aws re/start

# Amazon EC2 Auto Scaling

# 학습 내용

## 강의의 핵심

**배울 내용은 다음과 같습니다.**

- Amazon EC2 Auto Scaling 및 시작 템플릿 설명하기
- 시작 템플릿 구성 및 관리, 크기 조정 제어하기
- Amazon Web Services(AWS)에서 Amazon EC2 Auto Scaling 사용하기

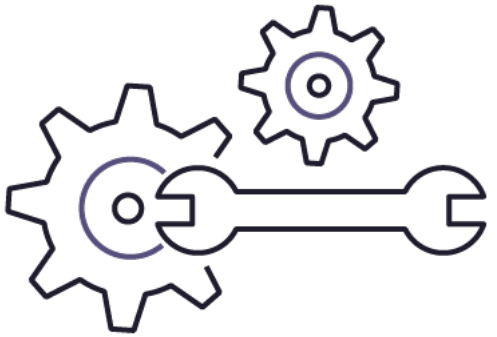
**주요 용어:**

- Automatic scaling
- 시작 템플릿
- Auto Scaling 그룹
- Auto Scaling 정책

# Amazon EC2 Auto Scaling

# Amazon EC2 Auto Scaling

- 다음을 기반으로 Amazon Elastic Compute Cloud(Amazon EC2) 인스턴스를 자동으로 시작하거나 종료합니다.
  - 상태 확인
  - Amazon CloudWatch에서 지원하는 사용자 정의 정책
  - 일정
  - 그 외 기준(예: 프로그래밍 방식)
  - 설정한 원하는 용량을 수동으로 사용
- 수요에 맞춰 확장하고 비용을 절감하도록 축소



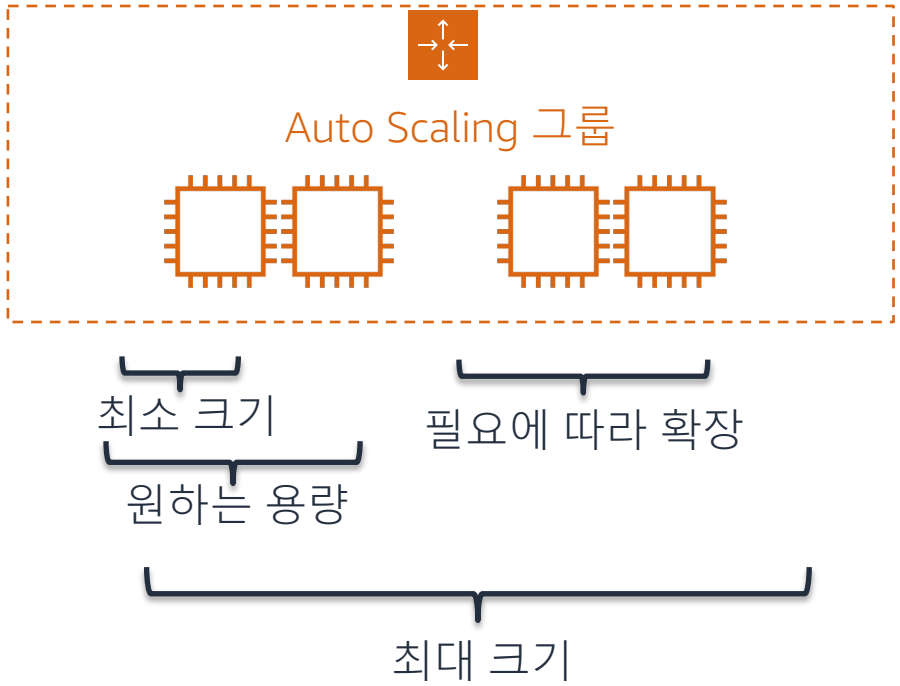
# Amazon EC2 Auto Scaling 실행

시작 템플릿 생성

**Auto Scaling** 그룹 생성

시작 대상: **EC2** 인스턴스 구성

관리 및 확장하는 방법:
**EC2** 인스턴스 그룹 크기 및
조정 정책 구성





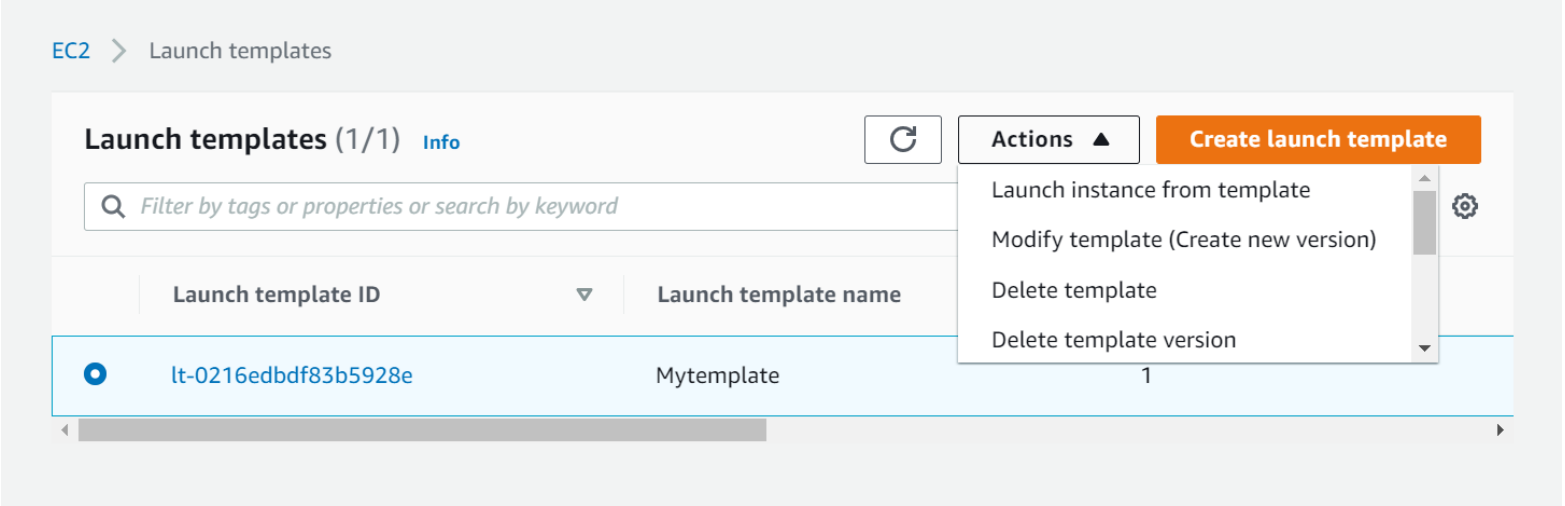

aws re/start

# Amazon EC2 Auto Scaling 실행

시작 템플릿

시작 템플릿을 생성하려면 다음을 지정해야 합니다.

- Amazon Machine Image(AMI)
- 인스턴스 유형
- VPC
- 보안 그룹
- 스토리지
- 인스턴스 키 페어
- IAM 역할
- 사용자 데이터
- 태깅






aws re/start

# Amazon EC2 Auto Scaling 실행

**Amazon EC2 Auto Scaling 그룹**

EC2 인스턴스의 논리적 그룹

다음 범위에서 자동으로 조정:
- 최소
- 원하는 값(선택 사항)
- 최대

Elastic Load Balancing과 통합(선택 사항)

그룹 크기를 유지하기 위한 상태 확인

인스턴스를 여러 가용 영역에 분산 및 밸런싱

# Amazon EC2 Auto Scaling 실행

**Amazon EC2 Auto Scaling 정책**

Amazon EC2 Auto Scaling 작업 수행을 위한 파라미터

정책 트리거 방법
- Amazon CloudWatch 경보
- 대상 추적
- 예약
- 수동

확장/축소 및 정도:
- ChangeInCapacity(+/- #)
- ExactCapacity(#)
- ChangeInPercent(+/- %)

# Amazon EC2 Auto Scaling 및 인스턴스 상태

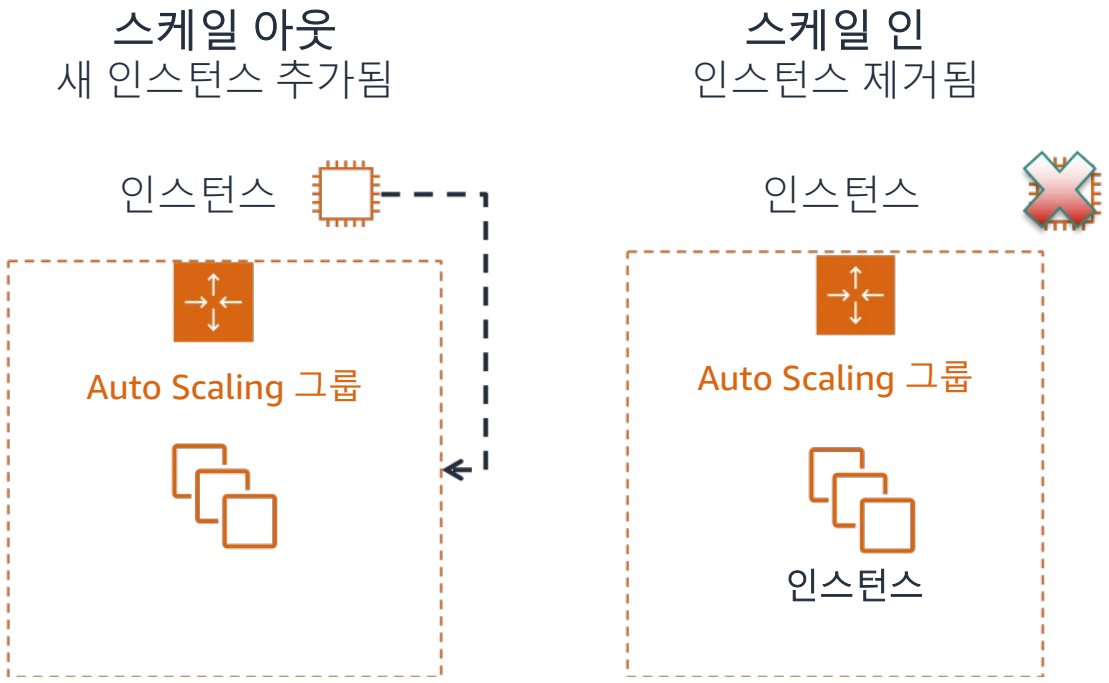
- 인스턴스의 상태 유지 관리
- 비정상으로 표시된 인스턴스 종료
- 기본적으로 EC2 인스턴스 상태 확인 사용
- 인스턴스가 로드 밸런서 뒤에 있는 경우 다음 검사를 구성할 수 있습니다.
  - 로드 밸런서의 인스턴스 확인
  - EC2 인스턴스 확인

- 외부 스크립트를 통해 인스턴스 재활용을 트리거할 수 있습니다.

  `aws autoscaling set-instance-health` 명령

# Amazon EC2 Auto Scaling 종료 정책

- 축소 시 어떤 인스턴스를 종료할지 결정
- 다음 요인에 따라 종료 순서 결정
  - 인스턴스 수가 가장 많은 가용 영역
  - 다중 정책(정책은 나열된 순서대로 실행됨)



# 종료 정책

| 종료 정책 | 설명 |
| --- | --- |
| OldestInstance | 가장 오래 실행된 인스턴스를 선택합니다. |
| NewestInstance | 가장 짧게 실행된 인스턴스를 선택합니다. |
| OldestLaunchTemplate | 가장 오래된 시작 템플릿이 있는 인스턴스를 종료합니다(기본값). |
| ClosestToNextInstanceHour | 다음 청구 가능 시간과 가장 가까운 인스턴스를 종료합니다(기본값). |

# 안정된 상태 그룹 생성

**안정된 상태 그룹:**

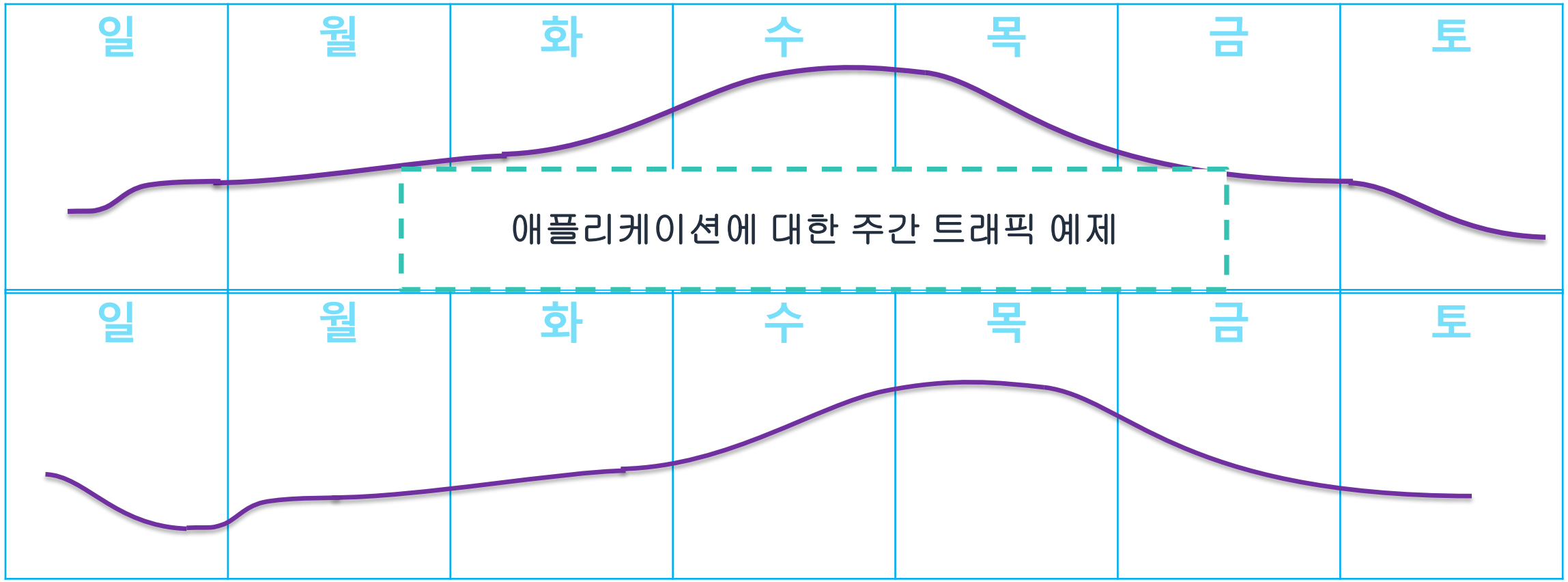
- Amazon EC2 Auto Scaling 그룹을 **최소**, **최대** 및 **원하는** 값으로 설정
- 인스턴스가 비정상적 상태가 되거나 가용 영역이 중단되면 인스턴스를 자동으로 다시 생성
- 인스턴스가 재활용되는 동안 잠재적인 중단 시간

**용례:**

각 가용 영역에서 안정된 상태의 Network Address Translation(NAT) 서버를 유지합니다.

# Amazon EC2

## 시기 지정형 크기 조정



# 동적 스케일링: Amazon EC2

**대상 추적 조정**

특정 지표에 대한 목표 값을 기반으로 그룹의 현재 용량을 늘리거나 줄입니다.

**단계 조정**

경보 위반의 크기를 기반으로 하는 일련의 조정 조절을 기반으로 그룹의 현재 용량을 늘리거나 줄입니다.

**단순 조정**

단일 조정 조절을 기반으로 그룹의 현재 용량을 늘리거나 줄입니다.

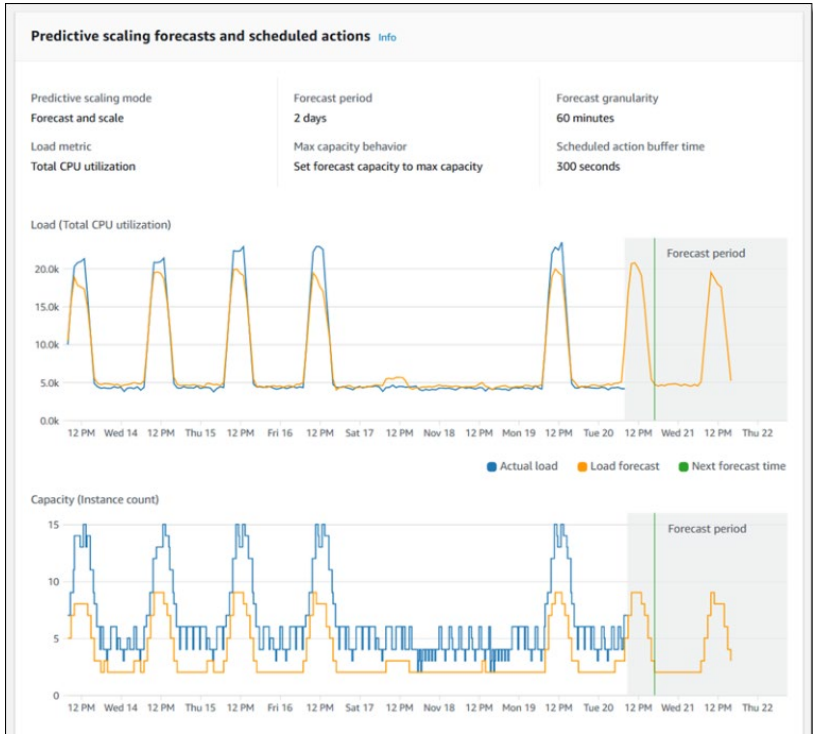
# 예측 스케일링: Amazon EC2

## 예측 스케일링

- 로드 예측
- 최소 용량 예약

## 용례

- 동적 스케일링 대상 추적
- 주기적인 최고치 기간이 있는 애플리케이션 환경



# 학습 내용 확인

엔지니어들은 Auto Scaling 그룹의 서버에 대한 새 구성을 만들었습니다. 그들은 새 서버의 구성이 실패할 경우 Auto Scaling 그룹에서 신속하게 제거할 수 있기를 원합니다.

**OldestInstance**, **NewestIntance**, **OldestLaunchTemplate** 또는 **ClosestToNextInstanceHour** 중 어떤 종료 정책이 원하는 결과를 가장 잘 얻을 수 있습니까?

첫 번째 질문에서 계속해서 엔지니어들은 Auto Scaling 그룹에서 새 서버를 어떻게 강제로 생성합니까?

그런 다음, 필요한 경우 업데이트된 구성이 있는 서버를 어떻게 강제로 제거합니까?

동일한 시나리오를 계속 진행하면 Auto Scaling 그룹에서 새로 생성된 모든 서버에 장애가 발생하는 것으로 나타났습니다.

엔지니어들은 무엇을 했어야 했습니까?

# 핵심 사항


- Amazon EC2 Auto Scaling은 정의된 조건에 따라 EC2 인스턴스를 자동으로 추가하거나 제거하여 애플리케이션 가용성을 유지하는 데 도움이 됩니다.
- Amazon EC2 Auto Scaling은 다음 3개 부분으로 구성되어 있습니다.
  - 시작 템플릿
  - Auto Scaling 그룹
  - 조정 정책
- 다음에 따라 크기 조정할 수 있습니다.
  - 인스턴스 상태
  - Amazon CloudWatch 경보
  - 시간표 또는 과거 사용량(예측)
- 시작 템플릿을 생성하려면 몇 가지 세부 정보를 지정해야 합니다.





aws re/start