

비용을 이해하기 위한 모니터링

aws training and certification

보다 유연하고 탄력적인 아키텍처를 만들려면
어디에서 비용을 지출하고 있는지 알아야 합니다.

AWS Cost Explorer

-  보고서를 생성합니다.
-  13개월 데이터
-  예측을 제공합니다.
-  지출 패턴을 참조합니다.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Cost Optimization Monitor – 서비스 사용량 및 비용 분석 정보를 제공하는 보고서를 생성할 수 있습니다. 기간, 계정, 리소스 또는 태그를 기준으로 분류할 수 있는 예상 비용을 제공합니다.

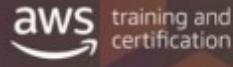
AWS Cost Explorer – 지난 13개월까지 데이터를 볼 수 있으므로 시간 흐름에 따른 AWS 리소스 소비 패턴을 확인할 수 있습니다.

AWS Cost Explorer를 사용한 예측 – 예측은 과거 사용량을 기반으로 사용자가 선택한 예측 기간 동안 AWS 서비스 사용량을 예측하는 것입니다. 보고서의 미래 시간 범위를 선택하여 예측을 생성합니다. 예측을 통해 AWS 청구 금액을 예상하고 사용할 것으로 예측되는 금액에 대해서 경보와 예산을 적용할 수 있습니다. 예측은 예상이므로 예상 청구 금액은 추정치이며 각 청구서 기간의 실제 요금과 다를 수 있습니다.

정확도 범위에 따라 신뢰 구간이 다릅니다. 신뢰 구간이 높을수록 예측이 정확할 가능성이 높습니다. AWS Cost Explorer 예측의 신뢰 구간은 80%입니다. AWS에 80% 신뢰 구간 내로 예측하는 데 충분한 데이터가 없는 경우, AWS Cost Explorer는 예측을 표시하지 않습니다.

<https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/ce-modify.html#ce-timerange>

Amazon CloudWatch를 사용하여 인프라 모니터링



Amazon CloudWatch

• 리소스에 대한 지표를 수집하고 추적합니다.

• 경보를 생성하고 알림을 전송할 수 있습니다.

• 설정한 규칙에 따라 리소스의 용량 변화를 트리거할 수 있습니다.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

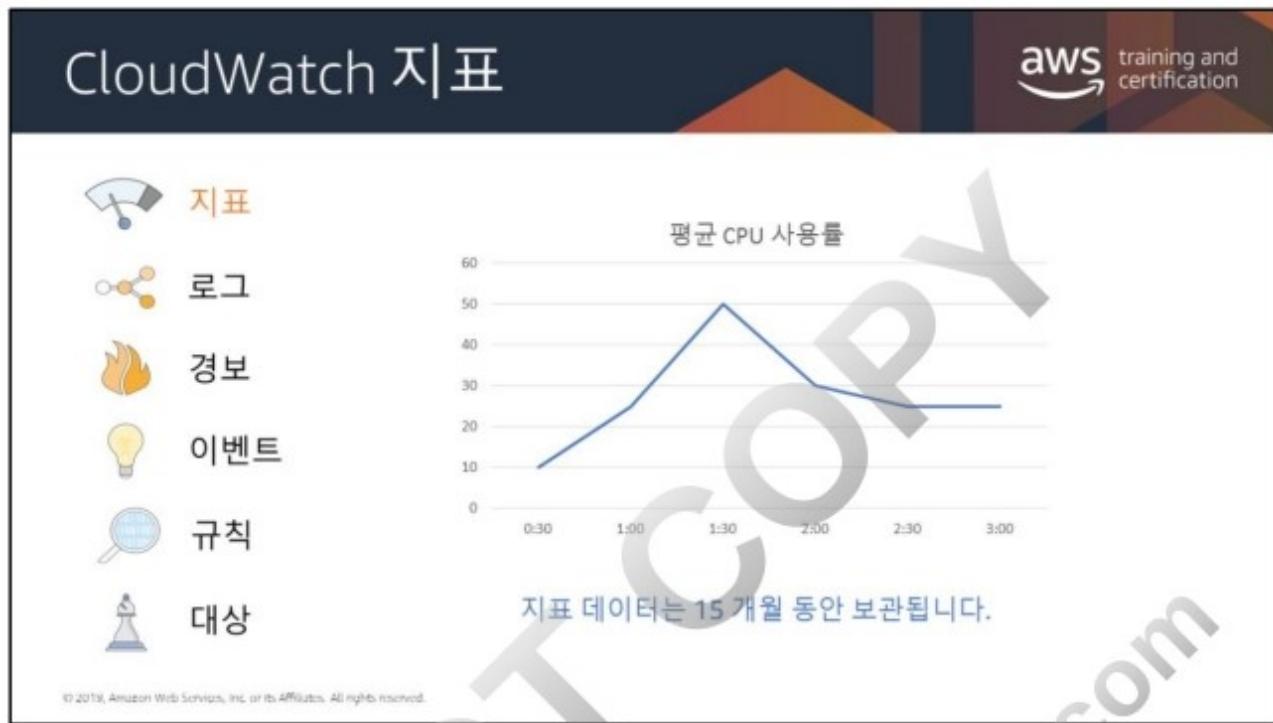
탄력적 아키텍처를 생성하는 여정의 첫 단계는 Amazon CloudWatch를 살펴보는 것입니다. CloudWatch는 AWS 리소스 및 애플리케이션에 대한 가시성을 확대하는 데 도움이 됩니다.

CloudWatch를 사용하여 리소스 및 애플리케이션에 대해 측정할 수 있는 변수인 지표를 수집하고 추적할 수 있습니다. CloudWatch 경보는 알림을 보내거나 정의한 규칙을 기준으로 모니터링하는 리소스를 자동으로 변경합니다. 예를 들어 Amazon EC2 인스턴스의 CPU 사용량과 디스크 읽기 및 쓰기를 모니터링한 다음, 이러한 데이터를 사용하여 증가된 로드를 처리하기 위해 추가 인스턴스를 시작해야 할지 결정할 수 있습니다. 또한 이러한 데이터를 사용하여 사용률이 낮은 인스턴스를 중지하고 비용을 절감할 수도 있습니다. AWS에서 기본으로 제공하는 지표 이외에도 사용자 지정 지표를 모니터링할 수 있습니다. CloudWatch를 사용하면 시스템 전체의 리소스 사용률, 애플리케이션 성능 및 운영 상태를 파악할 수 있습니다.

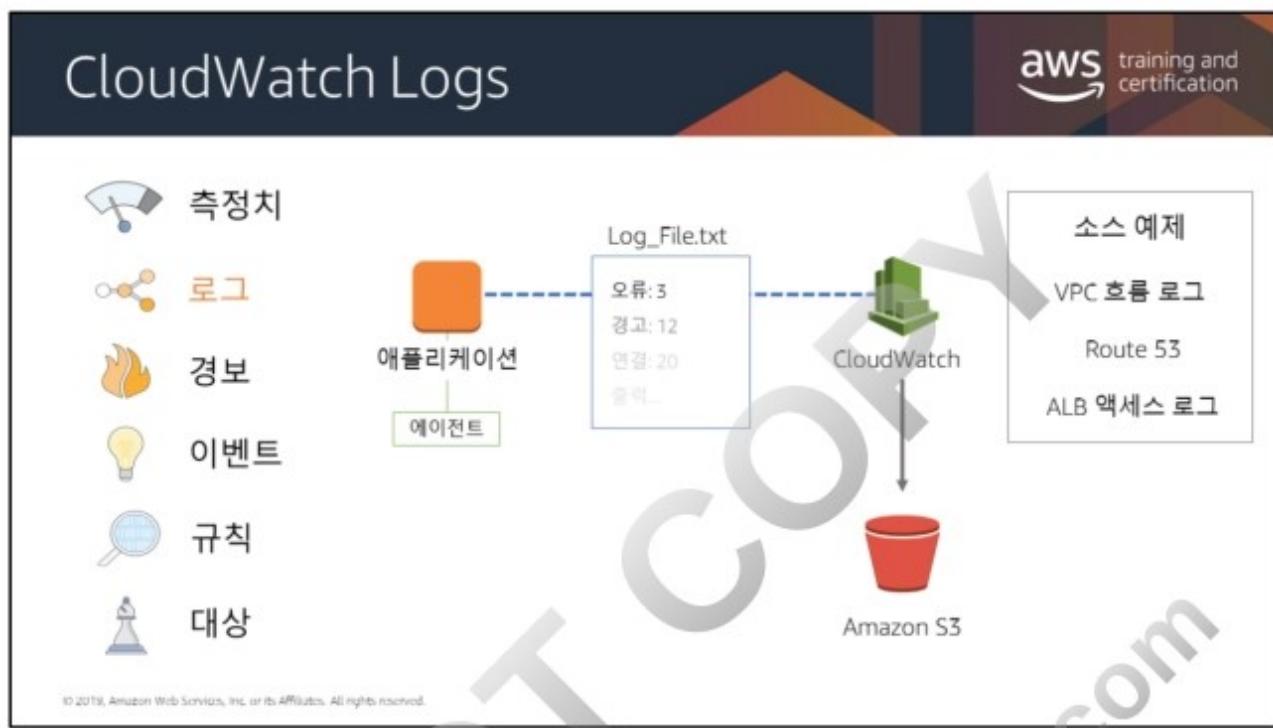
자세한 내용은 다음을 참조하십시오.

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/WhatIsCloudWatch.html>.





지표는 시스템 성능에 대한 데이터입니다. 많은 AWS 서비스가 리소스에 대한 지표를 기본적으로 제공합니다(예: Amazon EC2 인스턴스, Amazon EBS 볼륨, Amazon RDS DB 인스턴스). 또한 Amazon EC2 인스턴스 같은 일부 리소스에 대해 세부 모니터링을 활성화하거나 자체 애플리케이션 지표를 게시할 수도 있습니다. Amazon CloudWatch는 검색, 그래프 처리 및 경보를 위해 계정에 모든 지표(AWS 리소스 지표 및 사용자가 제공한 애플리케이션 지표 모두)를 로드할 수 있습니다.



CloudWatch Logs를 사용하면 소스(예: EC2 인스턴스, Amazon Route 53, AWS CloudTrail 및 기타 AWS 서비스)의 로그 파일을 모니터링, 저장 및 액세스할 수 있습니다.

예를 들어 Amazon EC2 인스턴스의 로그를 실시간으로 모니터링할 수 있습니다. 애플리케이션 로그에서 오류 발생 횟수를 추적하고 해당 비율이 사전에 정의된 수준을 초과할 경우 알림을 보낼 수 있습니다.

CloudWatch Logs는 로그 데이터 자체를 모니터링하기 때문에 코드 변경이 필요하지 않습니다.

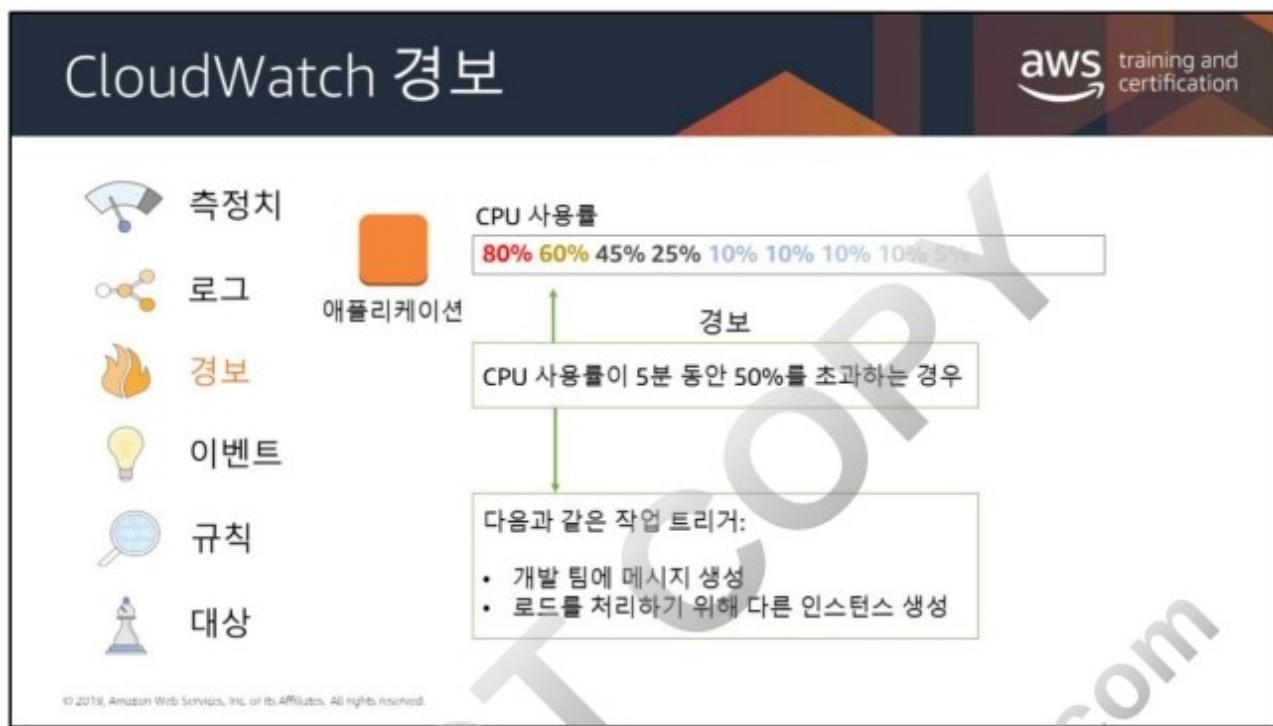
또한 CloudWatch Logs Insights를 사용하여 몇 초 만에 로그를 분석하면 빠른 대화형 쿼리 및 시각화를 얻을 수 있습니다. 선 또는 누적 영역형 차트를 사용하여 쿼리 결과를 시각화할 수 있으며, CloudWatch 대시보드에서 해당 쿼리를 추가할 수 있습니다.

자세한 내용은 다음을 참조하십시오.

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/WhatIsCloudWatchLogs.html>

<https://aws.amazon.com/blogs/aws/new-amazon-cloudwatch-logs-insights-fast-interactive-log-analytics/>

DO NOT COPY
zlagusdbs@gmail.com

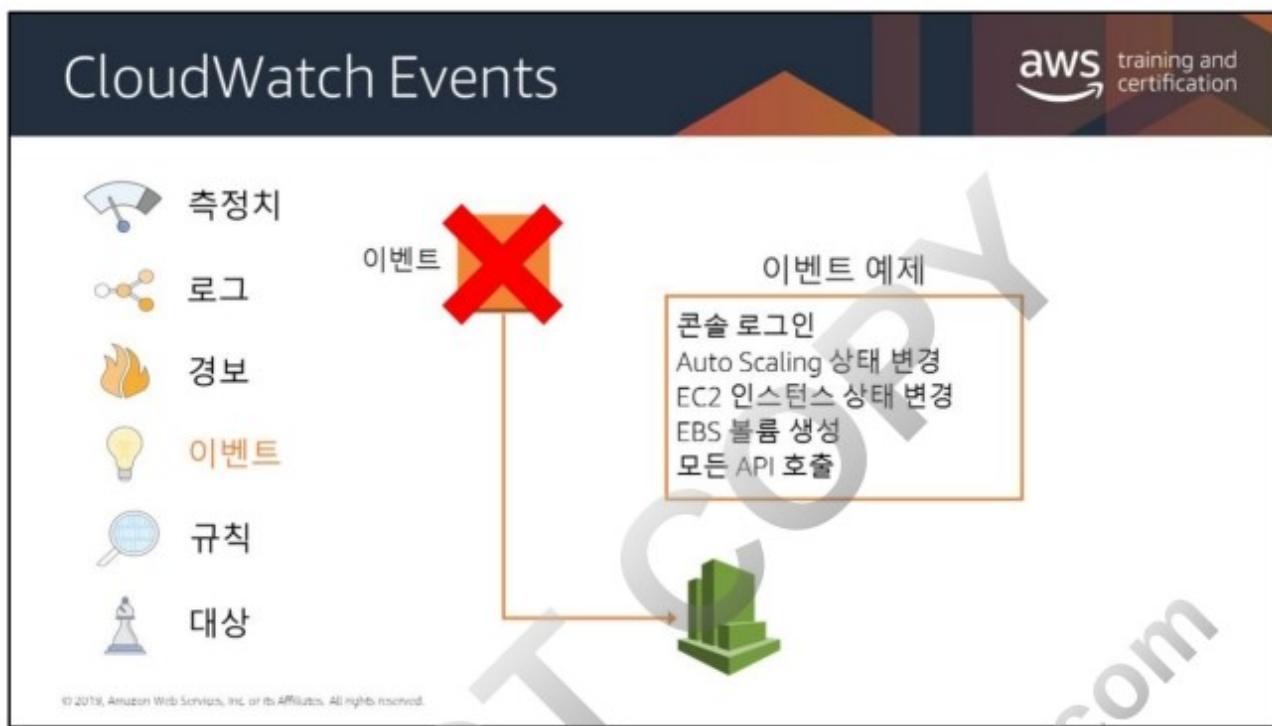


경보를 사용하여 작업을 자동으로 시작할 수 있습니다. 경보는 지정한 기간에 단일 지표를 감시하고 시간에 따른 임계값 대비 지표 값을 기준으로 지정된 작업을 하나 이상 수행합니다. 작업이란 Amazon SNS 주제 또는 Auto Scaling 정책으로 전송되는 알림을 말합니다. 대시보드에 경보를 추가할 수도 있습니다.

경보는 지속되는 상태 변경에 대해서만 작업을 호출합니다. CloudWatch 경보는 특정 상태가 되었다고 해서 작업을 호출하지는 않습니다. 상태가 변경되어야 하고 지정된 기간 동안 변경된 상태가 유지되어야 합니다.

이 예에서는 경보가 트리거되면 Auto Scaling 정책 실행, 알림 전송(인스턴스에 대한 알림을 운영팀에 전송) 등 다른 작업이 시작됩니다.

또한, 작업은 경보가 트리거되지 않아도 실행될 수 있습니다.



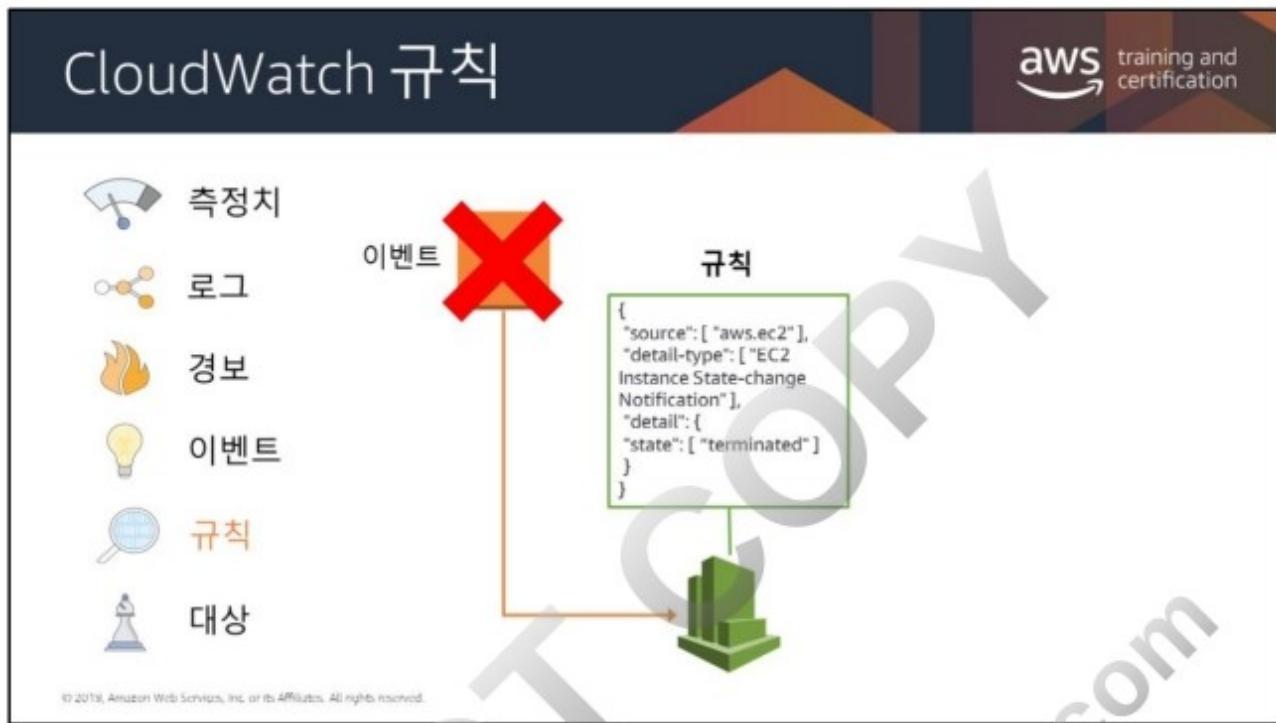
Amazon CloudWatch Events는 AWS 리소스의 변경 사항을 설명하는 시스템
이벤트의 스트림을 거의 실시간으로 제공합니다.

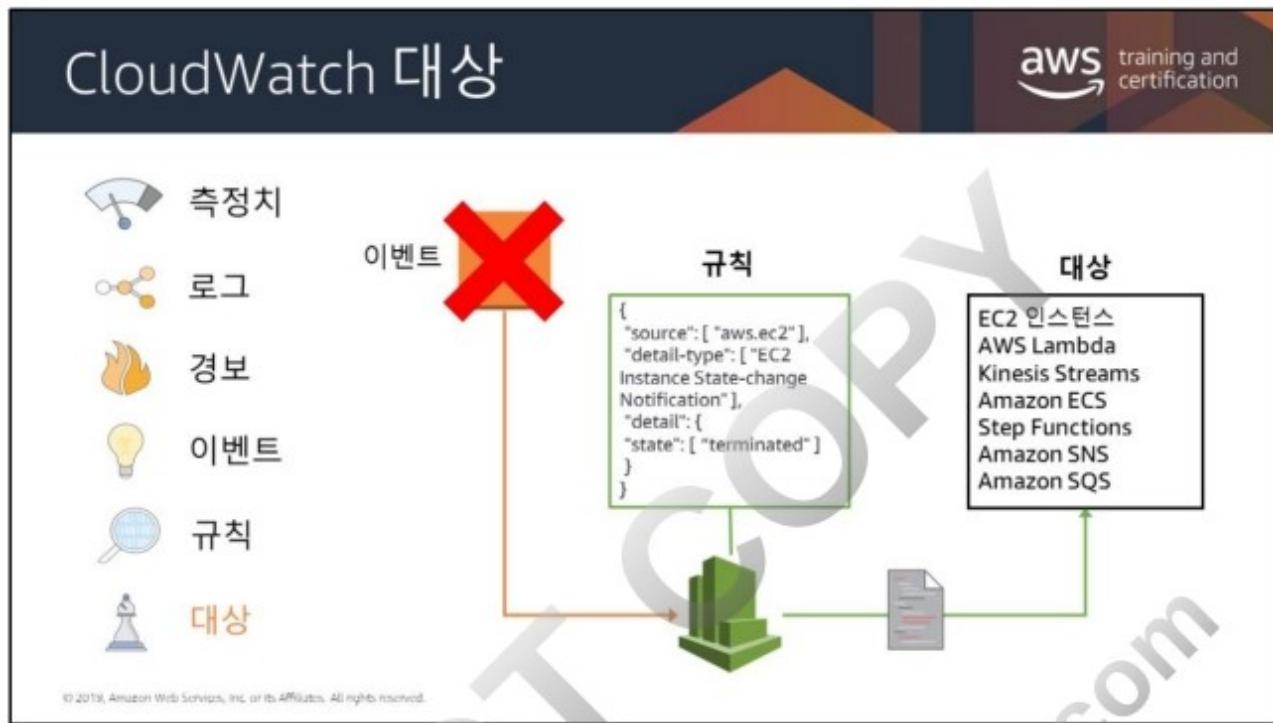
AWS 리소스는 상태 변경 시 이벤트를 생성할 수 있습니다. 예를 들어, Amazon EC2는 EC2 인스턴스의 상태가 보류에서 실행으로 변경될 때 이벤트를 생성하며, Amazon EC2 Auto Scaling은 인스턴스가 시작 또는 종료될 때 이벤트를 생성합니다.

신속하게 설정할 수 있는 단순 규칙을 사용하여 일치하는 이벤트를 검색하고
하나 이상의 대상 함수 또는 스트림으로 이를 라우팅할 수 있습니다.

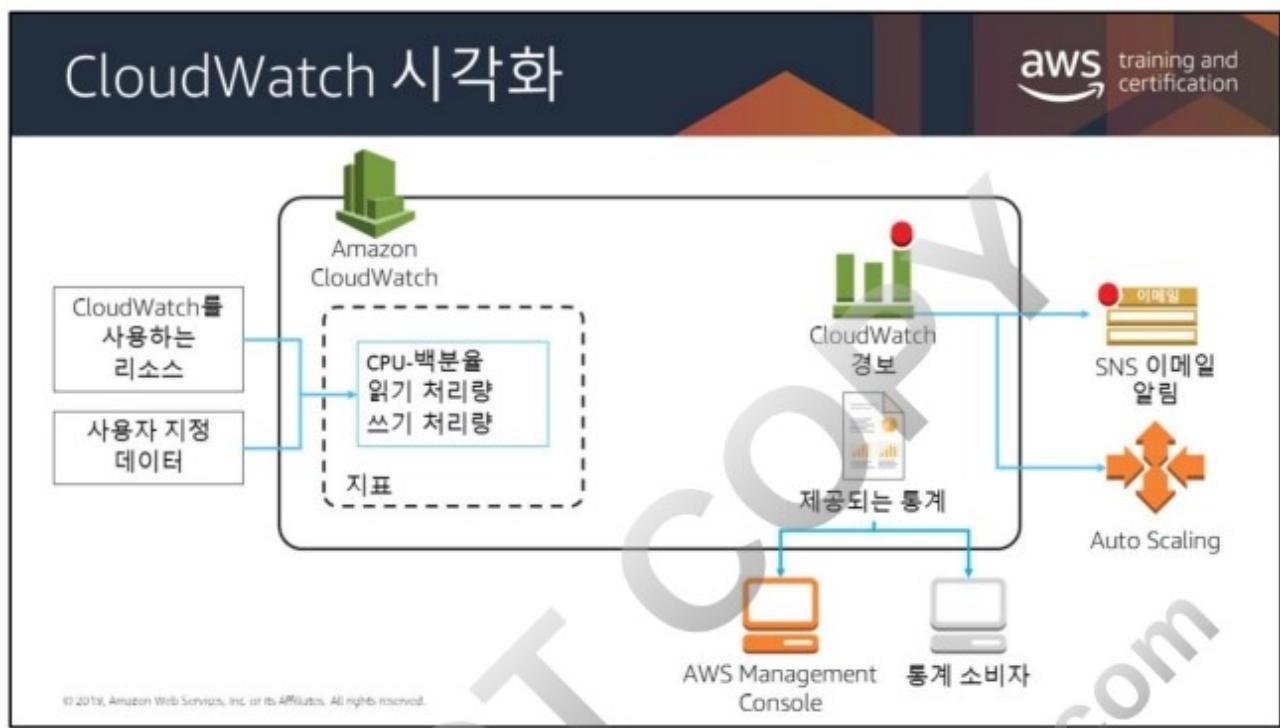
CloudWatch Events는 운영 변경 시 이를 알아차립니다. CloudWatch Events는
환경에 응답하기 위한 메시지를 전송하고, 함수를 활성화하고, 변경을 수행하고,
상태 정보를 기록하는 등 이러한 운영 변경에 응답하고 필요에 따라 교정 조치를
취합니다.

또한 CloudWatch Events를 사용하여 cron 또는 rate 표현식을 통해 특정 시간에
자체 트리거되는 자동 작업을 예약할 수 있습니다.





대상은 이벤트를 처리합니다. 대상에는 Amazon EC2 인스턴스, AWS Lambda 함수, Kinesis 스트림, Amazon ECS 작업, Step Functions 상태 시스템, Amazon SNS 주제, Amazon SQS 대기열 및 기본 제공 대상이 포함될 수 있습니다. 대상은 JSON 형식으로 이벤트를 수신합니다.



Amazon CloudWatch는 기본적으로 지표 리포지토리입니다. AWS 서비스(예: Amazon EC2)는 지표를 리포지토리에 저장하므로 이러한 지표를 기반으로 통계를 검색할 수 있습니다. 사용자 지정 지표를 리포지토리에 저장하면 해당 지표에 대한 통계도 검색할 수 있습니다.

자세한 내용은 다음을 참조하십시오.

https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/cloudwatch_architecture.html



AWS CloudTrail은 계정에 대한 AWS API 호출을 기록하고 로그 파일을 사용자에게 전달하는 웹 서비스입니다. API 호출자 자격 증명, API 호출 시간, API 호출자의 원본 IP 주소, 요청 파라미터 및 AWS 서비스가 반환한 응답 요소와 같은 정보가 기록됩니다.

CloudTrail에서는 AWS 관리 콘솔, AWS SDK, 명령줄 도구, 상위 수준 AWS 서비스(예: AWS CloudFormation)를 통해 이루어진 API 호출을 비롯하여 계정에 대한 AWS API 호출 내역을 확인할 수 있습니다. CloudTrail에서 작성되는 AWS API 호출 내역을 통해 보안 분석, 리소스 변경 사항 추적 및 규정 준수 감사를 수행할 수 있습니다.

CloudTrail은 리전 단위로 활성화됩니다. 여러 리전을 사용하는 경우, 리전별로 로그 파일이 전송될 장소를 선택할 수 있습니다. 예를 들어 리전별로 별도의 Amazon S3 버킷을 사용하거나 모든 리전의 로그 파일을 하나의 Amazon S3 버킷에 집계할 수 있습니다.

CloudTrail에서 지원하는 AWS 서비스 목록은

<http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-supported-services.html>를 참조하십시오.

CloudTrail APN 파트너에 대한 자세한 내용은 다음을 참조하십시오.

- Splunk: <http://aws.amazon.com/cloudtrail/partners/splunk/>
- AlertLogic: <https://aws.amazon.com/cloudtrail/partners/alert-logic/>
- SumoLogic: <https://aws.amazon.com/cloudtrail/partners/sumo-logic/>

네트워크 모니터링 VPC 흐름 로그

VPC Flow Logs

- VPC의 **트래픽 흐름 세부 정보**를 캡처합니다.
- 허용, 거부 또는 모든 트래픽
- **VPC, 서브넷** 및 **ENI**에 대해 활성화될 수 있습니다.
- 로그는 **CloudWatch Logs**로 게시됩니다.
- 로그는 **Amazon S3**로 게시됩니다.

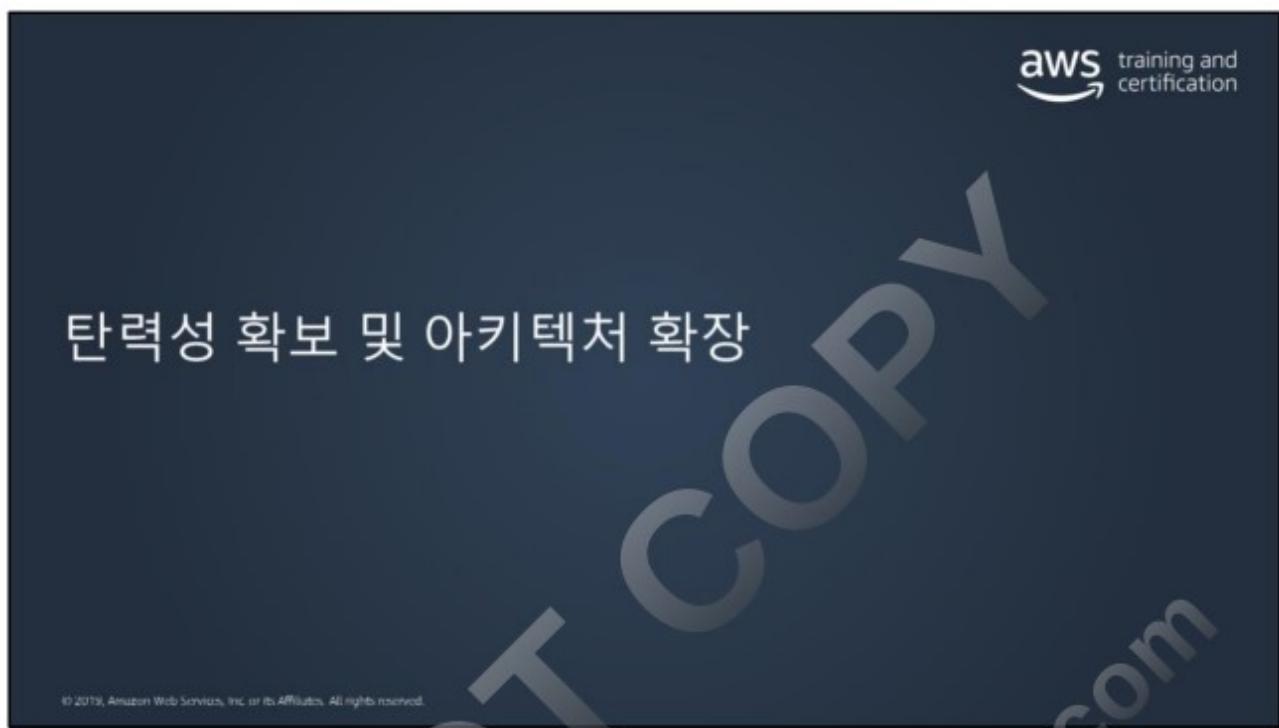
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

VPC Flow Logs는 VPC의 네트워크 인터페이스에서 송수신되는 IP 트래픽에 대한 정보를 캡처할 수 있게 해주는 기능입니다. 흐름 로그 데이터는 Amazon CloudWatch Logs를 통해 저장됩니다. 흐름 로그를 생성하고 난 후에는 Amazon CloudWatch Logs의 데이터를 확인하고 가져올 수 있습니다.

흐름 로그는 특정 트래픽이 인스턴스에 도달하지 않는 문제를 해결하는 등 다양한 작업에 도움을 주므로 과도하게 제한적인 보안 그룹 규칙을 진단할 수 있게 도와줍니다. 또한, 흐름 로그를 보안 도구로 사용하여 인스턴스에 도달하는 트래픽을 모니터링할 수 있습니다.

사용 사례:

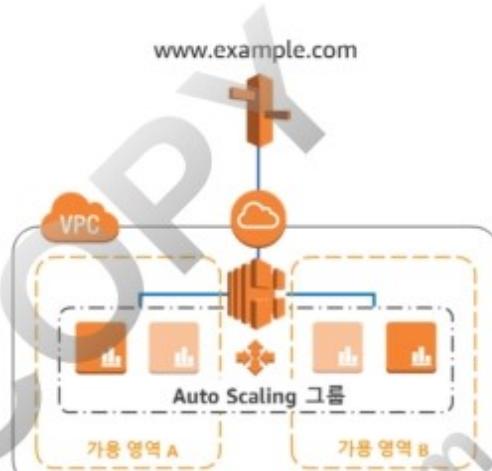
- 연결 문제 해결
- 네트워크 액세스 규칙 테스트
- 트래픽 모니터링
- 보안 인시던트 탐지 및 조사



Auto Scaling을 사용하여 탄력성 제공

 Amazon EC2 Auto Scaling

- 지정된 조건에 따라 인스턴스를 시작 또는 종료합니다.
- 지정된 경우, 새 인스턴스를 로드 밸런서에 자동으로 등록합니다.
- 여러 가용 영역에 걸쳐 시작할 수 있습니다.



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

조정 정책을 지정했다면, Auto Scaling에서는 애플리케이션의 늘어나거나 줄어드는 수요에 따라 인스턴스를 시작하거나 종료할 수 있습니다. Auto Scaling은 ELB와 통합되어 기존 Auto Scaling 그룹에 하나 이상의 로드 밸런서를 추가할 수 있습니다. 로드 밸런서를 연결한 후에는 로드 밸런서가 자동으로 인스턴스를 그룹에 등록하고 인스턴스 간에 수신 트래픽을 분산합니다.

하나의 가용 영역이 비정상 또는 사용 불가 상태가 되었을 때, Auto Scaling에서는 영향을 받지 않은 가용 영역에서 새 인스턴스를 시작합니다. 비정상 가용 영역이 정상 상태로 복귀하는 경우 Auto Scaling 그룹의 모든 가용 영역에 걸쳐 애플리케이션 인스턴스가 자동으로 고르게 재분배됩니다. 이는 최소의 인스턴스로 가용 영역에서 새 인스턴스를 시작하려고 하는 방식으로 Auto Scaling에 의해 수행됩니다. 하지만 시도가 실패하는 경우 성공할 때까지 Auto Scaling은 다른 가용 영역에서 시작을 계속 시도합니다.

자동 조정 방법

aws training and certification

예약

예측 가능한 워크로드에 적합

시간 또는 날짜를 기준으로 조정

사용 사례: 야간에 개발 및 테스트 인스턴스 종료

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

일정 기반 조정을 사용하면 알려진 부하 변경에 앞서 애플리케이션을 조정할 수 있습니다. 예를 들어 매주 수요일에 웹 애플리케이션 트래픽이 증가하고 목요일까지 높은 상태로 유지되다가 금요일에 줄어들기 시작합니다. 웹 애플리케이션의 예측 가능한 트래픽 패턴에 따라 조정 활동을 계획할 수 있습니다.

자동 조정 방법

aws training and certification

예약	동적
예측 가능한 워크로드에 적합	일반적 조정에 탁월
 시간 또는 날짜를 기준으로 조정	 대상 추적 지원
사용 사례: 야간에 개발 및 테스트 인스턴스 종료	사용 사례: CPU 사용률에 따라 조정

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.





Amazon EC2 Auto Scaling은 동일한 Auto Scaling 그룹(ASG) 내에서 여러 구매 옵션을 지원합니다. 단일 ASG 내에 스팟, 온디맨드 및 예약 인스턴스(RI)(청구서가 처리될 때까지는 온디맨드 인스턴스)를 포함할 수 있으므로 컴퓨팅 비용을 최대 90%까지 절감할 수 있습니다.

Amazon EC2 Fleet을 사용하면 원하는 ASG 용량을 구성하는 EC2 인스턴스 유형의 조합을 정의할 수 있습니다. 이 조합은 구매 옵션 중 각 유형의 비율로 정의됩니다. EC2 Auto Scaling은 ASG가 축소 또는 확장함에 따라 원하는 비용 최적화를 유지합니다. 혼합 플릿으로 구성된 ASG도 단일 플릿 ASG와 동일한 수명 주기 후크, 인스턴스 상태 확인, 예약 조정을 지원합니다.

인스턴스 유형 및 구매 모델을 혼합하여 ASG를 정의할 때 구성할 수 있는 옵션은 다음과 같습니다.

최고 스팟 가격: ASG 인스턴스의 최고 스팟 가격을 설정합니다.

스팟 할당 전략: 가용 영역 다양성에 따라 구성합니다. 단일 가용 영역에서 특정 인스턴스 유형에 대한 수요가 높을 때 특히 유용합니다.

(선택 사항) 온디맨드 기본: 초기 용량을 온디맨드 인스턴스로 구성합니다. 전체 용량을 구성하는 온디맨드 인스턴스 비율과 구분됩니다.

기본을 초과하는 온디맨드 비율: 초기 그룹에 추가하는 온디맨드 인스턴스 비율을 제어합니다.

혼합 플릿 구성은 RAM 및 vCPU 용량이 다른 다양한 EC2 인스턴스 유형과 조합할 수 있습니다. EC2 Auto Scaling은 원하는 용량에 맞춰 가장 낮은 가격의 조합을 자동으로 프로비저닝합니다.

DO NOT COPY
zlagusdbs@gmail.com

Auto Scaling 최소 용량

aws training and certification

Auto Scaling 그룹에서 다음을 정의:

- 원하는 용량
- 최소 용량
- 최대 용량

설정하기에 적절한 **최소** 용량은 어떻게 됩니까?

설정하기에 적절한 **최대** 용량은 어떻게 됩니까?

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



Auto Scaling 고려 사항

aws training and certification

- 여러 유형의 autoscaling을 결합해야 할 수 있음
- 단계 조정을 사용하여 아키텍처를 조정하려면 더 많은 작업이 필요할 수 있음
- 일부 아키텍처의 경우 둘 이상의 지표를 사용하여 조정해야 함(예: CPU 외의 추가 지표 사용)
- 조기에 빠르게 확장하고 시간이 지남에 따라 천천히 축소
- 수명 주기 후크 사용

Auto Scaling이 인스턴스를 시작 또는 종료할 때 사용자 지정 작업 수행

주의: 인스턴스가 시작 후 완전히 사용 가능한 상태가 되려면 몇 분 정도 걸릴 수 있습니다.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



읽기 전용 복제본으로 수평적 규모조정: Amazon RDS

The diagram illustrates the architecture of Amazon RDS Read Replicas. It shows a central orange square at the top connected to five blue cylinders below it, labeled 'R' (Read) four times and 'M' (Master) once. A red arrow labeled '읽기' (Read) points from the left towards the top, while a green arrow labeled '쓰기' (Write) points from the bottom towards the right. This visualizes how reads are handled by multiple replicas while writes are directed to the master database.

- 읽기 중심의 워크로드 처리를 위해 수평적으로 확장
- 보고서 오프로드
- 주의 사항:
 - 복제는 비동기식
 - 현재 Amazon Aurora, MySQL, MariaDB, PostgreSQL, Oracle에서 사용 가능

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

PostgreSQL 읽기 전용 복제본에는 특정 요구 사항이 있습니다. 자세한 내용은 다음을 참조하십시오.

http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_ReadRepl.html#USER_ReadRepl.PostgreSQL

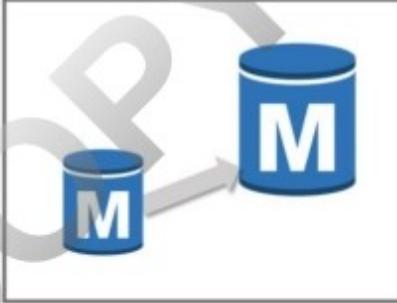
Oracle은 Dataguard로 읽기 전용 복제본을 지원합니다. 자세한 내용은 다음을 참조하십시오. <https://aws.amazon.com/about-aws/whats-new/2019/03/Amazon-RDS-for-Oracle-Now-Supports-In-region-Read-Replicas-with-Active-Data-Guard-for-Read-Scalability-and-Availability/>

Amazon RDS 규모조정: 버튼을 눌러 조정

aws training and certification

- 노드를 수직적으로 확장 또는 축소
- micro부터 24xlarge에 이르는 모든 크기 지원
- 종종 다운타임 없이 수직적으로 조정

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



Amazon RDS API를 사용하거나 콘솔에서 몇 번의 클릭만으로 컴퓨팅 및 메모리 리소스를 조정해 배포를 확장하거나 축소할 수 있습니다. 조정 작업은 일반적으로 몇 분 내에 완료됩니다. *일반적인 RDS는 조정 시 1~2분의 짧은 다운타임이 필요하지만 Aurora Serverless는 다운타임 없이 조정할 수 있습니다.

스토리지 요구 사항이 증가함에 따라 다운타임 없이 즉시 추가 스토리지를 프로비저닝할 수 있습니다. 또한 RDS의 PIOPS를 사용하면(SQL Server용 Amazon RDS 제외) IOPS 속도를 1,000 IOPS 단위로 1,000 IOPS에서 30,000 IOPS까지 지정하고 스토리지를 100GB에서 3TB까지 지정하여 DB 인스턴스의 처리량을 확장할 수 있습니다.

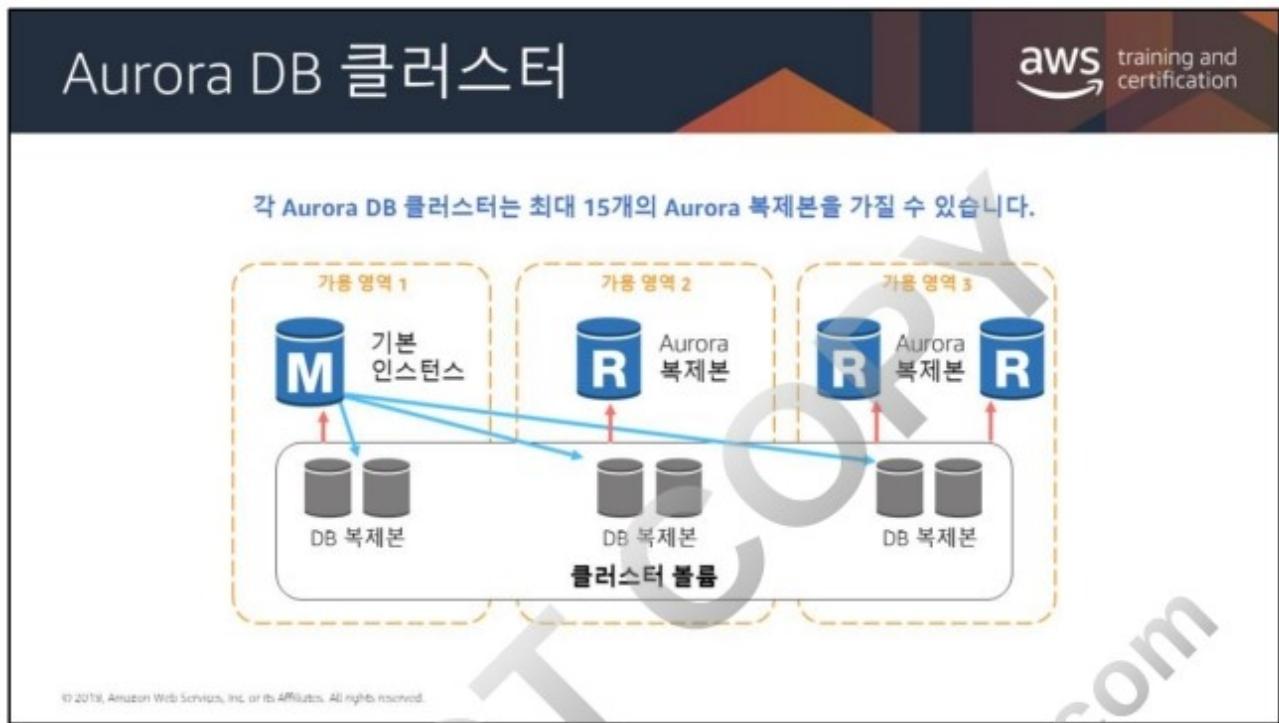
다운타임 없이 스토리지를 늘릴 수 있습니다. 그러나 인스턴스 유형을 변경하려면 다운타임이 필요합니다. 다음을 참조하십시오.

<https://aws.amazon.com/blogs/database/scaling-your-amazon-rds-instance-vertically-and-horizontally>

현재 SQL Server용 Amazon RDS에서는 기존 SQL Server DB 인스턴스의 스토리지 또는 IOPS 확장을 지원하지 않습니다.

대기 데이터베이스가 먼저 업그레이드된 다음 새로 크기가 조정된 데이터베이스로 장애 조치가 이루어지기 때문에 다중 가용 영역 환경에서 확장할 때 가동 중지 시간이 최소화됩니다. 단일 가용 영역 인스턴스의 경우 조정 작업 동안에는 인스턴스를 사용할 수 없습니다. DB 인스턴스 변경으로 인한 가동 중단 시간을 설명하는 표를 보려면

http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.DBInstance.Modifying.html#USER_ModifyInstance.Settings을 참조하십시오.



기본 인스턴스 – 읽기 및 쓰기 작업을 지원하고, 클러스터 볼륨의 모든 데이터 변경을 실행합니다. 각 Aurora DB 클러스터마다 기본 인스턴스가 하나씩 있습니다.

Aurora 복제본 – 읽기 작업만 지원합니다. 각 Aurora DB 클러스터마다 기본 인스턴스 이외에 최대 15개의 Aurora 복제본을 가질 수 있습니다. 다수의 Aurora 복제본이 읽기 워크로드를 분산시키면 Aurora 복제본을 별도의 가용 영역으로 이동시켜 데이터베이스 가용성을 높이는 것도 가능합니다.

읽기 전용 복제본은 마스터와 동일한 리전에 있을 수 있습니다.

Aurora Serverless

애플리케이션에 자동으로 응답합니다.

- 용량 조정
- 종료
- 시작

사용한 ACU 수에 따라 비용 지불

갑작스럽거나 예측할 수 없는 단기 워크로드에 적합합니다.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Amazon Aurora Serverless는 관계형 데이터베이스인 Amazon Aurora를 위한 온디맨드 Auto Scaling 구성입니다. Aurora Serverless DB Cluster는 데이터베이스 서버 인프라를 관리할 필요 없이 애플리케이션의 필요에 따라 자동으로 시작 및 종료하고 용량을 축소 또는 확장하는 DB 클러스터입니다.

Aurora Serverless는 사용 빈도가 낮거나 간헐적이거나 예측할 수 없는 워크로드를 위한 상대적으로 간단하고 비용 효율적인 옵션을 제공합니다. 자동으로 시작하고, 애플리케이션의 사용량에 맞춰 용량을 조정하고, 사용하지 않는 경우 종료되기 때문에 이러한 옵션을 제공할 수 있습니다. 최대 및 최소 Aurora 용량 단위(ACU)를 정의하고 사용한 ACU 수에 대해서만 지불합니다.

데이터베이스 샤딩으로 Amazon RDS 쓰기 조정

샤딩이 없으면 모든 데이터가 **하나의 파티션**에 상주합니다.

- 예: 하나의 데이터베이스에서 성이 A~Z에 속하는 사용자

샤딩은 데이터를 **큰 청크(샤드)**로 분할합니다.

- 예: 하나의 데이터베이스에서 성이 A~M에 속하는 사용자, 다른 데이터베이스에서 N~Z에 속하는 사용자
- 사용자

대부분의 경우에 샤딩은 **뛰어난 성능과 높은 운영 효율성**을 제공합니다.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

샤딩은 데이터베이스 서버를 여러 대 사용하여 쓰기 성능을 개선하는 기술입니다. 기본적으로 동일한 구조를 가진 데이터베이스는 쓰기 프로세스를 분산할 수 있도록 적절한 테이블 열을 키로 사용하여 준비되고 분할됩니다. AWS 클라우드에서 제공하는 RDBMS 서비스를 사용하면 이러한 샤딩을 수행하여 가용성과 운영 효율성을 높일 수 있습니다.

샤딩 백엔드 데이터베이스로 Amazon RDS를 사용할 수 있습니다. MySQL Server와 같은 샤딩 소프트웨어를 Spider Storage Engine과 결합하여 Amazon EC2 인스턴스에 설치합니다. 여러 RDS를 준비하고 이를 샤딩 백엔드 데이터베이스로 사용합니다. RDS를 여러 리전에 배포할 수 있습니다.

자세한 내용은 다음을 참조하십시오.

http://en.clouddesigntpattern.org/index.php/CDP:Sharding_Write_Pattern.

DynamoDB - 두 가지 조정

aws training and certification

Auto Scaling

모든 새 테이블의 기본값



상한 및 하한 지정

사용 사례: 일반 조정, 대부분의 애플리케이션에 적합한 솔루션.

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

콘솔을 사용하여 새 DynamoDB 테이블을 생성하는 경우, 테이블에 Auto Scaling이 기본적으로 활성화됩니다. DynamoDB Auto Scaling은 동적으로 변동하는 요청 볼륨에 대응하여 가동 중단 없이 읽기 및 쓰기 처리량 용량을 자동으로 조정합니다. DynamoDB Auto Scaling을 사용하면 사용자가 원하는 처리량 사용률 목표, 최소 및 최대 한도만 설정하면 Auto Scaling이 나머지를 자동으로 처리합니다.

DynamoDB Auto Scaling은 Amazon CloudWatch와 연동하여 지속적으로 실제 처리량 사용을 모니터링하다가 실제 사용률이 목표에서 벗어날 경우 자동으로 용량을 확장 또는 축소합니다. Auto Scaling은 신규 테이블, 기존 테이블 및 글로벌 보조 인덱스에 대해 활성화할 수 있습니다. 콘솔에서 몇 번의 클릭으로 Auto Scaling을 활성화할 수 있으며, 콘솔을 통해 모든 조정 활동을 확인할 수 있습니다. 또한 AWS 명령줄 인터페이스 및 AWS 소프트웨어 개발 키트를 사용해 프로그래밍 방식으로 DynamoDB Auto Scaling을 관리할 수도 있습니다.

DynamoDB Auto Scaling을 사용하는 데는 DynamoDB 및 CloudWatch 경보에 대해 이미 지불하고 있는 비용 외에 추가 비용이 들지 않습니다. DynamoDB Auto Scaling은 모든 AWS 리전에서 사용할 수 있으며, 즉시 적용됩니다.

DynamoDB - 두 가지 조정

aws training and certification

Auto Scaling

모든 새 테이블의 기본값



상한 및 하한 지정

사용 사례: 일반 조정, 대부분의 애플리케이션에 적합한 솔루션.

온디맨드

요청당 지불



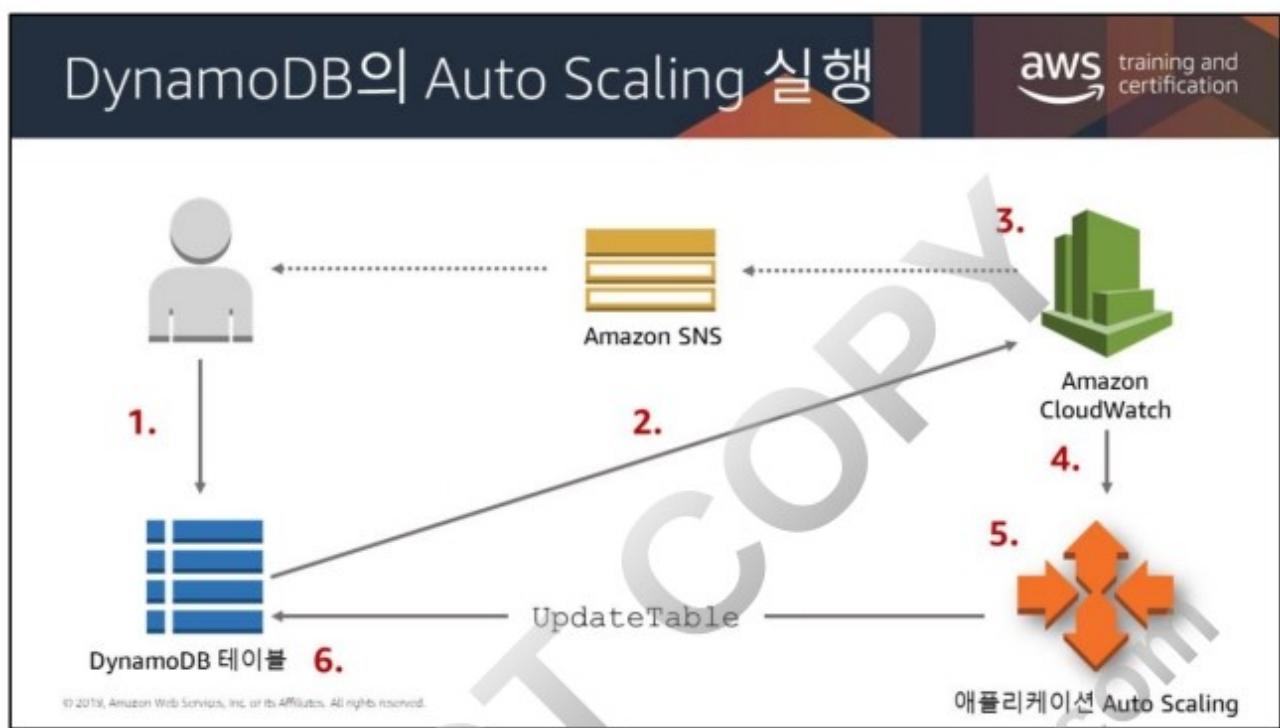
프로비저닝 없음

사용 사례: 갑작스럽고 예측할 수 없는 워크로드, 빠르게 용량이 필요한 경우.

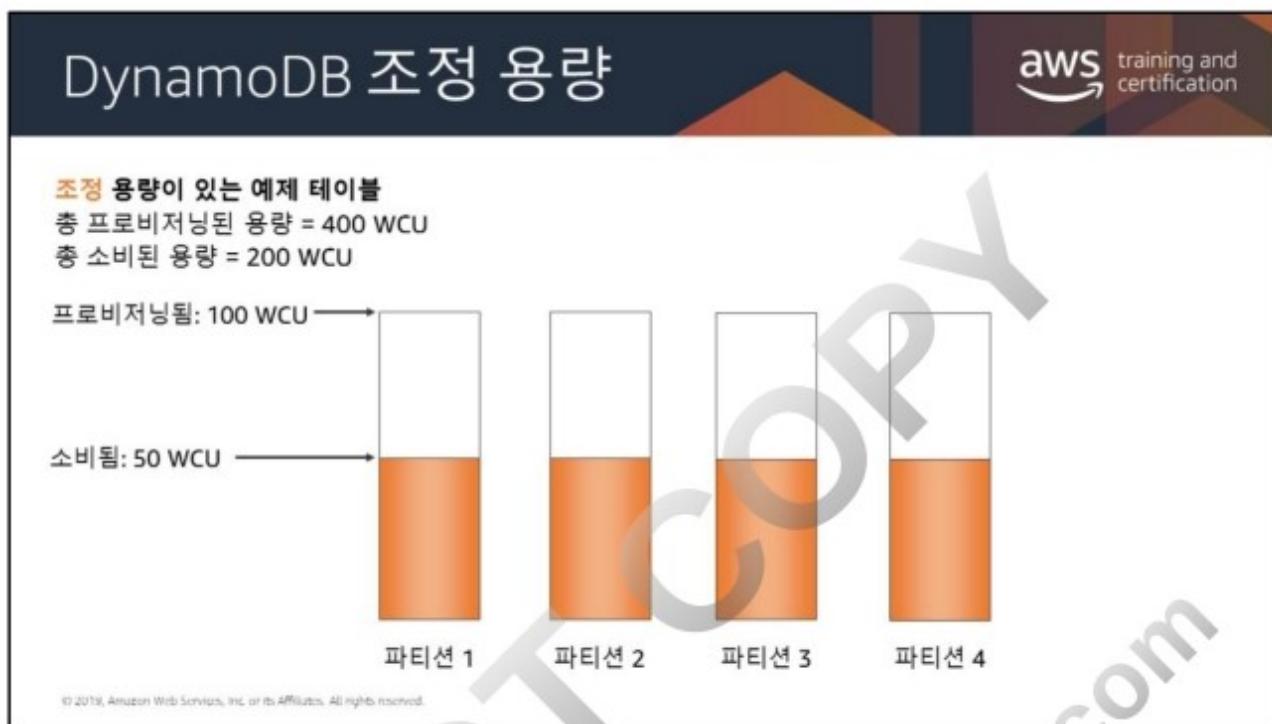
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Amazon DynamoDB On-Demand는 용량 계획 없이 초당 수천 건의 요청을 처리할 수 있는 DynamoDB의 유연한 결제 옵션입니다. 프로비저닝 요금 모델 대신 요청당 요금으로 전환합니다. DynamoDB On-Demand는 모든 트래픽 수준의 증가 또는 조정을 관찰할 수 있습니다. 트래픽 수준이 새로운 피크에 도달하는 경우, DynamoDB는 워크로드에 맞춰 신속하게 조정됩니다. 워크로드 예측이 어렵거나 짧은 시간 동안 대규모 스파이크가 있는 경우에 적합합니다. 하루에 한 번 프로비저닝 용량에서 온디맨드로 테이블을 변경할 수 있습니다. 온디맨드 용량에서 프로비저닝 용량으로는 자유롭게 변경할 수 있습니다.

<https://aws.amazon.com/blogs/aws/amazon-dynamodb-on-demand-no-capacity-planning-and-pay-per-request-pricing/>



1. DynamoDB 테이블의 애플리케이션 Auto Scaling 정책을 생성합니다.
2. DynamoDB가 사용 용량 지표를 Amazon CloudWatch에 게시합니다.
3. 테이블에서 사용한 용량이 특정 기간 동안의 목표 사용률을 초과하는 경우(또는 목표에 미달하는 경우), Amazon CloudWatch가 경보를 트리거합니다. 콘솔에서 경보를 보고 Amazon SNS를 사용하여 알림을 수신할 수 있습니다.
4. CloudWatch 경보가 애플리케이션 Auto Scaling을 호출하여 조정 정책을 평가합니다.
5. 애플리케이션 Auto Scaling이 UpdateTable 요청을 생성하여 테이블의 프로비저닝된 처리량을 조정합니다.
6. DynamoDB는 UpdateTable 요청을 처리하고 해당 테이블의 할당된 처리 용량을 동적으로 늘리거나 줄임으로써 목표 사용률에 근접하게 합니다.



항상 읽기와 및 쓰기 작업을 골고루 배포할 수 있는 것은 아닙니다. 데이터 액세스가 불균형할 때, “핫” 파티션은 다른 파티션보다 볼륨이 많은 읽기 및 쓰기 트래픽을 받을 수 있습니다. 극단적인 상황에서는 단일 파티션이 3,000 RCU나 1,000 WCU 이상을 수신하는 경우 조절이 발생할 수도 있습니다.

고르지 못한 액세스 패턴을 더 효과적으로 수용하기 위해 DynamoDB 조정 용량을 사용하면, 애플리케이션은 트래픽이 테이블의 프로비저닝된 용량이나 파티션 최대 용량을 초과하지 않을 경우 조절 없이 핫 파티션에 계속 읽기 및 쓰기 작업을 수행할 수 있습니다. 조정 용량은 더 많은 트래픽을 받는 파티션의 처리량 용량을 자동으로 증가시킵니다.

모든 DynamoDB 테이블에서 자동으로 조정 용량이 활성화되어 있기 때문에, 이를 명시적으로 활성화하거나 비활성화할 필요가 없습니다.

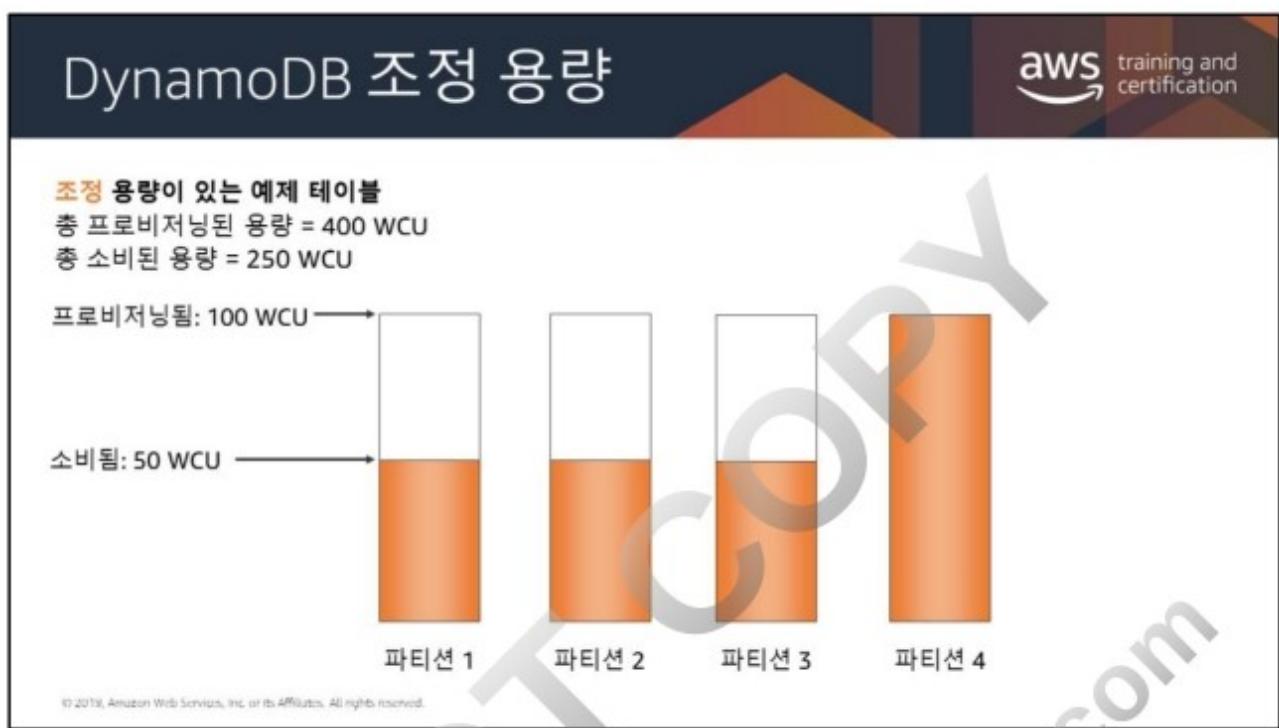
다음 다이어그램에는 조정 용량 작동 방식이 나와 있습니다. 예제 테이블은 4개의 파티션에 균일하게 공유된 400 WCU(쓰기 용량 단위)로 프로비저닝되어 있어 각 파티션은 초당 최대 100 WCU를 유지할 수 있습니다. 파티션 1, 2, 3은 각각 50 WCU/초의 쓰기 트래픽을 수신하는 반면, 파티션 4는 150 WCU/초를 수신합니다. 이 핫 파티션은 미사용 버스트 용량이 있는 동안 쓰기 트래픽을 수락할 수 있지만, 결국 100 WCU/초를 초과하는 트래픽을 제한합니다.

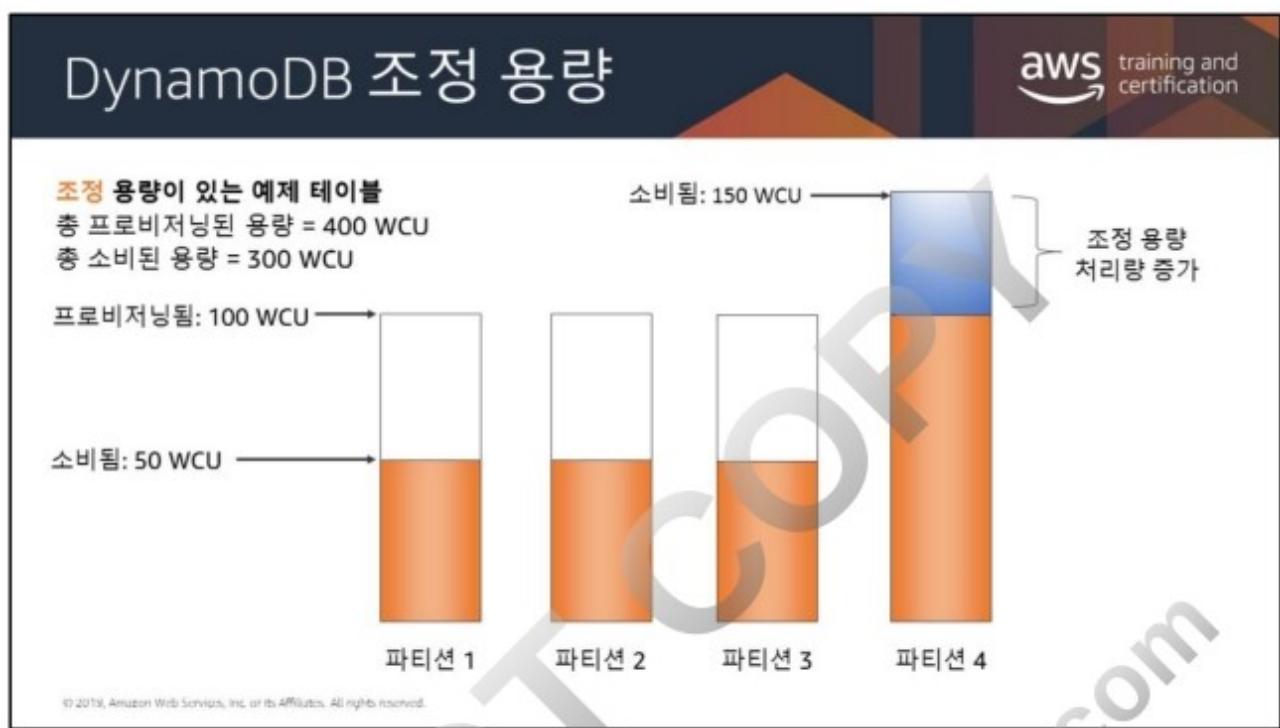
DynamoDB 조정 용량은 파티션 4의 용량을 늘리는 것으로 대응하므로 해당 파티션이 제한되지 않고 150 WCU/초의 높은 워크로드를 유지할 수 있습니다.

자세한 내용은 다음을 참조하십시오.

<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-partition-key-design.html>.

DO NOT COPY
zlagusdbs@gmail.com





파티션 키 값	균등성
사용자 ID(애플리케이션에 사용자가 많은 경우)	좋음
상태 코드(상태 코드가 단 몇 개만 가능한 경우)	나쁨
항목 생성 날짜, 가장 가까운 시간으로 반올림(예: 일, 시 또는 분)	나쁨
디바이스 ID(각 디바이스가 비교적 유사한 간격으로 데이터를 액세스하는 경우)	좋음
디바이스 ID(많은 디바이스가 추적되고 있긴 하지만, 하나만 사용량이 매우 많은 경우)	나쁨

ID 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

테이블 기본 키의 파티션 키 부분은 테이블의 데이터가 저장되는 논리적 파티션을 결정합니다. 이는 기본 물리적 파티션에 영향을 줍니다. 테이블의 프로비저닝된 I/O 용량은 이 물리적 파티션으로 고르게 분배됩니다. 따라서 I/O 요청을 고르게 분산시키지 않는 파티션 키 설계는 "핫" 파티션을 발생시킬 수 있으며, 이는 조절과 프로비저닝된 I/O 용량을 비효율적으로 사용하게 되는 문제를 초래합니다.

테이블의 프로비저닝된 처리량의 최적 사용량은 개별 항목의 워크로드 패턴과 파티션-키 설계가 결정합니다. 이것이 모든 파티션 키 값에 액세스하여 효율적인 처리량 수준을 달성해야 한다는 의미는 아닙니다. 또한 액세스된 파티션 키 값의 백분율이 높아야 한다는 의미는 더더욱 아닙니다. 그 의미는 워크로드가 액세스하는 고유 파티션 키 값이 많을 수록 요청이 여러 파티션 공간으로 더 많이 분산된다는 것입니다. 일반적으로 총 파티션 키 값 중 액세스한 파티션 키 값의 비율이 증가할수록 처리량을 보다 효율적으로 활용할 수 있습니다.

자세한 내용은 다음을 참조하십시오.

<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-partition-key-uniform-load.html>



실습 4: 고가용성 환경 생성

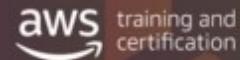
"복원력이 뛰어난 인프라를 원합니다."

사용된 기술:

- Amazon VPC
- Application Load Balancer
- Amazon EC2 Auto Scaling 그룹
- Amazon RDS

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

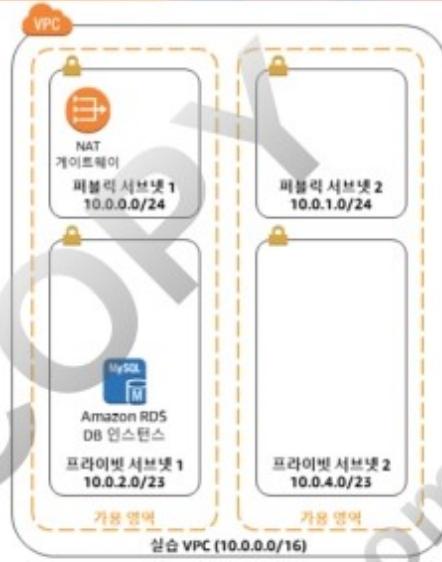
실습 4: 고가용성 환경 생성



실습 시작 시 제공됨:

- 2개의 가용영역에 걸친 VPC
- 2개의 퍼블릭 서브넷
- 2개의 프라이빗 서브넷
- 1개의 NAT 게이트웨이
- Amazon RDS DB 인스턴스

고가용성을 얻을 수 있습니다!



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

실습 4: 고가용성 환경 생성

aws training and certification

요청을 여러 서버에 분산하려면 다음을 사용합니다.

- Amazon EC2 Auto Scaling 그룹
- 로드 밸런서



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

실습 4: 고가용성 환경 생성

aws training and certification

로드 밸런서는 **퍼블릭 서브넷**에 분산됩니다.

애플리케이션 서버는 **프라이빗 서브넷**에 있습니다.



© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

실습 4: 고가용성 환경 생성

aws training and certification

3티어 아키텍처를 생성합니다.

보안 그룹은 각 계층 간에 추가 보안을 제공합니다.

```
graph LR; Internet[인터넷] -- "HTTP + HTTPS 트래픽 허용" --> ALB[애플리케이션 로드 밸런서]; ALB -- "HTTP 트래픽 허용" --> AS[앱 서버]; AS -- "MySQL 트래픽 허용" --> RDS[Amazon RDS MySQL DB 인스턴스];
```

인터넷 → HTTP + HTTPS 트래픽 허용 → 애플리케이션 로드 밸런서 보안 그룹 → HTTP 트래픽 허용 → 앱 서버 보안 그룹 → MySQL 트래픽 허용 → 데이터베이스 보안 그룹

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

실습 4: 고가용성 환경 생성

aws training and certification

최종 구성:

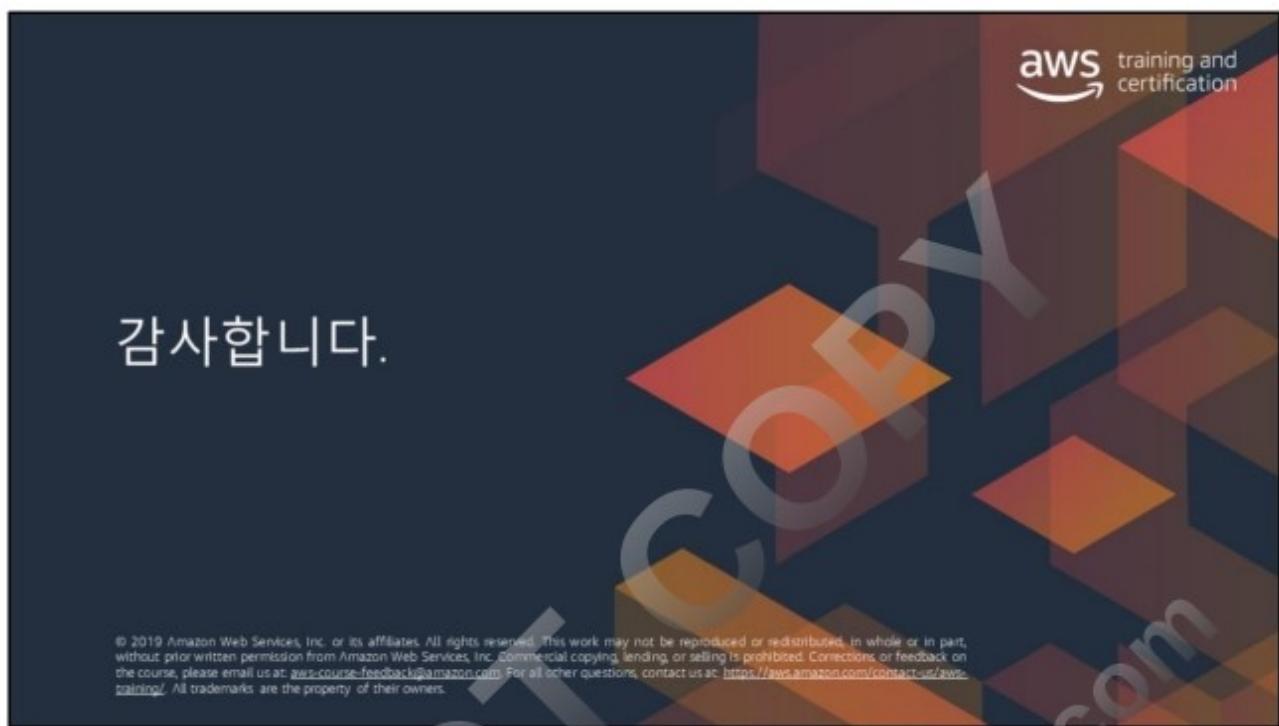
- 로드 밸런서
- 여러 애플리케이션 서버
- 다중 가용 영역 데이터베이스
- 각 가용 영역의 NAT 게이트웨이

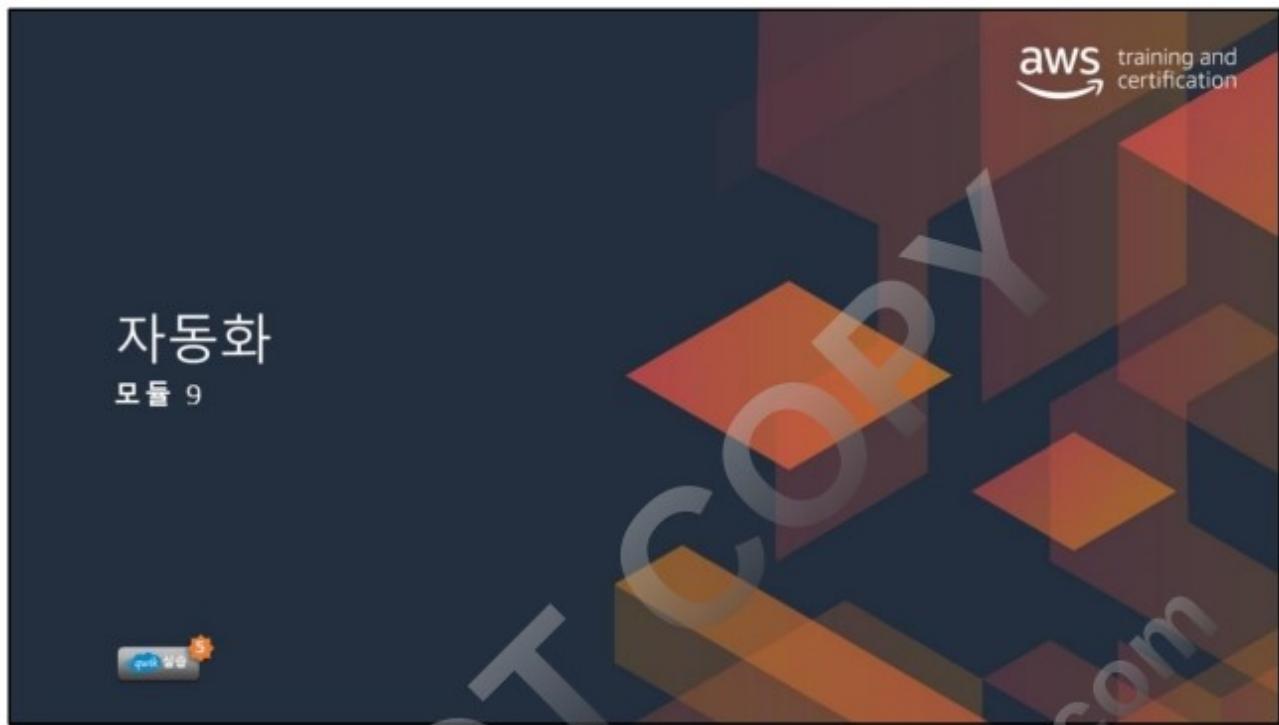
확장성, 안정성, 고가용성!

시간: 40분

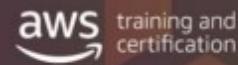
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

The diagram illustrates a VPC (Virtual Private Cloud) architecture designed for high availability and redundancy. It features two separate availability zones (AZ1 and AZ2), each enclosed in a dashed orange border. Within each zone, there is a NAT gateway (represented by an orange icon with a port symbol) and an application load balancer (represented by a blue icon with a port symbol). Both the NAT gateways and application load balancers are connected to their respective public subnets. Each zone also contains multiple application servers (represented by orange squares). An Auto Scaling group (represented by a dashed line connecting two servers) spans both zones. Additionally, there are MySQL DB instances (represented by blue icons with a 'M' symbol) located in AZ1, and an Amazon RDS multi-region backup (represented by a blue cylinder with an 'S' symbol) located in AZ2. The private subnets (10.0.2.0/23 and 10.0.4.0/23) are connected to their respective public subnets (10.0.0.0/24 and 10.0.1.0/24) via the NAT gateways. The entire VPC is identified as '실습 VPC (10.0.0.0/16)' at the bottom.





모듈 9



아키텍처 측면에서의 필요성

지속적 성장을 위해서는 자동화를 시작해야 합니다. 조직에 있는 다양한 아키텍처를 일관되게 배포, 관리, 업데이트할 방법이 필요합니다.

모듈 개요

- 자동화가 필요한 이유
- 인프라 자동화
- 배포 자동화

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

