



사용자 가이드

Application Auto Scaling



Application Auto Scaling: 사용자 가이드

Copyright © 2024 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon의 상표 및 브랜드 디자인은 Amazon 외 제품 또는 서비스와 함께, 브랜드 이미지를 떨어뜨리거나 고객에게 혼동을 일으킬 수 있는 방식으로 사용할 수 없습니다. Amazon이 소유하지 않은 기타 모든 상표는 해당 상표 소유자의 자산이며, 해당 상표의 소유자가 Amazon의 계열사이거나 Amazon과 제휴 관계에 있거나 Amazon의 후원을 받는 업체일 수 있습니다.

Table of Contents

Application Auto Scaling이란?	1
Application Auto Scaling의 기능	1
Application Auto Scaling 작업	2
개념	3
자세히 알아보기	4
통합되는 서비스	5
아마존 AppStream 2.0	7
서비스 연결 역할	7
서비스 보안 주체	8
Application Auto Scaling을 사용하여 AppStream 2.0 플릿을 확장 가능한 타겟으로 등록	8
관련 리소스	9
Amazon Aurora	9
서비스 연결 역할	9
서비스 보안 주체	9
Application Auto Scaling을 통해 Aurora DB 클러스터를 확장 가능 대상으로 등록	10
관련 리소스	11
Amazon Comprehend	11
서비스 연결 역할	11
서비스 보안 주체	11
Application Auto Scaling을 통해 Amazon Comprehend 리소스를 확장 가능 대상으로 등록	11
관련 리소스	13
Amazon DynamoDB	13
서비스 연결 역할	13
서비스 보안 주체	13
Application Auto Scaling을 통해 DynamoDB 리소스를 확장 가능 대상으로 등록	14
관련 리소스	16
Amazon ECS	16
서비스 연결 역할	16
서비스 보안 주체	17
Application Auto Scaling을 통해 ECS 서비스를 확장 가능 대상으로 등록	17
관련 리소스	18
아마존 ElastiCache	18
서비스 연결 역할	18
서비스 보안 주체	19

Application Auto Scaling을 사용하여 Redis 복제 그룹을 확장 가능한 타겟으로 등록	
ElastiCache	19
관련 리소스	20
Amazon Keyspaces(Apache Cassandra용)	21
서비스 연결 역할	21
서비스 보안 주체	21
Application Auto Scaling을 통해 Amazon Keyspaces 테이블을 확장 가능 대상으로 등록	21
관련 리소스	23
AWS Lambda	23
서비스 연결 역할	23
서비스 보안 주체	23
Application Auto Scaling을 통해 Lambda 서비스를 확장 가능 대상으로 등록	24
관련 리소스	24
Amazon Managed Streaming for Apache Kafka(MSK)	25
서비스 연결 역할	25
서비스 보안 주체	25
Application Auto Scaling을 통해 Amazon MSK 클러스터 스토리지를 확장 가능 대상으로 등록	25
관련 리소스	27
Amazon Neptune	27
서비스 연결 역할	27
서비스 보안 주체	27
Application Auto Scaling을 통해 Neptune 클러스터를 확장 가능 대상으로 등록	27
관련 리소스	28
아마존 SageMaker	28
서비스 연결 역할	29
서비스 보안 주체	29
Application Auto Scaling을 사용하여 SageMaker 엔드포인트 변형을 확장 가능한 타겟으로 등록	29
Application Auto Scaling을 통해 서버리스 엔드포인트의 동시성을 확장 가능 대상으로 등록하기	30
Application Auto Scaling을 통해 추론 구성 요소를 확장 가능 대상으로 등록	31
관련 리소스	32
스팟 플릿(Amazon EC2)	32
서비스 연결 역할	32
서비스 보안 주체	33

Application Auto Scaling을 통해 스팟 플릿을 확장 가능 대상으로 등록	33
관련 리소스	34
사용자 정의 리소스	34
서비스 연결 역할	34
서비스 보안 주체	34
Application Auto Scaling을 통해 사용자 지정 리소스를 확장 가능 대상으로 등록	35
관련 리소스	36
를 사용하여 스케일링을 구성합니다. AWS CloudFormation	37
Application Auto Scaling 및 AWS CloudFormation 템플릿	37
예제 템플릿 코드 조각	38
에 대해 자세히 알아보십시오. AWS CloudFormation	38
예약된 조정	39
예약된 조정 작동 방식	39
작동 방식	40
고려 사항	40
자주 사용되는 명령	41
관련 리소스	41
제한 사항	42
cron 표현식 사용	42
예약된 작업의 예	45
한 번만 발생하는 예약된 작업 생성	45
반복되는 간격으로 실행되는 예약된 작업 만들기	47
반복되는 일정으로 실행되는 예약된 작업 생성	47
시간대를 지정하는 일회성 예약된 작업 만들기	48
시간대를 지정하는 반복 예약 작업 생성	49
예약된 크기 조정 관리	50
지정된 서비스에 대한 크기 조정 활동 보기	50
지정된 서비스에 대해 예약된 모든 작업 설명	52
확장 가능한 대상에 대해 하나 이상의 예약된 작업 설명	54
확장 가능한 대상에 대해 예약된 크기 조정 해제	55
예약된 작업 삭제	56
자습서: AWS CLI을(를) 사용하여 예약된 조정 시작하기	56
1단계: 확장 가능 대상 등록	57
2단계: 2개의 예약된 작업 생성	58
3단계: 조정 활동 보기	61
4단계: 다음 단계	65

5단계: 정리	65
대상 추적 조정 정책	67
표적 추적 작동 방식	68
작동 방식	68
지표 선택	69
목표 값 정의	70
휴지 기간 정의	70
고려 사항	72
여러 조정 정책	72
자주 사용되는 명령	73
관련 리소스	73
제한 사항	73
대상 추적 조정 정책 생성	74
확장 가능 대상 등록	75
대상 추적 조정 정책 생성	75
대상 추적 조정 정책 설명	77
대상 추적 조정 정책 삭제	79
지표 수학 사용	79
예: 태스크당 Amazon SQS 대기열 백로그	80
제한 사항	84
단계별 조정 정책	85
스텝 스케일링 작동 방식	86
작동 방식	86
단계 조절	87
조정 조절 타입	89
휴지 기간	90
자주 사용되는 명령	90
고려 사항	91
관련 리소스	41
제한 사항	91
단계 조정 정책 삭제	91
확장 가능 대상 등록	92
단계 조정 정책 삭제	93
조정 정책을 트리거하는 경보 생성	96
단계 조정 정책 설명	97
단계 조정 정책 삭제	98

자습서: 과중한 워크로드를 처리하도록 Auto Scaling 구성	100
필수 조건	100
1단계: 확장 가능 대상 등록	101
2단계: 요구 사항에 따라 예약된 작업 설정	102
2단계: 대상 추적 조정 정책 추가	105
4단계: 다음 단계	107
5단계: 정리	108
조정 일시 중지	110
조정 활동	110
조정 활동을 일시 중지하고 재개합니다.	111
일시 중지된 조정 활동 보기	113
조정 활동 재개	114
조정 활동	116
확장 가능한 대상을 기준으로 조정 활동을 찾아보세요.	116
규모가 조정되지 않은 활동 포함	117
사유 코드	119
모니터링	121
모니터링: 사용 CloudWatch	122
CloudWatch 리소스 사용량 모니터링을 위한 지표	123
대상 추적 조정 정책을 위해 사전 정의된 지표	134
AWS CloudTrail	137
의 Application Auto Scaling 정보 CloudTrail	137
Application Auto Scaling 로그 파일 항목 이해	138
.....	138
관련 리소스	139
아마존 EventBridge	139
Application Auto Scaling 이벤트	140
태그 지정 지원	144
태그 예제	144
보안을 위한 태그	145
태그에 대한 액세스 통제	146
보안	147
데이터 보호	147
ID 및 액세스 관리	148
액세스 제어	149
Application Auto Scaling에서 IAM을 사용하는 방식	149

AWS 관리형 정책	155
서비스 연결 역할	164
자격 증명 기반 정책 예시	169
문제 해결	181
권한 검증	182
AWS PrivateLink	184
인터페이스 VPC 엔드포인트 생성	184
VPC 엔드포인트 정책 생성	185
복원력	185
인프라 보안	186
규정 준수 확인	186
할당량	188
사용 설명서 기록	189
.....	CXCviii

Application Auto Scaling이란?

Application Auto Scaling은 Amazon EC2 이외의 개별 AWS 서비스에 맞게 확장 가능한 리소스를 자동으로 확장하기 위한 솔루션이 필요한 개발자 및 시스템 관리자를 위한 웹 서비스입니다. Application Auto Scaling을 사용하면 다음 리소스에 대한 자동 크기 조정을 구성할 수 .

- AppStream 2.0 플릿
- Aurora 복제본
- Amazon Comprehend 문서 분류 및 엔터티 인식기 엔드포인트
- DynamoDB 테이블 및 글로벌 보조 인덱스
- Amazon ECS 서비스
- ElastiCache Redis 클러스터용 (복제 그룹)
- Amazon EMR 클러스터
- Amazon Keyspaces(Apache Cassandra용) 표
- Lambda 함수의 프로비저닝된 동시성
- Amazon Managed Streaming for Apache Kafka(MSK) 브로커 스토리지
- Amazon Neptune 클러스터
- SageMaker 엔드포인트 변형
- SageMaker 추론 구성 요소
- SageMaker 서버리스 프로비저닝 동시성
- 스팟 플릿 요청
- 자체 애플리케이션 또는 서비스에서 제공하는 사용자 지정 리소스. [자세한 내용은 리포지토리를 참조하십시오. GitHub](#)

위에 나열된 모든 AWS 서비스의 지역별 가용성을 보려면 [지역 표](#)를 참조하십시오.

Auto Scaling 그룹을 사용하여 Amazon EC2 인스턴스 플릿을 조정하는 방법에 대한 자세한 내용은 [Amazon EC2 Auto Scaling 사용 설명서](#)를 참조하세요.

Application Auto Scaling의 기능

Application Auto Scaling을 사용하면 사용자가 정의하는 조건에 맞게 확장 가능한 리소스를 자동으로 조정할 수 있습니다.

- 대상 추적 조정 — 특정 CloudWatch 지표의 목표 값을 기반으로 리소스를 조정합니다.
- 단계 크기 조정 - 경보 위반의 크기에 따라 다른 일련의 크기 조정 조절을 기반으로 리소스 크기를 조정합니다.
- 예약된 크기 조정 - 한 번만 또는 반복되는 일정에 따라 리소스 크기를 조정합니다.

Application Auto Scaling 작업

크기 조정 중인 리소스에 따라 다음 인터페이스를 사용하여 크기 조정을 구성할 수 있습니다.

- AWS Management Console - 크기 조정을 구성하는 데 사용할 수 있는 웹 인터페이스를 제공합니다. AWS 계정을 등록한 경우 에 로그인하여 Application Auto Scaling에 AWS Management Console 액세스하십시오. 그런 다음 소개에 나열된 리소스 중 하나에 대해 서비스 콘솔을 엽니다. 작업을 하려는 AWS 리전 리소스와 동일한 콘솔을 열어야 합니다.

Note

콘솔 액세스는 일부 리소스에서 사용할 수 없습니다. 자세한 정보는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)을 참조하세요.

- AWS Command Line Interface (AWS CLI) — 다양한 명령을 제공하며 Windows AWS 서비스, macOS 및 Linux에서 지원됩니다. 시작하려면 [AWS Command Line Interface](#) 섹션을 참조하세요. 명령 목록은 명령 참조의 [애플리케이션 자동 크기 조정을](#) 참조하십시오. AWS CLI
- AWS Tools for Windows PowerShell— 환경에서 스크립트를 작성하는 사용자를 위해 다양한 AWS 제품 세트에 대한 명령을 제공합니다. PowerShell 시작하려면 [AWS Tools for Windows PowerShell 사용자 가이드](#)를 참조하세요. 자세한 설명은 [AWS Tools for PowerShell Cmdlet 참조](#)를 참조하세요.
- AWS SDK — 언어별 API 작업을 제공하고 서명 계산, 요청 재시도 처리, 오류 처리와 같은 많은 연결 세부 정보를 처리합니다. [자세한 내용은 빌드 기반 도구를 참조하십시오. AWS](#)
- HTTPS API - HTTPS 요청을 사용하여 호출하는 하위 수준의 API 작업을 제공합니다. 자세한 내용은 [Application Auto Scaling API Reference](#)(Application Auto Scaling API 레퍼런스)를 참조하세요.
- AWS CloudFormation— CloudFormation 템플릿을 사용한 스케일링 구성을 지원합니다. 자세한 정보는 [를 사용하여 Application Auto Scaling 리소스를 구성합니다. AWS CloudFormation](#)을 참조하세요.

프로그래밍 방식으로 AWS 서비스연결하려면 엔드포인트를 사용합니다. Application Auto Scaling을 호출하기 위한 엔드포인트에 대한 자세한 내용은 [Application Auto Scaling AWS 일반 참조엔드포인트와 중국 내 Amazon Web Services ARN의 할당량 시작하기 > 일급 비밀 지역의 리전 사용 설명서](#).

Application Auto Scaling의 개념

이 주제에서는 Application Auto Scaling에 대해 알아보고 이를 사용해 시작하는 데 도움이 되는 주요 개념에 대해 설명합니다.

확장 가능 대상

확장하려는 리소스를 지정하기 위해 생성하는 엔터티입니다. 각 확장 가능 대상은 기본 서비스의 일부 용량 차원을 나타내는 서비스 네임스페이스, 리소스 ID 및 확장 가능한 차원으로 고유하게 식별됩니다. 예를 들어, Amazon ECS 서비스는 태스크 수의 Auto scaling을 지원하고, DynamoDB 테이블은 테이블 및 그 글로벌 보조 인덱스의 읽기 및 쓰기 용량 Auto scaling을 지원하며, Aurora 클러스터는 복제본 수의 조정을 지원합니다.

Tip

각 확장 가능 대상에는 최대 및 최소 용량도 있습니다. 조정 정책은 최소-최대 범위보다 높거나 낮지 않습니다. Application Auto Scaling에서는 인식하지 못하는 이 범위를 벗어나는 기본 리소스를 직접 out-of-band 변경할 수 있습니다. 그러나 조정 정책을 호출하거나 RegisterScalableTarget API를 호출되면 Application Auto Scaling이 현재 용량을 검색하여 최소 및 최대 용량과 비교합니다. 최소-최대 범위를 벗어나면 설정된 최솟값과 최댓값을 준수하도록 용량이 업데이트됩니다.

축소

Application Auto Scaling이 확장 가능 대상에 대한 용량을 자동으로 줄이면 확장 가능 대상이 축소합니다. 조정 정책이 설정되면 확장 가능 대상을 최소 용량보다 작게 스케일 인할 수 없습니다.

확장

Application Auto Scaling이 확장 가능 대상에 대한 용량을 자동으로 늘이면 확장 가능 대상이 확장합니다. 조정 정책이 설정되면 확장 가능 대상을 최대 용량보다 크게 스케일 아웃할 수 없습니다.

조정 정책

조정 정책은 Application Auto Scaling에서 특정 CloudWatch 지표를 추적하도록 지시합니다. 그런 다음 지표가 특정 임계값보다 높거나 낮을 때 수행할 조정 작업을 결정합니다. 예를 들어 클러스터 전체의 CPU 사용량이 증가하기 시작하면 확장하고 다시 떨어지면 축소할 수 있습니다.

Auto Scaling에 사용되는 지표는 대상 서비스에 게시되지만, 자체 지표를 대상 서비스에 게시한 다음 조정 CloudWatch 정책과 함께 사용할 수도 있습니다.

크기 조정 활동 간의 휴지 기간을 사용하면 다른 크기 조정 활동이 시작되기 전에 리소스가 안정화됩니다. Application Auto Scaling은 휴지 기간에 지표를 계속 평가합니다. 휴지 기간이 끝나면 필요한 경우 조정 정책이 다른 크기 조정 활동을 시작합니다. 휴지 기간이 적용되는 동안 현재 지표 값에 따라 더 큰 확장이 필요한 경우 조정 정책이 즉시 확장됩니다.

예약된 작업

예약된 작업은 특정 날짜 및 시간에 자동으로 리소스의 크기를 조정합니다. 확장 가능 대상에 대한 최소 및 최대 용량을 수정하여 작동하므로 최소 용량을 높게 또는 최대 용량을 낮게 설정하여 일정 에 따라 축소 및 확장하는 데 사용할 수 있습니다. 예를 들어 예약된 작업을 사용하여 금요일에 용량을 줄이고 다음 월요일에 용량을 늘려 주말에 리소스를 사용하지 않는 애플리케이션을 조정할 수 있습니다.

또한 예약된 작업을 사용하여 시간에 따른 최솟값과 최댓값을 최적화하여 일반적인 트래픽보다 높은 상황(예: 마케팅 캠페인 또는 계절적 변동)에 적응할 수 있습니다. 이렇게 하면 용량을 늘리기 위해 더 많이 확장해야 하는 경우 성능을 개선하고 리소스를 적게 사용하는 경우 비용을 절감할 수 있습니다.

자세히 알아보기

[AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#) — 이 섹션에서는 확장할 수 있는 서비스를 소개하고 확장 가능 대상을 등록하여 Auto scaling을 설정하도록 돕습니다. 또한 Application Auto Scaling이 대상 서비스의 리소스에 액세스하기 위해 생성하는 각 IAM 서비스 연결 역할에 대해서도 설명합니다.












[Application Auto Scaling의 대상 추적 조정 정책](#) — Application Auto Scaling의 주요 기능 중 하나는 대상 추적 조정 정책입니다. 대상 추적 정책이 원하는 용량을 자동으로 조정하여 구성된 지표 및 대상 값에 따라 일정 수준으로 활용도를 유지하는 방법에 대해 알아보십시오. 예를 들어 스팟 플릿의 평균 CPU 사용률을 50%로 유지하도록 대상 추적을 구성할 수 있습니다. 그러면 Application Auto Scaling이 모든 서버에서 집계된 CPU 사용률을 50%로 유지하는 데 필요한 대로 EC2 인스턴스를 시작하거나 종료합니다.

AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는






Application Auto Scaling은 다른 AWS 서비스와 통합되므로 애플리케이션의 요구에 맞게 확장 기능을 추가할 수 있습니다. Auto Scaling은 서비스의 선택적 기능으로서 거의 모든 경우에 기본적으로 비활성화됩니다.

다음 표에는 지원되는 Auto Scaling 구성 방법에 대한 정보를 포함하여 Application Auto Scaling과 함께 사용할 수 있는 AWS 서비스가 나열되어 있습니다. 사용자 지정 리소스와 함께 Application Auto Scaling을 사용할 수도 있습니다.

- 콘솔 액세스 - 대상 서비스의 콘솔에서 조정 정책을 구성하여 자동 크기 조정을 시작하기 위한 호환 AWS 서비스를 구성할 수 있습니다.
- CLI 액세스 - AWS CLI를 사용하여 자동 크기 조정을 시작하기 위한 호환 AWS 서비스를 구성할 수 있습니다.
- SDK 액세스 — AWS SDK를 사용하여 Auto Scaling을 시작하도록 호환 가능한 AWS 서비스를 구성할 수 있습니다.
- CloudFormation 액세스 - AWS CloudFormation 스택 템플릿을 사용하여 Auto Scaling을 시작하도록 호환 AWS 서비스를 구성할 수 있습니다. 자세한 정보는 [클라우드Formation을 사용하여 Application Auto Scaling 리소스를 구성합니다. AWS CloudFormation](#)을 참조하세요.

AWS 서비스	콘솔 액세스 ¹	CLI 액세스	SDK 액세스	CloudFormation 액세스
AppStream 2.0	 예	 예	 예	 예
Aurora	 예	 예	 예	 예
Amazon Comprehend	 아 니요	 예	 예	 예

AWS 서비스	콘솔 액세스 ¹	CLI 액세스	SDK 액세스	CloudFormation 액세스
Amazon DynamoDB	 예	 예	 예	 예
Amazon ECS	 예	 예	 예	 예
아마존 ElastiCache	 예	 예	 예	 예
Amazon EMR	 예	 예	 예	 예
Amazon Keyspaces	 예	 예	 예	 예
Lambda	 아 니요	 예	 예	 예
Amazon MSK	 예	 예	 예	 예
Amazon Neptune	 아 니요	 예	 예	 예

AWS 서비스	콘솔 액세스 ¹	CLI 액세스	SDK 액세스	CloudFormation 액세스
SageMaker	 예	 예	 예	 예
스팟 플릿	 예	 예	 예	 예
사용자 정의 리소스	 아 니요	 예	 예	 예

¹ 조정 정책 구성을 위한 콘솔 액세스. 대부분의 서비스는 콘솔에서 스케줄링된 스케일링을 구성하는 것을 지원하지 않습니다. 현재는 Amazon AppStream 2.0과 스팟 플릿만이 예정된 조정을 위한 콘솔 액세스를 제공합니다. ElastiCache

Amazon AppStream 2.0 및 Application Auto Scaling

타겟 트래킹 스케일링 정책, 단계별 스케일링 정책, 스케줄링된 스케일링을 사용하여 AppStream 2.0 플릿을 확장할 수 있습니다.

다음 정보를 사용하면 AppStream 2.0을 Application Auto Scaling과 통합하는 데 도움이 됩니다.

2.0용으로 생성된 서비스 연결 역할 AppStream

Application Auto [Scaling에서 AppStream 2.0 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_AppStreamFleet`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `appstream.application-autoscaling.amazonaws.com`

Application Auto Scaling을 사용하여 AppStream 2.0 플릿을 확장 가능한 타겟으로 등록

Application Auto Scaling을 사용하려면 AppStream 2.0 플릿에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 확장 가능한 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

AppStream 2.0 콘솔을 사용하여 Auto Scaling을 구성하면 AppStream 2.0에서 확장 가능한 타겟을 자동으로 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

AppStream 2.0 플릿의 경우 [register-scalable-target](#) 명령을 호출하십시오. 다음 예제에서는 `sample-fleet`이라는 플릿의 원하는 용량을 등록합니다. 최소 용량은 플릿 인스턴스 한 개, 최대 용량은 플릿 인스턴스 5개입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace appstream \
  --scalable-dimension appstream:fleet:DesiredCapacity \
  --resource-id fleet/sample-fleet \
  --min-capacity 1 \
  --max-capacity 5
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
```



```
"ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 AppStream 2.0 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Amazon AppStream AppStream 2.0 관리 가이드의 [2.0용 플릿 오토 스케일링](#)

Amazon Aurora 및 Application Auto Scaling

대상 추적 조정 정책, 단계 조정 정책 및 예약된 조정을 사용하여 Aurora DB 클러스터를 조정할 수 있습니다.

Aurora를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Aurora에 대해 생성된 서비스 연결 역할

Application Auto [Scaling](#)을 통해 [Aurora 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- AWSServiceRoleForApplicationAutoScaling_RDSCluster

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- rds.application-autoscaling.amazonaws.com

Application Auto Scaling을 통해 Aurora DB 클러스터를 확장 가능 대상으로 등록

Aurora 클러스터에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

Aurora 콘솔을 사용하여 자동 크기 조정을 구성하면 Aurora가 자동으로 확장 가능 대상을 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

Aurora 클러스터에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 `my-db-cluster`라는 클러스터에서 Aurora 복제본의 수를 등록합니다. 최소 용량은 Aurora 복제본 한 개, 최대 용량은 Aurora 복제본 8개입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace rds \
  --scalable-dimension rds:cluster:ReadReplicaCount \
  --resource-id cluster:my-db-cluster \
  --min-capacity 1 \
  --max-capacity 8
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity`, `MaxCapacity`를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 Aurora 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Amazon RDS 사용 설명서의 [Aurora 복제본에 Amazon Aurora Auto Scaling 사용](#)

Amazon Comprehend 및 Application Auto Scaling

대상 추적 조정 정책 및 예약된 조정을 사용하여 Amazon Comprehend 문서 분류 및 엔터티 인식기 엔드포인트를 확장할 수 있습니다.

Amazon Comprehend를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Amazon Comprehend에 대해 생성된 서비스 연결 역할

Amazon Comprehend 리소스를 Application Auto Scaling을 통해 [확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `comprehend.application-autoscaling.amazonaws.com`

Application Auto Scaling을 통해 Amazon Comprehend 리소스를 확장 가능한 대상으로 등록

Amazon Comprehend 문서 분류 또는 엔터티 인식기 엔드포인트에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려면 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

문서 분류 엔드포인트에 대해 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 엔드포인트의 ARN을 사용하여 문서 분류자 엔드포인트에 대해 모델에서 사용할 추론 단위 수를 등록합니다. 최소 추론 용량은 1, 최대 추론 용량은 3입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace comprehend \
  --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-
endpoint/EXAMPLE \
  --min-capacity 1 \
  --max-capacity 3
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

엔터티 인식기 엔드포인트에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 엔드포인트의 ARN을 사용하여 엔터티 인식기에 대해 모델에서 사용할 추론 단위 수를 등록합니다. 최소 추론 용량은 1, 최대 추론 용량은 3입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace comprehend \
  --scalable-dimension comprehend:entity-recognizer-endpoint:DesiredInferenceUnits \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-
endpoint/EXAMPLE \
  --min-capacity 1 \
  --max-capacity 3
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
```

```
"ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 Amazon Comprehend 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Amazon Comprehend 개발자 안내서의 [엔드포인트를 통한 자동 조정](#)

Amazon DynamoDB 및 Application Auto Scaling

대상 추적 조정 정책 및 예약된 조정을 사용하여 DynamoDB 테이블 및 글로벌 보조 인덱스를 조정할 수 있습니다.

DynamoDB를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

DynamoDB에 대해 생성된 서비스 연결 역할

Application [Auto Scaling](#)을 통해 [DynamoDB 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- AWSServiceRoleForApplicationAutoScaling_DynamoDBTable

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- dynamodb.application-autoscaling.amazonaws.com

Application Auto Scaling을 통해 DynamoDB 리소스를 확장 가능 대상으로 등록

DynamoDB 테이블 또는 글로벌 보조 인덱스에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

DynamoDB 콘솔을 사용하여 자동 크기 조정을 구성하면 DynamoDB가 자동으로 확장 가능 대상을 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

[register-scalable-target](#) 명령을 호출하여 테이블의 쓰기 용량을 확인합니다. 다음 예제에서는 `my-table`이라는 테이블의 프로비저닝된 쓰기 용량을 등록합니다. 최소 쓰기 용량은 5, 최대 쓰기 용량은 10입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/my-table \
  --min-capacity 5 \
  --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

[register-scalable-target](#) 명령을 호출하여 테이블의 읽기 용량을 확인합니다. 다음 예제에서는 `my-table`이라는 테이블의 프로비저닝된 읽기 용량을 등록합니다. 최소 읽기 용량은 5, 최대 읽기 용량은 10입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
```

```
--scalable-dimension dynamodb:table:ReadCapacityUnits \
--resource-id table/my-table \
--min-capacity 5 \
--max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

글로벌 보조 인덱스의 쓰기 용량에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 my-table-index이라는 글로벌 보조 인덱스의 프로비저닝된 쓰기 용량을 등록합니다. 최소 쓰기 용량은 5, 최대 쓰기 용량은 10입니다.

```
aws application-autoscaling register-scalable-target \
--service-namespace dynamodb \
--scalable-dimension dynamodb:index:WriteCapacityUnits \
--resource-id table/my-table/index/my-table-index \
--min-capacity 5 \
--max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

글로벌 보조 인덱스의 읽기 용량에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 my-table-index이라는 글로벌 보조 인덱스의 프로비저닝된 읽기 용량을 등록합니다. 최소 읽기 용량은 5, 최대 읽기 용량은 10입니다.

```
aws application-autoscaling register-scalable-target \
--service-namespace dynamodb \
--scalable-dimension dynamodb:index:ReadCapacityUnits \
--resource-id table/my-table/index/my-table-index \
--min-capacity 5 \
```

```
--max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 DynamoDB 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

- Amazon DynamoDB 개발자 안내서의 [DynamoDB 자동 크기 조정으로 처리 용량 관리](#)
- Amazon DynamoDB 개발자 안내서에서 [테이블의 자동 스케일링 설정을 평가하십시오.](#)
- [블로그에서 DynamoDB 테이블 및 인덱스에 대한 Auto Scaling을 구성하는 데 사용하는 AWS CloudFormation 방법](#) AWS

예약 규모 조정에 대한 자습서도 에서 찾을 수 있습니다. [자습서: AWS CLI을\(를\) 사용하여 예약된 조정 시작하기](#) 이 자습서에서는 예약된 시간에 DynamoDB 테이블이 조정되도록 조정을 구성하는 기본 단계에 대해 알아봅니다.

Amazon ECS 및 Application Auto Scaling

대상 추적 조정 정책, 단계 조정 정책 및 예약된 조정을 사용하여 ECS 서비스를 조정할 수 있습니다.

Amazon ECS를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Amazon ECS에 대해 생성된 서비스 연결 역할

Application [Auto Scaling](#)을 통해 [Amazon ECS 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용

자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_ECSService`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `ecs.application-autoscaling.amazonaws.com`

Application Auto Scaling을 통해 ECS 서비스를 확장 가능 대상으로 등록

Amazon ECS 서비스에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

Amazon ECS 콘솔을 사용하여 자동 크기 조정을 구성하면 Amazon ECS가 자동으로 확장 가능 대상을 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

Amazon ECS 서비스에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 default 클러스터에서 실행되는 `sample-app-service`라는 서비스에 대한 확장 가능 대상을 등록합니다. 최소 태스크 수는 1, 최대 태스크 수는 10입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/default/sample-app-service \
  --min-capacity 1 \
  --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 Amazon ECS 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

- Amazon Elastic 컨테이너 서비스 개발자 안내서의 서비스 [자동 조정](#)
- Amazon Elastic 컨테이너 서비스 모범 사례 가이드에서 서비스 [자동 조정 구성](#)

Note

Amazon ECS 배포가 진행 중인 동안 스케일 아웃 프로세스를 일시 중단하기 위한 지침은 다음 설명서를 참조하십시오.

Amazon Elastic 컨테이너 서비스 개발자 안내서의 서비스 [자동 조정 및 배포](#)

ElastiCache Redis용 및 Application Auto Scaling용

대상 추적 조정 정책 및 예약된 조정을 사용하여 Redis 복제 그룹에 맞게 ElastiCache 규모를 조정할 수 있습니다.

다음 정보를 사용하면 Application Auto ElastiCache Scaling과 통합하는 데 도움이 됩니다.

ElastiCache에 대한 서비스 연결 역할이 생성됨

Application Auto [Scaling](#)을 통해 [ElastiCache 리소스를 확장 가능한 대상으로 AWS 계정 등록할 때 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용

자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `elasticache.application-autoscaling.amazonaws.com`

Application Auto Scaling을 사용하여 Redis 복제 그룹을 확장 가능한 타겟으로 등록 ElastiCache

Application Auto Scaling에서는 ElastiCache 복제 그룹에 대한 조정 정책 또는 스케줄링된 작업을 생성하기 전에 확장 가능한 타겟이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

ElastiCache 콘솔을 사용하여 Auto Scaling을 구성하면 확장 가능한 타겟이 ElastiCache 자동으로 등록됩니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

ElastiCache 복제 그룹에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 `mycluster`라는 복제 그룹에 대해 원하는 수의 노드 그룹을 등록합니다. 최소 용량은 1이고 최대 용량은 5입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace elasticache \
  --scalable-dimension elasticache:replication-group:NodeGroups \
  --resource-id replication-group/mycluster \
  --min-capacity 1 \
  --max-capacity 5
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

다음 예제에서는 mycluster라는 복제 그룹에 대해 원하는 수의 노드 그룹당 복제본을 등록합니다. 최소 용량은 1이고 최대 용량은 5입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace elasticache \
  --scalable-dimension elasticache:replication-group:Replicas \
  --resource-id replication-group/mycluster \
  --min-capacity 1 \
  --max-capacity 5
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 ElastiCache 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Amazon 포 ElastiCache 레디스용 사용 설명서의 Redis ElastiCache 클러스터용 [Auto Scaling](#)

Amazon Keyspaces(Apache Cassandra용) 및 Application Auto Scaling

대상 추적 조정 정책 및 예약된 조정을 사용하여 Amazon Keyspaces 테이블을 조정할 수 있습니다.

Amazon Keyspaces를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Amazon Keyspaces에 대해 생성된 서비스 연결 역할

Amazon Keyspaces 리소스를 Application Auto Scaling을 통해 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 [서비스 연결 역할](#)이 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_CassandraTable`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `cassandra.application-autoscaling.amazonaws.com`

Application Auto Scaling을 통해 Amazon Keyspaces 테이블을 확장 가능 대상으로 등록

Amazon Keyspaces 테이블에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

Amazon Keyspaces 콘솔을 사용하여 자동 크기 조정을 구성하면 Amazon Keyspaces가 자동으로 확장 가능 대상을 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

Amazon Keyspaces 테이블에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 mytable이라는 테이블의 프로비저닝된 쓰기 용량을 등록합니다. 최소 쓰기 용량은 5, 최대 쓰기 용량은 10입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace cassandra \
  --scalable-dimension cassandra:table:WriteCapacityUnits \
  --resource-id keyspace/mykeyspace/table/mytable \
  --min-capacity 5 \
  --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

다음 예제에서는 mytable이라는 테이블의 프로비저닝된 읽기 용량을 등록합니다. 최소 읽기 용량은 5, 최대 읽기 용량은 10입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace cassandra \
  --scalable-dimension cassandra:table:ReadCapacityUnits \
  --resource-id keyspace/mykeyspace/table/mytable \
  --min-capacity 5 \
  --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 Amazon Keyspaces 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

[Amazon Keyspaces \(아파치 카산드라용\) 개발자 안내서에서 Amazon Keyspaces 자동 크기 조정을 통한 처리 용량 관리](#)

AWS Lambda 및 Application Auto Scaling

대상 추적 조정 정책 및 예약된 조정을 사용하여 AWS Lambda 프로비저닝된 동시성을 확장할 수 있습니다.

Lambda를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Lambda에 대해 생성된 서비스 연결 역할

Application [Auto Scaling](#)을 통해 [Lambda 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- lambda.application-autoscaling.amazonaws.com

Application Auto Scaling을 통해 Lambda 서비스를 확장 가능 대상으로 등록

Lambda 함수에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려면 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

Lambda 함수에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 my-function이라는 함수, BLUE라는 함수 별칭에 대해 프로비저닝된 동시성을 등록합니다. 최소 용량은 0, 최대 용량은 100입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace lambda \
  --scalable-dimension lambda:function:ProvisionedConcurrency \
  --resource-id function:my-function:BLUE \
  --min-capacity 0 \
  --max-capacity 100
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우, 다음 설명서에서 Lambda 함수 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

- [개발자 안내서에서 프로비저닝된 동시성 구성 AWS Lambda](#)
- 블로그에서 반복되는 [피크 사용량에 맞춰 Lambda 프로비저닝된 동시성 예약 AWS](#)

Amazon Managed Streaming for Apache Kafka(MSK) 및 Application Auto Scaling

대상 추적 조정 정책을 사용하여 Amazon MSK 클러스터 스토리지를 확장할 수 있습니다. 대상 추적 정책에 의한 축소가 비활성화되어 있습니다.

Amazon MSK를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Amazon MSK에 대해 생성된 서비스 연결 역할

Amazon MSK 리소스를 Application Auto Scaling을 통해 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 [서비스 연결 역할](#)이 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_KafkaCluster`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 말할 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `kafka.application-autoscaling.amazonaws.com`

Application Auto Scaling을 통해 Amazon MSK 클러스터 스토리지를 확장 가능한 대상으로 등록

Application Auto Scaling에서는 Amazon MSK 클러스터의 브로커당 스토리지 볼륨 크기에 대한 조정 정책을 생성하기 전에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 크기를 조정할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

Amazon MSK 콘솔을 사용하여 자동 크기 조정을 구성하면 Amazon MSK가 자동으로 확장 가능 대상을 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

Amazon MSK 클러스터에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 Amazon MSK 클러스터의 브로커당 스토리지 볼륨 크기를 등록합니다. 최소 용량은 100GiB, 최대 용량은 800GiB입니다.


```
aws application-autoscaling register-scalable-target \
  --service-namespace kafka \
  --scalable-dimension kafka:broker-storage:VolumeSize \
  --resource-id arn:aws:kafka:us-east-1:123456789012:cluster/demo-
cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5 \
  --min-capacity 100 \
  --max-capacity 800
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity`, `MaxCapacity`를 파라미터로 제공합니다.

 Note

Amazon MSK 클러스터가 확장 가능한 대상인 경우 축소가 비활성화되어 활성화할 수 없습니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우, 다음 설명서에서 Amazon MSK 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Apache Kafka용 Amazon 매니지드 스트리밍 개발자 가이드의 [자동 크기 조정](#)

Amazon Neptune 및 Application Auto Scaling

대상 추적 조정 정책 및 예약된 크기 조정을 사용하여 Neptune 클러스터의 크기를 조정할 수 있습니다.

Neptune을 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

Neptune에 대해 생성된 서비스 연결 역할

Application [Auto Scaling](#)을 통해 Neptune 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 [서비스 연결 역할](#)이 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_NeptuneCluster`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `neptune.application-autoscaling.amazonaws.com`

Application Auto Scaling을 통해 Neptune 클러스터를 확장 가능 대상으로 등록

Neptune 클러스터에 대한 크기 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려면 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

Neptune 클러스터의 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 mycluster라는 플릿의 원하는 용량을 등록합니다. 최소 용량은 한 개, 최대 용량은 여덟 개입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace neptune \
  --scalable-dimension neptune:cluster:ReadReplicaCount \
  --resource-id cluster:mycluster \
  --min-capacity 1 \
  --max-capacity 8
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 Neptune 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Neptune 사용 설명서의 [Amazon Neptune DB 클러스터의 복제본 수 자동 조정](#)

SageMaker 아마존과 애플리케이션 오토 스케일링

대상 추적 조정 정책, 단계 조정 정책 및 예약된 조정을 사용하여 SageMaker 엔드포인트 변형, 서버리스 엔드포인트의 프로비저닝된 동시성 및 추론 구성 요소를 확장할 수 있습니다.

다음 정보를 사용하면 Application Auto SageMaker Scaling과 통합하는 데 도움이 됩니다.

SageMaker에 대한 서비스 연결 역할이 생성됨

Application Auto [Scaling](#)을 통해 SageMaker 리소스를 확장 가능한 대상으로 AWS 계정 등록할 때 다음과 같은 [서비스 연결 역할](#)이 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `sagemaker.application-autoscaling.amazonaws.com`

Application Auto Scaling을 사용하여 SageMaker 엔드포인트 변형을 확장 가능한 타겟으로 등록

Application Auto Scaling을 사용하려면 먼저 확장 가능한 대상이 있어야 SageMaker 모델 (변형)에 대한 조정 정책 또는 스케줄링된 작업을 생성할 수 있습니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

SageMaker 콘솔을 사용하여 Auto Scaling을 구성하면 확장 가능한 타겟이 SageMaker 자동으로 등록됩니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

제품 변형에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 my-endpoint 엔드포인트에서 실행되는 my-variant라는 제품 변형에 대해 원하는 인스턴스 수를 등록합니다. 최소 용량은 1개, 최대 용량은 8개입니다.

```
aws application-autoscaling register-scalable-target \
```

```
--service-namespace sagemaker \
--scalable-dimension sagemaker:variant:DesiredInstanceCount \
--resource-id endpoint/my-endpoint/variant/my-variant \
--min-capacity 1 \
--max-capacity 8
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

Application Auto Scaling을 통해 서버리스 엔드포인트의 동시성을 확장 가능 대상으로 등록하기

서버리스 엔드포인트의 프로비저닝된 동시성에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상도 필요합니다.

SageMaker 콘솔을 사용하여 Auto Scaling을 구성하면 확장 가능한 타겟이 SageMaker 자동으로 등록됩니다.

아니면 다음 방법 중 하나를 사용하여 확장 가능 대상을 등록하세요.

- AWS CLI:

제품 변형에 대한 [register-scalable-target](#) 명령을 호출하십시오. 다음 예제에서는 my-endpoint 엔드포인트에서 실행되는 my-variant(이)라는 제품 변형에 대해 프로비저닝된 동시성을 등록합니다. 최소 용량은 1개이고 최대 용량은 10개입니다.

```
aws application-autoscaling register-scalable-target \
--service-namespace sagemaker \
--scalable-dimension sagemaker:variant:DesiredProvisionedConcurrency \
--resource-id endpoint/my-endpoint/variant/my-variant \
--min-capacity 1 \
```

```
--max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

Application Auto Scaling을 통해 추론 구성 요소를 확장 가능 대상으로 등록

추론 구성 요소에 대한 크기 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상도 필요합니다.

- AWS CLI:

추론 [register-scalable-target](#) 컴포넌트의 명령을 호출합니다. 다음 예제에서는 my-inference-component라는 추론 구성 요소에 대해 원하는 복제본 수를 등록합니다. 최소 용량은 0개이고 최대 용량은 3개입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace sagemaker \
  --scalable-dimension sagemaker:inference-component:DesiredCopyCount \
  --resource-id inference-component/my-inference-component \
  --min-capacity 0 \
  --max-capacity 3
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 Amazon SageMaker 개발자 안내서에서 SageMaker 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

- [Amazon SageMaker 모델을 자동으로 크기 조정](#)
- [서버리스 엔드포인트의 프로비저닝된 동시성을 자동으로 확장합니다.](#)
- [다중 모델 엔드포인트 배포를 위한 Auto Scaling 정책 설정](#)
- [비동기 엔드포인트 자동 크기 조정](#)

Note

2023년에는 실시간 추론 엔드포인트에 SageMaker 구축된 새로운 추론 기능을 도입했습니다. SageMaker 엔드포인트의 인스턴스 유형과 초기 인스턴스 수를 정의하는 엔드포인트 구성을 사용하여 엔드포인트를 생성합니다. 그런 다음, 엔드포인트에 모델을 배포하는 데 사용할 수 있는 SageMaker 호스팅 객체인 추론 구성 요소를 생성합니다. 추론 구성 요소 확장에 대한 자세한 내용은 [Amazon에서 기본 모델 배포 비용 및 지연 시간을 줄이는 데 도움이 되는 새로운 추론 기능 SageMaker 추가 및 SageMaker Amazon의 최신 기능을 사용하여 모델 배포 비용을 평균 50% 절감하는](#) 블로그를 참조하십시오. AWS

Amazon EC2 스팟 플릿 및 Application Auto Scaling

대상 추적 조정 정책, 단계 조정 정책 및 예약된 조정을 사용하여 스팟 플릿을 조정할 수 있습니다.

스팟 플릿을 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

스팟 플릿에 대해 생성된 서비스 연결 역할

Application Auto Scaling을 통해 스팟 플릿 리소스를 확장 가능한 대상으로 AWS 계정 등록할 때 다음과 같은 [서비스 연결 역할](#)이 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- `AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest`

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 말할 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- `ec2.application-autoscaling.amazonaws.com`

Application Auto Scaling을 통해 스팟 플릿을 확장 가능 대상으로 등록

스팟 플릿에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

스팟 플릿 콘솔을 사용하여 자동 크기 조정을 구성하면 스팟 플릿이 자동으로 확장 가능 대상을 등록합니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려는 경우 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

스팟 플릿에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 요청 ID를 사용해 스팟 플릿의 대상 용량을 등록합니다. 최소 용량은 인스턴스 2개, 최대 용량은 인스턴스 10개입니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --min-capacity 2 \
  --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
```

```
"ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 스팟 플릿 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

Amazon EC2 사용 설명서의 [스팟 플릿의 자동 크기 조정](#)

사용자 지정 리소스 및 Application Auto Scaling

대상 추적 조정 정책, 단계 조정 정책 및 예약된 조정을 사용하여 사용자 지정 리소스를 조정할 수 있습니다.

사용자 지정 리소스를 Application Auto Scaling과 통합하는 데 도움이 되는 정보는 다음과 같습니다.

사용자 지정 리소스에 대해 생성된 서비스 연결 역할

Application Auto [Scaling](#)에서 [사용자 지정 리소스를 확장 가능한 대상으로 AWS 계정 등록하면 다음과 같은 서비스 연결 역할이](#) 자동으로 생성됩니다. 이 역할을 통해 Application Auto Scaling이 사용자 계정 내에서 지원되는 작업을 수행할 수 있습니다. 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

- AWSServiceRoleForApplicationAutoScaling_CustomResource

서비스 연결 역할이 사용하는 서비스 보안 주체

앞부분에서 다른 서비스 연결 역할은 역할에 대해 정의된 신뢰 관계로 권한이 부여되는 서비스 보안 주체만 맡을 수 있습니다. Application Auto Scaling이 사용하는 서비스 연결 역할은 다음 서비스 보안 주체에 대한 액세스 권한을 부여합니다.

- custom-resource.application-autoscaling.amazonaws.com

Application Auto Scaling을 통해 사용자 지정 리소스를 확장 가능 대상으로 등록

사용자 지정 리소스에 대한 조정 정책 또는 예약된 작업을 생성하려면 먼저 Application Auto Scaling에 확장 가능 대상이 필요합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다. 확장 가능 대상은 리소스 ID, 확장 가능한 차원 및 네임스페이스의 조합으로 고유하게 식별됩니다.

AWS CLI 또는 AWS SDK 중 하나를 사용하여 Auto Scaling을 구성하려면 다음 옵션을 사용할 수 있습니다.

- AWS CLI:

사용자 지정 리소스에 대한 [register-scalable-target](#) 명령을 호출합니다. 다음 예제에서는 사용자 지정 리소스를 확장 가능한 대상으로 등록합니다. 원하는 최소 용량 수는 1, 원하는 최대 용량 수는 10 개입니다. `custom-resource-id.txt` 파일에는 Amazon API Gateway 엔드포인트를 통해 사용자 지정 리소스의 경로를 나타내는 리소스 ID를 식별하는 문자열이 포함되어 있습니다.

```
aws application-autoscaling register-scalable-target \
  --service-namespace custom-resource \
  --scalable-dimension custom-resource:ResourceType:Property \
  --resource-id file://~/custom-resource-id.txt \
  --min-capacity 1 \
  --max-capacity 10
```

`custom-resource-id.txt`의 콘텐츠:

```
https://example.execute-api.us-west-2.amazonaws.com/prod/scalableTargetDimensions/1-23456789
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

[RegisterScalableTarget](#) 작업을 호출하고 ResourceId, ScalableDimension, ServiceNamespace, MinCapacity, MaxCapacity를 파라미터로 제공합니다.

관련 리소스

Application Auto Scaling을 막 시작하는 경우 다음 설명서에서 사용자 지정 리소스 확장에 대한 유용한 추가 정보를 찾을 수 있습니다.

[GitHub리포지토리](#)

를 사용하여 Application Auto Scaling 리소스를 구성합니다.

AWS CloudFormation

Application Auto Scaling은 리소스를 모델링하고 설정하는 데 도움이 되는 서비스인 와 통합되어 있으므로 AWS 리소스와 인프라를 생성하고 관리하는 데 소요되는 시간을 줄일 수 있습니다. AWS CloudFormation원하는 모든 리소스를 설명하는 템플릿을 만들고 해당 AWS 리소스를 자동으로 AWS CloudFormation 프로비저닝 및 구성합니다.

를 사용하면 템플릿을 재사용하여 AWS CloudFormation Application Auto Scaling 리소스를 일관되고 반복적으로 설정할 수 있습니다. 리소스를 한 번 설명한 다음 여러 AWS 계정 지역과 지역에서 동일한 리소스를 반복해서 프로비저닝하십시오.

Application Auto Scaling 및 AWS CloudFormation 템플릿

Application Auto Scaling 및 관련 서비스에 대한 리소스를 프로비저닝하고 구성하려면 [AWS CloudFormation 템플릿](#)을 이해해야 합니다. 템플릿은 JSON 또는 YAML로 서식 지정된 텍스트 파일입니다. 이 템플릿은 AWS CloudFormation 스택에 프로비저닝하려는 리소스를 설명합니다. JSON이 나 YAML에 익숙하지 않은 경우 AWS CloudFormation Designer를 사용하여 템플릿을 시작하는 데 도움을 받을 수 있습니다. AWS CloudFormation 자세한 내용은 AWS CloudFormation 사용 설명서에서 [AWS CloudFormation Designer이란 무엇입니까?](#)를 참조하세요.

Application Auto Scaling 리소스에 대한 스택 템플릿을 생성할 때 다음을 제공해야 합니다.

- 대상 서비스에 대한 네임스페이스(예: **appstream**). 서비스 네임스페이스를 구하려면 [AWS::ApplicationAutoScaling::ScalableTarget](#)참조를 참조하십시오.
- 대상 리소스에 연결된 확장 가능 차원(예: **appstream:fleet:DesiredCapacity**). 확장 가능한 크기를 얻으려면 [AWS::ApplicationAutoScaling::ScalableTarget](#)참조를 참조하십시오.
- 대상 리소스의 리소스 ID(예: **fleet/sample-fleet**). 특정 리소스 ID의 구문 및 예제에 대한 자세한 내용은 [AWS::ApplicationAutoScaling::ScalableTarget](#)참조를 참조하십시오.
- 대상 리소스에 대한 서비스 연결 역할(예: **arn:aws:iam::012345678910:role/aws-service-role/appstream.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_AppStreamFleet**). 역할 ARN을 얻으려면 [서비스 연결 역할 ARN 참조](#) 표를 참조하세요.

Application Auto Scaling 리소스에 대한 자세한 내용은 AWS CloudFormation 사용 설명서의 [Application Auto Scaling](#) 레퍼런스를 참조하세요.

예제 템플릿 코드 조각

AWS CloudFormation 사용자 안내서의 다음 섹션에서 AWS CloudFormation 템플릿에 포함할 예제 스택을 찾을 수 있습니다.

- 조정 정책 및 예약된 작업의 예는 [Application Auto Scaling 리소스 구성](#)을 참조하십시오 AWS CloudFormation.
- 규모 조정 정책의 추가 예는 [AWS::ApplicationAutoScaling::ScalingPolicy](#)을 참조하십시오.

에 대해 자세히 알아보십시오. AWS CloudFormation

자세히 AWS CloudFormation을 알아보려면 다음 리소스를 참조하십시오.

- [AWS CloudFormation](#)
- [AWS CloudFormation 사용 설명서](#)
- [AWS CloudFormation API Reference](#)
- [AWS CloudFormation 명령줄 인터페이스 사용 설명서](#)

Application Auto Scaling의 예약된 조정

예약된 조정을 사용하면 특정 시간에 용량을 늘리거나 줄이는 예약된 작업을 생성하여 예측 가능한 부하 변화에 따라 애플리케이션에 대한 Auto Scaling을 설정할 수 있습니다. 이를 통해 예측 가능한 부하 변화에 맞춰 애플리케이션 규모를 사전에 조정할 수 있습니다.

예를 들어, 주중에는 부하가 증가하고 주말에는 부하가 감소하는 주간 트래픽 패턴이 정기적으로 발생한다고 가정해 보겠습니다. Application Auto Scaling에서 이 패턴에 맞게 규모 조정 일정을 구성할 수 있습니다.

- 수요일 아침에는 예약된 작업이 이전에 설정된 확장 가능한 대상의 최소 용량을 늘려 용량을 늘립니다.
- 금요일 저녁에는 예약된 또 다른 작업이 이전에 설정된 확장 가능한 대상의 최대 용량을 줄여 용량을 줄입니다.

이러한 예약된 규모 조정 작업을 통해 비용과 성능을 최적화할 수 있습니다. 애플리케이션은 주중 트래픽 피크를 처리할 수 있을 만큼 충분한 용량을 갖추게 되지만, 다른 시간에 불필요한 용량을 과도하게 프로비저닝하지는 않습니다.

예약된 조정 및 조정 정책을 함께 사용하면 규모 조정에 대한 예방적 및 대응적 접근 방식의 이점을 모두 얻을 수 있습니다. 예약된 작업이 실행된 후 조정 정책은 계속해서 용량을 추가로 조정할지를 결정할 수 있습니다. 이를 통해 애플리케이션의 로드를 처리할 수 있는 충분한 용량을 보유하도록 보장합니다. 애플리케이션이 수요에 맞게 조정되는 동안 현재 용량은 예약된 작업에서 설정한 최소 및 최대 용량 이내여야 합니다.

내용

- [Application Auto Scaling의 스케줄 조정 작동 방식](#)
- [cron 표현식을 사용하여 반복 크기 조정 작업 예약](#)
- [Application Auto Scaling의 예약된 작업 예제](#)
- [Application Auto Scaling의 예약된 크기 조정 관리](#)
- [자습서: AWS CLI을\(를\) 사용하여 예약된 조정 시작하기](#)

Application Auto Scaling의 스케줄 조정 작동 방식

이 항목에서는 예약 규모 조정 작동 방식을 설명하고 이를 효과적으로 사용하기 위해 이해해야 하는 주요 고려 사항을 소개합니다.

내용

- [작동 방식](#)
- [고려 사항](#)
- [예약된 작업 생성, 관리 및 삭제에 일반적으로 사용되는 명령](#)
- [관련 리소스](#)
- [제한 사항](#)

작동 방식

예약된 조정을 사용하려면 Application Auto Scaling이 특정 시간에 조정 작업을 수행하도록 하는 예약된 작업을 생성할 수 있습니다. 예약 작업을 생성할 때 확장 가능한 대, 조정 활동이 발생할 시간, 최소 용량 및 최대 용량을 지정합니다. 규모를 한 번만 조정하거나 반복되는 일정으로 조정하도록 예약된 작업을 생성할 수 있습니다.

지정된 시간에 Application Auto Scaling은 현재 용량을 지정된 최소 및 최대 용량과 비교하여 새 용량 값을 기반으로 조정합니다.

- 현재 용량이 지정된 최소 용량보다 적을 경우 Application Auto Scaling이 지정된 최소 용량으로 확장합니다(용량 증가).
- 현재 용량이 지정된 최대 용량보다 클 경우 Application Auto Scaling이 지정된 최대 용량으로 축소합니다(용량 감소).

고려 사항

예약된 작업을 만들 경우, 다음 사항에 유의해야 합니다.

- 예약된 작업은 MinCapacity 및 MaxCapacity를 지정된 날짜와 시간에 예약된 작업에 의해 지정된 값으로 설정합니다. 요청에는 이러한 크기 중 하나만 선택적으로 포함할 수 있습니다. 예를 들어 최소 용량만 지정된 예약된 작업을 생성할 수 있습니다. 그러나 경우에 따라 새 최소 용량이 최대 용량보다 크지 않거나 새 최대 용량이 최소 용량보다 작지 않을 것을 보장하기 위해 두 크기를 모두 포함해야 합니다.
- 기본적으로 사용자가 설정한 반복 일정의 시간대는 UTC(협정 세계시)입니다. 현지 표준 시간대 또는 네트워크의 다른 부분에 대한 표준 시간대와 일치하도록 시간을 변경할 수 있습니다. 일광 절약 시간을 준수하는 시간대를 지정하는 경우, 작업이 DST(일광 절약 시간제)에 맞게 자동으로 조정됩니다. 자세한 정보는 [cron 표현식을 사용하여 반복 크기 조정 작업 예약](#)을 참조하세요.

- 확장 가능한 대상에 대해 예약된 조정을 일시적으로 해제할 수 있습니다. 이렇게 하면 예약된 작업을 삭제할 필요 없이 활성 상태가 되는 것을 방지할 수 있습니다. 그런 다음 다시 사용하려는 경우, 예약된 조정을 재개할 수 있습니다. 자세한 정보는 [Application Auto Scaling의 조정 일시 중지 및 재개](#)를 참조하세요.
- 예약 작업의 실행 순서는 확장 가능한 대상 전체가 아니라 동일한 확장 가능한 대상에 대해 보장됩니다.
- 예약된 작업을 성공적으로 완료하려면 지정된 리소스가 대상 서비스에서 확장 가능한 상태여야 합니다. 그렇지 않을 경우 요청이 실패하고 Resource Id [ActualResourceId] is not scalable. Reason: The status of all DB instances must be 'available' or 'incompatible-parameters'와 같은 오류 메시지를 반환합니다.
- Application Auto Scaling과 대상 서비스가 분산되어 있기 때문에 예약된 작업이 트리거 되는 시간과 대상 서비스가 조정 작업을 인식하는 시간 간에는 몇 초의 지연이 있을 수 있습니다. 예약된 작업은 지정된 순서대로 실행되기 때문에, 예약된 작업의 시작 시간이 서로 가까운 경우 실행하는 데 더 많은 시간이 소요될 수 있습니다.

예약된 작업 생성, 관리 및 삭제에 일반적으로 사용되는 명령

일정 조정 작업에 일반적으로 사용되는 명령은 다음과 같습니다.

- [register-scalable-target](#)을 사용하여 리소스를 확장 가능한 대상 (Application Auto Scaling에서 확장할 수 있는 리소스)으로 AWS 등록하고 조정을 일시 중지하고 재개합니다.
- [put-scheduled-action](#)을 사용하여 기존 확장 가능 대상에 대해 예약된 작업을 추가하거나 수정할 수 있습니다.
- 지역 내 규모 조정 활동에 대한 [정보를 반환하는 데 사용할 수 있는 확장 활동을 설명합니다.](#) AWS
- 특정 지역의 예정된 [작업에 대한 정보를 반환하는 스케줄링된 작업을 설명합니다.](#) AWS
- [delete-scheduled-action](#)을 사용하여 예약된 작업을 삭제할 수 있습니다.

관련 리소스

예약 규모 조정 사용에 대한 자세한 예는 Compute 블로그의 반복되는 최대 사용량에 대한 [AWS Lambda 프로비저닝된 동시성 예약](#) 블로그 게시물을 참조하십시오. AWS

샘플 AWS 리소스를 사용한 예약된 작업 생성 방법을 안내하는 자습서는 [자습서: AWS CLI을\(를\) 사용하여 예약된 조정 시작하기](#) 섹션을 참조하세요.

Auto Scaling의 예약된 작업 생성에 대한 자세한 내용은 Amazon EC2 Auto Scaling 사용 설명서의 [Amazon EC2 Auto Scaling 예약된 조정](#)을 참조하세요.

제한 사항

다음은 예약된 조정을 사용할 때의 제한 사항입니다.

- 예약된 작업의 이름은 확장 가능한 대상별로 고유해야 합니다.
- Application Auto Scaling은 일정 표현식에 초 단위의 정밀성을 제공하지 않습니다. cron 표현식을 사용해 가장 정밀하게 설정할 수 있는 단위가 1분입니다.
- Amazon MSK 클러스터는 확장 가능한 대상이 될 수 없습니다. Amazon MSK에서는 예약된 조정이 지원되지 않습니다.
- 확장 가능한 리소스에 대한 예약된 작업을 확인, 추가, 업데이트 또는 제거할 수 있는 콘솔 액세스 권한은 사용하는 리소스에 따라 다릅니다. 자세한 정보는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)을 참조하세요.

cron 표현식을 사용하여 반복 크기 조정 작업 예약

Important

Amazon EC2 Auto Scaling용 cron 표현식에 대한 도움이 필요한 경우 Amazon EC2 Auto Scaling 사용 설명서의 [반복 일정](#) 주제를 참조하세요. Amazon EC2 Auto Scaling의 경우, Application Auto Scaling에서 사용하는 사용자 지정 cron 구문 대신 기존 cron 구문을 사용합니다.

cron 표현식을 사용하여 반복 일정에 따라 실행되는 예약 작업을 생성할 수 있습니다.

반복 일정을 생성하려면 cron 표현식과 시간대를 지정하여 예약된 작업이 반복되는 시기를 지정합니다. 지원되는 시간대 값은 [Joda-Time](#)에서 지원하는 IANA 표준 시간대의 표준 이름입니다(예: Etc/GMT+9 또는 Pacific/Tahiti). 선택적으로 시작 시간, 해지 시간 또는 두 가지 모두에 대한 날짜 및 시간을 지정할 수 있습니다. 를 사용하여 예약된 작업을 생성하는 예제 명령은 AWS CLI 을 참조하십시오. [시간대를 지정하는 반복 예약 작업 생성](#)

지원되는 cron 표현식 형식은 [Minute] [Hour] [Day_of_Month] [Month_of_Year] [Day_of_Week]와 같이 공백으로 구분된 여섯 개의 필드로 구성됩니다. 예를 들어 cron 표현식 30 6 ? * MON *은(는) 매주

월요일 오전 6:30에 발생하는 예약된 작업을 구성합니다. 별표는 필드의 모든 값을 일치시키기 위한 와일드카드로 사용됩니다.

Application Auto Scaling 예약 작업의 크론 구문에 대한 자세한 내용은 Amazon EventBridge 사용 설명서의 [크론 표현식 참조](#)를 참조하십시오.

반복 일정을 생성할 때에는 시작 시간과 종료 시간을 신중하게 선택해야 합니다. 다음 사항에 유의하십시오:

- 시작 시간을 지정하면 Application Auto Scaling이 이 시간에 작업을 수행한 다음 지정된 반복에 따라 작업을 수행합니다.
- 해지 시간을 지정하면 이 시간 이후에는 작업이 반복되지 않습니다. Application Auto Scaling은 이전 값을 추적하지 않고 종료 시간 이후에 이전 값으로 되돌립니다.
- 또는 AWS SDK를 사용하여 예약된 작업을 AWS CLI 생성하거나 업데이트할 때는 시작 시간과 종료 시간을 UTC로 설정해야 합니다.

예제

Application Auto Scaling 확장 가능 대상에 대한 반복 일정을 생성하는 경우 다음 표를 참조할 수 있습니다. 다음 예는 Application Auto Scaling을 사용하여 예약된 작업을 생성하거나 업데이트하기 위한 올바른 구문입니다.

분	시간	일	월	요일	연도	의미
0	10	*	*	?	*	매일 오전 10시(UTC)에 실행
15	12	*	*	?	*	매일 오후 12시 15분(UTC)에 실행
0	18	?	*	월-금	*	매주 월요일부터 금요일까지 오후 6시

분	시간	일	월	요일	연도	의미
						(UTC)에 실행
0	8	1	*	?	*	매월 1일 오전 8시 (UTC)에 실행
0/15	*	*	*	?	*	15분마다 실행
0/10	*	?	*	월-금	*	월요일부터 금요일까지 10분마다 실행
0/5	8~17	?	*	월-금	*	월요일부터 금요일까지 오전 8시부터 오후 5시 55분(UTC) 사이에 5분마다 실행

예외

7개의 필드가 포함된 문자열 값을 사용하여 cron 표현식을 만들 수도 있습니다. 이 경우 처음 세 필드를 사용하여 예약된 작업이 실행되어야 하는 시간(초 포함)을 지정할 수 있습니다. 전체 cron 표현식은 [Seconds] [Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year]와 같이 공백으로 구분된 필드가 있습니다. 그러나 이 접근 방식을 사용한다고 해서 예약된 작업이 지정한 정확한 초 단위의 시간에 실행된다는 보장은 없습니다. 또한 일부 서비스 콘솔은 cron 표현식의 초 필드를 지원하지 않을 수 있습니다.

Application Auto Scaling의 예약된 작업 예제

다음 예제는 AWS CLI [put-scheduled-action](#) 명령으로 예약된 작업을 만드는 방법을 보여줍니다. 새로운 용량을 지정할 때 최소 용량, 최대 용량 또는 둘 다 지정할 수 있습니다.

간략하게 나타내기 위해 이 주제의 예제에서는 Application Auto Scaling과 통합되는 몇 가지 서비스에 대한 CLI 명령을 설명합니다. 다른 확장 가능 대상을 지정하려면 `--service-namespace`에 네임스페이스, `--scalable-dimension`에 확장 가능 차원, `--resource-id`에 리소스 ID를 지정합니다. 각 서비스에 대한 자세한 내용과 예는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)의 주제를 참조하세요.

를 사용할 때는 명령이 AWS CLI 프로필에 맞게 구성된 상태에서 실행된다는 점을 기억하십시오. AWS 리전 다른 리전에서 명령을 실행하려는 경우 프로필의 기본 리전을 변경하거나 명령에 `--region` 파라미터를 사용합니다.

내용

- [한 번만 발생하는 예약된 작업 생성](#)
- [반복되는 간격으로 실행되는 예약된 작업 만들기](#)
- [반복되는 일정으로 실행되는 예약된 작업 생성](#)
- [시간대를 지정하는 일회성 예약된 작업 만들기](#)
- [시간대를 지정하는 반복 예약 작업 생성](#)

한 번만 발생하는 예약된 작업 생성

지정된 날짜 및 시간에 확장 가능한 대상을 한 번만 자동으로 조정하려면 `--schedule "at(yyyy-mm-ddThh:mm:ss)"` 옵션을 사용합니다.

Example 예: 일회성 스케일 아웃.

다음은 특정 날짜 및 시간에 용량을 확장하기 위해 예약된 작업을 생성하는 예제입니다.

`--schedule(2021년 3월 31일, 오후 10시(UTC 기준))`에 지정된 날짜 및 시간에, MinCapacity에 지정된 값이 현재 용량보다 큰 경우 Application Auto Scaling은 MinCapacity로 확장합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource \
```

```
--scalable-dimension custom-resource:ResourceType:Property \  
--resource-id file://~/custom-resource-id.txt \  
--scheduled-action-name scale-out \  
--schedule "at(2021-03-31T22:00:00)" \  
--scalable-target-action MinCapacity=3
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource --  
scalable-dimension custom-resource:ResourceType:Property --resource-id file://~/custom-  
resource-id.txt --scheduled-action-name scale-out --schedule "at(2021-03-31T22:00:00)"  
--scalable-target-action MinCapacity=3
```

Note

이 예약된 작업이 실행될 때 최대 용량이 최소 용량에 지정된 값보다 작은 경우 최소 용량뿐만 아니라 새로운 최소 및 최대 용량을 지정해야 합니다.

Example 예: 일회성 축소.

다음은 특정 날짜 및 시간에 용량을 축소하기 위해 예약된 작업을 생성하는 예제입니다.

--schedule(2021년 3월 31일, 오후 10시 30분(UTC 기준))에 지정된 날짜 및 시간에, MaxCapacity에 지정된 값이 현재 용량보다 큰 경우 Application Auto Scaling은 MaxCapacity로 축소합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource \  
--scalable-dimension custom-resource:ResourceType:Property \  
--resource-id file://~/custom-resource-id.txt \  
--scheduled-action-name scale-in \  
--schedule "at(2021-03-31T22:30:00)" \  
--scalable-target-action MinCapacity=0,MaxCapacity=0
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource --  
scalable-dimension custom-resource:ResourceType:Property --resource-id file://~/custom-
```

```
resource-id.txt --scheduled-action-name scale-in --schedule "at(2021-03-31T22:30:00)"
--scalable-target-action MinCapacity=0,MaxCapacity=0
```

반복되는 간격으로 실행되는 예약된 작업 만들기

반복되는 간격으로 조정을 예약하려면 `--schedule "rate(value unit)"` 옵션을 사용합니다. 값은 양의 정수여야 합니다. 단위는 `minute`, `minutes`, `hour`, `hours`, `day` 또는 `days`가 될 수 있습니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 [요금 표현식](#)을 참조하십시오.

다음은 `rate` 표현식을 사용하는 예약된 작업의 예제입니다.

지정된 일정에 따라(2021년 1월 30일 오후 12시에 시작하여 2021년 1월 31일 오후 10시(UTC 기준)에 끝나는 5시간마다) `MinCapacity`에 지정된 값이 현재 용량보다 큰 경우, Application Auto Scaling이 `MinCapacity`로 확장합니다. `MaxCapacity`에 지정된 값이 현재 용량보다 작을 경우 Application Auto Scaling은 `MaxCapacity`로 축소합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace ecs \
--scalable-dimension ecs:service:DesiredCount \
--resource-id service/my-cluster/my-service \
--scheduled-action-name my-recurring-action \
--schedule "rate(5 hours)" \
--start-time 2021-01-30T12:00:00 \
--end-time 2021-01-31T22:00:00 \
--scalable-target-action MinCapacity=3,MaxCapacity=10
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service --scheduled-action-name my-recurring-action --schedule "rate(5 hours)" --start-time 2021-01-30T12:00:00 --end-time 2021-01-31T22:00:00 --scalable-target-action MinCapacity=3,MaxCapacity=10
```

반복되는 일정으로 실행되는 예약된 작업 생성

반복되는 일정으로 조정을 예약하려면 `--schedule "cron(fields)"` 옵션을 사용합니다. 자세한 정보는 [cron 표현식을 사용하여 반복 크기 조정 작업 예약](#)을 참조하세요.

다음은 `cron` 표현식을 사용하는 예약된 작업의 예제입니다.

지정된 일정(UTC 기준 매일 오전 9시)에 MinCapacity에 지정된 값이 현재 용량보다 큰 경우 Application Auto Scaling은 MinCapacity로 확장합니다. MaxCapacity에 지정된 값이 현재 용량보다 작을 경우 Application Auto Scaling은 MaxCapacity로 축소합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace appstream \
  --scalable-dimension appstream:fleet:DesiredCapacity \
  --resource-id fleet/sample-fleet \
  --scheduled-action-name my-recurring-action \
  --schedule "cron(0 9 * * ? *)" \
  --scalable-target-action MinCapacity=10,MaxCapacity=50
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace appstream --
scalable-dimension appstream:fleet:DesiredCapacity --resource-id fleet/sample-fleet --
scheduled-action-name my-recurring-action --schedule "cron(0 9 * * ? *)" --scalable-
target-action MinCapacity=10,MaxCapacity=50
```

시간대를 지정하는 일회성 예약된 작업 만들기

예약된 작업은 기본적으로 UTC 표준 시간대로 설정됩니다. 다른 시간대를 지정하려면 `--timezone` 옵션을 선택하고 시간대의 정식 이름(예: `America/New_York`)을 지정합니다. 자세한 내용은 <https://www.joda.org/joda-time/timezones.html>을 참조하세요. `put-scheduled-action`을 호출할 때 지원되는 IANA 시간대에 대한 정보를 제공합니다.

다음은 특정 날짜 및 시간에 용량을 조정하기 위해 예약된 작업을 생성할 때 `--timezone` 옵션을 사용하는 예제입니다.

`--schedule`(2021년 1월 31일, 오후 5시(현지 시간))에 지정된 날짜 및 시간에, MinCapacity에 지정된 값이 현재 용량보다 큰 경우 Application Auto Scaling은 MinCapacity로 확장합니다. MaxCapacity에 지정된 값이 현재 용량보다 작을 경우 Application Auto Scaling은 MaxCapacity로 축소합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace comprehend \
  --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits \
```



```
--resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE \
--scheduled-action-name my-one-time-action \
--schedule "at(2021-01-31T17:00:00)" --timezone "America/New_York" \
--scalable-target-action MinCapacity=1,MaxCapacity=3
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace comprehend --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/EXAMPLE --scheduled-action-name my-one-time-action --schedule "at(2021-01-31T17:00:00)" --timezone "America/New_York" --scalable-target-action MinCapacity=1,MaxCapacity=3
```

시간대를 지정하는 반복 예약 작업 생성

다음은 용량을 조정하기 위해 반복 예약 작업을 생성할 때 `--timezone` 옵션을 사용하는 예입니다. 자세한 정보는 [cron 표현식을 사용하여 반복 크기 조정 작업 예약](#)을 참조하세요.

지정된 일정(현지 시간 기준 매주 월요일~금요일 오후 6시)에 MinCapacity에 지정된 값이 현재 용량보다 큰 경우 Application Auto Scaling은 MinCapacity로 확장합니다. MaxCapacity에 지정된 값이 현재 용량보다 작을 경우 Application Auto Scaling은 MaxCapacity로 축소합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace lambda \
--scalable-dimension lambda:function:ProvisionedConcurrency \
--resource-id function:my-function:BLUE \
--scheduled-action-name my-recurring-action \
--schedule "cron(0 18 ? * MON-FRI *)" --timezone "Etc/GMT+9" \
--scalable-target-action MinCapacity=10,MaxCapacity=50
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace lambda --scalable-dimension lambda:function:ProvisionedConcurrency --resource-id function:my-function:BLUE --scheduled-action-name my-recurring-action --schedule "cron(0 18 ? * MON-FRI *)" --timezone "Etc/GMT+9" --scalable-target-action MinCapacity=10,MaxCapacity=50
```

Application Auto Scaling의 예약된 크기 조정 관리

AWS CLI 여기에는 예약된 작업을 관리하는 데 도움이 되는 몇 가지 다른 명령이 포함되어 있습니다.

간략하게 나타내기 위해 이 주제의 예제에서는 Application Auto Scaling과 통합되는 몇 가지 서비스에 대한 CLI 명령을 설명합니다. 다른 확장 가능 대상을 지정하려면 `--service-namespace`에 네임스페이스, `--scalable-dimension`에 확장 가능 차원, `--resource-id`에 리소스 ID를 지정합니다. 각 서비스에 대한 자세한 내용과 예는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)의 주제를 참조하세요.

를 사용할 때는 명령이 AWS CLI프로필에 AWS 리전 맞게 구성된 상태에서 실행된다는 점을 기억하십시오. 다른 리전에서 명령을 실행하려는 경우 프로필의 기본 리전을 변경하거나 명령에 `--region` 파라미터를 사용합니다.

내용

- [지정된 서비스에 대한 크기 조정 활동 보기](#)
- [지정된 서비스에 대해 예약된 모든 작업 설명](#)
- [확장 가능한 대상에 대해 하나 이상의 예약된 작업 설명](#)
- [확장 가능한 대상에 대해 예약된 크기 조정 해제](#)
- [예약된 작업 삭제](#)

지정된 서비스에 대한 크기 조정 활동 보기

지정된 서비스 네임스페이스의 모든 확장 가능한 대상에 대한 크기 조정 작업을 보려면 [describe-scaling-activities](#) 명령을 사용합니다.

다음 예제에서는 dynamodb 서비스 네임스페이스와 관련된 크기 조정 활동을 검색합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-activities --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace dynamodb
```

이 명령이 성공하면 다음과 비슷한 출력이 표시됩니다.

```

{
  "ScalingActivities": [
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting write capacity units to 10.",
      "ResourceId": "table/my-table",
      "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
      "StartTime": 1561574415.086,
      "ServiceNamespace": "dynamodb",
      "EndTime": 1561574449.51,
      "Cause": "maximum capacity was set to 10",
      "StatusMessage": "Successfully set write capacity units to 10. Change
successfully fulfilled by dynamodb.",
      "StatusCode": "Successful"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting min capacity to 5 and max capacity to 10",
      "ResourceId": "table/my-table",
      "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
      "StartTime": 1561574414.644,
      "ServiceNamespace": "dynamodb",
      "Cause": "scheduled action name my-second-scheduled-action was triggered",
      "StatusMessage": "Successfully set min capacity to 5 and max capacity to
10",
      "StatusCode": "Successful"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting write capacity units to 15.",
      "ResourceId": "table/my-table",
      "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
      "StartTime": 1561574108.904,
      "ServiceNamespace": "dynamodb",
      "EndTime": 1561574140.255,
      "Cause": "minimum capacity was set to 15",
      "StatusMessage": "Successfully set write capacity units to 15. Change
successfully fulfilled by dynamodb.",
      "StatusCode": "Successful"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting min capacity to 15 and max capacity to 20",

```

```

        "ResourceId": "table/my-table",
        "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
        "StartTime": 1561574108.512,
        "ServiceNamespace": "dynamodb",
        "Cause": "scheduled action name my-first-scheduled-action was triggered",
        "StatusMessage": "Successfully set min capacity to 15 and max capacity to
20",
        "StatusCode": "Successful"
    }
]
}

```

확장 가능한 대상 중 하나에 대해서만 조정 활동을 검색하도록 이 명령을 변경하려면 `--resource-id` 옵션을 추가합니다.

지정된 서비스에 대해 예약된 모든 작업 설명

지정된 서비스 네임스페이스의 모든 확장 가능한 대상에 대한 크기 조정 작업을 설명하려면 [describe-scheduled-actions](#) 명령을 사용합니다.

다음 예제에서는 ec2 서비스 네임스페이스와 관련된 예약된 작업을 검색합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2
```

이 명령이 성공하면 다음과 비슷한 출력이 반환됩니다.

```

{
  "ScheduledActions": [
    {
      "ScheduledActionName": "my-one-time-action",
      "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:493a6261-fbb9-432d-855d-3c302c14bdb9:resource/ec2/
spot-fleet-request/sfr-107dc873-0802-4402-a901-37294EXAMPLE:scheduledActionName/my-one-
time-action",
      "ServiceNamespace": "ec2",
      "Schedule": "at(2021-01-31T17:00:00)",

```

```
    "Timezone": "America/New_York",
    "ResourceId": "spot-fleet-request/sfr-107dc873-0802-4402-
a901-37294EXAMPLE",
    "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
    "ScalableTargetAction": {
      "MaxCapacity": 1
    },
    "CreationTime": 1607454792.331
  },
  {
    "ScheduledActionName": "my-recurring-action",
    "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:493a6261-fbb9-432d-855d-3c302c14bdb9:resource/ec2/
spot-fleet-request/sfr-107dc873-0802-4402-a901-37294EXAMPLE:scheduledActionName/my-
recurring-action",
    "ServiceNamespace": "ec2",
    "Schedule": "rate(5 minutes)",
    "ResourceId": "spot-fleet-request/sfr-107dc873-0802-4402-
a901-37294EXAMPLE",
    "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
    "StartTime": 1604059200.0,
    "EndTime": 1612130400.0,
    "ScalableTargetAction": {
      "MinCapacity": 3,
      "MaxCapacity": 10
    },
    "CreationTime": 1607454949.719
  },
  {
    "ScheduledActionName": "my-one-time-action",
    "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:4bce34c7-bb81-4ecf-b776-5c726efb1567:resource/ec2/
spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE:scheduledActionName/my-one-
time-action",
    "ServiceNamespace": "ec2",
    "Schedule": "at(2020-12-08T9:36:00)",
    "Timezone": "America/New_York",
    "ResourceId": "spot-fleet-request/sfr-40edeb7b-9ae7-44be-
bef2-5c4c8EXAMPLE",
    "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
    "ScalableTargetAction": {
      "MinCapacity": 1,
      "MaxCapacity": 3
    },
  },
```

```

        "CreationTime": 1607456031.391
    }
]
}

```

확장 가능한 대상에 대해 하나 이상의 예약된 작업 설명

지정된 확장 가능 대상에 대해 예약된 작업에 대한 정보를 검색하려면 [describe-scheduled-actions](#) 명령을 사용하여 예약된 작업을 설명할 때 `--resource-id` 옵션을 추가합니다.

`--scheduled-action-names` 옵션을 포함하고 예약된 작업의 이름을 그 값으로 지정하면, 다음 예제에서처럼 명령이 그 이름이 일치하는 예약된 작업만을 반환합니다.

Linux, macOS 또는 Unix

```

aws application-autoscaling describe-scheduled-actions --service-namespace ec2 \
  --resource-id spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE \
  --scheduled-action-names my-one-time-action

```

Windows

```

aws application-autoscaling describe-scheduled-actions --service-namespace ec2 --
resource-id spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE --scheduled-
action-names my-one-time-action

```

출력의 예제는 다음과 같습니다.

```

{
  "ScheduledActions": [
    {
      "ScheduledActionName": "my-one-time-action",
      "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:4bce34c7-bb81-4ecf-b776-5c726efb1567:resource/ec2/
spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE:scheduledActionName/my-one-
time-action",
      "ServiceNamespace": "ec2",
      "Schedule": "at(2020-12-08T9:36:00)",
      "Timezone": "America/New_York",
      "ResourceId": "spot-fleet-request/sfr-40edeb7b-9ae7-44be-
bef2-5c4c8EXAMPLE",
      "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",

```

```

        "ScalableTargetAction": {
            "MinCapacity": 1,
            "MaxCapacity": 3
        },
        "CreationTime": 1607456031.391
    }
]
}

```

--scheduled-action-names 옵션에 대해 둘 이상의 값이 제공되면 이름이 일치하는 예약된 작업이 모두 출력에 포함됩니다.

확장 가능한 대상에 대해 예약된 크기 조정 해제

예약된 작업을 삭제하지 않고도 예약된 조정을 일시적으로 해제할 수 있습니다. 자세한 내용은 [Application Auto Scaling의 조정 일시 중지 및 재개](#) 섹션을 참조하세요.

다음 예제에서처럼 --suspended-state 옵션과 함께 [register-scalable-target](#) 명령을 사용하고 ScheduledScalingSuspended 속성의 값으로 true를 지정하여 확장 가능 대상에 대한 예약된 작업을 일시 중지합니다.

Linux, macOS 또는 Unix

```

aws application-autoscaling register-scalable-target --service-namespace rds \
  --scalable-dimension rds:cluster:ReadReplicaCount --resource-id cluster:my-db-cluster \
  --suspended-state '{"ScheduledScalingSuspended": true}'

```

Windows

```

aws application-autoscaling register-scalable-target --service-namespace rds --
scalable-dimension rds:cluster:ReadReplicaCount --resource-id cluster:my-db-cluster --
suspended-state "{\"ScheduledScalingSuspended\": true}"

```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```

{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}

```

예약된 조정을 재개하려면 이 명령을 다시 실행하고 `ScheduledScalingSuspended` 속성값으로 `false`를 지정합니다.

예약된 작업 삭제

예약된 작업을 완료한 경우 [delete-scheduled-action](#) 명령을 사용하여 이를 삭제할 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling delete-scheduled-action --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-37294EXAMPLE \
  --scheduled-action-name my-recurring-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-37294EXAMPLE --scheduled-action-name my-recurring-action
```

성공하면 이 명령이 프롬프트를 반환합니다.

자습서: AWS CLI을(를) 사용하여 예약된 조정 시작하기

다음 자습서에서는 라는 샘플 DynamoDB 테이블을 AWS CLI 확장하는 예약된 작업을 생성하도록 지원하여 를 사용하여 예약 조정을 시작하는 방법을 보여줍니다. `TestTable` 테스트에 사용할 `TestTable` 테이블이 DynamoDB에 아직 없는 경우 Amazon DynamoDB 개발자 안내서의 [1단계: DynamoDB 테이블 생성](#)에 나와 있는 `create-table` 명령을 실행하여 지금 하나 생성할 수 있습니다.

를 사용할 때는 프로필에 구성된 AWS 리전에서 명령이 실행된다는 점을 기억하십시오. AWS CLI다른 리전에서 명령을 실행하려는 경우 프로필의 기본 리전을 변경하거나 명령에 `--region` 파라미터를 사용합니다.

Note

이 자습서에서는 AWS 요금이 부과될 수 있습니다. [프리 티어](#) 사용량을 모니터링하고 DynamoDB 데이터베이스에서 사용하는 읽기 및 쓰기 용량 단위 수와 관련된 비용을 이해해야 합니다.

내용

- [1단계: 확장 가능 대상 등록](#)
- [2단계: 2개의 예약된 작업 생성](#)
- [3단계: 조정 활동 보기](#)
- [4단계: 다음 단계](#)
- [5단계: 정리](#)

1단계: 확장 가능 대상 등록

Application Auto Scaling에서 확장 가능 대상으로 DynamoDB 테이블을 등록하는 것부터 시작합니다.

Application Auto Scaling을 통해 확장 가능 대상을 등록하려면

1. 먼저 [describe-scalable-targets](#) 명령을 사용하여 이미 등록된 DynamoDB 리소스가 있는지 확인합니다. 이를 통해 TestTable 테이블이 새 테이블이 아닌 경우 등록되지 않은 것을 확인할 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scalable-targets \
  --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb
```

기존의 확장 가능 대상이 없는 경우에는 다음과 같이 응답합니다.

```
{
  "ScalableTargets": []
}
```

2. 다음 [register-scalable-target](#) 명령을 사용하여 TestTable이라는 DynamoDB 테이블의 쓰기 용량을 등록합니다. 5개 쓰기 용량 유닛에 원하는 최소 용량과 10개 쓰기 용량 유닛에 원하는 최대 용량을 설정합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable \
  --min-capacity 5 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb
  --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
TestTable --min-capacity 5 --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-
id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

2단계: 2개의 예약된 작업 생성

Application Auto Scaling은 조정 작업을 수행해야 할 때 그 시점을 예약할 수 있도록 해줍니다. 확장 가능 대상, 일정, 최소 및 최대 용량을 지정하면 지정된 시간에 Application Auto Scaling이 조정 가능 대상의 최솟값 및 최댓값을 업데이트합니다. 현재 용량이 이 범위를 벗어나면 그 결과로 조정 활동이 이루어집니다.

최소 및 최대 용량으로 업데이트를 조정하면 조정 정책의 수립 여부를 결정할 때 도움이 됩니다. 조정 정책을 사용하면 현재 리소스 사용률에 따라 동적으로 리소스를 조정할 수 있습니다. 조정 정책을 위한 일반적인 가드레일에는 최소 및 최대 용량에 해당되는 값들이 포함되어 있습니다.

이 연습에서는 스케일 아웃 및 스케일 인을 위한 2개의 일회성 작업을 생성해 보겠습니다.

예약된 작업을 생성하고 확인하려면

1. 첫 번째 예약된 작업을 생성하려면 다음 [put-scheduled-action](#) 명령을 사용합니다.

--schedule의 at 명령은 미래의 지정된 날짜 및 시간에 한 번 실행되도록 작업을 예약합니다. 시간은 UTC 기준 24시간 형식입니다. 지금부터 약 5분 후에 작업이 이루어지도록 예약합니다.

지정된 날짜 및 시간이 되면 Application Auto Scaling이 MinCapacity 및 MaxCapacity 값을 업데이트합니다. 테이블에 현재 5개의 쓰기 용량 유닛이 있다고 가정한다면 Application Auto Scaling이 MinCapacity로 확장되어 테이블을 원하는 새 범위(15~20개의 쓰기 용량 유닛) 내로 가져옵니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable \
  --scheduled-action-name my-first-scheduled-action \
  --schedule "at(2019-05-20T17:05:00)" \
  --scalable-target-action MinCapacity=15,MaxCapacity=20
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace dynamodb
--scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
TestTable --scheduled-action-name my-first-scheduled-action --schedule
"at(2019-05-20T17:05:00)" --scalable-target-action MinCapacity=15,MaxCapacity=20
```

이 명령이 제대로 실행되면 어떤 출력도 반환하지 않습니다.

2. 축소를 위해 Application Auto Scaling에서 사용할 두 번째 예약된 작업을 생성하려면 다음 [put-scheduled-action](#) 명령을 사용합니다.

지금부터 약 10분 후에 작업이 이루어지도록 예약합니다.

지정된 날짜 및 시간이 되면 Application Auto Scaling이 테이블의 MinCapacity 및 MaxCapacity를 업데이트하고 MaxCapacity로 축소하여 원래 원하는 범위(5 ~ 10개의 쓰기 용량 유닛)로 테이블을 돌려놓습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable \
  --scheduled-action-name my-second-scheduled-action \
```

```
--schedule "at(2019-05-20T17:10:00)" \  
--scalable-target-action MinCapacity=5,MaxCapacity=10
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace dynamodb  
--scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/  
TestTable --scheduled-action-name my-second-scheduled-action --schedule  
"at(2019-05-20T17:10:00)" --scalable-target-action MinCapacity=5,MaxCapacity=10
```

3. (선택 사항) 다음 [describe-scheduled-actions](#) 명령을 사용하여 지정된 서비스 네임스페이스에 대한 예약된 작업 목록을 가져올 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scheduled-actions \  
--service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace dynamodb
```

출력의 예제는 다음과 같습니다.

```
{  
  "ScheduledActions": [  
    {  
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",  
      "Schedule": "at(2019-05-20T18:35:00)",  
      "ResourceId": "table/TestTable",  
      "CreationTime": 1561571888.361,  
      "ScheduledActionARN": "arn:aws:autoscaling:us-  
east-1:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/  
dynamodb/table/TestTable:scheduledActionName/my-first-scheduled-action",  
      "ScalableTargetAction": {  
        "MinCapacity": 15,  
        "MaxCapacity": 20  
      },  
      "ScheduledActionName": "my-first-scheduled-action",  
      "ServiceNamespace": "dynamodb"  
    },  
  ],  
}
```

```

    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Schedule": "at(2019-05-20T18:40:00)",
      "ResourceId": "table/TestTable",
      "CreationTime": 1561571946.021,
      "ScheduledActionARN": "arn:aws:autoscaling:us-
east-1:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/TestTable:scheduledActionName/my-second-scheduled-action",
      "ScalableTargetAction": {
        "MinCapacity": 5,
        "MaxCapacity": 10
      },
      "ScheduledActionName": "my-second-scheduled-action",
      "ServiceNamespace": "dynamodb"
    }
  ]
}

```

3단계: 조정 활동 보기

이 단계에서는 예약된 작업에 의해 트리거된 조정 활동을 보고 DynamoDB가 테이블의 쓰기 용량을 변경했는지 확인합니다.

조정 활동을 보려면

1. 선택한 시간 동안 기다렸다가 다음 [describe-scaling-activities](#) 명령을 사용하여 예약된 작업이 수행 중인지 확인합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-activities \
  --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-
namespace dynamodb
```

다음은 예약된 작업이 진행되는 동안 첫 번째 예약된 작업의 출력 예제입니다.

조정 활동은 생성 날짜순으로 정렬되며, 가장 최근의 조정 활동이 가장 먼저 반환됩니다.

```
{
  "ScalingActivities": [
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting write capacity units to 15.",
      "ResourceId": "table/TestTable",
      "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
      "StartTime": 1561574108.904,
      "ServiceNamespace": "dynamodb",
      "Cause": "minimum capacity was set to 15",
      "StatusMessage": "Successfully set write capacity units to 15. Waiting
for change to be fulfilled by dynamodb.",
      "StatusCode": "InProgress"
    },
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting min capacity to 15 and max capacity to 20",
      "ResourceId": "table/TestTable",
      "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
      "StartTime": 1561574108.512,
      "ServiceNamespace": "dynamodb",
      "Cause": "scheduled action name my-first-scheduled-action was
triggered",
      "StatusMessage": "Successfully set min capacity to 15 and max capacity
to 20",
      "StatusCode": "Successful"
    }
  ]
}
```

다음은 두 예약된 작업이 모두 실행되고 난 후의 출력 예제입니다.

```
{
  "ScalingActivities": [
    {
      "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
      "Description": "Setting write capacity units to 10.",
      "ResourceId": "table/TestTable",
      "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
      "StartTime": 1561574415.086,
```

```

    "ServiceNamespace": "dynamodb",
    "EndTime": 1561574449.51,
    "Cause": "maximum capacity was set to 10",
    "StatusMessage": "Successfully set write capacity units to 10. Change
successfully fulfilled by dynamodb.",
    "StatusCode": "Successful"
  },
  {
    "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
    "Description": "Setting min capacity to 5 and max capacity to 10",
    "ResourceId": "table/TestTable",
    "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
    "StartTime": 1561574414.644,
    "ServiceNamespace": "dynamodb",
    "Cause": "scheduled action name my-second-scheduled-action was
triggered",
    "StatusMessage": "Successfully set min capacity to 5 and max capacity
to 10",
    "StatusCode": "Successful"
  },
  {
    "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
    "Description": "Setting write capacity units to 15.",
    "ResourceId": "table/TestTable",
    "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
    "StartTime": 1561574108.904,
    "ServiceNamespace": "dynamodb",
    "EndTime": 1561574140.255,
    "Cause": "minimum capacity was set to 15",
    "StatusMessage": "Successfully set write capacity units to 15. Change
successfully fulfilled by dynamodb.",
    "StatusCode": "Successful"
  },
  {
    "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
    "Description": "Setting min capacity to 15 and max capacity to 20",
    "ResourceId": "table/TestTable",
    "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
    "StartTime": 1561574108.512,
    "ServiceNamespace": "dynamodb",
    "Cause": "scheduled action name my-first-scheduled-action was
triggered",
    "StatusMessage": "Successfully set min capacity to 15 and max capacity
to 20",

```

```

        "StatusCode": "Successful"
    }
]
}

```

2. 예약된 작업을 성공적으로 실행한 후 DynamoDB 콘솔을 열고 작업할 테이블을 선택합니다. 용량 (Capacity) 탭에서 쓰기 용량 유닛(Write capacity units)을 확인합니다. 두 번째 조정 작업이 실행되고 나면 쓰기 용량 유닛이 15에서 10으로 조정되었을 것입니다.

다음 [describe-table](#) 명령을 사용하여 테이블의 현재 쓰기 용량을 확인할 수도 있습니다. 출력을 필터링하려면 `--query` 옵션을 포함합니다. 이 출력 필터링 기능에 대한 자세한 내용은 AWS Command Line Interface 사용 설명서의 AWS CLI [에서 명령 출력 제어를](#) 참조하십시오. AWS CLI

Linux, macOS 또는 Unix

```

aws dynamodb describe-table --table-name TestTable \
  --query 'Table.[TableName,TableStatus,ProvisionedThroughput]'

```

Windows

```

aws dynamodb describe-table --table-name TestTable --query "Table.
[TableName,TableStatus,ProvisionedThroughput]"

```

출력의 예제는 다음과 같습니다.

```

[
  "TestTable",
  "ACTIVE",
  {
    "NumberOfDecreasesToday": 1,
    "WriteCapacityUnits": 10,
    "LastIncreaseDateTime": 1561574133.264,
    "ReadCapacityUnits": 5,
    "LastDecreaseDateTime": 1561574435.607
  }
]

```


4단계: 다음 단계

예약된 조정과 조정 정책을 모두 사용하여 조정을 시도하려면 [자습서: 과중한 워크로드를 처리하도록 Auto Scaling 구성](#)의 단계를 따르세요.

5단계: 정리

시작하기 연습을 마쳤으면 다음과 같이 연결된 리소스를 정리할 수 있습니다.

예약된 작업을 삭제하려면

다음 [delete-scheduled-action](#) 명령은 지정된 예약된 작업을 삭제합니다. 이 예약된 작업을 나중에도 계속 사용하고 싶은 경우에는 이 단계를 건너뛸 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling delete-scheduled-action \  
  --service-namespace dynamodb \  
  --scalable-dimension dynamodb:table:WriteCapacityUnits \  
  --resource-id table/TestTable \  
  --scheduled-action-name my-second-scheduled-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace dynamodb --  
scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/TestTable --  
scheduled-action-name my-second-scheduled-action
```

확장 가능 대상의 등록을 취소하려면

다음 [deregister-scalable-target](#) 명령을 사용하여 확장 가능 대상의 등록을 취소합니다. 이미 생성한 조정 정책이 있거나 아직 삭제되지 않은 예약된 작업이 있는 경우에는 이 명령을 통해 삭제됩니다. 확장 가능 대상을 나중에도 사용할 수 있도록 등록 상태로 유지하고 싶은 경우에는 이 단계를 건너뛸 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling deregister-scalable-target \  
  --service-namespace dynamodb \  
  --scalable-dimension dynamodb:table:WriteCapacityUnits \  
  --resource-id table/TestTable
```

```
--resource-id table/TestTable
```

Windows

```
aws application-autoscaling deregister-scalable-target --service-namespace dynamodb --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/TestTable
```

DynamoDB 테이블을 삭제하려면

다음 [delete-table](#) 명령을 사용하여 이 자습서에서 사용한 테이블을 삭제합니다. 이 테이블을 나중에도 계속 사용하고 싶은 경우에는 이 단계를 건너뛸 수 있습니다.

Linux, macOS 또는 Unix

```
aws dynamodb delete-table --table-name TestTable
```

Windows

```
aws dynamodb delete-table --table-name TestTable
```

Application Auto Scaling의 대상 추적 조정 정책

대상 추적 조정 정책은 대상 지표값을 기준으로 애플리케이션을 자동으로 조정합니다. 이를 통해 애플리케이션은 수동 개입 없이 최적의 성능과 비용 효율성을 유지할 수 있습니다.

대상 추적을 사용할 때는 애플리케이션의 이상적인 평균 사용률 또는 처리량 수준을 나타내는 지표와 목표 값을 선택합니다. Application Auto Scaling은 지표가 대상에서 벗어날 때 조정 이벤트를 트리거하는 CloudWatch 경보를 생성하고 관리합니다. 이는 온도 조절기가 목표 온도를 유지하는 방법과 비슷합니다.

예를 들어, 현재 스팟 플릿에서 애플리케이션이 실행되고 있고 사용자가 애플리케이션 로드 변경이 있는 경우 플릿의 CPU 사용량을 50% 정도로 유지시키려 한다고 가정해 보겠습니다. 이로 인해 과도한 유휴 리소스를 유지하지 않고도 트래픽 급증을 처리할 수 있는 추가 용량을 확보할 수 있습니다.

평균 CPU 사용률 50%를 목표로 하는 대상 추적 조정 정책을 생성하면 이러한 요건을 충족할 수 있습니다. 그러면 CPU가 50%를 초과하면 Application Auto Scaling이 스케일 아웃(용량 증가)하여 증가된 로드를 처리합니다. CPU가 50% 미만으로 떨어지면 스케일 인(용량 감소)하여 사용률이 낮은 기간 동안 비용을 최적화합니다.

대상 추적 정책을 사용하면 CloudWatch 경보와 조정 조정을 수동으로 정의할 필요가 없습니다. Application Auto Scaling은 사용자가 설정한 대상에 따라 이를 자동으로 처리합니다.

사전 정의된 지표 또는 사용자 지정 지표를 기준으로 대상 추적 정책을 설정할 수 있습니다.

- 사전 정의된 지표 — 평균 CPU 사용률 또는 대상당 평균 요청 수와 같이 Application Auto Scaling에서 제공하는 지표입니다.
- 사용자 지정 지표 — 지표 수학을 사용하여 지표를 결합하거나, 기존 지표를 활용하거나, 게시된 자체 사용자 지정 지표를 사용할 수 있습니다. CloudWatch

확장 가능한 대상의 용량 변화에 반비례하여 변경되는 지표를 선택하세요. 따라서 용량을 두 배로 늘리면 지표가 50% 감소합니다. 이렇게 하면 지표 데이터가 비례적 조정 이벤트를 정확하게 트리거할 수 있습니다.

내용

- [Application Auto Scaling의 대상 추적 스케일링 작동 방식](#)
- [를 사용하여 Application Auto Scaling에 대한 대상 추적 조정 정책을 생성합니다. AWS CLI](#)
- [지표 수학을 사용하여 Application Auto Scaling에서 대상 추적 조정 정책 생성](#)

Application Auto Scaling의 대상 추적 스케일링 작동 방식

이 항목에서는 대상 추적 규모 조정 작동 방식을 설명하고 대상 추적 조정 정책의 주요 요소를 소개합니다.

내용

- [작동 방식](#)
- [지표 선택](#)
- [목표 값 정의](#)
- [휴지 기간 정의](#)
- [고려 사항](#)
- [여러 조정 정책](#)
- [조정 정책 생성, 관리 및 삭제를 위해 일반적으로 사용되는 명령](#)
- [관련 리소스](#)
- [제한 사항](#)

작동 방식

목표 추적 조정을 사용하려면 대상 추적 조정 정책을 생성하고 다음을 지정해야 합니다.

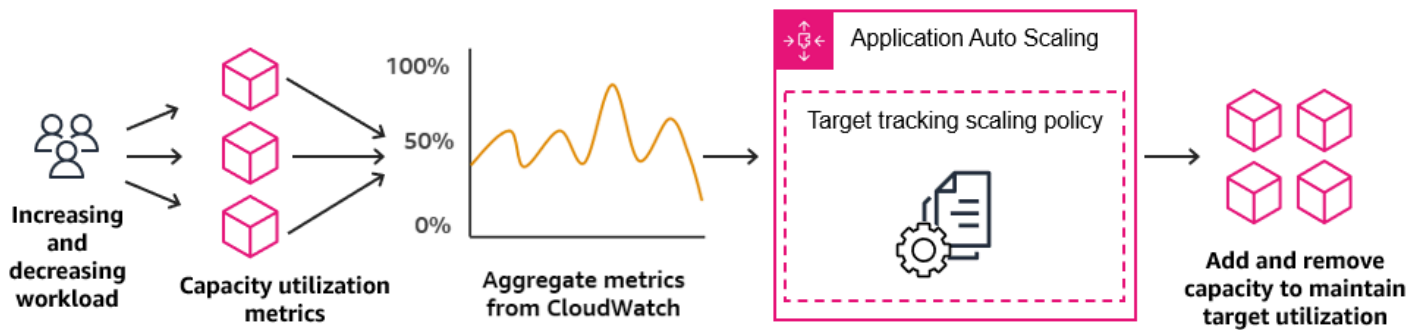
- 지표 - 평균 CPU 사용률 또는 대상당 평균 요청 수와 같은 추적할 CloudWatch 지표입니다.
- 대상 값 - 지표의 대상 값(예: CPU 사용률 50% 또는 대상당 분당 1,000개 요청)입니다.

Application Auto Scaling은 조정 정책을 호출하는 CloudWatch 경보를 생성 및 관리하고 지표와 대상 값을 기반으로 조정 조정을 계산합니다. 이는 필요에 따라 용량을 추가 및 제거하여 지표를 지정한 목표값으로, 혹은 목표값에 가깝게 유지합니다.

지표가 목표 값을 초과하면 Application Auto Scaling은 용량을 추가하여 메트릭 값과 대상 값 간의 차이를 줄임으로써 스케일 아웃합니다. 지표가 목표 값보다 낮으면 Application Auto Scaling은 용량을 제거하여 스케일 인합니다.

조정 작업은 용량의 급격한 변동을 방지하기 위해 작업 중간에 휴지 기간을 두고 수행됩니다. 선택적으로 조정 정책에 대한 휴지 기간을 구성할 수 있습니다.

다음 다이어그램은 설정이 완료될 때 대상 추적 조정 정책의 작동에 대한 개요를 보여줍니다.



대상 추적 조정 정책은 사용률이 증가할 때 용량을 추가하는 것이 사용률이 감소할 때 용량을 제거하는 것보다 더 적극적으로 적용된다는 점에 유의하세요. 예를 들어 정책의 지정된 지표가 대상 값에 도달하면 정책은 애플리케이션이 이미 많이 로드된 것으로 가정합니다. 따라서 최대한 빨리 지표 값에 비례하는 용량을 추가하여 응답합니다. 지표가 높을수록 더 많은 용량이 추가됩니다.

지표가 대상 값 미만으로 떨어지면 정책은 사용률이 결국 다시 증가할 것으로 예상합니다. 이 경우 사용률이 대상 값보다 충분히 낮은 임계값(일반적으로 10% 이상 낮음)에 도달하는 경우에만 사용률이 낮은 것으로 간주되어 용량을 제거하므로 조정 속도가 느려집니다. 이러한 더 보수적인 동작의 목적은 애플리케이션에서 더 이상 이전과 동일한 수준의 수요가 발생하지 않을 때만 용량을 제거하는 것입니다.

지표 선택

맞춤 또는 사전 정의된 지표를 사용하여 대상 추적 크기 조정 정책을 생성할 수 있습니다.

사전 정의된 지표 유형을 사용하여 대상 추적 크기 조정 정책을 생성하는 경우 [대상 추적 조정 정책을 위해 사전 정의된 지표](#)에 사전 정의된 지표 목록에서 지표를 하나 선택합니다.

지표를 선택하는 경우, 다음 사항에 유의하세요.

- 모든 사용자 지정 지표를 대상 추적에 사용할 수 있는 것은 아닙니다. 지표는 유효한 사용량 수치로서 확장 가능한 대상의 사용량을 설명해야 합니다. 지표 값은 확장 가능한 대상의 용량과 비례하여 증가하거나 감소해야만 지표 데이터에 따라 확장 가능한 대상을 늘리거나 줄일 수 있습니다.
- ALBRequestCountPerTarget 지표를 사용하려면 ResourceLabel 파라미터를 지정하여 지표와 연관된 대상 그룹을 식별해야 합니다.
- 예를 들어 지표가 실제 0 값을 내보내는 경우 CloudWatch (예:ALBRequestCountPerTarget), 일정 기간 동안 애플리케이션에 대한 트래픽이 없을 때 Application Auto Scaling은 0으로 확장할 수 있습니다. 확장 가능 대상에 요청이 라우팅되지 않은 경우 대상을 0으로 축소하려면, 확장 가능 대상의 최소 용량을 0으로 설정해야 합니다.

- 조정 정책에 사용할 새 지표를 게시하는 대신 지표 수학을 사용하여 기존 지표를 결합할 수 있습니다. 자세한 정보는 [지표 수학을 사용하여 Application Auto Scaling에서 대상 추적 조정 정책 생성](#)을 참조하세요.
- 사용하는 서비스가 서비스 콘솔에서 사용자 지정 지표 지정을 지원하는지 확인하려면 해당 서비스에 대한 설명서를 참조하세요.
- 사용률 변화에 따라 더 빠르게 조정할 수 있도록 1분 간격으로 제공되는 지표를 사용할 것을 권장합니다. 대상 추적은 사전 정의된 모든 지표와 사용자 지정 지표에 대해 1분 단위로 집계된 지표를 평가하지만, 기본 지표는 데이터를 게시하는 빈도가 낮을 수 있습니다. 예컨대, 모든 Amazon EC2 지표는 기본적으로 5분 간격으로 전송되지만, 1분 간격으로 구성할 수 있습니다(세부 모니터링이라고 함). 이 선택은 개별 서비스에 따라 달라집니다. 대부분의 경우 가능한 가장 짧은 간격을 사용하려고 합니다.

목표 값 정의

대상 추적 조정 정책을 생성할 경우, 목표 값을 지정해야 합니다. 목표 값은 애플리케이션의 최적 평균 사용률 또는 처리량(throughput)을 나타냅니다. 리소스를 비용 효율적으로 사용하려면 예상치 못한 트래픽 증가에 대비하여, 적절한 버퍼를 두고 목표 값을 가능한 한 높게 설정합니다. 애플리케이션이 정상적인 트래픽 흐름을 위해 최적으로 스케일 아웃되면 실제 지표 값은 목표 값과 같거나 그보다 조금 낮아야 합니다.

조정 정책이 Application Load Balancer 대상당 요청 수, 네트워크 I/O 또는 기타 수 지표와 같은 처리량(throughput)을 기반으로 하는 경우 대상 값은 1분 동안 단일 엔터티(예: Application Load Balancer 대상 그룹의 단일 대상)의 최적 평균 처리량(throughput)을 나타냅니다.

휴지 기간 정의

대상 추적 조정 정책에서 휴지 기간을 선택적으로 정의할 수 있습니다.

이전 조정 활동이 적용될 때까지 기다리는 시간을 휴지 기간이라고 합니다.

휴지 기간에는 다음과 같은 두 가지 유형이 있습니다.

- 확장 휴지 기간을 사용하는 경우, 지속적이지만 과도하지는 않게 확장하는 것이 목적입니다. Application Auto Scaling에서 조정 정책을 사용하여 성공적으로 스케일 아웃하면 휴지 기간이 계산되기 시작합니다. 조정 정책은 더 큰 스케일 아웃이 트리거되거나 휴지 기간이 종료되지 않는 한 원하는 용량을 다시 늘리지 않습니다. 확장 휴지 기간이 진행되는 동안 확장 활동을 시작하여 추가된 용량은 다음 확장 활동에 대해 원하는 용량의 일부로 계산됩니다.

- 스케일 인 휴지 기간을 사용하는 경우 애플리케이션의 가용성을 보호하기 위해 보수적으로 확장하므로 스케일 인 휴지 기간이 만료될 때까지 스케일 인 활동이 차단됩니다. 그러나 축소 휴지 기간 중에 다른 경보가 확장 활동을 트리거하면 Application Auto Scaling은 대상을 즉시 확장합니다. 이 경우 스케일 인 휴지 기간이 중지되고 완료되지 않습니다.

각 휴지 기간은 초 단위로 측정되며 정책 관련 조정 활동 조정에만 적용됩니다. 휴지 기간에 예약된 작업이 예약된 시간에 시작되면 휴지 기간이 만료될 때까지 기다리지 않고 조정 활동을 즉시 트리거할 수 있습니다.

나중에 미세 조정할 수 있는 기본값으로 시작할 수 있습니다. 예를 들어 대상 추적 조정 정책이 짧은 기간에 발생하는 변경 사항에 대해 지나치게 공격적이지 않도록 휴지 기간을 늘려야 할 수 있습니다.

기본값

Application Auto Scaling은 ElastiCache 복제 그룹의 기본값인 600을 제공하고 다음과 같은 확장 가능한 대상에 대해 기본값인 300을 제공합니다.

- AppStream 2.0 플릿
- Aurora DB 클러스터
- ECS 서비스
- Neptune 클러스터
- SageMaker 엔드포인트 변형
- SageMaker 추론 구성 요소
- SageMaker 서버리스 프로비저닝 동시성
- Spot Fleets
- 사용자 정의 리소스

다른 모든 확장 가능한 대상의 경우 기본값은 0 또는 null입니다.

- Amazon Comprehend 문서 분류 및 엔터티 인식기 엔드포인트
- DynamoDB 테이블 및 글로벌 보조 인덱스
- Amazon Keyspaces 테이블
- Lambda 프로비저닝된 동시성
- Amazon MSK 브로커 스토리지

Application Auto Scaling Scaling에서 휴지 기간을 평가할 때 Null 값은 0 값 동일하게 처리됩니다.

null 값을 비롯한 모든 기본값을 업데이트하여 휴지 기간을 직접 설정할 수 있습니다.

고려 사항

대상 추적 조정 정책과 관련한 작업을 수행할 때는 다음 고려 사항이 적용됩니다.

- 대상 추적 조정 정책과 함께 사용되는 CloudWatch 경보를 생성, 편집 또는 삭제하지 마십시오. Application Auto Scaling은 대상 추적 조정 정책과 관련된 CloudWatch 경보를 생성 및 관리하고 더 이상 필요하지 않을 경우 이를 삭제합니다.
- 지표에 데이터 포인트가 누락된 경우 CloudWatch 경보 상태가 `INSUFFICIENT_DATA` 로 변경됩니다. `INSUFFICIENT_DATA` 이 경우, 새 데이터 포인트를 찾을 때까지 Application Auto Scaling이 확장 가능 대상을 조정할 수 없습니다. 자세한 내용은 Amazon CloudWatch User Guide의 [CloudWatch 경보가 누락된 데이터를 처리하는 방법 구성](#)을 참조하십시오.
- 설계상 지표가 드물게 보고되는 경우, 지표 수확이 유용할 수 있습니다. 예를 들어, 가장 최근 값을 사용하려면 `m1`이 지표에 있는 `FILL(m1, REPEAT)` 함수를 사용하세요.
- 대상 값과 실제 지표 데이터 포인트 사이에는 차이가 발생할 수 있습니다. Application Auto Scaling이 추가하거나 제거할 용량을 결정할 때마다 항상 반올림 또는 내림을 통해 어림짐작으로 동작하기 때문입니다. 이는 용량을 부족하게 추가하거나 너무 많이 제거하는 일을 방지하기 위해서입니다. 하지만 용량이 작은 확장 가능한 대상의 경우 실제 지표 데이터 포인트는 대상 값과 멀어질 수도 있습니다.
용량이 큰 확장 가능한 대상의 경우 용량을 추가 또는 제거하면 대상 값과 실제 지표 데이터 포인트 사이의 차이를 줄일 수 있습니다.
- 대상 추적 조정 정책은 지정한 지표가 목표 값을 초과할 때 한해서 확장을 수행해야 합니다. 대상 추적 조정 정책에서는 지정한 지표가 목표 값보다 작을 때 확장할 수 없습니다.

여러 조정 정책

각각 다른 지표를 사용한다는 전제하에 확장 가능한 대상에 대해 다수의 대상 추적 조정 정책을 보유할 수 있습니다. Application Auto Scaling은 항상 가용성을 우선시하므로, 대상 추적 정책이 확장 또는 축소를 허용하는지에 따라 그 동작이 달라집니다. 대상 추적 정책 중 하나라도 확장을 허용할 경우 확장 가능한 대상을 확장하지만 모든 대상 추적 정책(축소 부분이 활성화됨)이 축소를 허용하는 경우에만 대상을 축소합니다.

여러 조정 정책이 확장 가능한 대상에 스케일 아웃 또는 스케일 인을 동시에 지시하는 경우 Application Auto Scaling은 스케일 아웃과 스케일 인 모두에 대해 가장 큰 용량을 제공하는 정책에 따라 조정합니

다. 이로써 다양한 시나리오를 수용할 만큼 폭넓은 유연성을 발휘할 뿐만 아니라 워크로드를 처리하는데 필요한 용량을 항상 충분히 확보할 수 있습니다.

대상 추적 조정 정책의 스케일 인 부분을 비활성화하여 스케일 아웃에 사용하는 것보다 다양한 스케일 인 방법을 사용할 수 있습니다. 예를 들어 확장을 위해 대상 추적 조정 정책을 사용하고 축소를 위해 단계 조정 정책을 사용할 수 있습니다.

그러나 대상 추적 조정 정책과 단계별 조정 정책을 함께 사용하는 경우, 정책 간 충돌로 인해 바람직하지 않은 동작이 발생할 수 있으므로 주의해야 합니다. 예를 들어 대상 추적 정책이 축소 준비되기 전에 단계 조정 정책이 축소 활동을 시작하는 경우 축소 활동이 차단되지 않습니다. 축소 작업이 완료된 후 대상 추적 정책이 확장 가능한 대상에 다시 확장하도록 지시할 수 있습니다.

기본적으로 주기적 워크로드의 경우 예약된 조정을 사용하여 일정에 따라 용량 변경을 자동화할 수도 있습니다. 예약된 각 작업에 대해 새 최소 용량 값과 새 최대 용량 값을 정의할 수 있습니다. 이러한 값은 조정 정책의 경계를 형성합니다. 예약된 조정과 대상 추적 조정을 함께 사용하면 용량이 즉시 필요할 때 사용률 수준이 급격히 증가하는 영향을 줄일 수 있습니다.

조정 정책 생성, 관리 및 삭제를 위해 일반적으로 사용되는 명령

조정 정책 작업에 일반적으로 사용되는 명령은 다음과 같습니다.

- [register-scalable-target](#)을 사용하여 리소스를 확장 가능한 대상 (Application Auto Scaling에서 확장할 수 있는 리소스) 으로 AWS 등록하고 조정을 일시 중지하고 재개합니다.
- [put-scaling-policy](#)를 사용하여 기존 확장 가능 대상에 대한 조정 정책을 추가하거나 수정할 수 있습니다.
- 지역 내 규모 조정 활동에 대한 [정보를 반환하는 데 사용할 수 있는 확장 활동을 설명합니다.](#) AWS
- [describe-scaling-activities](#)를 사용하여 조정 정책에 대한 정보를 AWS 리전으로 반환할 수 있습니다.
- [delete-scaling-policy](#)를 사용하여 조정 정책을 삭제할 수 있습니다.

관련 리소스

Auto Scaling의 대상 추적 조정 정책 생성에 대한 자세한 내용은 Amazon EC2 Auto Scaling 사용 설명서의 [Amazon EC2 Auto Scaling 대상 추적 조정 정책](#)을 참조하세요.

제한 사항

대상 추적 조정 정책을 사용할 때의 제한 사항은 다음과 같습니다.

- 확장 가능 대상은 Amazon EMR 클러스터가 될 수 없습니다. Amazon EMR에 대해서는 대상 추적 조정 정책이 지원되지 않습니다.
- Amazon MSK 클러스터가 확장 가능한 대상인 경우 축소가 비활성화되어 활성화할 수 없습니다.
- RegisterScalableTarget 또는 PutScalingPolicy API 작업을 사용하여 규모 조정 계획을 업데이트할 수 없습니다. AWS Auto Scaling
- 확장 가능한 리소스에 대한 대상 추적 조정 정책을 표시, 추가, 업데이트 또는 제거할 수 있는 콘솔 액세스 권한은 사용하는 리소스에 따라 다릅니다. 자세한 정보는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)을 참조하세요.

를 사용하여 Application Auto Scaling에 대한 대상 추적 조정 정책을 생성합니다. AWS CLI

다음 구성 작업에 를 사용하여 Application Auto Scaling에 AWS CLI 대한 대상 추적 조정 정책을 생성할 수 있습니다.

1. 확장 가능 대상을 등록합니다.
2. 확장 가능한 대상에 대상 추적 조정 정책을 추가합니다.

간략하게 나타내기 위해 이 주제의 예제에서는 Amazon EC2 스팟 플릿에 대한 CLI 명령을 설명합니다. 다른 확장 가능 대상을 지정하려면 --service-namespace에 네임스페이스, --scalable-dimension에 확장 가능 차원, --resource-id에 리소스 ID를 지정합니다. 각 서비스에 대한 자세한 내용과 예는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)의 주제를 참조하세요.

를 사용할 때는 명령이 AWS CLI 사용자 프로필에 AWS 리전 구성된 상태에서 실행된다는 점을 기억하십시오. 다른 리전에서 명령을 실행하려는 경우 프로필의 기본 리전을 변경하거나 명령에 --region 파라미터를 사용합니다.

내용

- [확장 가능 대상 등록](#)
- [대상 추적 조정 정책 생성](#)
- [대상 추적 조정 정책 설명](#)
- [대상 추적 조정 정책 삭제](#)

확장 가능 대상 등록

아직 하지 않았다면 확장 가능 대상을 등록합니다. [register-scalable-target](#) 명령을 사용하여 대상 서비스의 특정 리소스를 확장 가능 대상으로 등록합니다. 다음 예제에서는 Application Auto Scaling으로 스팟 플릿 요청을 등록합니다. Application Auto Scaling에서는 스팟 플릿의 인스턴스 수를 최소 2개와 최대 10개 사이에서 조정할 수 있습니다. *user input placeholder*를 사용자의 정보로 바꿉니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --min-capacity 2 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace ec2 --
scalable-dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-
request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --min-capacity 2 --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

대상 추적 조정 정책 생성

목표 추적 조정 정책을 만들려면 다음 예를 사용하여 시작하는 데 도움이 될 수 있습니다.

대상 추적 조정 정책을 생성하려면

1. 다음 cat 명령을 사용하여 조정 정책의 목표 값과 사전 정의된 지표 사양을 `config.json` 디렉터리에 이름이 지정된 JSON 파일에 저장합니다. 다음은 평균 CPU 사용률을 50%로 유지하는 대상 추적 구성의 예입니다.

```
$ cat ~/config.json
{
  "TargetValue": 50.0,
```

```
"PredefinedMetricSpecification":
{
  "PredefinedMetricType": "EC2SpotFleetRequestAverageCPUUtilization"
}
}
```

자세한 내용은 Application Auto Scaling API 참조의 PredefinedMetric [사양](#)을 참조하십시오.

또는 사용자 지정 지표 사양을 만들고 에서 각 파라미터의 값을 추가하여 사용자 지정 지표를 사용하여 규모를 조정할 수도 CloudWatch 있습니다. 다음은 지정된 지표의 평균 사용률을 100으로 유지하는 대상 추적 구성의 예입니다.

```
$ cat ~/config.json
{
  "TargetValue": 100.0,
  "CustomizedMetricSpecification":{
    "MetricName": "MyUtilizationMetric",
    "Namespace": "MyNamespace",
    "Dimensions": [
      {
        "Name": "MyOptionalMetricDimensionName",
        "Value": "MyOptionalMetricDimensionValue"
      }
    ],
    "Statistic": "Average",
    "Unit": "Percent"
  }
}
```

자세한 내용은 Application Auto Scaling API 참조의 CustomizedMetric [사양](#)을 참조하십시오.

- 다음 [put-scaling-policy](#) 명령을 생성한 config.json 파일과 함께 사용하여 cpu50-target-tracking-scaling-policy라는 조정 정책을 생성합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scaling-policy --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --policy-name cpu50-target-tracking-scaling-policy --policy-type
TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration file://config.json
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --policy-name cpu50-target-tracking-
scaling-policy --policy-type TargetTrackingScaling --target-tracking-scaling-
policy-configuration file://config.json
```

성공하면 이 명령은 사용자 대신 생성된 두 CloudWatch 경보의 ARN과 이름을 반환합니다.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-
id:scalingPolicy:policy-id:resource/ec2/spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE:policyName/cpu50-target-tracking-scaling-policy",
  "Alarms": [
    {
      "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-
b46e-434a-a60f-3b36d653feca",
      "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca"
    },
    {
      "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-
d19b-4a63-a812-6c67aaf2910d",
      "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
    }
  ]
}
```

대상 추적 조정 정책 설명

다음 [describe-scaling-policies](#) 명령을 사용하여 지정된 서비스 네임스페이스에 대한 모든 조정 정책을 설명할 수 있습니다.

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2
```

--query 파라미터를 사용하여 대상 추적 조정 정책으로 결과를 필터링할 수 있습니다. query의 구문에 대한 자세한 내용은 AWS Command Line Interface 사용 설명서의 [AWS CLI의 명령 출력 제어](#)를 참조하세요.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2 \
  --query 'ScalingPolicies[?PolicyType==`TargetTrackingScaling`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2 --query
  "ScalingPolicies[?PolicyType==`TargetTrackingScaling`]"
```

출력의 예제는 다음과 같습니다.

```
[
  {
    "PolicyARN": "PolicyARN",
    "TargetTrackingScalingPolicyConfiguration": {
      "PredefinedMetricSpecification": {
        "PredefinedMetricType": "EC2SpotFleetRequestAverageCPUUtilization"
      },
      "TargetValue": 50.0
    },
    "PolicyName": "cpu50-target-tracking-scaling-policy",
    "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
    "ServiceNamespace": "ec2",
    "PolicyType": "TargetTrackingScaling",
    "ResourceId": "spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE",
    "Alarms": [
      {
        "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca",
        "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca"
      },
      {
        "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d",

```

```

        "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
    }
  ],
  "CreationTime": 1515021724.807
}
]

```

대상 추적 조정 정책 삭제

대상 추적 조정 정책을 완료했다면 [delete-scaling-policy](#) 명령을 사용하여 이를 삭제할 수 있습니다.

다음 명령은 지정된 스폿 플릿 요청에 대해 지정된 대상 추적 조정 정책을 삭제합니다. 또한 Application Auto Scaling에서 사용자를 대신하여 생성한 CloudWatch 경보도 삭제합니다.

Linux, macOS 또는 Unix

```

aws application-autoscaling delete-scaling-policy --service-namespace ec2 \
--scalable-dimension ec2:spot-fleet-request:TargetCapacity \
--resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
--policy-name cpu50-target-tracking-scaling-policy

```

Windows

```

aws application-autoscaling delete-scaling-policy --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --policy-name cpu50-target-tracking-scaling-
policy

```

지표 수학을 사용하여 Application Auto Scaling에서 대상 추적 조정 정책 생성

메트릭 수학을 사용하면 여러 CloudWatch 메트릭을 쿼리하고 수학 식을 사용하여 이러한 메트릭을 기반으로 새 시계열을 만들 수 있습니다. CloudWatch 콘솔에서 결과 시계열을 시각화하고 대시보드에 추가할 수 있습니다. 지표 수학에 대한 자세한 내용은 Amazon [사용 CloudWatch 설명서의 지표 수학 사용](#)을 참조하십시오.

다음은 지표 수학 표현식에 적용되는 고려 사항입니다.

- 사용 가능한 모든 CloudWatch 지표를 쿼리할 수 있습니다. 각 지표는 지표 이름, 네임스페이스, 0개 이상의 측정기준으로 이루어진 고유한 조합입니다.
- 모든 산술 연산자 (+ - */^), 통계 함수 (예: AVG 또는 SUM) 또는 지원하는 기타 함수를 사용할 수 있습니다. CloudWatch
- 수학 표현식의 공식에서 지표 및 다른 수학 표현식의 결과를 모두 사용할 수 있습니다.
- 지표 규격에 사용된 표현식은 결국 단일 시계열을 반환해야 합니다.
- CloudWatch [콘솔 또는 데이터 API를 사용하여 메트릭 수학 표현식이 유효한지 확인할 수 있습니다.](#)
[CloudWatch GetMetric](#)

주제

- [예: 태스크당 Amazon SQS 대기열 백로그](#)
- [제한 사항](#)

예: 태스크당 Amazon SQS 대기열 백로그

태스크당 Amazon SQS 대기열 백로그를 계산하려면 대기열에서 검색 가능한 대략적인 메시지 수를 가져와 서비스에서 실행 중인 Amazon ECS 태스크의 수로 나눕니다. 자세한 내용은 AWS 컴퓨팅 블로그에서 [사용자 지정 지표를 사용하는 Amazon Elastic Container Service \(ECS\) Auto Scaling](#)을 참조하십시오.

표현식의 로직은 다음과 같습니다.

sum of (number of messages in the queue)/(number of tasks that are currently in the RUNNING state)

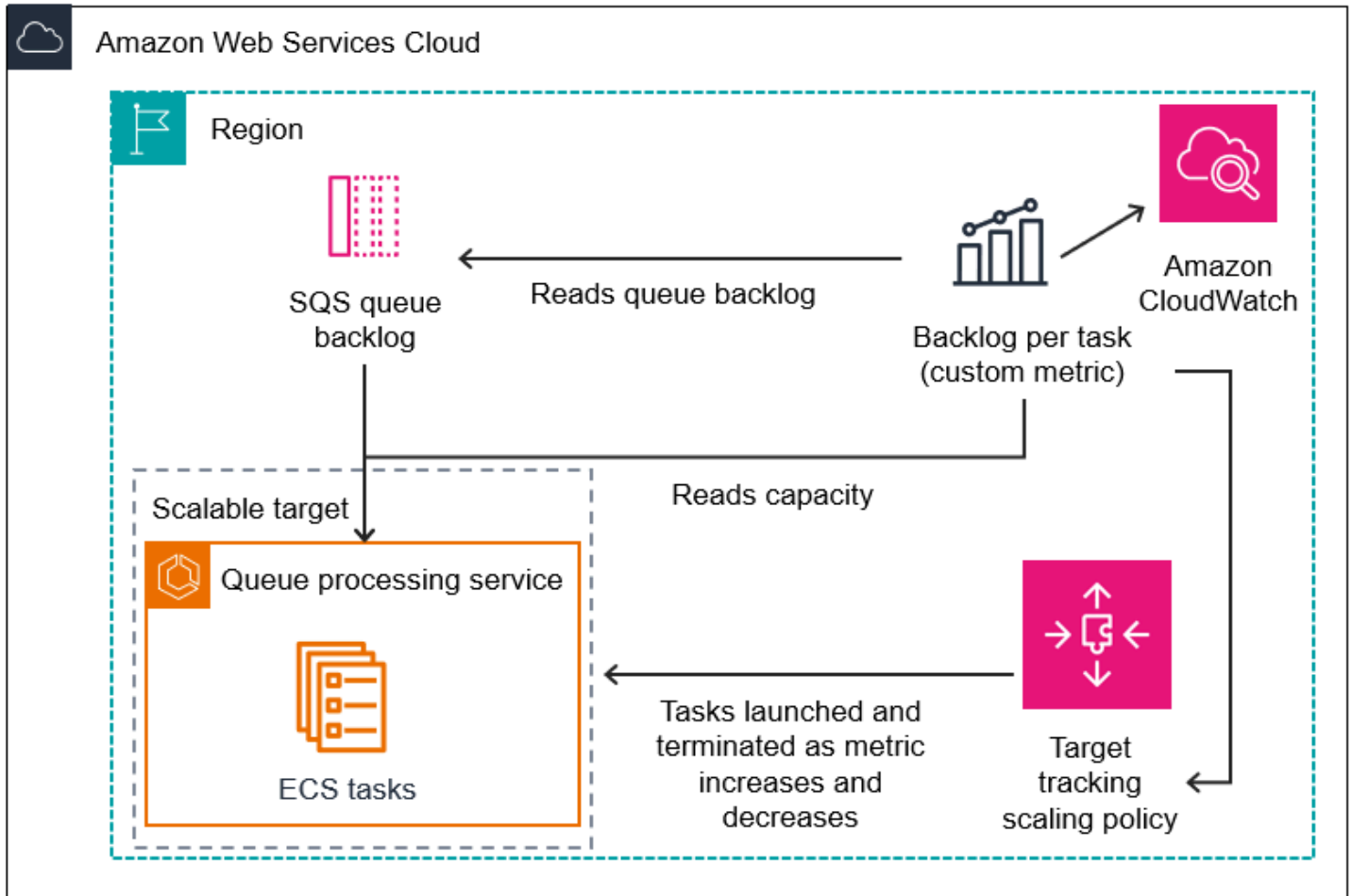
그러면 CloudWatch 측정치 정보는 다음과 같습니다.

ID	CloudWatch 지표	통계	기간
m1	ApproximateNumberOfMessages가시적	Sum	1분
m2	RunningTask개수	평균	1분

지표 수식 ID와 표현식은 다음과 같습니다.

ID	표현식
e1	(m1)/(m2)

다음 다이어그램은 이 지표의 아키텍처를 보여줍니다.



이 지표 수학을 사용하여 대상 추적 조정 정책 생성(AWS CLI)

1. 지표 수학 표현식을 맞춤된 지표 규격의 일부로서 config.json이라는 이름의 JSON 파일로 저장합니다.

다음 표가 시작하는 데 도움이 될 수 있습니다. *user input placeholder*를 사용자의 정보로 바꿉니다.

```
{
  "CustomizedMetricSpecification": {
```

```

    "Metrics": [
      {
        "Label": "Get the queue size (the number of messages waiting to be
processed)",
        "Id": "m1",
        "MetricStat": {
          "Metric": {
            "MetricName": "ApproximateNumberOfMessagesVisible",
            "Namespace": "AWS/SQS",
            "Dimensions": [
              {
                "Name": "QueueName",
                "Value": "my-queue"
              }
            ]
          },
          "Stat": "Sum"
        },
        "ReturnData": false
      },
      {
        "Label": "Get the ECS running task count (the number of currently
running tasks)",
        "Id": "m2",
        "MetricStat": {
          "Metric": {
            "MetricName": "RunningTaskCount",
            "Namespace": "ECS/ContainerInsights",
            "Dimensions": [
              {
                "Name": "ClusterName",
                "Value": "my-cluster"
              },
              {
                "Name": "ServiceName",
                "Value": "my-service"
              }
            ]
          },
          "Stat": "Average"
        },
        "ReturnData": false
      }
    ]
  }
}

```

```

        "Label": "Calculate the backlog per instance",
        "Id": "e1",
        "Expression": "m1 / m2",
        "ReturnData": true
    }
]
},
"TargetValue": 100
}

```

자세한 내용은 Application Auto Scaling API 참조의 TargetTracking ScalingPolicy [구성을](#) 참조하십시오.

Note

다음은 지표의 이름, 네임스페이스, 차원 및 통계를 찾는 CloudWatch 데 도움이 되는 몇 가지 추가 리소스입니다.

- AWS 서비스에 사용할 수 있는 지표에 대한 자세한 내용은 Amazon CloudWatch User Guide의 CloudWatch [지표를 게시하는AWS 서비스를](#) 참조하십시오.
- [가 포함된 지표의 정확한 지표 이름, 네임스페이스 및 차원 \(해당하는 경우\) 을 가져오려면 목록 CloudWatch 지표를 참조하십시오. AWS CLI](#)

2. 이 정책을 생성하려면 다음 예에 나와 있는 것처럼 JSON 파일을 입력으로 사용하여 [put-scaling-policy](#) 명령을 실행합니다.

```

aws application-autoscaling put-scaling-policy --policy-name sqs-backlog-target-tracking-scaling-policy \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service \
  --policy-type TargetTrackingScaling --target-tracking-scaling-policy-configuration file://config.json

```

성공하면 이 명령은 정책의 Amazon 리소스 이름 (ARN) 과 사용자를 대신하여 생성된 두 CloudWatch 경보의 ARN을 반환합니다.

```

{
  "PolicyARN": "arn:aws:autoscaling:us-west-2:012345678910:scalingPolicy:8784a896-b2ba-47a1-b08c-27301cc499a1:resource/ecs/service/my-cluster/my-service:policyName/sqs-backlog-target-tracking-scaling-policy",

```

```

    "Alarms": [
      {
        "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmHigh-9bc77b56-0571-4276-ba0f-d4178882e0a0",
        "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmHigh-9bc77b56-0571-4276-ba0f-d4178882e0a0"
      },
      {
        "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmLow-9b6ad934-6d37-438e-9e05-02836ddcbdc4",
        "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmLow-9b6ad934-6d37-438e-9e05-02836ddcbdc4"
      }
    ]
  }
}

```

Note

이 명령에서 오류가 발생하는 경우 AWS CLI 로컬에서 최신 버전으로 업데이트했는지 확인하십시오.

제한 사항

- 최대 요청 크기는 50KB입니다. [PutScaling정책 정의에서 메트릭 수학을 사용할 때 Policy](#) API 요청의 총 페이로드 크기입니다. 이 한도를 초과할 경우 Application Auto Scaling은 요청을 거부합니다.
- 대상 추적 조정 정책과 함께 지표 수학을 사용할 경우 다음 서비스는 지원되지 않습니다.
 - Amazon Keyspaces(Apache Cassandra용)
 - DynamoDB
 - Amazon EMR
 - Amazon MSK
 - Amazon Neptune

Application Auto Scaling의 단계 조정 정책

단계별 조정 정책은 경보를 기반으로 CloudWatch 미리 정의된 증분으로 애플리케이션 용량을 조정합니다. 경보 임계값 위반 시 스케일 아웃(용량 증가) 및 스케일 인(용량 감소)을 처리하도록 별도의 조정 정책을 정의할 수 있습니다.

단계별 조정 정책을 사용하면 조정 프로세스를 호출하는 CloudWatch 경보를 생성하고 관리할 수 있습니다. 경보 위반이 발생하면 Application Auto Scaling에서 해당 경보와 관련된 조정 정책을 시작합니다.

단계 조정 정책은 단계 조정이라고 하는 일련의 조정을 사용하여 용량을 조정합니다. 조정 크기는 경보 위반 규모에 따라 다릅니다.

- 첫 번째 임계값을 초과한 위반의 경우 Application Auto Scaling은 첫 번째 단계 조정을 적용합니다.
- 두 번째 임계값을 초과한 위반의 경우 Application Auto Scaling은 두 번째 단계 조정을 적용합니다.

이를 통해 조정 정책은 경보 지표의 사소한 변경 및 주요 변경 모두에 적절하게 대응할 수 있습니다.

정책은 규모 조정 작업 진행 중에도 추가 경보 위반에 계속 응답합니다. 즉, Application Auto Scaling은 모든 경보 위반이 발생하는 즉시 이를 평가합니다. 휴지 기간은 여러 개의 경보 위반이 빠르게 연속해서 발생하는 오버스케일링을 방지하는 데 사용됩니다.

대상 추적과 마찬가지로 단계 조정은 트래픽 변화에 따라 애플리케이션 용량을 자동으로 조정하는 데 도움이 될 수 있습니다. 하지만 꾸준한 규모 조정이 필요할 때에는 대상 추적 정책을 구현하고 관리하는 것이 더 쉬운 경향이 있습니다.

다음과 같은 확장 가능한 대상에 단계 조정 정책을 사용할 수 있습니다.

- AppStream 2.0 플릿
- Aurora DB 클러스터
- ECS 서비스
- EMR 클러스터
- SageMaker 엔드포인트 변형
- SageMaker 추론 구성 요소
- SageMaker 서버리스 프로비저닝 동시성
- Spot Fleets
- 사용자 정의 리소스

내용

- [Application Auto Scaling의 단계별 조정 작동 방식](#)
- [를 사용하여 Application Auto Scaling에 대한 단계별 조정 정책을 생성합니다. AWS CLI](#)

Application Auto Scaling의 단계별 조정 작동 방식

이 주제에서는 단계별 조정 작동 방식을 설명하고 단계별 조정 정책의 주요 요소를 소개합니다.

내용

- [작동 방식](#)
- [단계 조절](#)
- [조정 조절 타입](#)
- [휴지 기간](#)
- [조정 정책 생성, 관리 및 삭제를 위해 일반적으로 사용되는 명령](#)
- [고려 사항](#)
- [관련 리소스](#)
- [제한 사항](#)

작동 방식

스텝 스케일링을 사용하려면 확장 가능한 대상의 메트릭을 모니터링하는 CloudWatch 경보를 생성해야 합니다. 경보 위반을 결정하는 지표, 임계값, 평가 기간 수를 정의합니다. 또한 경보 임계값 위반 시 용량을 조정하는 방법을 정의하고 이를 확장 가능한 대상과 연결하는 방법을 정의하는 단계 조정 정책을 생성합니다.

정책에 단계 조정을 추가합니다. 경보의 위반 규모에 따라 다양한 단계 조정을 정의할 수 있습니다. 예:

- 경보 지표가 60%에 도달하면 10개 용량 단위로 스케일 아웃합니다.
- 경보 지표가 75%에 도달하면 30개 용량 단위로 스케일 아웃합니다.
- 경보 지표가 85%에 도달하면 40개 용량 단위로 스케일 아웃합니다.

지정된 수의 평가 기간 동안 경보 임계값이 위반되면 Application Auto Scaling은 정책에 정의된 단계 조정을 적용합니다. 경보 상태가 OK로 돌아갈 때까지 추가 경보 위반에 대해 조정을 계속할 수 있습니다.

조정 작업은 용량의 급격한 변동을 방지하기 위해 작업 중간에 휴지 기간을 두고 수행됩니다. 선택적으로 조정 정책에 대한 휴지 기간을 구성할 수 있습니다.

단계 조절

단계 조정 정책을 생성할 때 경보 위반의 크기에 따라 대상 용량이 동적으로 자동 조정되도록 하나 이상의 단계 조정을 지정합니다. 각 단계별 조정은 다음을 지정합니다.

- 지표 값의 하한값입니다.
- 지표 값의 상한값입니다.
- 조정 타입에 근거하여 축소하거나 스케일 아웃하는 양입니다.

CloudWatch 경보와 관련된 지표의 통계를 기반으로 지표 데이터 포인트를 집계합니다. CloudWatch 경보를 위반하면 적절한 조정 정책이 호출됩니다. Application Auto Scaling은 지정된 집계 유형을 원시 지표 데이터가 아닌 가장 최근의 지표 데이터 요소에 적용합니다. CloudWatch 이 집계된 지표 값을 단계별 조정으로 정의된 상한값 및 하한값과 비교하여 어느 단계의 조정을 수행할 것인지 결정합니다.

위반 임계값과 연계하여 상한값과 하한값을 지정합니다. 예를 들어 지표가 50% 를 초과하는 경우에 대한 CloudWatch 경보 및 확장 정책을 만들었다고 가정해 보겠습니다. 그런 다음 지표가 50% 미만일 때를 대비한 두 번째 경보와 스케일 인 정책을 만들었습니다. 각 정책에 대해 PercentChangeInCapacity 조정 유형을 지정하여 일련의 단계 조정을 설정했습니다.

예: 스케일 아웃 정책에 대한 단계 조절

하한	상한	조절
0	10	0
10	20	10
20	null	30

예: 축소 정책에 대한 단계 조절

하한	상한	조절
-10	0	0
-20	-10	-10

하한	상한	조정
null	-20	-30

이렇게 하면 다음과 같은 조정 구성이 생성됩니다.

Metric value								
-infinity	30%	40%	60%	70%	infinity			

-30%		-10%		Unchanged		+10%		+30%

이제 용량이 10인 확장 가능한 대상에 이 조정 구성을 사용한다고 가정해 보겠습니다. 다음 사항은 확장 가능한 대상의 용량과 관련된 조정 구성의 동작을 요약합니다.

- 원래 용량이 유지되고 집계된 지표 값은 40을 초과하고 60 미만입니다.
- 지표 값이 60에 도달하면 Application Auto Scaling에서 확장 가능한 대상의 용량이 1 증가하여 11이 됩니다. 이는 확장 정책의 두 번째 단계 조정을 기반으로 합니다(10의 10퍼센트). 새로운 용량이 추가되면 Application Auto Scaling은 현재 용량을 11로 늘립니다. 이 용량 증가 후에도 지표 값이 70으로 증가하면 Application Auto Scaling이 대상 용량을 3 증가하여 14가 됩니다. 이는 확장 정책의 세 번째 단계 조정을 기반으로 합니다(11의 30퍼센트인 3.3, 3으로 반내림).
- 지표 값이 40이 될 경우 Application Auto Scaling은 스케일 인 정책의 두 번째 단계 조절에 따라 확장 가능한 대상 용량을 1(14의 10%인 1.4, 1로 반내림) 줄여서 13으로 만듭니다. 이 용량 감소 후에도 지표 값이 30으로 떨어질 경우 Application Auto Scaling은 축소 정책의 세 번째 단계 조절에 따라 대상 용량을 3(13의 30%인 3.9, 3으로 반내림)만큼 더 줄여 10으로 만듭니다.

조정 정책에 대한 단계 조절을 지정할 때는 다음 사항에 유의합니다.

- 단계 조절의 범위는 중복되거나 격차가 있어서는 안 됩니다.
- 1단계 조절에만 null 하한값(negative infinity)이 포함될 수 있습니다. 1단계 조절에 음의 하한값이 포함될 경우, null 하한값으로 단계 조절을 해야 합니다.
- 1단계 조절에만 null 상한값(positive infinity)이 포함될 수 있습니다. 1단계 조절에 양의 상한값이 포함될 경우, null 상한값으로 단계 조절을 해야 합니다.
- 상한 및 하한값은 동일한 단계 조절에서 null이 될 수 없습니다.

- 지표 값이 위반 임계값을 초과할 경우, 하한값은 포함되고 상한값은 제외됩니다. 지표 값이 위반 임계값 미만일 경우, 하한값은 제외되고 상한값은 포함됩니다.

조정 조절 타입

선택한 조정 조절 타입에 따라 최적의 조정 작업을 수행하는 조정 정책을 정의할 수 있습니다. 조절 유형을 확장 가능 대상의 현재 용량의 백분율로 지정하거나 절대 숫자로 지정할 수 있습니다.

Application Auto Scaling은 다음과 같이 단계 조정 정책을 위한 조절 유형을 지원합니다.

- **ChangeIn용량** - 확장 가능한 대상의 현재 용량을 지정된 값만큼 늘리거나 줄입니다. 양의 값은 용량을 늘리고, 음의 값은 용량을 줄입니다. 예: 현재 용량이 3이고 조절이 5인 경우 Application Auto Scaling은 용량에 5를 추가하여 총 8로 변경합니다.
- **ExactCapacity**—확장 가능한 대상의 현재 용량을 지정된 값으로 변경합니다. 이 조절 유형에는 음이 아닌 값을 지정합니다. 예: 현재 용량이 3이고 조절이 5인 경우 Application Auto Scaling은 용량을 5로 변경합니다.
- **PercentChangeInCapacity**—확장 가능한 대상의 현재 용량을 지정된 비율만큼 늘리거나 줄입니다. 양의 값은 용량을 늘리고, 음의 값은 용량을 줄입니다. 예: 현재 용량이 10이고 조절이 10퍼센트인 경우 Application Auto Scaling은 용량에 1을 추가하여 총 11로 변경합니다.

Note

결과 값이 정수가 아닌 경우 Application Auto Scaling은 다음과 같이 반올림(반내림)합니다.

- 1보다 큰 값은 반내림합니다. 예컨대, 12.7은 12로 반내림합니다.
- 0과 1 사이의 값은 1로 반올림합니다. 예컨대, .67은 1로 반올림합니다.
- 0과 -1 사이의 값은 1로 반내림합니다. 예컨대, -.58은 -1으로 반올림합니다.
- -1보다 작은 값은 반올림합니다. 예컨대, -6.67은 -6으로 반올림합니다.

PercentChangeInCapacity를 사용하면 **MinAdjustmentMagnitude** 매개변수를 사용하여 확장할 최소 양을 지정할 수도 있습니다. 예를 들어 25%를 추가하는 정책을 생성하고 최소량으로 2를 지정한다고 가정해 보십시오. 확장 가능 대상의 용량이 4이고 조정 정책이 실행되는 경우, 4의 25%는 1이 됩니다. 그러나 최소 증분을 2로 지정했기 때문에 Application Auto Scaling은 2를 추가합니다.

휴지 기간

단계 조정 정책에서 휴지 기간을 선택적으로 정의할 수 있습니다.

이전 조정 활동이 적용될 때까지 기다리는 시간을 휴지 기간이라고 합니다.

단계별 조정 구성을 위해 휴지 기간 사용을 계획하는 두 가지 방법은 다음과 같습니다.

- 스케일 아웃 정책을 위한 휴지 기간을 사용하는 경우, 지속적이지만 과도하지는 않게 스케일 아웃하는 것이 목적입니다. Application Auto Scaling에서 조정 정책을 사용하여 성공적으로 스케일 아웃하면 휴지 기간이 계산되기 시작합니다. 조정 정책은 더 큰 스케일 아웃이 트리거되거나 휴지 기간이 종료되지 않는 한 원하는 용량을 다시 늘리지 않습니다. 확장 휴지 기간이 진행되는 동안 확장 활동을 시작하여 추가된 용량은 다음 확장 활동에 대해 원하는 용량의 일부로 계산됩니다.
- 스케일 인 정책을 위한 휴지 기간을 사용하는 경우 애플리케이션의 가용성을 보호하기 위해 보수적으로 확장하므로 스케일 인 휴지 기간이 만료될 때까지 스케일 인 활동이 차단됩니다. 그러나 축소 휴지 기간 중에 다른 경보가 확장 활동을 트리거하면 Application Auto Scaling은 대상을 즉시 확장합니다. 이 경우 스케일 인 휴지 기간이 중지되고 완료되지 않습니다.

예를 들어 트래픽 피크가 발생하면 경보가 트리거되고 Application Auto Scaling은 증가된 부하를 처리할 수 있도록 용량을 자동으로 추가합니다. 스케일 아웃 정책에 대해 휴지 기간을 설정한 경우 경보가 정책을 트리거하여 용량을 2로 늘리면 조정 활동이 성공적으로 완료되고 스케일 아웃 휴지 기간이 시작됩니다. 휴지 기간에 경보가 다시 트리거되지만 더 적극적인 단계 조정(3)에서 발생하는 경우, 이전의 증가치 2는 현재 용량의 일부로 간주됩니다. 따라서 용량은 1만 추가됩니다. 이렇게 하면 필요한 용량보다 더 추가하지 않아도 휴지 기간이 만료될 때까지 기다리는 것보다 빠르게 확장할 수 있습니다.

휴지 기간은 초 단위로 측정되며 정책 관련 조정 활동 조정에만 적용됩니다. 휴지 기간에 예약된 작업이 예약된 시간에 시작되면 휴지 기간이 만료될 때까지 기다리지 않고 조정 활동을 즉시 트리거할 수 있습니다.

값을 지정하지 않을 경우 기본값은 300입니다.

조정 정책 생성, 관리 및 삭제를 위해 일반적으로 사용되는 명령

조정 정책 작업에 일반적으로 사용되는 명령은 다음과 같습니다.

- [register-scalable-target](#)을 사용하여 리소스를 확장 가능한 대상 (Application Auto Scaling에서 확장할 수 있는 리소스) 으로 AWS 등록하고 조정을 일시 중지하고 재개합니다.
- [put-scaling-policy](#)를 사용하여 기존 확장 가능 대상에 대한 조정 정책을 추가하거나 수정할 수 있습니다.

- [describe-scaling-activities](#)를 사용하여 조정 작업에 대한 정보를 AWS 리전으로 반환할 수 있습니다.
- [describe-scaling-activities](#)를 사용하여 조정 정책에 대한 정보를 AWS 리전으로 반환할 수 있습니다.
- [delete-scaling-policy](#)를 사용하여 조정 정책을 삭제할 수 있습니다.

고려 사항

단계 조정 정책과 관련한 작업을 수행할 때는 다음 고려 사항이 적용됩니다.

- 단계 조정을 사용할 수 있을 만큼 애플리케이션의 단계 조정을 정확하게 예측할 수 있는지 생각해 보세요. 조정 지표가 스케일 아웃 가능한 대상의 용량에 비례하여 증가하거나 감소하는 경우, 대상 추적 조정 정책을 대신 사용하는 것이 좋습니다. 단계별 조정을 고급 구성에 대한 추가 정책으로 사용할 수도 있습니다. 예컨대, 사용률이 일정 수준에 도달할 때 더 공격적인 대응을 구성할 수 있습니다.
- 플래핑을 방지하려면 스케일 아웃 임계값과 스케일 인 임계값 사이에서 적절한 마진을 선택해야 합니다. 플래핑은 스케일 인과 스케일 아웃의 무한 루프입니다. 즉, 확장 작업을 진행하면 지표 값이 변경되고 반대 방향으로 다른 확장 작업이 시작됩니다.

관련 리소스

Auto Scaling의 단계 조정 정책 생성에 대한 자세한 내용은 Amazon EC2 Auto Scaling 사용 설명서의 [Amazon EC2 Auto Scaling 단계 및 단순 조정 정책](#)을 참조하세요.

제한 사항

- 확장 가능한 리소스에 대한 단계 조정 정책을 확인, 추가, 업데이트 또는 제거할 수 있는 콘솔 액세스 권한은 사용하는 리소스에 따라 다릅니다. 자세한 정보는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)을 참조하세요.

를 사용하여 Application Auto Scaling에 대한 단계별 조정 정책을 생성합니다. AWS CLI

다음 구성 작업에 를 사용하여 Application Auto AWS CLI Scaling에 대한 단계별 조정 정책을 생성할 수 있습니다.

1. 확장 가능 대상을 등록합니다.
2. 확장 가능한 대상에 단계 조정 정책을 추가합니다.

3. 정책에 대한 CloudWatch 경보를 생성합니다.

간략하게 나타내기 위해 이 주제의 예제에서는 Amazon ECS 서비스에 대한 CLI 명령을 설명합니다. 다른 확장 가능 대상을 지정하려면 `--service-namespace`에 네임스페이스, `--scalable-dimension`에 확장 가능 차원, `--resource-id`에 리소스 ID를 지정합니다. 각 서비스에 대한 자세한 내용과 예는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)의 주제를 참조하세요.

를 사용할 때는 명령이 AWS CLI프로필에 AWS 리전 맞게 구성된 상태에서 실행된다는 점을 기억하십시오. 다른 리전에서 명령을 실행하려는 경우 프로필의 기본 리전을 변경하거나 명령에 `--region` 파라미터를 사용합니다.

내용

- [확장 가능 대상 등록](#)
- [단계 조정 정책 삭제](#)
- [조정 정책을 트리거하는 경보 생성](#)
- [단계 조정 정책 설명](#)
- [단계 조정 정책 삭제](#)

확장 가능 대상 등록

아직 하지 않았다면 확장 가능 대상을 등록합니다. [register-scalable-target](#) 명령을 사용하여 대상 서비스의 특정 리소스를 확장 가능 대상으로 등록합니다. 다음 예제에서는 Application Auto Scaling을 사용하여 Amazon ECS 서비스를 등록합니다. Application Auto Scaling에서는 태스크 수를 최소 2개와 최대 10개 사이에서 조정할 수 있습니다. *user input placeholder*를 사용자의 정보로 바꿉니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --min-capacity 2 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
--min-capacity 2 --max-capacity 10
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

단계 조정 정책 삭제

확장 가능한 대상에 대한 단계별 조정 정책을 만들려면 다음 예를 사용하면 시작하는 데 도움이 됩니다.

Scale out

스케일 아웃 (용량 증가) 을 위한 단계별 조정 정책을 만들려면

- 다음 `cat` 명령을 사용하여 홈 `config.json` 디렉터리에 이름이 지정된 JSON 파일에 단계 조정 정책 구성을 저장합니다. 다음은 다음 단계 조정 (CloudWatch 경보 임계값 70으로 가정) 에 따라 확장 가능한 대상의 용량을 늘리는 조정 유형의 `PercentChangeInCapacity` 예제 구성입니다.
 - 지표 값이 70보다 크거나 같지만 85 미만인 경우 용량을 10% 늘립니다.
 - 지표 값이 85보다 크거나 같지만 95% 미만인 경우 용량을 20% 늘립니다.
 - 지표 값이 95보다 크거나 같으면 용량을 30% 늘립니다.

```
$ cat ~/config.json
{
  "AdjustmentType": "PercentChangeInCapacity",
  "MetricAggregationType": "Average",
  "Cooldown": 60,
  "MinAdjustmentMagnitude": 1,
  "StepAdjustments": [
    {
      "MetricIntervalLowerBound": 0.0,
      "MetricIntervalUpperBound": 15.0,
      "ScalingAdjustment": 10
    },
    {
      "MetricIntervalLowerBound": 15.0,
      "MetricIntervalUpperBound": 25.0,
```

```

    "ScalingAdjustment": 20
  },
  {
    "MetricIntervalLowerBound": 25.0,
    "ScalingAdjustment": 30
  }
]
}

```

자세한 내용은 Application Auto Scaling API 레퍼런스를 참조하십시오
[StepScalingPolicyConfiguration](#).

2. 다음 [put-scaling-policy](#) 명령을 생성한 config.json 파일과 함께 사용하여 my-step-scaling-policy라는 조정 정책을 생성합니다.

Linux, macOS 또는 Unix

```

aws application-autoscaling put-scaling-policy --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --policy-name my-step-scaling-policy --policy-type StepScaling \
  --step-scaling-policy-configuration file://config.json

```

Windows

```

aws application-autoscaling put-scaling-policy --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-
service --policy-name my-step-scaling-policy --policy-type StepScaling --step-
scaling-policy-configuration file://config.json

```

출력에는 해당 정책의 고유 이름 역할을 하는 ARN이 포함됩니다. 정책에 대한 CloudWatch 경 보를 만들려면 이 정보가 필요합니다.

```

{
  "PolicyARN":
  "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-
a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-step-
scaling-policy"
}

```

Scale in

스케일 인 (용량 감소) 을 위한 단계별 조정 정책을 만들려면

1. 다음 `cat` 명령을 사용하여 `config.json` 디렉터리에 이름이 지정된 JSON 파일에 단계 조정 정책 구성을 저장합니다. 다음은 다음 단계 조정 (CloudWatch 경보 임계값이 `ChangeInCapacity 50`이라고 가정) 에 따라 확장 가능한 대상의 용량을 줄이는 조정 유형의 예제 구성입니다.
 - 지표 값이 50보다 작거나 같지만 40보다 크면 용량을 1씩 줄이십시오.
 - 지표 값이 40보다 작거나 같지만 30보다 크면 용량을 2만큼 줄이십시오.
 - 지표 값이 30보다 작거나 같으면 용량을 3만큼 줄이십시오.

```
$ cat ~/config.json
{
  "AdjustmentType": "ChangeInCapacity",
  "MetricAggregationType": "Average",
  "Cooldown": 60,
  "StepAdjustments": [
    {
      "MetricIntervalUpperBound": 0.0,
      "MetricIntervalLowerBound": -10.0,
      "ScalingAdjustment": -1
    },
    {
      "MetricIntervalUpperBound": -10.0,
      "MetricIntervalLowerBound": -20.0,
      "ScalingAdjustment": -2
    },
    {
      "MetricIntervalUpperBound": -20.0,
      "ScalingAdjustment": -3
    }
  ]
}
```

자세한 내용은 Application Auto Scaling API 레퍼런스를 참조하십시오

[StepScalingPolicyConfiguration](#).

2. 다음 `put-scaling-policy` 명령을 생성한 `config.json` 파일과 함께 사용하여 `my-step-scaling-policy`라는 조정 정책을 생성합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scaling-policy --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --policy-name my-step-scaling-policy --policy-type StepScaling \
  --step-scaling-policy-configuration file://config.json
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-
service --policy-name my-step-scaling-policy --policy-type StepScaling --step-
scaling-policy-configuration file://config.json
```

출력에는 해당 정책의 고유 이름 역할을 하는 ARN이 포함됩니다. 정책에 대한 CloudWatch 경보를 만들려면 이 정보가 필요합니다.

```
{
  "PolicyARN":
  "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-
a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-step-
scaling-policy"
}
```

조정 정책을 트리거하는 경보 생성

마지막으로 다음 CloudWatch [put-metric-alarm](#) 명령을 사용하여 단계별 조정 정책에 사용할 경보를 생성합니다. 이 예제에서는 평균 CPU 사용률을 기반으로 하는 경보를 사용합니다. 이 경보는 최소 2회 이상 연속되는 60초 평가 기간에 70%의 임계값에 도달하면 ALARM 상태가 되도록 구성합니다. 다른 CloudWatch 지표를 지정하거나 자체 사용자 지정 지표를 사용하려면 이름을 `에` 지정하고 네임스페이스를 `in`에 `--metric-name` 지정하십시오. `--namespace`

Linux, macOS 또는 Unix

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-ECS:service/my-
cluster/my-service \
  --metric-name CPUUtilization --namespace AWS/ECS --statistic Average \
  --period 60 --evaluation-periods 2 --threshold 70 \
```



```
--comparison-operator GreaterThanOrEqualToThreshold \
--dimensions Name=ClusterName,Value=default Name=ServiceName,Value=sample-app-service
\
--alarm-actions PolicyARN
```

Windows

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-ECS:service/
my-cluster/my-service --metric-name CPUUtilization --namespace AWS/ECS --statistic
Average --period 60 --evaluation-periods 2 --threshold 70 --comparison-operator
GreaterThanOrEqualToThreshold --dimensions Name=ClusterName,Value=default
Name=ServiceName,Value=sample-app-service --alarm-actions PolicyARN
```

단계 조정 정책 설명

다음 [describe-scaling-policies](#) 명령을 사용하여 지정된 서비스 네임스페이스에 대한 모든 조정 정책을 설명할 수 있습니다.

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs
```

--query 파라미터를 사용하여 단계 조정 정책으로 결과를 필터링할 수 있습니다. query의 구문에 대한 자세한 내용은 AWS Command Line Interface 사용 설명서의 [AWS CLI의 명령 출력 제어](#)를 참조하세요.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs \
--query 'ScalingPolicies[?PolicyType==`StepScaling`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs --query
"ScalingPolicies[?PolicyType==`StepScaling`]"
```

출력의 예제는 다음과 같습니다.

```
[
  {
    "PolicyARN": "PolicyARN",
    "StepScalingPolicyConfiguration": {
```

```

    "MetricAggregationType": "Average",
    "Cooldown": 60,
    "StepAdjustments": [
      {
        "MetricIntervalLowerBound": 0.0,
        "MetricIntervalUpperBound": 15.0,
        "ScalingAdjustment": 1
      },
      {
        "MetricIntervalLowerBound": 15.0,
        "MetricIntervalUpperBound": 25.0,
        "ScalingAdjustment": 2
      },
      {
        "MetricIntervalLowerBound": 25.0,
        "ScalingAdjustment": 3
      }
    ],
    "AdjustmentType": "ChangeInCapacity"
  },
  "PolicyType": "StepScaling",
  "ResourceId": "service/my-cluster/my-service",
  "ServiceNamespace": "ecs",
  "Alarms": [
    {
      "AlarmName": "Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-
service",
      "AlarmARN": "arn:aws:cloudwatch:region:012345678910:alarm:Step-Scaling-
AlarmHigh-ECS:service/my-cluster/my-service"
    }
  ],
  "PolicyName": "my-step-scaling-policy",
  "ScalableDimension": "ecs:service:DesiredCount",
  "CreationTime": 1515024099.901
}
]

```

단계 조정 정책 삭제

더 이상 필요 없는 단계적 조정 정책은 삭제할 수 있습니다. 조정 정책과 CloudWatch 경보를 모두 삭제하려면 다음 작업을 완료하십시오.

조정 정책을 삭제하려면

다음 [delete-scaling-policy](#) 명령을 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling delete-scaling-policy --service-namespace ecs \  
--scalable-dimension ecs:service:DesiredCount \  
--resource-id service/my-cluster/my-service \  
--policy-name my-step-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace ecs --scalable-  
dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service --  
policy-name my-step-scaling-policy
```

CloudWatch 경보를 삭제하려면

[delete-alarms](#) 명령을 사용합니다. 한 번에 하나 이상의 경보를 삭제할 수 있습니다. 예컨대, 다음 명령을 사용하여 Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-service 및 Step-Scaling-AlarmLow-ECS:service/my-cluster/my-service 경보를 삭제합니다.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-ECS:service/my-  
cluster/my-service Step-Scaling-AlarmLow-ECS:service/my-cluster/my-service
```

자습서: 과중한 워크로드를 처리하도록 Auto Scaling 구성

Important

이 자습서를 살펴보기 전에 먼저 [자습서: AWS CLI을\(를\) 사용하여 예약된 조정 시작하기](#) 입문 자습서를 검토하는 것이 좋습니다.

이 자습서에서는 애플리케이션이 보통의 워크로드보다 과부하 상황일 때 시간 범위를 기준으로 확장 및 축소하는 방법에 대해 알아봅니다. 이 기능은 정기적으로 또는 계절에 따라 갑자기 많은 수의 방문자가 발생할 수 있는 애플리케이션이 있는 경우에 유용합니다.

예약된 조정과 함께 대상 추적 조정 정책을 사용하여 추가 로드를 처리할 수 있습니다. 예약된 조정은 지정한 일정에 따라 사용자를 대신하여 자동으로 MinCapacity 및 MaxCapacity를 변경합니다. 대상 추적 조정 정책이 리소스에서 활성화되면 새로운 최소 및 최대 용량 범위 내에서 현재 리소스 사용률에 따라 동적으로 확장될 수 있습니다.

이 자습서를 완료하면 다음을 수행하는 방법을 알 수 있습니다.

- 예약된 조정을 사용하여 과부하가 발생하기 전에 용량을 추가한 다음 더 이상 필요하지 않을 때 추가 용량을 제거합니다.
- 대상 추적 조정 정책을 사용하여 현재 리소스 사용률에 따라 애플리케이션을 조정합니다.

내용

- [필수 조건](#)
- [1단계: 확장 가능 대상 등록](#)
- [2단계: 요구 사항에 따라 예약된 작업 설정](#)
- [2단계: 대상 추적 조정 정책 추가](#)
- [4단계: 다음 단계](#)
- [5단계: 정리](#)

필수 조건

이 자습서에서는 다음을 이미 완료했다고 가정합니다.

- 를 만들었습니다 AWS 계정.
- 설치 및 구성했습니다 AWS CLI.
- Application Auto Scaling을 통해 리소스를 확장 가능한 타겟으로 등록 및 등록 취소하는 데 필요한 권한을 부여했습니다. 또한 조정 정책 및 예약된 작업을 생성하는 데 필요한 권한을 부여했습니다. 자세한 정보는 [Application Auto Scaling에 사용되는 Identity and Access Management](#)을 참조하세요.
- 이 자습서에 사용할 수 있는 비프로덕션 환경에서 지원되는 리소스를 만들었습니다. 아직 없는 경우 지금 하나 만듭니다. Application Auto Scaling과 함께 작동하는 AWS 서비스 및 리소스에 대한 자세한 내용은 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#) 섹션을 참조하세요.

Note

이 자습서를 완료하는 동안 리소스의 최소 및 최대 용량 값을 0으로 설정하여 현재 용량을 0으로 재설정하는 두 가지 단계가 있습니다. Application Auto Scaling과 함께 사용 중인 리소스에 따라 이 단계에서 현재 용량을 0으로 재설정하지 못할 수 있습니다. 문제를 해결하는 데 도움이 되도록 출력에는 최소 용량이 지정된 값보다 작을 수 없다는 메시지가 표시되며 AWS 리소스가 수용할 수 있는 최소 용량 값을 제공합니다.

1단계: 확장 가능 대상 등록

Application Auto Scaling에서 확장 가능 대상으로서 리소스를 등록하는 것부터 시작합니다. 확장 가능 대상은 Application Auto Scaling에서 확장하거나 축소할 수 있는 리소스입니다.

Application Auto Scaling을 통해 확장 가능 대상을 등록하려면

- 다음 [register-scalable-target](#) 명령을 사용하여 새로운 확장 가능 대상을 등록합니다. `--min-capacity` 및 `--max-capacity` 값을 0으로 설정하여 현재 용량을 0으로 재설정합니다.

`--service-namespace`의 샘플 텍스트를 Application Auto Scaling과 함께 사용 중인 AWS 서비스의 네임스페이스로, `--scalable-dimension`을 등록하려는 리소스와 연결된 확장 가능한 차원으로, `--resource-id`를 리소스의 식별자로 바꿉니다. 이러한 값은 사용되는 리소스와 리소스 ID가 구성되는 방식에 따라 다릅니다. 자세한 내용은 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#) 섹션의 주제를 참조하세요. 이러한 주제에는 Application Auto Scaling에 확장 가능 대상을 등록하는 방법을 보여주는 예제 명령이 포함되어 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 0 --max-capacity 0
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace namespace \
  --scalable-dimension dimension --resource-id identifier --min-capacity 0 --max-
  capacity 0
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-
id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

2단계: 요구 사항에 따라 예약된 작업 설정

[put-scheduled-action](#) 명령을 사용하여 비즈니스 요구에 맞게 구성된 예약된 작업을 만들 수 있습니다. 이 자습서에서는 용량을 0으로 줄임으로써 근무 시간 외에 리소스 소비를 중지하는 구성에 중점을 둡니다.

아침에 확장하는 예약된 작업을 만들려면

1. 확장 가능한 대상을 확장하려면 다음 [put-scheduled-action](#) 명령을 사용합니다. cron 표현식을 사용한 반복되는 일정(UTC 기준)의 `--schedule` 파라미터를 포함합니다.

지정된 일정(매일 오전 9시(UTC 기준))에 따라 Application Auto Scaling은 MinCapacity 및 MaxCapacity 값을 1~5 용량 단위의 원하는 범위로 업데이트합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action \
  --service-namespace namespace \
  --scalable-dimension dimension \
```

```
--resource-id identifier \  
--scheduled-action-name my-first-scheduled-action \  
--schedule "cron(0 9 * * ? *)" \  
--scalable-target-action MinCapacity=1,MaxCapacity=5
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace namespace --  
scalable-dimension dimension --resource-id identifier --scheduled-action-name my-  
first-scheduled-action --schedule "cron(0 9 * * ? *)" --scalable-target-action  
MinCapacity=1,MaxCapacity=5
```

이 명령이 제대로 실행되면 어떤 출력도 반환하지 않습니다.

2. 예약된 작업이 있는지 확인하려면 다음 [describe-scheduled-actions](#)을 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scheduled-actions \  
--service-namespace namespace \  
--query 'ScheduledActions[?ResourceId==`identifier`]'
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-  
namespace namespace --query "ScheduledActions[?ResourceId==`identifier`]"
```

출력의 예제는 다음과 같습니다.

```
[  
  {  
    "ScheduledActionName": "my-first-scheduled-action",  
    "ScheduledActionARN": "arn",  
    "Schedule": "cron(0 9 * * ? *)",  
    "ScalableTargetAction": {  
      "MinCapacity": 1,  
      "MaxCapacity": 5  
    },  
    ...  
  }  
]
```

밤에 축소하는 예약된 작업을 만들려면

1. 앞의 절차를 반복하여 하루가 끝날 때 Application Auto Scaling에서 축소하는 데 사용하는 또 다른 예약된 작업을 생성합니다.

다음 [put-scheduled-action](#) 명령에서 지시한 바와 같이, 지정된 스케줄에 따라(매일 오후 8:00(UTC 기준)) Application Auto Scaling이 대상의 MinCapacity 및 MaxCapacity를 0으로 업데이트합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scheduled-action \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --scheduled-action-name my-second-scheduled-action \
  --schedule "cron(0 20 * * ? *)" \
  --scalable-target-action MinCapacity=0,MaxCapacity=0
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace namespace --scalable-dimension dimension --resource-id identifier --scheduled-action-name my-second-scheduled-action --schedule "cron(0 20 * * ? *)" --scalable-target-action MinCapacity=0,MaxCapacity=0
```

2. 예약된 작업이 있는지 확인하려면 다음 [describe-scheduled-actions](#)을 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scheduled-actions \
  --service-namespace namespace \
  --query 'ScheduledActions[?ResourceId==`identifier`]'
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace namespace --query "ScheduledActions[?ResourceId==`identifier`]"
```

출력의 예제는 다음과 같습니다.


```
[
  {
    "ScheduledActionName": "my-first-scheduled-action",
    "ScheduledActionARN": "arn",
    "Schedule": "cron(0 9 * * ? *)",
    "ScalableTargetAction": {
      "MinCapacity": 1,
      "MaxCapacity": 5
    },
    ...
  },
  {
    "ScheduledActionName": "my-second-scheduled-action",
    "ScheduledActionARN": "arn",
    "Schedule": "cron(0 20 * * ? *)",
    "ScalableTargetAction": {
      "MinCapacity": 0,
      "MaxCapacity": 0
    },
    ...
  }
]
```

2단계: 대상 추적 조정 정책 추가

기본 일정을 준비했으므로 현재 리소스 사용률에 따라 확장할 대상 추적 조정 정책을 추가합니다.

대상 추적을 사용하면 Application Auto Scaling이 정책의 대상 값을 지정된 지표의 현재 값과 비교합니다. 값이 일정 기간에 동일하지 않은 경우 Application Auto Scaling은 안정적인 성능을 유지하기 위해 용량을 추가하거나 제거합니다. 애플리케이션의 로드와 지표 값이 증가함에 따라 Application Auto Scaling은 MaxCapacity를 초과하지 않고 가능한 한 빨리 용량을 추가합니다. 로드가 최소이기 때문에 Application Auto Scaling이 용량을 제거하는 경우 MinCapacity 이하로 떨어지지 않고 용량을 제거할 수 있습니다. 사용량에 따라 용량을 조정하여 애플리케이션에 필요한 만큼만 요금을 지불합니다.

애플리케이션에 로드가 없기 때문에 지표에 데이터가 충분하지 않은 경우 Application Auto Scaling은 용량을 추가하거나 제거하지 않습니다. 즉, Application Auto Scaling은 정보가 충분하지 않은 상황에서 가용성을 우선시합니다.

여러 조정 정책을 추가할 수 있지만 충돌하는 단계 조정 정책을 추가하지 않아야 합니다. 이로 인해 바람직하지 않은 동작이 발생할 수 있습니다. 예컨대, 대상 추적 정책이 축소 준비되기 전에 단계별 조정

정책이 축소 활동을 시작하는 경우, 축소 활동이 차단되지 않습니다. 축소 작업이 완료된 후 대상 추적 정책이 Application Auto Scaling에 다시 확장하도록 지시할 수 있습니다.

대상 추적 조정 정책을 생성하려면

1. [put-scaling-policy](#) 명령을 사용하여 정책을 생성합니다.

대상 추적에 가장 자주 사용되는 지표는 사전 정의되어 있으므로 에서 CloudWatch 전체 지표 사양을 제공하지 않고도 사용할 수 있습니다. 사용 가능한 미리 정의된 지표에 대한 자세한 내용은 [Application Auto Scaling의 대상 추적 조정 정책](#) 섹션을 참조하세요.

이 명령을 실행하기 전에 미리 정의된 지표가 대상 값을 기대하는지 확인합니다. 예를 들어 CPU 사용률이 50%에 도달하면 확장하도록 대상 값을 50.0으로 지정합니다. 또는 사용률이 70%에 도달했을 때 Lambda 프로비저닝된 동시성을 확장하려면 대상 값을 0.7로 지정합니다. 특정 리소스의 대상 값에 대한 자세한 내용은 대상 추적을 구성하는 방법에 대해 서비스에서 제공하는 설명서를 참조하세요. 자세한 내용은 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#) 섹션을 참조하세요.

Linux, macOS 또는 Unix

```
aws application-autoscaling put-scaling-policy \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration '{ "TargetValue": 50.0,
  "PredefinedMetricSpecification": { "PredefinedMetricType": "predefinedmetric" } }'
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --policy-name my-scaling-
policy --policy-type TargetTrackingScaling --target-tracking-scaling-policy-
configuration "{ \"TargetValue\": 50.0, \"PredefinedMetricSpecification\":
{ \"PredefinedMetricType\": \"predefinedmetric\" } }"
```

성공하면 이 명령은 사용자 대신 생성된 두 CloudWatch 경보의 ARN과 이름을 반환합니다.

2. 예약된 작업이 있는지 확인하려면 다음 [describe-scaling-policies](#) 명령을 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace namespace \
  --query 'ScalingPolicies[?ResourceId==`identifier`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace namespace \
  --query "ScalingPolicies[?ResourceId==`identifier`]"
```

출력의 예제는 다음과 같습니다.

```
[
  {
    "PolicyARN": "arn",
    "TargetTrackingScalingPolicyConfiguration": {
      "PredefinedMetricSpecification": {
        "PredefinedMetricType": "predefinedmetric"
      },
      "TargetValue": 50.0
    },
    "PolicyName": "my-scaling-policy",
    "PolicyType": "TargetTrackingScaling",
    "Alarms": [],
    ...
  }
]
```

4단계: 다음 단계

크기 조정 활동이 발생하면 확장 가능 대상에 대한 조정 활동의 출력에서 이에 대한 레코드를 볼 수 있습니다. 예를 들면 다음과 같습니다.

```
Successfully set desired count to 1. Change successfully fulfilled by ecs.
```

Application Auto Scaling을 사용하여 크기 조정 활동을 모니터링하려면 다음 [describe-scaling-activities](#) 명령을 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-activities
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace namespace
  --scalable-dimension dimension --resource-id identifier
```

5단계: 정리

적극적으로 조정하는 동안 생성된 자원에 대한 요금이 계정에서 발생하지 않도록 하려면 다음과 같이 연결된 조정 구성을 정리할 수 있습니다.

규모 조정 구성을 삭제해도 기본 AWS 리소스는 삭제되지 않습니다. 또한 원래 용량으로 반환하지 않습니다. 리소스를 생성한 서비스 콘솔을 사용하여 리소스를 삭제하거나 리소스 용량을 조정할 수 있습니다.

예약된 작업을 삭제하려면

다음 [delete-scheduled-action](#) 명령은 지정된 예약된 작업을 삭제합니다. 생성한 예약된 작업을 나중에도 계속 사용하고 싶은 경우에는 이 단계를 건너뛸 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling delete-scheduled-action \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --scheduled-action-name my-second-scheduled-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace namespace
  --scalable-dimension dimension --resource-id identifier --scheduled-action-name my-second-scheduled-action
```

조정 정책을 삭제하려면

다음 [delete-scaling-policy](#) 명령은 지정된 대상 추적 조정 정책을 삭제합니다. 생성한 조정 정책을 나중에도 계속 사용하고 싶은 경우에는 이 단계를 건너뛸 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling delete-scaling-policy \  
  --service-namespace namespace \  
  --scalable-dimension dimension \  
  --resource-id identifier \  
  --policy-name my-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace namespace --  
scalable-dimension dimension --resource-id identifier --policy-name my-scaling-policy
```

확장 가능 대상의 등록을 취소하려면

다음 [deregister-scalable-target](#) 명령을 사용하여 확장 가능 대상의 등록을 취소합니다. 이미 생성한 조정 정책이 있거나 아직 삭제되지 않은 예약된 작업이 있는 경우에는 이 명령을 통해 삭제됩니다. 확장 가능 대상을 나중에도 사용할 수 있도록 등록 상태로 유지하고 싶은 경우에는 이 단계를 건너뛸 수 있습니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling deregister-scalable-target \  
  --service-namespace namespace \  
  --scalable-dimension dimension \  
  --resource-id identifier
```

Windows

```
aws application-autoscaling deregister-scalable-target --service-namespace namespace --  
scalable-dimension dimension --resource-id identifier
```

Application Auto Scaling의 조정 일시 중지 및 재개

이 주제에서는 애플리케이션의 조정 가능 대상에 대한 조정 활동을 하나 이상 일시 중지한 후 재개하는 방법을 설명합니다. 이러한 일시 중지-재개 기능은 조정 정책 및 예약 작업에 의해 트리거된 조정 활동을 일시적으로 중지하는 데 사용됩니다. 예를 들어 변경을 수행하거나 구성 문제를 조사하는 동안 자동 조정 작업이 방해받지 않도록 하려는 경우에 유용합니다. 조정 정책 및 예약 작업을 보관해 두었다가, 준비되었을 때 조정 활동을 재개할 수 있습니다.

다음 예제 CLI 명령에서는 config.json 파일에 다음과 같은 JSON 형식 파라미터를 전달합니다. JSON 데이터 구조를 묶을 따옴표를 사용하여 명령줄에 이러한 파라미터를 전달할 수도 있습니다. 자세한 내용은 AWS Command Line Interface 사용 설명서의 [AWS CLI에서 문자열에 따옴표 사용](#)을 참조하세요.

내용

- [조정 활동](#)
- [조정 활동을 일시 중지하고 재개합니다.](#)

Note

Amazon ECS 배포가 진행 중인 동안 스케일 아웃 프로세스를 일시 중단하기 위한 지침은 다음 설명서를 참조하십시오.

Amazon Elastic 컨테이너 서비스 개발자 안내서의 서비스 [자동 조정 및 배포](#)

조정 활동

Application Auto Scaling은 다음과 같은 조정 활동을 일시 중지된 상태로 유지하도록 지원합니다.

- 조정 정책에 의해 트리거되는 모든 축소 활동.
- 조정 정책에 의해 트리거되는 모든 확장 활동.
- 예약된 작업을 포함하는 모든 조정 활동.

다음은 개별 조정 활동이 일시 중지될 때 어떤 일이 발생하는지에 대해 설명합니다. 각 조정 활동이 개별적으로 일시 중지되고 재개될 수 있습니다. 조정 활동을 일시 중지하는 이유에 따라 여러 조정 활동을 함께 일시 중지해야 할 수도 있습니다.

DynamicScalingInSuspended

- Application Auto Scaling은 대상 추적 조정 정책 또는 단계 조정 정책이 트리거될 때 용량을 제거하지 않습니다. 이렇게 하면 규모 조정 정책이나 관련 경보를 삭제하지 않고도 규모 조정 정책과 관련된 축소 활동을 일시적으로 비활성화할 수 있습니다. CloudWatch 축소를 재개하면 Application Auto Scaling이 현재 위반된 경보 임계값으로 정책을 평가합니다.

DynamicScalingOutSuspended

- Application Auto Scaling은 대상 추적 조정 정책 또는 단계 조정 정책이 트리거될 때 용량을 추가하지 않습니다. 이렇게 하면 조정 정책이나 관련 경보를 삭제하지 않고도 조정 정책과 관련된 확장 활동을 일시적으로 비활성화할 수 있습니다. CloudWatch 확장을 재개하면 Application Auto Scaling이 현재 위반된 경보 임계값으로 정책을 평가합니다.

ScheduledScalingSuspended

- Application Auto Scaling은 일시 중지 기간에 실행되도록 예약된 조정 작업을 시작하지 않습니다. 예약된 조정을 재개하면 Application Auto Scaling은 실행 시간이 아직 경과하지 않은 예약된 작업만 평가합니다.

조정 활동을 일시 중지하고 재개합니다.

Application Auto Scaling 확장 가능 대상에 대한 개별 조정 활동 또는 모든 조정 활동을 일시 중지 및 재개할 수 있습니다.

Note

간략하게 나타내기 위해 이 예에서는 DynamoDB 테이블의 조정을 일시 중지 및 재개하는 방법을 보여줍니다. 다른 확장 가능 대상을 지정하려면 `--service-namespace`에 네임스페이스, `--scalable-dimension`에 확장 가능 차원, `--resource-id`에 리소스 ID를 지정합니다. 각 서비스에 대한 자세한 내용과 예는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)의 주제를 참조하세요.

조정 활동을 일시 중지하려면

다음과 같이 명령줄 창을 열고 [register-scalable-target](#) 명령을 `--suspended-state` 옵션과 함께 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

조정 정책에 의해 트리거되는 축소 활동만 일시 중지하려면 다음을 config.json에 지정합니다.

```
{
  "DynamicScalingInSuspended":true
}
```

조정 정책에 의해 트리거되는 확장 활동만 일시 중지하려면 다음을 config.json에 지정합니다.

```
{
  "DynamicScalingOutSuspended":true
}
```

예약된 작업을 포함하는 조정 활동만 일시 중지하려면 다음을 config.json에 지정합니다.

```
{
  "ScheduledScalingSuspended":true
}
```


모든 조정 활동을 일시 중지하려면

다음과 같이 [register-scalable-target](#) 명령을 `--suspended-state` 옵션과 함께 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

이 예제에서는 config.json 파일에 다음과 같은 JSON 형식 파라미터가 포함된 것으로 가정합니다.

```
{
  "DynamicScalingInSuspended":true,
  "DynamicScalingOutSuspended":true,
  "ScheduledScalingSuspended":true
}
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

일시 중지된 조정 활동 보기

[describe-scalable-targets](#) 명령을 사용하여 확장 가능 대상에 대해 일시 중지된 상태인 조정 활동을 확인합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table
```

Windows

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table
```

출력의 예제는 다음과 같습니다.

```
{
  "ScalableTargets": [
    {
      "ServiceNamespace": "dynamodb",
      "ScalableDimension": "dynamodb:table:ReadCapacityUnits",
      "ResourceId": "table/my-table",
      "MinCapacity": 1,
      "MaxCapacity": 20,
      "SuspendedState": {
        "DynamicScalingOutSuspended": true,
        "DynamicScalingInSuspended": true,
        "ScheduledScalingSuspended": true
      },
      "CreationTime": 1558125758.957,
      "RoleARN": "arn:aws:iam::123456789012:role/aws-service-role/dynamodb.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_DynamoDBTable"
    }
  ]
}
```

조정 활동 재개

조정 활동을 재개할 준비가 되면 [register-scalable-target](#) 명령을 사용하여 재개할 수 있습니다.

다음과 같은 예제 명령은 지정된 확장 가능 대상에 대한 모든 조정 활동을 재개합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --suspended-state file://config.json
```

이 예제에서는 config.json 파일에 다음과 같은 JSON 형식 파라미터가 포함된 것으로 가정합니다.

```
{
  "DynamicScalingInSuspended":false,
  "DynamicScalingOutSuspended":false,
  "ScheduledScalingSuspended":false
}
```

이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Application Auto Scaling 확장 활동

Application Auto Scaling은 조정 정책의 CloudWatch 지표를 모니터링하고 임계값이 초과되면 조정 활동을 시작합니다. 또한 사용자가 확장 가능 대상의 최대 또는 최소 크기를 수정하면 수동으로 또는 일정에 따라 확장 활동을 시작합니다.

확장 활동이 발생하면 Application Auto Scaling에서는 다음 중 하나를 수행합니다.

- 확장 가능 대상의 용량 늘리기(스케일 아웃이라고 함)
- 확장 가능 대상의 용량 줄이기(스케일 인이라고 함)

최근 6주 동안의 확장 활동을 조회할 수 있습니다.

확장 가능한 대상을 기준으로 조정 활동을 찾아보세요.

특정 확장 가능 대상에 대한 확장 활동을 확인하려면 다음 [describe-scaling-activities](#) 명령을 사용합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-activities --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
```

다음은 StatusCode에 활동의 현재 상태가 있고 StatusMessage에 확장 활동의 상태가 있는 응답의 예입니다.

```
{
  "ScalingActivities": [
    {
      "ScalableDimension": "ecs:service:DesiredCount",
      "Description": "Setting desired count to 1.",
      "ResourceId": "service/my-cluster/my-service",
      "ActivityId": "e6c5f7d1-dbbb-4a3f-89b2-51f33e766399",
```

```

        "StartTime": 1462575838.171,
        "ServiceNamespace": "ecs",
        "EndTime": 1462575872.111,
        "Cause": "monitor alarm web-app-cpu-lt-25 in state ALARM triggered policy
web-app-cpu-lt-25",
        "StatusMessage": "Successfully set desired count to 1. Change successfully
fulfilled by ecs.",
        "StatusCode": "Successful"
    }
]
}

```

응답의 필드에 대한 설명은 Application Auto Scaling API 참조를 참조하십시오 [ScalingActivity](#).

다음 상태 코드는 확장 활동이 발생하는 활동 이벤트가 언제 완료됨 상태에 도달하는지를 나타냅니다.

- Successful – 확장이 완료되었음
- Overridden – 더 새로운 확장 이벤트를 통해 원하는 용량으로 업데이트되었음
- Unfulfilled – 확장 시간을 초과했거나 대상 서비스에서 요청을 이행할 수 없음
- Failed – 예외가 발생하여 확장 실패

Note

확장 활동에 Pending 또는 InProgress의 상태가 있을 수도 있습니다. 대상 서비스에서 응답하기 전에 모든 확장 활동에 Pending 상태가 있습니다. 대상에서 응답하면 확장 활동의 상태가 InProgress로 변경됩니다.

규모가 조정되지 않은 활동 포함

기본적으로 확장 활동에는 Application Auto Scaling에서 확장 여부를 결정하는 시간이 반영되지 않습니다.

주어진 지정된 지표의 최대 임계값을 Amazon ECS 서비스에서 초과하는데 작업 수가 이미 허용되는 최대 작업 수에 도달한 예를 가정하겠습니다. 이 경우에는 원하는 태스크 수를 Application Auto Scaling에서 스케일 아웃하지 않습니다.


확장되지 않은 활동(확장된 활동 아님)을 응답에 포함하려면 [describe-scaling-activities](#) 명령에 `--include-not-scaled-activities` 옵션을 추가합니다.

Linux, macOS 또는 Unix

```
aws application-autoscaling describe-scaling-activities --include-not-scaled-activities \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service
```

Windows

```
aws application-autoscaling describe-scaling-activities --include-not-scaled-activities \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-id \
  service/my-cluster/my-service
```

 Note

이 명령에서 오류가 발생하는 경우 AWS CLI 로컬에서 최신 버전으로 업데이트했는지 확인하십시오.

확장되지 않은 활동이 응답에 포함되는지 확인하도록 전부는 아니지만 일부 실패한 확장 활동에 대한 출력에 `NotScaledReasons` 요소가 표시됩니다.

```
{
  "ScalingActivities": [
    {
      "ScalableDimension": "ecs:service:DesiredCount",
      "Description": "Attempting to scale due to alarm triggered",
      "ResourceId": "service/my-cluster/my-service",
      "ActivityId": "4d759079-a31f-4d0c-8468-504c56e2eecf",
      "StartTime": 1664928867.915,
      "ServiceNamespace": "ecs",
      "Cause": "monitor alarm web-app-cpu-gt-75 in state ALARM triggered policy web-app-cpu-gt-75",
      "StatusCode": "Failed",
      "NotScaledReasons": [
        {
          "Code": "AlreadyAtMaxCapacity",
          "MaxCapacity": 4
        }
      ]
    }
  ]
}
```

```
    ]
}
```

응답의 필드에 대한 설명은 Application Auto Scaling API 참조를 참조하십시오 [ScalingActivity](#).

확장되지 않은 활동이 반환되는 경우 Code에 나열된 사유 코드에 따라 CurrentCapacity, MaxCapacity, MinCapacity 등의 속성이 응답에 표시될 수도 있습니다.

대량의 중복 항목을 방지하기 위해 스케일링되지 않은 첫 번째 활동만 스케일링 활동 기록에 기록됩니다. 규모를 조정하지 않은 후속 활동은 규모를 조정하지 않는 한 새 항목을 생성하지 않습니다. 단, 규모를 조정하지 않는 한 새 항목이 생성되지 않습니다.

사유 코드

다음은 확장되지 않은 활동의 사유 코드입니다.

사유 코드	정의			
AutoScalingAnticipatedFlapping	플래핑이 발생하기 때문에 자동 확장 알고리즘에서 확장 작업을 진행하지 않기로 결정했습니다. 플래핑은 스케일 인과 스케일 아웃의 무한 루프입니다. 즉, 확장 작업을 진행하면 지표 값이 변경되어 반대 방향으로 다른 확장 작업이 시작됩니다.			
TargetServicePutResourceAsInscalable	대상 서비스에서 일시적으로 리소스를 확장할 수 없는 상태로 전환했습니다. 확장 정책에 구성된			

사유 코드	정의			
	자동 확장 조건이 충족되면 Application Auto Scaling에서 다시 시도합니다.			
AlreadyAtMaxCapacity	사용자가 지정한 최대 용량에 따라 확장이 차단되었습니다. Application Auto Scaling에서 스케일 아웃하려면 최대 용량을 늘려야 합니다.			
AlreadyAtMinCapacity	사용자가 지정한 최소 용량에 따라 확장이 차단되었습니다. Application Auto Scaling에서 스케일 인하려면 최소 용량을 줄여야 합니다.			
AlreadyAtDesiredCapacity	수정한 용량이 현재 용량이 동일하다고 자동 확장 알고리즘에서 계산했습니다.			

Application Auto Scaling 모니터링

모니터링은 Application Auto Scaling 및 기타 AWS 솔루션의 안정성, 가용성 및 성능을 유지하는 데 있어 중요한 부분입니다. 다중 지점 장애가 발생할 경우 이를 보다 쉽게 디버깅할 수 있도록 AWS 솔루션의 모든 부분에서 모니터링 데이터를 수집해야 합니다. AWS Application Auto Scaling을 관찰하고, 문제가 발생할 경우 보고하고, 적절한 경우 자동 조치를 취할 수 있는 모니터링 도구를 제공합니다.

다음 기능을 사용하여 AWS 리소스를 관리할 수 있습니다.

AWS CloudTrail

를 사용하면 사용자 또는 사용자를 대신하여 Application Auto Scaling API에 대한 호출을 추적할 수 있습니다 AWS 계정. AWS CloudTrail CloudTrail 지정한 Amazon S3 버킷의 로그 파일에 정보를 저장합니다. 어떤 사용자 및 계정이 Application Auto Scaling을 호출했는지, 어떤 소스 IP 주소에 호출이 이루어졌는지, 언제 호출이 발생했는지 확인할 수 있습니다. 자세한 정보는 [를 사용하여 애플리케이션 Auto Scaling API 호출을 기록합니다. AWS CloudTrail](#)을 참조하세요.

Note

워크로드에 대한 데이터를 기록하고 수집하는 데 도움이 되는 다른 AWS 서비스에 대한 자세한 내용은 AWS 규범 지침의 [애플리케이션 소유자를 위한 로깅 및 모니터링 안내서](#)를 참조하십시오.

아마존 CloudWatch

CloudWatch Amazon은 로그를 분석하고 AWS 리소스 및 호스팅된 애플리케이션의 지표를 실시간으로 모니터링할 수 있도록 지원합니다. 지표를 수집 및 추적하고, 맞춤 대시보드를 생성할 수 있으며, 지정된 지표가 지정한 임계값에 도달하면 사용자에게 알리거나 조치를 취하도록 경보를 설정할 수 있습니다. 예를 들어 리소스 사용률을 CloudWatch 추적하여 사용률이 매우 높거나 지표의 경보가 INSUFFICIENT_DATA 상태에 들어갔을 때 알림을 받을 수 있습니다. 자세한 정보는 [를 사용하여 확장 가능한 리소스의 사용량을 모니터링합니다. CloudWatch](#)을 참조하세요.

CloudWatch 또한 Application Auto Scaling에 대한 AWS API 사용 지표도 추적합니다. 이러한 지표를 사용하여 API 호출량이 정의한 임계값을 위반할 때 경고하는 경보를 구성할 수 있습니다. 자세한 내용은 Amazon [AWS 사용 CloudWatch 설명서의 사용량 지표](#)를 참조하십시오.

아마존 EventBridge

EventBridge Amazon은 다양한 소스의 데이터에 애플리케이션을 쉽게 연결할 수 있게 해주는 서버리스 이벤트 버스 서비스입니다. EventBridge 자체 애플리케이션, SaaS (Software-as-a-Service) 애플리케이션 및 서비스의 실시간 데이터 스트림을 제공하고 해당 데이터를 AWS Lambda와 같은 대상으로 라우팅합니다. 이 방법을 통해 서비스에서 발생하는 이벤트를 모니터링하고 이벤트 기반 아키텍처를 구축할 수 있습니다. 자세한 정보는 [Amazon을 사용하여 애플리케이션 Auto Scaling 이벤트를 모니터링합니다 EventBridge](#)을 참조하세요.

AWS Health Dashboard

AWS Health Dashboard (PHD) 는 정보를 표시하고 리소스 상태 변화에 따라 호출되는 알림도 제공합니다. AWS 이 정보는 최근 이벤트와 예정된 이벤트를 카테고리별로 보여주는 대시보드와 지난 90일간의 모든 이벤트를 보여주는 전체 이벤트 로그의 두 가지 방법으로 표시됩니다. 자세한 내용은 [시작하기](#)를 참조하십시오. AWS Health Dashboard

를 사용하여 확장 가능한 리소스의 사용량을 모니터링합니다.

CloudWatch

CloudWatchAmazon을 사용하면 확장 가능한 리소스 전반에서 애플리케이션을 거의 지속적으로 파악할 수 있습니다. CloudWatch AWS 리소스에 대한 모니터링 서비스입니다. 를 CloudWatch 사용하여 지표를 수집 및 추적하고, 경보를 설정하고, AWS 리소스 변화에 자동으로 대응할 수 있습니다. 또한 대시보드를 생성하여 필요한 특정 지표 또는 지표 집합을 모니터링할 수 있습니다.

Application Auto Scaling과 통합되는 서비스와 상호 작용하면 해당 서비스가 다음 표에 나와 있는 메트릭을 로 CloudWatch 전송합니다. CloudWatch에서는 지표가 먼저 서비스 네임스페이스별로 그룹화된 다음 각 네임스페이스 내의 다양한 차원 조합별로 그룹화됩니다. 이러한 지표는 리소스 사용량을 모니터링하고 애플리케이션의 용량을 계획하는 데 도움이 될 수 있습니다. 애플리케이션의 워크로드가 일정하지 않은 경우, 이는 Auto Scaling 사용을 고려해야 한다는 뜻입니다. 이러한 지표에 대한 자세한 설명은 관심 지표에 대한 설명서를 참조하세요.

내용

- [CloudWatch 리소스 사용량 모니터링을 위한 지표](#)
- [대상 추적 조정 정책을 위해 사전 정의된 지표](#)

CloudWatch 리소스 사용량 모니터링을 위한 지표

다음 표에는 리소스 사용 모니터링을 지원하는 데 사용할 수 있는 CloudWatch 지표가 나열되어 있습니다. 목록이 완전하지는 않지만 좋은 출발점이 될 것입니다. CloudWatch 콘솔에 이러한 지표가 표시되지 않는 경우 리소스 설정을 완료했는지 확인하십시오. 자세한 내용은 [Amazon CloudWatch 사용 설명서](#)를 참조하십시오.

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
AppStream 2.0			
플릿	AWS/ AppStream	이름: Available Capacity 차원: 플 릿	AppStream 2.0 메트릭
플릿	AWS/ AppStream	이름: CapacityU tilization 차원: 플 릿	AppStream 2.0 메트릭
Aurora			
복제본	AWS/ RDS	이름: CPUUtiliz ation 크기: DBCluster Identifier, 역할 (리 더)	Aurora 클러스터 수준 지표

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
복제본	AWS/RDS	이름: DatabaseConnections 크기: DBClusterIdentifier, 역할 (리더)	Aurora 클러스터 수준 지표
Amazon Comprehend			
문서 분류 엔드포인트	AWS/Comprehend	이름: InferenceUtilization 치수: EndpointArn	Amazon Comprehend 엔드포인트 지표
엔터티 인식기 엔드포인트	AWS/Comprehend	이름: InferenceUtilization 치수: EndpointArn	Amazon Comprehend 엔드포인트 지표
DynamoDB			

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
테이블 및 전역 보조 인덱스	AWS/ DynamoDB	이름: ProvisionedReadCapacityUnits 치수: TableName , GlobalSecondaryIndexName	DynamoDB 지표
테이블 및 전역 보조 인덱스	AWS/ DynamoDB	이름: ProvisionedWriteCapacityUnits 치수: TableName , GlobalSecondaryIndexName	DynamoDB 지표

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
테이블 및 전역 보조 인덱스	AWS/ DynamoDB	이름: ConsumedReadCapacityUnits 치수: TableName , GlobalSecondaryIndexName	DynamoDB 지표
테이블 및 전역 보조 인덱스	AWS/ DynamoDB	이름: ConsumedWriteCapacityUnits 치수: TableName , GlobalSecondaryIndexName	DynamoDB 지표
Amazon ECS			
서비스	AWS/ ECS	이름: CPUUtilization 치수: ClusterName, ServiceName	Amazon ECS 지표

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
서비스	AWS/ ECS	이름: MemoryUtilization 치수: ClusterName, ServiceName	Amazon ECS 지표
서비스	AWS/ ApplicationELB	이름: RequestCountPerTarget 치수: TargetGroup	Application Load Balancer 지표
ElastiCache			
클러스터(복제 그룹)	AWS/ ElastiCache	이름: DatabaseMemoryUsageCountedForEvictPercentage 치수: ReplicationGroupId	ElastiCache Redis 메트릭의 경우

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
클러스터(복제 그룹)	AWS/ ElastiCache	이름: DatabaseCapacityUsageCountedForEvictionPercentage 치수: ReplicationGroupId	ElastiCache Redis 메트릭의 경우
클러스터(복제 그룹)	AWS/ ElastiCache	이름: EngineCPUUtilization 크기: ReplicationGroupId, 역할 (기본)	ElastiCache Redis 메트릭의 경우
클러스터(복제 그룹)	AWS/ ElastiCache	이름: EngineCPUUtilization 크기: ReplicationGroupId, 역할 (레플리카)	ElastiCache Redis 메트릭의 경우
Amazon EMR			

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
클러스터	AWS/ 리듀스 ElasticMap	이름: 원 사 퍼센트 MemoryAvai lable 치수: ClusterId	Amazon EMR 지표
Amazon Keyspaces			
표	AWS/ Cassa ndra	이름: Provision edReadCap acityUnits 크기: 키 스페이스, TableName	Amazon Keyspaces 지표
표	AWS/ Cassa ndra	이름: Provision edWriteCa pacityUni ts 크기: 키 스페이스, TableName	Amazon Keyspaces 지표

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
표	AWS/Cassandra	이름: ConsumedReadCapacityUnits 크기: 키스페이스, TableName	Amazon Keyspaces 지표
표	AWS/Cassandra	이름: ConsumedWriteCapacityUnits 크기: 키스페이스, TableName	Amazon Keyspaces 지표
Lambda			
프로비저닝된 동시성	AWS/Lambda	이름: ProvisionedConcurrencyUtilization 크기: FunctionName, 리소스	Lambda 함수 지표
Amazon MSK			

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
브로커 스토리지	AWS/ Kafka	이름: KafkaData LogsDiskUsed 차원: 클러스터 이름	Amazon MSK 지표
브로커 스토리지	AWS/ Kafka	이름: KafkaData LogsDiskUsed 차원: 클러스터 이름, 브로커 ID	Amazon MSK 지표
Neptune			
클러스터	AWS/ Neptune	이름: CPUUtilization 크기: DBClusterIdentifier, 역할 (리더)	Neptune 지표
SageMaker			

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
엔드포인트 변형	AWS/SageMaker	이름: InvocationsPerInstance 치수: EndpointName, VariantName	호출 지표
추론 구성 요소	AWS/SageMaker	이름: InferencesPerCopy 치수: InferenceComponentName	호출 지표
서버리스 엔드포인트의 프로비저닝된 동시성	AWS/SageMaker	이름: ServerlessProvisionedConcurrencyUtilization 치수: EndpointName, VariantName	서버리스 엔드포인트 지표
스팟 플릿(Amazon EC2)			

확장 가능한 리소스	네임스페이스	CloudWatch 메트릭	설명서 링크
Spot Fleets	AWS/EC2Spot	이름: CPUUtilization 치수: FleetRequestId	스팟 플릿 지표
Spot Fleets	AWS/EC2Spot	이름: NetworkIn 치수: FleetRequestId	스팟 플릿 지표
Spot Fleets	AWS/EC2Spot	이름: NetworkOut 치수: FleetRequestId	스팟 플릿 지표
Spot Fleets	AWS/ApplicationELB	이름: RequestCountPerTarget 치수: TargetGroup	Application Load Balancer 지표

대상 추적 조정 정책을 위해 사전 정의된 지표

다음 표에는 [Application Auto Scaling API 참조의](#) 사전 정의된 지표 유형과 해당 CloudWatch 지표 이름이 나열되어 있습니다. 사전 정의된 각 지표는 기본 지표 값의 집계를 나타냅니다. CloudWatch 별도로 명시되지 않는 한 결과는 1분 동안의 평균 리소스 사용량을 백분율로 표시합니다. 사전 정의된 지표는 대상 추적 조정 정책을 설정하는 컨텍스트 내에서만 사용됩니다.

이러한 지표에 대한 자세한 내용은 서비스의 설명서([CloudWatch 리소스 사용량 모니터링을 위한 지표](#)의 표에서 확인 가능)를 참조하세요.

사전 정의된 지표 유형	CloudWatch 지표 이름
AppStream 2.0	
AppStreamAverageCapacityUtilization	CapacityUtilization
Aurora	
RDSReaderAverageCPUUtilization	CPUUtilization
RDSReaderAverageDatabaseConnections	DatabaseConnections ¹
Amazon Comprehend	
ComprehendInferenceUtilization	InferenceUtilization
DynamoDB	
DynamoDBReadCapacityUtilization	ProvisionedReadCapacityUnits, ConsumedReadCapacityUnits ²
DynamoDBWriteCapacityUtilization	ProvisionedWriteCapacityUnits, ConsumedWriteCapacityUnits ²
Amazon ECS	
ECSServiceAverageCPUUtilization	CPUUtilization

사전 정의된 지표 유형	CloudWatch 지표 이름
ECSServiceAverageMemoryUtilization	MemoryUtilization
ALBRequestCountPerTarget	RequestCountPerTarget ¹
ElastiCache	
ElastiCacheDatabaseMemoryUsageCountedForEvictPercentage	DatabaseMemoryUsageCountedForEvict 백분율
ElastiCacheDatabaseCapacityUsageCountedForEvictPercentage	DatabaseCapacityUsageCountedForEvict 백분율
ElastiCachePrimaryEngineCPUUtilization	EngineCPUUtilization
ElastiCacheReplicaEngineCPUUtilization	EngineCPUUtilization
Amazon Keyspaces	
CassandraReadCapacityUtilization	ProvisionedReadCapacityUnits, ConsumedReadCapacityUnits ²
CassandraWriteCapacityUtilization	ProvisionedWriteCapacityUnits, ConsumedWriteCapacityUnits ²
Lambda	
LambdaProvisionedConcurrencyUtilization	ProvisionedConcurrency 사용률
Amazon MSK	
KafkaBrokerStorageUtilization	KafkaDataLogsDisk 증고
Neptune	

사전 정의된 지표 유형	CloudWatch 지표 이름
NeptuneReaderAverageCPUUtilization	CPUUtilization
SageMaker	
SageMakerVariantInvocationsPerInstance	InvocationsPer인스턴스 ¹
SageMakerInferenceComponentInvocationsPerCopy	InvocationsPer복사 ¹
SageMakerVariantProvisionedConcurrencyUtilization	ServerlessProvisionedConcurrencyUtilization
스팟 플릿	
EC2SpotFleetRequestAverageCPUUtilization	CPUUtilization ³
EC2SpotFleetRequestAverageNetworkIn ³	NetworkIn ^{1 3}
EC2SpotFleetRequestAverageNetworkOut ³	NetworkOut ^{1 3}
ALBRequestCountPerTarget	RequestCountPerTarget ¹

¹ 지표는 백분율이 아닌 개수를 기반으로 합니다.

² DynamoDB 및 Amazon Keyspace의 경우 사전 정의된 지표는 프로비저닝된 처리량 사용량을 기반으로 한 조정을 지원하는 두 CloudWatch 지표의 집계입니다.

³ 확장 가능한 최상의 성능을 얻으려면 Amazon EC2 세부 모니터링을 사용해야 합니다.

를 사용하여 애플리케이션 Auto Scaling API 호출을 기록합니다. AWS CloudTrail

Application Auto Scaling은 사용자 AWS CloudTrail, 역할 또는 Application Auto Scaling API를 AWS 서비스 사용하여 수행한 작업에 대한 기록을 제공하는 서비스인 서비스와 통합됩니다. CloudTrail Application Auto Scaling에 대한 모든 API 호출을 이벤트로 캡처합니다. 캡처된 호출에는 Application Auto Scaling API에 대한 AWS Management Console 및 코드 호출이 포함됩니다. 트레일을 생성하면 Application Auto Scaling을 위한 CloudTrail 이벤트를 포함하여 Amazon S3 버킷에 이벤트를 지속적으로 전송할 수 있습니다. 트레일을 구성하지 않아도 CloudTrail 콘솔의 이벤트 기록에서 가장 최근 이벤트를 계속 볼 수 있습니다. 에서 수집한 CloudTrail 정보를 사용하여 Application Auto Scaling에 이루어진 요청, 요청이 이루어진 IP 주소, 요청한 사람, 요청 시기 및 추가 세부 정보를 확인할 수 있습니다.

자세한 CloudTrail 내용은 [AWS CloudTrail 사용 설명서를](#) 참조하십시오.

의 Application Auto Scaling 정보 CloudTrail

CloudTrail 계정을 만들 AWS 계정 때 활성화됩니다. Application Auto Scaling 활동이 발생하면 해당 활동이 CloudTrail 이벤트 기록의 다른 AWS 서비스 이벤트와 함께 이벤트에 기록됩니다. 에서 최근 이벤트를 보고, 검색하고, 다운로드할 수 있습니다 AWS 계정. 자세한 내용은 이벤트 [기록으로 CloudTrail 이벤트 보기를](#) 참조하십시오.

Application Auto Scaling을 위한 이벤트를 포함하여 귀하의 AWS 계정 이벤트에 대한 지속적인 기록을 보려면 트레일을 생성하십시오. 트레일을 사용하면 CloudTrail Amazon S3 버킷으로 로그 파일을 전송할 수 있습니다. 콘솔에서 추적을 생성하면 기본적으로 모든 AWS 리전에 추적이 적용됩니다. 트레일은 AWS 파티션에 있는 모든 지역의 이벤트를 기록하고 지정한 Amazon S3 버킷으로 로그 파일을 전송합니다. 또한 CloudTrail 로그에서 수집된 이벤트 데이터를 추가로 분석하고 이에 따라 조치를 취하도록 다른 Amazon Web Services를 구성할 수 있습니다. 자세한 내용은 다음을 참조하십시오.

- [추적 생성 개요](#)
- [CloudTrail 지원되는 서비스 및 통합](#)
- [에 대한 Amazon SNS 알림 구성 CloudTrail](#)
- [여러 지역에서 CloudTrail 로그 파일 수신 및 여러 계정으로부터 CloudTrail 로그 파일 수신](#)

모든 애플리케이션 Auto Scaling 작업은 [애플리케이션 Auto Scaling API 참조에](#) 의해 CloudTrail 기록되고 문서화됩니다. 예를 들어, PutScalingPolicyDeleteScalingPolicy, 및 DescribeScalingPolicies 작업에 대한 호출은 CloudTrail 로그 파일에 항목을 생성합니다.

모든 이벤트 및 로그 항목에는 요청을 생성한 사용자에 대한 정보가 들어 있습니다. 신원 정보를 이용하면 다음을 쉽게 알아볼 수 있습니다.

- 요청이 루트 또는 AWS Identity and Access Management (IAM) 사용자 자격 증명으로 이루어졌는지 여부.
- 역할 또는 연동 사용자를 위한 임시 보안 인증으로 요청을 생성하였는지.
- 다른 AWS 서비스에서 요청했는지 여부.

자세한 내용은 [CloudTrail UserIdentity](#) 요소를 참조하십시오.

Application Auto Scaling 로그 파일 항목 이해

트레일은 지정한 Amazon S3 버킷에 이벤트를 로그 파일로 전송할 수 있는 구성입니다. CloudTrail 로그 파일에는 하나 이상의 로그 항목이 포함되어 있습니다. 이벤트는 모든 소스의 단일 요청을 나타내며 요청된 작업, 작업 날짜 및 시간, 요청 매개 변수 등에 대한 정보를 포함합니다. CloudTrail 로그 파일은 공개 API 호출의 정렬된 스택 트레이스가 아니므로 특정 순서로 표시되지 않습니다.

다음 예제는 DescribeScalableTargets 작업을 보여주는 CloudTrail 로그 항목을 보여줍니다.

```
{
  "eventVersion": "1.05",
  "userIdentity": {
    "type": "Root",
    "principalId": "123456789012",
    "arn": "arn:aws:iam::123456789012:root",
    "accountId": "123456789012",
    "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
    "sessionContext": {
      "attributes": {
        "mfaAuthenticated": "false",
        "creationDate": "2018-08-21T17:05:42Z"
      }
    }
  },
  "eventTime": "2018-08-16T23:20:32Z",
  "eventSource": "autoscaling.amazonaws.com",
  "eventName": "DescribeScalableTargets",
  "awsRegion": "us-west-2",
  "sourceIPAddress": "72.21.196.68",
  "userAgent": "EC2 Spot Console",
```

```

    "requestParameters": {
      "serviceNamespace": "ec2",
      "scalableDimension": "ec2:spot-fleet-request:TargetCapacity",
      "resourceIds": [
        "spot-fleet-request/sfr-05ceaf79-3ba2-405d-e87b-612857f1357a"
      ]
    },
    "responseElements": null,
    "additionalEventData": {
      "service": "application-autoscaling"
    },
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
  }
}

```

관련 리소스

CloudWatch 로그를 사용하면 에서 캡처한 특정 이벤트를 모니터링하고 알림을 받을 수 있습니다. CloudTrail CloudWatch Logs로 전송되는 이벤트는 트레일에 기록되도록 구성된 이벤트이므로 모니터링하려는 이벤트 유형을 기록하도록 트레일 또는 트레일을 구성했는지 확인하십시오. CloudWatch 로그는 로그 파일의 정보를 모니터링하고 특정 임계값이 충족되면 알려줄 수 있습니다. 또한 매우 내구력 있는 스토리지에 로그 데이터를 저장할 수 있습니다. 자세한 내용은 [Amazon CloudWatch Logs 사용 설명서와 사용 설명서의 Amazon Logs를 사용한 CloudTrail CloudWatch 로그 파일 모니터링 항목](#)을 참조하십시오. AWS CloudTrail

Amazon을 사용하여 애플리케이션 Auto Scaling 이벤트를 모니터링합니다 EventBridge

EventBridgeAmazon은 이전에 CloudWatch Events라고 불렸으며, Application Auto Scaling과 관련된 이벤트를 모니터링하고 다른 이벤트를 사용하는 대상 작업을 시작할 수 있도록 도와줍니다. AWS 서비스의 AWS 서비스 이벤트가 거의 EventBridge 실시간으로 전송됩니다.

를 사용하여 EventBridge 수신 이벤트와 일치하는 규칙을 만들고 이를 처리 대상으로 라우팅할 수 있습니다.

자세한 내용은 Amazon EventBridge 사용 설명서의 [Amazon EventBridge 시작하기](#)를 참조하십시오.

Application Auto Scaling 이벤트

다음 예제에서는 Application Auto Scaling에 대한 이벤트를 보여줍니다. 이벤트는 최선의 작업에 근거하여 생성됩니다.

현재 Application Auto Scaling에는 최대로 확장된 이벤트와 이를 통한 CloudTrail API 호출만 사용할 수 있습니다.

이벤트 유형

- [상태 변경 이벤트: 최대 용량으로 조정](#)
- [를 통한 API 호출 이벤트 CloudTrail](#)

상태 변경 이벤트: 최대 용량으로 조정

다음 예제 이벤트는 Application Auto Scaling이 확장 가능 대상의 용량을 최대 크기 한도까지 증가(스케일 아웃)했음을 보여줍니다. 수요가 다시 증가하면 대상이 이미 최대 크기로 조정되었기 때문에 Application Auto Scaling에서 더 큰 크기로 조정할 수 없습니다.

detail 객체에서 resourceId, serviceNamespace 및 scalableDimension 속성은 확장 가능 대상을 식별합니다. newDesiredCapacity 및 oldDesiredCapacity 속성의 값은 스케일 아웃 이벤트 후의 새 용량과 스케일 아웃 이벤트 전의 원래 용량을 나타냅니다. maxCapacity는 확장 가능 대상의 최대 크기 한도입니다.

```
{
  "version": "0",
  "id": "11112222-3333-4444-5555-666677778888",
  "detail-type": "Application Auto Scaling Scaling Activity State Change",
  "source": "aws.application-autoscaling",
  "account": "123456789012",
  "time": "2019-06-12T10:23:40Z",
  "region": "us-west-2",
  "resources": [],
  "detail": {
    "startTime": "2022-06-12T10:20:43Z",
    "endTime": "2022-06-12T10:23:40Z",
    "newDesiredCapacity": 8,
    "oldDesiredCapacity": 5,
    "minCapacity": 2,
    "maxCapacity": 8,
    "resourceId": "table/my-table",
    "scalableDimension": "dynamodb:table:WriteCapacityUnits",
```

```

    "serviceNamespace": "dynamodb",
    "statusCode": "Successful",
    "scaledToMax": true,
    "direction": "scale-out"
  }

```

모든 확장 가능 대상의 모든 scaledToMax 상태 변경 이벤트를 캡처하는 규칙을 생성하려면 다음 샘플 이벤트 패턴을 사용합니다.

```

{
  "source": [
    "aws.application-autoscaling"
  ],
  "detail-type": [
    "Application Auto Scaling Scaling Activity State Change"
  ],
  "detail": {
    "scaledToMax": [
      true
    ]
  }
}

```

를 통한 API 호출 이벤트 CloudTrail

트레일은 이벤트를 Amazon S3 버킷에 로그 파일로 전송하는 데 AWS CloudTrail 사용하는 구성입니다. CloudTrail 로그 파일에는 로그 항목이 포함되어 있습니다. 이벤트는 개별 로그 항목을 나타내며 요청된 작업, 작업 날짜 및 시간, 요청 파라미터에 대한 정보를 포함합니다. 시작하는 CloudTrail 방법을 알아보려면 AWS CloudTrail 사용 안내서의 [트레일 만들기를](#) 참조하십시오.

CloudTrail 를 AWS API Call via CloudTrail 통해 전달되는 이벤트는 가치가 detail-type 있습니다.

다음 예제 이벤트는 콘솔 사용자가 Application Auto Scaling [RegisterScalableTarget](#) 작업을 호출했음을 보여주는 CloudTrail 로그 파일 항목을 나타냅니다.

```

{
  "version": "0",
  "id": "99998888-7777-6666-5555-444433332222",
  "detail-type": "AWS API Call via CloudTrail",
  "source": "aws.autoscaling",
  "account": "123456789012",

```

```
"time": "2022-07-13T16:50:15Z",
"region": "us-west-2",
"resources": [],
"detail": {
  "eventVersion": "1.08",
  "userIdentity": {
    "type": "IAMUser",
    "principalId": "123456789012",
    "arn": "arn:aws:iam::123456789012:user/Bob",
    "accountId": "123456789012",
    "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
    "sessionContext": {
      "sessionIssuer": {
        "type": "Role",
        "principalId": "123456789012",
        "arn": "arn:aws:iam::123456789012:role/Admin",
        "accountId": "123456789012",
        "userName": "Admin"
      },
      "webIdFederationData": {},
      "attributes": {
        "creationDate": "2022-07-13T15:17:08Z",
        "mfaAuthenticated": "false"
      }
    }
  },
  "eventTime": "2022-07-13T16:50:15Z",
  "eventSource": "autoscaling.amazonaws.com",
  "eventName": "RegisterScalableTarget",
  "awsRegion": "us-west-2",
  "sourceIPAddress": "AWS Internal",
  "userAgent": "EC2 Spot Console",
  "requestParameters": {
    "resourceId": "spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE",
    "serviceNamespace": "ec2",
    "scalableDimension": "ec2:spot-fleet-request:TargetCapacity",
    "minCapacity": 2,
    "maxCapacity": 10
  },
  "responseElements": null,
  "additionalEventData": {
    "service": "application-autoscaling"
  },
  "requestID": "e9caf887-8d88-11e5-a331-3332aa445952",
```

```

    "eventID": "49d14f36-6450-44a5-a501-b0fdcdfaeb98",
    "readOnly": false,
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "123456789012",
    "eventCategory": "Management",
    "sessionCredentialFromConsole": "true"
  }
}

```

모든 확장 가능한 대상에 대한 모든 [DeleteScaling정책](#) 및 [DeregisterScalableTarget](#) API 호출을 기반으로 규칙을 만들려면 다음 샘플 이벤트를 사용하십시오.

```

{
  "source": [
    "aws.autoscaling"
  ],
  "detail-type": [
    "AWS API Call via CloudTrail"
  ],
  "detail": {
    "eventSource": [
      "autoscaling.amazonaws.com"
    ],
    "eventName": [
      "DeleteScalingPolicy",
      "DeregisterScalableTarget"
    ],
    "additionalEventData": {
      "service": [
        "application-autoscaling"
      ]
    }
  }
}

```

사용에 대한 자세한 내용은 [CloudTrail](#) 을 참조하십시오 [를 사용하여 애플리케이션 Auto Scaling API 호출을 기록합니다. AWS CloudTrail.](#)

Application Auto Scaling에 대한 태그 지정 지원

AWS CLI 또는 SDK를 사용하여 Application Auto Scaling의 확장 가능한 대상에 태그를 지정할 수 있습니다. 확장 가능한 대상은 Application Auto Scaling이 확장할 수 있는 리소스 AWS 또는 사용자 지정 리소스를 나타내는 엔티티입니다.

각 태그는 Application Auto Scaling API를 사용하는 사용자 정의 키와 값으로 구성된 레이블입니다. 태그를 지정하면 조직의 요구에 따라 특정 확장 가능 대상에 세분화된 액세스를 구성할 수 있습니다. 자세한 정보는 [Application Auto Scaling 작업을 사용한 ABAC](#)을 참조하세요.

새 확장 가능 대상을 등록할 때 태그를 추가하거나 기존 확장 가능 대상에 태그를 추가할 수 있습니다.

태그 관리에 일반적으로 사용되는 명령은 다음과 같습니다.

- [register-scalable-target](#)은 새로운 확장 가능 대상을 등록할 때 태그를 지정하는 데 사용합니다.
- [tag-resource](#)는 기존 확장 가능 대상에 태그를 추가할 때 사용합니다.
- [list-tags-for-resource](#)는 확장 가능 대상의 태그를 반환할 때 사용합니다.
- [untag-resource](#)는 태그를 삭제할 때 사용합니다.

태그 예제

다음과 같이 [register-scalable-target](#) 명령을 `--tags` 옵션과 함께 사용합니다. 이 예제에서는 확장 가능 대상에 2개의 태그, 즉 태그 키 이름이 **environment**이고 태그 값이 **production**인 태그와 태그 키 이름이 **iscontainerbased**이고 태그 값이 **true**인 태그를 지정합니다.

`--min-capacity``--max-capacity` 및 의 샘플 값과 샘플 텍스트를 Application Auto Scaling에서 사용 중인 AWS 서비스의 네임스페이스, `--scalable-dimension` 등록 중인 리소스와 연결된 확장 가능한 차원, 리소스의 식별자로 `--resource-id` 바꾸십시오. `--service-namespace` 각 서비스에 대한 자세한 내용과 예는 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#)의 주제를 참조하세요.

```
aws application-autoscaling register-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 1 --max-capacity 10 \
  --tags environment=production,iscontainerbased=true
```


이 명령이 성공하면 확장 가능 대상의 ARN이 반환됩니다.

```
{
  "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Note

이 명령에서 오류가 발생하는 경우 AWS CLI 로컬에서 최신 버전으로 업데이트했는지 확인하십시오.

보안을 위한 태그

태그를 사용하여 요청자(예: IAM 사용자 또는 역할)에게 특정 작업을 수행할 권한이 있는지 확인합니다. 다음 조건 키를 하나 이상 사용하여 IAM 정책의 조건 요소에 태그 정보를 제공합니다.

- 특정 태그가 있는 확장 가능 대상에 대한 사용자 작업을 허용(또는 거부)하려면 `aws:ResourceTag/tag-key: tag-value`를 사용합니다.
- 요청에 특정 태그가 존재하도록 (또는 존재하지 않도록) 요구하려면 `aws:RequestTag/tag-key: tag-value`를 사용합니다.
- 요청에 특정 태그 키가 존재하도록 (또는 존재하지 않도록) 요구하려면 `aws:TagKeys [tag-key, ...]`를 사용합니다.

예를 들어 다음의 IAM 정책은 사용자에게 `DeregisterScalableTarget`, `DeleteScalingPolicy` 및 `DeleteScheduledAction` 작업에 대한 권한을 부여합니다. 그러나 적용되는 확장 가능 대상에 `environment=production` 태그가 있는 경우 작업을 거부하기도 합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "application-autoscaling:DeregisterScalableTarget",
        "application-autoscaling:DeleteScalingPolicy",
        "application-autoscaling:DeleteScheduledAction"
      ]
    }
  ]
}
```

```

    ],
    "Resource": "*"
  }
},
{
  "Effect": "Deny",
  "Action": [
    "application-autoscaling:DeregisterScalableTarget",
    "application-autoscaling:DeleteScalingPolicy",
    "application-autoscaling:DeleteScheduledAction"
  ],
  "Resource": "*",
  "Condition": {
    "StringEquals": {"aws:ResourceTag/environment": "production"}
  }
}
]
}

```

태그에 대한 액세스 통제

태그를 사용하여 요청자(예: IAM 사용자 또는 역할)에게 확장 가능 대상에 대한 태그를 추가, 수정 또는 삭제할 수 있는 권한이 있는지 확인합니다.

예를 들어 확장 가능 대상에서 **temporary** 키가 지정된 태그만 제거하도록 허용하는 IAM 정책을 생성할 수 있습니다.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "application-autoscaling:UntagResource",
      "Resource": "*",
      "Condition": {
        "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }
      }
    }
  ]
}

```

Application Auto Scaling의 보안

클라우드 AWS 보안이 최우선 과제입니다. AWS 고객은 가장 보안에 민감한 조직의 요구 사항을 충족하도록 구축된 데이터 센터 및 네트워크 아키텍처의 혜택을 누릴 수 있습니다.

보안은 기업과 기업 간의 AWS 공동 책임입니다. [공동 책임 모델](#)은 이 사항을 클라우드의 보안 및 클라우드 내 보안으로 설명합니다.

- 클라우드 보안 - AWS 클라우드에서 AWS 서비스를 실행하는 인프라를 보호하는 역할을 합니다. AWS 또한 안전하게 사용할 수 있는 서비스를 제공합니다. Application Auto Scaling에 적용되는 규정 준수 프로그램에 대해 자세히 알아보려면 규정 준수 [프로그램별 범위 내 서비스 규정 준수](#) 참조하십시오.
- 클라우드에서의 보안 — 사용하는 AWS 서비스에 따라 책임이 결정됩니다. 또한 귀하는 데이터의 민감도, 회사 요구 사항, 관련 법률 및 규정을 비롯한 기타 요소에 대해서도 책임이 있습니다.

이 설명서는 Application Auto Scaling 사용 시 책임 분담 모델을 적용하는 방법을 이해하는 데 도움이 됩니다. 다음 항목에서는 보안 및 규정 준수 목표를 충족하도록 Application Auto Scaling을 구성하는 방법을 보여줍니다. 또한 Application Auto Scaling 리소스를 모니터링하고 보호하는 데 도움이 되는 다른 AWS 서비스를 사용하는 방법도 알아봅니다.

내용

- [Application Auto Scaling에서의 데이터 보호](#)
- [Application Auto Scaling에 사용되는 Identity and Access Management](#)
- [인터페이스 VPC 엔드포인트를 사용하여 Application Auto Scaling에 액세스](#)
- [Application Auto Scaling의 복원성](#)
- [Application Auto Scaling의 인프라 보안](#)
- [Application Auto Scaling의 규정 준수 확인](#)

Application Auto Scaling에서의 데이터 보호

AWS [공동 책임 모델](#) 이 모델에 설명된 대로 AWS 은 (는) 모두를 실행하는 글로벌 인프라를 보호하는 역할을 AWS 클라우드합니다. 사용자는 인프라에서 호스팅되는 콘텐츠를 관리해야 합니다. 사용하는 AWS 서비스 의 보안 구성과 관리 작업에 대한 책임도 사용자에게 있습니다. 데이터 프라이버시에 대한 자세한 내용은 [데이터 프라이버시 FAQ](#)를 참조하세요. 유럽의 데이터 보호에 대한 자세한 내용은 AWS 보안 블로그의 [AWS 공동 책임 모델 및 GDPR](#) 블로그 게시물을 참조하세요.

데이터 보호를 위해 AWS 계정 자격 증명을 보호하고 AWS IAM Identity Center OR AWS Identity and Access Management (IAM) 을 사용하여 개별 사용자를 설정하는 것이 좋습니다. 이렇게 하면 개별 사용자에게 자신의 직무를 충실히 이행하는 데 필요한 권한만 부여됩니다. 또한 다음과 같은 방법으로 데이터를 보호하는 것이 좋습니다.

- 각 계정에 멀티 팩터 인증 설정(MFA)을 사용하세요.
- SSL/TLS를 사용하여 리소스와 통신하세요. AWS TLS 1.2는 필수이며 TLS 1.3를 권장합니다.
- 를 사용하여 API 및 사용자 활동 로깅을 설정합니다. AWS CloudTrail
- 포함된 모든 기본 보안 제어와 함께 AWS 암호화 솔루션을 사용하십시오 AWS 서비스.
- Amazon S3에 저장된 민감한 데이터를 검색하고 보호하는 데 도움이 되는 Amazon Macie와 같은 고급 관리형 보안 서비스를 사용하세요.
- 명령줄 인터페이스 또는 API를 AWS 통해 액세스할 때 FIPS 140-2로 검증된 암호화 모듈이 필요한 경우 FIPS 엔드포인트를 사용하십시오. 사용 가능한 FIPS 엔드포인트에 대한 자세한 내용은 [FIPS\(Federal Information Processing Standard\) 140-2](#)를 참조하세요.

고객의 이메일 주소와 같은 기밀 정보나 중요한 정보는 태그나 이름 필드와 같은 자유 양식 필드에 입력하지 않는 것이 좋습니다. 여기에는 콘솔 AWS CLI, API 또는 AWS 서비스 AWS SDK를 사용하여 Application Auto Scaling 또는 기타 작업을 수행하는 경우가 포함됩니다. 이름에 사용되는 태그 또는 자유 형식 텍스트 필드에 입력하는 모든 데이터는 청구 또는 진단 로그에 사용될 수 있습니다. 외부 서버에 URL을 제공할 때 해당 서버에 대한 요청을 검증하기 위해 보안 인증 정보를 URL에 포함해서는 안 됩니다.

Application Auto Scaling에 사용되는 Identity and Access Management

AWS Identity and Access Management (IAM) 은 관리자가 리소스에 대한 액세스를 안전하게 제어하는 데 도움이 되는 도구입니다. AWS IAM 관리자는 어떤 사용자가 Application Auto Scaling 리소스를 사용할 수 있는 인증(로그인) 및 권한(권한 있음)을 받을 수 있는지를 제어합니다. IAM은 추가 AWS 서비스 비용 없이 사용할 수 있습니다.

IAM 설명서 전체 내용은 [IAM 사용 설명서](#)를 참조하세요.

액세스 제어

요청을 인증하는 데 유효한 자격 증명이 있더라도 권한이 없다면 Application Auto Scaling 리소스를 생성하거나 액세스할 수 없습니다. 예를 들어 조정 정책을 생성하고 예약된 조정을 구성하는 등의 권한이 있어야 합니다.

다음 섹션에서는 IAM 관리자가 Application Auto Scaling API 작업을 수행할 수 있는 사용자를 제어함으로써 IAM을 사용하여 AWS 리소스를 보호하는 방법을 자세히 설명합니다.

내용

- [Application Auto Scaling에서 IAM을 사용하는 방식](#)
- [AWS Application Auto Scaling에 대한 관리형 정책](#)
- [Application Auto Scaling에 대한 서비스 연결 역할](#)
- [Application Auto Scaling 자격 증명 기반 정책 예제](#)
- [Application Auto Scaling 액세스 문제 해결](#)
- [대상 리소스에 대한 Application Auto Scaling API 호출에 대한 권한 검증](#)

Application Auto Scaling에서 IAM을 사용하는 방식

Note

2017년 12월에 Application Auto Scaling 통합 서비스에 여러 서비스 연결 역할을 사용할 수 있도록 하는 Application Auto Scaling 업데이트가 있었습니다. 사용자가 조정을 구성할 수 있도록 특정 IAM 권한 및 Application Auto Scaling 서비스 연결 역할(또는 Amazon EMR 자동 크기 조정의 서비스 역할)이 필요합니다.

IAM을 사용하여 Application Auto Scaling에 대한 액세스를 관리하려면 먼저 어떤 IAM 기능을 Application Auto Scaling에 사용할 수 있는지를 학습하세요.

Application Auto Scaling에서 사용할 수 있는 IAM 기능

IAM 특성	Application Auto Scaling 지원
ID 기반 정책	예
정책 작업	예

IAM 특성	Application Auto Scaling 지원
정책 리소스	예
정책 조건 키(서비스별)	예
리소스 기반 정책	아니요
ACL	아니요
ABAC(정책 내 태그)	부분
임시 보안 인증	예
서비스 역할	예
서비스 연결 역할	예

Application Auto Scaling 및 기타 기능이 대부분의 IAM 기능과 어떻게 AWS 서비스 작동하는지 자세히 알아보려면 IAM 사용 설명서에서 [IAM과AWS 서비스 연동되는](#) 기능을 참조하십시오.

Application Auto Scaling 자격 증명 기반 정책

ID 기반 정책 지원	예
-------------	---

자격 증명 기반 정책은 IAM 사용자, 사용자 그룹 또는 역할과 같은 자격 증명에 연결할 수 있는 JSON 권한 정책 문서입니다. 이러한 정책은 사용자와 역할이 어떤 리소스와 어떤 조건에서 어떤 작업을 수행할 수 있는지를 제어합니다. 자격 증명 기반 정책을 생성하는 방법을 알아보려면 IAM 사용 설명서의 [IAM 정책 생성](#)을 참조하세요.

IAM 자격 증명 기반 정책을 사용하면 허용되거나 거부되는 작업과 리소스뿐 아니라 작업이 허용되거나 거부되는 조건을 지정할 수 있습니다. 자격 증명 기반 정책에서는 보안 주체가 연결된 사용자 또는 역할에 적용되므로 보안 주체를 지정할 수 없습니다. JSON 정책에서 사용하는 모든 요소에 대해 알아보려면 IAM 사용 설명서의 [IAM JSON 정책 요소 참조](#)를 참조하세요.

Application Auto Scaling의 자격 증명 기반 정책 예제

Application Auto Scaling 자격 증명 기반 정책 예제를 보려면 [Application Auto Scaling 자격 증명 기반 정책 예제](#) 섹션을 참조하세요.

작업

정책 작업 지원

예

IAM 정책 구문에는 IAM을 지원하는 모든 서비스의 모든 API 작업을 지정할 수 있습니다. Application Auto Scaling의 경우 접두사와 함께 API 작업 이름 application-autoscaling:을 (를) 사용합니다. 예를 들어, application-autoscaling:RegisterScalableTarget, application-autoscaling:PutScalingPolicy 및 application-autoscaling:DeregisterScalableTarget입니다.

단일 명령문에서 여러 작업을 지정하려면 다음 예제와 같이 쉼표로 구분합니다.

```
"Action": [
    "application-autoscaling:DescribeScalingPolicies",
    "application-autoscaling:DescribeScalingActivities"
```

와일드카드(*)를 사용하여 여러 작업을 지정할 수 있습니다. 예를 들어, Describe라는 단어로 시작하는 모든 작업을 지정하려면 다음 작업을 포함합니다.

```
"Action": "application-autoscaling:Describe*"
```

애플리케이션 Auto Scaling 작업 목록은 서비스 권한 부여 참조의 AWS [Application Auto Scaling에서 정의한 작업을](#) 참조하십시오.

리소스

정책 리소스 지원

예

IAM 정책 구문에서 Resource 요소는 명령문이 다루는 하나 이상의 객체를 지정합니다. Application Auto Scaling의 경우 Amazon 리소스 이름(ARN)을 사용하여 지정한 확장 가능 대상에 각 IAM 정책 구문이 적용됩니다.

확장 가능 대상을 위한 ARN 리소스 형식:

```
arn:aws:application-autoscaling:region:account-id:scalable-target/unique-identifier
```

예를 들어 구문에서 다음과 같이 ARN을 사용하여 특정 확장 가능 대상을 나타낼 수 있습니다. 고유 ID(1234abcd56ab78cd901ef1234567890ab123)는 Application Auto Scaling을 통해 확장 가능 대상에 할당된 값입니다.

```
"Resource": "arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
```

다음과 같이 고유 식별자를 와일드카드(*)로 대체하여 특정 계정에 속하는 모든 인스턴스를 지정할 수 있습니다.

```
"Resource": "arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/*"
```

모든 리소스를 지정하려 하거나 특정 API 작업이 ARN을 지원하지 않는 경우 다음과 같이 와일드카드(*)를 Resource 요소로 사용합니다.

```
"Resource": "*"
```

자세한 내용은 서비스 권한 부여 참조의 AWS [Application Auto Scaling에서 정의한 리소스 유형](#)을 참조하십시오.

조건 키

서비스별 정책 조건 키 지원	예
-----------------	---

IAM 정책에 Application Auto Scaling 리소스에 대한 액세스를 제어하는 조건을 지정할 수 있습니다. 이 정책문은 조건이 true일 때만 유효합니다.

Application Auto Scaling은 자격 증명 기반 정책에서 Application Auto Scaling API 작업을 수행 가능한 사용자를 결정하는 데 사용할 수 있는 다음과 같은 서비스 정의 조건 키를 지원합니다.

- application-autoscaling:scalable-dimension
- application-autoscaling:service-namespace

조건 키를 사용할 수 있는 Application Auto Scaling API 작업에 대해 알아보려면 서비스 권한 부여 참조의 AWS [Application Auto Scaling에서 정의한 작업](#)을 참조하십시오. Application Auto Scaling 조건 키 사용에 대한 자세한 내용은 [AWS Application Auto Scaling의 조건 키](#)를 참조하십시오.

모든 서비스에 사용할 수 있는 글로벌 조건 키를 보려면 IAM 사용 설명서의 [AWS 글로벌 조건 컨텍스트 키](#)를 참조하세요.

리소스 기반 정책

리소스 기반 정책 지원	아니요
--------------	-----

Amazon 심플 스토리지 AWS 서비스와 같은 다른 서비스는 리소스 기반 권한 정책을 지원합니다. 예를 들어, 권한 정책을 S3 버킷에 연결하여 해당 버킷에 대한 액세스 권한을 관리할 수 있습니다.

Application Auto Scaling은 리소스 기반 정책을 지원하지 않습니다.

액세스 제어 목록(ACL)

ACL 지원	아니요
--------	-----

Application Auto Scaling은 액세스 제어 목록(ACL)을 지원하지 않습니다.

Application Auto Scaling 작업을 사용한 ABAC

ABAC(정책 내 태그) 지원	부분
------------------	----

속성 기반 액세스 제어(ABAC)는 속성을 기반으로 권한을 정의하는 권한 부여 전략입니다. AWS에서는 이러한 속성을 태그라고 합니다. IAM 개체 (사용자 또는 역할) 및 여러 AWS 리소스에 태그를 첨부할 수 있습니다. ABAC의 첫 번째 단계로 개체 및 리소스에 태그를 지정합니다. 그런 다음 보안 주체의 태그가 액세스하려는 리소스의 태그와 일치할 때 작업을 허용하도록 ABAC 정책을 설계합니다.

ABAC는 빠르게 성장하는 환경에서 유용하며 정책 관리가 번거로운 상황에 도움이 됩니다.

태그를 기반으로 액세스를 제어하려면 `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` 또는 `aws:TagKeys` 조건 키를 사용하여 정책의 [조건 요소](#)에 태그 정보를 제공합니다.

태그를 지원하는 리소스에는 ABAC를 사용할 수 있지만 모든 리소스가 태그를 지원하는 것은 아닙니다. 예약된 작업 및 크기 조정 정책은 태그를 지원하지 않지만, 확장 가능 대상은 태그를 지원합니다. 자세한 정보는 [Application Auto Scaling에 대한 태그 지정 지원](#)을 참조하세요.

ABAC에 대한 자세한 정보는 IAM 사용 설명서의 [ABAC란 무엇인가요?](#)를 참조하세요. ABAC 설정 단계가 포함된 자습서를 보려면 IAM 사용 설명서의 [속성 기반 액세스 제어\(ABAC\) 사용](#)을 참조하세요.

Application Auto Scaling과 함께 임시 자격 증명 사용

임시 보안 인증 지원

예

임시 자격 증명을 사용하여 로그인하면 작동하지 AWS 서비스 않는 것도 있습니다. 임시 자격 증명을 사용하는 방법을 AWS 서비스 비롯한 추가 정보는 [IAM 사용 설명서의 IAM과AWS 서비스 연동되는 내용](#)을 참조하십시오.

사용자 이름과 암호를 제외한 다른 방법을 AWS Management Console 사용하여 로그인하면 임시 자격 증명을 사용하는 것입니다. 예를 들어 회사의 SSO (Single Sign-On) 링크를 AWS 사용하여 액세스 하는 경우 이 프로세스에서 자동으로 임시 자격 증명을 생성합니다. 또한 콘솔에 사용자로 로그인한 다음 역할을 전환할 때 임시 보안 인증을 자동으로 생성합니다. 역할 전환에 대한 자세한 정보는 IAM 사용 설명서의 [역할로 전환\(콘솔\)](#)을 참조하세요.

또는 API를 사용하여 임시 자격 증명을 수동으로 생성할 수 있습니다 AWS CLI . AWS 그런 다음 해당 임시 자격 증명을 사용하여 액세스할 수 AWS있습니다. AWS 장기 액세스 키를 사용하는 대신 임시 자격 증명을 동적으로 생성할 것을 권장합니다. 자세한 정보는 [IAM의 임시 보안 인증](#) 섹션을 참조하세요.

서비스 역할

서비스 역할 지원

예

Amazon EMR 클러스터에서 자동 크기 조정을 사용하는 경우 이 기능을 사용하면 Application Auto Scaling이 사용자를 대신하여 [서비스 역할](#)을 수입할 수 있습니다. 서비스 연결 역할과 마찬가지로, 서비스 역할을 사용하면 서비스가 다른 서비스의 리소스에 액세스하고 사용자를 대신하여 작업을 완료 할 수 있습니다. 서비스 역할은 IAM 계정에 나타나고, 해당 계정이 소유합니다. 즉, IAM 관리자가 이 역할에 대한 권한을 변경할 수 있습니다. 그러나 권한을 변경하면 서비스의 기능이 손상될 수 있습니다.

Application Auto Scaling은 Amazon EMR에 대해서만 서비스 역할을 지원합니다. EMR 서비스 역할에 대한 설명서는 Amazon EMR 관리 가이드의 [인스턴스 그룹에 대한 사용자 지정 정책과 함께 자동 크기 조정 사용](#)을 참조하세요.

Note

서비스 연결 역할이 도입됨에 따라 Amazon ECS 및 스팟 플릿과 같은 여러 레거시 서비스 역할이 더 이상 필요하지 않습니다.

서비스 링크 역할

서비스 링크 역할 지원

예

서비스 연결 역할은 예 연결된 서비스 역할의 한 유형입니다. AWS 서비스는 사용자를 대신하여 작업을 수행하기 위해 역할을 수임할 수 있습니다. 서비스 연결 역할은 사용자에게 AWS 계정 표시되며 해당 서비스가 소유합니다. IAM 관리자는 서비스 링크 역할의 권한을 볼 수 있지만 편집은 할 수 없습니다.

자세한 정보는 Application Auto Scaling 서비스 연결 역할에 대한 자세한 정보는 [Application Auto Scaling에 대한 서비스 연결 역할](#)(를) 참조하세요.

AWS Application Auto Scaling에 대한 관리형 정책

AWS 관리형 정책은 에서 생성하고 관리하는 독립 실행형 정책입니다. AWS 관리형 정책은 많은 일반 사용 사례에 대한 권한을 제공하도록 설계되었으므로 사용자, 그룹 및 역할에 권한을 할당하기 시작할 수 있습니다.

AWS 관리형 정책은 모든 AWS 고객이 사용할 수 있으므로 특정 사용 사례에 대해 최소 권한을 부여하지 않을 수도 있다는 점에 유의하세요. 사용 사례에 고유한 [고객 관리형 정책](#)을 정의하여 권한을 줄이는 것이 좋습니다.

관리형 정책에 정의된 권한은 변경할 수 없습니다. AWS 관리형 정책에 정의된 권한을 업데이트 하는 경우 해당 업데이트는 정책이 연결된 모든 주체 ID (사용자, 그룹, 역할) 에 영향을 미칩니다. AWS 새 API 작업이 시작되거나 기존 서비스에 새 AWS 서비스 API 작업을 사용할 수 있게 되면 AWS 관리형 정책을 업데이트할 가능성이 가장 높습니다.

자세한 내용은 IAM 사용자 설명서의 [AWS 관리형 정책](#)을 참조하세요.

AWS 관리형 정책: AppStream 2.0 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingAppStreamFleetPolicy](#)

이 정책은 Application Auto Scaling이

[AWSServiceRoleForApplicationAutoScaling_AppStreamFleet](#) Amazon에 전화를 걸어 사용자를 AppStream 대신하여 CloudWatch 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: appstream:DescribeFleets
- 작업: appstream:UpdateFleet
- 작업: cloudwatch:DescribeAlarms
- 작업: cloudwatch:PutMetricAlarm
- 작업: cloudwatch>DeleteAlarms

AWS 관리형 정책: Aurora 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingRDSClusterPolicy](#)

이 정책은 Application Auto Scaling이 사용자를

[AWSServiceRoleForApplicationAutoScaling_RDSCluster](#) 대신하여 Aurora를 CloudWatch 호출하고 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: rds:AddTagsToResource
- 작업: rds:CreateDBInstance
- 작업: rds>DeleteDBInstance
- 작업: rds:DescribeDBClusters
- 작업: rds:DescribeDBInstance
- 작업: cloudwatch:DescribeAlarms
- 작업: cloudwatch:PutMetricAlarm

- 작업: `cloudwatch:DeleteAlarms`

AWS 관리형 정책: 아마존 Comprehend 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingComprehendEndpointPolicy](#)

이 정책은 Application Auto Scaling이 Amazon

[AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint](#) Comprehend를 CloudWatch 호출하고 사용자를 대신하여 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `comprehend:UpdateEndpoint`
- 작업: `comprehend:DescribeEndpoint`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`
- 작업: `cloudwatch:DeleteAlarms`

AWS 관리형 정책: DynamoDB 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingDynamoDBTablePolicy](#)

이 정책은 Application Auto Scaling이

[AWSServiceRoleForApplicationAutoScaling_DynamoDBTable](#) DynamodBand를 CloudWatch 호출하고 사용자 대신 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `dynamodb:DescribeTable`
- 작업: `dynamodb:UpdateTable`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`

- 작업: `cloudwatch:DeleteAlarms`

AWS 관리형 정책: Amazon ECS 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingECSServicePolicy](#)

이 정책은 Application Auto Scaling이 Amazon ECS를 CloudWatch 호출하고 사용자를 대신하여 조정을 수행할 수 있도록 이름이 지정된 [AWSServiceRoleForApplicationAutoScaling_ECSService](#) 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `ecs:DescribeServices`
- 작업: `ecs:UpdateService`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`
- 작업: `cloudwatch:DeleteAlarms`

AWS 관리형 정책: ElastiCache 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingElastiCacheRGPolicy](#)

이 정책은 Application Auto Scaling이 사용자를 [AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG](#) 대신하여 조정을 ElastiCache CloudWatch 호출하고 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책을 통해 Application Auto Scaling은 지정된 리소스에서 다음 작업을 완료할 수 있습니다.

- 작업: 모든 리소스에 대한 `elasticache:DescribeReplicationGroups`
- 작업: 모든 리소스에 대한 `elasticache:ModifyReplicationGroupShardConfiguration`
- 작업: 모든 리소스에 대한 `elasticache:IncreaseReplicaCount`
- 작업: 모든 리소스에 대한 `elasticache:DecreaseReplicaCount`
- 작업: 모든 리소스에 대한 `elasticache:DescribeCacheClusters`

- 작업: 모든 리소스에 대한 `elasticache:DescribeCacheParameters`
- 작업: 모든 리소스에 대한 `cloudwatch:DescribeAlarms`
- 작업: `arn:aws:cloudwatch:*:*:alarm:TargetTracking*` 리소스에 대한 `cloudwatch:PutMetricAlarm`
- 작업: `arn:aws:cloudwatch:*:*:alarm:TargetTracking*` 리소스에 대한 `cloudwatch>DeleteAlarms`
- 작업: `cloudwatch>DeleteAlarms`

AWS 관리형 정책: Amazon 키스페이스 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingCassandraTablePolicy](#)

이 정책은 Application Auto Scaling이 Amazon [AWSServiceRoleForApplicationAutoScaling_CassandraTable](#) Keyspace를 CloudWatch 호출하고 사용자를 대신하여 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책을 통해 Application Auto Scaling은 지정된 리소스에서 다음 작업을 완료할 수 있습니다.

- 조치: 다음 리소스에 `cassandra:Select` 대한 조치:
 - `arn:*:cassandra:*:*:/keyspace/system/table/*`
 - `arn:*:cassandra:*:*:/keyspace/system_schema/table/*`
 - `arn:*:cassandra:*:*:/keyspace/system_schema_mcs/table/*`
- 작업: 모든 리소스에 대한 `cassandra:Alter`
- 작업: 모든 리소스에 대한 `cloudwatch:DescribeAlarms`
- 작업: 모든 리소스에 대한 `cloudwatch:PutMetricAlarm`
- 작업: 모든 리소스에 대한 `cloudwatch>DeleteAlarms`

AWS 관리형 정책: Lambda 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingLambdaConcurrencyPolicy](#)

이 정책은 Application Auto Scaling이 [AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency](#) Lambda를 CloudWatch 호출하고 사용자를 대신하여 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `lambda:PutProvisionedConcurrencyConfig`
- 작업: `lambda:GetProvisionedConcurrencyConfig`
- 작업: `lambda>DeleteProvisionedConcurrencyConfig`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`
- 작업: `cloudwatch>DeleteAlarms`

AWS 관리형 정책: 아마존 MSK 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingKafkaClusterPolicy](#)

이 정책은 Application Auto Scaling이 Amazon [AWSServiceRoleForApplicationAutoScaling_KafkaCluster](#)MSK를 CloudWatch 호출하고 사용자를 대신하여 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `kafka:DescribeCluster`
- 작업: `kafka:DescribeClusterOperation`
- 작업: `kafka:UpdateBrokerStorage`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`
- 작업: `cloudwatch>DeleteAlarms`

AWS 관리형 정책: Neptune 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingNeptuneClusterPolicy](#)

이 정책은 Application Auto Scaling이

[AWSServiceRoleForApplicationAutoScaling_NeptuneCluster](#) Neptune을 CloudWatch 호출하고 사용자 대신 조정을 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책을 통해 Application Auto Scaling은 지정된 리소스에서 다음 작업을 완료할 수 있습니다.

- 작업: 모든 리소스에 대한 `rds:ListTagsForResource`
- 작업: 모든 리소스에 대한 `rds:DescribeDBInstances`
- 작업: 모든 리소스에 대한 `rds:DescribeDBClusters`
- 작업: 모든 리소스에 대한 `rds:DescribeDBClusterParameters`
- 작업: 모든 리소스에 대한 `cloudwatch:DescribeAlarms`
- 작업: Amazon Neptune 데이터베이스 엔진에 접두사 `autoscaled-reader`가 있는 리소스에 대해 `rds:AddTagsToResource("Condition":{"StringEquals":{"rds:DatabaseEngine":"neptune"}})`
- 작업: Amazon Neptune 데이터베이스 엔진의 모든 DB 클러스터(`rds:CreateDBInstance`)에 접두사 `autoscaled-reader`가 있는 리소스에 대해 `"Resource":"arn:*:rds:*:*:db:autoscaled-reader*", "arn:aws:rds:*:*:cluster:*"("Condition":{"StringEquals":{"rds:DatabaseEngine":"neptune"}})`
- 작업: `arn:aws:rds:*:*:db:autoscaled-reader*` 리소스에 대한 `rds>DeleteDBInstance`
- 작업: `arn:aws:cloudwatch:*:*:alarm:TargetTracking*` 리소스에 대한 `cloudwatch:PutMetricAlarm`
- 작업: `arn:aws:cloudwatch:*:*:alarm:TargetTracking*` 리소스에 대한 `cloudwatch>DeleteAlarms`

AWS 관리형 정책: SageMaker 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingSageMakerEndpointPolicy](#)

이 정책은 Application Auto Scaling이 사용자를

[AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint](#) 대신하여 조정을 SageMaker CloudWatch 호출하고 수행할 수 있도록 이름이 지정된 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책을 통해 Application Auto Scaling은 지정된 리소스에서 다음 작업을 완료할 수 있습니다.

- 작업: 모든 리소스에 대한 `sagemaker:DescribeEndpoint`
- 작업: 모든 리소스에 대한 `sagemaker:DescribeEndpointConfig`
- 작업: 모든 리소스에 대한 `sagemaker:DescribeInferenceComponent`
- 작업: 모든 리소스에 대한 `sagemaker:UpdateEndpointWeightsAndCapacities`
- 작업: 모든 리소스에 대한 `sagemaker:UpdateInferenceComponentRuntimeConfig`
- 작업: 모든 리소스에 대한 `cloudwatch:DescribeAlarms`
- 작업: 모든 리소스에 대한 `cloudwatch:GetMetricData`
- 작업: `arn:aws:cloudwatch:*:*:alarm:TargetTracking*` 리소스에 대한 `cloudwatch:PutMetricAlarm`
- 작업: `arn:aws:cloudwatch:*:*:alarm:TargetTracking*` 리소스에 대한 `cloudwatch>DeleteAlarms`

AWS 관리형 정책: EC2 스팟 플릿 및 CloudWatch

정책 이름: [AWSApplicationAutoscalingEC2SpotFleetRequestPolicy](#)

이 정책은 Application Auto Scaling이 Amazon [AWSServiceRoleForApplicationAutoScaling_ECEC2](#) 를 호출하고 사용자를 대신하여 조정을 수행할 수 있도록 하기 위한 서비스 연결 역할인 `2 SpotFleet CloudWatch` 요청과 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `ec2:DescribeSpotFleetRequests`
- 작업: `ec2:ModifySpotFleetRequest`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`
- 작업: `cloudwatch>DeleteAlarms`

AWS 관리형 정책: 사용자 지정 리소스 및 CloudWatch

정책 이름: [AWSApplicationAutoScalingCustomResourcePolicy](#)

이 정책은 Application Auto Scaling이 API Gateway를 통해 사용 가능한 사용자 지정 리소스를 CloudWatch 호출하고 사용자 대신 조정을 수행할 수 있도록 이름이 지정된 [AWSServiceRoleForApplicationAutoScaling_CustomResource](#) 서비스 연결 역할에 연결됩니다.

권한 세부 정보

권한 정책에 따라 Application Auto Scaling은 모든 관련 리소스 (“리소스”: “*”) 에서 다음 작업을 완료할 수 있습니다.

- 작업: `execute-api:Invoke`
- 작업: `cloudwatch:DescribeAlarms`
- 작업: `cloudwatch:PutMetricAlarm`
- 작업: `cloudwatch>DeleteAlarms`

Application Auto Scaling에서 AWS 관리형 정책을 업데이트했습니다.

이 서비스가 이러한 변경 사항을 추적하기 시작한 이후 Application Auto Scaling의 AWS 관리형 정책 업데이트에 대한 세부 정보를 확인하십시오. 이 페이지의 변경 사항에 대한 자동 알림을 받아보려면 Application Auto Scaling 문서 기록 페이지에서 RSS 피드를 구독하세요.

변경 사항	설명	날짜
Application Auto Scaling은 SageMaker 서비스 연결 역할에 권한을 추가합니다.	이제 이 정책은 향후 통합을 위해 SageMaker 리소스의 자동 크기 조정을 위한 호환성을 지원하는 UpdateInferenceComponentRuntimeConfig API 작업 SageMaker DescribeInferenceComponent 및 API 작업을 호출할 수 있는 권한을 서비스에 부여합니다. 또한 이 정책은 이제 CloudWatch PutMetricAlarm 및 DeleteAlarms API 작업을 대상 추적 조정 정책과 함께 사	2023년 11월 13일

변경 사항	설명	날짜
	용되는 CloudWatch 경보로만 제한합니다.	
Application Auto Scaling에 Neptune 정책 추가	Application Auto Scaling은 Neptune에 대한 새로운 관리형 정책을 추가했습니다. 이 정책은 Application Auto Scaling이 Neptune을 CloudWatch 호출하고 사용자 대신 조정을 수행하도록 허용하는 서비스 연결 역할에 연결됩니다.	2021년 10월 6일
Application Auto Scaling은 Redis 정책을 추가합니다 ElastiCache .	Application Auto Scaling에서 에 대한 새 관리형 정책을 추가했습니다 ElastiCache. 이 정책은 Application Auto Scaling 이 사용자를 대신하여 조정을 ElastiCache CloudWatch 호출하고 수행할 수 있도록 하는 서비스 연결 역할에 연결됩니다.	2021년 8월 19일
Application Auto Scaling에서 변경 내용 추적 시작	Application Auto Scaling은 AWS 관리형 정책의 변경 사항을 추적하기 시작했습니다.	2021년 8월 19일

Application Auto Scaling에 대한 서비스 연결 역할

Application Auto Scaling은 사용자를 대신하여 다른 AWS 서비스를 호출하는 데 필요한 권한에 대해 [서비스 연결 역할](#)을 사용합니다. 서비스 연결 역할은 서비스에 직접 연결된 고유한 유형의 AWS Identity and Access Management (IAM) 역할입니다. AWS 서비스 연결 역할은 연결된 서비스만 서비스 연결 역할을 맡을 수 있기 때문에 AWS 서비스에 권한을 위임할 수 있는 안전한 방법을 제공합니다.

Application Auto Scaling과 통합되는 서비스의 경우 Application Auto Scaling은 사용자를 위한 서비스 연결 역할을 생성합니다. 각 서비스에는 서비스 연결 역할이 하나씩 있습니다. 각각의 서비스 연결 역할은 그 역할을 맡도록 지정된 서비스 주체를 신뢰합니다. 자세한 내용은 [서비스 연결 역할 ARN 참조](#) 섹션을 참조하세요.

Application Auto Scaling은 서비스 연결 역할에 필요한 모든 권한을 포함합니다. 이러한 관리형 권한은 Application Auto Scaling에서 생성 및 관리되며 각 리소스 유형에 대해 허용되는 작업을 정의합니다. 각 역할이 부여하는 권한에 대한 자세한 내용은 [AWS Application Auto Scaling에 대한 관리형 정책](#) 섹션을 참조하세요.

내용

- [서비스 연결 역할 생성에 필요한 권한](#)
- [서비스 연결 역할 생성\(자동\)](#)
- [서비스 연결 역할 생성\(수동\)](#)
- [서비스 연결 역할 편집](#)
- [서비스 연결 역할 삭제](#)
- [Application Auto Scaling 서비스 연결 역할 지원 리전](#)
- [서비스 연결 역할 ARN 참조](#)

서비스 연결 역할 생성에 필요한 권한

Application Auto Scaling을 사용하려면 사용자가 처음으로 특정 서비스를 AWS 계정 RegisterScalableTarget 호출할 때 서비스 연결 역할을 생성할 수 있는 권한이 필요합니다. 역할이 이미 존재하지 않으면, Application Auto Scaling이 사용자의 계정에 대상 서비스에 대한 서비스 연결 역할을 생성합니다. 서비스 연결 역할은 사용자를 대신하여 대상 서비스를 호출할 수 있도록 Application Auto Scaling에 권한을 부여합니다.

역할 자동 생성이 성공하려면 사용자가 iam:CreateServiceLinkedRole 작업에 대한 권한을 보유해야 합니다.

```
"Action": "iam:CreateServiceLinkedRole"
```

다음은 스팟 플릿에 대한 서비스 연결 역할을 생성하는 권한을 부여하는 자격 증명 기반 정책입니다. 다음과 같이 정책 Resource 필드의 서비스 연결 역할을 ARN으로 지정하고 서비스 연결 역할에 대한 서비스 보안 주체를 조건으로 지정할 수 있습니다. 각 서비스에 대한 ARN은 [서비스 연결 역할 ARN 참조](#) 섹션을 참조하세요.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
```

```

    "Action": "iam:CreateServiceLinkedRole",
    "Resource": "arn:aws:iam::*:role/aws-service-role/ec2.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest",
    "Condition": {
      "StringLike": {
        "iam:AWSServiceName": "ec2.application-autoscaling.amazonaws.com"
      }
    }
  ]
}

```

Note

iam:AWSServiceName IAM 조건 키는 역할이 연결된 서비스 보안 주체(이 예제 정책에서는 *ec2.application-autoscaling.amazonaws.com*)를 지정합니다. 서비스 보안 주체를 알기 어렵습니다. 서비스에 대한 서비스 보안 주체를 보려면 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#) 섹션을 참조하세요.

서비스 연결 역할 생성(자동)

서비스 링크 역할은 수동으로 생성할 필요가 없습니다. RegisterScalableTarget을 호출하면 Application Auto Scaling이 적절한 서비스 연결 역할을 자동으로 생성합니다. 예를 들어 Amazon ECS 서비스에 대해 자동 조정을 설정하면 Application Auto Scaling이 AWSServiceRoleForApplicationAutoScaling_ECSService 역할을 생성합니다.

서비스 연결 역할 생성(수동)

서비스 연결 역할을 생성하려면 IAM 콘솔 또는 IAM API를 사용할 수 있습니다. AWS CLI 자세한 내용은 IAM 사용 설명서의 [서비스 연결 역할 생성](#) 섹션을 참조하세요.

서비스 연결 역할을 생성하려면(AWS CLI)

다음 [create-service-linked-role](#) CLI 명령을 사용하여 Application Auto Scaling 서비스 연결 역할을 생성합니다. 요청에서 서비스 이름 “prefix”를 지정합니다.

서비스 이름 접두사를 찾으려면 [AWS 서비스 Application Auto Scaling과 함께 사용할 수 있는](#) 단원에서 각 서비스에 대한 서비스 연결 역할의 서비스 보안 주체에 대한 정보를 참조하세요. 서비스 이름과 서비스 보안 주체는 동일한 접두사를 공유합니다. 예를 들어, AWS Lambda 서비스 연결 역할을 생성하려면 `lambda.application-autoscaling.amazonaws.com`

```
aws iam create-service-linked-role --aws-service-name prefix.application-
autoscaling.amazonaws.com
```

서비스 연결 역할 편집

Application Auto Scaling이 생성한 서비스 연결 역할로는 설명만 편집할 수 있습니다. 자세한 내용은 IAM 사용 설명서의 [서비스 연결 역할 편집](#)을 참조하세요.

서비스 연결 역할 삭제

지원되는 서비스에서 Application Auto Scaling을 더 이상 사용하지 않는 경우 해당 서비스 연결 역할을 삭제하는 것이 좋습니다.

먼저 관련 AWS 리소스를 삭제한 후에만 서비스 연결 역할을 삭제할 수 있습니다. 따라서 리소스에 대한 Application Auto Scaling 권한을 실수로 취소하는 것을 방지할 수 있습니다. 자세한 정보는 확장 가능한 리소스의 [설명서](#)를 참조하세요. 예를 들어, Amazon ECS 서비스를 삭제하려면 Amazon Elastic Container Service 개발자 안내서의 [서비스 삭제](#) 섹션을 참조하세요.

IAM을 사용하여 서비스 연결 역할을 삭제할 수 있습니다. 자세한 내용은 [IAM 사용 설명서](#)의 서비스에 연결 역할 삭제 섹션을 참조하세요.

서비스 연결 역할을 삭제한 후 RegisterScalableTarget을 호출하면 Application Auto Scaling이 그 역할을 다시 생성합니다.

Application Auto Scaling 서비스 연결 역할 지원 리전

Application Auto Scaling은 서비스를 사용할 수 있는 모든 AWS 지역에서 서비스 연결 역할을 사용할 수 있도록 지원합니다.

서비스 연결 역할 ARN 참조

다음 표에는 Application Auto Scaling과 함께 AWS 서비스 작동하는 각 서비스 연결 역할의 Amazon 리소스 이름 (ARN) 이 나와 있습니다.

Service	ARN
AppStream 2.0	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/appstream.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_AppStreamFleet

Service	ARN
Aurora	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/rds.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_RDSCluster
Comprehend	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/comprehend.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint
DynamoDB	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/dynamodb.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_DynamoDBTable
ECS	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/ecs.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ECSService
ElastiCache	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/elasticache.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG
Keyspaces	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/cassandra.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_CassandraTable
Lambda	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/lambda.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency
MSK	arn:aws:iam:: <i>012345678910</i> :role/aws-service-role/kafka.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_KafkaCluster

Service	ARN
Neptune	arn:aws:iam:: 012345678910 :role/aws-service-role/neptune.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_NeptuneCluster
SageMaker	arn:aws:iam:: 012345678910 :role/aws-service-role/sagemaker.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint
Spot Fleets	arn:aws:iam:: 012345678910 :role/aws-service-role/ec2.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest
사용자 정의 리소스	arn:aws:iam:: 012345678910 :role/aws-service-role/custom-resource.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_CustomResource

Note

지정된 서비스 연결 역할이 아직 없더라도 AWS CloudFormation 스택 템플릿의 [AWS::ApplicationAutoScaling::ScalableTarget](#) 리소스 RoleARN 속성에 대해 서비스 연결 역할의 ARN을 지정할 수 있습니다. Application Auto Scaling이 자동으로 역할을 생성합니다.

Application Auto Scaling 자격 증명 기반 정책 예제

기본적으로 새로 가입한 사용자는 아무 것도 할 수 있는 AWS 계정 권한이 없습니다. IAM 관리자는 Application Auto Scaling API 작업을 수행할 수 있는 IAM 자격 증명(예: 사용자 또는 역할) 권한을 부여하는 IAM 정책을 생성하고 할당해야 합니다.

다음 예제 JSON 정책 문서를 사용하여 IAM 정책을 생성하는 방법을 알아보려면 IAM 사용 설명서의 [JSON 탭에서 정책 생성](#)을 참조하세요.

내용

- [Application Auto Scaling API 작업에 필요한 권한](#)
- [대상 서비스에 대한 API 작업에 필요한 권한 및 CloudWatch](#)
- [에서 작업할 수 있는 권한 AWS Management Console](#)

Application Auto Scaling API 작업에 필요한 권한

다음 정책은 Application Auto Scaling API를 호출할 때 일반적인 사용 사례에 대한 권한을 부여합니다. 자격 증명 기반 정책을 작성할 때 이 섹션을 참조하세요. 각 정책은 Application Auto Scaling API 작업의 모두 또는 일부에 대한 권한을 부여합니다. 또한 최종 사용자에게 대상 서비스 및 CloudWatch (자세한 내용은 다음 섹션 참조)에 대한 권한이 있는지 확인해야 합니다.

다음 자격 증명 기반 정책에서는 모든 Application Auto Scaling API 작업에 대한 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "application-autoscaling:*"
      ],
      "Resource": "*"
    }
  ]
}
```

다음 자격 증명 기반 정책에서는 예약된 작업이 아니라 확장 정책을 구성하는 데 필요한 모든 Application Auto Scaling API 작업에 대한 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "application-autoscaling:RegisterScalableTarget",
        "application-autoscaling:DescribeScalableTargets",
        "application-autoscaling:DeregisterScalableTarget",

```

```

        "application-autoscaling:PutScalingPolicy",
        "application-autoscaling:DescribeScalingPolicies",
        "application-autoscaling:DescribeScalingActivities",
        "application-autoscaling>DeleteScalingPolicy"
    ],
    "Resource": "*"
}
]
}

```

다음 자격 증명 기반 정책에서는 확장 정책이 아니라 예약된 작업을 구성하는 데 필요한 모든 Application Auto Scaling API 작업에 대한 권한을 부여합니다.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "application-autoscaling:RegisterScalableTarget",
        "application-autoscaling:DescribeScalableTargets",
        "application-autoscaling:DeregisterScalableTarget",
        "application-autoscaling:PutScheduledAction",
        "application-autoscaling:DescribeScheduledActions",
        "application-autoscaling:DescribeScalingActivities",
        "application-autoscaling>DeleteScheduledAction"
      ],
      "Resource": "*"
    }
  ]
}

```

대상 서비스에 대한 API 작업에 필요한 권한 및 CloudWatch

대상 서비스에서 Application Auto Scaling을 성공적으로 구성하고 사용하려면 최종 사용자에게 Amazon CloudWatch 및 조정을 구성할 각 대상 서비스에 대한 권한을 부여해야 합니다. 다음 정책을 사용하여 대상 서비스 및 작업에 필요한 최소 권한을 CloudWatch 부여하십시오.

내용

- [AppStream 2.0 플릿](#)
- [Aurora 복제본](#)

- [Amazon Comprehend 문서 분류 및 엔터티 인식기 엔드포인트](#)
- [DynamoDB 테이블 및 글로벌 보조 인덱스](#)
- [ECS 서비스](#)
- [ElastiCache 복제 그룹](#)
- [Amazon EMR 클러스터](#)
- [Amazon Keyspaces 테이블](#)
- [Lambda 함수](#)
- [Amazon Managed Streaming for Apache Kafka\(MSK\) 브로커 스토리지](#)
- [Neptune 클러스터](#)
- [SageMaker 엔드포인트](#)
- [스팟 플릿\(Amazon EC2\)](#)
- [사용자 정의 리소스](#)

AppStream 2.0 플릿

다음 ID 기반 정책은 필요한 모든 AppStream 2.0 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "appstream:DescribeFleets",
        "appstream:UpdateFleet",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

Aurora 복제본

다음 ID 기반 정책은 필요한 모든 Aurora 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "rds:AddTagsToResource",
        "rds:CreateDBInstance",
        "rds>DeleteDBInstance",
        "rds:DescribeDBClusters",
        "rds:DescribeDBInstances",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

Amazon Comprehend 문서 분류 및 엔터티 인식기 엔드포인트

다음 자격 증명 기반 정책은 필요한 모든 Amazon Comprehend CloudWatch 및 API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "comprehend:UpdateEndpoint",
        "comprehend:DescribeEndpoint",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

DynamoDB 테이블 및 글로벌 보조 인덱스

다음 ID 기반 정책은 필요한 모든 DynamoDB 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "dynamodb:DescribeTable",
        "dynamodb:UpdateTable",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

ECS 서비스

다음 자격 증명 기반 정책은 필요한 모든 ECS 및 API 작업에 권한을 부여합니다. CloudWatch

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "ecs:DescribeServices",
        "ecs:UpdateService",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

ElastiCache 복제 그룹

다음 ID 기반 정책은 필요한 모든 ElastiCache 작업과 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "elasticache:ModifyReplicationGroupShardConfiguration",
        "elasticache:IncreaseReplicaCount",
        "elasticache:DecreaseReplicaCount",
        "elasticache:DescribeReplicationGroups",
        "elasticache:DescribeCacheClusters",
        "elasticache:DescribeCacheParameters",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

Amazon EMR 클러스터

다음 ID 기반 정책은 필요한 모든 Amazon EMR 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "elasticmapreduce:ModifyInstanceGroups",
        "elasticmapreduce:ListInstanceGroups",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

```
]
}
```

Amazon Keyspaces 테이블

다음 ID 기반 정책은 필요한 모든 Amazon Keyspace와 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "cassandra:Select",
        "cassandra:Alter",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

Lambda 함수

다음 ID 기반 정책은 필요한 모든 Lambda 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "lambda:PutProvisionedConcurrencyConfig",
        "lambda:GetProvisionedConcurrencyConfig",
        "lambda>DeleteProvisionedConcurrencyConfig",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```



```
]
}
```

Amazon Managed Streaming for Apache Kafka(MSK) 브로커 스토리지

다음 ID 기반 정책은 필요한 모든 Amazon MSK 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "kafka:DescribeCluster",
        "kafka:DescribeClusterOperation",
        "kafka:UpdateBrokerStorage",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}
```

Neptune 클러스터

다음 ID 기반 정책은 필요한 모든 Neptune 및 CloudWatch API 작업에 권한을 부여합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "rds:AddTagsToResource",
        "rds:CreateDBInstance",
        "rds:DescribeDBInstances",
        "rds:DescribeDBClusters",
        "rds:DescribeDBClusterParameters",
        "rds>DeleteDBInstance",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",

```

```

        "cloudwatch:DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}

```

SageMaker 엔드포인트

다음 ID 기반 정책은 필요한 모든 SageMaker 작업과 CloudWatch API 작업에 권한을 부여합니다.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "sagemaker:DescribeEndpoint",
        "sagemaker:DescribeEndpointConfig",
        "sagemaker:DescribeInferenceComponent",
        "sagemaker:UpdateEndpointWeightsAndCapacities",
        "sagemaker:UpdateInferenceComponentRuntimeConfig",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch:DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}

```

스팟 플릿(Amazon EC2)

다음 ID 기반 정책은 필요한 모든 스팟 플릿 및 CloudWatch API 작업에 권한을 부여합니다.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "ec2:DescribeSpotFleetRequests",
        "ec2:ModifySpotFleetRequest",

```

```

        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
    ],
    "Resource": "*"
}
]
}

```

사용자 정의 리소스

다음 자격 증명 기반 정책에서는 API Gateway API 실행 작업에 대한 권한을 부여합니다. 또한 이 정책은 필요한 모든 CloudWatch 작업에 대한 권한을 부여합니다.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "execute-api:Invoke",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch>DeleteAlarms"
      ],
      "Resource": "*"
    }
  ]
}

```

에서 작업할 수 있는 권한 AWS Management Console

독립형 Application Auto Scaling 콘솔은 없습니다. Application Auto Scaling과 통합되는 대부분의 서비스에는 콘솔에서 조정을 구성하는 데 도움이 되는 전용 기능이 있습니다.

대부분의 경우 각 서비스는 Application Auto Scaling API 작업에 대한 권한을 포함하여 콘솔에 대한 액세스를 정의하는 AWS 관리형 (사전 정의된) IAM 정책을 제공합니다. 자세한 내용은 콘솔을 사용할 서비스에 대한 설명서를 참조하세요.

또한 사용자 지정 IAM 정책을 생성하여 AWS Management Console에서 특정 Application Auto Scaling API 작업을 보고 작업할 수 있는 세분화된 권한을 사용자에게 제공할 수 있습니다. 이전 섹션의 예제 정책을 사용할 수 있지만, 이 예제는 AWS CLI 또는 SDK를 사용하여 이루어진 요청에 맞게 설

계되었습니다. 콘솔에서는 추가적인 API 작업을 통해 해당 기능을 구현하므로 이러한 정책이 예상과 다르게 작동할 수 있습니다. 예를 들어 단계별 조정을 구성하려면 사용자에게 CloudWatch 경보를 만들고 관리할 수 있는 추가 권한이 필요할 수 있습니다.

Tip

콘솔에서 작업을 수행하는 데 필요한 API 작업을 파악하려는 경우 AWS CloudTrail 등의 서비스를 사용할 수 있습니다. 자세한 내용은 [AWS CloudTrail 사용 설명서](#)를 참조하세요.

다음 자격 증명 기반 정책에서는 스팟 플릿에 대한 확장 정책을 구성하는 권한을 부여합니다. 스팟 플릿에 대한 IAM 권한 외에, Amazon EC2 콘솔에서 플릿 확장 설정에 액세스하는 콘솔 사용자에게 동적 확장을 지원하는 서비스에 대한 적절한 권한이 있어야 합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "application-autoscaling:*",
        "ec2:DescribeSpotFleetRequests",
        "ec2:ModifySpotFleetRequest",
        "cloudwatch:DeleteAlarms",
        "cloudwatch:DescribeAlarmHistory",
        "cloudwatch:DescribeAlarms",
        "cloudwatch:DescribeAlarmsForMetric",
        "cloudwatch:GetMetricStatistics",
        "cloudwatch:ListMetrics",
        "cloudwatch:PutMetricAlarm",
        "cloudwatch:DisableAlarmActions",
        "cloudwatch:EnableAlarmActions",
        "sns:CreateTopic",
        "sns:Subscribe",
        "sns:Get*",
        "sns:List*"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": "iam:CreateServiceLinkedRole",
```

```

    "Resource": "arn:aws:iam::*:role/aws-service-role/ec2.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest",
    "Condition": {
      "StringLike": {
        "iam:AWSServiceName": "ec2.application-autoscaling.amazonaws.com"
      }
    }
  }
]
}

```

이 정책을 통해 콘솔 사용자는 Amazon EC2 콘솔에서 조정 정책을 확인 및 수정하고 콘솔에서 CloudWatch 경보를 생성 및 관리할 수 있습니다. CloudWatch

API 작업을 조정하여 사용자 액세스를 제한할 수 있습니다. 예를 들어, application-autoscaling:*을 application-autoscaling:Describe*로 바꾸면 사용자는 읽기 전용 액세스 권한을 갖게 됩니다.

또한 필요에 따라 CloudWatch 권한을 조정하여 기능에 대한 사용자 액세스를 제한할 수 있습니다. CloudWatch 자세한 내용은 Amazon 사용 [설명서의 CloudWatch 콘솔에 필요한 CloudWatch 권한을](#) 참조하십시오.

Application Auto Scaling 액세스 문제 해결

AccessDeniedException 또는 Application Auto Scaling으로 작업할 때 이와 유사한 어려움이 있으면 이 단원의 정보를 참조하세요.

Application Auto Scaling에서 작업을 수행할 권한이 없음

AWS API 작업을 AccessDeniedException 호출할 때 메시지가 표시되면 사용 중인 AWS Identity and Access Management (IAM) 자격 증명에 해당 호출을 수행하는 데 필요한 권한이 없는 것입니다.

다음 예제 오류는 mateojackson 사용자가 확장 가능 대상에 대한 세부 정보를 보려고 하지만, application-autoscaling:DescribeScalableTargets 권한이 없는 경우에 발생합니다.

```

An error occurred (AccessDeniedException) when calling the DescribeScalableTargets
operation: User: arn:aws:iam::123456789012:user/mateojackson is not authorized to
perform: application-autoscaling:DescribeScalableTargets

```

이 오류 또는 이와 유사한 오류가 발생하면 관리자에게 문의하여 도움을 받아야 합니다.

계정 관리자는 Application Auto Scaling에서 대상 서비스 및 리소스에 액세스하는 데 사용하는 모든 API 작업에 액세스할 수 있는 권한이 있는지 확인해야 CloudWatch 합니다. 작업 중인 리소스에 따라 다른 권한이 필요합니다. 또한 Application Auto Scaling에서는 사용자가 해당 리소스에 대해 조정을 처음 구성할 때 서비스 연결 역할을 생성할 수 있는 권한이 필요합니다.

관리자인데, IAM 정책에서 오류를 반환했거나 예상대로 작동하지 않음

Application Auto Scaling 작업 외에도 IAM 정책은 대상 서비스 및 CloudWatch 를 호출할 수 있는 권한을 부여해야 합니다. 사용자 또는 애플리케이션에 이러한 추가 권한이 없는 경우 액세스가 예기치 않게 거부될 수 있습니다. 계정에서 사용자 및 애플리케이션에 대한 IAM 정책을 작성하려면 [Application Auto Scaling 자격 증명 기반 정책 예제](#)의 정보를 참조하세요.

유효성 검사가 수행되는 방법에 대한 자세한 내용은 [대상 리소스에 대한 Application Auto Scaling API 호출에 대한 권한 검증](#) 섹션을 참조하세요.

일부 권한 문제는 Application Auto Scaling에서 사용하는 서비스 연결 역할을 생성하는 문제로 인해 발생할 수도 있습니다. 이러한 서비스 연결 역할 생성에 대한 자세한 내용은 [Application Auto Scaling에 대한 서비스 연결 역할](#) 섹션을 참조하세요.

대상 리소스에 대한 Application Auto Scaling API 호출에 대한 권한 검증

Application Auto Scaling API 작업에 대해 승인된 요청을 하려면 API 호출자에게 대상 서비스 및 내부 AWS 리소스에 액세스할 수 있는 권한이 있어야 합니다. CloudWatch Application Auto Scaling은 요청을 CloudWatch 진행하기 전에 대상 서비스와 관련된 요청에 대한 권한을 검증합니다. 이를 위해 일련의 호출을 실행하여 대상 리소스에 대한 IAM 권한을 검증합니다. 응답이 반환되면 Application Auto Scaling이 읽습니다. IAM 권한이 지정된 작업을 허용하지 않는 경우 Application Auto Scaling은 요청에 실패하고 누락된 권한에 대한 정보가 포함된 오류를 사용자에게 반환합니다. 이렇게 하면 사용자가 배포하려는 조정 구성이 의도한 대로 작동하고 요청이 실패하면 유용한 오류가 반환됩니다.

작동 방식의 예로, 다음 정보는 Application Auto Scaling에서 CloudWatch Aurora 및 를 사용하여 권한 검증을 수행하는 방법에 대한 세부 정보를 제공합니다.

사용자가 Aurora DB 클러스터에 대해 RegisterScalableTarget API를 호출하는 경우 Application Auto Scaling에서는 다음과 같은 검사를 모두 수행하여 IAM 사용자에게 필수 권한(굵게 표시)이 있는지 확인합니다.

- **rds:CreateDBInstance**: 사용자에게 이 권한이 있는지를 확인하기 위해 CreateDBInstance API 작업으로 요청을 보내, 사용자가 지정한 Aurora DB 클러스터에 잘못된 파라미터(빈 인스턴스 ID)가 있는 DB 인스턴스를 생성하려고 합니다. 권한이 부여된 사용자의 경우 API가 요청을 감사한

후 `InvalidParameterValue` 오류 코드 응답을 반환합니다. 그러나 권한이 없는 사용자의 경우 `AccessDenied` 오류가 발생하고 Application Auto Scaling 요청에 실패합니다. 누락된 권한이 나열된 `ValidationException` 오류가 사용자에게 제공됩니다.

- `rds:DeleteDBInstance`: 빈 인스턴스 ID를 `DeleteDBInstance` API 작업에 보냅니다. 권한이 부여된 사용자의 경우 이 요청으로 `InvalidParameterValue` 오류가 발생합니다. 권한이 없는 사용자의 경우 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다(첫 번째 글머리 기호에 설명된 것과 동일한 처리).
- `rds: AddTags ToResource`: `AddTagsToResource` API 작업에는 Amazon 리소스 이름 (ARN) 이 필요하므로 잘못된 계정 ID (12345) 와 더미 인스턴스 ID (존재하지 않는 db) 를 사용하여 “더미” 리소스를 지정하여 ARN () 을 구성해야 합니다. `arn:aws:rds:us-east-1:12345:db:non-existing-db` 권한이 부여된 사용자의 경우 이 요청으로 `InvalidParameterValue` 오류가 발생합니다. 권한이 없는 사용자의 경우 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다.
- `rds:DescribeDBCluster`: Auto Scaling을 위해 등록되는 리소스의 클러스터 이름을 설명합니다. 권한이 부여된 사용자의 경우 유효한 설명 결과를 얻습니다. 권한이 없는 사용자의 경우 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다.
- `rds:DescribeDBInstance`. 확장 가능한 대상을 등록하기 위해 사용자가 제공한 클러스터 이름을 필터링하는 `db-cluster-id` 필터로 `DescribeDBInstance` API를 호출합니다. 권한이 부여된 사용자의 경우 DB 클러스터의 모든 DB 인스턴스를 설명할 수 있습니다. 권한이 없는 사용자의 경우 이 호출로 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다.
- 클라우드워치: 알람: 매개변수 없이 API를 호출합니다. `PutMetric PutMetricAlarm` 경보 이름이 누락되기 때문에 요청으로 권한이 부여된 사용자에게 `ValidationError`가 발생합니다. 권한이 없는 사용자의 경우 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다.
- `cloudwatch: DescribeAlarms`: 최대 레코드 수 값을 1로 설정하여 `DescribeAlarms` API를 호출합니다. 권한 있는 사용자의 경우 응답에서 하나의 경보에 대한 정보가 예상됩니다. 권한이 없는 사용자의 경우 이 호출로 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다.
- `cloudwatch:DeleteAlarms`: `PutMetricAlarm` 위와 마찬가지로 요청할 파라미터는 제공하지 않습니다. `DeleteAlarms` 요청에서 경보 이름이 누락되었기 때문에 이 호출은 권한이 부여된 사용자에게 `ValidationError`로 실패합니다. 권한이 없는 사용자의 경우 `AccessDenied`가 발생하고 사용자에게 유효성 검사 예외를 보냅니다.

이러한 유효성 검사 예외 중 하나가 발생할 때마다 기록됩니다. 를 사용하여 검증에 실패한 호출을 수동으로 식별하는 조치를 취할 수 있습니다. AWS CloudTrail자세한 내용은 [AWS CloudTrail 사용 설명서](#)를 참조하세요.

Note

를 사용하여 CloudTrail Application Auto Scaling 이벤트에 대한 알림을 받는 경우 이러한 알림에는 기본적으로 사용자 권한을 검증하기 위한 Application Auto Scaling 호출이 포함됩니다. 이러한 알림을 필터링하려면 이러한 유효성 검사를 위한 `application-autoscaling.amazonaws.com`을 포함할 `invokedBy` 필드를 사용합니다.

인터페이스 VPC 엔드포인트를 사용하여 Application Auto Scaling에 액세스

를 사용하여 AWS PrivateLink VPC와 Application Auto Scaling 간에 프라이빗 연결을 생성할 수 있습니다. 인터넷 게이트웨이, NAT 디바이스, VPN 연결 또는 연결을 사용하지 않고도 마치 VPC에 있는 것처럼 Application Auto Scaling에 액세스할 수 있습니다. AWS Direct Connect VPC의 인스턴스는 Application Auto Scaling에 액세스하는 데 퍼블릭 IP 주소가 필요하지 않습니다.

AWS PrivateLink에서 제공되는 인터페이스 엔드포인트를 생성하여 이 프라이빗 연결을 설정합니다. 인터페이스 엔드포인트에 대해 사용 설정하는 각 서브넷에서 엔드포인트 네트워크 인터페이스를 생성합니다. 이는 Application Auto Scaling으로 향하는 트래픽의 진입점 역할을 하는 요청자 관리 네트워크 인터페이스입니다.

자세한 내용은 가이드의 [액세스를 참조하십시오 AWS 서비스 . AWS PrivateLink](#) AWS PrivateLink

내용

- [인터페이스 VPC 엔드포인트 생성](#)
- [VPC 엔드포인트 정책 생성](#)

인터페이스 VPC 엔드포인트 생성

다음 서비스 이름을 사용하여 Application Auto Scaling에 대한 엔드포인트를 생성합니다.

```
com.amazonaws.region.application-autoscaling
```

자세한 내용은 가이드의 [인터페이스 VPC 엔드포인트를 사용한 AWS 서비스 액세스](#)를 참조하십시오. AWS PrivateLink

다른 설정은 변경할 필요가 없습니다. Application Auto Scaling은 서비스 엔드포인트 또는 프라이빗 인터페이스 VPC 엔드포인트 중 사용 중인 엔드포인트를 사용하여 다른 AWS 서비스를 호출합니다.

VPC 엔드포인트 정책 생성

VPC 엔드포인트에 정책을 연결하여 Application Auto Scaling API에 대한 액세스를 제어할 수 있습니다. 이 정책은 다음을 지정합니다.

- 작업을 수행할 수 있는 보안 주체.
- 수행할 수 있는 작업.
- 작업을 수행할 수 있는 리소스.

다음 예에서는 엔드포인트를 통해 조정 정책을 삭제할 수 있는 모든 사용자 권한을 거부하는 VPC 엔드포인트 정책을 보여줍니다. 또한 이 정책 예에서는 모든 사용자에게 다른 모든 작업을 수행할 수 있는 권한을 부여합니다.

```
{
  "Statement": [
    {
      "Action": "*",
      "Effect": "Allow",
      "Resource": "*",
      "Principal": "*"
    },
    {
      "Action": "application-autoscaling:DeleteScalingPolicy",
      "Effect": "Deny",
      "Resource": "*",
      "Principal": "*"
    }
  ]
}
```

자세한 정보는 AWS PrivateLink 가이드의 [VPC 엔드포인트 정책](#)을 참조하세요.

Application Auto Scaling의 복원성

AWS 글로벌 인프라는 AWS 지역 및 가용 영역을 중심으로 구축됩니다.

AWS 지역은 물리적으로 분리되고 격리된 여러 가용 영역을 제공하며, 이러한 가용 영역은 지연 시간이 짧고 처리량이 높으며 중복성이 높은 네트워킹으로 연결됩니다.

가용 영역을 사용하면 중단 없이 영역 간에 자동으로 장애 극복 조치가 이루어지는 애플리케이션 및 데이터베이스를 설계하고 운영할 수 있습니다. 가용 영역은 기존의 단일 또는 다중 데이터 센터 인프라보다 가용성, 내결함성, 확장성이 뛰어납니다.

AWS [지역 및 가용 영역에 대한 자세한 내용은 글로벌 인프라를 참조하십시오AWS](#).

Application Auto Scaling의 인프라 보안

Application Auto Scaling은 매니지드 서비스로서 AWS 글로벌 네트워크 보안을 통해 보호됩니다. AWS 보안 서비스 및 인프라 AWS 보호 방법에 대한 자세한 내용은 [AWS 클라우드 보안을](#) 참조하십시오. 인프라 보안 모범 사례를 사용하여 AWS 환경을 설계하려면 Security Pillar AWS Well-Architected Framework의 [인프라 보호](#)를 참조하십시오.

AWS 게시된 API 호출을 사용하여 네트워크를 통해 Application Auto Scaling에 액세스할 수 있습니다. 고객은 다음을 지원해야 합니다.

- 전송 계층 보안(TLS) TLS 1.2는 필수이며 TLS 1.3을 권장합니다.
- DHE(Ephemeral Diffie-Hellman) 또는 ECDHE(Elliptic Curve Ephemeral Diffie-Hellman)와 같은 완전 전송 보안(PFS)이 포함된 암호 제품군 Java 7 이상의 최신 시스템은 대부분 이러한 모드를 지원합니다.

또한 요청은 액세스 키 ID 및 IAM 주체와 관련된 비밀 액세스 키를 사용하여 서명해야 합니다. 또는 [AWS Security Token Service\(AWS STS\)](#)를 사용하여 임시 보안 인증을 생성하여 요청에 서명할 수 있습니다.

Application Auto Scaling의 규정 준수 확인

특정 규정 준수 프로그램의 범위 내에 AWS 서비스 있는지 알아보려면AWS 서비스 규정 준수 [프로그램의AWS 서비스 범위별, 규정](#) 참조하여 관심 있는 규정 준수 프로그램을 선택하십시오. 일반 정보는 [AWS 규정 준수 프로그램AWS 보증 프로그램 규정AWS](#) 참조하십시오.

를 사용하여 AWS Artifact타사 감사 보고서를 다운로드할 수 있습니다. 자세한 내용은 의 보고서 <https://docs.aws.amazon.com/artifact/latest/ug/downloading-documents.html> 참조하십시오 AWS Artifact.

사용 시 규정 준수 AWS 서비스 책임은 데이터의 민감도, 회사의 규정 준수 목표, 관련 법률 및 규정에 따라 결정됩니다. AWS 규정 준수에 도움이 되는 다음 리소스를 제공합니다.

- [보안 및 규정 준수 킷스타트 가이드](#) - 이 배포 가이드에서는 아키텍처 고려 사항을 설명하고 보안 및 규정 준수에 AWS 중점을 둔 기본 환경을 배포하기 위한 단계를 제공합니다.
- [Amazon Web Services의 HIPAA 보안 및 규정 준수를 위한 설계 — 이 백서에서는 기업이 HIPAA 적격 애플리케이션을 만드는 AWS 데 사용할 수 있는 방법을 설명합니다.](#)

Note

모든 AWS 서비스 사람이 HIPAA 자격을 갖춘 것은 아닙니다. 자세한 내용은 [HIPAA 적격 서비스 참조](#)를 참조하십시오.

- [AWS 규정 준수 리소스AWS](#) — 이 워크북 및 가이드 모음은 해당 산업 및 지역에 적용될 수 있습니다.
- [AWS 고객 규정 준수 가이드](#) — 규정 준수의 관점에서 공동 책임 모델을 이해하십시오. 이 가이드에서는 보안을 유지하기 위한 모범 사례를 AWS 서비스 요약하고 여러 프레임워크 (미국 표준 기술 연구소 (NIST), 결제 카드 산업 보안 표준 위원회 (PCI), 국제 표준화기구 (ISO) 등) 에서 보안 제어에 대한 지침을 매핑합니다.
- AWS Config 개발자 안내서의 [규칙을 사용하여 리소스 평가](#) — 이 AWS Config 서비스는 리소스 구성이 내부 관행, 업계 지침 및 규정을 얼마나 잘 준수하는지 평가합니다.
- [AWS Security Hub](#) — 이를 AWS 서비스 통해 내부 AWS보안 상태를 포괄적으로 파악할 수 있습니다. Security Hub는 보안 제어를 사용하여 AWS 리소스를 평가하고 보안 업계 표준 및 모범 사례에 대한 규정 준수를 확인합니다. 지원되는 서비스 및 제어 목록은 [Security Hub 제어 참조](#)를 참조하십시오.
- [Amazon GuardDuty](#) — 환경에 의심스럽고 악의적인 활동이 있는지 AWS 계정모니터링하여 워크로드, 컨테이너 및 데이터에 대한 잠재적 위협을 AWS 서비스 탐지합니다. GuardDuty 특정 규정 준수 프레임워크에서 요구하는 침입 탐지 요구 사항을 충족하여 PCI DSS와 같은 다양한 규정 준수 요구 사항을 해결하는 데 도움이 될 수 있습니다.
- [AWS Audit Manager](#) — 이를 AWS 서비스 통해 AWS 사용량을 지속적으로 감사하여 위협을 관리하고 규정 및 업계 표준을 준수하는 방법을 단순화할 수 있습니다.

Application Auto Scaling에 대한 할당량

AWS 계정 각각에 대해 기본 할당량 (이전에는 한도라고 함) 이 있습니다. AWS 서비스다르게 표시되지 않는 한 리전별로 각 할당량이 적용됩니다. 일부 할당량에 대한 증가를 요청할 수 있으며 다른 할당량은 늘릴 수 없습니다.

Application Auto Scaling에 대한 할당량을 보려면 [Service Quotas 콘솔](#)을 엽니다. 탐색 창에서 AWS 서비스를 선택하고 Application Auto Scaling을 선택합니다.

할당량 증가를 요청하려면 Service Quotas 사용 설명서의 [할당량 증가 요청](#)을 참조하십시오.

Application Auto Scaling과 관련된 할당량은 다음과 같습니다. AWS 계정

명칭	기본값	조정 가능
리소스 유형별 확장 가능한 목표	아마존 다이내모DB: 5,000 아마존 ECS: 3,000 아마존 키스페이스: 1,500 기타 리소스 유형: 500	예
확장 가능한 대상별 조정 정책 (단계 조정 및 대상 추적 정책 모두)	50	아니요
확장 가능한 대상별 예약된 작업	200	아니요
조정 정책 단계당 단계별 조정	20	아니요

워크로드를 조정할 때 서비스 할당량을 염두에 두세요. 예를 들어, 서비스에서 허용되는 최대 용량 단위 수에 도달하면 확장이 중지됩니다. 수요가 감소하고 현재 용량이 감소하면 Application Auto Scaling이 다시 확장할 수 있습니다. 이 용량 한도에 다시 도달하지 않도록 증가를 요청할 수 있습니다. 각 서비스에는 리소스의 최대 용량에 대한 자체 기본 할당량이 있습니다. 다른 Amazon Web Services의 기본 할당량에 대한 자세한 내용은 Amazon Web Services 일반 참조의 [서비스 엔드포인트 및 할당량](#)을 참조하세요.

Application Auto Scaling의 문서 기록

다음 표에서는 2018년 1월 이후의 Application Auto Scaling 설명서에 대한 중요 추가 사항을 설명합니다. 이 설명서에 대한 업데이트 알림을 받으려면 RSS 피드에 가입하면 됩니다.

변경 사항	설명	날짜
설명서 변경 사항	할당량 설명서의 리소스 유형 당 최대 확장 가능 대상 수 항목을 업데이트했습니다. Application Auto Scaling에 대한 할당량을 참조하세요.	2024년 1월 16일
SageMaker 추론 구성 요소 지원	Application Auto Scaling을 사용하여 추론 구성 요소의 복사본 규모를 조정합니다.	2023년 11월 29일
IAM 서비스 연결 역할 권한 업데이트	AWSApplicationAutoScalingSageMakerEndpointPolicy 정책에 대한 Application Auto Scaling 업데이트. 자세한 내용은 AWS 관리형 정책에 대한 Application Auto Scaling 업데이트 를 참조하세요.	2023년 11월 13일
SageMaker 서버리스 프로비저닝된 동시성 지원	Application Auto Scaling을 사용하여 서버리스 엔드포인트의 프로비저닝된 동시성을 조정합니다.	2023년 5월 9일
태그를 사용하여 확장 가능 대상 분류	이제 Application Auto Scaling 확장 가능 대상에 태그 형식으로 메타데이터를 지정할 수 있습니다. Application Auto Scaling에 대한 태그 지정 지원 을 참조하세요.	2023년 3월 20일

<u>CloudWatch 미터법 수학적 지원</u>	이제 대상 추적 조정 정책을 생성할 때 지표 수학을 사용할 수 있습니다. 메트릭 수학을 사용하면 여러 CloudWatch 메트릭을 쿼리하고 수학적 식을 사용하여 이러한 메트릭을 기반으로 새 시계열을 만들 수 있습니다. <u>지표 수학을 사용하여 Application Auto Scaling에서 대상 추적 조정 정책 생성</u> 을 참조하세요.	2023년 3월 14일
<u>설명서 변경 사항</u>	Application Auto Scaling 사용 설명서의 새 주제는 Application Auto Scaling에서 AWS CloudShell 을(를) 사용하는데 도움이 됩니다. <u>명령줄에서 Application Auto Scaling을 사용하려면 사용을 AWS CloudShell</u> 참조하십시오.	2023년 2월 17일
<u>확장하지 않는 이유</u>	이제 Application Auto Scaling에서 Application Auto Scaling API를 사용하여 리소스를 확장하지 않는 머신 판독 가능한 이유를 검색할 수 있습니다. <u>Application Auto Scaling 확장 활동을 참조</u> 하세요.	2023년 1월 4일
<u>설명서 변경 사항</u>	할당량 설명서의 리소스 유형당 최대 확장 가능 대상 수 항목을 업데이트했습니다. <u>Application Auto Scaling에 대한 할당량을 참조</u> 하세요.	2022년 5월 6일

[Amazon Neptune 클러스터에 대한 지원 추가](#)

Application Auto Scaling을 사용하여 Amazon Neptune DB 클러스터의 복제본 수를 조정할 수 있습니다. 자세한 내용은 [Amazon Neptune 및 Application Auto Scaling](#)을 참조하세요. 주제 [Application Auto Scaling의 AWS 관리형 정책 업데이트](#)가 Neptune과의 통합을 위한 새로운 관리형 정책을 나열하도록 업데이트되었습니다.

2021년 10월 6일

[이제 Application Auto Scaling에서 AWS 관리형 정책의 변경 사항을 보고합니다.](#)

2021년 8월 19일부터 관리형 정책에 대한 변경 사항은 관리형 정책에 대한 [Application Auto Scaling 업데이트](#) 항목에 보고됩니다. AWS 나열된 첫 번째 변경 ElastiCache 사항은 Redis에 필요한 권한 추가입니다.

2021년 8월 19일

[Redis 복제 ElastiCache 그룹에 대한 지원 추가](#)

Application Auto Scaling을 사용하여 ElastiCache Redis용 복제 그룹 (클러스터)의 노드 그룹 수와 노드 그룹당 복제본 수를 확장할 수 있습니다. 자세한 내용은 [ElastiCache 내용은 Redis 및 Application Auto Scaling](#)을 참조하십시오.

2021년 8월 19일

설명서 변경 사항

Application Auto Scaling 사용 설명서의 새로운 IAM 주제는 Application Auto Scaling에 대한 액세스 문제를 해결하는 데 도움이 됩니다. 자세한 내용은 [Application Auto Scaling에 사용되는 Identity and Access Management](#)를 참조하세요. 또한 대상 서비스 및 CloudWatch Amazon에서의 작업에 대한 새 예제 IAM 권한 정책이 추가되었습니다. 자세한 내용은 [AWS CLI 또는 SDK 사용을 위한 예제 정책을 참조하십시오](#).

2021년 2월 23일

현지 시간대 지원 추가

이제 현지 시간대에서 예약된 작업을 만들 수 있습니다. 일광 절약 시간을 준수하는 경우 DST(일광 절약 시간제)에 맞게 자동으로 조정됩니다. 자세한 내용은 [예약된 조정](#)을 참조하세요.

2021년 2월 2일

설명서 변경 사항

Application Auto Scaling 사용 설명서의 새로운 [자습서](#)는 Application Auto Scaling을 사용할 때 대상 추적 조정 정책 및 예약된 조정을 사용하여 애플리케이션의 가용성을 높이는 방법을 이해하는 데 도움이 됩니다. 또한 새 [항목에서는](#) 주의가 필요할 수 있는 문제가 CloudWatch 감지되었을 때 알림을 트리거하는 방법을 설명합니다.

2020년 10월 15일

Amazon Managed Streaming for Apache Kafka 클러스터 스토리지에 대한 지원 추가	대상 추적 조정 정책을 사용하여 Amazon MSK 클러스터와 연결된 브로커 스토리지의 양을 조정합니다.	2020년 9월 30일
Amazon Comprehend 엔터티 인식기 엔드포인트에 대한 지원 추가	Application Auto Scaling을 사용하여 Amazon Comprehend 엔터티 인식기 엔드포인트에 프로비저닝된 추론 단위 수를 조정합니다.	2020년 9월 28일
Amazon Keyspaces(Apache Cassandra용) 표에 대한 지원 추가	Application Auto Scaling을 사용하여, Amazon Keyspaces 테이블의 프로비저닝된 처리량(읽기 및 쓰기 용량)을 조정합니다.	2020년 4월 23일
새로운 “보안” 장	Application Auto Scaling 사용 설명서의 새로운 보안 장은 Systems Manager를 사용할 때 공동 책임 모델 을 적용하는 방법을 이해하는 데 도움이 됩니다. 이번 업데이트에서는 사용 설명서의 "인증 및 액세스 제어" 장이 새롭고 더욱 유용한 내용인 Application Auto Scaling의 Identity and Access Management 단원으로 대체되었습니다.	2020년 1월 16일
마이너 업데이트	다양한 개선 및 수정 사항입니다.	2020년 1월 15일

알림 기능 추가	이제 Application Auto Scaling에서 이벤트를 Amazon에 EventBridge 보내고 특정 작업이 발생하면 사용자에게 알림을 보냅니다. AWS Health Dashboard 자세한 내용은 Application Auto Scaling 모니터링 섹션을 참조하세요.	2019년 12월 20일
AWS Lambda 함수 지원 추가	Application Auto Scaling을 사용하여 Lambda 함수의 프로비저닝된 동시성을 조정합니다.	2019년 12월 3일
Amazon Comprehend 문서 분류 엔드포인트에 대한 지원 추가	Application Auto Scaling을 사용하여 Amazon Comprehend 문서 분류 엔드포인트의 처리 용량을 조정합니다.	2019년 11월 25일
타겟 트래킹 스케일링 정책에 AppStream 2.0 지원 추가	대상 추적 조정 정책을 사용하여 AppStream 2.0 플릿의 크기를 조정할 수 있습니다.	2019년 11월 25일
Amazon VPC 엔드포인트에 대한 지원	이제 VPC와 Application Auto Scaling 간에 프라이빗 연결을 설정할 수 있습니다. 마이그레이션 고려 사항 및 지침은 Application Auto Scaling 및 인터페이스 VPC 엔드포인트 를 참조하세요.	2019년 11월 22일
조정 일시 중지 및 재개	조정 일시 중지 및 재개에 대한 지원이 추가되었습니다. 자세한 내용은 Application Auto Scaling의 조정 일시 중지 및 재개 를 참조하세요.	2019년 8월 29일

새로운 섹션	설정 섹션이 Application Auto Scaling 설명서에 추가되었습니다. 사용 설명서 전체에서 일부 개선 및 수정이 이루어졌습니다.	2019년 6월 28일
설명서 변경 사항	Application Auto Scaling 설명서의 예약된 조정 , 단계 조정 정책 및 대상 추적 조정 정책 섹션이 개선되었습니다.	2019년 3월 11일
사용자 지정 리소스에 대한 지원 추가	Application Auto Scaling을 사용하여 자체 애플리케이션 또는 서비스에서 제공하는 사용자 지정 리소스를 조정합니다. 자세한 내용은 GitHub리포지토리를 참조하십시오 .	2018년 7월 9일
SageMaker 엔드포인트 변형에 대한 지원 추가	Application Auto Scaling을 사용해 변형을 위해 제공되는 엔드포인트 인스턴스 수를 조정합니다.	2018년 2월 28일

다음 표에서는 2018년 1월 이전에 Application Auto Scaling 설명서에서 변경된 중요 사항에 대해 설명합니다.

변경 사항	설명	날짜
Aurora 복제본 지원 추가	Application Auto Scaling을 사용하여 원하는 수로 조정합니다. 자세한 내용은 Amazon RDS 사용 설명서의 Aurora 복제본에 Amazon Aurora Auto Scaling 사용 을 참조하세요.	2017년 11월 17일
예약된 조정에 대한 지원 추가	예약된 조정을 사용하여 미리 설정된 시간 또는 간격에 따	2017년 11월 8일

변경 사항	설명	날짜
	<p>라 리소스를 조정합니다. 자세한 내용은 Application Auto Scaling의 예약된 조정을 참조하세요.</p>	
<p>대상 추적 조정 정책에 대한 지원 추가</p>	<p>대상 추적 조정 정책을 사용하여 몇 가지 단계를 통해 애플리케이션에 대한 동적 조정을 설정합니다. 자세한 내용은 Application Auto Scaling의 대상 추적 조정 정책을 참조하세요.</p>	<p>2017년 7월 12일</p>
<p>DynamoDB 테이블 및 글로벌 보조 인덱스에 대해 프로비저닝된 읽기 및 쓰기 용량 지원 추가</p>	<p>Application Auto Scaling을 사용하여 프로비저닝된 처리량(읽기 및 쓰기 용량)을 조정합니다. 자세한 내용은 Amazon DynamoDB 개발자 안내서의 Auto Scaling으로 처리 용량 관리를 참조하세요.</p>	<p>2017년 6월 14일</p>
<p>AppStream 2.0 플릿에 대한 지원 추가</p>	<p>Application Auto Scaling을 사용하여 플릿의 크기를 조정합니다. 자세한 내용은 Amazon AppStream AppStream 2.0 관리 가이드의 2.0용 플릿 오토 스케일링을 참조하십시오.</p>	<p>2017년 3월 23일</p>
<p>Amazon EMR 클러스터에 대한 지원 추가</p>	<p>Application Auto Scaling을 사용하여 코어 및 태스크 노드를 조정합니다. 자세한 내용은 Amazon EMR 관리 안내서의 Amazon EMR에서 자동 조정 사용을 참조하세요.</p>	<p>2016년 11월 18일</p>

변경 사항	설명	날짜
스팟 플릿에 대한 지원 추가	Application Auto Scaling을 사용하여 대상 용량을 조정합니다. 자세한 내용은 Amazon EC2 사용 설명서의 스팟 플릿의 자동 조정을 참조하십시오.	2016년 9월 1일
Amazon ECS 서비스에 대한 지원 추가	Application Auto Scaling을 사용하여 원하는 수로 조정합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 서비스 Auto Scaling 을 참조하세요.	2016년 8월 9일

기계 번역으로 제공되는 번역입니다. 제공된 번역과 원본 영어의 내용이 상충하는 경우에는 영어 버전이 우선합니다.