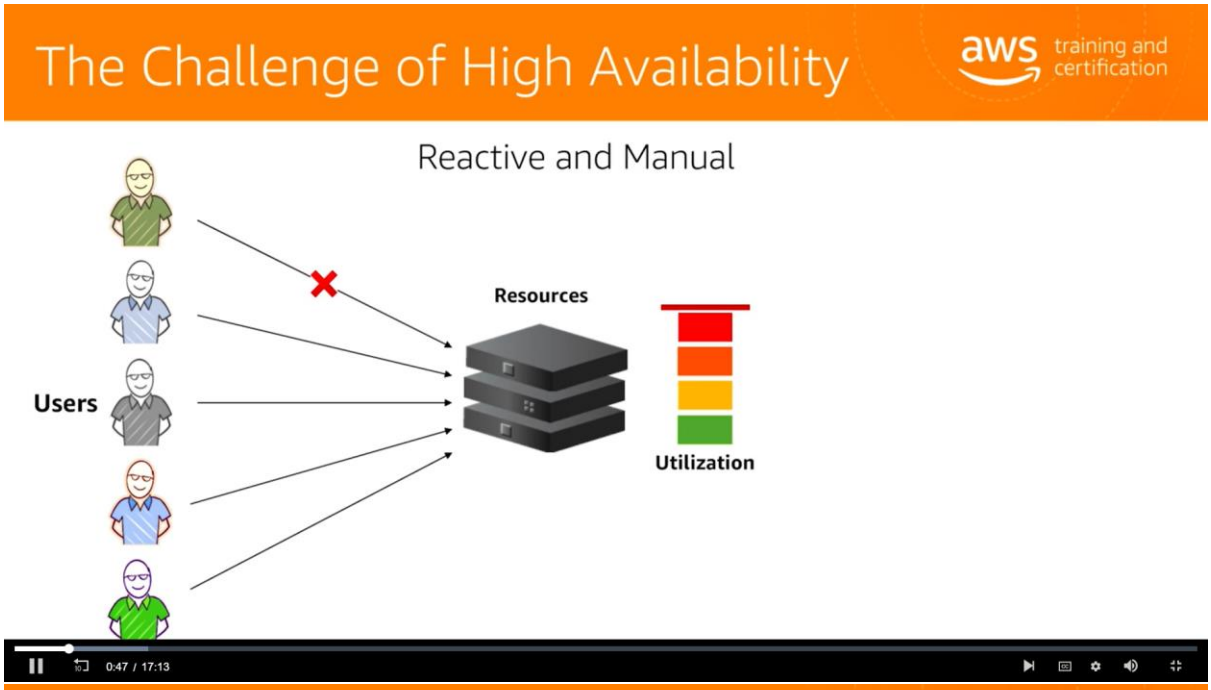
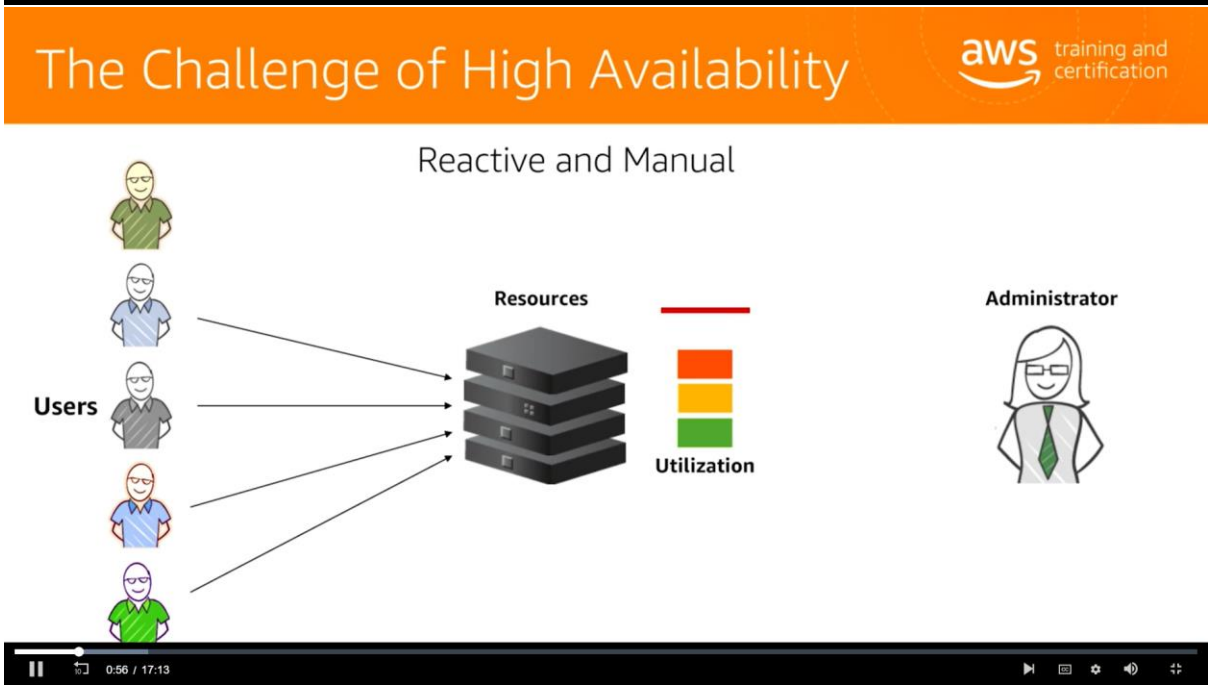

The Challenge of High Availability
aws training and certification
Reactive and Manual
Users
Resources
Utilization
0:47 / 17:13


The Challenge of High Availability
aws training and certification
Reactive and Manual
Users
Resources
Utilization
Administrator
0:56 / 17:13

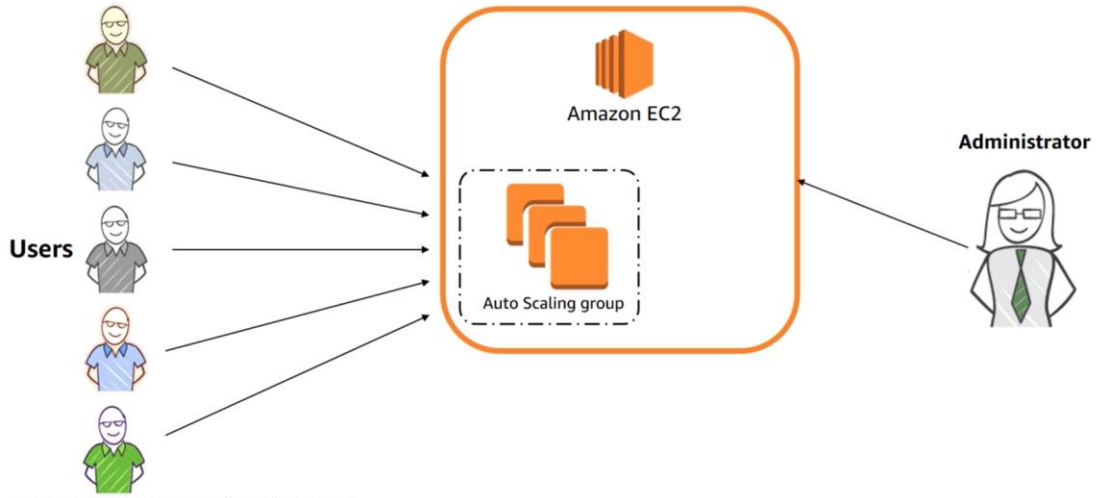
Automatic Scaling on AWS

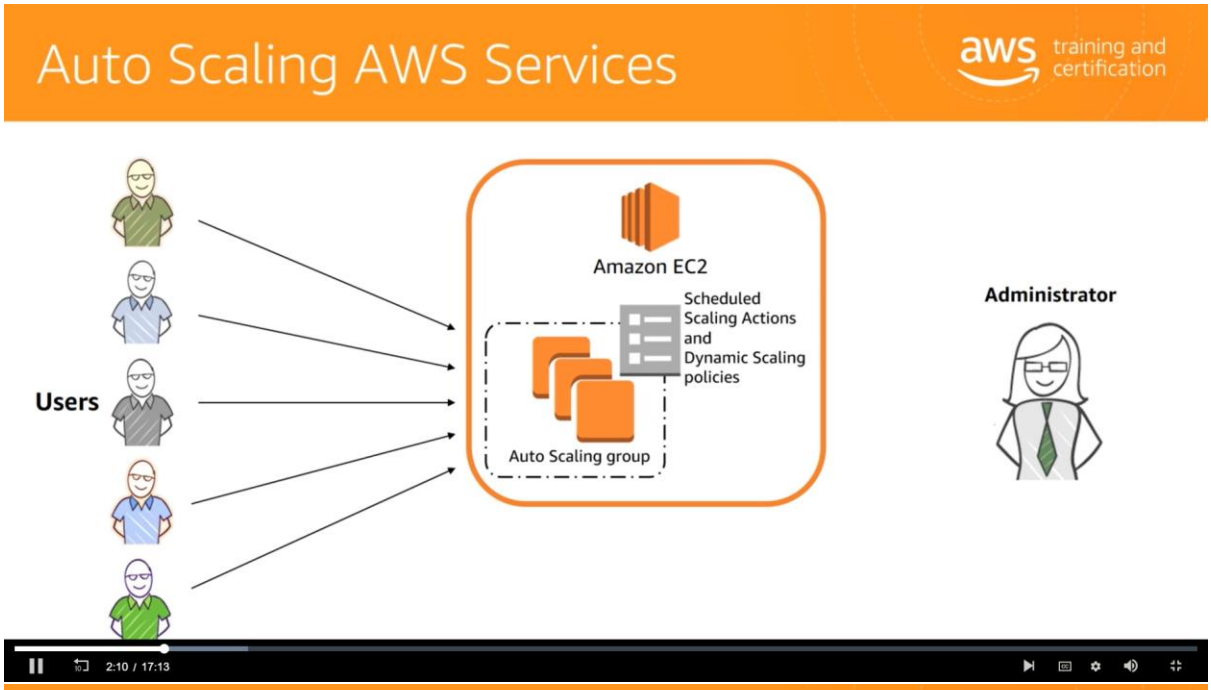
aws training and certification

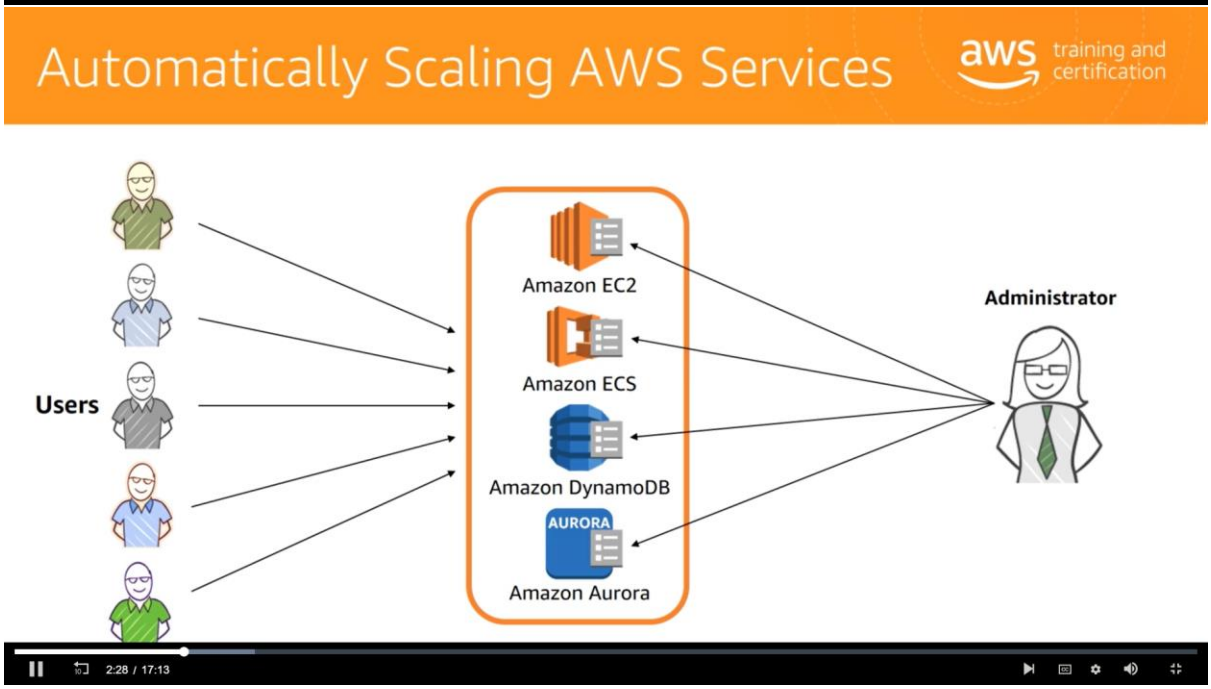
1:16 / 17:13

Auto Scaling AWS Services

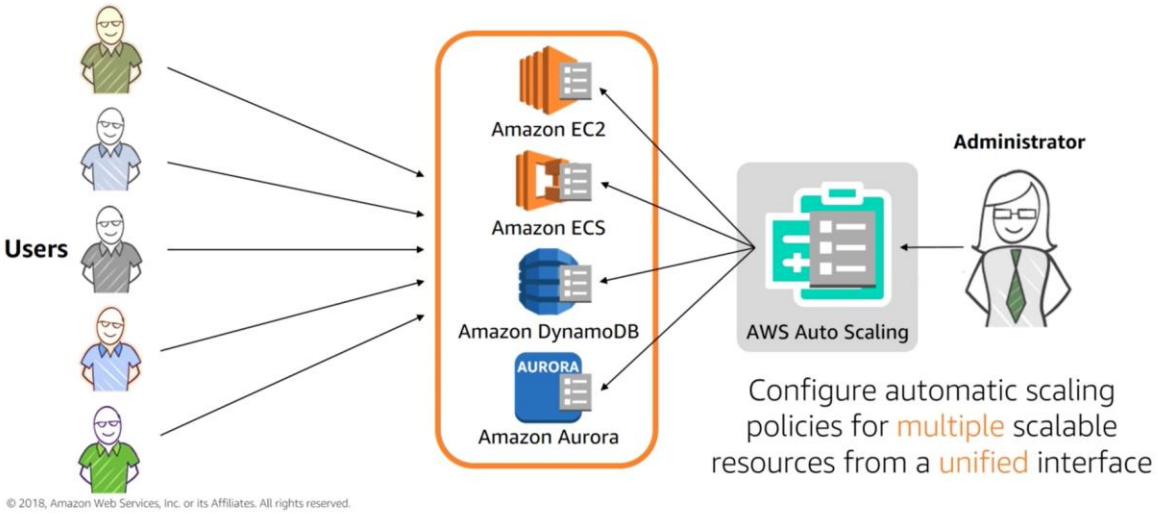
aws training and certification




Auto Scaling AWS Services
aws training and certification
Users
Amazon EC2
Scheduled Scaling Actions and Dynamic Scaling policies
Auto Scaling group
Administrator
2:10 / 17:13


Automatically Scaling AWS Services
aws training and certification
Users
Amazon EC2
Amazon ECS
Amazon DynamoDB
AURORA
Amazon Aurora
Administrator
2:28 / 17:13

Introducing AWS Auto Scaling



Overview

AWS Auto Scaling:

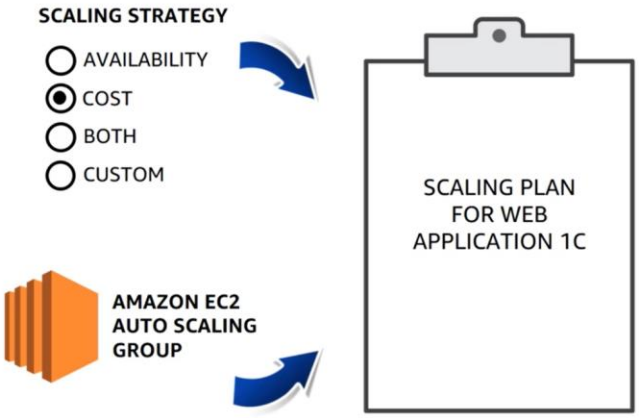
- Leverages existing Amazon EC2 Auto Scaling and Application Auto Scaling services
- Enables you to select an application as defined by AWS CloudFormation stack or in AWS Elastic Beanstalk

Configuring AWS Auto Scaling

aws training and certification

Scaling Plan Overview

aws training and certification



AWS Auto Scaling Benefits

aws training and certification

SCALE EFFICIENTLY WITH ONE INTERFACE

SCALE INTELLIGENTLY WITH SCALING PLANS

MAINTAIN PERFORMANCE WITH CONTINUAL MONITORING

FREE SERVICE: PAY ONLY FOR RESOURCES YOU need

Unified Scaling

aws training and certification

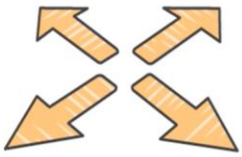
	Amazon EC2	Amazon EC2	Amazon ECS	Amazon DynamoDB	Amazon Aurora
Scalable Target	Auto Scaling Group	Spot Fleets	Service	Table or GSI	Cluster
Scalable Dimension	Amazon EC2 instances	Amazon EC2 Spot instances	Tasks	Provisioned Capacity	Aurora Replicas

Predictable Scaling

aws training and certification

Built-In Scaling Strategies

aws training and certification



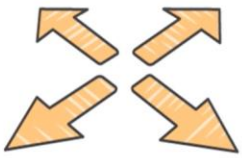
Optimize for Availability

Balance Availability and Cost

Optimize for Cost

Built-In Scaling Strategies

aws training and certification



Optimize for Availability

Low resource utilization target

Balance Availability and Cost

Moderate resource utilization target

Optimize for Cost

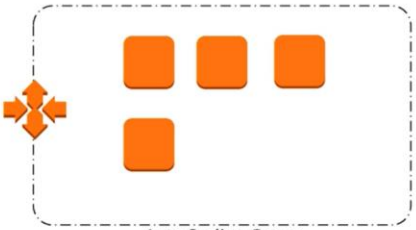
High resource utilization target

Target Tracking Scaling Policies

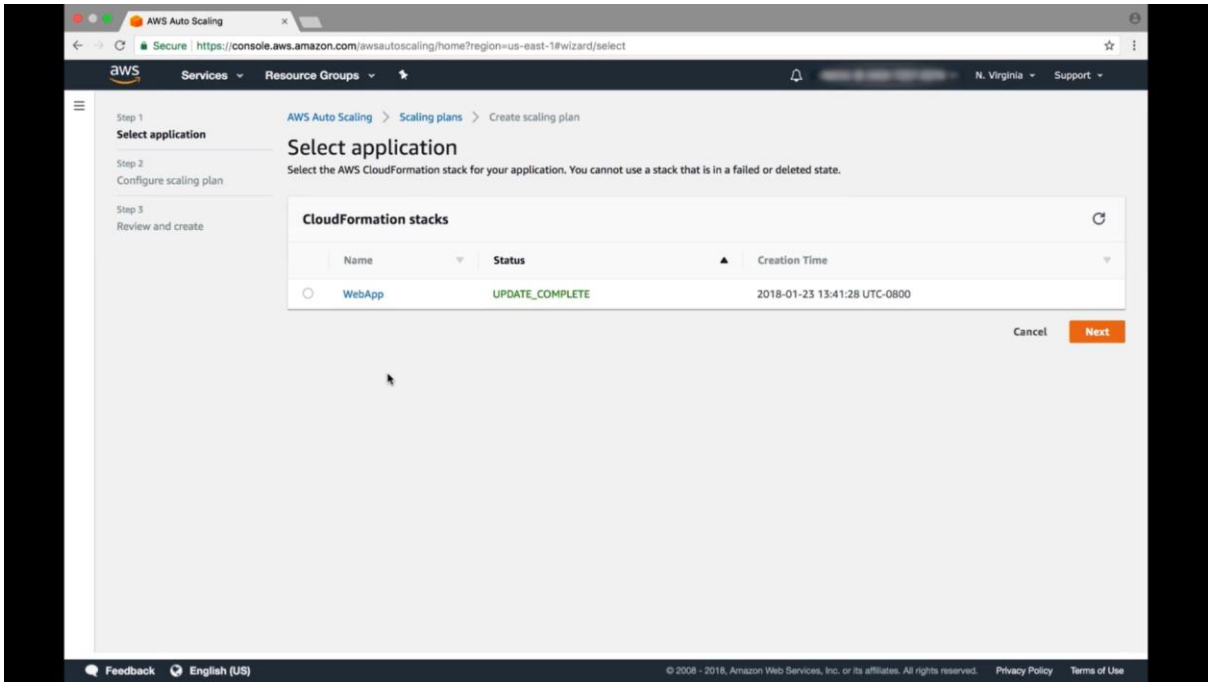
aws training and certification

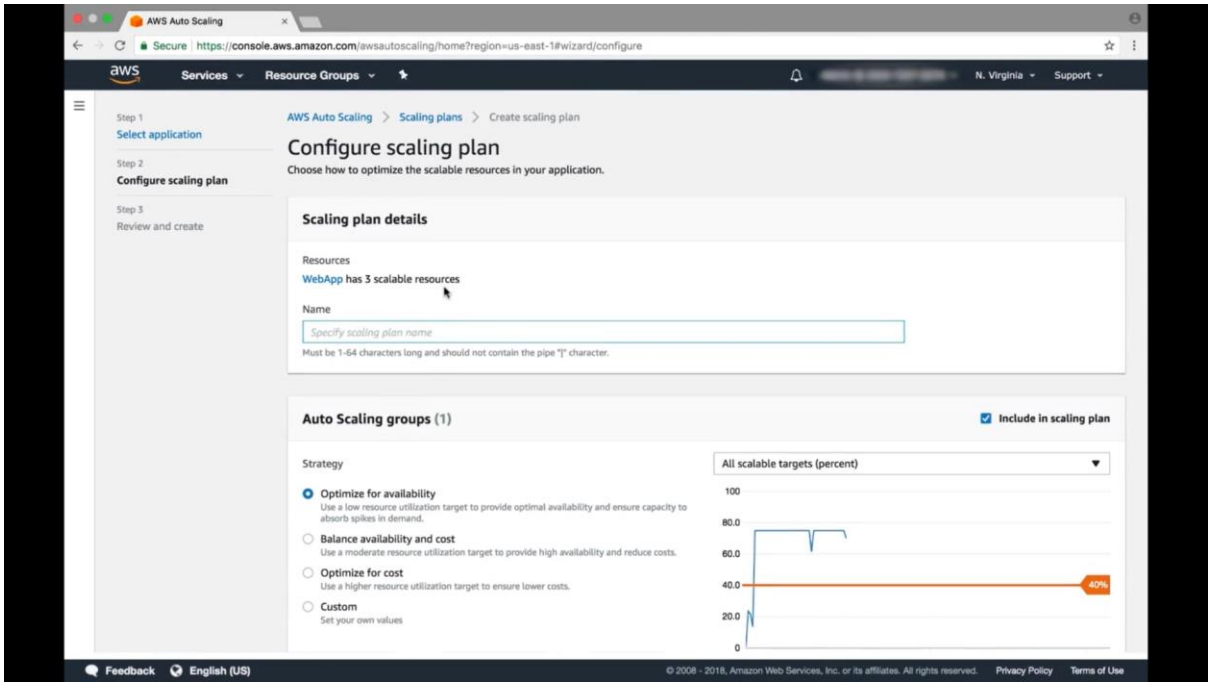
METRIC: CPU UTILIZATION

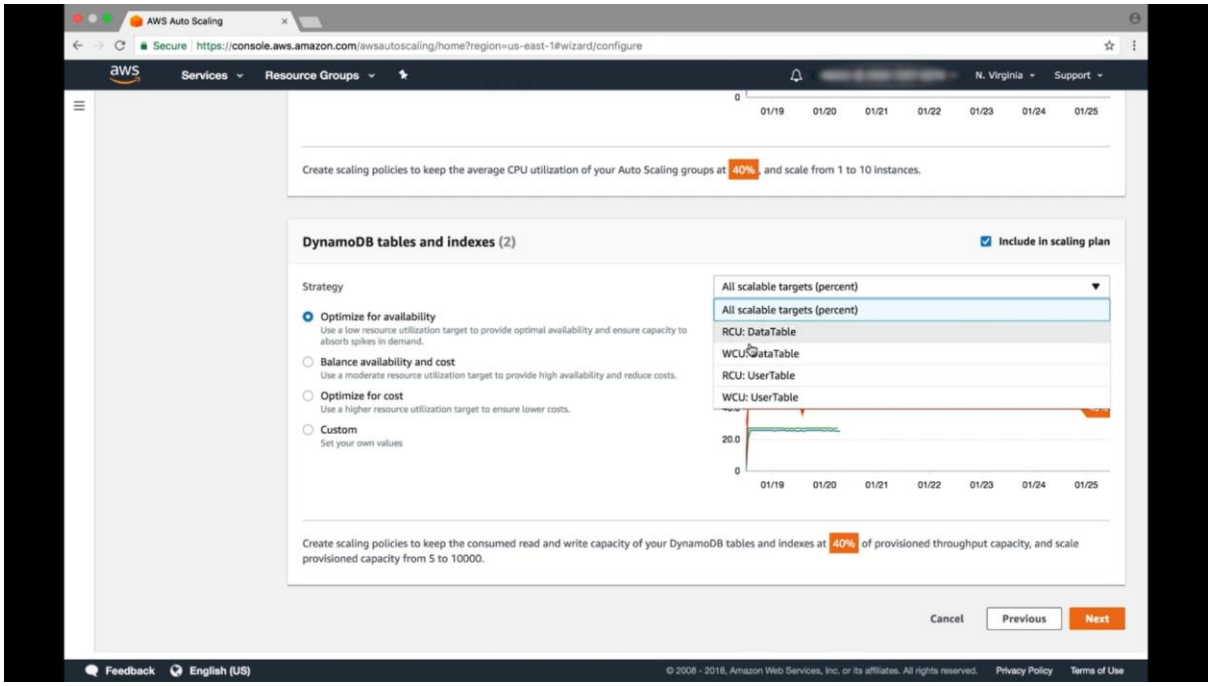
TARGET VALUE: 40%

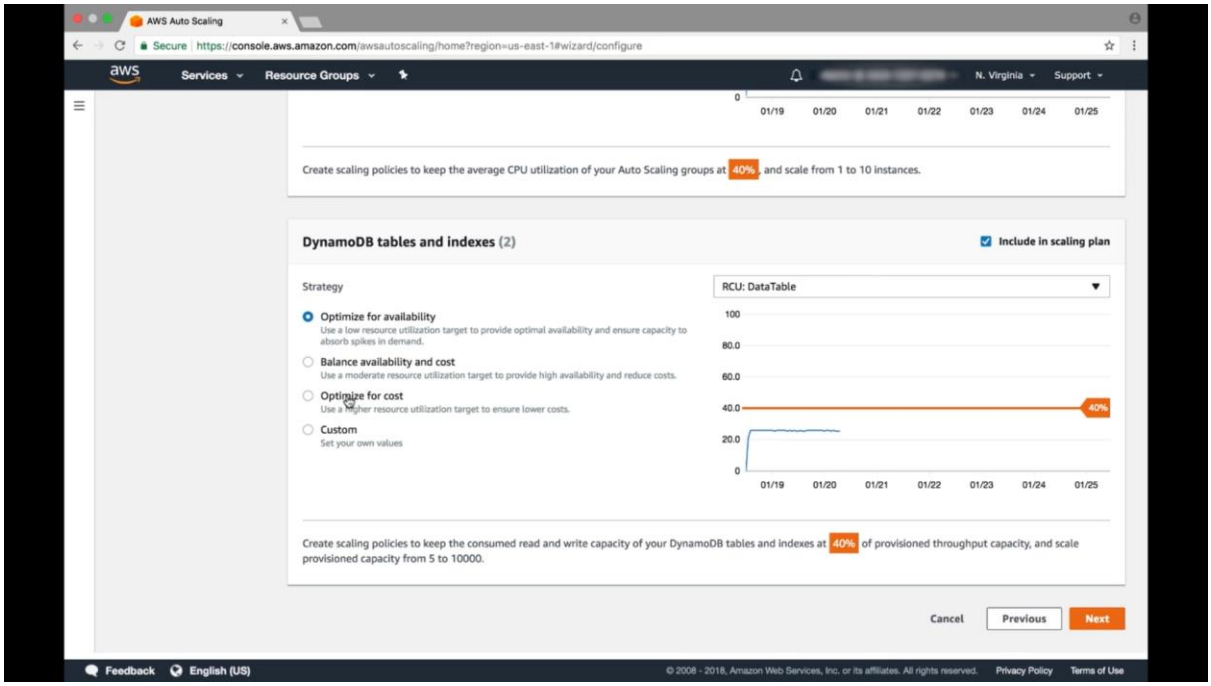


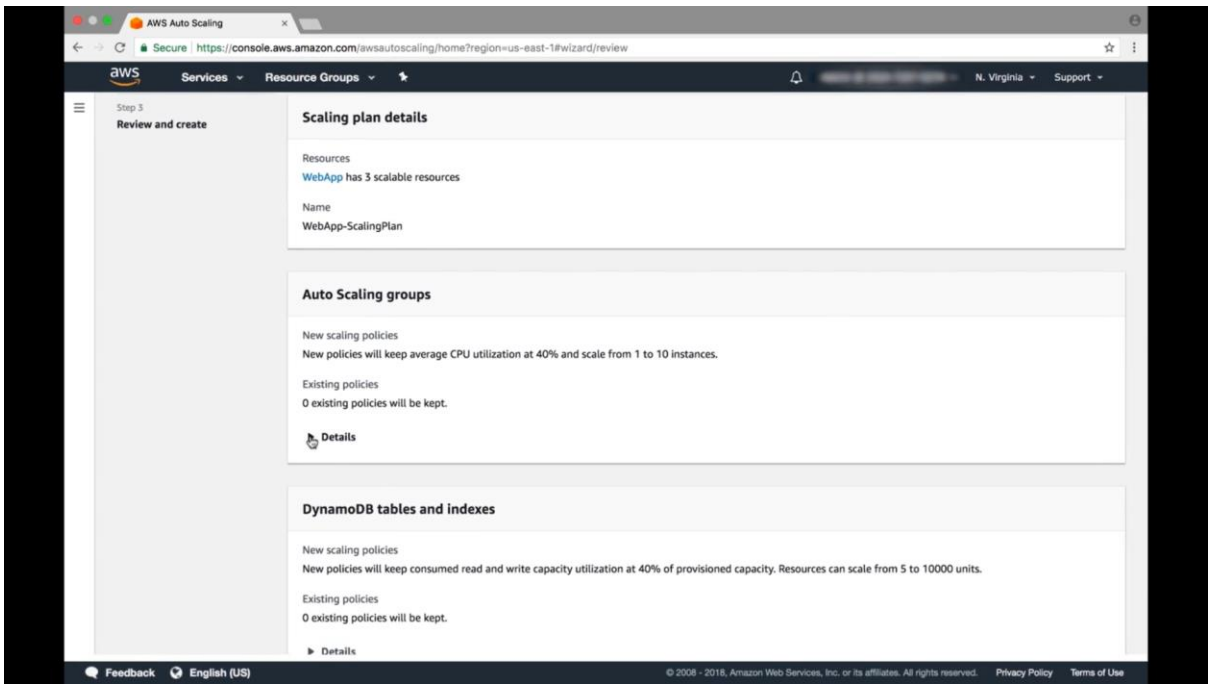
Auto Scaling Group

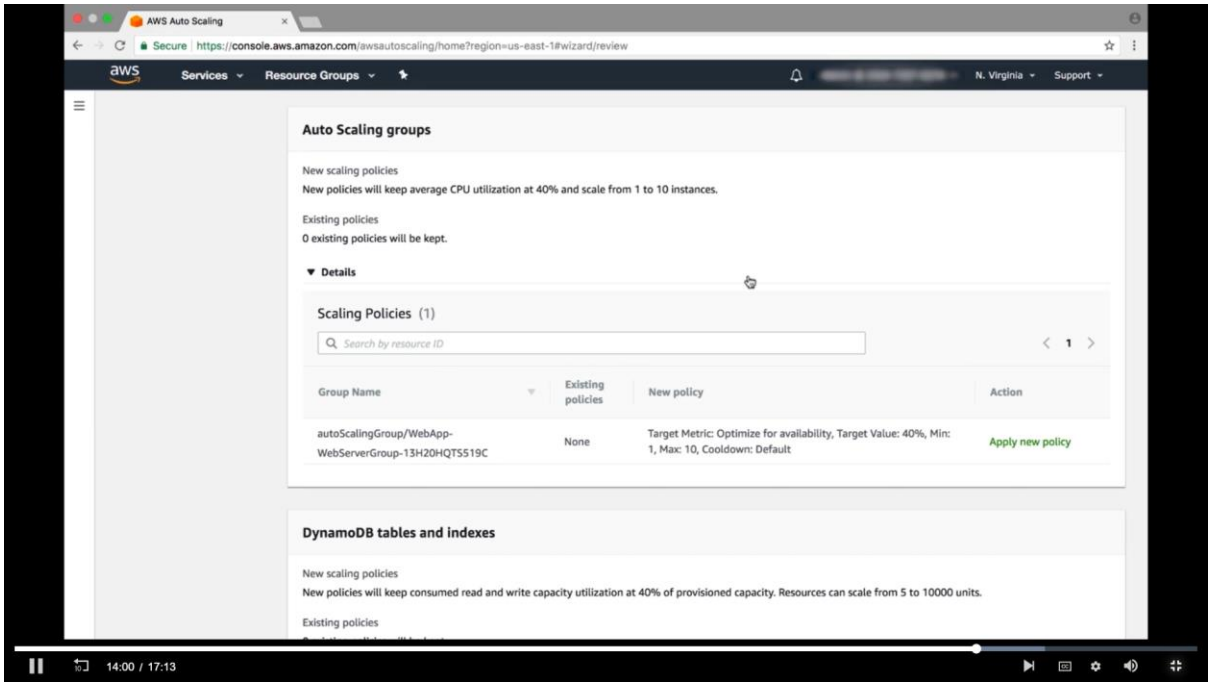


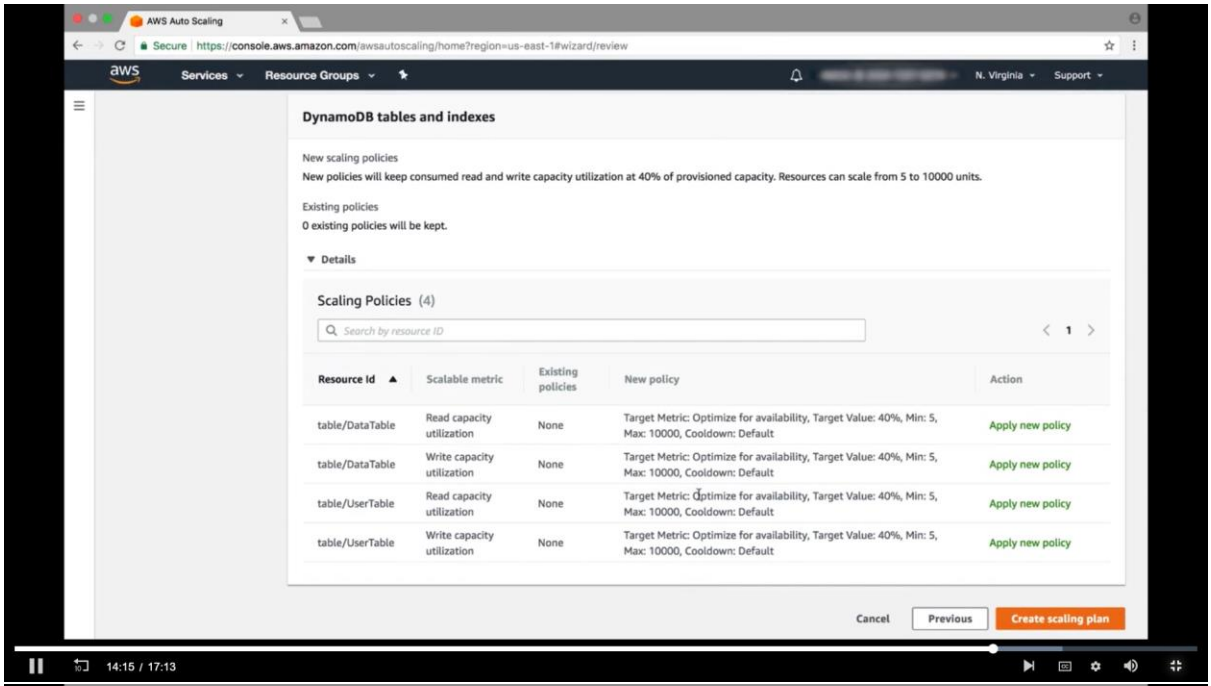

AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#wizard/configure
aws
Services
Resource Groups
N. Virginia
Support
Step 1
Select application
Step 2
Configure scaling plan
Step 3
Review and create
AWS Auto Scaling > Scaling plans > Create scaling plan
Configure scaling plan
Choose how to optimize the scalable resources in your application.
Scaling plan details
Resources
WebApp has 3 scalable resources
Name
Specify scaling plan name
Must be 1-64 characters long and should not contain the pipe "|" character.
Auto Scaling groups (1)
Include in scaling plan
Strategy
All scalable targets (percent)
Optimize for availability
Use a low resource utilization target to provide optimal availability and ensure capacity to absorb spikes in demand.
Balance availability and cost
Use a moderate resource utilization target to provide high availability and reduce costs.
Optimize for cost
Use a higher resource utilization target to ensure lower costs.
Custom
Set your own values
100
80.0
60.0
40.0
20.0
0
40%
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

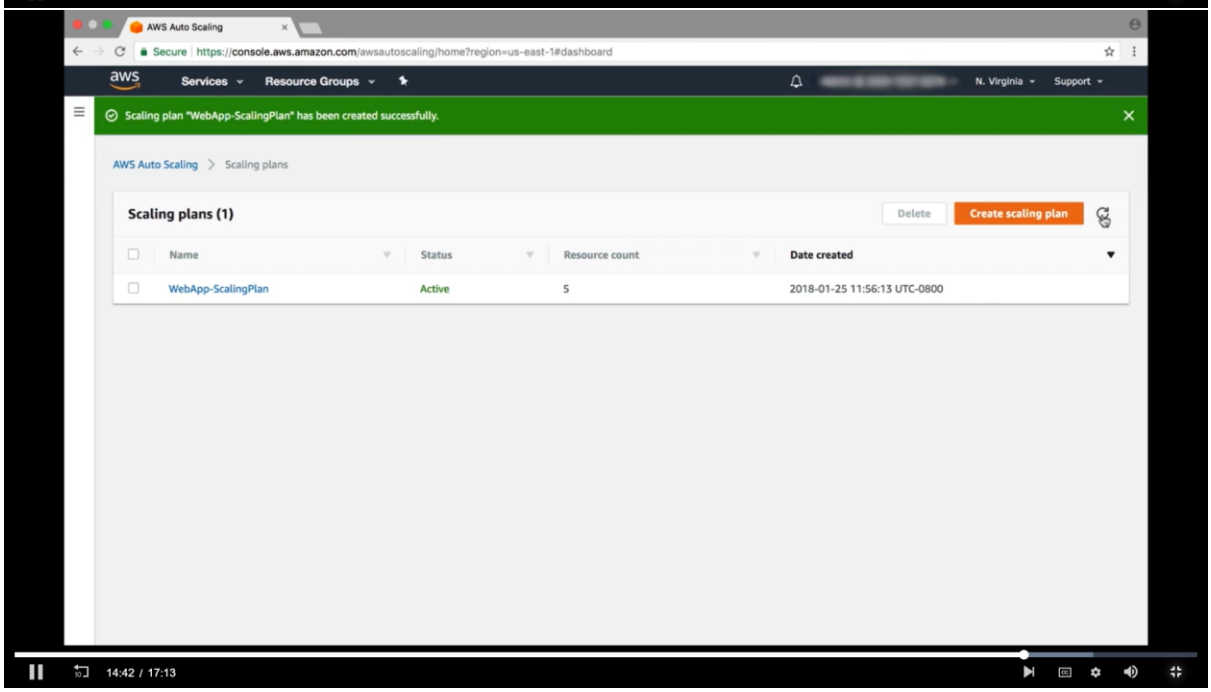

AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#wizard/configure
aws
Services
Resource Groups
N. Virginia
Support
0
01/19
01/20
01/21
01/22
01/23
01/24
01/25
Create scaling policies to keep the average CPU utilization of your Auto Scaling groups at 40% and scale from 1 to 10 instances.
DynamoDB tables and indexes (2)
Include in scaling plan
Strategy
All scalable targets (percent)
All scalable targets (percent)
RCU: DataTable
WCU: DataTable
RCU: UserTable
WCU: UserTable
Optimize for availability
Use a low resource utilization target to provide optimal availability and ensure capacity to absorb spikes in demand.
Balance availability and cost
Use a moderate resource utilization target to provide high availability and reduce costs.
Optimize for cost
Use a higher resource utilization target to ensure lower costs.
Custom
Set your own values
20.0
0
Create scaling policies to keep the consumed read and write capacity of your DynamoDB tables and indexes at 40% of provisioned throughput capacity, and scale provisioned capacity from 5 to 10000.
Cancel
Previous
Next
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

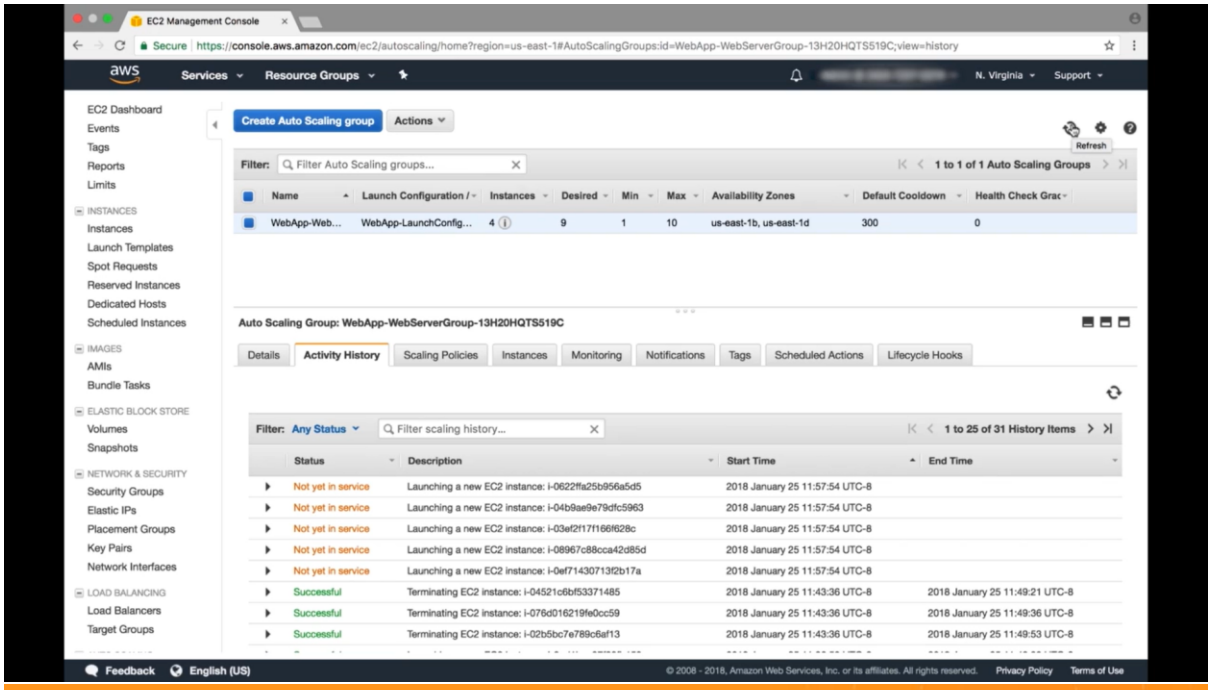

AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#wizard/configure
Services
Resource Groups
N. Virginia
Support
0
01/19 01/20 01/21 01/22 01/23 01/24 01/25
Create scaling policies to keep the average CPU utilization of your Auto Scaling groups at 40% and scale from 1 to 10 instances.
DynamoDB tables and indexes (2)
Include in scaling plan
Strategy
RCU: DataTable
Optimize for availability
Use a low resource utilization target to provide optimal availability and ensure capacity to absorb spikes in demand.
Balance availability and cost
Use a moderate resource utilization target to provide high availability and reduce costs.
Optimize for cost
Use a higher resource utilization target to ensure lower costs.
Custom
Set your own values
100
80.0
60.0
40.0
20.0
0
40%
01/19 01/20 01/21 01/22 01/23 01/24 01/25
Create scaling policies to keep the consumed read and write capacity of your DynamoDB tables and indexes at 40% of provisioned throughput capacity, and scale provisioned capacity from 5 to 10000.
Cancel
Previous
Next
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#wizard/review
Services
Resource Groups
N. Virginia
Support
Step 3
Review and create
Scaling plan details
Resources
WebApp has 3 scalable resources
Name
WebApp-ScalingPlan
Auto Scaling groups
New scaling policies
New policies will keep average CPU utilization at 40% and scale from 1 to 10 instances.
Existing policies
0 existing policies will be kept.
Details
DynamoDB tables and indexes
New scaling policies
New policies will keep consumed read and write capacity utilization at 40% of provisioned capacity. Resources can scale from 5 to 10000 units.
Existing policies
0 existing policies will be kept.
Details
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#wizard/review
aws
Services
Resource Groups
N. Virginia
Support
Auto Scaling groups
New scaling policies
New policies will keep average CPU utilization at 40% and scale from 1 to 10 instances.
Existing policies
0 existing policies will be kept.
Details
Scaling Policies (1)
Search by resource ID
1
Group Name
Existing policies
New policy
Action
autoScalingGroup/WebApp-WebServerGroup-13H20HQTS519C
None
Target Metric: Optimize for availability, Target Value: 40%, Min: 1, Max: 10, Cooldown: Default
Apply new policy
DynamoDB tables and indexes
New scaling policies
New policies will keep consumed read and write capacity utilization at 40% of provisioned capacity. Resources can scale from 5 to 10000 units.
Existing policies
14:00 / 17:13


AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#wizard/review
Services
Resource Groups
N. Virginia
Support
DynamoDB tables and indexes
New scaling policies
New policies will keep consumed read and write capacity utilization at 40% of provisioned capacity. Resources can scale from 5 to 10000 units.
Existing policies
0 existing policies will be kept.
Details
Scaling Policies (4)
Search by resource ID
Resource Id
Scalable metric
Existing policies
New policy
Action
table/DataTable
Read capacity utilization
None
Target Metric: Optimize for availability, Target Value: 40%, Min: 5, Max: 10000, Cooldown: Default
Apply new policy
table/DataTable
Write capacity utilization
None
Target Metric: Optimize for availability, Target Value: 40%, Min: 5, Max: 10000, Cooldown: Default
Apply new policy
table/UserTable
Read capacity utilization
None
Target Metric: Optimize for availability, Target Value: 40%, Min: 5, Max: 10000, Cooldown: Default
Apply new policy
table/UserTable
Write capacity utilization
None
Target Metric: Optimize for availability, Target Value: 40%, Min: 5, Max: 10000, Cooldown: Default
Apply new policy
Cancel
Previous
Create scaling plan
14:15 / 17:13


AWS Auto Scaling
Secure | https://console.aws.amazon.com/awsautoscaling/home?region=us-east-1#dashboard
Services
Resource Groups
N. Virginia
Support
Scaling plan "WebApp-ScalingPlan" has been created successfully.
AWS Auto Scaling
Scaling plans
Scaling plans (1)
Delete
Create scaling plan
Name
Status
Resource count
Date created
WebApp-ScalingPlan
Active
5
2018-01-25 11:56:13 UTC-0800
14:42 / 17:13



Considerations

aws training and certification

When should you consider using AWS Auto Scaling?

- If you have applications that use one or more scalable resources, and have a variable load

Considerations

aws training and certification

When should you consider using AWS Auto Scaling?

- If you have applications that use one or more scalable resources, and have a variable load
- If you want more guidance on defining your application scaling plan

Key Takeaways

aws training and certification

- Monitor your applications continually and automatically adjust capacity.
- Set up scaling efficiently through automatic resource discovery via a single unified interface.
- Make smart decisions regarding application high availability with built-in scaling strategies and predictable scaling.
- Maintain application performance automatically with smart scaling policies.