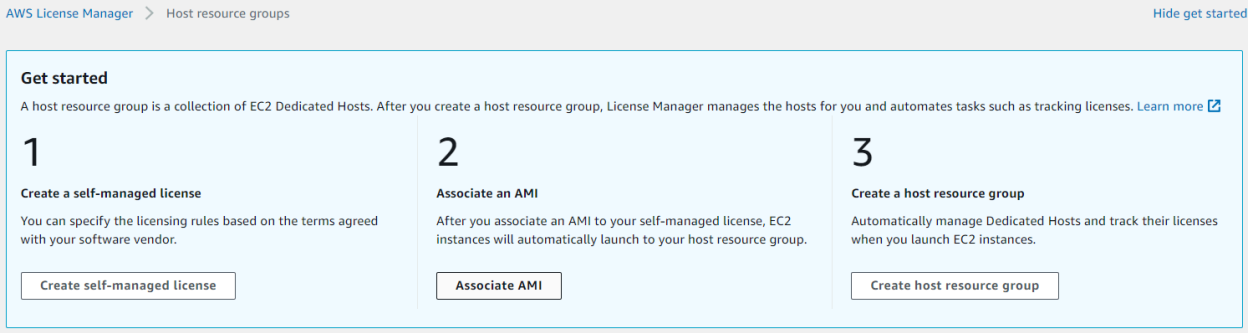
Granted License - These licenses are acquired from the AWS Marketplace or an independent software vendor (ISV).



AWS License Manager Dashboard - consolidated view of all licenses across the account/s



Hosted Resource Group - tracks the licenses from a collection of Dedicated hosts using the defined Self-managed license associated with an AMI.



License type conversions - allows a customer to switch License type from BYOL to AWS-provided license and vice versa.

References:

https://docs.aws.amazon.com/organizations/latest/userguide/orgs_introduction.html
https://docs.aws.amazon.com/organizations/latest/userguide/orgs_manage_policies_scps.html
https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/ce-what-is.html
https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/cost-alloc-tags.html
https://docs.aws.amazon.com/license-manager/latest/userguide/license-manager.html
https://aws.amazon.com/aws-cost-management/aws-cost-and-usage-reporting/
https://docs.aws.amazon.com/cur/latest/userguide/what-is-cur.html

Monitoring and Logging on AWS

Amazon CloudWatch is a monitoring tool that you can use for various AWS resources. With CloudWatch, you can collect metrics, logs, and apply event-driven actions using alarms and events.

CloudWatch Metrics for EC2

Amazon EC2 integrates with CloudWatch to monitor and collect performance data from EC2 instances called metrics. It can also collect metrics of the application hosted in the EC2 instance and on-premises server using **CloudWatch Agent**. Amazon CloudWatch consumes the default metrics available for your EC2 instances, but you also have the capabilities to publish your own custom metric. Here's a comparison of the default metrics (AWS-provided) and a custom metric.

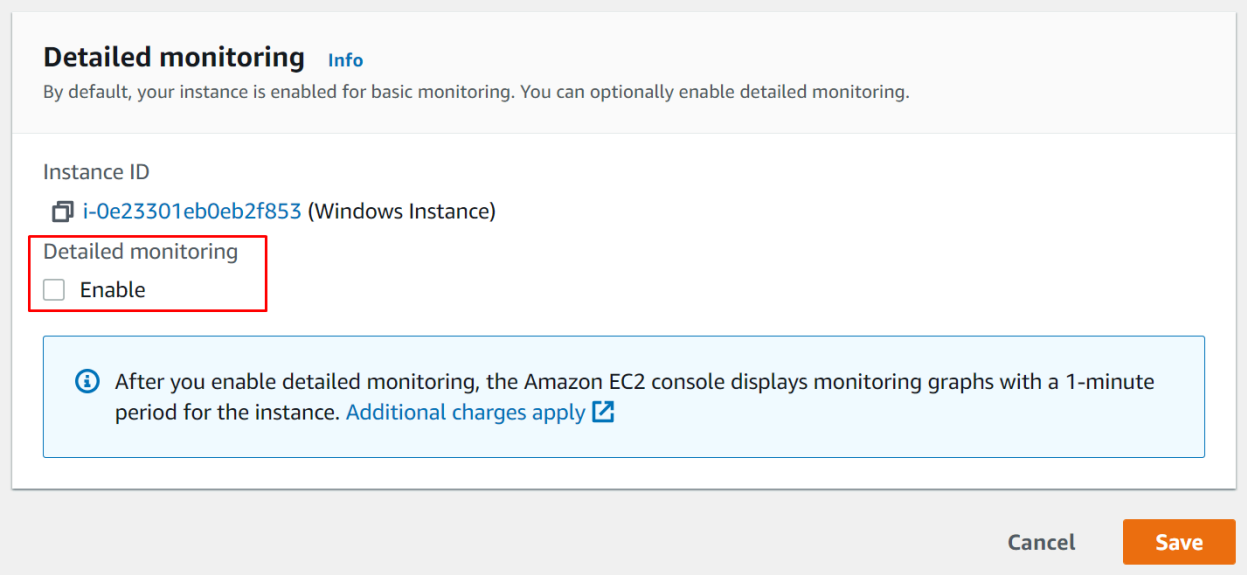
	Monitoring Interval	Supported Metrics
Default EC2 Metrics	5 minutes - basic monitoring 1-minute - detailed monitoring	CPU, Network, Disk, Status check
Custom Metrics	1 minute down to a second	Memory, Application Metrics

Metric resolution is simply the granularity of the data being collected. Below are the two types of metric resolution.

- Standard resolution - 1-minute granularity
- High resolution - 1-second granularity

While AWS services provide standard resolution metrics, you can configure your custom metric in high resolution.

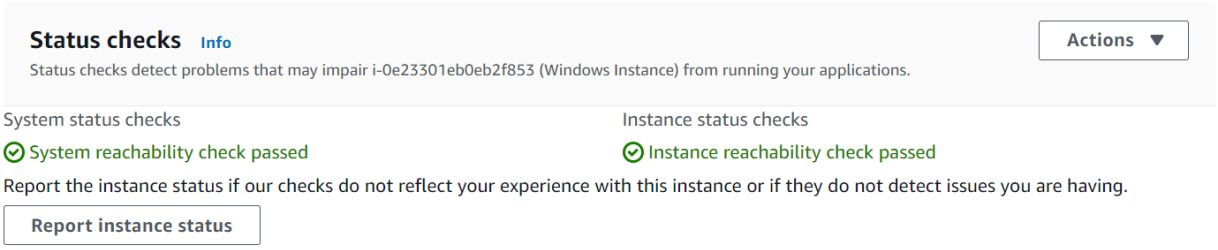
By default, Amazon CloudWatch does basic monitoring of all of your instances. CloudWatch collects the metrics at a 5-minute interval. Optionally, you can enable Amazon CloudWatch Detailed Monitoring to collect metrics in 1-minute intervals for an additional cost.



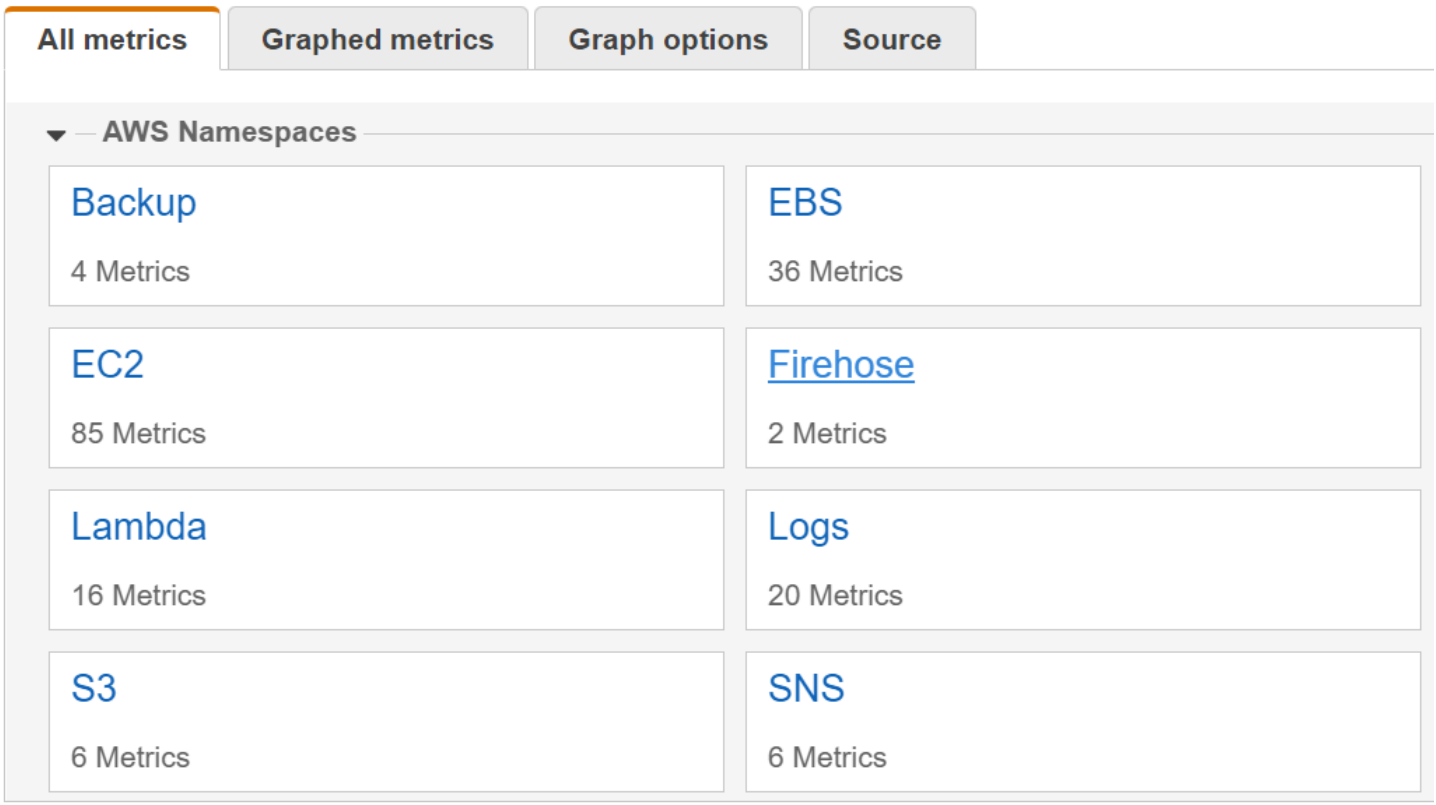
Instance Status Checks

Amazon EC2 publishes status check metrics to CloudWatch. EC2 performs status checks every minute to evaluate and identify potential failures on the instance and its underlying hardware and software. AWS does the following types of status checks to your instances.

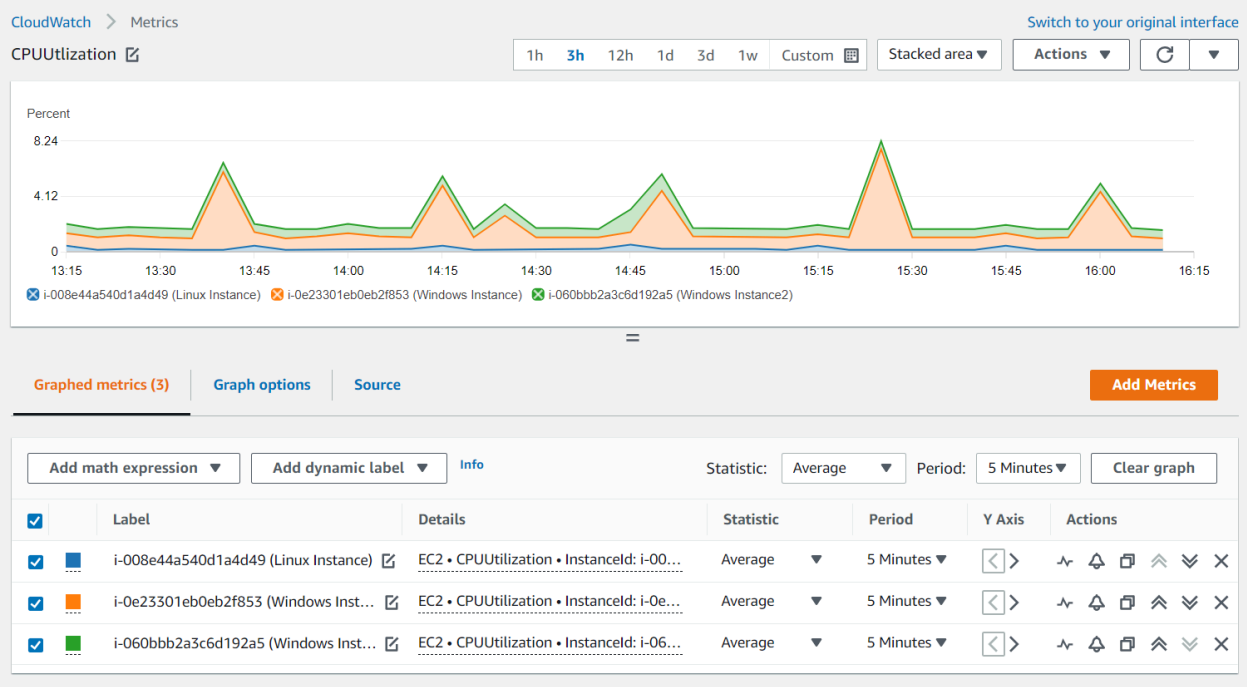
- **System Status Check** - checks the underlying infrastructure to where an instance is hosted
- **Instance Status Check** - checks if the operating system of instance accepts traffic from its network interface



CloudWatch categorizes metrics using Namespaces. AWS provides the default Namespaces per AWS service, but you can also create custom Namespaces for your custom metrics.



Below is an example of how CloudWatch presents the metrics collected from your instances. The graph illustrates the CPU Utilization metrics of the instances for the past 3 hrs. You can configure a graph to change the time scope, widget type, or modify the current metrics.

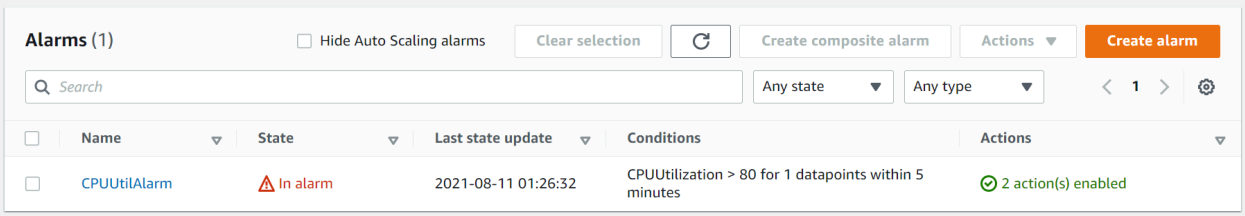


Creating CloudWatch Alarm

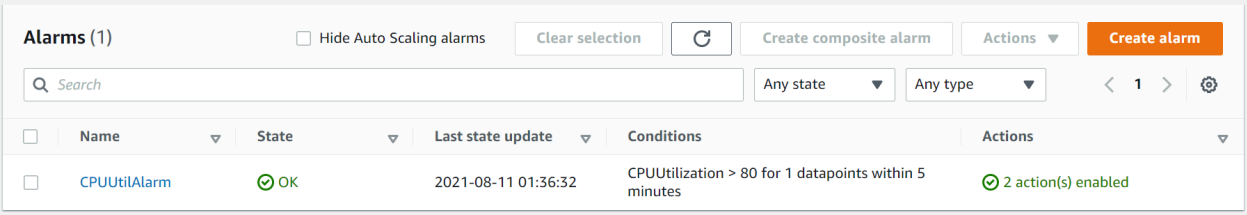
One of the valuable tools that you can use for monitoring is the CloudWatch Alarm. You can use CloudWatch Alarm to create alarms from the metrics collected from your AWS resources. The alarm has three states on its lifecycle.

- **OK** - metric is within the threshold/alarm condition isn't met
- **ALARM** - metric is outside the threshold/alarm condition is met
- **INSUFFICIENT_DATA** - no or insufficient data collected/alarm is just created

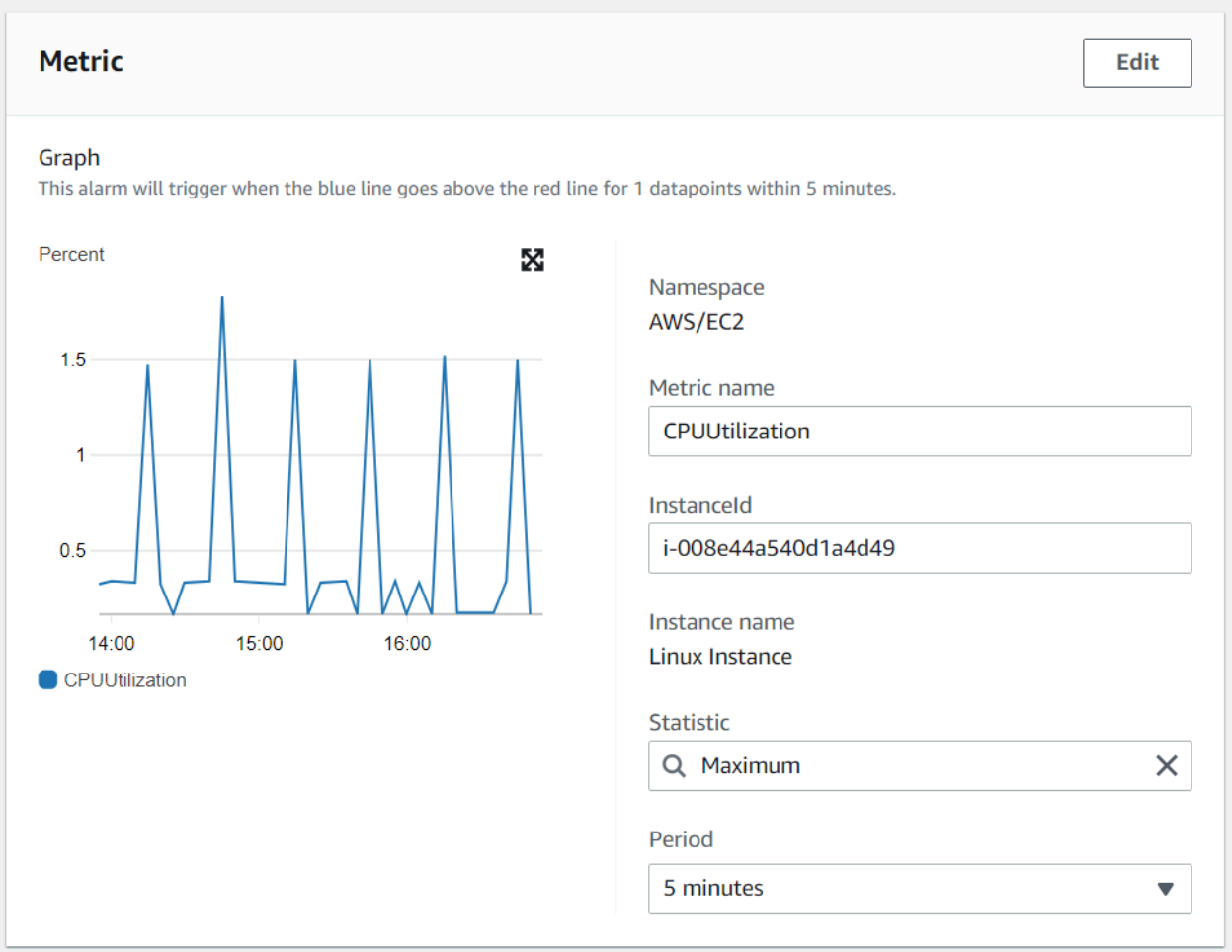
Example of CloudWatch Alarm with In Alarm state:



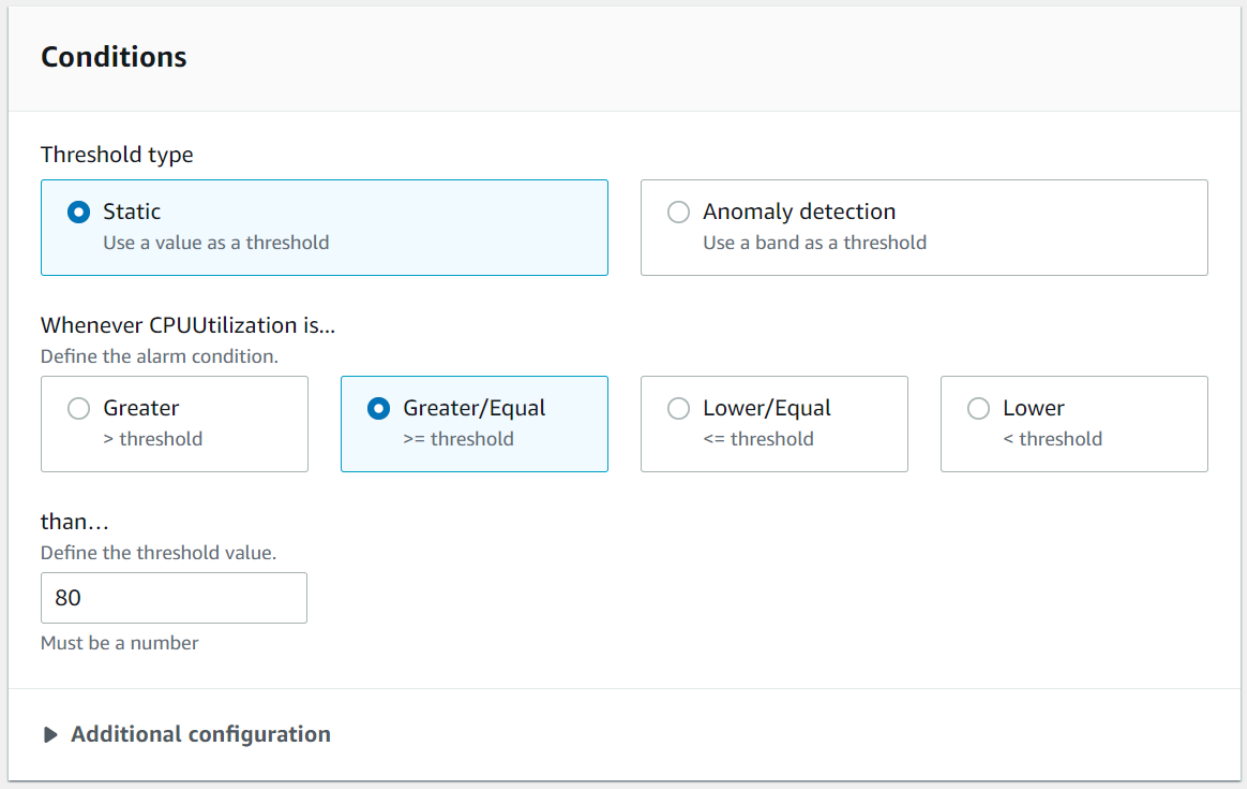
Example of CloudWatch Alarm with OK state:



When creating an alarm, you create an alarm trigger by specifying a metric and a condition.
You first select an AWS service where you want to create an alarm. Looking at the example below, the CloudWatch Alarm will check the maximum CPUUtilization metric of the EC2 instance within 5 minutes.

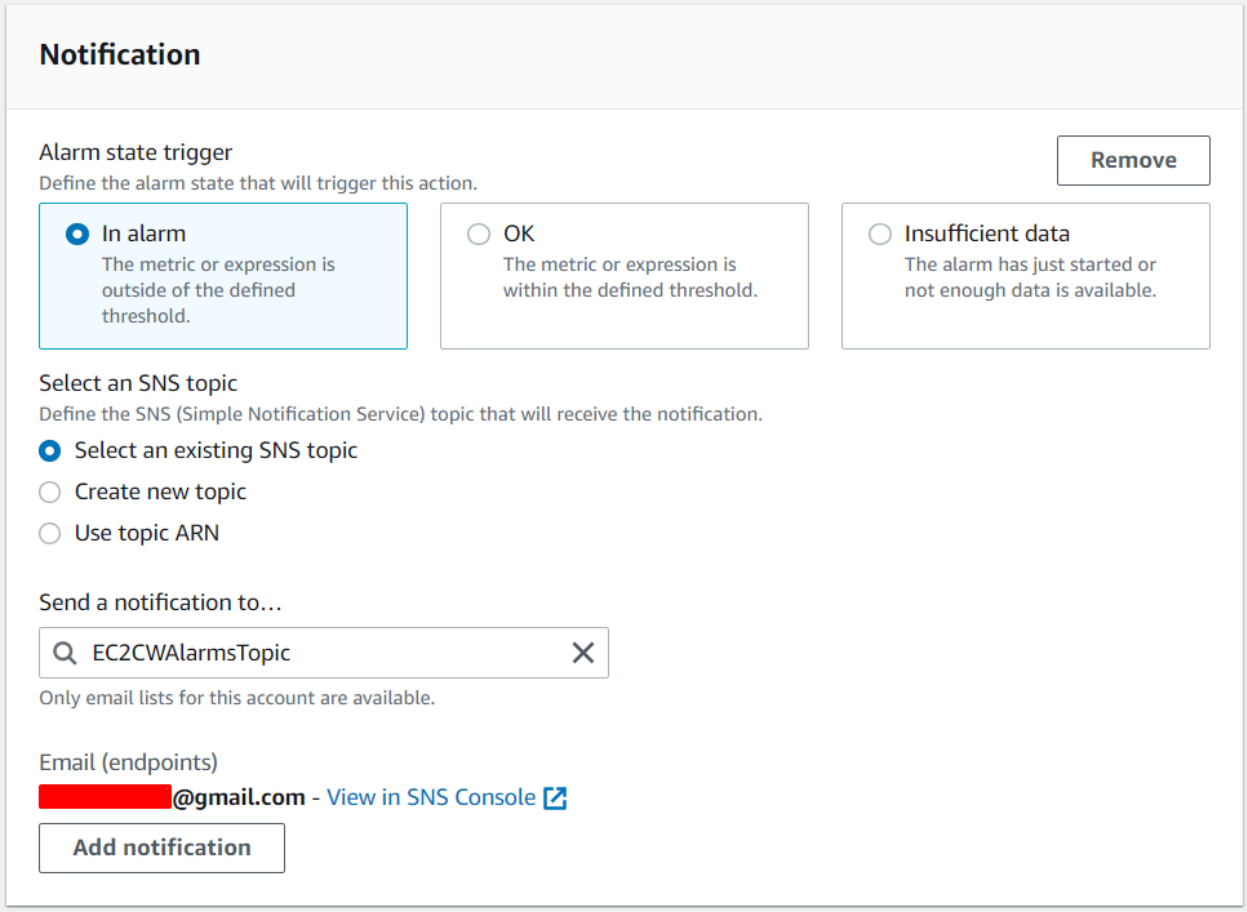


You can select Static or Anomaly detection for the threshold type. While Static uses the value from the defined threshold, the Anomaly detection evaluates past data to determine a potential anomaly.



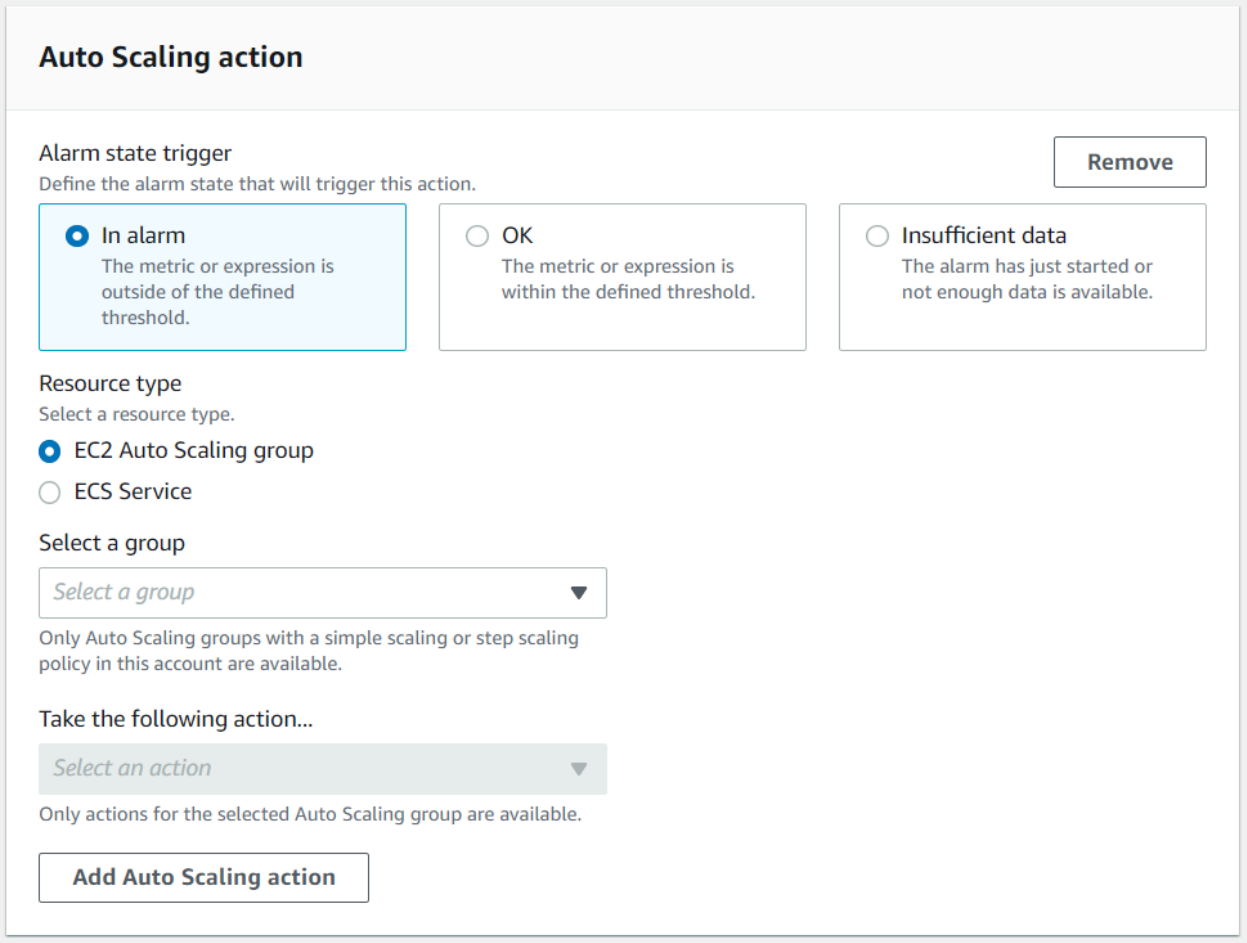
Once the metrics and the condition is set, you can define the action you want to do when the alarm is triggered. You can choose any action below.

Trigger an SNS Notification
Amazon CloudWatch will publish a message in the SNS Topic you selected, sending an email notification.



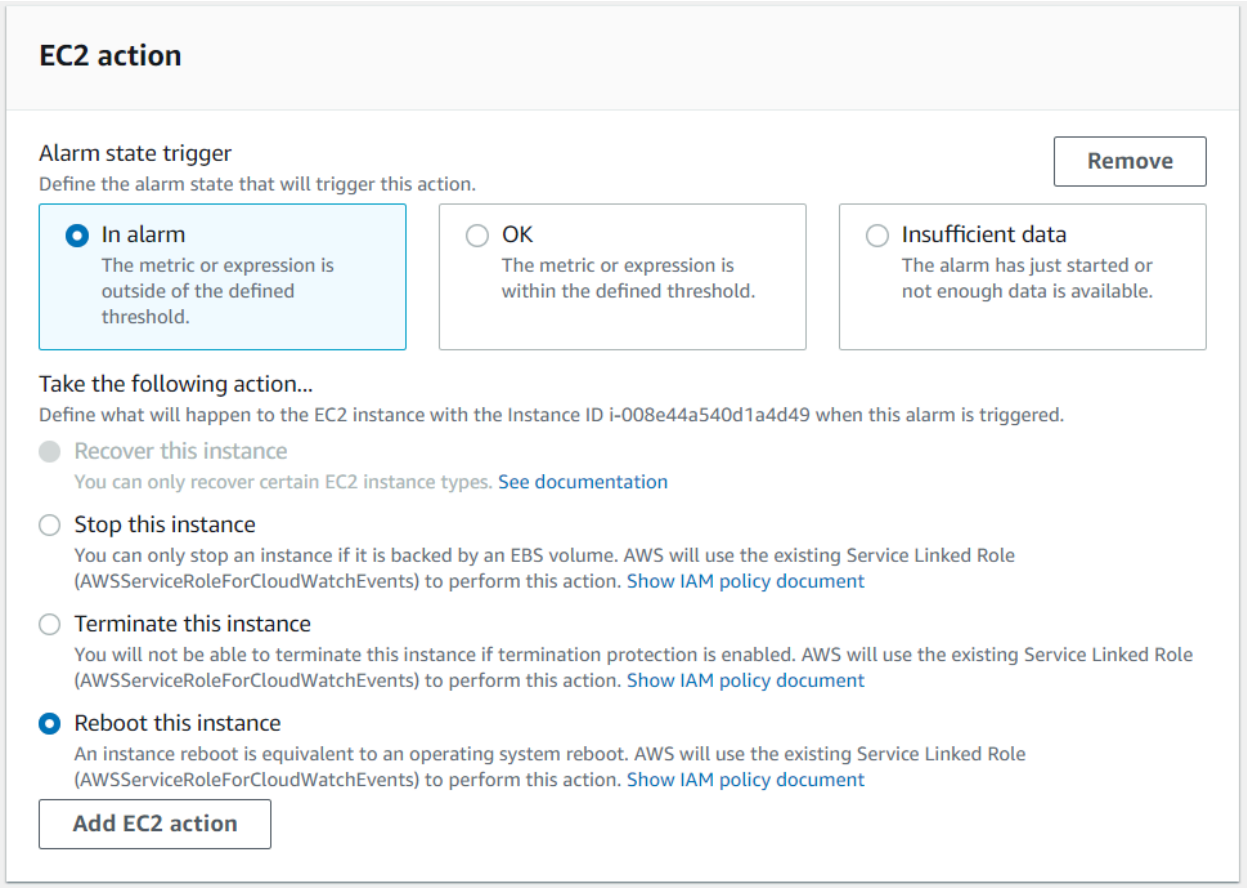
Trigger an Auto Scaling Action

You can trigger an auto scaling action only to your account's existing EC2 Auto Scaling group that uses a simple and step scaling policy. You can also configure trigger action for an ECS service.



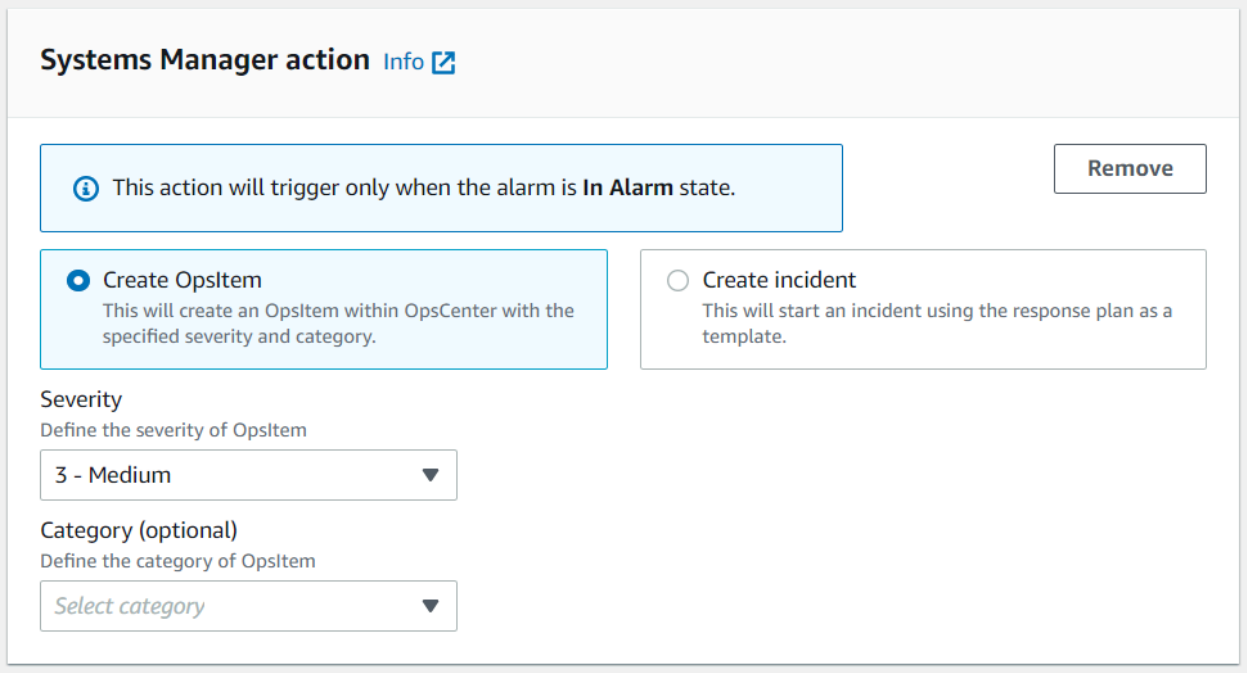
Trigger an EC2 Action

You can configure Amazon CloudWatch to trigger an EC2 action for every alarm state that triggers. You can stop, terminate, or reboot the instance.



Trigger a Systems Manager Action

Configure a Systems Manager action to create an OpsItem or Incident every time the alarm goes in In Alarm state.

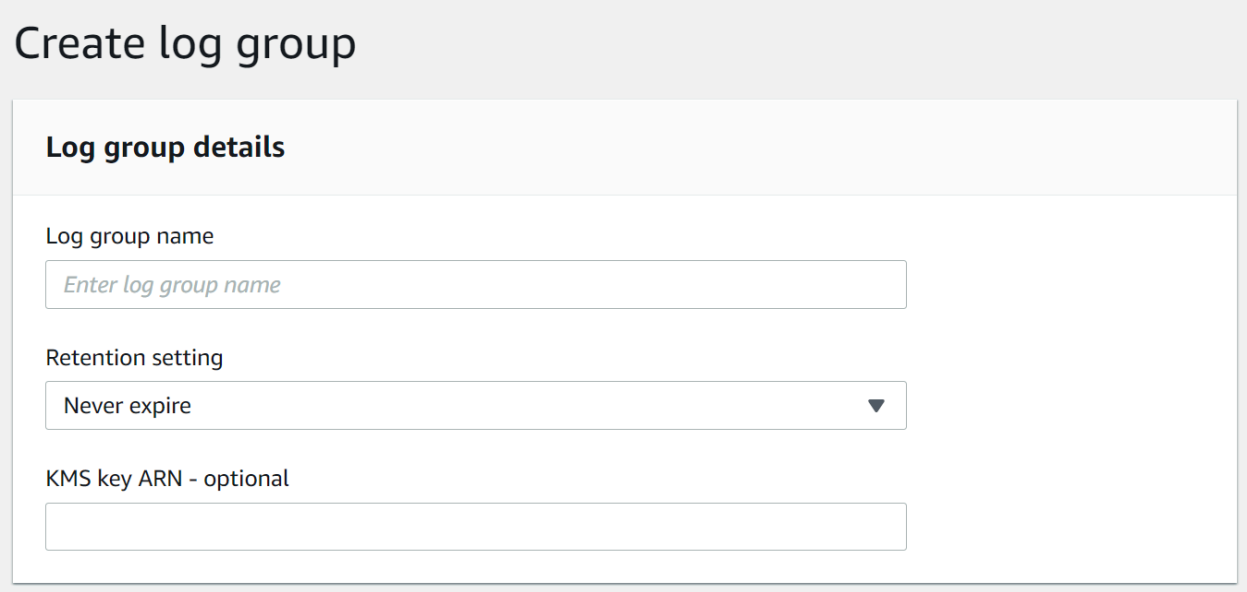


Working with CloudWatch Logs

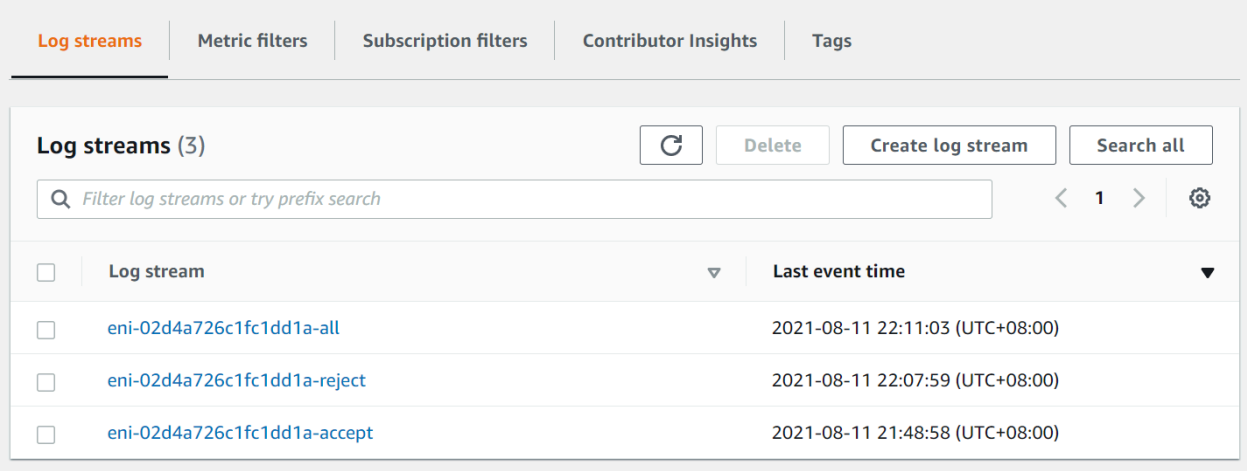
CloudWatch Logs is a component of CloudWatch that allows you to collect various logs from your application and different AWS services. If you are looking for an AWS native solution to centrally consolidate all of your logs, CloudWatch Logs is the right way to go.

Log groups

CloudWatch collects data called log events from various resources. **Log group** serves as a container for the log events. You define the log group name, log retention, and an optional KMS key when you create a log group.



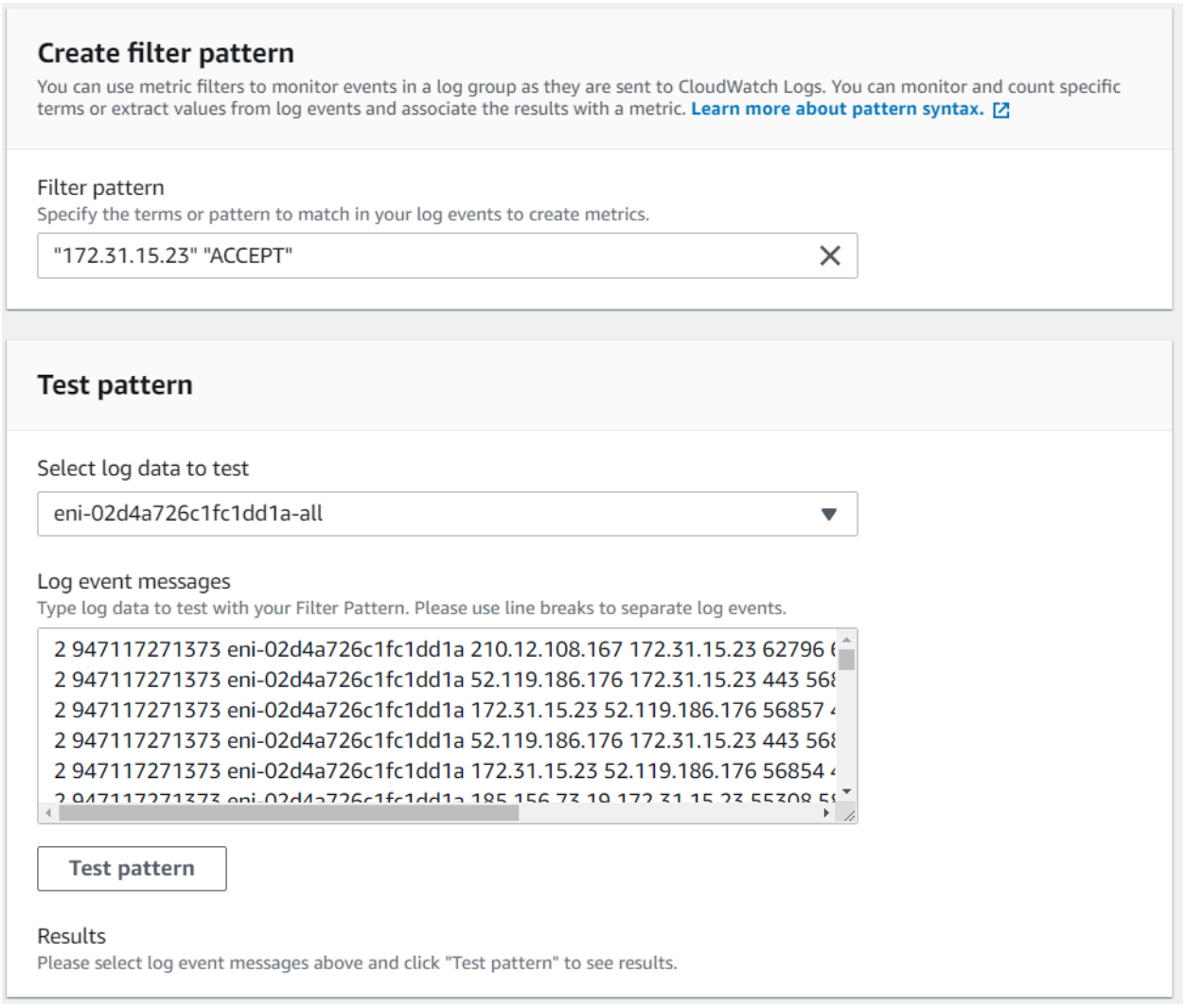
Log events are collected and are dumped to log groups via log streams. **Log streams** represent the sequence of data being collected from multiple sources. A log group can contain multiple log streams. Here is an example of a log group that has log streams to collect log events from the VPC Flow Logs.



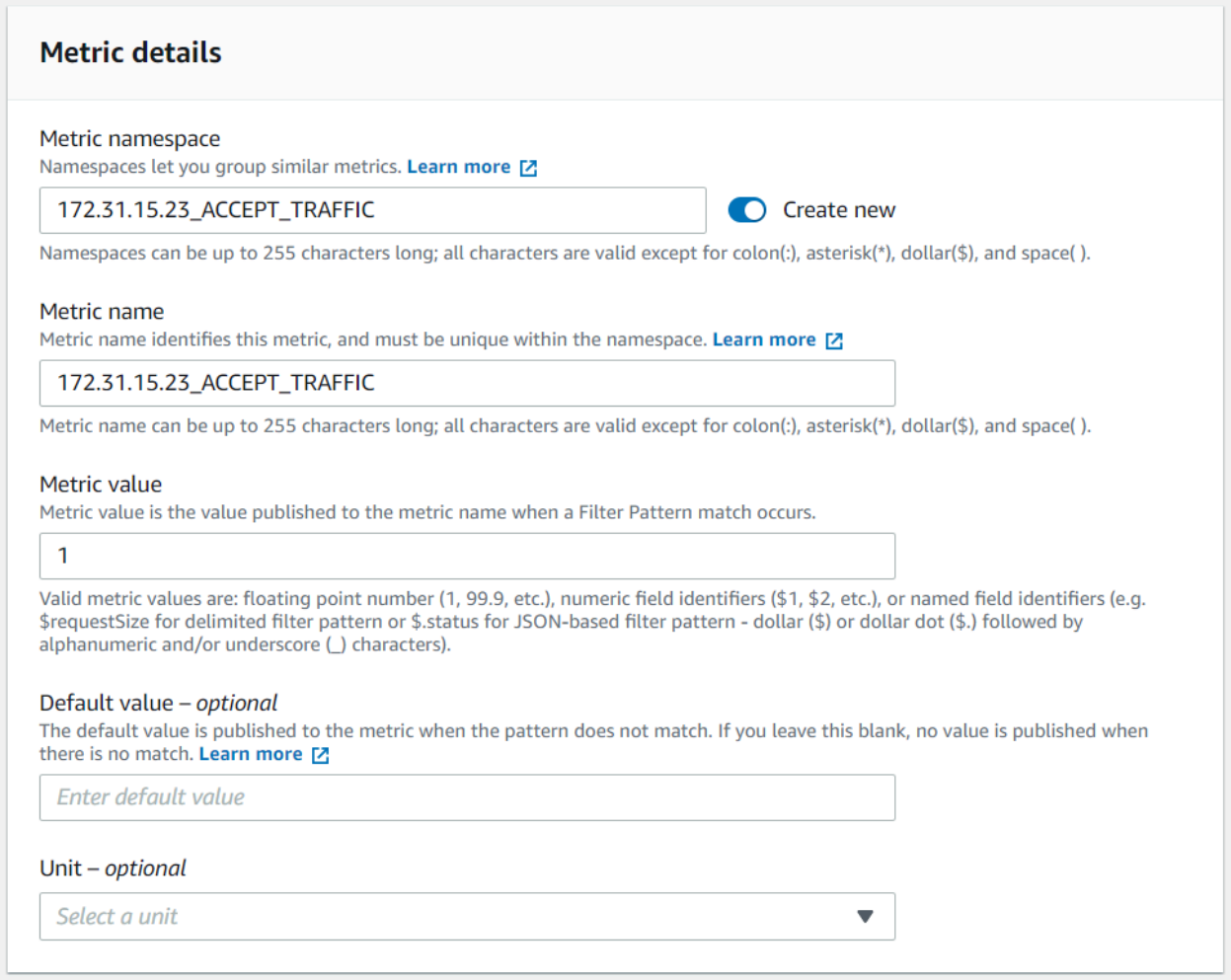
Utilizing Metric Filter

Metric Filter allows you to drill down into your logs. You can monitor specific values or events from your logs and push them into CloudWatch metrics as custom Metrics. By defining a filter pattern, you can filter the log events and choose only the needed data.

The example below filters the log events to display only events that have "*172.31.15.23*" and "*ACCEPT*" values. You can test the pattern on custom log data, or better yet, on the log events from your existing log stream.



Once the filter pattern is defined, you now set details for your metrics. Once the log events are pushed into the CloudWatch Metrics, it will appear as a custom Namespace. You can now also create CloudWatch Alarm using the custom metrics.



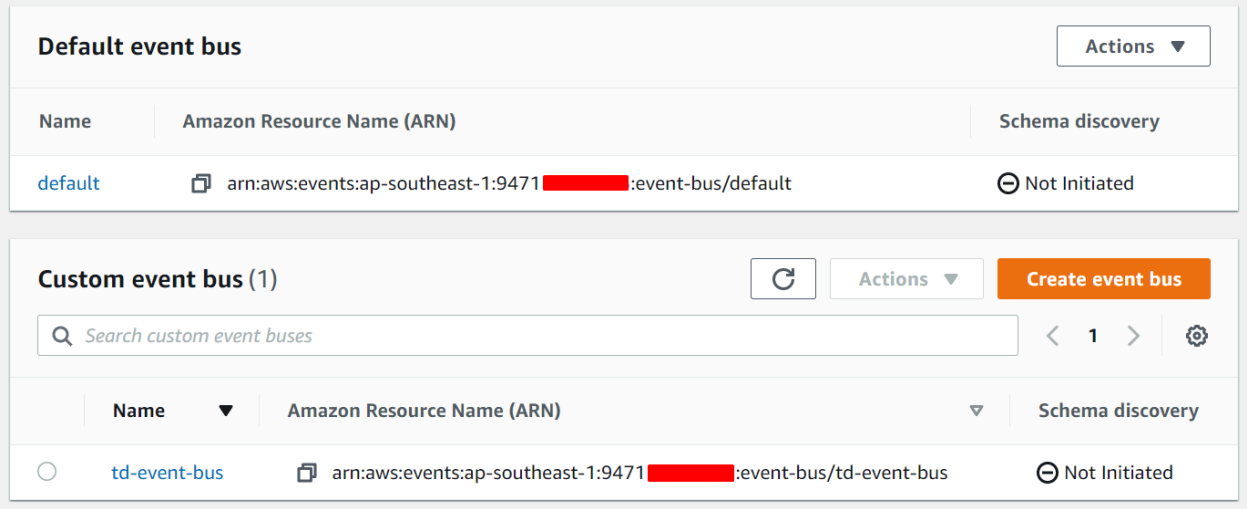
Event-driven Architecture with Amazon EventBridge

Amazon EventBridge is a serverless event bus service in AWS that enables you to build event-driven applications using events from different sources. Amazon EventBridge, as the name implies, serves as a bridge between your application and various event sources.

To further understand how Amazon EventBridge works, let's break down its components.

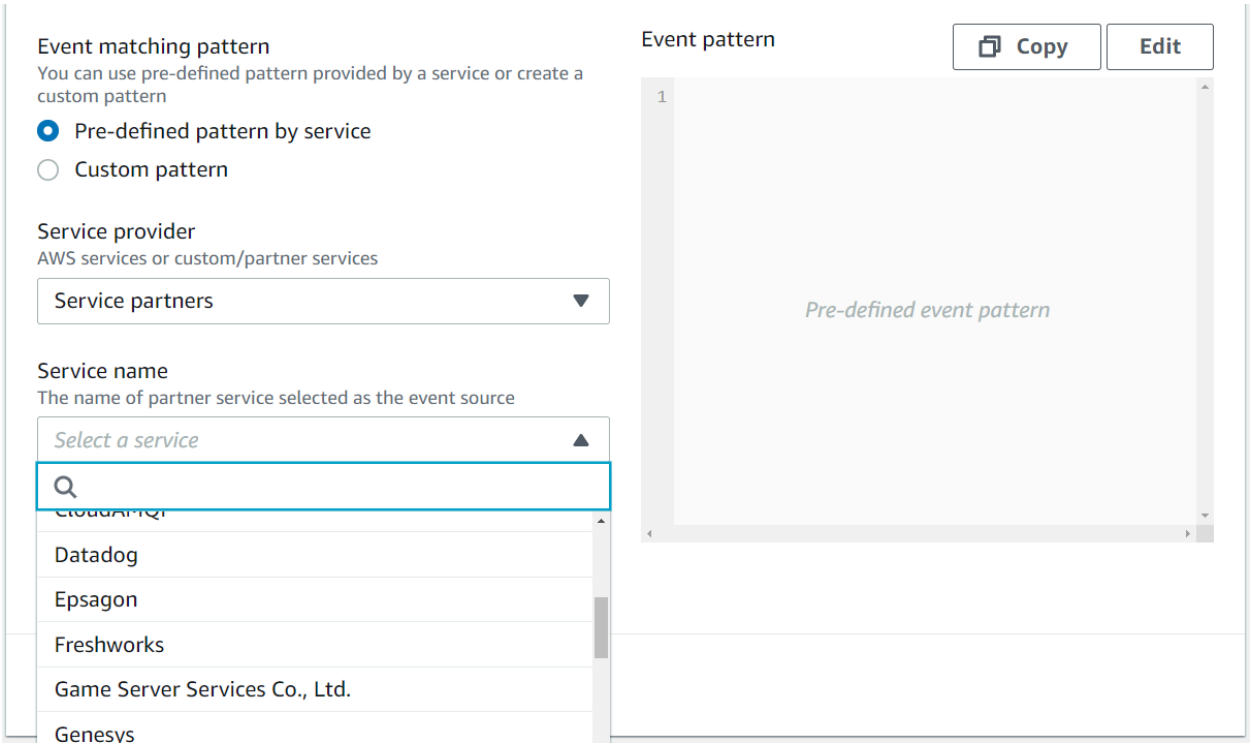
Event Bus

The Event Bus resource serves as a receiver of events from different sources. AWS has a default event bus used by various AWS services to send events. There's also an option that extends the capability of EventBridge to accept events from other sources, including various SaaS providers via a custom event bus.



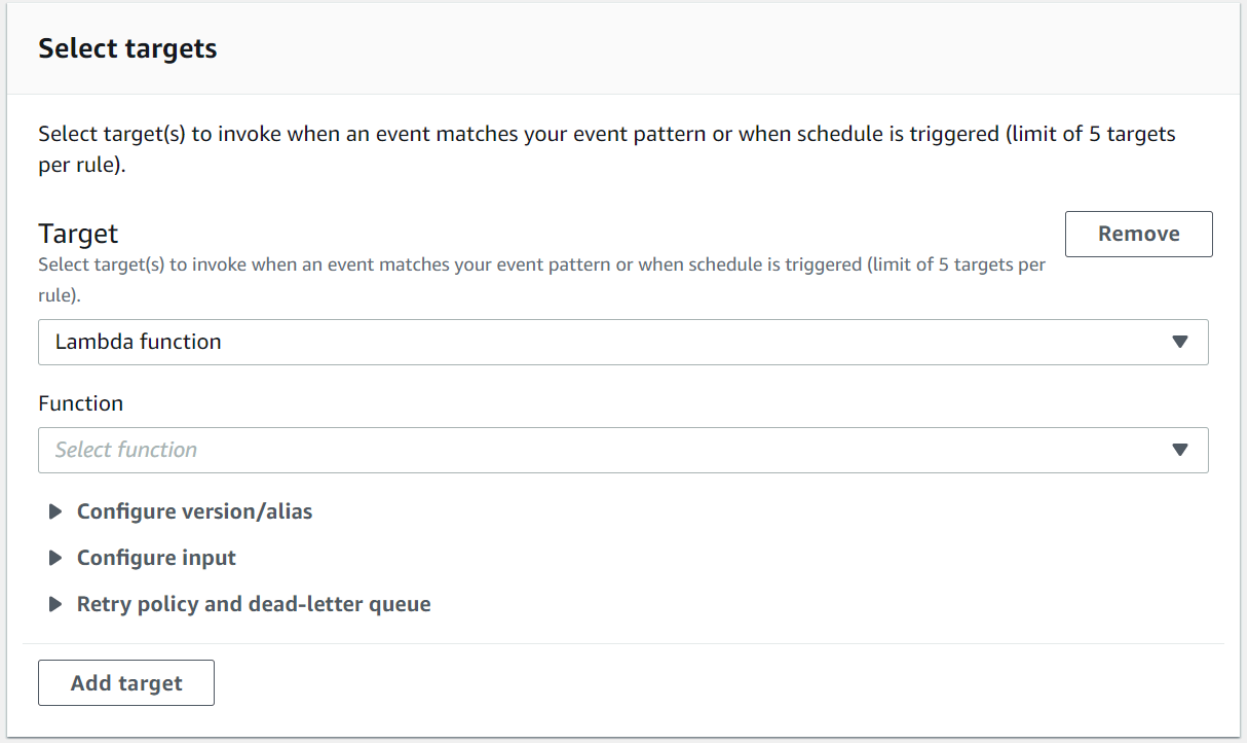
Rules

Rules are used for filtering events from the event bus or creating scheduled invokes for your targets. When you create a rule to match events, you specify a custom pattern or use AWS' predefined pattern. The predefined patterns include patterns for AWS services and SaaS providers like what you see from the example below. AWS supports tons of Saas applications, including Atlassian, Datadog, New Relic, PagerDuty, and Zendesk, to name a few.



Target

Amazon EventBridge evaluates all the events from the event bus according to the defined rule. Once an event matches a rule, Amazon EventBridge will invoke the defined target. Amazon EventBridge supports multiple AWS Services targets like Amazon EC2 actions, AWS Lambda function, SNS topic, and others.



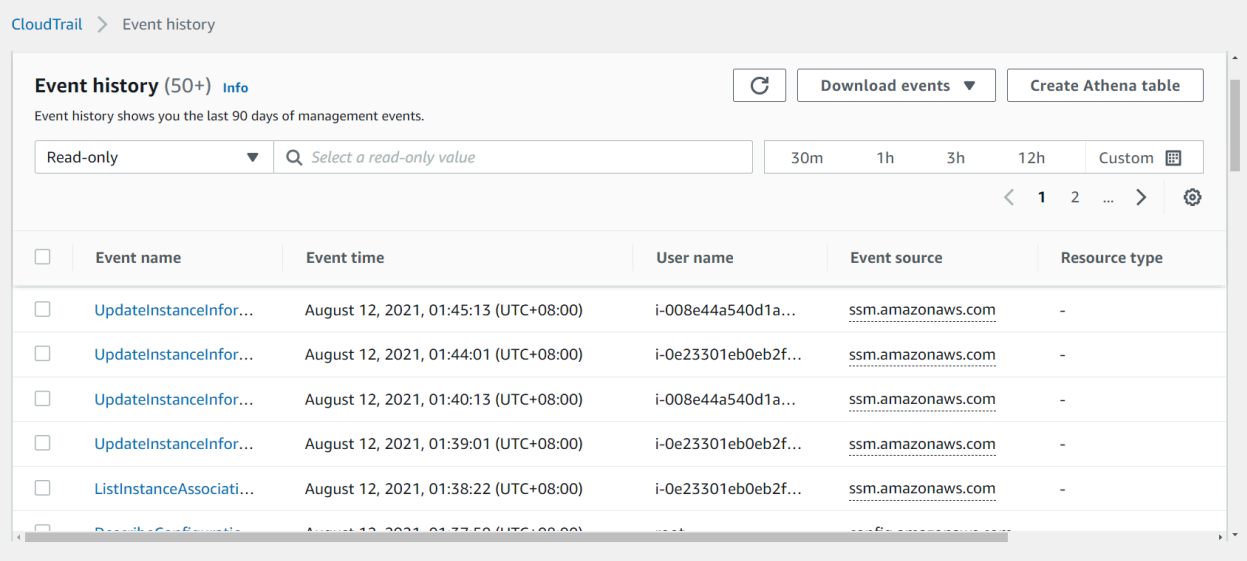
Exploring Events on CloudTrail

AWS CloudTrail records all the API calls made by a user, role, or service in AWS. You can use CloudTrail to review all the activity within your AWS environment, which helps in auditing, compliance requirements, or troubleshooting.

CloudTrail Event History

CloudTrail Event History keeps a record of events for the past 90 days. You use the provided filter to navigate through the data or create an Athena Table for complex queries. Below are the following attributes that you can use to filter CloudTrail events.

- AWS access key
- Event ID
- Event name
- Event sources
- Read-only
- Resource name
- Resource type
- User name



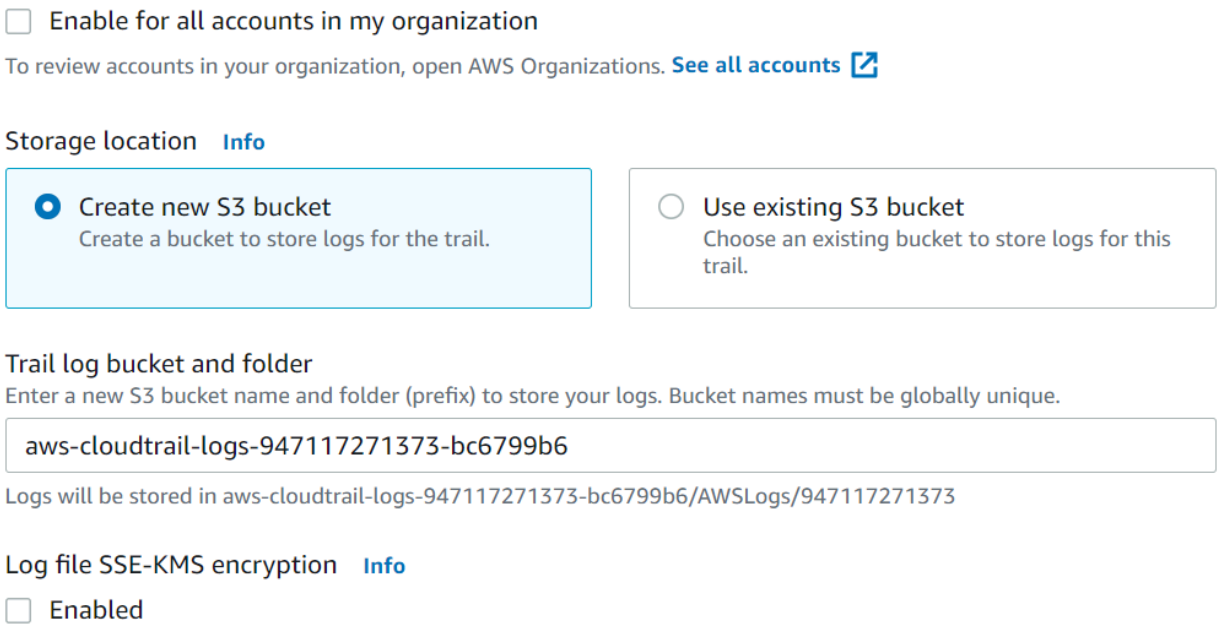
Below is an example of a "ConsoleLogin" event record. It shows essential data like User Identity, Account ID, and Login status.

Event record Info

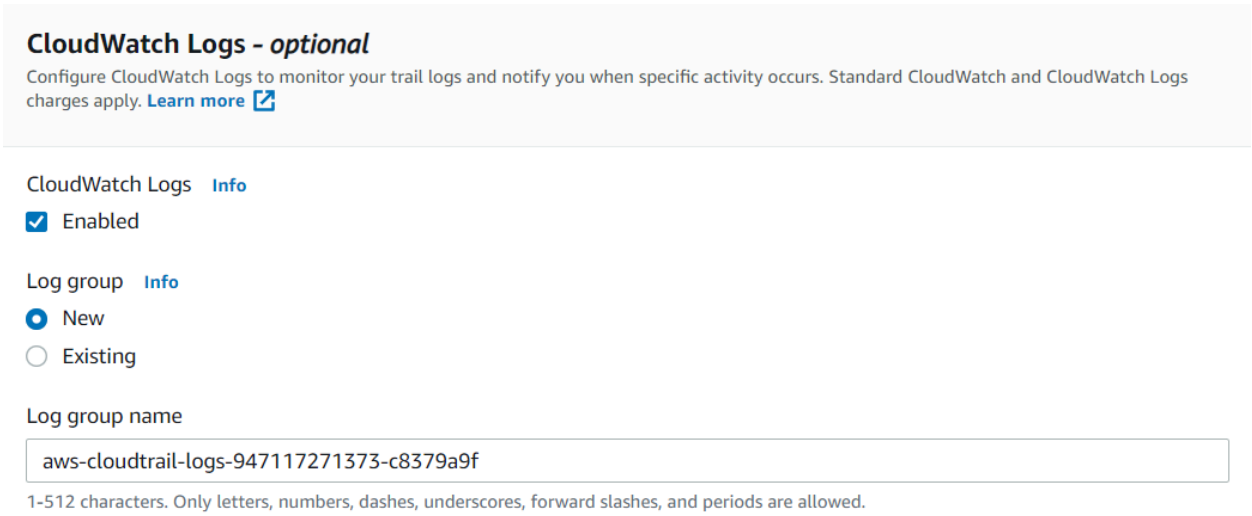

```
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "Root",
        "principalId": "947117271373",
        "arn": "arn:aws:iam::9471[redacted]:root",
        "accountId": "9471[redacted]"
    },
    "eventTime": "2021-08-11T12:49:22Z",
    "eventSource": "signin.amazonaws.com",
    "eventName": "ConsoleLogin",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "112.198.123.183",
    "userAgent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/92.0.4515.131 Safari/537.36",
    "requestParameters": null,
    "responseElements": {
        "ConsoleLogin": "Success"
```


Storing CloudTrail Events to S3 Bucket

CloudTrail allows you to create **Trails** for you to export specific CloudTrail events to S3. You can set up KMS encryption, log file validation, and SNS notification when creating trails. Once the logs are stored in an S3 bucket, you can also run queries on the CloudTrail logs using Amazon Athena.



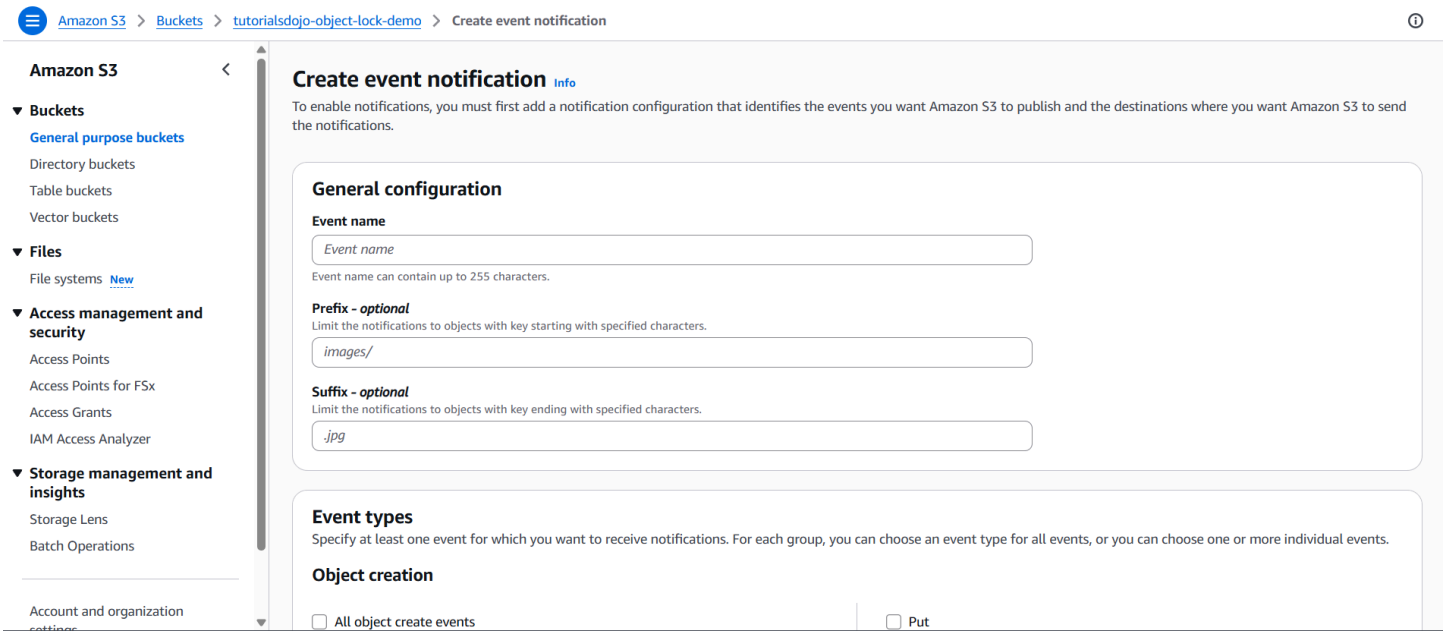
Moreover, you have an option to push the logs in an Amazon CloudWatch log group as well.



Amazon S3 Event Notifications

Amazon S3 Event Notifications enable automatic responses whenever specific events occur in an S3 bucket. This feature helps build event-driven architectures by allowing AWS services to react immediately to object-level activities such as uploads, deletions, restores, replication events, or lifecycle transitions.

S3 Event Notifications are commonly used to automate workflows, trigger serverless processing, generate alerts, or integrate with downstream systems.



Supported Event Destinations

Amazon S3 can publish notifications to the following AWS services:

- AWS Lambda
- Amazon Simple Queue Service (Amazon SQS)
- Amazon Simple Notification Service (Amazon SNS)
- Amazon EventBridge

These integrations allow applications to process S3 events asynchronously and at scale.

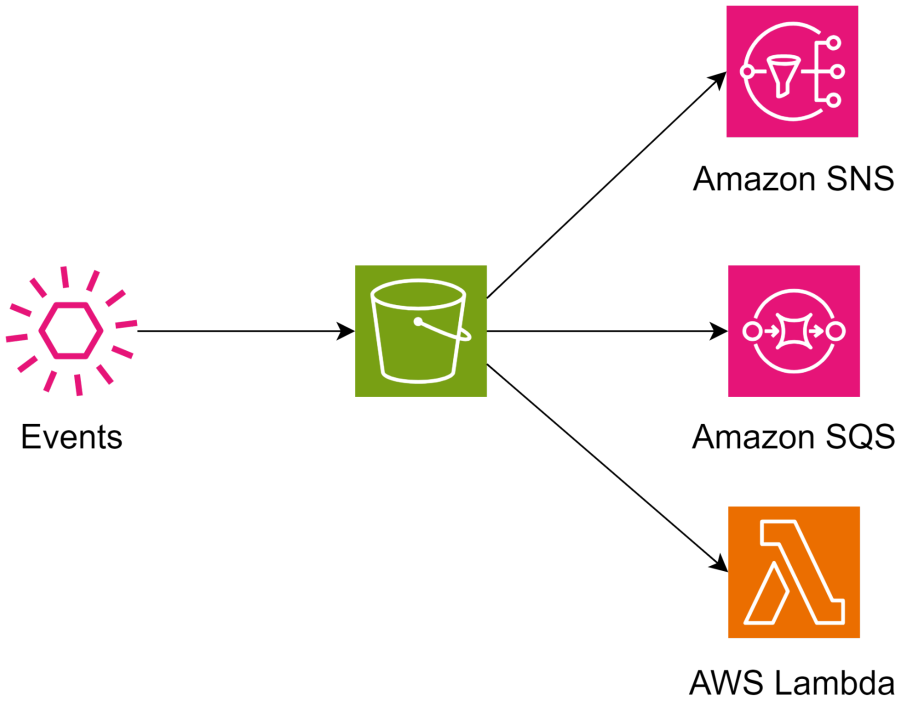
Common Amazon S3 Event Types

Below are some commonly used S3 event types:

Event Type	Description
`s3:ObjectCreated:*`	Triggered when an object is uploaded
`s3:ObjectRemoved:*`	Triggered when an object is deleted
`s3:ObjectRestore:*`	Triggered during Glacier restore operations

`s3:Replication:*`	Triggered for replication operations
`s3:LifecycleTransition`	Triggered when lifecycle rules move objects between storage classes
`s3:IntelligentTiering`	Triggered for Intelligent-Tiering archival events

How Amazon S3 Event Notifications Work



When an event occurs in an S3 bucket, Amazon S3 generates a JSON-formatted notification message. The message contains important metadata such as:

- Bucket name
- Object key
- Event type
- Event timestamp
- Requester identity
- Source IP address

The notification is then delivered to the configured destination service.

References:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/monitoring_automated_manual.html
https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/working_with_metrics.html
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/WhatIsCloudWatchLogs.html
https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html
https://aws.amazon.com/aws-cost-management/aws-cost-and-usage-reporting/
https://docs.aws.amazon.com/AmazonS3/latest/userguide/EventNotifications.html

AWS User Notifications

AWS User Notifications is a service that centralizes alerts and updates from multiple AWS services into a single, consistent channel. Instead of relying on separate service-specific consoles or emails, User Notifications lets you configure delivery preferences (such as email, chat, or console alerts) and subscribe to topics like cost anomalies, operational events, or security findings. This helps ensure that critical information, whether related to billing, compliance, or workload health, is surfaced proactively and in a unified format. By consolidating notifications, AWS enables teams to respond faster, reduce missed alerts, and maintain better visibility across distributed environments.

There are two categories of notifications available in AWS User Notifications:

- **AWS-managed notifications** – Automatically generated by AWS.
- **User-configured notifications (UCNs)** – Created through custom notification rules. These can be set up for **Amazon CloudWatch alarms, Support cases,** and other events you define.

Amazon Managed Service for Prometheus

Amazon Managed Service for Prometheus represents a serverless, Prometheus-compatible monitoring solution specifically designed for container metrics, enabling organizations to securely oversee container environments at scale. This managed service allows users to leverage the same open-source Prometheus data model and query language for monitoring containerized workload performance, while simultaneously gaining enhanced scalability, availability, and security without the burden of managing underlying infrastructure. The service also provides automatic scaling capabilities for ingestion, storage, and querying of operational metrics as workloads dynamically scale up and down, and it seamlessly integrates with AWS security services to facilitate fast and secure data access.

The service architecture is built for high availability through Multi-AZ deployments, where data ingested into a workspace is automatically replicated across three Availability Zones within the same Region. Users can work with container clusters running on Amazon Elastic Kubernetes Service as well as self-managed Kubernetes environments. Regarding data retention, metrics ingested into a workspace are typically stored for 150 days and are then automatically deleted, though users have the flexibility to adjust the retention period by configuring their workspace up to a maximum of 1,095 days (three years).

Amazon Managed Service for Prometheus delivers several compelling advantages for organizations:

- **Query and Scalability Capabilities:** The service enables teams to analyze container performance data through PromQL-based querying across extensive metric datasets from their Kubernetes infrastructure. As monitoring demands increase, the platform dynamically adjusts its capacity while preserving query performance regardless of deployment size.
- **Cost and Operational Efficiency:** By eliminating the need to maintain dedicated Prometheus infrastructure, organizations avoid expenses related to server maintenance, version updates, and capacity planning. The fully managed approach removes the burden of infrastructure administration.
- **Comprehensive Monitoring Scenarios:** Teams can establish centralized observability for containerized applications regardless of where they run, whether in AWS, private data centers, or alternative cloud platforms, with built-in connectivity to Amazon Managed Grafana for creating visualizations and alert mechanisms. The service accepts metric data from diverse environments through AWS Distro for OpenTelemetry and standard Prometheus-compatible exporters.

References:
https://aws.amazon.com/prometheus/
https://docs.aws.amazon.com/prometheus/latest/userguide/what-is-Amazon-Managed-Service-Prometheus.html
https://docs.aws.amazon.com/notifications/latest/userguide/what-is-service.html

AWS Health Dashboard

The AWS Health Dashboard is a personalized, real-time service that shows the performance and availability of the AWS services underlying your AWS resources. It offers timely and pertinent information to assist in managing ongoing events and preparing for scheduled activities. The dashboard sends out warnings and gives advice on fixing things when AWS is going through events that might affect your tasks. The issues reported on the AWS Health Dashboard are categorized into three types: open recent issues, scheduled changes, and event history, providing a comprehensive overview of past and present operational events that affect your AWS services. This tool is designed to give end users a reliable, up-to-the-minute overview of their applications' health, reducing the need for reactive problem-solving and minimizing downtime.

AWS Trusted Advisor

AWS Trusted Advisor serves as a personal cloud consultant for users, offering real-time guidance to ensure adherence to AWS best practices. It examines the user's AWS environment and provides recommendations

when it identifies potential cost savings, performance improvements, or security enhancements. The tool focuses on four key areas: cost optimization, security, fault tolerance, and performance improvement. Users can improve their system's efficiency and security by utilizing AWS Trusted Advisor while significantly reducing operational costs. This makes it an essential tool for any organization looking to maximize the benefits of their AWS deployment.

AWS tools and SDKs

AWS SDKs simplify the process for developers to connect their applications, libraries, or scripts with AWS services. They offer APIs for numerous AWS services, enabling developers to engage with these services using their chosen programming languages. AWS SDKs cater to various languages, such as Java, Python, PHP, .NET, Ruby, and JavaScript, both for browser-based and server-side applications with Node.js. Additionally, AWS provides the AWS Command Line Interface (CLI), a comprehensive tool for managing AWS services from the command line and automating them with scripts.

AWS tools and SDKs include security and access management features, monitoring and logging, debugging, and optimizing performance. When developers use these tools and kits, they can spend more time making their applications work the way they want and less time taking care of the systems running their applications. This leads to quicker creation of applications and better use of resources.

References:

https://docs.aws.amazon.com/health/latest/ug/what-is-aws-health.html
https://docs.aws.amazon.com/health/latest/ug/getting-started-health-dashboard.html
https://docs.aws.amazon.com/health/latest/APIReference/Welcome.html
https://docs.aws.amazon.com/awssupport/latest/user/trusted-advisor.html
https://aws.amazon.com/what-is/sdk/
https://docs.aws.amazon.com/sdkref/latest/guide/overview.html

COMPARISON OF AWS SERVICES

S3 vs EBS vs EFS

	Amazon S3	Amazon EBS	Amazon EFS
Type of storage	Object storage. You can store virtually any kind of data in any format.	Persistent block-level storage for EC2 instances.	POSIX-compliant file storage for EC2 instances.
Features	Accessible to anyone or any service with the right permissions	Deliver performance for workloads that require the lowest-latency access to data from a single EC2 instance	Has a file system interface, file system access semantics (such as strong consistency and file locking), and concurrently-accessible storage for multiple EC2 instances
Max Storage Style	Virtually unlimited	16 TiB for one volume	Unlimited system size
Max File Size	Individual Amazon S3 objects can range in size to a maximum of 5 terabytes.	Equivalent to the maximum size of your volumes	47.9 TiB for a single file
Performance (Latency)	Low, for mixed request types, and integration with CloudFront	Lowest, consistent; SSD-backed storages include the highest performance Provisioned OPS SSD and General Purpose SSD that balance price and performance.	Low, consistent; use Max I/O mode for higher performance
Performance (Throughput)	Multiple GBs per second; supports multi-part upload	Up to 2 GB per second. HDD-backed volumes include throughput intensive workloads and Cold HDD for less frequently accessed data.	10+ GB per second. Bursting Throughput mode scales with the scales with the size of the file system. Provisioned throughput mode offers higher dedicated throughput than bustring throughput.

Durability	Stored redundantly across multiple AZs; has 99.999999999% durability	Stored redundantly in a single AZ	Stored redundantly across multiple AZs
Availability	S3 Standard – 99.99% availability S3 Standard-IA – 99.9% availability S3 One Zone-IA – 99.5% availability. S3 Intelligent Tiering – 99.9%	Has 99.999% availability	99.9% SLA. Runs in multi – AZ
Scalability	Highly scalable	Manually increase/decrease your memory size. Attach and detach additional volumes to and from your EC2 instance to scale	EFS file systems are elastic, and automatically grow and shrink as you add and remove files.
Data Accessing	One to millions of connections over the wed; S3 provides a REST web services interface	Single EC2 instance in a single AZ Amazon EBS Multi-Attach a single Provisioned IOPS SSD (io1 or io2 Block Express) volume to up to 16 Nitro-based instances that are in the same Availability Zone.	One to thousands of EC2 instances or on-premises servers, from multiple AZs, regions, VPCs, and accounts concurrently
Access Control	Uses bucket policies and IAM user policies. Has Block Public Access settings to help manage public access to resources.	IAM Policies, Roles, and Security Groups	Only resources that can access endpoints in your VPC, called a mount target, can access your file system; POSIX-compliant user and group-level permissions.
Encryption Methods	Supports SSL endpoints using the HTTPS protocol, Client-Side and Server-Side Encryption (SSE-S3, SSE-C, SSE – KMS)	Encrypts both data-at-rest and data-in-transit through EBS encryption that uses AWS KMS keys.	Encrypt data at rest and in transit. Data at rest encryption uses AWS KMS. Data in transit uses TLS.

Backup and Restoration	Use versioning or cross-region replication	All EBS volume types offer durable snapshot capabilities.	EFS to EFS replication through third party tools or AWS DataSynch
Pricing	Billing prices are based on the location of your bucket. Lower costs equal lower prices. You get cheaper prices the more you use S3 storage.	You pay Gb-month of provisioned storage, provisioned IOPS-month, GB-month of snapshot data stored in S3	You pay more the amount of file system storage used per month. When using the Provisioned Throughput mode you pay for the throughput you provision per month.
Use cases	Web serving and content management, media and entertainment, backups, big data analytics, data lake	Boot volumes, transactional and NoSQL databases, data warehousing & ETL	Web serving and content management, enterprise applications, media and entertainment, home directories, database backups, developer tools, container storage, big data analytics
Service endpoint	Can be accessed within and outside a VPC (via S3 bucket URL)	Accessed within one's VPC	Accessed within one's VPC

Amazon S3 vs Glacier

- Amazon S3 is a durable, secure, simple, and fast storage service, while Amazon S3 Glacier is used for archiving solutions.
- Use S3 if you need low latency or frequent access to your data. Use S3 Glacier for low storage cost, and you do not require millisecond access to your data.
- You have three retrieval options when it comes to Glacier, each varying in the cost and speed it retrieves an object for you. You retrieve data in milliseconds from S3.
- Both S3 and Glacier are designed for durability of 99.999999999% of objects across multiple Availability Zones.
- S3 and Glacier are designed for availability of 99.99%.
- S3 can be used to host static web content, while Glacier cannot.
- In S3, users create buckets. In Glacier, users create archives and vaults.
- You can store a virtually unlimited amount of data in both S3 and Glacier.
- A single Glacier archive can contain 40TB of data.
- S3 supports Versioning.
- You can run analytics and querying on S3.
- You can configure a lifecycle policy for your S3 objects to automatically transfer them to Glacier. You can also upload objects directly to either S3 or Glacier.
- S3 Standard-IA and One Zone-IA have a minimum capacity charge per object of 128KB. Glacier's minimum is 40KB.
- Objects stored in S3 have a minimum storage duration of 30 days (except for S3 Standard). Objects that are archived to Glacier have a minimum 90 days of storage. Objects that are deleted, overwritten, or transitioned to a different storage class before the minimum duration will incur the normal usage charge plus a pro-rated request charge for the remainder of the minimum storage duration.
- Glacier has a per GB retrieval fee.
- You can transition objects from some S3 storage classes to another. Glacier objects can only be transitioned to the Glacier Deep Archive storage class.
- S3 (standard, intelligent-tiering, standard-IA, and one zone-IA) and Glacier are backed by an SLA.

S3 Standard vs S3 Standard-IA vs S3 One Zone-IA vs S3 Intelligent Tiering vs S3 Express One Zone

	S3 Standard	S3 Standard-Infrequent Access (IA)	S3 One Zone-Infrequent Access (IA)	S3 Intelligent Tiering	S3 Express One Zone
Features	General-purpose storage of frequently accessed data.	For long-lived, rapid but less frequently accessed data; data is stored redundantly in multiple AZs.	For long-lived, rapid but less frequently accessed data; data is stored redundantly in only one AZ of your choice.	For long-lived data that have unpredictable access patterns.	High performance storage for most frequently accessed data.
Durability	99.999999999% (11 9's)				
Availability	99.99%	99.9%	99.5%	99.9%	99.95%
Availability SLA	99.9%	99%	99%	99%	99.9%
Number of Availability Zones	At least 3	At least 3	Only 1	At least 3	Only 1
Minimum capacity charge per object	N/A	128KB	128KB	N/A	512KB
Minimum storage duration charge	N/A	30 days	30 Days	30 Days	1 hour
Inserting data	Directly PUT into S3 Standard	Directly PUT into S3 Standard-IA or set Lifecycle policies to transition objects from the S3 Standard to the S3 Standard-IA storage class.	Directly PUT into S3 One Zone-IA or set Lifecycle policies to transition objects from the S3 Standard to the S3 One Zone-IA storage class.	Directly PUT into S3 Intelligent-Tiering or set Lifecycle policies to transition objects from the S3 Standard to the S3 Intelligent-Tiering storage class.	Directly PUT into S3 Express One Zone.
Retrieval fee	N/A	per GB retrieved	per GB retrieved	N/A	N/A

First byte latency	milliseconds				single-digit milliseconds
Storage transition	S3 Standard to all other S3 storage types, including Glacier	S3 Standard-IA to S3 One Zone-IA or S3 Glacier	S3 One Zone-IA to S3 Glacier	S3 Intelligent to S3 One Zone-IA or S3 Glacier	You can either upload new objects directly or copy data from other storage classes.
Use cases	Cloud applications, dynamic websites, content distribution, mobile and gaming applications, and big data analytics.	Ideally suited for long-term file storage, older sync and share storage, and other aging data.	For infrequently-accessed storage, like backup copies, disaster recovery copies, or other easily recreatable data.	Data with unknown or changing access patterns, optimize storage costs automatically, and unpredictable workloads.	For most frequently accessed data.

Additional Notes:

- Data stored in the S3 One Zone-IA storage class will be lost in the event of AZ destruction.
- S3 Standard-IA costs less than S3 Standard in terms of storage price, while still providing the same high durability, throughput, and low latency of S3 Standard.
- S3 One Zone-IA has 20% less cost than Standard-IA.
- It is recommended to use multipart upload for objects larger than 100MB.

AWS DataSync vs Storage Gateway

	AWS DataSync	AWS Storage Gateway
Description	AWS DataSync is an online data transfer service that simplifies, automates, and accelerates the process of copying large amounts of data to and from AWS storage services over the Internet or over AWS Direct Connect.	AWS Storage Gateway is a hybrid cloud storage service that gives you on-premises access to virtually unlimited cloud storage by linking it to S3. Storage Gateway provides 3 types of storage interfaces for your on-premises applications: file, volume, and tape.
How it Works	Uses an agent which is a virtual machine (VM) that is owned by the user and is used to read or write data from your storage systems. You can activate the agent from the Management Console. The agent will then read from a source location, and sync your data to Amazon S3, Amazon EFS, or Amazon Fsx for Windows File Server.	Uses a Storage Gateway Appliance – a VM from Amazon – which is installed and hosted on your data center. After the setup, you can use the AWS console to provision your storage options: File Gateway, Cached Volumes, or Stored Volumes, in which data will be saved to Amazon S3. You can also purchase the hardware appliance to facilitate the transfer instead of installing the VM
Protocols	DataSync connects to existing storage systems and data sources with standard storage protocols (NFS, SMB), or using the Amazon S3 API	Storage Gateway provides a standard set of storage protocols such as iSCSI, SMB, and NFS.
Storage	AWS DataSync can copy data between Network File Systems (NFS), SMB file servers or self-managed object storages. It can also move data between your on-premises storage and Amazon S3, Amazon EFS, or Amazon FSx,	File Gateway enables you to store and retrieve objects in Amazon S3 using file protocols such as NFS and SMB. Volume Gateway stores your data locally in the gateway and syncs them to Amazon S3. It also allows you to take point-in-time copies of your volumes with EBS snapshots which you can restore and mount to your appliance as iSCSI device. Tape Gateway data is immediately stored in Amazon S3 and can be archived to Amazon S3 Glacier Flexible Retrieval or Amazon S3 Glacier Deep Archive.
Pricing	You are charged standard request, storage, and data transfer rates to read from and write to AWS services, such as Amazon S3,	You are charged based on the type and amount of storage you use, the requests

	Amazon EFS, AmazonFSx for Windows File Server, and AWS KMS.	you make, and the amount of data transferred out of AWS.
Combination	You can use a combination of DataSync and File Gateway to minimize your on-premises' operational costs while seamlessly connecting on-premises applications to your cloud storage. AWS DataSync enables you to automate and accelerate online data transfers to AWS storage services. File Gateway then provides your on-premises applications with low latency access to the migrated data.	

S3 Transfer Acceleration vs Direct Connect vs VPN

S3 Transfer Acceleration (TA)

- Amazon S3 Transfer Acceleration makes public Internet transfers to S3 faster, as it leverages Amazon CloudFront's globally distributed AWS Edge Locations.
- There is no guarantee that you will experience increased transfer speeds. If S3 Transfer Acceleration is not likely to be faster than a regular S3 transfer of the same object to the same destination AWS Region, AWS will not charge for the use of S3 TA for that transfer.
- This is not the best transfer service to use if transfer disruption is not tolerable.
- S3 TA provides the same security benefits as regular transfers to Amazon S3. This service also supports multi-part upload.
- ***S3 TA vs Direct Connect***
    - AWS Direct Connect is a good choice for customers who have a private networking requirement or who have access to AWS Direct Connect exchanges. S3 Transfer Acceleration is best for submitting data from distributed client locations over the public Internet, or where variable network conditions make throughput poor.
- ***S3 TA vs VPN***
    - You typically use (IPsec) VPN if you want your resources contained in a private network. VPN tools such as OpenVPN allow you to setup stricter access controls if you have a private S3 bucket. You can complement this further with the increased speeds from S3 TA.

AWS Direct Connect

- Using AWS Direct Connect, data that would have previously been transported over the Internet can now be delivered through a **private physical network connection** between AWS and your datacenter or corporate network. Customers' traffic will remain in AWS global network backbone, after it enters AWS global network backbone.
- Benefits of Direct Connect vs internet-based connections
    - reduced costs
    - increased bandwidth
    - a more consistent network experience
- 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.
- You can run IPv4 and IPv6 on the same virtual interface.
- Direct Connect does not support multicast.
- A Direct Connect connection is **not redundant**. Therefore, a second line needs to be established if redundancy is required. Enable *Bidirectional Forwarding Detection* (BFD) when configuring your connections to ensure fast detection and failover.
- AWS Direct Connect offers SLA.
- Direct Connect vs IPsec VPN
    - 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.

- You can combine one or more Direct Connect dedicated network connections with the Amazon VPC VPN. This combination provides an IPsec-encrypted private connection that also includes the benefits of Direct Connect.

AWS VPN

- AWS VPN is comprised of two services:
  - AWS Site-to-Site VPN enables you to securely connect your on-premises network or branch office site to your Amazon VPC.
  - AWS Client VPN enables you to securely connect users to AWS or on-premises networks.
- Data transferred between your VPC and datacenter routes over an encrypted VPN connection to help maintain the confidentiality and integrity of data in transit.
- If data that passes through Direct Connect moves in a dedicated private network line, AWS VPN instead encrypts the data before passing it through the Internet.
- VPN connection throughput can depend on multiple factors, such as the capability of your customer gateway, the capacity of your connection, average packet size, the protocol being used, TCP vs. UDP, and the network latency between your customer gateway and the virtual private gateway.
- All the VPN sessions are **full-tunnel VPN**. (cannot split tunnel)
- AWS Site-to-Site VPN enable you to create **failover** and CloudHub solutions **with AWS Direct Connect**.
- AWS Client VPN is designed to connect devices to your applications. It allows you to choose from **OpenVPN-based client**.

Amazon EBS: SSD vs HDD

On a given volume configuration, certain I/O characteristics drive the performance behavior for your EBS volumes. SSD-backed volumes, such as General Purpose SSD (gp2) and Provisioned IOPS SSD (io1, io2 Block Express), deliver consistent performance whether an I/O operation is random or sequential. HDD-backed volumes like Throughput Optimized HDD (st1) and Cold HDD (sc1) deliver optimal performance only when I/O operations are large and sequential.

In the exam, always consider the difference between SSD and HDD as shown on the table below. This will allow you to easily eliminate specific EBS-types in the options which are not SSD or not HDD, depending on whether the question asks for a storage type which has ***small, random*** I/O operations or ***large, sequential*** I/O operations.

FEATURES	SSD (Solid State Drive)	SSD (Hard Disk Drive)
Best workloads with:	small, random I/O operations	large, sequential I/O operations
Can be used as bootable volume?	Yes.	No
Suitable Use Cases	• Best for transactional workloads • Critical business applications that require sustained IOPS performance • Large database workloads such as MongoDB, Oracle, Microsoft SQL Server and many others…	• Best for large streaming workloads requiring consistent, fast throughput at a low price. • Big data, Data warehouses, Log processing. • Throughput-oriented storage for large volumes of data that is infrequently accessed.
Cost	moderate/ high	low
Dominant Performance Attribute	IOPS	Throughput (MiB/s)

Provisioned IOPS SSD (io1, io2 Block Express) volumes are designed to meet the needs of I/O-intensive workloads, particularly database workloads, that are sensitive to storage performance and consistency. Unlike gp2, which uses a bucket and credit model to calculate performance, an io1 volume allows you to specify a consistent IOPS rate when you create the volume, and Amazon EBS delivers within 10 percent of the

provisioned IOPS performance 99.9 percent of the time over a given year. Provisioned IOPS SSD io2 Block Express is an upgrade of Provisioned IOPS SSD io1. It offers higher 99.999% durability and higher IOPS per GiB ratio with 500 IOPS per GiB, all at the same cost as io1 volumes.

Volume Name	General Purpose SSD		Provisioned IOPS SSD	
Volume type	gp3	gp2	Io2 Block Express	io1
Description	General Purpose SSD volume that balances price performance for a wide variety of transactional workloads	General Purpose SSD volume that balances price performance for a wide variety of transactional workloads	High performance SSD volume designed for **business-critical** latency-sensitive applications	High performance SSD volume designed for latency-sensitive transactional workloads
Use Cases	virtual desktops, medium sized single instance databases such as MSFT SQL Server and Oracle DB, low-latency interactive apps, dev & test, boot volumes	Boot volumes, low-latency interactive apps, dev & test	Workloads that require sub-millisecond latency, and sustained IOPS performance or more than 64,000 IOPS or 1,000 MiB/s of throughput	Workloads that require sustained IOPS performance or more than 16,000 IOPS and I/O-intensive database workloads
Volume Size	1 GB – 16 TB	1 GB – 16 TB	4 GB – 16 TB	4 GB – 16 TB
Durability	99.8% - 99.9% durability	99.8% - 99.9% durability	99.999%	99.8% - 99.9%
Max IOPS / Volume	16,000	16,000	64,000	64,000
Max Throughput / Volume	1000 MB/s	250 MB/s	1,000 MB/s	1,000 MB/s
Max IOPS / Instance	260,000	260,000	160,000	260,000
Max IOPS / GB	N/A	N/A	500 IOPS/GB	50 IOPS/GB
Max Throughput / Instance	7,500 MB/s	7,500 MB/s	4,750 MB/s	7,500 MB/s
Latency	single digit millisecond	single digit millisecond	single digit millisecond	single digit millisecond
Multi-Attach	No	No	Yes	Yes

Volume Name	Throughput Optimized HDD	Cold HDD
Volume type	st1	sc1
Description	Low cost HDD volume designed for frequently accessed, throughput-intensive workloads	Throughput-oriented storage for data that is infrequently accessed Scenarios where the lowest storage cost is important
Use Cases	Big data, data warehouses, log processing	Colder data requiring fewer scans per day
Volume Size	125 GB – 16 TB	125 GB – 16 TB
Durability	99.8% - 99.9% durability	99.8% - 99.9% durability
Max IOPS / Volume	500	250
Max Throughput / Volume	500 MB/s	250 MB/s
Max IOPS / Instance	260,000	260,000
Max IOPS / GB	N/A	N/A
Max Throughput / Instance	7,500 MB/s	7,500 MB/s
Multi-Attach	No	No

Amazon RDS vs Amazon DynamoDB

	RDS	DynamoDB
Type of database	Managed relational (SQL) database	Fully managed key-value and document (NoSQL) database
Features	Has several database instance types for different kinds of workloads and supports six database engines – Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database, and SQL Server.	Delivers single-digit millisecond performance at any scale.
Storage Size	• 128 TB for Aurora engine. • 64 TB for MySQL, MariaDB, Oracle and PostgreSQL engines. • 16 TB for SQL Server engine.	Supports tables of virtually any size.
Number of tables per unit	Depends on the database engine	256
Performance	General Purpose Storage is an SSD-backed storage option that delivers at consistent baseline of 3 IOPS per provisioned GB with the ability to burst up to 3,000 IOPS. Provisioned IOPS Storage is an SSD-backed storage option designed to deliver a consistent IOPS rate that you specify when creating a database instance, up to 40,000 IOPS per database Instance. Amazon RDS provisions that IOPS rate for the lifetime of the database instance. Optimized for OLTP database workloads. Magnetic – Amazon RDS also supports magnetic storage for backward compatibility.	Single-digit millisecond read and write performance. Can handle more than 10 trillion requests per day with peaks greater than 20 million requests per second, over petabytes of storage. DynamoDB Accelerator (DAX) is an in-memory cache that can improve the read performance of your DynamoDB tables by up to 10 times – taking the time required for reads from milliseconds to microseconds, even at millions of requests per second. You specify the read and write throughput for each of your tables.
Availability and durability	Amazon RDS Multi-AZ deployments synchronously	DynamoDB global tables replicate your data automatically across 3 Availability

	replicates your data to a standby instance in a different Availability Zone Amazon RDS will automatically replace the compute instance powering your deployment in the event of a hardware failure.	Zones of your choice of AWS Regions and automatically scale capacity to accommodate your workloads.
Backups	The automated backup feature enables point-in-time recovery for your database instance. Database snapshots are user-initiated backups of your instance stored in Amazon S3 that are kept until you explicitly delete them.	Point-in-time recovery (PITR) provides continuous backups of your DynamoDB table data, and you can restore that table to any point in time up to the second during the preceding 35 days. On-demand backups and restore allows you to create full backups of your DynamoDB tables' data for data archiving.
Scalability	The Amazon Aurora engine will automatically grow the size of your database volume. The MySQL, MariaDB, SQL Server, Oracle, and PostgreSQL engines allow you to scale on-the-fly with zero downtime. RDS also supports storage auto scaling Reads replicas are available in Amazon RDS for MySQL, MariaDB, and PostgreSQL as well as Amazon Aurora.	Support tables of virtually any size with horizontal scaling. For tables using on-demand capacity mode, DynamoDB instantly accommodates your workloads a they ramp up or down to any previously reached traffic level. For tables using provisioned capacity, DynamoDB delivers automatic scaling of throughput and storage based on your previously set capacity.
Security	Isolate your database in your own virtual network. Connect to your on-premises IT infrastructure using industry-standard encrypted IPsec VPNs. You can configure firewall settings and control network access to your database instances. Integrates with IAM.	Integrates with IAM.
Encryption	Encrypt your databases using keys you manage through AWS KMS. With encryption enabled, data	DynamoDB encrypts data at rest by default using encryption keys stored in AWS KMS.

	stored at rest is encrypted, as are its automated backups, read replicas, and snapshots. Supports Transparent Data Encryption in SQL Server and Oracle. Supports the use of SSL to secure data in transit.	
Maintenance	Amazon RDS will update databases with the latest patches. You can exert optional control over when and if your database instance is patched.	No maintenance since DynamoDB is serverless.
Pricing	A monthly charge for each database instance that you launch. Option to reserve a DB instance for a One or three year term and receive discounts in pricing, compared to On-Demand instance pricing.	Charges for reading, writing, and storing data in your DynamoDb tables, along with any optional features you choose to enable. There are specific billing options for each of DynamoDB's capacity modes.
Use cases	Traditional applications, ERP, CRM, and e-commerce.	Internet-scale applications, real-time bidding, shopping carts, and customer Preferences, content management, Personalization, and mobile applications.

Additional notes:

- DynamoDB has built-in support for ACID transactions.
- DynamoDB uses filter expressions because it does not support complex queries.
- Multi-AZ deployments for the MySQL, MariaDB, Oracle, and PostgreSQL engines utilize synchronous physical replication. Multi-AZ deployments for the SQL Server engine use synchronous logical replication.

Amazon RDS vs Amazon Aurora

	Aurora	RDS
Type of database	Relational database	
Features	• MySQL and PostgreSQL compatible. • 5x faster than standard MySQL databases and 3x faster than standard PostgreSQL databases. • Use Parallel Query to run transactional and analytical workloads in the same Aurora database, while maintaining high performance. • You can distribute and load balance your unique workloads across different sets of Aurora DB instances using custom endpoints. • Aurora Serverless allows for on-demand, autoscaling of your Aurora DB instance capacity.	• Has several database instance types for different kinds of workloads and supports five database engines - MySQL, PostgreSQL, MariaDB, Oracle, and SQL Server. • Can use either General Purpose Storage and Provisioned IOPS storage to deliver a consistent IOPS performance
Maximum storage capacity	• 128 TB	• 64 TB for MySQL, MariaDB, Oracle, and PostgreSQL engines • 16 TB for SQL Server engine
DB instance classes	• Memory Optimized classes - for workloads that need to process large data sets in memory. • Burstable classes - provides the instance the ability to burst to a higher level of CPU performance when required by the workload.	• Standard classes - for a wide range of workloads, you can use general purpose instance. It offers a balance of compute, memory, and networking resources. • Memory Optimized classes - for workloads that need to process large data sets in memory. • Burstable classes - provides the instance the ability to burst to a

		higher level of CPU performance when required by the workload.
Availability and durability	• Amazon Aurora uses RDS Multi-AZ technology to automate failover to one of up to 15 Amazon Aurora Replicas across three Availability Zones • Amazon Aurora Global Database uses storage-based replication to replicate a database across multiple AWS Regions, with typical latency of less than 1 second. • Self-healing: data blocks and disks are continuously scanned for errors and replaced automatically.	• Amazon RDS Multi-AZ deployments synchronously replicates your data to a standby instance in a different Availability Zone. • Amazon RDS will automatically replace the compute instance powering your deployment in the event of a hardware failure.
Backups	• Point-in-time recovery to restore your database to any second during your retention period, up to the last five minutes. • Automatic backup retention period up to thirty-five days. • Backtrack to the original database state without needing to restore data from a backup.	• The automated backup feature enables point-in-time recovery for your database instance. • Database snapshots are user-initiated backups of your instance stored in Amazon S3 that are kept until you explicitly delete them.

Scalability	• Aurora automatically increases the size of your volumes as your database grows larger (increments of 10 GB). • Aurora also supports replica auto-scaling, where it automatically adds and removes DB replicas in response to changes in performance metrics. • Cross-region replicas provide fast local reads to your users, and each region can have an additional 15 Aurora replicas to further scale local reads.	• The MySQL, MariaDB, SQL Server, Oracle, and PostgreSQL engines scale your storage automatically as your database workload grows with zero downtime. • Read replicas are available for Amazon RDS for MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server. Amazon RDS creates a second DB instance using a snapshot of the source DB instance and uses the engines' native asynchronous replication to update the read replica whenever there is a change to the source. • Can scale compute and memory resources (vertically) of up to a maximum of 32 vCPUs and 244 GiB of RAM.
Security	• Isolate the database in your own virtual network via VPC. • Connect to your on-premises IT infrastructure using encrypted IPsec VPNs or Direct Connect and VPC Endpoints. • Configure security group firewall and network access rules to your database instances. • Integrates with IAM.	
Encryption	• Encrypt your databases using keys you manage through AWS KMS. With Amazon Aurora encryption, data stored at rest is encrypted, as are its automated backups, snapshots, and replicas in the same cluster. • Supports the use of SSL (AES-256) to secure data in transit.	• Encrypt your databases using keys you manage through AWS KMS. With Amazon RDS encryption, data stored at rest is encrypted, as are its automated backups, read replicas, and snapshots. • Supports Transparent Data Encryption in SQL Server and Oracle. • Supports the use of SSL to secure data in transit

DB Authentication	• Password authentication • Password and IAM database authentication	• Password authentication • Password and IAM database authentication • Password and Kerberos authentication
Maintenance	• Amazon Aurora automatically updates the database with the latest patches. • Amazon Aurora Serverless enables you to run your database in the cloud without managing/maintaining any database infrastructure.	• Amazon RDS will update databases with the latest major and minor patches on scheduled maintenance windows. You can exert optional control over when and if your database instance is patched.
Monitoring	• Use Enhanced Monitoring to collect metrics from the operating system instance. • Use Performance Insights to detect database performance problems and take corrective action. • Uses Amazon SNS to receive a notification on database events.	
Pricing	• A monthly charge for each database instance that you launch if you use on-demand. This includes both the instance compute capacity and the amount of storage being used. • Option to reserve a DB instance for a one or three-year term (reserve instances) and receive discounts in pricing.	

Use Cases	• Enterprise applications - a great option for any enterprise application that uses relational database since it handles provisioning, patching, backup, recovery, failure detection, and repair. • SaaS applications - without worrying about the underlying database that powers the application, you can concentrate on building high-quality applications. • Web and mobile gaming - since games need a database with high throughput, storage scalability, and must be highly available. Aurora suits the variable use pattern of these apps perfectly.	• Web and mobile applications - since the application needs a database with high throughput, storage scalability, and must be highly available. RDS also fulfills the needs of such highly demanding apps. • E-commerce applications - a managed database service that offers PCI compliance. You can just focus on building high-quality customer experiences without thinking of the underlying database. • Mobile and online games - game developers don't need to worry about provisioning, scaling, and monitoring of database servers since RDS manages the database infrastructure.

Multi-AZ deployments vs. Multi-Region deployments vs. Read Replicas

Multi-AZ deployments	Multi-Region deployments	Read Replicas
Main purpose is high availability	Main purpose is disaster recovery and local performance	Main purpose is scalability
Non-Aurora: synchronous replication; Aurora: asynchronous replication	Asynchronous replication	Asynchronous replication
Non-Aurora: only the primary instance is active; Aurora: all instances are active	All regions are accessible and can be used for reads	All read replicas are accessible and can be used for readscaling
Non-Aurora: automated backups are taken from standby; Aurora: automated backups are taken from shared storage layer	Automated backups can be taken in each region	No backups configured by default
Always span at least two Availability Zones within a single region	Each region can have a Multi-AZ deployment	Can be within an Availability Zone, Cross-AZ, or Cross-Region
Non-Aurora: database engine version upgrades happen on primary; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent in each region; all instances are updated together	Non-Aurora: database engine version upgrade is independent from source instance; Aurora: all instances are updated together
Automatic failover to standby (non-Aurora) or read replica (Aurora) when a problem is detected.	Aurora allows promotion of a secondary region to be the master	Can be manually promoted to a standalone database instance (non-Aurora) or to be the primary instance (Aurora)

Amazon Container Services (Amazon ECS) vs AWS Lambda

Amazon Container Service (ECS)	AWS Lambda
Amazon 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. ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure.	AWS Lambda is a function-as-a-service offering that runs your code in response to events and automatically manages the compute resources for you, since Lambda is a serverless compute service. With Lambda, you do not have to worry about managing servers, and directly focus on your application code.
With ECS, deploying containerized applications is easily accomplished. This service fits well in running batch jobs or in a microservice architecture.	Lambda automatically scales your function to meet demands. It is noteworthy, however, that Lambda has a maximum execution duration per request of 900 seconds or 15 minutes.
You have a central repository where you can upload your Docker Images from ECS container for safekeeping called Amazon ECR.	To allow your Lambda function to access other services such as Cloudwatch Logs, you would need to create an execution role that has the necessary permissions to do so.
Applications in ECS can be written in a stateful or stateless matter.	You can easily integrate your function with different services such as API Gateway, DynamoDB, CloudFront, etc. using the Lambda console.
The Amazon ECS CLI supports Docker Compose, which allows you to simplify your local development experience as well as easily set up and run your containers on Amazon ECS.	You can test your function code locally in the Lambda console before launching it into production.
Since your applications still run on EC2 instances, server management is your responsibility. This gives you more granular control over your system.	Currently, Lambda supports only a number of programming languages such as Java, Go, PowerShell, Node.js, C#, Python, and Ruby. ECS is not limited by programming languages since it mainly caters to Docker.
It is up to you to manage scaling and load balancing of your EC2 instances as well, unlike in AWS Lambda where functions scale automatically.	Lambda functions must be stateless since you do not have volumes for data storage.
You are charged for the costs incurred by your EC2 instances in your clusters. Most of the time, Amazon ECS costs more than using AWS Lambda	You are charged based on the number of requests for your functions and the duration, the time it takes for your code to execute. To minimize costs, you can

since your active EC2 instances will be charged by the hour.	throttle the number of concurrent executions running at a time, and the execution time limit of the function.
One version of Amazon ECS, know as AWS Fargate, will fully manage your infrastructure so you can just focus on deploying containers. AWS Fargate has a different pricing model from the standard EC2 cluster.	With Lambda@Edge, AWS Lambda can run your code across AWS locations globally in response to Amazon CloudFront events, such as requests for content to or from origin servers and viewers. This makes it easier to deliver content to end users with lower latency.
ECS will automatically recover unhealthy containers to ensure that you have the desired number of containers supporting your application.	

Security Group vs NACL

Security Group	Network Access Control List
Acts as a firewall for associated Amazon EC2 instances.	Acts as a firewall for associated subnets.
Controls both inbound and outbound traffic at the instance level.	Controls both inbound and outbound traffic at the subnet level.
You can secure your VPC instances using only security groups.	Network ACLs are an additional layer of defense.
Supports allow rules only.	Supports allow rules and deny rules.
Stateful (Return traffic is automatically allowed, regardless of any rules).	Stateless (Return traffic must be explicitly allowed by rules).
Evaluates all rules before deciding whether to allow traffic.	Evaluates rules in number order when deciding whether to allow traffic, starting with the lowest numbered rule.
Applies only to the instance that is associated to it.	Applies to all instances in the subnet it is associated with.
Has separate rules for inbound and outbound traffic.	Has separate rules for inbound and outbound traffic.
A newly created security group denies all inbound traffic by default.	A newly created nACL denies all inbound traffic by default.
A newly created security group has an outbound rule that allows all outbound traffic by default	A newly created nACL denies all outbound traffic by default.
Instances associated with a security group can't talk to each other unless you add rules allowing it.	Each subnet in your VPC must be associated with a network ACL. If none is associated, the default nACL is selected.
Security groups are associated with network interfaces.	You can associate a network ACL with multiple subnets; however, a subnet can be associated with only one network ACL at a time.

Your VPC has a default security group with the following rules:

1. Allow inbound traffic from instances assigned to the same security group.
2. Allow all outbound IPv4 traffic and IPv6 traffic if you have allocated an IPv6 CIDR block.

Your VPC has a default network ACL with the following rules:

1. Allows all inbound and outbound IPv4 traffic and, if applicable, IPv6 traffic.
2. Each network ACL also includes a non modifiable and non removable rule whose rule number is an asterisk. This rule ensures that if a packet doesn't match any of the other numbered rules, it's denied.

Application Load Balancer vs Network Load Balancer vs Gateway Load Balancer

FEATURE	Application Load Balancer	Network Load Balancer	Gateway Load Balancer
Protocols	HTTP, HTTPS, gRPC	TCP, UDP, TLS	IP
Platforms	VPC	VPC	VPC
Health checks	HTTP, HTTPS, gRPC...	TCP, HTTP, HTTPS	TCP, HTTP, HTTPS
Cloudwatch Metrics	Yes	Yes	Yes
Logging	Yes	Yes	Yes
Zonal Failover	Yes	Yes	Yes
Connection Draining (deregistration delay)	Yes	Yes	Yes
Load Balancing to multiple ports on the same instance	Yes	Yes	Yes
IP addresses as targets	Yes	Yes (TCP, TLS)	Yes
Load Balancer deletion protection	Yes		
Configuration idle connection timeout	Yes		
Cross-zone load balancing	Yes	Yes	Yes
Sticky sessions	Yes	Yes	Yes
Static IP		Yes	
Elastic IP address		Yes	
Preserve Source IP address	Yes	Yes	Yes
Resource-based IAM permissions	Yes	Yes	Yes
Slow start	Yes		
Web sockets	Yes	Yes	Yes

PrivateLink Support		Yes (TCP, TLS)	Yes (GWLBE)
Source IP address CIDR-based routing	Yes		
Layer 7			
Path-based routing	Yes		
Host-based routing	Yes		
Native HTTP/2	Yes		
Redirects	Yes		
Fixed response	Yes		
Lambda Functions as targets	Yes		
HTTP header-based routing	Yes		
HTTP method-based routing	Yes		
Query string parameter-based routing	Yes		
Security			
SSL offloading	Yes	Yes	
Server Name Indication (SNI)	Yes	Yes	
Back-end server encryption	Yes	Yes	
User authentication	Yes		
Session Resumption	Yes	Yes	
Terminates flow/ proxy behavior	Yes	Yes	Yes

Common features between the load balancers:

- Has instance health check features
- Has built-in CloudWatch monitoring
- Logging features
- Support zonal failover
- Supports connection draining
- Support cross-zone load balancing (evenly distributes traffic across registered instances in enabled AZs)
- Resource-based IAM permission policies
- Tag-based IAM permissions
- Flow stickiness - all packets are sent to one target and return the traffic that comes from the same target.

EC2 Instance Health Check vs ELB Health Check vs Auto Scaling and Custom Health Check

EC2 instance health check	Elastic Load Balancer health check	Auto Scaling and Custom health checks

• Amazon EC2 performs automated checks on every running EC2 instance to identify hardware and software issues.	• To discover the availability of your registered EC2 instances, a load balancer periodically sends pings, attempts connections, or sends requests to test the EC2 instances.	• All instances in your Auto Scaling group start in the healthy state. Instances are assumed to be healthy unless EC2 Auto Scaling receives notification that they are unhealthy. This notification can come from one or more of the following sources:
		  ○ Amazon EC2 (default)
		  ○ Elastic Load Balancing
		  ○ A custom health check.

• Status checks are performed every minute and each returns a pass or a fail status.
  ○ If all checks pass, the overall status of the instance is OK.
  ○ If one or more checks fail, the overall status is impaired.
	• The status of the instances that are healthy at the time of the health check is InService. The status of any instances that are unhealthy at the time of the health check is OutOfService.	• After Amazon EC2 Auto Scaling marks an instance as unhealthy, it is scheduled for replacement. If you do not want instances to be replaced, you can suspend the health check process for any individual Auto Scaling group.

• Status checks are built into EC2, so they cannot be disabled or deleted.	• When configuring a health check, you would need to provide the following:
	  ○ a specific port
	  ○ protocol to use
	  ■ HTTP/HTTPS health check succeeds if the instance returns a 200 response code within the health check interval.
		• If an instance is in any state other than running or if the system status is impaired, Amazon EC2 Auto Scaling considers the instance to be unhealthy and launches a replacement instance.

	■ A TCP health check succeeds if the TCP connection succeeds. ■ An SSL health check succeeds if the SSL handshake succeeds. ○ ping path	
• You can create or delete alarms that are triggered based on the result of the status checks.	• ELB health checks do not support WebSockets.	• If you attached a load balancer or target group to your Auto Scaling group, Amazon EC2 Auto Scaling determines the health status of the instances by checking both the EC2 status checks and the Elastic Load Balancing health checks.
• There are two types of status checks	• The load balancer routes requests only to the healthy instances. When an instance becomes impaired, the load balancer resumes routing requests to the instance only when it has been restored to a healthy state.	• Amazon EC2 Auto Scaling waits until the health check grace period ends before checking the health status of the instance. Ensure that the health check grace period covers the expected startup time for your application.
• System Status Checks • These checks detect underlying problems with your instance that require AWS involvement to repair. When a system status check fails, you can choose to wait for AWS to fix the issue, or you can resolve it yourself.	• The load balancer checks the health of the registered instances using either ○ the default health check configuration provided by Elastic Load Balancing or ○ a health check configuration that you configure (auto scaling or custom health checks for example)..	• Health check grace period does not start until lifecycle hook actions are completed and the instance enters the InService state.
• Instance Status Checks • Monitor the software and network configuration of your individual instance.	• Network Load Balancers use active and passive health checks to determine whether a target is available to handle requests.	• With custom health checks, you can send an instance's health information directly

Amazon EC2 checks the health of an instance by sending an address resolution protocol (ARP) request to the ENI. These checks detect problems that require your involvement to repair.	○ With active health checks, the load balancer periodically sends a request to each registered target to check its status. After each health check is completed, the load balancer node closes the connection that was established. ○ With passive health checks, the load balancer observes how targets respond to connections, which enables it to detect an unhealthy target before it is reported as unhealthy by active health checks. You cannot disable, configure, or monitor passive health checks.	from your system to Amazon EC2 Auto Scaling.

ELB Health Checks vs Route 53 Health Checks For Target Health Monitoring

Health Check Service	AWS Elastic Load Balancing	Amazon Route53
What is it for?	This health check periodically sends a request to a target instance, server or function to verify its status i.e. available to accept traffic requests.	This health check monitors the state of a record's target, which can be an EC2 instance, a server, or an AWS service that has an endpoint.
Target health check settings	You enter the port and common path of your targets that the load balancer will send the health check request to.	You enter the domain name or the IP address, port and patch that Route 53 will use to send the health check requests to if the record is a non-alias record, or by setting Evaluate target health to "Yes" if the record is an alias record.
Area span	Load balancers can monitor targets that span multiple availability zones but not multiple regions.	Route 53 monitors your targets regardless of their location, as long as they are reachable by Route 53.
Health check frequency	You specify a value between 5 seconds and 300 seconds.	Choose either every 10 seconds or every 30 seconds.
Response timeout	You can enter a value between 2 seconds and 60 seconds.	Cannot be configured.
Criteria to pass health check	You specify a threshold that a target a target should pass/ fail a health check to determine its status.	If more than 18% of health checkers report that an endpoint is healthy. Route 53 consider is healthy. If 18% of health checkers or fewer report that an endpoint is healthy, Route 53 considers it unhealthy. Route 53 health check servers are located in different locations worldwide.
Accessibility	Make sure targets are reachable by the load balancer. New targets can be easily added and removed from the load balancer.	Make sure endpoints are reachable and resolvable when users hit your URL. Due to DNS

		caching, it may take a while for new target endpoints to reflect end users.
Primary purpose	High availability and fault tolerance for your services.	DNS failover routing.

AWS CloudTrail vs Amazon CloudWatch

- CloudWatch is a monitoring service for AWS resources and applications. CloudTrail is a web service that records API activity in your AWS account. They are both useful monitoring tools in AWS.
- By default, CloudWatch offers free basic monitoring for your resources, such as EC2 instances, EBS volumes, and RDS DB instances. CloudTrail is also enabled by default when you create your AWS account.
- With CloudWatch, you can collect and track metrics, collect and monitor log files, and set alarms. CloudTrail, on the other hand, logs information on who made a request, the services used, the actions performed, the parameters for the actions, and the response elements returned by the AWS service. CloudTrail Logs are then stored in an S3 bucket or a CloudWatch Logs log group that you specify.
- You can enable detailed monitoring from your AWS resources to send metric data to CloudWatch more frequently, with an additional cost.
- CloudTrail delivers one free copy of management event logs for each AWS region. Management events include management operations performed on resources in your AWS account, such as when a user logs in to your account. Logging data events are charged. Data events include resource operations performed on or within the resource itself, such as S3 object-level API activity or Lambda function execution activity.
- CloudTrail helps you ensure compliance and regulatory standards.
- An Amazon CloudWatch Log reports on application logs, while an AWS CloudTrail Log provides you specific information on what occurred in your AWS account.
- Amazon EventBridge is a near real-time stream of system events describing changes to your AWS resources. CloudTrail focuses more on AWS API calls made in your AWS account.
- Typically, CloudTrail delivers an event within 15 minutes of the API call. CloudWatch delivers metric data in 5-minute periods for basic monitoring and 1-minute periods for detailed monitoring. The CloudWatch Logs Agent will send log data every five seconds by default.

CloudWatch Agent vs SSM Agent vs Custom Daemon Scripts

CloudWatch Agent	SSM Agent (AWS Systems Manager)	Custom Daemon Scripts
CloudWatch agent allows you to collect more system-level metrics from your EC2 and on-premises servers than just the standard CloudWatch metrics. Using the CloudWatch Agent, you can monitor operating system metrics such as: • `mem_used_percent` - percentage of memory currently in use • `disk_used_percent` - percentage of disk space utilized • `swap_used_percent` - percentage of swap space used • `cpu_usage_idle` - percentage of idle CPU time • `cpu_usage_user` - percentage of CPU time spent on user processes • `diskio_reads` and `diskio_writes` - disk I/O operations It also enables you to retrieve custom metrics from your applications or services using the *StatsD* and *collectd* protocols. *StatsD* is supported on both Linux servers and	SSM Agent is Amazon software that runs on your EC2 instances and your hybrid instances that are configured for Systems Manager. SSM Agent processes requests from the Systems Manager service in the cloud and configures your machine as specified in the request. You can manage servers without having to log in to them using automation. SSM Agent sends status and execution information back to the Systems Manager service by using the *EC2 Messaging* service. SSM Agent runs on Amazon EC2 instances using root permissions (Linux) or SYSTEM permissions (Windows). CloudWatch agent replaces SSM agent in sending metric logs to CloudWatch Logs.	You use custom scripts (such as cron or bash scripts) if the two previously mentioned agents do not fit your needs. CloudWatch agent is useful for collecting system-level metrics and logs. You can create custom scripts that perform some modifications before the metrics are sent out. SSM Agent is also useful for automation purposes, though Systems Manager does not have a document for every case scenario. You may also have some compliance requirements that would require SSM Agent to be disabled (recall that SSM agent runs at root level permissions).

servers running Windows Server. *collectd* is supported only on Linux servers. You can use CloudWatch agent to collect logs from your servers and send them to CloudWatch Logs. Metrics collected by the CloudWatch agent are billed as custom metrics. You can install CloudWatch Agent using three ways: • via Command Line • via SSM Agent • via AWS CloudFormation		

Latency Routing vs Geoproximity Routing vs Geolocation Routing

Latency Routing	
Definition	Lets Route 53 serve user requests from the AWS Region that provides the lowest latency. It does not, however, guarantee that users in the same geographic region will be served from the same location. Latency-based routing is based on latency measurements performed over a period of time, and the measurements reflect changes in network connectivity and routing.
How it works:	To use latency-based routing, you create latency records for your resources in multiple AWS Regions. When Route 53 receives a DNS query for your domain or subdomain, it determines which AWS Regions you've created latency records for, determines which region gives the user the lowest latency, and then selects a latency record for that region. Route 53 responds with the value from the selected record, such as the IP address for a web server. Record sets can be created using any record type supported by Route 53, except NS or SOA records.
Use Case	Use when you have resources in multiple AWS Regions and you want to route traffic to the region that provides the best latency.
Geoproximity Routing	
Definition	Lets Amazon Route 53 route traffic to your resources based on the geographic location of your users and your resources. You can also optionally choose to route more traffic or less to a given resource by specifying a value, known as a bias. A bias expands or shrinks the size of the geographic region from which traffic is routed to a resource.
How it works:	To use geoproximity routing, you must use Route 53 traffic flow. You create traffic flow policies for your resources and specify one of the following values for each policy: • If you're using AWS resources, you can set the AWS Region where your resource is created • If you're using non-AWS resources, you can enter the latitude and longitude of the resource

Use Case	Use when you want to route traffic based on the location of your resources and, optionally, shift traffic from resources in one location to resources in another.
Geolocation Routing	
Definition	Resources serve traffic based on the geographic location of your users, meaning the location that DNS queries originate from.
How it works:	Geolocation works by mapping IP addresses to locations. Some IP addresses aren't mapped to geographic locations, so Amazon Route 53 will receive some DNS queries from locations that it can't identify. You can create a default record that handles both queries from IP addresses that aren't mapped to any location and queries that come from locations that you haven't created geolocation records for. If you don't create a default record, Route 53 returns a "no answer" response for queries from those locations. No two records should specify the same geographic location.
Use Case	Use when you want to route traffic based on the location of your users. • You can localize your content and present some or all of your website in the language of your users. • You can restrict the distribution of content to only the locations in which you have distribution rights. • Useful for balancing load across endpoints in a predictable, easy-to-manage way so that each user location is consistently routed to the same endpoint.

Service Control Policies vs IAM Policies

Service Control Policies (SCP)	IAM Policies
• SCPs are mainly used along with AWS Organizations organizational units (OUs).	• IAM Policies operate at the Principal level.

• SCPs do not replace IAM Policies such that they do not provide actual permissions. To perform an action, you would still need to grant appropriate IAM Policy permissions.	• There are two types of IAM policies
	  ○ Identity-based policies – attached to an IAM user, group, or role.
	  ○ Resource-based policies – attached to an AWS resource such as an S3 bucket.

• Even if a Principal is allowed to perform a certain action (granted through IAM Policies), an attached SCP will override that capability if it enforces a Deny on that action. SCP takes precedence over IAM Policies.	• IAM Policies can grant/deny a Principal permissions to perform certain actions to certain resources. This can be used together with SCP to ensure stricter controls in AWS Organizations.
• SCPs can be applied to the root of an organization or to individual accounts in an OU.	• An IAM policy can be applied only to IAM users, groups, or roles, and it can never restrict the root identity of the AWS account.
• When you apply an SCP to an OU or an individual AWS account, you choose to either enable (whitelist), or disable (blacklist) the specified AWS service. Access to any service that isn't explicitly allowed by the SCPs associated with an account, its parent OUs, or the management account is denied to the AWS accounts or OUs associated with the SCP.	• IAM Policies cannot be attached to OUs.
• Any account has only those permissions permitted by every parent above it. If a permission is blocked at any level above the account, either implicitly (by not being included in an Allow policy statement) or explicitly (by being included in a Deny policy	• An IAM Policy can allow or deny actions. An explicit allow overrides an implicit deny. An explicit deny overrides an explicit allow.

statement), a user or role in the affected account can't use that permission, even if there is an attached IAM policy granting Administrator permissions to the user.	
• SCPs affect only principals that are managed by accounts that are part of the organization.	

S3 Pre-Signed URLS vs CloudFront Signed URLS vs Origin Access Control

S3 Pre-signed URLs	CloudFront Signed URLs	Origin Access Control (OAC)
All S3 buckets and objects by default are **private**. Only the object owner has permission to access these objects. Pre-signed URLs use the owner's security credentials to grant others time-limited permission to download or upload objects.	You can control user access to your private content in two ways • Restrict access to files in CloudFront edge caches • Restrict access to files in your Amazon S3 bucket (unless you've configured it as a website endpoint)	You can configure an S3 bucket as the origin of a CloudFront distribution. OAC prevents users from viewing your S3 files by simply using the direct URL for the file. Instead, they would need to access it through a CloudFront URL.
When creating a pre-signed URL, you (as the owner) need to provide the following: • Your security credentials • An S3 bucket name • An object key • Specify the HTTP method (GET to download the object or PUT to upload an object) • Expiration date and time of the URL.	You can configure CloudFront to require that users access your files using either **signed URLs** or **signed cookies**. You then develop your application either to create and distribute signed URLs to authenticated users or to send Set-Cookie headers that set signed cookies on the viewers for authenticated users. When you create signed URLs or signed cookies to control access to your files, you can specify the following restrictions: • An expiration date and time for the URL • (Optional) The date and time the URL becomes valid • (Optional) The IP address or range of addresses of the computers that can be used to access your content You can use signed URLs or signed cookies for any CloudFront distribution, regardless of whether the origin is an Amazon S3 bucket or an HTTP server.	To require that users access your content through CloudFront URLs, you perform the following tasks: • Create a special CloudFront user called an **origin access control**. • Give the origin access Control permission to read the files in your bucket. • Remove permission for anyone else to use Amazon S3 URLs to read the files (through bucket policies or ACLs). You cannot set OAC if your S3 bucket is configured as a website endpoint.

.

SNI Custom SSL vs Dedicated IP Custom SSL

Server Name Indication (SNI) Custom SSL	Dedicated IP Custom SSL
• Relies on the SNI extension of the TLS protocol, which allows multiple domains to serve SSL traffic over the same IP address.	• Mainly useful for browsers that do not support SNI.
• Offers the same level of security when using Dedicated IP Custom SSL.	• For this feature, the Amazon content delivery network allocates dedicated IP addresses to serve your SSL content at each Edge location.
• If you configure CloudFront to serve HTTPS requests using SNI, CloudFront associates your alternate domain name with an IP address for each edge location. The IP address to your domain name is determined during the SSL/TLS handshake negotiation, and isn't dedicated to your distribution.	• You will need to upload a SSL certificate and associate it with your CloudFront distributions.
• Some older browsers do not support SNI and will not be able to establish a connection with CloudFront to load the HTTPS version of your content.	• You can associate more than two custom SSL certificate with your AWS Account by submitting a CloudFront Limit Increase Form.
• You can use SNI Custom SSL with no upfront or monthly fees for certificate management.	• This method works for every HTTPS request, regardless of the browser or other viewer that the user is using.
	• Because of the added cost associated with dedicating IP addresses per SSL certificate, AWS charges a fixed monthly fee of $600 for each custom SSL certificate you associate with your content delivery network distributions, pro-rated by the hour.
	• You can switch to using a custom SSL/TLS certificate with SNI instead and eliminate the charge that is associated with dedicated IP addresses.

Redis (cluster mode enabled vs disabled) vs Memcached

	Redis (cluster mode enabled)	Redis (cluster mode disabled)	Memcached
Data Types	string, sets, sorted sets, lists, hashes, bitmaps, hyperloglog, geospatial indexes	string, sets, sorted sets, lists, hashes, bitmaps, hyperloglog, geospatial indexes	string, objects (like databases)
Data Partitioning (distribute your data among multiple nodes)	Supported	Unsupported	Supported
Modifiable cluster	Only versions 3.2.10 and later	Yes	Yes
Online resharding	Only versions 3.2.10 and later	No	No
Encryption	3.2.6, 4.0.10 and later	3.2.6, 4.0.10 and later	Unsupported
Sub-millisecond latency	Yes	Yes	Yes
FedRAMP, PCI DSS and HIPAA compliant	3.2.6, 4.0.10 and later	3.2.6, 4.0.10 and later	No
Multi-threaded (make use of multiple processing cores)	No	No	No
Node type upgrading	No	Yes	No
Engine upgrading	Yes		
Cluster replication (create multiple copies of a primary cluster)	Supported	Supported	Unsupported
Multi-AZ for automatic failover	Required	Optional	Unsupported
Transactions (execute a group of commands as an isolated and atomic operation)	Supported	Supported	Unsupported
Pub/Sub capability	Yes	Yes	No

Backup and restore (keep your data on disk with a point in time snapshot)	Supported	Supported	Unsupported
Lua Scripting (execute transactional Lua scripts)	Supported	Supported	Unsupported
Use case	• You need to partition your data across two to 250 or 500 nodes if the Redis engine version is 5.0.6 or higher.(clustered mode only). • You need geospatial indexing (clustered mode or non-clustered mode). • You don't need to support multiple databases. • Plus features of non-clustered mode	• You need complex data types, such as strings, hashes, lists, sets, sorted sets, and bitmaps. • You need to sort or rank in-memory datasets. • You need persistence of your key store. • You need to replicate your data from the primary to one or more read replicas for read-intensive applications. • You need automatic failover if your primary node fails. • You need pub/sub capabilities. • You need backup and restore capabilities. • You need to support multiple databases.	• You need the simplest model possible. • You need to run large nodes with multiple cores or threads. • You need the ability to scale out and in, adding and removing nodes as demand on your system increases and decreases. • You need to cache objects, such as a database. • Needs Auto Discovery to simplify the way an application connects to a cluster.

FINAL REMARKS AND TIPS

Cloud Operations and Systems Management are never easy. There is a lot of care and perseverance needed to make sure your development team has a working environment to launch and deploy their applications. Beyond that, you also need to ensure that they continue working correctly as expected through continuous monitoring. In mission-critical enterprise systems, many underlying components, such as networks and access controls, cannot afford failures. CloudOps is often disregarded compared to other fields, but it is undeniable that it is the backbone of every IT infrastructure.

Since you are at the end of our eBook, we hope that the concepts we covered here were able to help you prepare for your AWS Certified CloudOps Engineer - Associate certification exam. We crafted this eBook as a way for you to review the important concepts that might appear in the actual exam, but it also serves to train you for the actual CloudOps Engineer role. As more and more customers are moving to the cloud to use it for their products and services, system management has become quite a daunting task that only a few people like you can face head-on. As evidently shown in our exam overview section, there are many domains that you should study for and even more services to familiarize yourself with. We know that these headwinds are not going to be easy, but we also know that you can succeed in your endeavors and provide yourself with more opportunities for career advancement. We are with you in every step of your AWS journey.

And with that, we at Tutorials Dojo thank you for supporting us through this eBook. If you wish to validate what you have learned so far, now is a great time to check out our **AWS Certified CloudOps Engineer Associate Practice Exams** and **AWS Hands-On Labs**. You can also try the free sampler version of our full practice test course here. It will fill in the gaps in your knowledge that you are not aware of and give you a sense of the actual exam environment. That way, you'll know what to expect in the actual exam, and you can pace yourself through the questions better. If you have any issues, concerns, or constructive feedback on our eBook, feel free to contact us at support@tutorialsdojo.com.

Good luck with your exam, and we'd love to hear back from you soon.

Your Learning Partners,
Jon Bonso, and the Tutorials Dojo Team

ABOUT THE AUTHOR

Jon Bonso (10x AWS Certified)

Born and raised in the Philippines, Jon is the Co-Founder of **Tutorials Dojo**. Now based in Sydney, Australia, he has over a decade of diversified experience in Banking, Financial Services, and Telecommunications. He's 10x AWS Certified, an AWS Community Builder, and has worked with various cloud services such as Google Cloud, and Microsoft Azure. Jon is passionate about what he does and dedicates a lot of time creating educational courses. He has given IT seminars to different universities in the Philippines for free and has launched educational websites using his own money and without any external funding.