에서 증강 생성 옵션 및 아키텍처 검색 AWS

# AWS 권장 가이드

# AWS 권장 가이드: 에서 증강 생성 옵션 및 아키텍처 검색 AWS

# Table of Contents

# 에서 증강 생성 옵션 및 아키텍처 검색 AWS

Mithil Shah, Rajeev Muralidhar 및 Natacha Fort, Amazon Web Services

2024년 10월(문서 기록)

생성형 AI는 간단한 텍스트 프롬프트에서 이미지, 비디오, 텍스트, 오디오와 같은 새로운 콘텐츠와 아티팩트를 생성할 수 있는 AI 모델의 하위 집합을 말합니다. 생성형 AI 모델은 광범위한 주제와 작업을 포괄하는 방대한 양의 데이터를 기반으로 훈련됩니다. 이를 통해 명시적으로 훈련되지 않은 작업이라도 다양한 작업을 수행할 수 있는 뛰어난 다양성을 입증할 수 있습니다. 단일 모델이 여러 작업을 수행할 수 있기 때문에 이러한 모델을 파운데이션 모델(FMs)이라고 하는 경우가 많습니다.

생성형 AI 모델의 주목할 만한 애플리케이션 중 하나는 질문에 답하는 데 능숙하다는 것입니다. 그러나 이러한 모델을 사용하여 사용자 지정 문서를 기반으로 질문에 답변할 때 발생하는 특정 문제가 있습니다. 사용자 지정 문서에는 독점 정보, 내부 웹 사이트, 내부 설명서, Confluence 페이지, SharePoint 페이지 등이 포함될 수 있습니다. 한 가지 옵션은 검색 증강 생성(RAG)을 사용하는 것입니다. RAG를 사용하면 파운데이션 모델은 응답을 생성하기 전에 훈련 데이터 소스(예: 사용자 지정 문서) 외부에 있는 신뢰할 수 있는 데이터 소스를 참조합니다.

이 가이드에서는 검색 증강 생성(RAG) 시스템을 포함하여 사용자 지정 설명서의 질문에 답변하는 데 사용할 수 있는 고유한 생성형 AI 옵션을 설명합니다. 또한 Amazon Web Services()에서 RAG 시스템을 구축하는 방법에 대한 개요도 제공합니다AWS. RAG 옵션 및 아키텍처를 검토하여의 완전 관리형 서비스와 AWS 사용자 지정 RAG 아키텍처 중에서 선택할 수 있습니다.

## 대상 독자

이 가이드의 대상은 RAG 솔루션을 구축하고, 사용 가능한 아키텍처를 검토하고, 각 옵션의 이점과 단점을 이해하려는 생성형 AI 아키텍트와 관리자입니다.

## 목표

이 가이드는 다음을 수행하는 데 도움이 됩니다.

- 사용자 지정 문서의 질문에 답변하는 데 사용할 수 있는 생성형 AI 옵션 이해
- 에서 RAG 시스템의 아키텍처 옵션 검토 AWS
- 각 RAG 옵션의 장단점 이해
- AWS 환경에 맞는 RAG 아키텍처 선택

# 사용자 지정 문서 쿼리를 위한 생성형 AI 옵션

조직에는 다양한 정형 및 비정형 데이터 소스가 있는 경우가 많습니다. 이 가이드는 생성형 AI를 사용하여 비정형 데이터의 질문에 답변하는 방법을 중점적으로 다룹니다.

조직의 비정형 데이터는 다양한 소스에서 가져올 수 있습니다. 여기에는 PDFs, 텍스트 파일, 내부 Wiki, 기술 문서, 공개 웹 사이트, 지식 기반 등이 포함될 수 있습니다. 비정형 데이터에 대한 질문에 답변할 수 있는 파운데이션 모델을 원하는 경우 다음 옵션을 사용할 수 있습니다.

- 사용자 지정 문서 및 기타 훈련 데이터를 사용하여 새 파운데이션 모델 훈련
- 사용자 지정 문서의 데이터를 사용하여 기존 파운데이션 모델 미세 조정
- 질문을 할 때 컨텍스트 내 학습을 사용하여 파운데이션 모델에 문서를 전달합니다.
- 검색 증강 생성(RAG) 접근 방식 사용

사용자 지정 데이터가 포함된 새로운 파운데이션 모델을 처음부터 훈련시키는 것은 야심찬 작업입니다. BloombergGPT 모델Bloomberg과 같이 성공적으로 수행한 회사는 몇 개입니다. 또 다른 예는 텍스트와 함께 6LG AI Research,000 억 개의 아트워크와 2 억 5천만 개의 고해상도 이미지를 사용하여 훈련된의 멀티모달 EXAONE 모델입니다. The Cost of AI: Should You Build or Buy Your Foundation Model(LinkedIn)에 따르면 교육 Meta Llama 2 비용은 약 480 만 USD입니다. 모델을 처음부터 훈련하기 위한 두 가지 주요 사전 조건은 리소스(재무, 기술, 시간)에 대한 액세스와 명확한 투자 수익입니다. 이것이 적합한 것 같지 않으면 다음 옵션은 기존 파운데이션 모델을 미세 조정하는 것입니다.

기존 모델을 미세 조정하려면 Amazon Titan, Mistral 또는 Llama 모델과 같은 모델을 만든 다음 모델을 사용자 지정 데이터에 맞게 조정해야 합니다. 미세 조정에는 다양한 기법이 있으며, 대부분 모델의 모든 파라미터를 수정하는 대신 몇 개의 파라미터만 수정하는 것이 포함됩니다. 이를 파라미터 효율적인 미세 조정이라고 합니다. 미세 조정에는 두 가지 기본 방법이 있습니다.

- 감독 미세 조정은 레이블이 지정된 데이터를 사용하며 새로운 종류의 작업에 맞게 모델을 훈련하는 데 도움이 됩니다. 예를 들어 PDF 양식을 기반으로 보고서를 생성하려는 경우 충분한 예제를 제공하여 모델을 교육해야 할 수 있습니다.
- 비지도 미세 조정은 작업에 구애받지 않으며 파운데이션 모델을 자체 데이터에 맞게 조정합니다. 문서의 컨텍스트를 이해하도록 모델을 훈련합니다. 그런 다음 미세 조정된 모델은 조직의 사용자 지정 스타일을 사용하여 보고서와 같은 콘텐츠를 생성합니다.

그러나 미세 조정은 질문 답변 사용 사례에는 적합하지 않을 수 있습니다. 자세한 내용은이 안내서의 RAG 비교 및 미세 조정을 참조하세요.

질문을 하면 파운데이션 모델을 문서로 전달하고 모델의 컨텍스트 내 학습을 사용하여 문서에서 답변을 반환할 수 있습니다. 이 옵션은 단일 문서의 임시 쿼리에 적합합니다. 그러나이 솔루션은 여러 문서를 쿼리하거나 Microsoft SharePoint 또는 Atlassian Confluence와 같은 시스템 및 애플리케이션을 쿼리하는 데는 적합하지 않습니다.

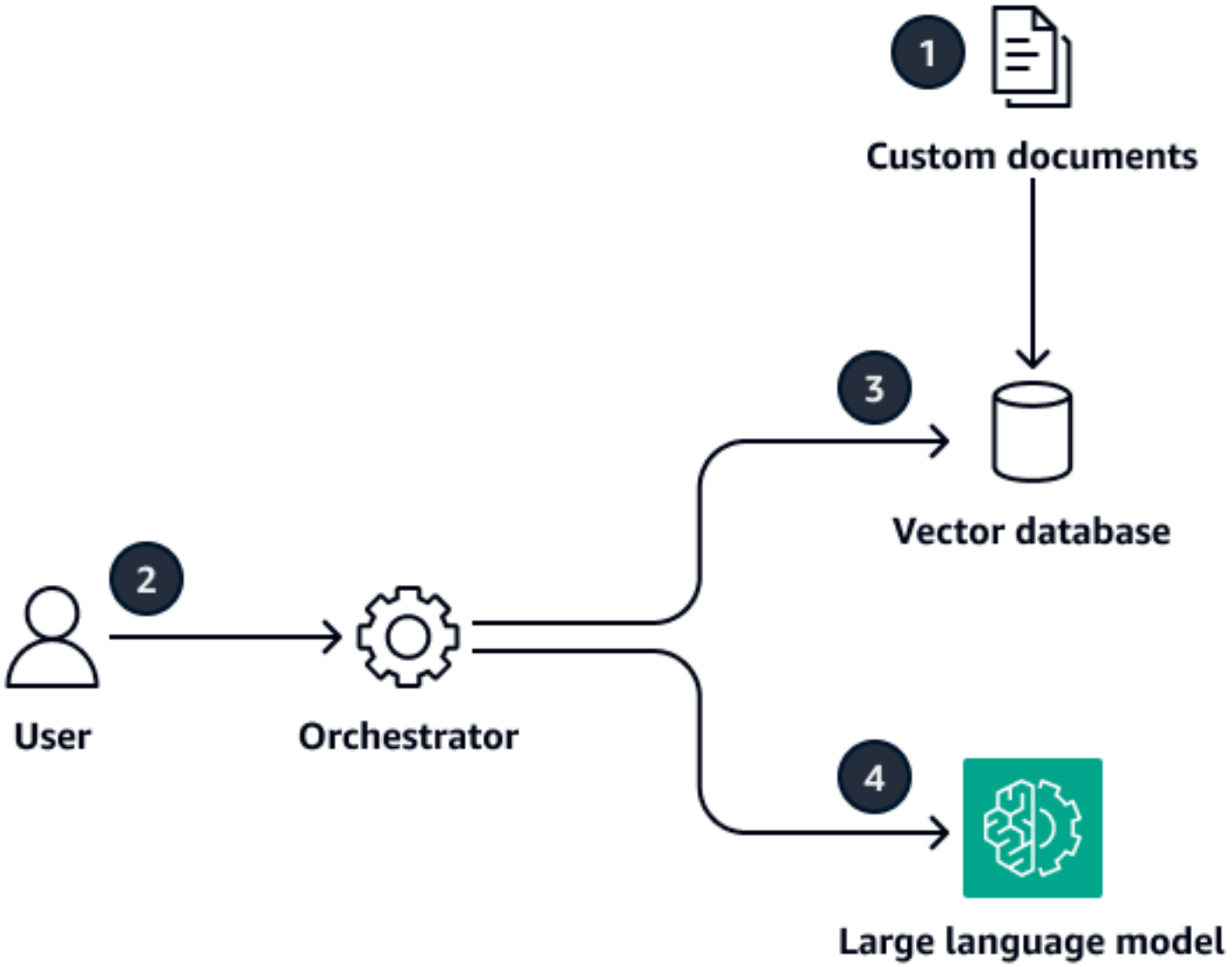
마지막 옵션은 RAG를 사용하는 것입니다. RAG를 사용하면 파운데이션 모델은 응답을 생성하기 전에 사용자 지정 문서를 참조합니다. RAG는 모델을 재학습할 필요 없이 모델의 기능을 조직의 내부 지식 기반까지 확장합니다. 다양한 컨텍스트에서 관련성, 정확성 및 유용성을 유지하도록 모델 출력을 개선하는 비용 효율적인 접근 방식입니다.

이 섹션의 주제:



## 검색 증강 생성 이해

검색 증강 생성(RAG)은 회사의 내부 문서와 같은 외부 데이터로 대규모 언어 모델(LLM)을 보강하는 데 사용되는 기술입니다. 이를 통해 모델은 특정 사용 사례에 맞는 정확하고 유용한 출력을 생성하는 데 필요한 컨텍스트를 얻을 수 있습니다. RAG는 엔터프라이즈에서 LLMs을 사용하는 실용적이고 효과적인 접근 방식입니다. 다음 다이어그램은 RAG 접근 방식의 작동 방식에 대한 개략적인 개요를 보여줍니다.




일반적으로 RAG 프로세스는 4단계입니다. 첫 번째 단계는 한 번 수행되고 나머지 세 단계는 필요한 만큼 여러 번 수행됩니다.

1. 임베딩을 생성하여 내부 문서를 벡터 데이터베이스로 수집합니다. 임베딩은 데이터의 의미 또는 컨텍스트 의미를 캡처하는 문서의 텍스트를 숫자로 표현한 것입니다. 벡터 데이터베이스는 기본적으로 이러한 임베딩의 데이터베이스이며 벡터 저장소 또는 벡터 인덱스라고도 합니다. 이 단계에서는 데이터 정리, 형식 지정 및 청킹이 필요하지만 일회성 선결제 활동입니다.
2. 사람이 자연어로 쿼리를 제출합니다.
3. 오케스트레이터는 벡터 데이터베이스에서 유사성 검색을 수행하고 관련 데이터를 검색합니다. 오케스트레이터는 검색된 데이터( 컨텍스트라고도 함)를 쿼리가 포함된 프롬프트에 추가합니다.
4. 오케스트레이터는 쿼리와 컨텍스트를 LLM으로 전송합니다. LLM은 추가 컨텍스트를 사용하여 쿼리에 대한 응답을 생성합니다.

사용자의 관점에서 RAG는 모든 LLM과 상호 작용하는 것처럼 보입니다. 그러나 시스템은 문제의 콘텐츠에 대해 훨씬 더 많이 알고 있으며 조직의 지식 기반에 맞게 미세 조정된 답변을 제공합니다.

RAG 접근 방식의 작동 방식에 대한 자세한 내용은 웹 사이트의 RAG란 무엇입니까?를 참조하세요 AWS .

## 프로덕션 수준 RAG 시스템의 구성 요소

프로덕션 수준 RAG 시스템을 구축하려면 RAG 워크플로의 여러 측면을 고려해야 합니다. 개념적으로 프로덕션 수준 RAG 워크플로에는 특정 구현에 관계없이 다음과 같은 기능과 구성 요소가 필요합니다.

- 커넥터 - 다양한 엔터프라이즈 데이터 소스를 벡터 데이터베이스와 연결합니다. 구조화된 데이터 소스의 예로는 트랜잭션 및 분석 데이터베이스가 있습니다. 비정형 데이터 소스의 예로는 객체 스토어, 코드 베이스 및 서비스형 소프트웨어(SaaS) 플랫폼이 있습니다. 각 데이터 소스에는 서로 다른 연결 패턴, 라이선스 및 구성이 필요할 수 있습니다.
- 데이터 처리 - 데이터는 PDFs, 스캔한 이미지, 문서, 프레젠테이션 및 Microsoft SharePoint 파일과 같은 다양한 형태와 형태로 제공됩니다. 데이터 처리 기법을 사용하여 인덱싱할 데이터를 추출, 처리 및 준비해야 합니다.
- 임베딩 - 관련성 검색을 수행하려면 문서와 사용자 쿼리를 호환되는 형식으로 변환해야 합니다. 임베딩 언어 모델을 사용하면 문서를 수치 표현으로 변환할 수 있습니다. 이는 기본적으로 기본 파운데이션 모델의 입력입니다.
- 벡터 데이터베이스 - 벡터 데이터베이스는 임베딩, 관련 텍스트 및 메타데이터의 인덱스입니다. 인덱스는 검색 및 검색에 최적화되어 있습니다.
- 리트리버 - 사용자 쿼리의 경우 리트리버는 벡터 데이터베이스에서 관련 컨텍스트를 가져오고 비즈니스 요구 사항에 따라 응답의 순위를 매깁니다.
- 파운데이션 모델 - RAG 시스템의 파운데이션 모델은 일반적으로 LLM입니다. 파운데이션 모델은 컨텍스트와 프롬프트를 처리하여 사용자에 대한 응답을 생성하고 형식을 지정합니다.
- 가드레일 - 가드레일은 쿼리, 프롬프트, 검색된 컨텍스트 및 LLM 응답이 정확하고 책임감 있으며 윤리적이고 할루시네이션 및 편향이 없도록 설계되었습니다.
- 오케스트레이터 - 오케스트레이터는 end-to-end 워크플로를 예약하고 관리할 책임이 있습니다.
- 사용자 경험 - 일반적으로 사용자는 채팅 기록 표시 및 응답에 대한 사용자 피드백 수집 등 다양한 기능을 갖춘 대화형 채팅 인터페이스와 상호 작용합니다.
- 자격 증명 및 사용자 관리 - 애플리케이션에 대한 사용자 액세스를 세밀하게 제어하는 것이 중요합니다. 에서 AWS 클라우드정책, 역할 및 권한은 일반적으로 AWS Identity and Access Management (IAM)을 통해 관리됩니다.

분명히 RAG 시스템을 계획, 개발, 릴리스 및 관리하기 위한 많은 작업이 있습니다. Amazon Bedrock 또는 Amazon Q Business와 같은 완전관리형 서비스는 차별화되지 않은 일부 과중한 작업을 관리하는 데 도움이 될 수 있습니다. 그러나 사용자 지정 RAG 아키텍처는 리트리버 또는 벡터 데이터베이스와 같은 구성 요소를 더 잘 제어할 수 있습니다.

## 검색 증강 생성 및 미세 조정 비교

다음 표에서는 미세 조정 및 RAG 기반 접근 방식의 장단점을 설명합니다.

| 접근 방식 | 장점 | 단점 |
| --- | --- | --- |
| 미세 조정 | • 미세 조정된 모델이 비지도 접근 방식을 사용하여 훈련된 경우 조직의 스타일과 더 일치하는 콘텐츠를 생성할 수 있습니다.<br>• 독점 또는 규제 데이터를 기반으로 훈련된 미세 조정된 모델은 조직이 사내 또는 업계별 데이터 및 규정 준수 표준을 따르는 데 도움이 될 수 있습니다. | • 미세 조정은 모델 크기에 따라 몇 시간에서 며칠이 걸릴 수 있습니다. 따라서 사용자 지정 문서가 자주 변경되는 경우에는 좋은 솔루션이 아닙니다.<br>• 미세 조정을 수행하려면 순위가 낮은 조정(LoRA) 및 파라미터 효율성이 높은 미세 조정(PEFT)과 같은 기술을 이해해야 합니다. 미세 조정에는 데이터 과학자가 필요할 수 있습니다.<br>• 일부 모델에서는 미세 조정을 사용하지 못할 수 있습니다.<br>• 미세 조정된 모델은 응답에서 소스에 대한 참조를 제공하지 않습니다.<br>• 미세 조정된 모델을 사용하여 질문에 답하면 할루시네이션 위험이 증가할 수 있습니다. |

| 접근 방식 | 장점 | 단점 |
| --- | --- | --- |
| RAG | • RAG를 사용하면 미세 조정 없이 사용자 지정 문서에 대한 질문 응답 시스템을 구축할 수 있습니다.<br>• RAG는 몇 분 안에 최신 문서를 통합할 수 있습니다.<br>• AWS 는 완전 관리형 RAG 솔루션을 제공합니다. 따라서 데이터 과학자나 기계 학습에 대한 전문 지식이 필요하지 않습니다.<br>• RAG 모델은 응답에서 정보 소스에 대한 참조를 제공합니다.<br>• RAG는 벡터 검색의 컨텍스트를 생성된 답변의 기반으로 사용하므로 할루시네이션 위험이 줄어듭니다. | • RAG는 전체 문서의 정보를 요약할 때 제대로 작동하지 않습니다. |

사용자 지정 문서를 참조하는 질문 응답 솔루션을 구축해야 하는 경우 RAG 기반 접근 방식부터 시작하는 것이 좋습니다. 요약과 같은 추가 작업을 수행하기 위해 모델이 필요한 경우 미세 조정을 사용합니다.

미세 조정 및 RAG 접근 방식을 단일 모델로 결합할 수 있습니다. 이 경우 RAG 아키텍처는 변경되지 않지만 답변을 생성하는 LLM도 사용자 지정 문서로 미세 조정됩니다. 이는 두 월드의 장점을 결합하며 사용 사례에 가장 적합한 솔루션일 수 있습니다. 지도 미세 조정을 RAG와 결합하는 방법에 대한 자세한 내용은의 RAFT: Adapting Language Model to Domain Specific RAG research를 참조하세요 University of California, Berkeley.

## 검색 증강 생성 사용 사례

다음은 RAG 접근 방식을 사용하는 일반적인 사용 사례입니다.

- 검색 엔진 - RAG 지원 검색 엔진은 검색 결과에 더 정확하고 up-to-date의 추천 코드 조각을 제공할 수 있습니다.
- 질문 응답 시스템 - RAG는 질문 응답 시스템의 응답 품질을 개선할 수 있습니다. 검색 기반 모델은 유사성 검색을 사용하여 답변이 포함된 관련 구절 또는 문서를 찾습니다. 그런 다음 해당 정보를 기반으로 간결하고 관련 있는 응답을 생성합니다.
- 소매 또는 전자 상거래 - RAG는 보다 관련성이 높고 개인화된 제품 추천을 제공하여 전자 상거래에서 사용자 경험을 개선할 수 있습니다. RAG는 사용자 기본 설정 및 제품 세부 정보에 대한 정보를 검색하고 통합하여 고객에게 더 정확하고 유용한 추천을 생성할 수 있습니다.
- 산업 또는 제조 - 제조에서 RAG를 사용하면 공장 운영과 같은 중요한 정보에 빠르게 액세스할 수 있습니다. 또한 의사 결정 프로세스, 문제 해결 및 조직 혁신에도 도움이 될 수 있습니다. 엄격한 규제 프레임워크 내에서 운영하는 제조업체의 경우 RAG는 업계 표준 또는 규제 기관과 같은 내부 및 외부 소스에서 업데이트된 규정 및 규정 준수 표준을 신속하게 검색할 수 있습니다.
- 의료 - RAG는 정확하고 시기 적절한 정보에 대한 액세스가 중요한 의료 산업에서 잠재력을 가지고 있습니다. RAG는 외부 소스에서 관련 의료 지식을 검색하고 통합하여 의료 애플리케이션에서 보다 정확하고 컨텍스트 인식 응답을 제공할 수 있습니다. 이러한 애플리케이션은 궁극적으로 모델이 아닌 호출을 수행하는 인간 임상의가 액세스할 수 있는 정보를 강화합니다.
- 법적 - 복잡한 법률 문서가 쿼리 컨텍스트를 제공하는 인수 합병과 같은 법적 시나리오에서 RAG를 강력하게 적용할 수 있습니다. 이를 통해 법률 전문가는 복잡한 규제 문제를 신속하게 해결할 수 있습니다.

# 의 완전 관리형 검색 증강 생성 옵션 AWS

에서 Retrieval Augmented Generation(RAG) 워크플로를 관리하려면 사용자 지정 RAG 파이프라인을 사용하거나에서 AWS 제공하는 일부 완전 관리형 서비스 기능을 사용할 AWS수 있습니다. 여기에는 RAG 기반 시스템의 많은 핵심 구성 요소가 포함되어 있기 때문에 완전 관리형 서비스는 차별화되지 않은 과중한 작업을 관리하는 데 도움이 될 수 있습니다. 그러나 이러한 서비스는 사용자 지정 기회를 덜 제공합니다.

완전 관리형는 커넥터를 AWS 서비스 사용하여 웹 사이트, Atlassian Confluence 또는 Microsoft SharePoint와 같은 외부 데이터 소스에서 데이터를 수집합니다. 지원되는 데이터 소스는 마다 다릅니다 AWS 서비스.

이 섹션에서는에서 RAG 워크플로를 빌드하기 위한 AWS다음과 같은 완전 관리형 옵션을 살펴봅니다.

- Knowledge Bases for Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI Canvas

이러한 옵션 중에서 선택하는 방법에 대한 자세한 내용은이 가이드에서 검색 증강 생성 옵션 선택 AWS의 섹션을 참조하세요.

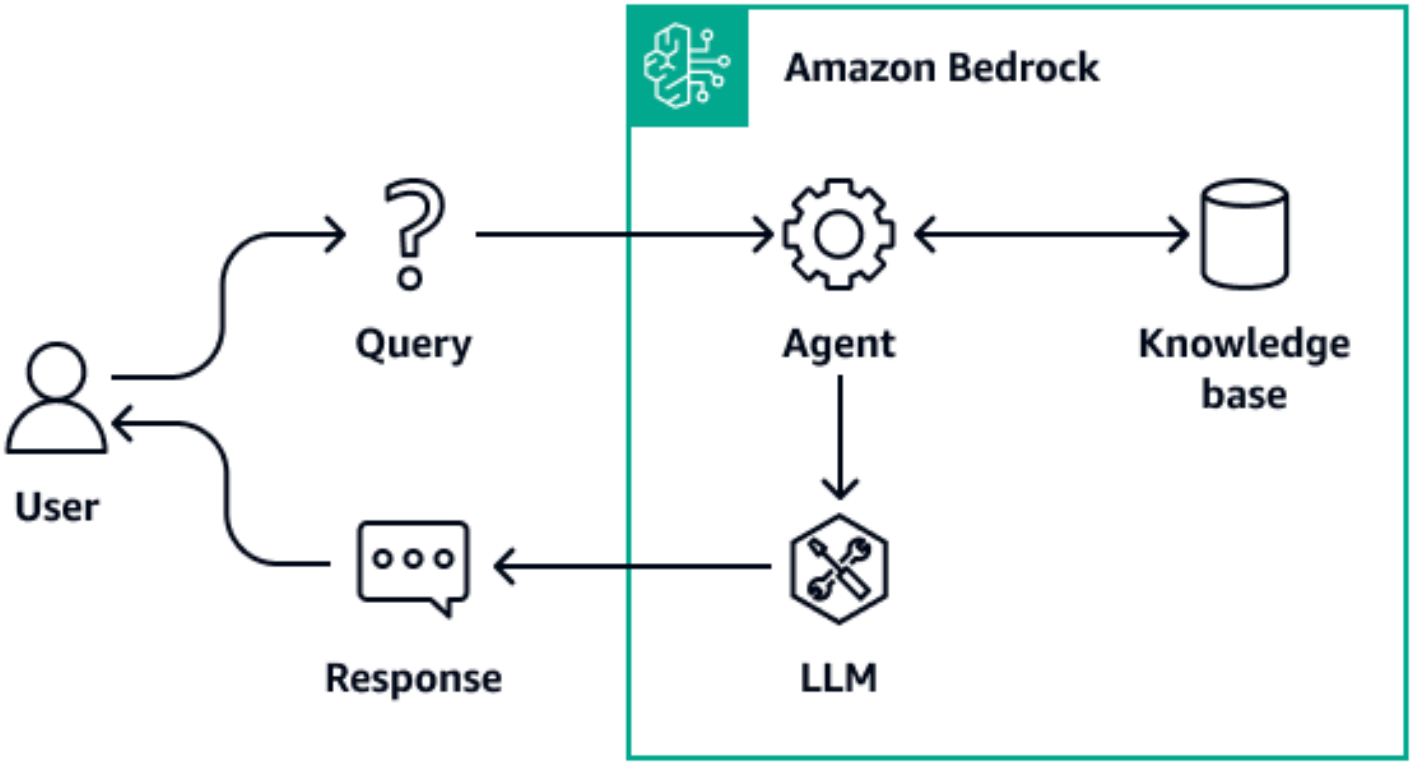
## Knowledge Bases for Amazon Bedrock

Amazon Bedrock은 선도적인 AI 스타트업과 Amazon의 고성능 파운데이션 모델(FM)을 통합 API를 통해 사용할 수 있게 해주는 완전 관리형 서비스입니다. 지식 기반은 수집부터 검색 및 프롬프트 보강에 이르기까지 전체 RAG 워크플로를 구현하는 데 도움이 되는 Amazon Bedrock 기능입니다. 데이터 소스에 대한 사용자 지정 통합을 구축하거나 데이터 흐름을 관리할 필요가 없습니다. 세션 컨텍스트 관리는 생성형 AI 애플리케이션이 멀티턴 대화를 쉽게 지원할 수 있도록 내장되어 있습니다.

데이터 위치를 지정하면 Amazon Bedrock의 지식 기반이 내부적으로 문서를 가져와 텍스트 블록으로 청크하고 텍스트를 임베딩으로 변환한 다음 선택한 벡터 데이터베이스에 임베딩을 저장합니다. Amazon Bedrock은 임베딩을 관리하고 업데이트하여 벡터 데이터베이스를 데이터와 동기화된 상태로 유지합니다. 지식 기반 작동 방식에 대한 자세한 내용은 Amazon Bedrock 지식 기반 작동 방식을 참조하세요.

Amazon Bedrock 에이전트에 지식 기반을 추가하면 에이전트는 사용자 입력을 기반으로 적절한 지식 기반을 식별합니다. 에이전트는 관련 정보를 검색하고 입력 프롬프트에 정보를 추가합니다. 업데이트

된 프롬프트는 모델에 응답을 생성하기 위한 추가 컨텍스트 정보를 제공합니다. 투명성을 개선하고 할루시네이션을 최소화하기 위해 지식 기반에서 검색된 정보는 소스로 추적할 수 있습니다.




Amazon Bedrock은 RAG에 대해 다음 두 가지 APIs를 지원합니다.

- RetrieveAndGenerate -이 API를 사용하여 지식 기반을 쿼리하고 검색하는 정보에서 응답을 생성할 수 있습니다. 내부적으로 Amazon Bedrock은 쿼리를 임베딩으로 변환하고, 지식 기반을 쿼리하고, 검색 결과를 컨텍스트 정보로 사용하여 프롬프트를 보강하고, LLM 생성 응답을 반환합니다. 또한 Amazon Bedrock은 대화의 단기 메모리를 관리하여 보다 상황별 결과를 제공합니다.
- 검색 -이 API를 사용하여 지식 기반에서 직접 검색된 정보로 지식 기반을 쿼리할 수 있습니다. 이 API에서 반환된 정보를 사용하여 검색된 텍스트를 처리하거나, 관련성을 평가하거나, 응답 생성을 위한 별도의 워크플로를 개발할 수 있습니다. 내부적으로 Amazon Bedrock은 쿼리를 임베딩으로 변환하고, 지식 기반을 검색하고, 관련 결과를 반환합니다. 검색 결과 위에 추가 워크플로를 빌드할 수 있습니다. 예를 들어 LangChain `AmazonKnowledgeBasesRetriever` 플러그인을 사용하여 RAG 워크플로를 생성형 AI 애플리케이션에 통합할 수 있습니다.

APIs 사용에 대한 샘플 아키텍처 패턴 및 step-by-step 지침은 이제 지식 기반에서 Amazon Bedrock에서 완전 관리형 RAG 환경 제공(AWS 블로그 게시물)을 참조하세요. `RetrieveAndGenerate` API를 사용하여 지능형 채팅 기반 애플리케이션을 위한 RAG 워크플로를 빌드하는 방법에 대한 자세한 내용

은 Amazon Bedrock 지식 기반을 사용하여 컨텍스트 챗봇 애플리케이션 빌드( 블로그 게시물)를 참조하세요.AWS

## 지식 기반용 데이터 소스

독점 데이터를 지식 기반에 연결할 수 있습니다. 데이터 소스 커넥터를 구성한 후 데이터를 지식 기반과 동기화하거나 최신 상태로 유지하고 데이터를 쿼리에 사용할 수 있도록 할 수 있습니다. Amazon Bedrock 지식 기반은 다음 데이터 소스에 대한 연결을 지원합니다.

- Amazon Simple Storage Service(Amazon S3) - 콘솔 또는 API를 사용하여 Amazon S3 버킷을 Amazon Bedrock 지식 기반에 연결할 수 있습니다. 지식 기반은 버킷의 파일을 수집하고 인덱싱합니다. 이러한 유형의 데이터 소스는 다음 기능을 지원합니다.
  - 문서 메타데이터 필드 - 별도의 파일을 포함하여 Amazon S3 버킷의 파일에 대한 메타데이터를 지정할 수 있습니다. 그런 다음 이러한 메타데이터 필드를 사용하여 응답의 관련성을 필터링하고 개선할 수 있습니다.
  - 포함 또는 제외 필터 - 크롤링할 때 특정 콘텐츠를 포함하거나 제외할 수 있습니다.
  - 증분 동기화 - 콘텐츠 변경 사항이 추적되고 마지막 동기화 이후 변경된 콘텐츠만 크롤링됩니다.
- Confluence - 콘솔 또는 API를 사용하여 Amazon Bedrock 지식 기반에 Atlassian Confluence 인스턴스를 연결할 수 있습니다. 이러한 유형의 데이터 소스는 다음 기능을 지원합니다.
  - 기본 문서 필드 자동 감지 - 메타데이터 필드가 자동으로 감지되고 크롤링됩니다. 이러한 필드를 필터링에 사용할 수 있습니다.
  - 포함 또는 제외 콘텐츠 필터 - 스페이스, 페이지 제목, 블로그 제목, 설명, 첨부 파일 이름 또는 확장에 접두사 또는 정규식 패턴을 사용하여 특정 콘텐츠를 포함하거나 제외할 수 있습니다.
  - 증분 동기화 - 콘텐츠 변경 사항이 추적되고 마지막 동기화 이후 변경된 콘텐츠만 크롤링됩니다.
  - OAuth 2.0 인증, Confluence API 토큰을 사용한 인증 - 인증 자격 증명이에 저장됩니다 AWS Secrets Manager.
- Microsoft SharePoint - 콘솔 또는 API를 사용하여 SharePoint 인스턴스를 지식 기반에 연결할 수 있습니다. 이러한 유형의 데이터 소스는 다음 기능을 지원합니다.
  - 기본 문서 필드 자동 감지 - 메타데이터 필드가 자동으로 감지되고 크롤링됩니다. 이러한 필드를 필터링에 사용할 수 있습니다.
  - 포함 또는 제외 콘텐츠 필터 - 기본 페이지 제목, 이벤트 이름 및 파일 이름( 확장명 포함)에 접두사 또는 정규식 패턴을 사용하여 특정 콘텐츠를 포함하거나 제외할 수 있습니다.
  - 증분 동기화 - 콘텐츠 변경 사항이 추적되고 마지막 동기화 이후 변경된 콘텐츠만 크롤링됩니다.
  - OAuth 2.0 인증 - 인증 자격 증명이에 저장됩니다 AWS Secrets Manager.

- Salesforce - 콘솔 또는 API를 사용하여 Salesforce 인스턴스를 지식 기반에 연결할 수 있습니다. 이러한 유형의 데이터 소스는 다음 기능을 지원합니다.
  - 기본 문서 필드 자동 감지 - 메타데이터 필드가 자동으로 감지되고 크롤링됩니다. 이러한 필드를 필터링에 사용할 수 있습니다.
  - 포함 또는 제외 콘텐츠 필터 - 접두사 또는 정규식 패턴을 사용하여 특정 콘텐츠를 포함하거나 제외할 수 있습니다. 필터를 적용할 수 있는 콘텐츠 유형 목록은 Amazon Bedrock 설명서의 포함/제외 필터를 참조하세요.
  - 증분 동기화 - 콘텐츠 변경 사항이 추적되고 마지막 동기화 이후 변경된 콘텐츠만 크롤링됩니다.
  - OAuth 2.0 인증 - 인증 자격 증명이에 저장됩니다 AWS Secrets Manager.
- 웹 크롤러 - Amazon Bedrock 웹 크롤러는 사용자가 제공하는 URLs. 다음 기능이 지원됩니다.
  - 크롤링할 여러 URL을 선택합니다.
  - `Allow` 및와 같은 표준 robots.txt 지시문을 준수합니다. `Disallow`
  - 패턴과 일치하는 URLs 제외
  - 크롤링 속도 제한
  - Amazon CloudWatch에서 크롤링된 각 URL의 상태를 확인합니다.

Amazon Bedrock 지식 기반에 연결할 수 있는 데이터 소스에 대한 자세한 내용은 지식 기반용 데이터 소스 커넥터 생성을 참조하세요.

## 지식 기반을 위한 벡터 데이터베이스

지식 기반과 데이터 소스 간의 연결을 설정할 때 벡터 스토어라고도 하는 벡터 데이터베이스를 구성해야 합니다. 벡터 데이터베이스는 Amazon Bedrock이 데이터를 나타내는 임베딩을 저장, 업데이트 및 관리하는 곳입니다. 각 데이터 소스는 다양한 유형의 벡터 데이터베이스를 지원합니다. 데이터 소스에 사용할 수 있는 벡터 데이터베이스를 확인하려면 데이터 소스 유형을 참조하세요.

Amazon Bedrock이 Amazon OpenSearch Serverless에서 벡터 데이터베이스를 자동으로 생성하도록 하려면 지식 기반을 생성할 때이 옵션을 선택할 수 있습니다. 그러나 자체 벡터 데이터베이스를 설정하도록 선택할 수도 있습니다. 자체 벡터 데이터베이스를 설정하는 경우 지식 기반에 대한 자체 벡터 저장소의 사전 조건을 참조하세요. 각 벡터 데이터베이스 유형에는 자체 사전 조건이 있습니다.

데이터 소스 유형에 따라 Amazon Bedrock 지식 기반은 다음 벡터 데이터베이스를 지원합니다.

- Amazon OpenSearch Serverless
- Amazon Aurora PostgreSQL-Compatible Edition

- Pinecone(Pinecone 설명서)
- Redis Enterprise Cloud(Redis 설명서)
- MongoDB Atlas(MongoDB 설명서)

# Amazon Q Business

Amazon Q Business는 질문에 답변하고, 요약을 제공하고, 콘텐츠를 생성하고, 엔터프라이즈 데이터를 기반으로 작업을 완료하도록 구성할 수 있는 완전 관리형 생성형 AI 기반 어시스턴트입니다. 이를 통해 최종 사용자는 인용을 사용하여 엔터프라이즈 데이터 소스로부터 권한 인식 응답을 즉시 받을 수 있습니다.

## 주요 기능

Amazon Q Business의 다음 기능은 프로덕션급 RAG 기반 생성형 AI 애플리케이션을 구축하는 데 도움이 될 수 있습니다.

- 내장 커넥터 - Amazon Q Business는 , Adobe Experience Manager (AEM), 및 용 커넥터와 같은 40개 이상의 커넥터 유형을 지원합니다SalesforceJiraMicrosoft SharePoint. 전체 목록은 지원되는 커넥터를 참조하세요. 지원되지 않는 커넥터가 필요한 경우 Amazon AppFlow를 사용하여 데이터 소스에서 Amazon Simple Storage Service(Amazon S3)로 데이터를 가져온 다음 Amazon Q Business를 Amazon S3 버킷에 연결할 수 있습니다. Amazon AppFlow에서 지원하는 데이터 소스의 전체 목록은 지원되는 애플리케이션을 참조하세요.
- 기본 제공 인덱싱 파이프라인 - Amazon Q Business는 벡터 데이터베이스의 데이터를 인덱싱하기 위한 기본 제공 파이프라인을 제공합니다. AWS Lambda 함수를 사용하여 인덱싱 파이프라인에 대한 사전 처리 로직을 추가할 수 있습니다.
- 인덱스 옵션 - Amazon Q Business에서 네이티브 인덱스를 생성하고 프로비저닝할 수 있으며 Amazon Q Business 리트리버를 사용하여 해당 인덱스에서 데이터를 가져올 수 있습니다. 또는 미리 구성된 Amazon Kendra 인덱스를 리트리버로 사용할 수 있습니다. 자세한 내용은 Amazon Q Business 애플리케이션을 위한 리트리버 생성을 참조하세요.
- 파운데이션 모델 - Amazon Q Business는 Amazon Bedrock에서 지원되는 파운데이션 모델을 사용합니다. 전체 목록은 Amazon Bedrock에서 지원되는 파운데이션 모델을 참조하세요.
- 플러그인 - Amazon Q Business는에서 티켓 정보 및 티켓 생성을 요약하는 자동화된 방법과 같이 플러그인을 사용하여 대상 시스템과 통합할 수 있는 기능을 제공합니다Jira. 플러그인이 구성되면 플러그인은 최종 사용자 생산성을 높이는 데 도움이 되는 읽기 및 쓰기 작업을 지원할 수 있습니다. Amazon Q Business는 내장 플러그인과 사용자 지정 플러그인이라는 두 가지 유형의 플러그인을 지원합니다.

- 가드레일 - Amazon Q Business는 글로벌 제어 및 주제 수준 제어를 지원합니다. 예를 들어 이러한 제어는 프롬프트에서 개인 식별 정보(PII), 침해 또는 민감한 정보를 탐지할 수 있습니다. 자세한 내용은 Amazon Q Business의 관리자 제어 및 가드레일을 참조하세요.
- 자격 증명 관리 - Amazon Q Business를 사용하면 RAG 기반 생성형 AI 애플리케이션에 대한 사용자 및 사용자의 액세스를 관리할 수 있습니다. 자세한 내용은 Amazon Q Business의 자격 증명 및 액세스 관리를 참조하세요. 또한 Amazon Q Business 커넥터는 문서 자체와 함께 문서에 연결된 액세스 제어 목록(ACL) 정보를 인덱싱합니다. 그런 다음 Amazon Q Business는 Amazon Q Business 사용자 스토어에 인덱싱한 ACL 정보를 저장하여 사용자 및 그룹 매핑을 생성하고 최종 사용자의 문서 액세스를 기반으로 채팅 응답을 필터링합니다. 자세한 내용은 데이터 소스 커넥터 개념을 참조하세요.
- 문서 보강 - 문서 보강 기능을 사용하면 인덱스에 수집되는  문서 및 문서 속성과 수집 방식을 모두 제어할 수 있습니다. 이는 두 가지 접근 방식을 통해 수행할 수 있습니다.
  - 기본 작업 구성 - 기본 작업을 사용하여 데이터에서 문서 속성을 추가, 업데이트 또는 삭제합니다. 예를 들어 PII와 관련된 문서 속성을 삭제하도록 선택하여 PII 데이터를 스크러빙할 수 있습니다.
  - Lambda 함수 구성 - 사전 구성된 Lambda 함수를 사용하여 데이터에 대해 보다 사용자 지정된 고급 문서 속성 조작 로직을 수행합니다. 예를 들어, 엔터프라이즈 데이터는 스캔된 이미지로 저장될 수 있습니다. 이 경우 Lambda 함수를 사용하여 스캔한 문서에서 광학 문자 인식(OCR)을 실행하여 해당 문서에서 텍스트를 추출할 수 있습니다. 그 후 스캔한 각 문서는 수집 중에 텍스트 문서로 처리됩니다. 마지막으로 채팅 중에 Amazon Q는 응답을 생성할 때 스캔한 문서에서 추출한 텍스트 데이터를 고려합니다.

  솔루션을 구현할 때 두 문서 보강 접근 방식을 모두 결합하도록 선택할 수 있습니다. 기본 작업을 사용하여 데이터의 첫 번째 구문 분석을 수행한 다음 Lambda 함수를 사용하여 더 복잡한 작업을 수행할 수 있습니다. 자세한 내용은 Amazon Q Business의 문서 보강을 참조하세요.
- 통합 - Amazon Q Business 애플리케이션을 생성한 후 Slack 또는와 같은 다른 애플리케이션에 통합할 수 있습니다Microsoft Teams. 예를 들어 Amazon Q BusinessforAmazon Slack 게이트웨이 배포 및 Amazon Q Business용 Microsoft Teams 게이트웨이 배포(AWS 블로그 게시물)를 참조하세요.

## 최종 사용자 사용자 지정

Amazon Q Business는 조직의 데이터 소스 및 인덱스에 저장되지 않을 수 있는 문서 업로드를 지원합니다. 업로드된 문서는 저장되지 않습니다. 문서가 업로드되는 대화에만 사용할 수 있습니다. Amazon Q Business는 업로드를 위한 특정 문서 유형을 지원합니다. 자세한 내용은 Amazon Q Business에서 파일 및 채팅 업로드를 참조하세요.

Amazon Q Business에는 문서 속성별 필터링 기능이 포함되어 있습니다. 관리자와 최종 사용자 모두 이 기능을 사용할 수 있습니다. 관리자는 속성을 사용하여 최종 사용자의 채팅 응답을 사용자 지정하고

제어할 수 있습니다. 예를 들어, 데이터 소스 유형이 문서에 연결된 속성인 경우 채팅 응답을 특정 데이터 소스에서만 생성하도록 지정할 수 있습니다. 또는 최종 사용자가 선택한 속성 필터를 사용하여 채팅 응답 범위를 제한하도록 허용할 수 있습니다.

최종 사용자는 광범위한 Amazon Q Business 애플리케이션 환경 내에서 특별히 구축된 경량 Amazon Q Apps를 생성할 수 있습니다. Amazon Q 앱을 사용하면 마케팅 팀을 위해 특별히 구축된 앱과 같은 특정 도메인에 대한 태스크 자동화를 수행할 수 있습니다.

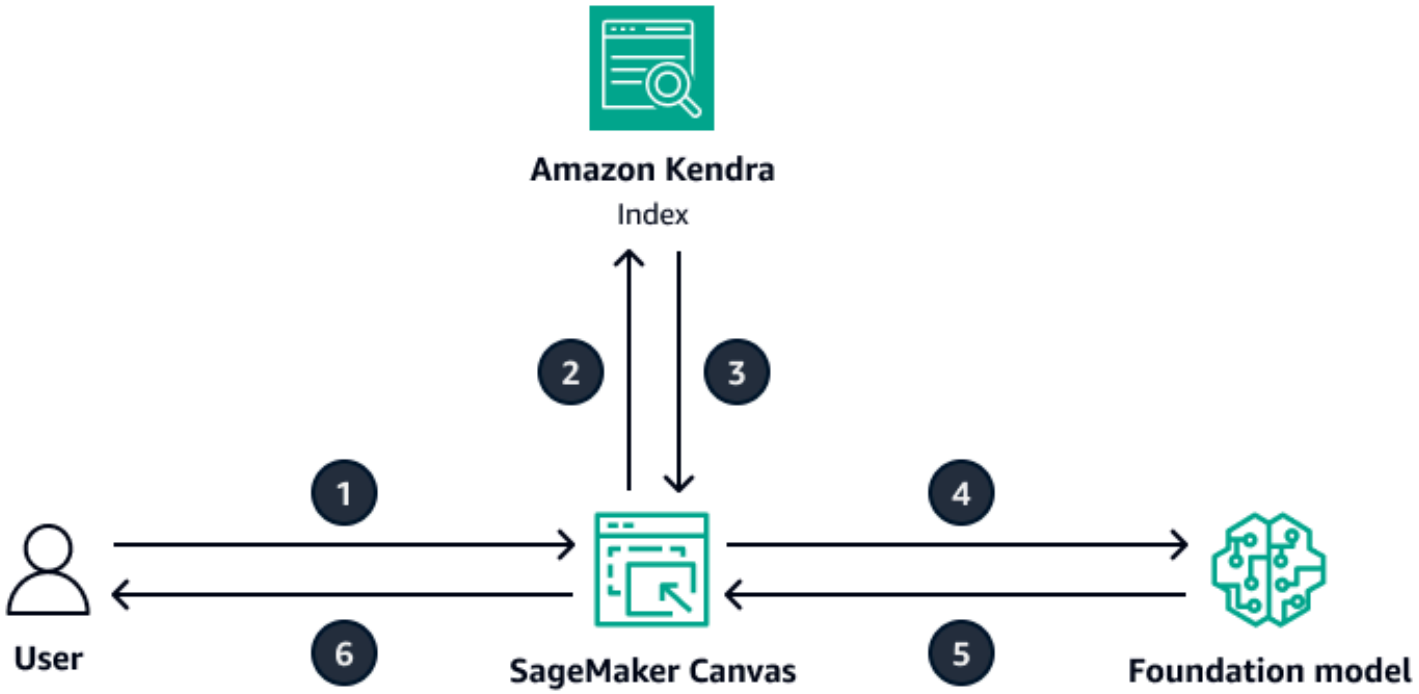
## Amazon SageMaker AI Canvas

Amazon SageMaker AI Canvas를 사용하면 코드를 작성할 필요 없이 기계 학습을 사용하여 예측을 생성할 수 있습니다. 통합 환경에서 end-to-end ML 수명 주기를 간소화하여 데이터를 준비하고 ML 모델을 빌드 및 배포할 수 있는 코드 없는 시각적 인터페이스를 제공합니다. 데이터 준비, 모델 개발, 편향 감지, 설명 가능성 및 모니터링의 복잡성은 직관적인 인터페이스 뒤에서 추상화됩니다. 사용자는 SageMaker AI Canvas를 사용하여 모델을 개발, 운영 및 모니터링하기 위해 SageMaker AI 또는 기계 학습 운영(MLOps) 전문가가 될 필요가 없습니다.

SageMaker AI Canvas를 사용하면 코드 없는 문서 쿼리 기능을 통해 RAG 기능이 제공됩니다. Amazon Kendra 인덱스를 기본 엔터프라이즈 검색으로 사용하여 SageMaker AI Canvas에서 채팅 환경을 강화할 수 있습니다. 자세한 내용은 문서 쿼리를 사용하여 문서에서 정보 추출을 참조하세요.

SageMaker AI Canvas를 Amazon Kendra 인덱스에 연결하려면 일회성 설정이 필요합니다. 도메인 구성의 일부로 클라우드 관리자는 사용자가 SageMaker Canvas와 상호 작용할 때 쿼리할 수 있는 Kendra 인덱스를 하나 이상 선택할 수 있습니다. 문서 쿼리 기능을 활성화하는 방법에 대한 지침은 Amazon SageMaker AI Canvas 사용 시작하기를 참조하세요.

SageMaker AI Canvas는 Amazon Kendra와 선택한 파운데이션 모델 간의 기본 통신을 관리합니다. SageMaker AI Canvas가 지원하는 파운데이션 모델에 대한 자세한 내용은 SageMaker AI Canvas의 생성형 AI 파운데이션 모델을 참조하세요. 다음 다이어그램은 클라우드 관리자가 SageMaker AI Canvas를 Amazon Kendra 인덱스에 연결한 후 문서 쿼리 기능이 작동하는 방식을 보여줍니다.




이 다이어그램은 다음 워크플로를 보여줍니다.

1. 사용자가 SageMaker AI Canvas에서 새 채팅을 시작하고, 문서 쿼리를 켜고, 대상 인덱스를 선택한 다음 질문을 제출합니다.
2. SageMaker AI Canvas는 쿼리를 사용하여 Amazon Kendra 인덱스에서 관련 데이터를 검색합니다.
3. SageMaker AI Canvas는 Amazon Kendra 인덱스에서 데이터와 해당 소스를 검색합니다.
4. SageMaker AI Canvas는 Amazon Kendra 인덱스에서 검색된 컨텍스트를 포함하도록 프롬프트를 업데이트하고 파운데이션 모델에 프롬프트를 제출합니다.
5. 파운데이션 모델은 원래 질문과 검색된 컨텍스트를 사용하여 답변을 생성합니다.
6. SageMaker AI Canvas는 사용자에게 생성된 답변을 제공합니다. 여기에는 응답을 생성하는 데 사용된 문서와 같은 데이터 소스에 대한 참조가 포함됩니다.

# 의 사용자 지정 검색 증강 생성 아키텍처 AWS

이전 섹션에서는 검색 증강 생성(RAG) AWS 서비스 에 완전 관리형를 사용하는 방법을 설명합니다. 그러나 일부 사용 사례에서는 리트리버 또는 LLM(생성기라고도 함)과 같은 시스템 구성 요소를 더 잘 제어해야 합니다. 예를 들어 자체 벡터 데이터베이스를 선택하거나 지원되지 않는 데이터 소스에 액세스할 수 있는 유연성이 필요할 수 있습니다. 이러한 사용 사례의 경우 사용자 지정 RAG 아키텍처를 구축할 수 있습니다.

이 섹션은 다음 주제를 포함합니다:

- RAG 워크플로에 대한 검색
- RAG 워크플로용 생성기

이 섹션의 리트리버 및 생성기 옵션 중에서 선택하는 방법에 대한 자세한 내용은이 가이드에서 검색 증강 생성 옵션 선택 AWS의 섹션을 참조하세요.

## RAG 워크플로에 대한 검색

이 섹션에서는 리트리버를 빌드하는 방법을 설명합니다. Amazon Kendra와 같은 완전관리형 의미 체계 검색 솔루션을 사용하거나 AWS 벡터 데이터베이스를 사용하여 사용자 지정 의미 체계 검색을 구축할 수 있습니다.

리트리버 옵션을 검토하기 전에 벡터 검색 프로세스의 세 단계를 이해해야 합니다.

1. 인덱싱해야 하는 문서를 더 작은 부분으로 구분합니다. 이를 청킹이라고 합니다.
2. 임베딩이라는 프로세스를 사용하여 각 청크를 수학 벡터로 변환합니다. 그런 다음 벡터 데이터베이스의 각 벡터를 인덱싱합니다. 문서를 인덱싱하는 데 사용하는 접근 방식은 검색 속도와 정확도에 영향을 미칩니다. 인덱싱 접근 방식은 벡터 데이터베이스와 벡터 데이터베이스가 제공하는 구성 옵션에 따라 달라집니다.
3. 동일한 프로세스를 사용하여 사용자 쿼리를 벡터로 변환합니다. 리트리버는 벡터 데이터베이스에서 사용자의 쿼리 벡터와 유사한 벡터를 검색합니다. 유사성은 유클리드 거리, 코사인 거리 또는 점 제품과 같은 지표를 사용하여 계산됩니다.

이 가이드에서는 다음 AWS 서비스 또는 타사 서비스를 사용하여 사용자 지정 검색 계층을 구축하는 방법을 설명합니다. AWS

- Amazon Kendra

- Amazon OpenSearch Service
- Amazon Aurora PostgreSQL 및 pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

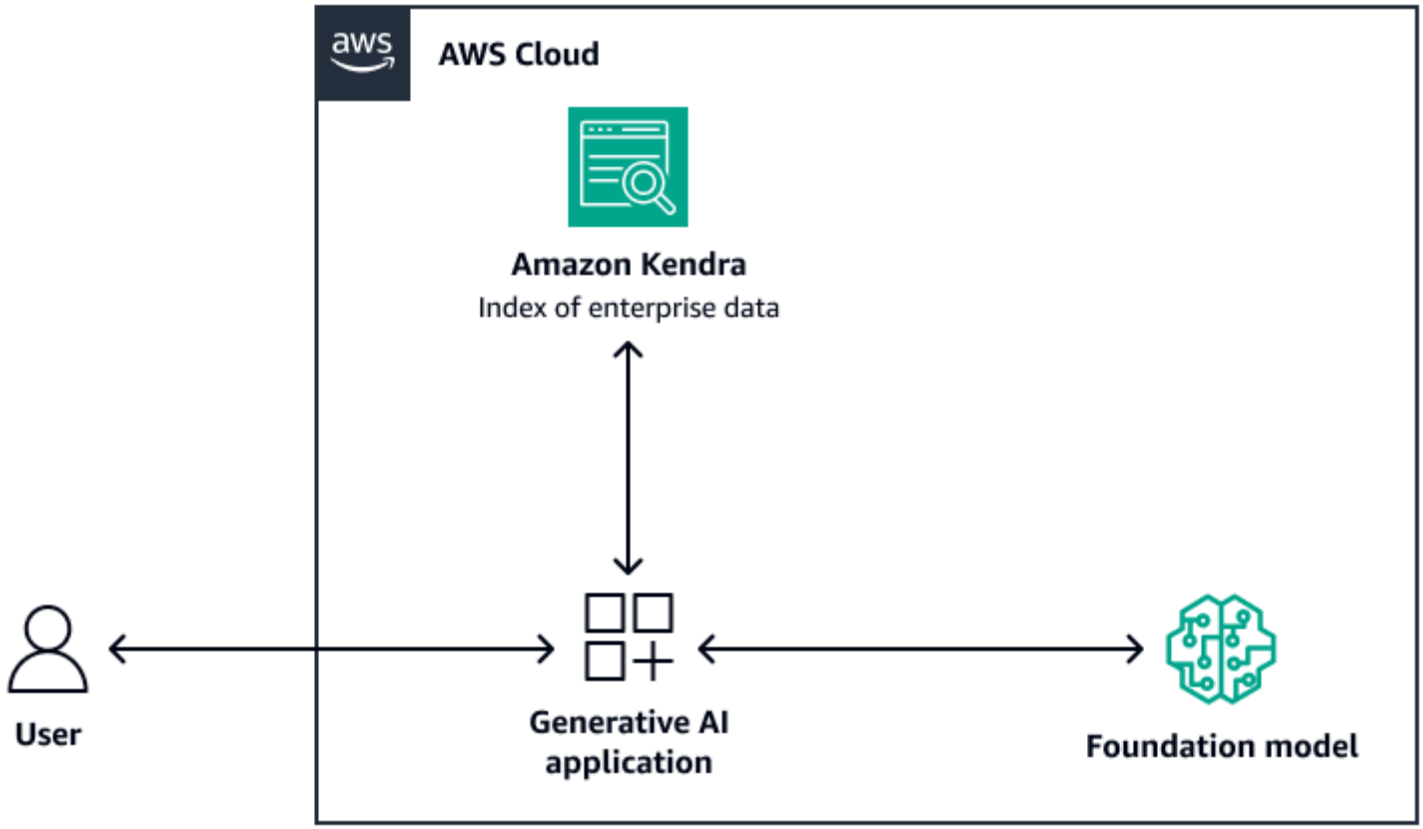
## Amazon Kendra

Amazon Kendra는 자연어 처리 및 고급 기계 학습 알고리즘을 사용하여 데이터의 검색 질문에 대한 특정 답변을 반환하는 완전관리형 지능형 검색 서비스입니다. Amazon Kendra를 사용하면 여러 소스에서 문서를 직접 수집하고 성공적으로 동기화된 후 문서를 쿼리할 수 있습니다. 동기화 프로세스는 수집된 문서에서 벡터 검색을 생성하는 데 필요한 인프라를 생성합니다. 따라서 Amazon Kendra에는 벡터 검색 프로세스의 기존 세 단계가 필요하지 않습니다. 초기 동기화 후 정의된 일정을 사용하여 지속적인 수집을 처리할 수 있습니다.

다음은 RAG용 Amazon Kendra 사용의 장점입니다.

- Amazon Kendra가 전체 벡터 검색 프로세스를 처리하기 때문에 벡터 데이터베이스를 유지 관리할 필요가 없습니다.
- Amazon Kendra에는 데이터베이스, 웹 사이트 크롤러, Amazon S3 버킷, Microsoft SharePoint 인스턴스 및 Atlassian Confluence 인스턴스와 같은 널리 사용되는 데이터 소스에 대한 사전 구축된 커넥터가 포함되어 있습니다. Box 및 용 커넥터와 같이 AWS 파트너가 개발한 커넥터를 사용할 수 있습니다GitLab.
- Amazon Kendra는 최종 사용자가 액세스할 수 있는 문서만 반환하는 액세스 제어 목록(ACL) 필터링을 제공합니다.
- Amazon Kendra는 날짜 또는 소스 리포지토리와 같은 메타데이터를 기반으로 응답을 부스팅할 수 있습니다.

다음 이미지는 Amazon Kendra를 RAG 시스템의 검색 계층으로 사용하는 샘플 아키텍처를 보여줍니다. 자세한 내용은 Amazon Kendra, LangChain및 대규모 언어 모델을 사용하여 엔터프라이즈 데이터에 고정밀 생성형 AI 애플리케이션을 빠르게 구축(AWS 블로그 게시물)을 참조하세요.




파운데이션 모델의 경우 Amazon Bedrock 또는 Amazon SageMaker AI JumpStart를 통해 배포된 LLM을 사용할 수 있습니다. 와 AWS Lambda 함께 LangChain를 사용하여 사용자, Amazon Kendra 및 LLM 간의 흐름을 오케스트레이션할 수 있습니다. Amazon Kendra, LangChain및 다양한 LLMs 사용하는 RAG 시스템을 빌드하려면 Amazon Kendra LangChain Extensions GitHub 리포지토리를 참조하세요.

## Amazon OpenSearch Service

Amazon OpenSearch Service는 벡터 검색을 수행하기 위해 k-Nearest Neighbors(k-NN) 검색을 위한 기본 제공 ML 알고리즘을 제공합니다. OpenSearch Service는 Amazon EMR Serverless용 벡터 엔진도 제공합니다. 이 벡터 엔진을 사용하여 확장 가능하고 성능이 뛰어난 벡터 스토리지 및 검색 기능을 갖춘 RAG 시스템을 구축할 수 있습니다. OpenSearch Serverless를 사용하여 RAG 시스템을 빌드하는 방법에 대한 자세한 내용은 Amazon OpenSearch Serverless 및 Amazon Bedrock Claude 모델용 벡터 엔진을 사용하여 확장 가능한 서버리스 RAG 워크플로 빌드(AWS 블로그 게시물)를 참조하세요.

다음은 벡터 검색에 OpenSearch Service를 사용할 때의 이점입니다.

- OpenSearch Serverless를 사용하여 확장 가능한 벡터 검색을 구축하는 등 벡터 데이터베이스를 완벽하게 제어할 수 있습니다.
- 청킹 전략을 제어할 수 있습니다.

- Non-Metric Space Library(NMSLIB), Faiss 및 Apache Lucene 라이브러리의 근사치 ANN(근사치) 알고리즘을 사용하여 k-NN 검색을 지원합니다. https://github.com/facebookresearch/faiss https://lucene.apache.org/ 사용 사례에 따라 알고리즘을 변경할 수 있습니다. OpenSearch Service를 통해 벡터 검색을 사용자 지정하는 옵션에 대한 자세한 내용은 Amazon OpenSearch Service 벡터 데이터베이스 기능 설명(AWS 블로그 게시물)을 참조하세요.
- OpenSearch Serverless는 Amazon Bedrock 지식 기반과 벡터 인덱스로 통합됩니다.

## Amazon Aurora PostgreSQL 및 pgvector

Amazon Aurora PostgreSQL 호환 버전은 PostgreSQL 배포를 설정, 운영 및 확장하는 데 도움이 되는 완전 관리형 관계형 데이터베이스 엔진입니다. pgvector는 벡터 유사성 검색 기능을 제공하는 오픈 소스 PostgreSQL 확장입니다. 이 확장은 Aurora PostgreSQL 호환 및 PostgreSQL용 Amazon Relational Database Service(RDS) 모두에 사용할 수 있습니다. Aurora PostgreSQL 호환 및 pgvector를 사용하는 RAG 기반 시스템을 구축하는 방법에 대한 자세한 내용은 다음 AWS 블로그 게시물을 참조하세요.

- Amazon SageMaker AI 및 pgvector를 사용하여 PostgreSQL에서 AI 기반 검색 구축
- 자연어 처리, 챗봇 및 감정 분석을 위해 pgvector 및 Amazon Aurora PostgreSQL 활용

다음은 pgvector 및 Aurora PostgreSQL 호환을 사용할 때의 이점입니다.

- 가장 가까운 정확한 이웃 검색을 지원합니다. 또한 L2 거리, 내부 제품 및 코사인 거리와 같은 유사성 지표를 지원합니다.
- 플랫 압축을 사용하는 반전 파일(IVFFlat) 및 계층적 탐색 가능 스몰 월드(HNSW) 인덱싱을 지원합니다.
- 벡터 검색을 동일한 PostgreSQL 인스턴스에서 사용할 수 있는 도메인별 데이터에 대한 쿼리와 결합할 수 있습니다.
- Aurora PostgreSQL 호환은 I/O에 최적화되어 있으며 계층형 캐싱을 제공합니다. 사용 가능한 인스턴스 메모리를 초과하는 워크로드의 경우 pgvector는 벡터 검색에 대한 초당 쿼리를 최대 8 회 늘릴 수 있습니다.

## Amazon Neptune Analytics

Amazon Neptune Analytics는 분석을 위한 메모리 최적화 그래프 데이터베이스 엔진입니다. 그래프 순회 내에서 최적화된 그래프 분석 알고리즘, 지연 시간이 짧은 그래프 쿼리 및 벡터 검색 기능의 라이브러리를 지원합니다. 또한 벡터 유사성 검색이 내장되어 있습니다. 그래프를 생성하고, 데이터를 로드하

고, 쿼리를 호출하고, 벡터 유사성 검색을 수행할 수 있는 하나의 엔드포인트를 제공합니다. Neptune Analytics를 사용하는 RAG 기반 시스템을 빌드하는 방법에 대한 자세한 내용은 지식 그래프를 사용하여 Amazon Bedrock 및 Amazon Neptune으로 GraphRAG 애플리케이션 빌드(AWS 블로그 게시물)를 참조하세요.

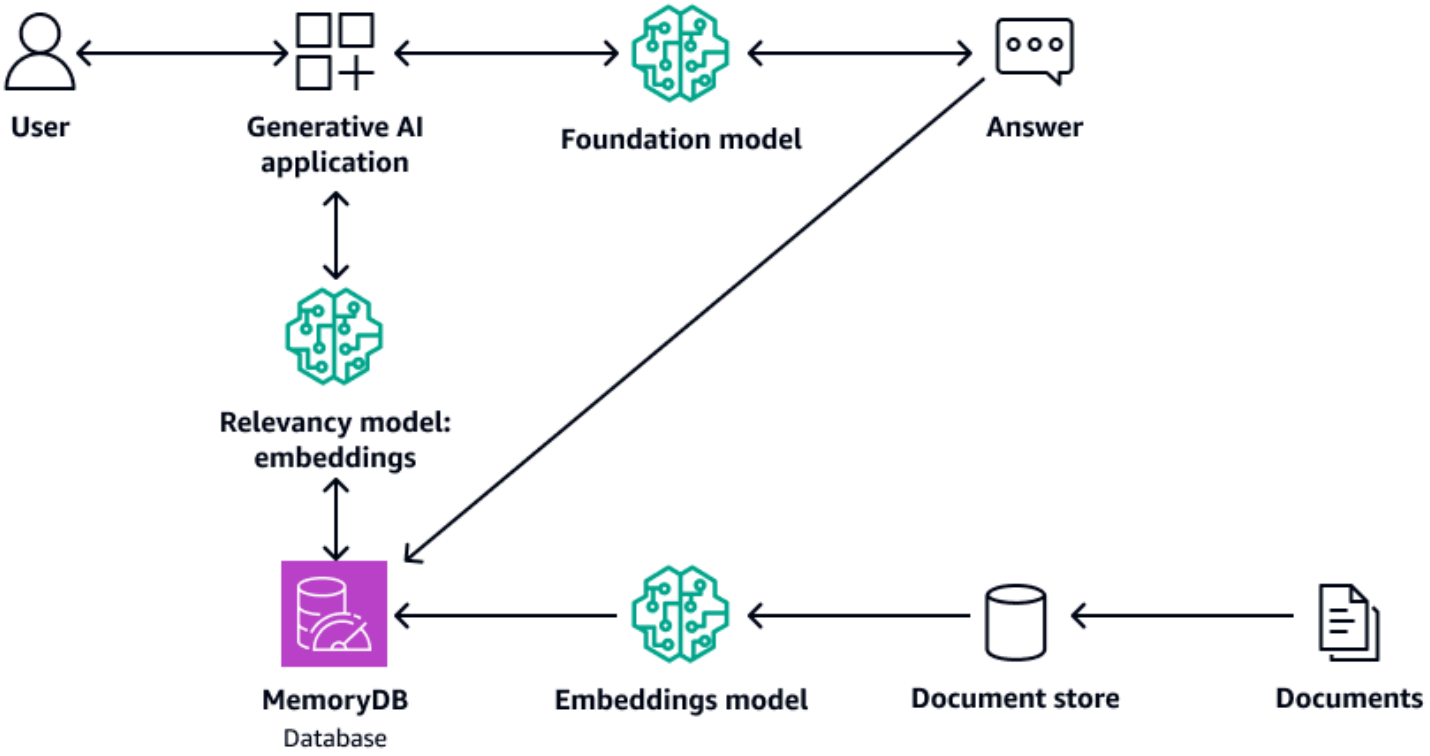
Neptune Analytics 사용의 장점은 다음과 같습니다.

- 그래프 쿼리에 임베딩을 저장하고 검색할 수 있습니다.
- Neptune Analytics를와 통합하면 LangChain이 아키텍처는 자연어 그래프 쿼리를 지원합니다.
- 이 아키텍처는 대용량 그래프 데이터 세트를 메모리에 저장합니다.

## Amazon MemoryDB

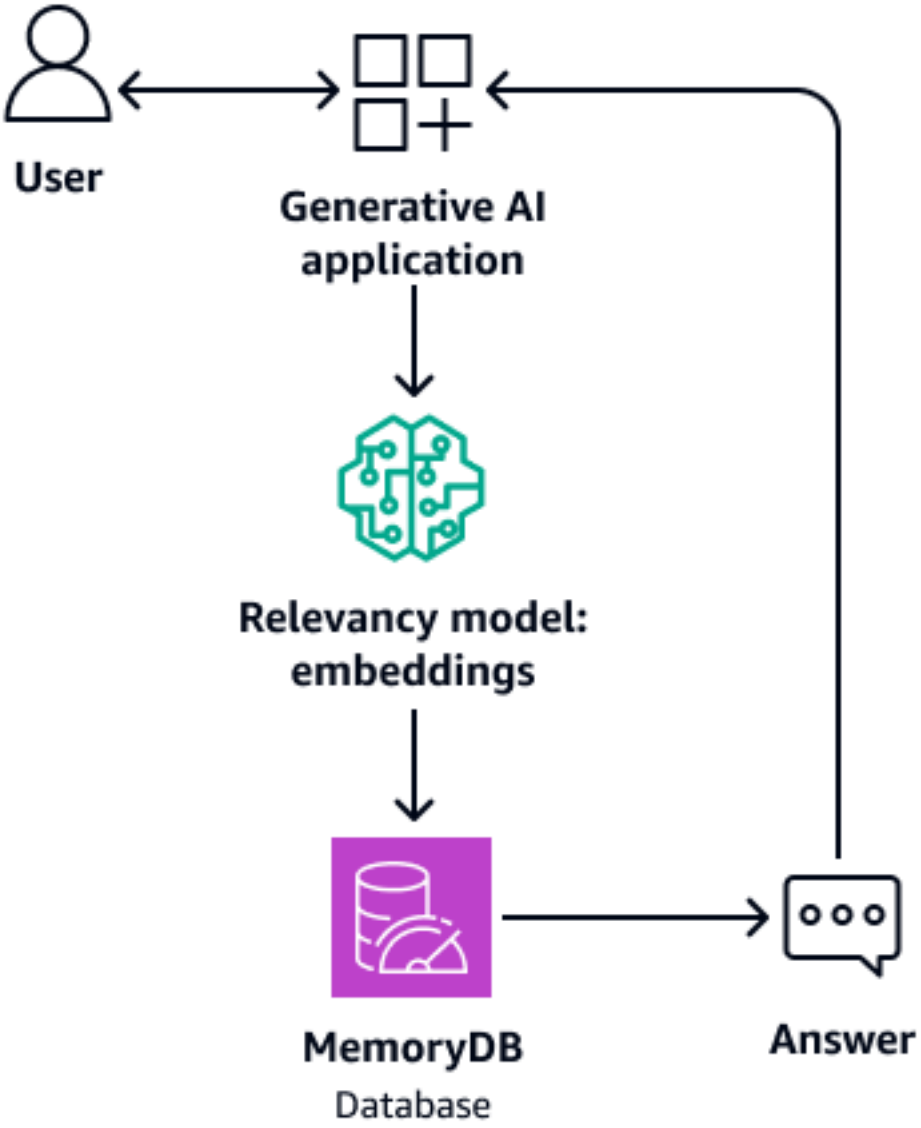
Amazon MemoryDB는 내구성이 뛰어난 인 메모리 데이터베이스 서비스로, 매우 빠른 성능을 제공합니다. 모든 데이터는 마이크로초 읽기, 한 자릿수 밀리초 쓰기 지연 시간 및 높은 처리량을 지원하는 메모리에 저장됩니다. MemoryDB에 대한 벡터 검색은 MemoryDB의 기능을 확장하며 기존 MemoryDB 기능과 함께 사용할 수 있습니다. 자세한 내용은 GitHub의 LLM 및 RAG 리포지토리를 사용한 질문 답변을 참조하세요.

다음 다이어그램은 MemoryDB를 벡터 데이터베이스로 사용하는 샘플 아키텍처를 보여줍니다.



다음은 MemoryDB 사용의 장점입니다.

- 플랫 인덱싱 알고리즘과 HNSW 인덱싱 알고리즘을 모두 지원합니다. 자세한 내용은 이제 AWS 뉴스 블로그에서 Amazon MemoryDB에 대한 벡터 검색을 참조하세요.
- 파운데이션 모델의 버퍼 메모리 역할을 할 수도 있습니다. 즉, 이전에 답변한 질문은 검색 및 생성 프로세스를 다시 거치는 대신 버퍼에서 검색됩니다. 다음 다이어그램에서는 이러한 프로세스를 보여줍니다.



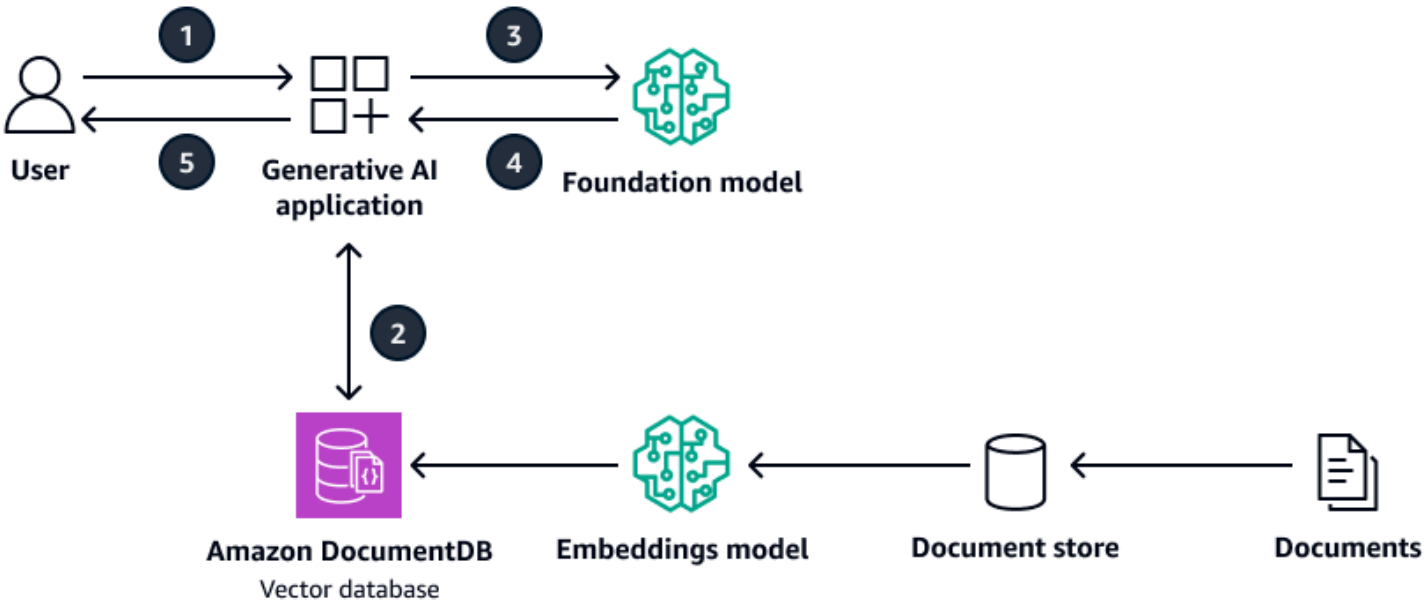

- 이 아키텍처는 인 메모리 데이터베이스를 사용하기 때문에 의미 체계 검색을 위한 한 자릿수 밀리초 쿼리 시간을 제공합니다.

- 95~99% 리콜로 초당 최대 33,000개의 쿼리를 제공하고 99% 리콜 이상으로 초당 26,500개의 쿼리를 제공합니다. 자세한 내용은에서 AWS Amazon MemoryDB 비디오에 대한 re:Invent 2023 - 초저지연 벡터 검색을 참조하세요YouTube.

## Amazon DocumentDB

Amazon DocumentDB(MongoDB 호환)는 빠르고 신뢰할 수 있는 완전 관리형 데이터베이스 서비스입니다. 클라우드에서 MongoDB호환 데이터베이스를 쉽게 설정, 운영 및 확장할 수 있습니다. Amazon DocumentDB에 대한 벡터 검색은 JSON 기반 문서 데이터베이스의 유연성과 풍부한 쿼리 기능을 벡터 검색의 기능과 결합합니다. 자세한 내용은 GitHub의 LLM 및 RAG 리포지토리를 사용한 질문 답변을 참조하세요.

다음 다이어그램은 Amazon DocumentDB를 벡터 데이터베이스로 사용하는 샘플 아키텍처를 보여줍니다.




이 다이어그램은 다음 워크플로를 보여줍니다.

1. 사용자가 생성형 AI 애플리케이션에 쿼리를 제출합니다.
2. 생성형 AI 애플리케이션은 Amazon DocumentDB 벡터 데이터베이스에서 유사성 검색을 수행하고 관련 문서 추출을 검색합니다.
3. 생성형 AI 애플리케이션은 검색된 컨텍스트로 사용자 쿼리를 업데이트하고 대상 파운데이션 모델에 프롬프트를 제출합니다.

4. 파운데이션 모델은 컨텍스트를 사용하여 사용자의 질문에 대한 응답을 생성하고 응답을 반환합니다.
5. 생성형 AI 애플리케이션은 사용자에게 응답을 반환합니다.

다음은 Amazon DocumentDB 사용의 장점입니다.

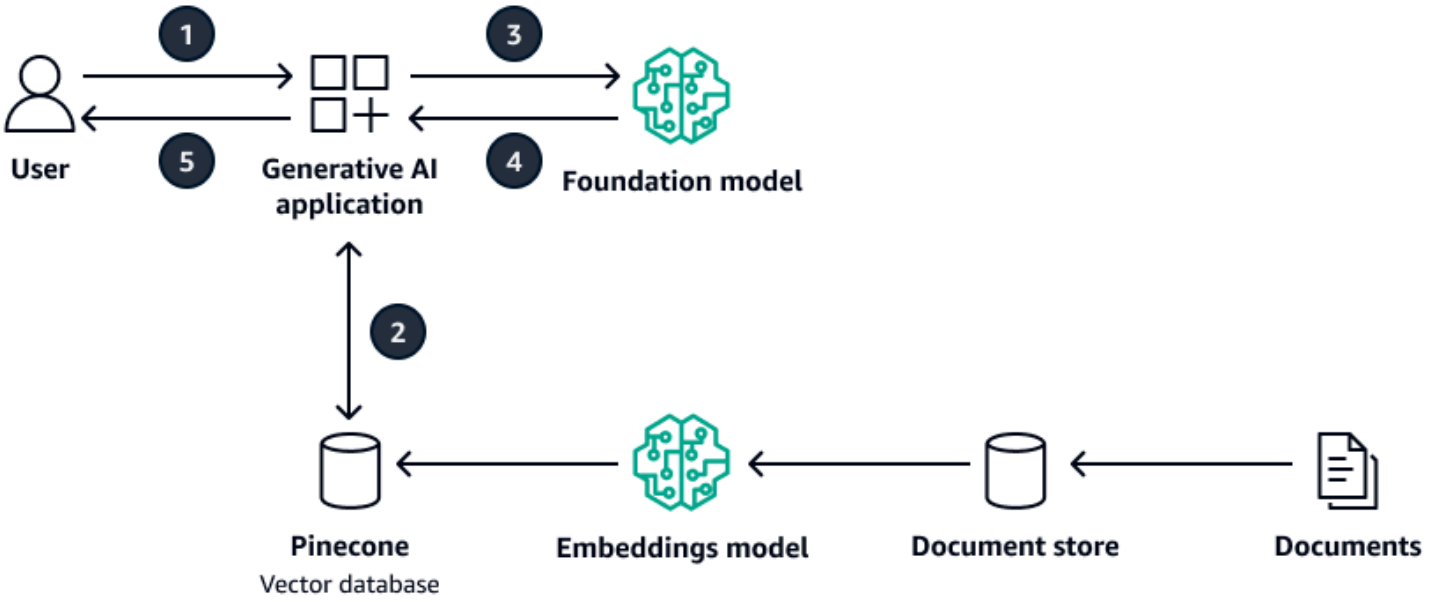
- HNSW 및 IVFFlat 인덱싱 방법을 모두 지원합니다.
- 벡터 데이터에서 최대 2,000개의 차원을 지원하고 유클리드, 코사인 및 점 제품 거리 지표를 지원합니다.
- 밀리초의 응답 시간을 제공합니다.

## Pinecone

Pinecone는 프로덕션 애플리케이션에 벡터 검색을 추가하는 데 도움이 되는 완전 관리형 벡터 데이터베이스입니다. 를 통해 사용할 수 있습니다AWS Marketplace. 결제는 사용량을 기준으로 하며, 요금은 포드 가격에 포드 수를 곱하여 계산됩니다. 를 사용하는 RAG 기반 시스템을 구축하는 방법에 대한 자세한 내용은 다음 AWS 블로그 게시물을 Pinecone참조하세요.

- Amazon SageMaker AI JumpStart의 Pinecone 벡터 데이터베이스 및 Llama-2를 사용하여 RAG를 통해 할루시네이션 완화
- Amazon SageMaker AI Studio를 사용하여 빠른 실험을 Pinecone 위해 Llama 2, LangChain및를 사용하여 RAG 질문 응답 솔루션 구축

다음 다이어그램은를 벡터 데이터베이스Pinecone로 사용하는 샘플 아키텍처를 보여줍니다.




이 다이어그램은 다음 워크플로를 보여줍니다.

1. 사용자가 생성형 AI 애플리케이션에 쿼리를 제출합니다.
2. 생성형 AI 애플리케이션은 Pinecone 벡터 데이터베이스에서 유사성 검색을 수행하고 관련 문서 추출을 검색합니다.
3. 생성형 AI 애플리케이션은 검색된 컨텍스트로 사용자 쿼리를 업데이트하고 대상 파운데이션 모델에 프롬프트를 제출합니다.
4. 파운데이션 모델은 컨텍스트를 사용하여 사용자의 질문에 대한 응답을 생성하고 응답을 반환합니다.
5. 생성형 AI 애플리케이션은 사용자에게 응답을 반환합니다.

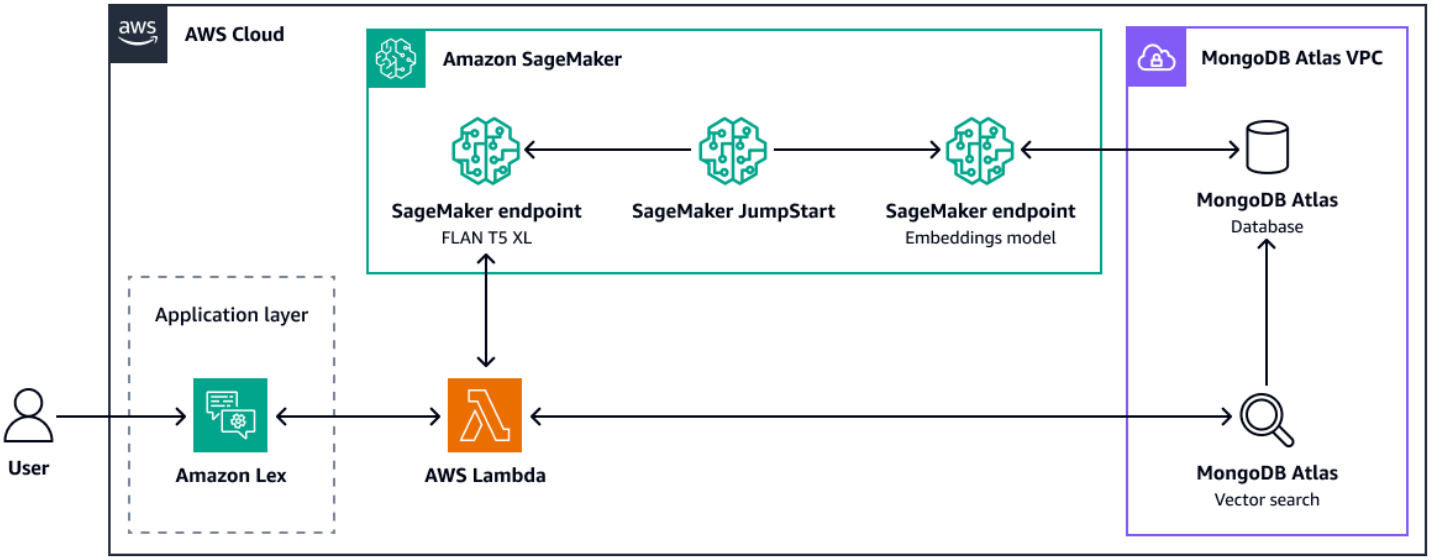
다음은 사용의 장점입니다. Pinecone

- 완전 관리형 벡터 데이터베이스이며 자체 인프라 관리의 오버헤드를 덜어줍니다.
- 필터링, 라이브 인덱스 업데이트 및 키워드 부스팅(하이브리드 검색)의 추가 기능을 제공합니다.

## MongoDB Atlas

MongoDB Atlas는 배포 배포 및 관리의 모든 복잡성을 처리하는 완전관리형 클라우드 데이터베이스입니다 AWS. 에 대한 벡터 검색을 MongoDB Atlas 사용하여 MongoDB 데이터베이스에 벡터 임베딩을 저장할 수 있습니다. Amazon Bedrock 지식 기반은 벡터 스토리지MongoDB Atlas를 지원합니다. 자세한 내용은 MongoDB 설명서의 Amazon Bedrock 지식 기반 통합 시작하기를 참조하세요.

RAG에 대한 MongoDB Atlas 벡터 검색을 사용하는 방법에 대한 자세한 내용은 를 사용한 검색 증강 생성LangChain, Amazon SageMaker AI JumpStart 및 MongoDB Atlas 의미 체계 검색(AWS 블로그 게시물)을 참조하세요. 다음 다이어그램은이 블로그 게시물에 자세히 설명된 솔루션 아키텍처를 보여줍니다.




다음은 MongoDB Atlas 벡터 검색 사용의 장점입니다.

- 의 기존 구현MongoDB Atlas을 사용하여 벡터 임베딩을 저장하고 검색할 수 있습니다.
- MongoDB 쿼리 API를 사용하여 벡터 임베딩을 쿼리할 수 있습니다.
- 벡터 검색 및 데이터베이스를 독립적으로 조정할 수 있습니다.
- 벡터 임베딩은 소스 데이터(문서) 근처에 저장되므로 인덱싱 성능이 향상됩니다.

## Weaviate

Weaviate는 텍스트 및 이미지와 같은 멀티모달 미디어 유형을 지원하는 인기 있는 오픈 소스, 지연 시간이 짧은 벡터 데이터베이스입니다. 데이터베이스는 벡터 검색과 구조화된 필터링을 결합하는 객체와 벡터를 모두 저장합니다. Weaviate 및 Amazon Bedrock을 사용하여 RAG 워크플로를 빌드하는 방법에 대한 자세한 내용은 Amazon Bedrock의 Cohere 파운데이션 모델로 엔터프라이즈 지원 생성형 AI 솔루션 빌드 및의 Weaviate 벡터 데이터베이스 AWS Marketplace(AWS 블로그 게시물)를 참조하세요.

다음은 사용의 장점입니다Weaviate.

- 오픈 소스이며 강력한 커뮤니티의 지원을 받습니다.
- 하이브리드 검색(벡터 및 키워드 모두)을 위해 구축되었습니다.

- 서비스 AWS 형 관리형 소프트웨어(SaaS) 또는 Kubernetes 클러스터로에 배포할 수 있습니다.

# RAG 워크플로용 생성기

대규모 언어 모델(LLMs)은 방대한 양의 데이터를 기반으로 사전 훈련된 매우 큰 딥 러닝 모델입니다. 매우 유연합니다. LLMs 질문 답변, 문서 요약, 언어 번역, 문장 완성과 같은 다양한 작업을 수행할 수 있습니다. 콘텐츠 생성과 사람들이 검색 엔진 및 가상 어시스턴트를 사용하는 방식을 방해할 가능성이 있습니다. 완벽하지는 않지만 LLMs 비교적 작은 프롬프트 또는 입력 수를 기반으로 예측할 수 있는 놀라운 능력을 보여줍니다.

LLMs RAG 솔루션의 중요한 구성 요소입니다. 사용자 지정 RAG 아키텍처의 경우 기본 옵션 역할을 AWS 서비스 하는 두 가지가 있습니다.

- Amazon Bedrock은 통합 API를 통해 선도적인 AI 회사 및 Amazon의 LLMs을 사용할 수 있도록 하는 완전관리형 서비스입니다.
- Amazon SageMaker AI JumpStart는 파운데이션 모델, 내장 알고리즘 및 사전 구축된 ML 솔루션을 제공하는 ML 허브입니다. SageMaker AI JumpStart를 사용하면 파운데이션 모델을 포함하여 사전 훈련된 모델에 액세스할 수 있습니다. 자체 데이터를 사용하여 사전 훈련된 모델을 미세 조정할 수도 있습니다.

## Amazon Bedrock

Amazon Bedrock은 Anthropic, , , Stability AI, MetaCohere, 및 AI21 Labs Mistral AIAmazon의 업계 최고의 모델을 제공합니다. 전체 목록은 Amazon Bedrock에서 지원되는 파운데이션 모델을 참조하세요. Amazon Bedrock을 사용하면 자체 데이터로 모델을 사용자 지정할 수도 있습니다.

모델 성능을 평가하여 RAG 사용 사례에 가장 적합한 모델을 결정할 수 있습니다. 최신 모델을 테스트하고 어떤 기능과 기능이 최상의 결과를 제공하고 최저 가격을 제공하는지 테스트할 수도 있습니다. Anthropic Claude Sonnet 모델은 다양한 작업에서 우수하고 높은 수준의 신뢰성과 예측 가능성을 제공하므로 RAG 애플리케이션에 일반적으로 사용됩니다.

## SageMaker AI JumpStart

SageMaker AI JumpStart는 다양한 문제 유형에 대해 사전 훈련된 오픈 소스 모델을 제공합니다. 배포 전에 이러한 모델을 점진적으로 훈련하고 미세 조정할 수 있습니다. Amazon SageMaker AI Studio의 SageMaker AI JumpStart 랜딩 페이지를 통해 사전 훈련된 모델, 솔루션 템플릿 및 예제에 액세스하거나 SageMaker AI Python SDK를 사용할 수 있습니다.

SageMaker AI JumpStart는 콘텐츠 작성, 코드 생성, 질문 답변, 카피라이팅, 요약, 분류, 정보 검색 등과 같은 사용 사례를 위한 state-of-the-art 파운데이션 모델을 제공합니다. JumpStart 파운데이션 모델을 사용하여 자체 생성형 AI 솔루션을 구축하고 사용자 지정 솔루션을 추가 SageMaker AI 기능과 통합할 수 있습니다. 자세한 내용은 Amazon SageMaker AI JumpStart 시작하기를 참조하세요.

SageMaker AI JumpStart는 ML 수명 주기에 액세스, 사용자 지정 및 통합할 수 있도록 공개적으로 사용 가능한 파운데이션 모델을 온보딩하고 유지 관리합니다. 자세한 내용은 공개적으로 사용 가능한 파운데이션 모델을 참조하세요. SageMaker AI JumpStart에는 타사 공급자의 독점 파운데이션 모델도 포함되어 있습니다. 자세한 내용은 독점 파운데이션 모델을 참조하세요.

# 에서 검색 증강 생성 옵션 선택 AWS

이 가이드의 완전 관리형 RAG 옵션 및 사용자 지정 RAG 아키텍처 섹션에서는 RAG 기반 검색 솔루션을 구축하기 위한 다양한 접근 방식을 설명합니다 AWS. 이 섹션에서는 사용 사례에 따라 이러한 옵션 중에서 선택하는 방법을 설명합니다. 경우에 따라 두 개 이상의 옵션이 작동할 수 있습니다. 이 시나리오에서 선택은 구현의 용이성, 조직에서 사용할 수 있는 기술, 회사의 정책 및 표준에 따라 달라집니다.

완전 관리형 및 사용자 지정 RAG 옵션을 다음 순서로 고려하고 사용 사례에 맞는 첫 번째 옵션을 선택하는 것이 좋습니다.

1. 다음과 같은 경우가 아니면 Amazon Q Business를 사용합니다.
   - 이 서비스는에서 사용할 수 없으며 AWS 리전데이터를 사용할 수 있는 리전으로 이동할 수 없습니다.
   - RAG 워크플로를 사용자 지정해야 하는 구체적인 이유가 있습니다.
   - 기존 벡터 데이터베이스 또는 특정 LLM을 사용하려는 경우
2. 다음과 같은 경우가 아니면 Amazon Bedrock에 대한 지식 기반을 사용합니다.
   - 지원되지 않는 벡터 데이터베이스가 있는 경우
   - RAG 워크플로를 사용자 지정해야 하는 구체적인 이유가 있습니다.
3. 다음과 같은 경우를 제외하고 Amazon Kendra를 원하는 생성기와 결합합니다.
   - 자체 벡터 데이터베이스를 선택하려고 함
   - 청킹 전략을 사용자 지정하려는 경우
4. 리트리버를 더 잘 제어하고 자체 벡터 데이터베이스를 선택하려는 경우:
   - 기존 벡터 데이터베이스가 없고 지연 시간이 짧거나 그래프 쿼리가 필요하지 않은 경우 Amazon OpenSearch Service를 사용하는 것이 좋습니다.
   - 기존 PostgreSQL 벡터 데이터베이스가 있는 경우 Amazon Aurora PostgreSQL 및 pgvector 옵션을 사용하는 것이 좋습니다.
   - 짧은 지연 시간이 필요한 경우 Amazon MemoryDB 또는 Amazon DocumentDB와 같은 인 메모리 옵션을 고려하세요.
   - 벡터 검색을 그래프 쿼리와 결합하려면 Amazon Neptune Analytics를 고려하세요.
   - 이미 타사 벡터 데이터베이스를 사용하고 있거나이 데이터베이스에서 특정 이점을 발견한 경우 Pinecone, MongoDB Atlas및를 고려하세요Weaviate.
5. LLM을 선택하려면:
   - Amazon Q Business를 사용하는 경우 LLM을 선택할 수 없습니다.

- Amazon Bedrock을 사용하는 경우 지원되는 파운데이션 모델 중 하나를 선택할 수 있습니다.
- Amazon Kendra 또는 사용자 지정 벡터 데이터베이스를 사용하는 경우이 가이드에 설명된 생성기 중 하나를 사용하거나 사용자 지정 LLM을 사용할 수 있습니다.

> **Note**
>
> 또한 사용자 지정 문서를 사용하여 기존 LLM을 미세 조정하여 응답의 정확도를 높일 수 있습니다. 자세한 내용은 이 안내서의 RAG 및 미세 조정 비교 섹션을 참조하세요.

6. 사용하려는 Amazon SageMaker AI Canvas의 기존 구현이 있거나 다른 LLMs의 RAG 응답을 비교하려는 경우 Amazon SageMaker AI Canvas를 고려하세요.

# 결론

이 가이드에서는 검색 증강 생성(RAG) 시스템을 빌드하기 위한 다양한 옵션을 설명합니다 AWS. Amazon Q Business 및 Amazon Bedrock 지식 기반과 같은 완전관리형 서비스로 시작할 수 있습니다. RAG 워크플로를 더 잘 제어하려면 사용자 지정 리트리버를 선택할 수 있습니다. 생성기의 경우 API를 사용하여 Amazon Bedrock에서 지원되는 LLM을 호출하거나 Amazon SageMaker AI JumpStart를 사용하여 자체 LLM을 배포할 수 있습니다. RAG 옵션 선택의 권장 사항을 검토하여 사용 사례에 가장 적합한 옵션을 결정합니다. 사용 사례에 가장 적합한 옵션을 선택한 후 이 가이드에 제공된 참조를 사용하여 RAG 기반 애플리케이션 구축을 시작합니다.

# 문서 기록

아래 표에 이 가이드의 주요 변경 사항이 설명되어 있습니다. 향후 업데이트에 대한 알림을 받으려면 RSS 피드를 구독하십시오.

| 변경 사항 | 설명 | 날짜 |
| --- | --- | --- |
| 최초 게시 | — | 2024년 10월 28일 |

# AWS 권장 가이드 용어집

다음은 AWS 권장 가이드에서 제공하는 전략, 가이드 및 패턴에서 일반적으로 사용되는 용어입니다. 용어집 항목을 제안하려면 용어집 끝에 있는 피드백 제공 링크를 사용하십시오.

## 숫자

7가지 전략

애플리케이션을 클라우드로 이전하기 위한 7가지 일반적인 마이그레이션 전략 이러한 전략은 Gartner가 2011년에 파악한 5가지 전략을 기반으로 하며 다음으로 구성됩니다.

- 리팩터링/리아키텍트 - 클라우드 네이티브 기능을 최대한 활용하여 애플리케이션을 이동하고 해당 아키텍처를 수정함으로써 민첩성, 성능 및 확장성을 개선합니다. 여기에는 일반적으로 운영 체제와 데이터베이스 이식이 포함됩니다. 예: 온프레미스 Oracle 데이터베이스를 Amazon Aurora PostgreSQL 호환 에디션으로 마이그레이션합니다.
- 리플랫포밍(리프트 앤드 리셰이프) - 애플리케이션을 클라우드로 이동하고 일정 수준의 최적화를 도입하여 클라우드 기능을 활용합니다. 예: 온프레미스 Oracle 데이터베이스를 AWS 클라우드의 Amazon Relational Database Service(Amazon RDS) for Oracle로 마이그레이션합니다.
- 재구매(드롭 앤드 숍) - 일반적으로 기존 라이선스에서 SaaS 모델로 전환하여 다른 제품으로 전환합니다. 예: 고객 관계 관리(CRM) 시스템을 Salesforce.com으로 마이그레이션합니다.
- 리호스팅(리프트 앤드 시프트) - 애플리케이션을 변경하지 않고 클라우드로 이동하여 클라우드 기능을 활용합니다. 예: 온프레미스 Oracle 데이터베이스를 AWS 클라우드클라우드의 EC2 인스턴스에 있는 Oracle로 마이그레이션합니다.
- 재배치(하이퍼바이저 수준의 리프트 앤 시프트) - 새 하드웨어를 구매하거나, 애플리케이션을 다시 작성하거나, 기존 운영을 수정하지 않고도 인프라를 클라우드로 이동합니다. 온프레미스 플랫폼에서 동일한 플랫폼의 클라우드 서비스로 서버를 마이그레이션합니다. 예: Microsoft Hyper-V 애플리케이션을 로 마이그레이션합니다 AWS.
- 유지(보관) - 소스 환경에 애플리케이션을 유지합니다. 대규모 리팩터링이 필요하고 해당 작업을 나중으로 연기하려는 애플리케이션과 비즈니스 차원에서 마이그레이션할 이유가 없어 유지하려는 레거시 애플리케이션이 여기에 포함될 수 있습니다.
- 사용 중지 - 소스 환경에서 더 이상 필요하지 않은 애플리케이션을 폐기하거나 제거합니다.

# A

ABAC

속성 기반 액세스 제어를 참조하세요.

추상화된 서비스

관리형 서비스를 참조하세요.

ACID

원자성, 일관성, 격리성, 내구성을 참조하세요.

능동-능동 마이그레이션

양방향 복제 도구 또는 이중 쓰기 작업을 사용하여 소스 데이터베이스와 대상 데이터베이스가 동기화된 상태로 유지되고, 두 데이터베이스 모두 마이그레이션 중 연결 애플리케이션의 트랜잭션을 처리하는 데이터베이스 마이그레이션 방법입니다. 이 방법은 일회성 전환이 필요한 대신 소규모의 제어된 배치로 마이그레이션을 지원합니다. 더 유연하지만 액티브 패시브 마이그레이션보다 더 많은 작업이 필요합니다.

능동-수동 마이그레이션

소스 데이터베이스와 대상 데이터베이스가 동기화된 상태로 유지되지만 소스 데이터베이스만 연결 애플리케이션의 트랜잭션을 처리하고 데이터는 대상 데이터베이스로 복제되는 데이터베이스 마이그레이션 방법입니다. 대상 데이터베이스는 마이그레이션 중 어떤 트랜잭션도 허용하지 않습니다.

집계 함수

행 그룹에서 작동하고 그룹에 대한 단일 반환 값을 계산하는 SQL 함수입니다. 집계 함수의 예로 `SUM` 및 `MAX`가 있습니다.

AI

인공 지능을 참조하세요.

AIOps

인공 지능 운영을 참조하세요.

익명화

데이터세트에서 개인 정보를 영구적으로 삭제하는 프로세스입니다. 익명화는 개인 정보 보호에 도움이 될 수 있습니다. 익명화된 데이터는 더 이상 개인 데이터로 간주되지 않습니다.

안티 패턴

솔루션이 다른 솔루션보다 비생산적이거나 비효율적이거나 덜 효과적이어서 반복되는 문제에 자주 사용되는 솔루션입니다.

애플리케이션 제어

맬웨어로부터 시스템을 보호하기 위해 승인된 애플리케이션만 사용하도록 허용하는 보안 접근 방식입니다.

애플리케이션 포트폴리오

애플리케이션 구축 및 유지 관리 비용과 애플리케이션의 비즈니스 가치를 비롯하여 조직에서 사용하는 각 애플리케이션에 대한 세부 정보 모음입니다. 이 정보는 포트폴리오 탐색 및 분석 프로세스의 핵심이며 마이그레이션, 현대화 및 최적화할 애플리케이션을 식별하고 우선순위를 정하는 데 도움이 됩니다.

인공 지능

컴퓨터 기술을 사용하여 학습, 문제 해결, 패턴 인식 등 일반적으로 인간과 관련된 인지 기능을 수행하는 것을 전문으로 하는 컴퓨터 과학 분야입니다. 자세한 내용은 What is Artificial Intelligence?를 참조하십시오.

인공 지능 운영(AIOps)

기계 학습 기법을 사용하여 운영 문제를 해결하고, 운영 인시던트 및 사용자 개입을 줄이고, 서비스 품질을 높이는 프로세스입니다. AWS 마이그레이션 전략에서 AIOps가 사용되는 방법에 대한 자세한 내용은 운영 통합 가이드를 참조하십시오.

비대칭 암호화

한 쌍의 키, 즉 암호화를 위한 퍼블릭 키와 복호화를 위한 프라이빗 키를 사용하는 암호화 알고리즘입니다. 퍼블릭 키는 복호화에 사용되지 않으므로 공유할 수 있지만 프라이빗 키에 대한 액세스는 엄격히 제한되어야 합니다.

원자성, 일관성, 격리성, 내구성(ACID)

오류, 정전 또는 기타 문제가 발생한 경우에도 데이터베이스의 데이터 유효성과 운영 신뢰성을 보장하는 소프트웨어 속성 세트입니다.

ABAC(속성 기반 액세스 제어)

부서, 직무, 팀 이름 등의 사용자 속성을 기반으로 세분화된 권한을 생성하는 방식입니다. 자세한 내용은 AWS Identity and Access Management (IAM) 설명서의 용 ABAC AWS를 참조하세요.

신뢰할 수 있는 데이터 소스

가장 신뢰할 수 있는 정보 소스로 간주되는 기본 버전의 데이터를 저장하는 위치입니다. 익명화, 편집 또는 가명화와 같은 데이터 처리 또는 수정의 목적으로 신뢰할 수 있는 데이터 소스의 데이터를 다른 위치로 복사할 수 있습니다.

가용 영역

다른 가용 영역의 장애로부터 격리 AWS 리전 되고 동일한 리전의 다른 가용 영역에 저렴하고 지연 시간이 짧은 네트워크 연결을 제공하는 내의 고유한 위치입니다.

AWS 클라우드 채택 프레임워크(AWS CAF)

조직이 클라우드로 성공적으로 전환 AWS 하기 위한 효율적이고 효과적인 계획을 개발하는 데 도움이 되는의 지침 및 모범 사례 프레임워크입니다. AWS CAF는 지침을 비즈니스, 사람, 거버넌스, 플랫폼, 보안 및 운영이라는 6가지 중점 영역으로 구성합니다. 비즈니스, 사람 및 거버넌스 관점은 비즈니스 기술과 프로세스에 초점을 맞추고, 플랫폼, 보안 및 운영 관점은 전문 기술과 프로세스에 중점을 둡니다. 예를 들어, 사람 관점은 인사(HR), 직원 배치 기능 및 인력 관리를 담당하는 이해관계자를 대상으로 합니다. 이러한 관점에서 AWS CAF는 성공적인 클라우드 채택을 위해 조직을 준비하는 데 도움이 되는 인력 개발, 교육 및 커뮤니케이션에 대한 지침을 제공합니다. 자세한 내용은 AWS CAF 웹사이트와 AWS CAF 백서를 참조하세요.

AWS 워크로드 검증 프레임워크(AWS WQF)

데이터베이스 마이그레이션 워크로드를 평가하고, 마이그레이션 전략을 권장하고, 작업 견적을 제공하는 도구입니다. AWS WQF는 AWS Schema Conversion Tool (AWS SCT)에 포함되어 있습니다. 데이터베이스 스키마 및 코드 객체, 애플리케이션 코드, 종속성 및 성능 특성을 분석하고 평가 보고서를 제공합니다.

# B

악성 봇

개인 또는 조직을 방해하거나 해를 입히기 위한 봇입니다.

BCP

비즈니스 연속성 계획을 참조하세요.

동작 그래프

리소스 동작과 시간 경과에 따른 상호 작용에 대한 통합된 대화형 뷰입니다. Amazon Detective에서 동작 그래프를 사용하여 실패한 로그온 시도, 의심스러운 API 직접 호출 및 유사한 작업을 검사할 수 있습니다. 자세한 내용은 Detective 설명서의 Data in a behavior graph를 참조하십시오.

빅 엔디안 시스템

가장 중요한 바이트를 먼저 저장하는 시스템입니다. 엔디안도 참조하세요.

바이너리 분류

바이너리 결과(가능한 두 클래스 중 하나)를 예측하는 프로세스입니다. 예를 들어, ML 모델이 “이 이메일이 스팸인가요, 스팸이 아닌가요?”, '이 제품은 책인가요, 자동차인가요?' 등의 문제를 예측해야 할 수 있습니다.

블룸 필터

요소가 세트의 멤버인지 여부를 테스트하는 데 사용되는 메모리 효율성이 높은 확률론적 데이터 구조입니다.

블루/그린(Blue/Green) 배포

동일하지만 별개의 두 환경을 생성하는 배포 전략입니다. 하나의 환경(파란색)에서 현재 애플리케이션 버전을 실행하고 새 애플리케이션 버전은 다른 환경(녹색)에서 실행합니다. 이 전략을 사용하면 영향을 최소화하면서 신속하게 롤백할 수 있습니다.

bot

인터넷을 통해 자동화된 태스크를 실행하고 인적 활동이나 상호 작용을 시뮬레이션하는 소프트웨어 애플리케이션입니다. 인터넷에서 정보를 인덱싱하는 웹 크롤러와 같이 유용하거나 이로운 봇도 있습니다. 악성 봇이라고 하는 다른 일부 봇은 개인 또는 조직을 방해하거나 해를 입히기 위한 봇입니다.

봇넷

맬웨어에 감염되고 봇 허더 또는 봇 운영자와 같은 단일 당사자가 제어하는 봇 네트워크입니다. 봇넷은 봇의 규모와 봇의 영향 범위를 확대하는 가장 잘 알려진 메커니즘입니다.

브랜치

코드 리포지토리의 포함된 영역입니다. 리포지토리에 생성되는 첫 번째 브랜치가 기본 브랜치입니다. 기존 브랜치에서 새 브랜치를 생성한 다음 새 브랜치에서 기능을 개발하거나 버그를 수정할 수 있습니다. 기능을 구축하기 위해 생성하는 브랜치를 일반적으로 기능 브랜치라고 합니다. 기능을 출시할 준비가 되면 기능 브랜치를 기본 브랜치에 다시 병합합니다. 자세한 내용은 About branches(GitHub 설명서)를 참조하십시오.

긴급 액세스 권한

예외적인 상황에서 승인된 프로세스를 통해 사용자가 일반적으로 액세스할 권한이 없는에 액세스할 수 AWS 계정 있는 빠른 방법입니다. 자세한 내용은 AWS Well-Architected 지침의 Implement break-glass procedures 지표를 참조하세요.

브라운필드 전략

사용자 환경의 기존 인프라 시스템 아키텍처에 브라운필드 전략을 채택할 때는 현재 시스템 및 인프라의 제약 조건을 중심으로 아키텍처를 설계합니다. 기존 인프라를 확장하는 경우 브라운필드 전략과 그린필드 전략을 혼합할 수 있습니다.

버퍼 캐시

가장 자주 액세스하는 데이터가 저장되는 메모리 영역입니다.

사업 역량

기업이 가치를 창출하기 위해 하는 일(예: 영업, 고객 서비스 또는 마케팅)입니다. 마이크로서비스 아키텍처 및 개발 결정은 비즈니스 역량에 따라 이루어질 수 있습니다. 자세한 내용은 백서의 AWS에서 컨테이너화된 마이크로서비스 실행의 비즈니스 역량 중심의 구성화 섹션을 참조하십시오.

비즈니스 연속성 계획(BCP)

대규모 마이그레이션과 같은 중단 이벤트가 운영에 미치는 잠재적 영향을 해결하고 비즈니스가 신속하게 운영을 재개할 수 있도록 지원하는 계획입니다.

# C

CAF

AWS Cloud Adoption Framework를 참조하세요.

카나리 배포

최종 사용자에게 제공하는 느린 증분 릴리스 버전입니다. 확신이 들면 새 버전을 배포하고 현재 버전을 완전히 교체합니다.

CCoE

클라우드 혁신 센터를 참조하세요.

CDC

데이터 캡처 변경을 참조하세요.

변경 데이터 캡처(CDC)

데이터베이스 테이블과 같은 데이터 소스의 변경 내용을 추적하고 변경 사항에 대한 메타데이터를 기록하는 프로세스입니다. 대상 시스템의 변경 내용을 감사하거나 복제하여 동기화를 유지하는 등의 다양한 용도로 CDC를 사용할 수 있습니다.

카오스 엔지니어링

시스템의 복원력을 테스트하기 위해 의도적으로 장애나 중단 이벤트를 도입합니다. AWS Fault Injection Service (AWS FIS)를 사용하여 AWS 워크로드에 스트레스를 주고 응답을 평가하는 실험을 수행할 수 있습니다.

CI/CD

지속적 통합 및 지속적 전송을 참조하세요.

분류

예측을 생성하는 데 도움이 되는 분류 프로세스입니다. 분류 문제에 대한 ML 모델은 이산 값을 예측합니다. 이산 값은 항상 서로 다릅니다. 예를 들어, 모델이 이미지에 자동차가 있는지 여부를 평가해야 할 수 있습니다.

클라이언트측 암호화

대상이 데이터를 AWS 서비스 수신하기 전에 로컬에서 데이터를 암호화합니다.

클라우드 혁신 센터(CCoE)

클라우드 모범 사례 개발, 리소스 동원, 마이그레이션 타임라인 설정, 대규모 혁신을 통한 조직 선도 등 조직 전체에서 클라우드 채택 노력을 추진하는 다분야 팀입니다. 자세한 내용은 AWS 클라우드 엔터프라이즈 전략 블로그의 CCoE 게시물을 참조하세요.

클라우드 컴퓨팅

원격 데이터 스토리지와 IoT 디바이스 관리에 일반적으로 사용되는 클라우드 기술 클라우드 컴퓨팅은 일반적으로 엣지 컴퓨팅 기술에 연결되어 있습니다.

클라우드 운영 모델

IT 조직에서 하나 이상의 클라우드 환경을 구축, 성숙화 및 최적화하는 데 사용되는 운영 모델입니다. 자세한 내용은 클라우드 운영 모델 구축을 참조하십시오.

클라우드 채택 단계

조직이 AWS 클라우드로 마이그레이션할 때 일반적으로 거치는 4단계는 다음과 같습니다.

- 프로젝트 - 개념 증명 및 학습 목적으로 몇 가지 클라우드 관련 프로젝트 실행
- 기반 – 클라우드 채택 확장을 위한 기초 투자(예: 랜딩 존 생성, CCoE 정의, 운영 모델 구축)
- 마이그레이션 - 개별 애플리케이션 마이그레이션
- Re-invention - 제품 및 서비스 최적화와 클라우드 혁신

이러한 단계는 Stephen Orban이 블로그 게시물 The Journey Toward Cloud-First and the Stages of Adoption on the AWS 클라우드 Enterprise Strategy 블로그에서 정의했습니다. AWS 마이그레이션 전략과 어떤 관련이 있는지에 대한 자세한 내용은 마이그레이션 준비 가이드를 참조하세요.

CMDB

구성 관리 데이터베이스를 참조하세요.

코드 리포지토리

소스 코드와 설명서, 샘플, 스크립트 등의 기타 자산이 버전 관리 프로세스를 통해 저장되고 업데이트되는 위치입니다. 일반적인 클라우드 리포지토리로 GitHub 또는 Bitbucket Cloud가 포함됩니다. 코드의 각 버전을 브랜치라고 합니다. 마이크로서비스 구조에서 각 리포지토리는 단일 기능 전용입니다. 단일 CI/CD 파이프라인은 여러 리포지토리를 사용할 수 있습니다.

콜드 캐시

비어 있거나, 제대로 채워지지 않았거나, 오래되었거나 관련 없는 데이터를 포함하는 버퍼 캐시입니다. 주 메모리나 디스크에서 데이터베이스 인스턴스를 읽어야 하기 때문에 성능에 영향을 미치며, 이는 버퍼 캐시에서 읽는 것보다 느립니다.

콜드 데이터

거의 액세스되지 않고 일반적으로 과거 데이터인 데이터. 이런 종류의 데이터를 쿼리할 때는 일반적으로 느린 쿼리가 허용됩니다. 이 데이터를 성능이 낮고 비용이 저렴한 스토리지 계층 또는 클래스로 옮기면 비용을 절감할 수 있습니다.

컴퓨터 비전(CV)

기계 학습을 사용하여 디지털 이미지 및 비디오와 같은 시각적 형식에서 정보를 분석하고 추출하는 AI 필드입니다. 예를 들어 Amazon SageMaker AI는 CV에 대한 이미지 처리 알고리즘을 제공합니다.

구성 드리프트

워크로드의 경우 구성이 예상되는 상태에서 변경됩니다. 이로 인해 워크로드가 규정을 준수하지 않을 수 있으며, 이는 일반적으로 점진적이고 의도되지 않은 작업입니다.

구성 관리 데이터베이스(CMDB)

하드웨어 및 소프트웨어 구성 요소와 해당 구성을 포함하여 데이터베이스와 해당 IT 환경에 대한 정보를 저장하고 관리하는 리포지토리입니다. 일반적으로 마이그레이션의 포트폴리오 탐색 및 분석 단계에서 CMDB의 데이터를 사용합니다.

규정 준수 팩

규정 준수 및 보안 검사를 사용자 지정하기 위해 조합할 수 있는 AWS Config 규칙 및 수정 작업 모음입니다. YAML 템플릿을 사용하여 적합성 팩을 AWS 계정 및 리전 또는 조직 전체에 단일 엔터티로 배포할 수 있습니다. 자세한 내용은 AWS Config 설명서의 적합성 팩을 참조하세요.

지속적 통합 및 지속적 전달(CI/CD)

소프트웨어 릴리스 프로세스의 소스, 빌드, 테스트, 스테이징 및 프로덕션 단계를 자동화하는 프로세스입니다. CI/CD는 일반적으로 파이프라인으로 설명됩니다. CI/CD를 통해 프로세스를 자동화하고, 생산성을 높이고, 코드 품질을 개선하고, 더 빠르게 제공할 수 있습니다. 자세한 내용은 지속적 전달의 이점을 참조하십시오. CD는 지속적 배포를 의미하기도 합니다. 자세한 내용은 지속적 전달(Continuous Delivery)과 지속적인 개발을 참조하십시오.

CV

컴퓨터 비전을 참조하세요.

# D

저장 데이터

스토리지에 있는 데이터와 같이 네트워크에 고정되어 있는 데이터입니다.

데이터 분류

중요도와 민감도를 기준으로 네트워크의 데이터를 식별하고 분류하는 프로세스입니다. 이 프로세스는 데이터에 대한 적절한 보호 및 보존 제어를 결정하는 데 도움이 되므로 사이버 보안 위험 관리 전략의 중요한 구성 요소입니다. 데이터 분류는 AWS Well-Architected Framework의 보안 원칙 구성 요소입니다. 자세한 내용은 데이터 분류를 참조하십시오.

데이터 드리프트

프로덕션 데이터와 ML 모델 학습에 사용된 데이터 간의 상당한 차이 또는 시간 경과에 따른 입력 데이터의 의미 있는 변화. 데이터 드리프트는 ML 모델 예측의 전반적인 품질, 정확성 및 공정성을 저하시킬 수 있습니다.

전송 중 데이터

네트워크를 통과하고 있는 데이터입니다. 네트워크 리소스 사이를 이동 중인 데이터를 예로 들 수 있습니다.

데이터 메시

중앙 집중식 관리 및 거버넌스를 통해 분산되고 탈중앙화된 데이터 소유권을 제공하는 아키텍처 프레임워크입니다.

데이터 최소화

꼭 필요한 데이터만 수집하고 처리하는 원칙입니다. 에서 데이터를 최소화하면 개인 정보 보호 위험, 비용 및 분석 탄소 발자국을 줄일 AWS 클라우드 수 있습니다.

데이터 경계

신뢰할 수 있는 자격 증명만 예상 네트워크에서 신뢰할 수 있는 리소스에 액세스하도록 하는 데 도움이 되는 AWS 환경의 예방 가드레일 세트입니다. 자세한 내용은 데이터 경계 구축을 참조하세요 AWS.

데이터 사전 처리

원시 데이터를 ML 모델이 쉽게 구문 분석할 수 있는 형식으로 변환하는 것입니다. 데이터를 사전 처리한다는 것은 특정 열이나 행을 제거하고 누락된 값, 일관성이 없는 값 또는 중복 값을 처리함을 의미할 수 있습니다.

데이터 출처

라이프사이클 전반에 걸쳐 데이터의 출처와 기록을 추적하는 프로세스(예: 데이터 생성, 전송, 저장 방법).

데이터 주체

데이터를 수집 및 처리하는 개인입니다.

데이터 웨어하우스

분석과 같은 비즈니스 인텔리전스를 지원하는 데이터 관리 시스템입니다. 데이터 웨어하우스에는 보통 많은 양의 기록 데이터가 포함되며 일반적으로 쿼리 및 분석에 사용됩니다.

데이터 정의 언어(DDL)

데이터베이스에서 테이블 및 객체의 구조를 만들거나 수정하기 위한 명령문 또는 명령입니다.

데이터베이스 조작 언어(DML)

데이터베이스에서 정보를 수정(삽입, 업데이트 및 삭제)하기 위한 명령문 또는 명령입니다.

DDL

데이터 정의 언어를 참조하세요.

딥 앙상블

예측을 위해 여러 딥 러닝 모델을 결합하는 것입니다. 딥 앙상블을 사용하여 더 정확한 예측을 얻거나 예측의 불확실성을 추정할 수 있습니다.

딥 러닝

여러 계층의 인공 신경망을 사용하여 입력 데이터와 관심 대상 변수 간의 매핑을 식별하는 ML 하위 분야입니다.

심층 방어

네트워크와 그 안의 데이터 기밀성, 무결성 및 가용성을 보호하기 위해 컴퓨터 네트워크 전체에 일련의 보안 메커니즘과 제어를 신중하게 계층화하는 정보 보안 접근 방식입니다. 이 전략을 채택하면 AWS Organizations 구조의 여러 계층에 여러 제어를 AWS추가하여 리소스를 보호할 수 있습니다. 예를 들어, 심층 방어 접근 방식은 다단계 인증, 네트워크 세분화 및 암호화를 결합할 수 있습니다.

위임된 관리자

에서 AWS Organizations호환되는 서비스는 AWS 멤버 계정을 등록하여 조직의 계정을 관리하고 해당 서비스에 대한 권한을 관리할 수 있습니다. 이러한 계정을 해당 서비스의 위임된 관리자라고 합니다. 자세한 내용과 호환되는 서비스 목록은 AWS Organizations 설명서의 AWS Organizations와 함께 사용할 수 있는 AWS 서비스를 참조하십시오.

배포

대상 환경에서 애플리케이션, 새 기능 또는 코드 수정 사항을 사용할 수 있도록 하는 프로세스입니다. 배포에는 코드 베이스의 변경 사항을 구현한 다음 애플리케이션 환경에서 해당 코드베이스를 구축하고 실행하는 작업이 포함됩니다.

개발 환경

환경을 참조하세요.

탐지 제어

이벤트 발생 후 탐지, 기록 및 알림을 수행하도록 설계된 보안 제어입니다. 이러한 제어는 기존의 예방적 제어를 우회한 보안 이벤트를 알리는 2차 방어선입니다. 자세한 내용은 AWS에서 보안 제어 구현의 탐지 제어를 참조하세요.

개발 가치 흐름 매핑 (DVSM)

소프트웨어 개발 라이프사이클에서 속도와 품질에 부정적인 영향을 미치는 제약 조건을 식별하고 우선 순위를 지정하는 데 사용되는 프로세스입니다. DVSM은 원래 린 제조 방식을 위해 설계된 가치 흐름 매핑 프로세스를 확장합니다. 소프트웨어 개발 프로세스를 통해 가치를 창출하고 이동하는 데 필요한 단계와 팀에 중점을 둡니다.

디지털 트윈

건물, 공장, 산업 장비 또는 생산 라인과 같은 실제 시스템을 가상으로 표현한 것입니다. 디지털 트윈은 예측 유지 보수, 원격 모니터링, 생산 최적화를 지원합니다.

차원 테이블

스타 스키마에서 팩트 테이블의 정량적 데이터에 대한 데이터 속성을 포함하는 더 작은 테이블을 말합니다. 차원 테이블 속성은 일반적으로 텍스트 필드나 텍스트처럼 동작하는 개별 숫자입니다. 이러한 속성은 보통 쿼리 제약, 필터링 및 결과 세트 레이블 지정에 사용됩니다.

재해

워크로드 또는 시스템이 기본 배포 위치에서 비즈니스 목표를 달성하지 못하게 방해하는 이벤트입니다. 이러한 이벤트는 자연재해, 기술적 오류, 의도하지 않은 구성 오류 또는 멀웨어 공격과 같은 사람의 행동으로 인한 결과일 수 있습니다.

재해 복구(DR)

재해로 인한 가동 중지 시간 및 데이터 손실을 최소화하기 위해 사용하는 전략 및 프로세스입니다. 자세한 내용은 AWS Well-Architected Framework의 Disaster Recovery of Workloads on AWS: Recovery in the Cloud를 참조하세요.

DML

데이터베이스 조작 언어를 참조하세요.

도메인 기반 설계

구성 요소를 각 구성 요소가 제공하는 진화하는 도메인 또는 핵심 비즈니스 목표에 연결하여 복잡한 소프트웨어 시스템을 개발하는 접근 방식입니다. 이 개념은 에릭 에반스에 의해 그의 저서인 도메인 기반 디자인: 소프트웨어 중심의 복잡성 해결(Boston: Addison-Wesley Professional, 2003)에서 소개되었습니다. Strangler Fig 패턴과 함께 도메인 기반 설계를 사용하는 방법에 대한 자세한 내용은 컨테이너 및 Amazon API Gateway를 사용하여 기존의 Microsoft ASP.NET(ASMX) 웹 서비스를 점진적으로 현대화하는 방법을 참조하십시오.

DR

재해 복구를 참조하세요.

드리프트 감지

기준이 되는 구성과의 편차 추적을 말합니다. 예를 들어 AWS CloudFormation 를 사용하여 시스템 리소스의 드리프트를 감지하거나를 사용하여 AWS Control Tower 거버넌스 요구 사항 준수에 영향을 미칠 수 있는 랜딩 존의 변경 사항을 감지할 수 있습니다.

DVSM

개발 가치 흐름 매핑을 참조하세요.

# E

EDA

탐색 데이터 분석을 참조하세요.

EDI

전자 데이터 교환을 참조하세요.

엣지 컴퓨팅

IoT 네트워크의 엣지에서 스마트 디바이스의 컴퓨팅 성능을 개선하는 기술 엣지 컴퓨팅은 클라우드 컴퓨팅에 비해 보다 통신 지연 시간을 줄이고 응답 시간을 개선할 수 있습니다.

전자 데이터 교환(EDI)

조직 간 비즈니스 문서의 자동화된 교환을 나타냅니다. 자세한 내용은 전자 데이터 교환(EDI)이란 무엇인가요?를 참조하세요.

암호화

사람이 읽을 수 있는 일반 텍스트 데이터를 사이퍼텍스트로 변환하는 컴퓨팅 프로세스입니다.

암호화 키

암호화 알고리즘에 의해 생성되는 무작위 비트의 암호화 문자열입니다. 키의 길이는 다양할 수 있으며 각 키는 예측할 수 없고 고유하게 설계되었습니다.

엔디안

컴퓨터 메모리에 바이트가 저장되는 순서입니다. 빅 엔디안 시스템은 가장 중요한 바이트를 먼저 저장합니다. 리틀 엔디안 시스템은 가장 덜 중요한 바이트를 먼저 저장합니다.

엔드포인트

서비스 엔드포인트를 참조하세요.

엔드포인트 서비스

Virtual Private Cloud(VPC)에서 호스팅하여 다른 사용자와 공유할 수 있는 서비스입니다. 를 사용하여 엔드포인트 서비스를 생성하고 다른 AWS 계정 또는 AWS Identity and Access Management (IAM) 보안 주체에 권한을 AWS PrivateLink 부여할 수 있습니다. 이러한 계정 또는 보안 주체는 인터페이스 VPC 엔드포인트를 생성하여 엔드포인트 서비스에 비공개로 연결할 수 있습니다. 자세한 내용은 Amazon Virtual Private Cloud(VPC) 설명서의 엔드포인트 서비스 생성을 참조하십시오.

엔터프라이즈 리소스 계획(ERP)

엔터프라이즈의 주요 비즈니스 프로세스(예: 회계, MES, 프로젝트 관리)를 자동화하고 관리하는 시스템입니다.

봉투 암호화

암호화 키를 다른 암호화 키로 암호화하는 프로세스입니다. 자세한 내용은 AWS Key Management Service (AWS KMS) 설명서의 봉투 암호화를 참조하세요.

환경

실행 중인 애플리케이션의 인스턴스입니다. 다음은 클라우드 컴퓨팅의 일반적인 환경 유형입니다.

- 개발 환경 - 애플리케이션 유지 관리를 담당하는 핵심 팀만 사용할 수 있는 실행 중인 애플리케이션의 인스턴스입니다. 개발 환경은 변경 사항을 상위 환경으로 승격하기 전에 테스트하는 데 사용됩니다. 이러한 유형의 환경을 테스트 환경이라고도 합니다.
- 하위 환경 - 초기 빌드 및 테스트에 사용되는 환경을 비롯한 애플리케이션의 모든 개발 환경입니다.
- 프로덕션 환경 - 최종 사용자가 액세스할 수 있는 실행 중인 애플리케이션의 인스턴스입니다. CI/CD 파이프라인에서 프로덕션 환경이 마지막 배포 환경입니다.
- 상위 환경 - 핵심 개발 팀 이외의 사용자가 액세스할 수 있는 모든 환경입니다. 프로덕션 환경, 프로덕션 이전 환경 및 사용자 수용 테스트를 위한 환경이 여기에 포함될 수 있습니다.

에픽

애자일 방법론에서 작업을 구성하고 우선순위를 정하는 데 도움이 되는 기능적 범주입니다. 에픽은 요구 사항 및 구현 작업에 대한 개괄적인 설명을 제공합니다. 예를 들어, AWS CAF 보안 에픽에는 ID 및 액세스 관리, 탐지 제어, 인프라 보안, 데이터 보호 및 인시던트 대응이 포함됩니다. AWS 마이그레이션 전략의 에픽에 대한 자세한 내용은 프로그램 구현 가이드를 참조하십시오.

ERP

엔터프라이즈 리소스 계획을 참조하세요.

탐색 데이터 분석(EDA)

데이터 세트를 분석하여 주요 특성을 파악하는 프로세스입니다. 데이터를 수집 또는 집계한 다음 초기 조사를 수행하여 패턴을 찾고, 이상을 탐지하고, 가정을 확인합니다. EDA는 요약 통계를 계산하고 데이터 시각화를 생성하여 수행됩니다.

# F

팩트 테이블

스타 스키마의 중앙 테이블입니다. 비즈니스 운영에 대한 정량적 데이터를 저장합니다. 일반적으로 팩트 테이블은 측정값이 있는 열 및 차원 테이블에 대한 외래 키가 있는 열과 같이 두 가지 열 유형을 포함합니다.

빠른 실패

개발 수명 주기를 줄이기 위해 빈번한 증분 테스트를 사용하는 철학입니다. 애자일 접근 방식의 핵심입니다.

장애 격리 경계

에서 장애의 영향을 제한하고 워크로드의 복원력을 개선하는 데 도움이 되는 가용 영역, AWS 리전 컨트롤 플레인 또는 데이터 플레인과 같은 AWS 클라우드경계입니다. 자세한 내용은 AWS 장애 격리 경계를 참조하세요.

기능 브랜치

브랜치를 참조하세요.

기능

예측에 사용하는 입력 데이터입니다. 예를 들어, 제조 환경에서 기능은 제조 라인에서 주기적으로 캡처되는 이미지일 수 있습니다.

기능 중요도

모델의 예측에 특성이 얼마나 중요한지를 나타냅니다. 이는 일반적으로 SHAP(Shapley Additive Descriptions) 및 통합 그레디언트와 같은 다양한 기법을 통해 계산할 수 있는 수치 점수로 표현됩니다. 자세한 내용은 기계 학습 모델 해석 가능성을 참조하세요 AWS.

기능 변환

추가 소스로 데이터를 보강하거나, 값을 조정하거나, 단일 데이터 필드에서 여러 정보 세트를 추출하는 등 ML 프로세스를 위해 데이터를 최적화하는 것입니다. 이를 통해 ML 모델이 데이터를 활용

할 수 있습니다. 예를 들어, 날짜 '2021-05-27 00:15:37'을 '2021년', '5월', '목', '15일'로 분류하면 학습 알고리즘이 다양한 데이터 구성 요소와 관련된 미묘한 패턴을 학습하는 데 도움이 됩니다.

퓨샷 프롬프팅

유사한 태스크를 수행하도록 요청하기 전에 LLM에 태스크와 원하는 출력을 보여주는 몇 가지 예제를 제공합니다. 이 기법은 모델이 프롬프트에 포함된 예제(샷)에서 학습하는 컨텍스트 내 학습을 적용합니다. 퓨샷 프롬프팅은 특정 형식 지정, 추론 또는 분야별 지식이 필요한 태스크에 효과적일 수 있습니다. 제로샷 프롬프팅도 참조하세요.

FGAC

세분화된 액세스 제어를 참조하세요.

세분화된 액세스 제어(FGAC)

여러 조건을 사용하여 액세스 요청을 허용하거나 거부합니다.

플래시컷 마이그레이션

단계적 접근 방식을 사용하는 대신 변경 데이터 캡처를 통해 지속적 데이터 복제를 사용하여 최단 시간에 데이터를 마이그레이션하는 데이터베이스 마이그레이션 방법입니다. 목표는 가동 중지 시간을 최소화하는 것입니다.

FM

파운데이션 모델을 참조하세요.

파운데이션 모델(FM)

일반화되고 레이블이 지정되지 않은 데이터의 대규모 데이터세트에서 훈련된 대규모 딥 러닝 신경망입니다. FM은 언어 이해, 텍스트 및 이미지 생성, 자연어 대화와 같은 다양한 일반 태스크를 수행할 수 있습니다. 자세한 내용은 파운데이션 모델이란?을 참조하세요.

# G

생성형 AI

대량의 데이터에서 훈련되었으며 간단한 텍스트 프롬프트를 사용하여 이미지, 비디오, 텍스트, 오디오와 같은 새 콘텐츠와 아티팩트를 생성할 수 있는 AI 모델의 하위 세트입니다. 자세한 내용은 생성형 AI란 무엇인가요?를 참조하세요.

지리적 차단

지리적 제한을 참조하세요.

지리적 제한(지리적 차단)

Amazon CloudFront에서 특정 국가의 사용자가 콘텐츠 배포에 액세스하지 못하도록 하는 옵션입니다. 허용 목록 또는 차단 목록을 사용하여 승인된 국가와 차단된 국가를 지정할 수 있습니다. 자세한 내용은 CloudFront 설명서의 콘텐츠의 지리적 배포 제한을 참조하십시오.

Gitflow 워크플로

하위 환경과 상위 환경이 소스 코드 리포지토리의 서로 다른 브랜치를 사용하는 방식입니다. Gitflow 워크플로는 레거시로 간주되며 트렁크 기반 워크플로는 선호되는 현대적 접근 방식입니다.

골든 이미지

시스템 또는 소프트웨어의 새 인스턴스를 배포하기 위한 템플릿으로 사용되는 해당 시스템 또는 소프트웨어의 스냅샷입니다. 예를 들어 제조 분야에서는 골든 이미지를 사용하여 여러 디바이스에서 소프트웨어를 프로비저닝할 수 있으며 이를 통해 딥이스 제조 작업의 속도, 확장성 및 생산성을 개선할 수 있습니다.

브라운필드 전략

새로운 환경에서 기존 인프라의 부재 시스템 아키텍처에 대한 그린필드 전략을 채택할 때 브라운필드라고도 하는 기존 인프라와의 호환성 제한 없이 모든 새로운 기술을 선택할 수 있습니다. 기존 인프라를 확장하는 경우 브라운필드 전략과 그린필드 전략을 혼합할 수 있습니다.

가드레일

조직 단위(OU) 전체에서 리소스, 정책 및 규정 준수를 관리하는 데 도움이 되는 중요 규칙입니다. 예방 가드레일은 규정 준수 표준에 부합하도록 정책을 시행하며, 서비스 제어 정책과 IAM 권한 경계를 사용하여 구현됩니다. 탐지 가드레일은 정책 위반 및 규정 준수 문제를 감지하고 해결을 위한 알림을 생성하며, 이는 AWS Config,, Amazon GuardDuty AWS Security Hub CSPM, , AWS Trusted Advisor Amazon Inspector 및 사용자 지정 AWS Lambda 검사를 사용하여 구현됩니다.

# H

HA

고가용성을 참조하세요.

이기종 데이터베이스 마이그레이션

다른 데이터베이스 엔진을 사용하는 대상 데이터베이스로 소스 데이터베이스 마이그레이션(예: Oracle에서 Amazon Aurora로) 이기종 마이그레이션은 일반적으로 리아키텍트 작업의 일부이며 스

키마를 변환하는 것은 복잡한 작업일 수 있습니다.AWS 는 스키마 변환에 도움이 되는 AWS SCT를 제공합니다.

높은 가용성(HA)

문제나 재해 발생 시 개입 없이 지속적으로 운영할 수 있는 워크로드의 능력. HA 시스템은 자동으로 장애 조치되고, 지속적으로 고품질 성능을 제공하고, 성능에 미치는 영향을 최소화하면서 다양한 부하와 장애를 처리하도록 설계되었습니다.

히스토리언 현대화

제조 산업의 요구 사항을 더 잘 충족하도록 운영 기술(OT) 시스템을 현대화하고 업그레이드하는 데 사용되는 접근 방식입니다. 히스토리언은 공장의 다양한 출처에서 데이터를 수집하고 저장하는 데 사용되는 일종의 데이터베이스입니다.

홀드아웃 데이터

기계 학습 모델을 훈련하는 데 사용되는 데이터세트에서 보류되는 레이블이 지정된 기록 데이터의 일부입니다. 홀드아웃 데이터를 사용하여 모델 예측을 홀드아웃 데이터와 비교해 모델 성능을 평가할 수 있습니다.

동종 데이터베이스 마이그레이션

동일한 데이터베이스 엔진을 공유하는 대상 데이터베이스로 소스 데이터베이스 마이그레이션(예: Microsoft SQL Server에서 Amazon RDS for SQL Server로) 동종 마이그레이션은 일반적으로 리호스팅 또는 리플랫포밍 작업의 일부입니다. 네이티브 데이터베이스 유틸리티를 사용하여 스키마를 마이그레이션할 수 있습니다.

핫 데이터

자주 액세스하는 데이터(예: 실시간 데이터 또는 최근 번역 데이터). 일반적으로 이 데이터에는 빠른 쿼리 응답을 제공하기 위한 고성능 스토리지 계층 또는 클래스가 필요합니다.

핫픽스

프로덕션 환경의 중요한 문제를 해결하기 위한 긴급 수정입니다. 핫픽스는 긴급하기 때문에 일반적인 DevOps 릴리스 워크플로 외부에서 실행됩니다.

하이퍼케어 기간

전환 직후 마이그레이션 팀이 문제를 해결하기 위해 클라우드에서 마이그레이션된 애플리케이션을 관리하고 모니터링하는 기간입니다. 일반적으로 이 기간은 1~4일입니다. 하이퍼케어 기간이 끝나면 마이그레이션 팀은 일반적으로 애플리케이션에 대한 책임을 클라우드 운영 팀에 넘깁니다.

# I

IaC

코드형 인프라를 참조하세요.

자격 증명 기반 정책

AWS 클라우드 환경 내에서 권한을 정의하는 하나 이상의 IAM 보안 주체에 연결된 정책입니다.

유휴 애플리케이션

90일 동안 평균 CPU 및 메모리 사용량이 5~20%인 애플리케이션입니다. 마이그레이션 프로젝트에서는 이러한 애플리케이션을 사용 중지하거나 온프레미스에 유지하는 것이 일반적입니다.

IIoT

산업용 사물 인터넷을 참조하세요.

변경 불가능한 인프라

기존 인프라를 업데이트, 패치 또는 수정하는 대신 프로덕션 워크로드에 대한 새 인프라를 배포하는 모델입니다. 변경 불가능한 인프라는 변경 가능한 인프라보다 본질적으로 더 일관되고 안정적이며 예측 가능합니다. 자세한 내용은 AWS Well-Architected Framework의 변경 불가능한 인프라를 사용하여 배포 모범 사례를 참조하세요.

인바운드(수신) VPC

AWS 다중 계정 아키텍처에서 애플리케이션 외부에서 네트워크 연결을 수락, 검사 및 라우팅하는 VPC입니다. AWS Security Reference Architecture에서는 애플리케이션과 더 넓은 인터넷 간의 양방향 인터페이스를 보호하기 위해 인바운드, 아웃바운드 및 검사 VPC로 네트워크 계정을 설정할 것을 권장합니다.

증분 마이그레이션

한 번에 전체 전환을 수행하는 대신 애플리케이션을 조금씩 마이그레이션하는 전환 전략입니다. 예를 들어, 처음에는 소수의 마이크로서비스나 사용자만 새 시스템으로 이동할 수 있습니다. 모든 것이 제대로 작동하는지 확인한 후에는 레거시 시스템을 폐기할 수 있을 때까지 추가 마이크로서비스 또는 사용자를 점진적으로 이동할 수 있습니다. 이 전략을 사용하면 대규모 마이그레이션과 관련된 위험을 줄일 수 있습니다.

Industry 4.0

연결성, 실시간 데이터, 자동화, 분석 및 AI/ML의 발전을 통해 제조 프로세스의 현대화를 나타내기 위해 2016년에 Klaus Schwab에서 도입한 용어입니다.

인프라

애플리케이션의 환경 내에 포함된 모든 리소스와 자산입니다.

코드형 인프라(IaC)

구성 파일 세트를 통해 애플리케이션의 인프라를 프로비저닝하고 관리하는 프로세스입니다. IaC는 새로운 환경의 반복 가능성, 신뢰성 및 일관성을 위해 인프라 관리를 중앙 집중화하고, 리소스를 표준화하고, 빠르게 확장할 수 있도록 설계되었습니다.

산업용 사물 인터넷(IIoT)

제조, 에너지, 자동차, 의료, 생명과학, 농업 등의 산업 부문에서 인터넷에 연결된 센서 및 디바이스의 사용 자세한 내용은 산업용 사물 인터넷(IoT) 디지털 트랜스포메이션 전략 구축을 참조하십시오.

검사 VPC

AWS 다중 계정 아키텍처에서는 VPC(동일하거나 다른 AWS 리전), 인터넷 및 온프레미스 네트워크 간의 네트워크 트래픽 검사를 관리하는 중앙 집중식 VPCs입니다. AWS Security Reference Architecture에서는 애플리케이션과 더 넓은 인터넷 간의 양방향 인터페이스를 보호하기 위해 인바운드, 아웃바운드 및 검사 VPC로 네트워크 계정을 설정할 것을 권장합니다.

사물 인터넷(IoT)

인터넷이나 로컬 통신 네트워크를 통해 다른 디바이스 및 시스템과 통신하는 센서 또는 프로세서가 내장된 연결된 물리적 객체의 네트워크 자세한 내용은 IoT란?을 참조하십시오.

해석력

모델의 예측이 입력에 따라 어떻게 달라지는지를 사람이 이해할 수 있는 정도를 설명하는 기계 학습 모델의 특성입니다. 자세한 내용은 기계 학습 모델 해석 가능성을 참조하세요 AWS.

IoT

사물 인터넷을 참조하세요.

IT 정보 라이브러리(ITIL)

IT 서비스를 제공하고 이러한 서비스를 비즈니스 요구 사항에 맞게 조정하기 위한 일련의 모범 사례 ITIL은 ITSM의 기반을 제공합니다.

IT 서비스 관리(ITSM)

조직의 IT 서비스 설계, 구현, 관리 및 지원과 관련된 활동 클라우드 운영을 ITSM 도구와 통합하는 방법에 대한 자세한 내용은 운영 통합 가이드를 참조하십시오.

ITIL

IT 정보 라이브러리를 참조하세요.

ITSM

IT 서비스 관리를 참조하세요.

# L

레이블 기반 액세스 제어(LBAC)

사용자 및 데이터 자체에 각각 보안 레이블 값을 명시적으로 할당하는 필수 액세스 제어(MAC)를 구현한 것입니다. 사용자 보안 레이블과 데이터 보안 레이블 간의 교차 부분에 따라 사용자가 볼 수 있는 행과 열이 결정됩니다.

랜딩 존

랜딩 존은 확장 가능하고 안전한 잘 설계된 다중 계정 AWS 환경입니다. 조직은 여기에서부터 보안 및 인프라 환경에 대한 확신을 가지고 워크로드와 애플리케이션을 신속하게 시작하고 배포할 수 있습니다. 랜딩 존에 대한 자세한 내용은 안전하고 확장 가능한 다중 계정 AWS 환경 설정을 참조하십시오.

대규모 언어 모델(LLM)

방대한 양의 데이터에서 사전 훈련된 딥 러닝 AI 모델입니다. LLM은 질문에 대한 답변, 문서 요약, 텍스트를 다른 언어로 번역, 문장 완성과 같은 여러 태스크를 수행할 수 있습니다. 자세한 내용은 대규모 언어 모델(LLM)이란 무엇인가요?를 참조하세요.

대규모 마이그레이션

300대 이상의 서버 마이그레이션입니다.

LBAC

레이블 기반 액세스 제어를 참조하세요.

최소 권한

작업을 수행하는 데 필요한 최소 권한을 부여하는 보안 모범 사례입니다. 자세한 내용은 IAM 설명서의 최소 권한 적용을 참조하십시오.

리프트 앤드 시프트

7R을 참조하세요.

리틀 엔디안 시스템

가장 덜 중요한 바이트를 먼저 저장하는 시스템입니다. 엔디안도 참조하세요.

LLM

대규모 언어 모델을 참조하세요.

하위 환경

환경을 참조하세요.

# M

기계 학습(ML)

패턴 인식 및 학습에 알고리즘과 기법을 사용하는 인공 지능의 한 유형입니다. ML은 사물 인터넷(IoT) 데이터와 같은 기록된 데이터를 분석하고 학습하여 패턴을 기반으로 통계 모델을 생성합니다. 자세한 내용은 기계 학습을 참조하십시오.

기본 브랜치

브랜치를 참조하세요.

맬웨어

컴퓨터 보안 또는 프라이버시를 위협하도록 설계된 소프트웨어입니다. 맬웨어는 컴퓨터 시스템을 방해하거나 민감한 정보를 유출하거나 무단 액세스 권한을 확보할 수 있습니다. 맬웨어의 예로 바이러스, 웜, 랜섬웨어, 트로이 목마, 스파이웨어, 키로거 등이 있습니다.

관리형 서비스

AWS 서비스 는 인프라 계층, 운영 체제 및 플랫폼을 AWS 운영하고, 사용자는 엔드포인트에 액세스하여 데이터를 저장하고 검색합니다. 관리형 서비스의 예로 Amazon Simple Storage Service(Amazon S3) 및 Amazon DynamoDB가 있습니다. 이를 추상화된 서비스라고도 합니다.

제조 실행 시스템(MES)

원자재를 생산 현장에서 완제품으로 변환하는 생산 프로세스를 추적, 모니터링, 문서화 및 제어하기 위한 소프트웨어 시스템입니다.

MAP

Migration Acceleration Program을 참조하세요.

메커니즘

도구를 생성하고 도구 채택을 유도한 다음 조정을 위해 결과를 검사하는 전체 프로세스입니다. 메커니즘은 작동 시 자체적으로 강화하고 개선하는 주기입니다. 자세한 내용은 AWS Well-Architected Framework의 메커니즘 구축을 참조하세요.

멤버 계정

조직의 일부인 관리 계정을 AWS 계정 제외한 모든 계정. AWS Organizations하나의 계정은 한 번에 하나의 조직 멤버만 될 수 있습니다.

MES

제조 실행 시스템을 참조하세요.

메시지 큐 원격 분석 전송(MQTT)

리소스 제약이 있는 IoT 디바이스에 대한 게시 및 구독 패턴을 기반으로 하는 경량 Machine-to-Machine(M2M) 통신 프로토콜입니다.

마이크로서비스

잘 정의된 API를 통해 통신하고 일반적으로 소규모 자체 팀이 소유하는 소규모 독립 서비스입니다. 예를 들어, 보험 시스템에는 영업, 마케팅 등의 비즈니스 역량이나 구매, 청구, 분석 등의 하위 영역에 매핑되는 마이크로 서비스가 포함될 수 있습니다. 마이크로서비스의 이점으로 민첩성, 유연한 확장, 손쉬운 배포, 재사용 가능한 코드, 복원력 등이 있습니다. 자세한 내용은 AWS 서버리스 서비스를 사용하여 마이크로서비스 통합을 참조하세요.

마이크로서비스 아키텍처

각 애플리케이션 프로세스를 마이크로서비스로 실행하는 독립 구성 요소를 사용하여 애플리케이션을 구축하는 접근 방식입니다. 이러한 마이크로서비스는 경량 API를 사용하여 잘 정의된 인터페이스를 통해 통신합니다. 애플리케이션의 특정 기능에 대한 수요에 맞게 이 아키텍처의 각 마이크로서비스를 업데이트, 배포 및 조정할 수 있습니다. 자세한 내용은 에서 마이크로서비스 구현을 참조하세요 AWS.

Migration Acceleration Program(MAP)

조직이 클라우드로 전환하기 위한 강력한 운영 기반을 구축하고 초기 마이그레이션 비용을 상쇄하는 데 도움이 되는 컨설팅 지원, 교육 및 서비스를 제공하는 AWS 프로그램입니다. MAP에는 레거시 마이그레이션을 체계적인 방식으로 실행하기 위한 마이그레이션 방법론과 일반적인 마이그레이션 시나리오를 자동화하고 가속화하는 도구 세트가 포함되어 있습니다.

대규모 마이그레이션

애플리케이션 포트폴리오의 대다수를 웨이브를 통해 클라우드로 이동하는 프로세스로, 각 웨이브에서 더 많은 애플리케이션이 더 빠른 속도로 이동합니다. 이 단계에서는 이전 단계에서 배운 모범 사례와 교훈을 사용하여 팀, 도구 및 프로세스의 마이그레이션 팩토리를 구현하여 자동화 및 민첩한 제공을 통해 워크로드 마이그레이션을 간소화합니다. 이것은 AWS 마이그레이션 전략의 세 번째 단계입니다.

마이그레이션 팩토리

자동화되고 민첩한 접근 방식을 통해 워크로드 마이그레이션을 간소화하는 다기능 팀입니다. 마이그레이션 팩토리 팀에는 일반적으로 스프린트에서 일하는 운영, 비즈니스 분석가 및 소유자, 마이그레이션 엔지니어, 개발자, DevOps 전문가가 포함됩니다. 엔터프라이즈 애플리케이션 포트폴리오의 20~50%는 공장 접근 방식으로 최적화할 수 있는 반복되는 패턴으로 구성되어 있습니다. 자세한 내용은 이 콘텐츠 세트의 클라우드 마이그레이션 팩토리 가이드와 마이그레이션 팩토리에 대한 설명을 참조하십시오.

마이그레이션 메타데이터

마이그레이션을 완료하는 데 필요한 애플리케이션 및 서버에 대한 정보 각 마이그레이션 패턴에는 서로 다른 마이그레이션 메타데이터 세트가 필요합니다. 마이그레이션 메타데이터의 예로는 대상 서브넷, 보안 그룹 및 AWS 계정이 있습니다.

마이그레이션 패턴

사용되는 마이그레이션 전략, 마이그레이션 대상, 마이그레이션 애플리케이션 또는 서비스를 자세히 설명하는 반복 가능한 마이그레이션 작업입니다. 예: AWS Application Migration Service를 사용하여 Amazon EC2로 마이그레이션을 리호스팅합니다.

Migration Portfolio Assessment(MPA)

AWS 클라우드로 마이그레이션하는 비즈니스 사례를 검증하기 위한 정보를 제공하는 온라인 도구입니다. MPA는 상세한 포트폴리오 평가(서버 적정 규모 조정, 가격 책정, TCO 비교, 마이그레이션 비용 분석)와 마이그레이션 계획(애플리케이션 데이터 분석 및 데이터 수집, 애플리케이션 그룹화, 마이그레이션 우선순위 지정, 웨이브 계획)을 제공합니다. MPA 도구(로그인 필요)는 모든 AWS 컨설턴트와 APN 파트너 컨설턴트가 무료로 사용할 수 있습니다.

마이그레이션 준비 상태 평가(MRA)

AWS CAF를 사용하여 조직의 클라우드 준비 상태에 대한 인사이트를 얻고, 강점과 약점을 식별하고, 식별된 격차를 해소하기 위한 행동 계획을 수립하는 프로세스입니다. 자세한 내용은 마이그레이션 준비 가이드를 참조하십시오. MRA는 AWS 마이그레이션 전략의 첫 번째 단계입니다.

마이그레이션 전략

워크로드를 AWS 클라우드로 마이그레이션하는 데 사용되는 접근 방식입니다. 자세한 내용은 이 용어집의 7R 항목과 조직을 동원하여 대규모 마이그레이션 가속화를 참조하세요.

ML

기계 학습을 참조하세요.

현대화

비용을 절감하고 효율성을 높이고 혁신을 활용하기 위해 구식(레거시 또는 모놀리식) 애플리케이션과 해당 인프라를 클라우드의 민첩하고 탄력적이고 가용성이 높은 시스템으로 전환하는 것입니다. 자세한 내용은 AWS 클라우드에서 애플리케이션을 현대화하기 위한 전략을 참조하세요.

현대화 준비 상태 평가

조직 애플리케이션의 현대화 준비 상태를 파악하고, 이점, 위험 및 종속성을 식별하고, 조직이 해당 애플리케이션의 향후 상태를 얼마나 잘 지원할 수 있는지를 확인하는 데 도움이 되는 평가입니다. 평가 결과는 대상 아키텍처의 청사진, 현대화 프로세스의 개발 단계와 마일스톤을 자세히 설명하는 로드맵 및 파악된 격차를 해소하기 위한 실행 계획입니다. 자세한 내용은 AWS 클라우드에서 애플리케이션의 현대화 준비 상태 평가를 참조하세요.

모놀리식 애플리케이션(모놀리식 유형)

긴밀하게 연결된 프로세스를 사용하여 단일 서비스로 실행되는 애플리케이션입니다. 모놀리식 애플리케이션에는 몇 가지 단점이 있습니다. 한 애플리케이션 기능에 대한 수요가 급증하면 전체 아키텍처 규모를 조정해야 합니다. 코드 베이스가 커지면 모놀리식 애플리케이션의 기능을 추가하거나 개선하는 것도 더 복잡해집니다. 이러한 문제를 해결하기 위해 마이크로서비스 아키텍처를 사용할 수 있습니다. 자세한 내용은 마이크로서비스로 모놀리식 유형 분해를 참조하십시오.

MPA

Migration Portfolio Assessment를 참조하세요.

MQTT

메시지 큐 원격 분석 전송을 참조하세요.

멀티클래스 분류

여러 클래스에 대한 예측(2개 이상의 결과 중 하나 예측)을 생성하는 데 도움이 되는 프로세스입니다. 예를 들어, ML 모델이 '이 제품은 책인가요, 자동차인가요, 휴대폰인가요?' 또는 '이 고객이 가장 관심을 갖는 제품 범주는 무엇인가요?'라고 물을 수 있습니다.

변경 가능한 인프라

프로덕션 워크로드에 대한 기존 인프라를 업데이트하고 수정하는 모델입니다. 일관성, 신뢰성 및 예측 가능성을 높이기 위해 AWS Well-Architected Framework에서는 변경 불가능한 인프라를 모범 사례로 사용할 것을 권장합니다.

# O

OAC

오리진 액세스 제어를 참조하세요.

OAI

오리진 액세스 ID를 참조하세요.

OCM

조직 변경 관리를 참조하세요.

오프라인 마이그레이션

마이그레이션 프로세스 중 소스 워크로드가 중단되는 마이그레이션 방법입니다. 이 방법은 가동 중지 증가를 수반하며 일반적으로 작고 중요하지 않은 워크로드에 사용됩니다.

OI

운영 통합을 참조하세요.

OLA

운영 수준 계약을 참조하세요.

온라인 마이그레이션

소스 워크로드를 오프라인 상태로 전환하지 않고 대상 시스템에 복사하는 마이그레이션 방법입니다. 워크로드에 연결된 애플리케이션은 마이그레이션 중에도 계속 작동할 수 있습니다. 이 방법은 가동 중지 차단 또는 최소화를 수반하며 일반적으로 중요한 프로덕션 워크로드에 사용됩니다.

OPC-UA

Open Process Communications - Unified Architecture(OPC-UA)를 참조하세요.

Open Process Communications - Unified Architecture(OPC-UA)

산업 자동화를 위한 Machine-to-Machine(M2M) 통신 프로토콜입니다. OPC-UA는 데이터 암호화, 인증 및 권한 부여 체계에 관한 상호 운용성 표준을 제공합니다.

운영 수준 협약(OLA)

서비스 수준에 관한 계약(SLA)을 지원하기 위해 직무 IT 그룹이 서로에게 제공하기로 약속한 내용을 명확히 하는 계약입니다.

운영 준비 상태 검토(ORR)

인시던트 및 잠재적 장애의 범위를 이해, 평가 또는 예방하거나 줄이는 데 도움이 되는 질문 체크리스트 및 관련 모범 사례입니다. 자세한 내용은 AWS Well-Architected Framework의 운영 준비 상태 검토(ORR)를 참조하세요.

운영 기술(OT)

물리적 환경에서 작동하여 산업 운영, 장비 및 인프라를 제어하는 하드웨어 및 소프트웨어 시스템입니다. 제조 분야에서 OT 및 정보 기술(IT) 시스템의 통합은 Industry 4.0 트랜스포메이션의 주요 중점 사항입니다.

운영 통합(OI)

클라우드에서 운영을 현대화하는 프로세스로 준비 계획, 자동화 및 통합을 수반합니다. 자세한 내용은 운영 통합 가이드를 참조하십시오.

조직 트레일

조직 AWS 계정 내 모든에 대한 모든 이벤트를 로깅 AWS CloudTrail 하는에서 생성된 추적입니다 AWS Organizations. 이 트레일은 조직에 속한 각 AWS 계정 에 생성되고 각 계정의 활동을 추적합니다. 자세한 내용은 CloudTrail 설명서의 Creating a trail for an organization을 참조하십시오.

조직 변경 관리(OCM)

사람, 문화 및 리더십 관점에서 중대하고 파괴적인 비즈니스 혁신을 관리하기 위한 프레임워크입니다. OCM은 변화 채택을 가속화하고, 과도기적 문제를 해결하고, 문화 및 조직적 변화를 주도함으로써 조직이 새로운 시스템 및 전략을 준비하고 전환할 수 있도록 지원합니다. AWS 마이그레이션 전략에서는 클라우드 채택 프로젝트에 필요한 변경 속도 때문에이 프레임워크를 인력 가속화라고 합니다. 자세한 내용은 사용 가이드를 참조하십시오.

오리진 액세스 제어(OAC)

CloudFront에서 Amazon Simple Storage Service(S3) 콘텐츠를 보호하기 위해 액세스를 제한하는 고급 옵션입니다. OAC는 AWS KMS (SSE-KMS)를 사용한 모든 서버 측 암호화 AWS 리전와 S3 버킷에 대한 동적 PUT 및 DELETE 요청에서 모든 S3 버킷을 지원합니다.

오리진 액세스 ID(OAI)

CloudFront에서 Amazon S3 콘텐츠를 보호하기 위해 액세스를 제한하는 옵션입니다. OAI를 사용하면 CloudFront는 Amazon S3가 인증할 수 있는 보안 주체를 생성합니다. 인증된 보안 주체는 특

정 CloudFront 배포를 통해서만 S3 버킷의 콘텐츠에 액세스할 수 있습니다. 더 세분화되고 향상된 액세스 제어를 제공하는 OAC도 참조하십시오.

ORR

운영 준비 상태 검토를 참조하세요.

OT

운영 기술을 참조하세요.

아웃바운드(송신) VPC

AWS 다중 계정 아키텍처에서 애플리케이션 내에서 시작된 네트워크 연결을 처리하는 VPC입니다. AWS Security Reference Architecture에서는 애플리케이션과 더 넓은 인터넷 간의 양방향 인터페이스를 보호하기 위해 인바운드, 아웃바운드 및 검사 VPC로 네트워크 계정을 설정할 것을 권장합니다.

# P

권한 경계

사용자나 역할이 가질 수 있는 최대 권한을 설정하기 위해 IAM 보안 주체에 연결되는 IAM 관리 정책입니다. 자세한 내용은 IAM 설명서의 권한 경계를 참조하십시오.

개인 식별 정보(PII)

직접 보거나 다른 관련 데이터와 함께 짝을 지을 때 개인의 신원을 합리적으로 추론하는 데 사용할 수 있는 정보입니다. PII의 예로는 이름, 주소, 연락처 정보 등이 있습니다.

PII

개인 식별 정보를 참조하세요.

플레이북

클라우드에서 핵심 운영 기능을 제공하는 등 마이그레이션과 관련된 작업을 캡처하는 일련의 사전 정의된 단계입니다. 플레이북은 스크립트, 자동화된 런북 또는 현대화된 환경을 운영하는 데 필요한 프로세스나 단계 요약의 형태를 취할 수 있습니다.

PLC

프로그래밍 가능 로직 컨트롤러를 참조하세요.

PLM

제품 수명 주기 관리를 참조하세요.

정책

권한 정의(ID 기반 정책 참조), 액세스 조건 지정(리소스 기반 정책 참조), AWS Organizations 내 조직의 모든 계정에 대한 최대 권한 정의(서비스 제어 정책 참조)와 같은 작업을 수행할 수 있는 객체입니다.

다국어 지속성

데이터 액세스 패턴 및 기타 요구 사항을 기반으로 독립적으로 마이크로서비스의 데이터 스토리지 기술 선택. 마이크로서비스가 동일한 데이터 스토리지 기술을 사용하는 경우 구현 문제가 발생하거나 성능이 저하될 수 있습니다. 요구 사항에 가장 적합한 데이터 저장소를 사용하면 마이크로서비스를 더 쉽게 구현하고 성능과 확장성을 높일 수 있습니다.

포트폴리오 평가

마이그레이션을 계획하기 위해 애플리케이션 포트폴리오를 검색 및 분석하고 우선순위를 정하는 프로세스입니다. 자세한 내용은 마이그레이션 준비 상태 평가를 참조하십시오.

조건자

보통 `WHERE` 절에 있는 `true` 또는 `false`를 반환하는 쿼리 조건입니다.

푸시다운 조건자

전송 전에 쿼리의 데이터를 필터링하는 데이터베이스 쿼리 최적화 기법입니다. 이렇게 하면 관계형 데이터베이스에서 검색하고 처리해야 하는 데이터의 양이 줄고 쿼리 성능이 향상됩니다.

예방적 제어

이벤트 발생을 방지하도록 설계된 보안 제어입니다. 이 제어는 네트워크에 대한 무단 액세스나 원치 않는 변경을 방지하는 데 도움이 되는 1차 방어선입니다. 자세한 내용은 Implementing security controls on AWS의 Preventative controls를 참조하십시오.

보안 주체

작업을 수행하고 리소스에 액세스할 수 AWS 있는의 엔터티입니다. 이 엔터티는 일반적으로 , AWS 계정 IAM 역할 또는 사용자의 루트 사용자입니다. 자세한 내용은 IAM 설명서의 역할 용어 및 개념의 보안 주체를 참조하십시오.

개인 정보 보호 중심 설계

전체 개발 프로세스에서 개인 정보를 고려하는 시스템 엔지니어링에서의 접근 방식입니다.

프라이빗 호스팅 영역

Amazon Route 53에서 하나 이상의 VPC 내 도메인과 하위 도메인에 대한 DNS 쿼리에 응답하는 방법에 대한 정보가 담긴 컨테이너입니다. 자세한 내용은 Route 53 설명서의 프라이빗 호스팅 영역 작업을 참조하십시오.

선제적 제어

규정 미준수 리소스의 배포를 방지하도록 설계된 보안 제어입니다. 이러한 제어는 리소스를 프로비저닝하기 전에 리소스를 스캔합니다. 리소스가 제어를 준수하지 않으면 프로비저닝되지 않습니다. 자세한 내용은 AWS Control Tower 설명서의 제어 참조 가이드를 참조하고 보안 제어 구현의 사전 예방적 제어를 참조하세요.  AWS

제품 수명 주기 관리(PLM)

설계, 개발 및 출시부터 성장 및 성숙도를 거쳐 거부 및 제거에 이르기까지 전체 수명 주기 동안 제품의 데이터 및 프로세스 관리를 나타냅니다.

프로덕션 환경

환경을 참조하세요.

프로그래밍 가능 로직 컨트롤러(PLC)

제조 분야에서 기계를 모니터링하고 제조 프로세스를 자동화하는 매우 안정적이고 적응력이 뛰어난 컴퓨터입니다.

프롬프트 체이닝

한 LLM 프롬프트의 출력을 다음 프롬프트의 입력으로 사용하여 더 나은 응답을 생성합니다. 이 기법은 복잡한 태스크를 하위 태스크로 나누거나 예비 응답을 반복적으로 세부 조정하거나 확장하는 데 사용됩니다. 이를 통해 모델 응답의 정확성과 관련성을 개선하고 보다 세분화되고 개인화된 결과를 얻을 수 있습니다.

가명화

데이터세트의 개인 식별자를 자리 표시자 값으로 바꾸는 프로세스입니다. 가명화는 개인 정보를 보호하는 데 도움이 될 수 있습니다. 가명화된 데이터는 여전히 개인 데이터로 간주됩니다.

게시/구독(pub/sub)

여러 마이크로서비스에서 비동기 통신을 지원하여 확장성과 응답성을 개선하는 패턴입니다. 예를 들어 마이크로서비스 기반 MES에서 마이크로서비스는 다른 마이크로서비스가 구독할 수 있는 채널에 이벤트 메시지를 게시할 수 있습니다. 시스템은 게시 서비스를 변경하지 않고도 새 마이크로서비스를 추가할 수 있습니다.

# Q

쿼리 계획

SQL 관계형 데이터베이스 시스템의 데이터에 액세스하는 데 사용되는 명령어와 같은 일련의 단계입니다.

쿼리 계획 회귀

데이터베이스 서비스 최적화 프로그램이 데이터베이스 환경을 변경하기 전보다 덜 최적의 계획을 선택하는 경우입니다. 통계, 제한 사항, 환경 설정, 쿼리 파라미터 바인딩 및 데이터베이스 엔진 업데이트의 변경으로 인해 발생할 수 있습니다.

# R

RACI 매트릭스

Responsible, Accountable, Consulted, Informed(RACI)를 참조하세요.

RAG

검색 증강 생성을 참조하세요.

랜섬웨어

결제가 완료될 때까지 컴퓨터 시스템이나 데이터에 대한 액세스를 차단하도록 설계된 악성 소프트웨어입니다.

RASCI 매트릭스

Responsible, Accountable, Consulted, Informed(RACI)를 참조하세요.

RCAC

행 및 열 액세스 제어를 참조하세요.

읽기 전용 복제본

읽기 전용 용도로 사용되는 데이터베이스의 사본입니다. 쿼리를 읽기 전용 복제본으로 라우팅하여 기본 데이터베이스의 로드를 줄일 수 있습니다.

리아키텍팅

7R을 참조하세요.

Recovery Point Objective(RPO)

마지막 데이터 복구 시점 이후 허용되는 최대 시간입니다. 이에 따라 마지막 복구 시점과 서비스 중단 사이에 허용되는 데이터 손실로 간주되는 범위가 결정됩니다.

Recovery Time Objective(RTO)

서비스 중단과 서비스 복원 사이의 허용 가능한 지연 시간입니다.

리팩터링

7R을 참조하세요.

리전

지리적 영역의 AWS 리소스 모음입니다. 각 AWS 리전 는 내결함성, 안정성 및 복원력을 제공하기 위해 서로 격리되고 독립적입니다. 자세한 내용은 계정에서 사용할 수 있는 AWS 리전 지정을 참조하세요.

회귀

숫자 값을 예측하는 ML 기법입니다. 예를 들어, '이 집은 얼마에 팔릴까?'라는 문제를 풀기 위해 ML 모델은 선형 회귀 모델을 사용하여 주택에 대해 알려진 사실(예: 면적)을 기반으로 주택의 매매 가격을 예측할 수 있습니다.

리호스팅

7R을 참조하세요.

릴리스

배포 프로세스에서 변경 사항을 프로덕션 환경으로 승격시키는 행위입니다.

재배치

7R을 참조하세요.

리플랫포밍

7R을 참조하세요.

재구매

7R을 참조하세요.

복원력

중단에 저항하거나 중단을 복구할 수 있는 애플리케이션의 기능입니다. 고가용성 및 재해 복구는 AWS 클라우드에서 복원력을 계획할 때 일반적인 고려 사항입니다. 자세한 내용은 AWS 클라우드 복원력을 참조하세요.

리소스 기반 정책

Amazon S3 버킷, 엔드포인트, 암호화 키 등의 리소스에 연결된 정책입니다. 이 유형의 정책은 액세스가 허용된 보안 주체, 지원되는 작업 및 충족해야 하는 기타 조건을 지정합니다.

RACI(Responsible, Accountable, Consulted, Informed) 매트릭스

마이그레이션 활동 및 클라우드 운영에 참여하는 모든 당사자의 역할과 책임을 정의하는 매트릭스입니다. 매트릭스 이름은 매트릭스에 정의된 책임 유형에서 파생됩니다. 실무 담당자 (R), 의사 결정권자 (A), 업무 수행 조언자 (C), 결과 통보 대상자 (I). 지원자는 (S) 선택사항입니다. 지원자를 포함하면 매트릭스를 RASCI 매트릭스라고 하고, 지원자를 제외하면 RACI 매트릭스라고 합니다.

대응 제어

보안 기준에서 벗어나거나 부정적인 이벤트를 해결하도록 설계된 보안 제어입니다. 자세한 내용은 AWS에서 보안 제어 구현의 대응 제어를 참조하세요.

retain

7R을 참조하세요.

사용 중지

7R을 참조하세요.

검색 증강 세대(RAG)

응답을 생성하기 전에 LLM이 훈련 데이터 소스 외부에 있는 신뢰할 수 있는 데이터 소스를 참조하는 생성형 AI 기술입니다. 예를 들어 RAG 모델은 조직의 지식 기반 또는 사용자 지정 데이터에 대한 시맨틱 검색을 수행할 수 있습니다. 자세한 내용은 검색 증강 생성(RAG)이란 무엇인가요?를 참조하세요.

교체

공격자가 자격 증명에 액세스하는 것을 더욱 어렵게 만들기 위해 보안 암호를 주기적으로 업데이트하는 프로세스입니다.

행 및 열 액세스 제어(RCAC)

액세스 규칙이 정의된 기본적이고 유연한 SQL 표현식을 사용합니다. RCAC는 행 권한과 열 마스크로 구성됩니다.

RPO

목표 복구 시점(RPO)을 참조하세요.

RTO

목표 복구 시간(RTO)을 참조하세요.

런북

특정 작업을 수행하는 데 필요한 일련의 수동 또는 자동 절차입니다. 일반적으로 오류율이 높은 반복 작업이나 절차를 간소화하기 위해 런북을 만듭니다.

# S

SAML 2.0

많은 ID 제공업체(idP)에서 사용하는 개방형 표준입니다. 이 기능을 사용하면 연동 SSO(Single Sign-On)를 AWS Management Console 사용할 수 있으므로 사용자는 조직의 모든 사용자에 대해 IAM에서 사용자를 생성하지 않고도에 로그인하거나 AWS API 작업을 호출할 수 있습니다. SAML 2.0 기반 페더레이션에 대한 자세한 내용은 IAM 설명서의 SAML 2.0 기반 페더레이션 정보를 참조하십시오.

SCADA

감독 제어 및 데이터 획득을 참조하세요.

SCP

서비스 제어 정책을 참조하세요.

보안 암호

에는 암호 또는 사용자 자격 증명과 같이 암호화된 형식으로 저장하는 AWS Secrets Manager기밀 또는 제한된 정보가 있습니다. 보안 암호 값과 메타데이터로 구성됩니다. 보안 암호 값은 바이너리, 단일 문자열 또는 여러 문자열일 수 있습니다. 자세한 내용은 AWS Secrets Manager 설명서의 Secrets Manager 보안 암호란 무엇인가요?를 참조하세요.

보안 중심 설계

전체 개발 프로세스에서 보안을 고려하는 시스템 엔지니어링에서의 접근 방식입니다.

보안 제어

위협 행위자가 보안 취약성을 악용하는 능력을 방지, 탐지 또는 감소시키는 기술적 또는 관리적 가드레일입니다. 보안 제어는 예방, 감지, 대응, 선제적과 같은 기본적인 네 가지 보안 제어 유형으로 구분됩니다.

보안 강화

공격 표면을 줄여 공격에 대한 저항력을 높이는 프로세스입니다. 더 이상 필요하지 않은 리소스 제거, 최소 권한 부여의 보안 모범 사례 구현, 구성 파일의 불필요한 기능 비활성화 등의 작업이 여기에 포함될 수 있습니다.

보안 정보 및 이벤트 관리(SIEM) 시스템

보안 정보 관리(SIM)와 보안 이벤트 관리(SEM) 시스템을 결합하는 도구 및 서비스입니다. SIEM 시스템은 서버, 네트워크, 디바이스 및 기타 소스에서 데이터를 수집, 모니터링 및 분석하여 위협과 보안 침해를 탐지하고 알림을 생성합니다.

보안 응답 자동화

보안 이벤트에 자동으로 응답하거나 이를 해결하도록 설계된 사전 정의되고 프로그래밍된 작업입니다. 이러한 자동화는 보안 모범 사례를 구현하는 데 도움이 되는 탐지 또는 대응 AWS 보안 제어 역할을 합니다. 자동화된 응답 작업의 예로 VPC 보안 그룹 수정, Amazon EC2 인스턴스 패치 적용 또는 자격 증명 교체 등이 있습니다.

서버 측 암호화

대상에서 데이터를 수신하는 AWS 서비스 에 의한 데이터 암호화.

서비스 제어 정책(SCP)

AWS Organizations에 속한 조직의 모든 계정에 대한 권한을 중앙 집중식으로 제어하는 정책입니다. SCP는 관리자가 사용자 또는 역할에 위임할 수 있는 작업에 대해 제한을 설정하거나 가드레일을 정의합니다. SCP를 허용 목록 또는 거부 목록으로 사용하여 허용하거나 금지할 서비스 또는 작업을 지정할 수 있습니다. 자세한 내용은 AWS Organizations 설명서의 서비스 제어 정책을 참조하세요.

서비스 엔드포인트

에 대한 진입점의 URL입니다 AWS 서비스. 엔드포인트를 사용하여 대상 서비스에 프로그래밍 방식으로 연결할 수 있습니다. 자세한 내용은 AWS 일반 참조의 AWS 서비스 엔드포인트를 참조하십시오.

서비스 수준에 관한 계약(SLA)

IT 팀이 고객에게 제공하기로 약속한 내용(예: 서비스 가동 시간 및 성능)을 명시한 계약입니다.

서비스 수준 지표(SLI)

오류 발생률, 가용성 또는 처리량과 같은 서비스의 성능 측면에 대한 측정값입니다.

서비스 수준 목표(SLO)

서비스 수준 지표로 측정되는 서비스의 상태를 나타내는 목표 지표입니다.

공동 책임 모델

클라우드 보안 및 규정 준수를 AWS 위해와 공유하는 책임을 설명하는 모델입니다. AWS 는 클라우드의 보안을 담당하는 반면, 사용자는 클라우드의 보안을 담당합니다. 자세한 내용은 공동 책임 모델을 참조하십시오.

SIEM

보안 정보 및 이벤트 관리 시스템을 참조하세요.

단일 장애점(SPOF)

애플리케이션을 중단시킬 수 있는 애플리케이션의 중요한 단일 구성 요소에서 발생하는 장애입니다.

SLA

서비스 수준 계약을 참조하세요.

SLI

서비스 수준 지표를 참조하세요.

SLO

서비스 수준 목표를 참조하세요.

분할 앤 시드 모델

현대화 프로젝트를 확장하고 가속화하기 위한 패턴입니다. 새로운 기능과 제품 릴리스가 정의되면 핵심 팀이 분할되어 새로운 제품 팀이 만들어집니다. 이를 통해 조직의 역량과 서비스 규모를 조정하고, 개발자 생산성을 개선하고, 신속한 혁신을 지원할 수 있습니다. 자세한 내용은 AWS 클라우드에서 애플리케이션을 현대화하기 위한 단계별 접근 방식을 참조하세요.

SPOF

단일 장애점을 참조하세요.

스타 스키마

하나의 큰 팩트 테이블을 사용하여 트랜잭션 또는 측정된 데이터를 저장하고 하나 이상의 더 작은 차원 테이블을 사용하여 데이터 속성을 저장하는 데이터베이스 조직 구조입니다. 이 구조는 데이터 웨어하우스에서 또는 비즈니스 인텔리전스 목적으로 사용하도록 설계되었습니다.

Strangler Fig 패턴

레거시 시스템을 폐기할 수 있을 때까지 시스템 기능을 점진적으로 다시 작성하고 교체하여 모놀리식 시스템을 현대화하기 위한 접근 방식. 이 패턴은 무화과 덩굴이 나무로 자라 결국 숙주를 압도하고 대체하는 것과 비슷합니다. Martin Fowler가 모놀리식 시스템을 다시 작성할 때 위험을 관리하는 방법으로 이 패턴을 도입했습니다. 이 패턴을 적용하는 방법의 예는 컨테이너 및 Amazon API Gateway를 사용하여 기존의 Microsoft ASP.NET(ASMX) 웹 서비스를 점진적으로 현대화하는 방법을 참조하십시오.

서브넷

VPC의 IP 주소 범위입니다. 서브넷은 단일 가용 영역에 상주해야 합니다.

감독 제어 및 데이터 획득(SCADA)

제조 분야에서 하드웨어와 소프트웨어를 사용하여 물리적 자산과 프로덕션 작업을 모니터링하는 시스템입니다.

대칭 암호화

동일한 키를 사용하여 데이터를 암호화하고 복호화하는 암호화 알고리즘입니다.

합성 테스트

사용자 상호 작용을 시뮬레이션하여 잠재적 문제를 감지하거나 성능을 모니터링하는 방식으로 진행되는 시스템 테스트입니다. Amazon CloudWatch Synthetics를 사용하여 이러한 테스트를 생성할 수 있습니다.

시스템 프롬프트

LLM에 컨텍스트, 명령 또는 지침을 제공하여 동작을 지시하는 기법입니다. 시스템 프롬프트는 컨텍스트를 설정하고 사용자와의 상호 작용을 위한 규칙을 설정하는 데 도움이 됩니다.

# T

tags

AWS 리소스를 구성하기 위한 메타데이터 역할을 하는 키-값 페어입니다. 태그를 사용하면 리소스를 손쉽게 관리, 식별, 정리, 검색, 필터링할 수 있습니다. 자세한 내용은 AWS 리소스에 태그 지정을 참조하십시오.

대상 변수

지도 ML에서 예측하려는 값으로, 결과 변수라고도 합니다. 예를 들어, 제조 설정에서 대상 변수는 제품 결함일 수 있습니다.

작업 목록

런북을 통해 진행 상황을 추적하는 데 사용되는 도구입니다. 작업 목록에는 런북의 개요와 완료해야 할 일반 작업 목록이 포함되어 있습니다. 각 일반 작업에 대한 예상 소요 시간, 소유자 및 진행 상황이 작업 목록에 포함됩니다.

테스트 환경

환경을 참조하세요.

훈련

ML 모델이 학습할 수 있는 데이터를 제공하는 것입니다. 훈련 데이터에는 정답이 포함되어야 합니다. 학습 알고리즘은 훈련 데이터에서 대상(예측하려는 답)에 입력 데이터 속성을 매핑하는 패턴을 찾고, 이러한 패턴을 캡처하는 ML 모델을 출력합니다. 그런 다음 ML 모델을 사용하여 대상을 모르는 새 데이터에 대한 예측을 할 수 있습니다.

Transit Gateway

VPC와 온프레미스 네트워크를 상호 연결하는 데 사용할 수 있는 네트워크 전송 허브입니다. 자세한 내용은 AWS Transit Gateway 설명서의 전송 게이트웨이란 무엇입니까?를 참조하세요.

트렁크 기반 워크플로

개발자가 기능 브랜치에서 로컬로 기능을 구축하고 테스트한 다음 해당 변경 사항을 기본 브랜치에 병합하는 접근 방식입니다. 이후 기본 브랜치는 개발, 프로덕션 이전 및 프로덕션 환경에 순차적으로 구축됩니다.

신뢰할 수 있는 액세스

사용자를 대신하여 AWS Organizations 및 해당 계정에서 조직에서 작업을 수행하도록 지정하는 서비스에 대한 권한 부여. 신뢰할 수 있는 서비스는 필요할 때 각 계정에 서비스 연결 역할을 생성하여 관리 작업을 수행합니다. 자세한 내용은 설명서의 다른 AWS 서비스와 AWS Organizations 함께 사용을 참조하세요 AWS Organizations .

튜닝

ML 모델의 정확도를 높이기 위해 훈련 프로세스의 측면을 여러 변경하는 것입니다. 예를 들어, 레이블링 세트를 생성하고 레이블을 추가한 다음 다양한 설정에서 이러한 단계를 여러 번 반복하여 모델을 최적화하는 방식으로 ML 모델을 훈련할 수 있습니다.

피자 두 판 팀

피자 두 판이면 충분한 소규모 DevOps 팀. 피자 두 판 팀 규모는 소프트웨어 개발에 있어 가능한 최상의 공동 작업 기회를 보장합니다.

# U

불확실성

예측 ML 모델의 신뢰성을 저해할 수 있는 부정확하거나 불완전하거나 알려지지 않은 정보를 나타내는 개념입니다. 불확실성에는 두 가지 유형이 있습니다. 인식론적 불확실성은 제한적이고 불완전한 데이터에 의해 발생하는 반면, 우연한 불확실성은 데이터에 내재된 노이즈와 무작위성에 의해 발생합니다. 자세한 내용은 Quantifying uncertainty in deep learning systems 가이드를 참조하십시오.

차별화되지 않은 작업

애플리케이션을 만들고 운영하는 데 필요하지만 최종 사용자에게 직접적인 가치를 제공하거나 경쟁 우위를 제공하지 못하는 작업을 헤비 리프팅이라고도 합니다. 차별화되지 않은 작업의 예로는 조달, 유지보수, 용량 계획 등이 있습니다.

상위 환경

환경을 참조하세요.

# V

정리

스토리지를 회수하고 성능을 향상시키기 위해 증분 업데이트 후 정리 작업을 수반하는 데이터베이스 유지 관리 작업입니다.

버전 제어

리포지토리의 소스 코드 변경과 같은 변경 사항을 추적하는 프로세스 및 도구입니다.

VPC 피어링

프라이빗 IP 주소를 사용하여 트래픽을 라우팅할 수 있게 하는 두 VPC 간의 연결입니다. 자세한 내용은 Amazon VPC 설명서의 VPC 피어링이란?을 참조하십시오.

취약성

시스템 보안을 손상시키는 소프트웨어 또는 하드웨어 결함입니다.

# W

웜 캐시

자주 액세스하는 최신 관련 데이터를 포함하는 버퍼 캐시입니다. 버퍼 캐시에서 데이터베이스 인스턴스를 읽을 수 있기 때문에 주 메모리나 디스크에서 읽는 것보다 빠릅니다.

웜 데이터

자주 액세스하지 않는 데이터입니다. 이런 종류의 데이터를 쿼리할 때는 일반적으로 적절히 느린 쿼리가 허용됩니다.

창 함수

현재 레코드와 어떤 식으로든 관련된 행 그룹에서 계산을 수행하는 SQL 함수입니다. 창 함수는 이동 평균을 계산하거나 현재 행의 상대적 위치를 기반으로 행 값에 액세스하는 등의 태스크를 처리하는 데 유용합니다.

워크로드

고객 대면 애플리케이션이나 백엔드 프로세스 같이 비즈니스 가치를 창출하는 리소스 및 코드 모음입니다.

워크스트림

마이그레이션 프로젝트에서 특정 작업 세트를 담당하는 직무 그룹입니다. 각 워크스트림은 독립적이지만 프로젝트의 다른 워크스트림을 지원합니다. 예를 들어, 포트폴리오 워크스트림은 애플리케이션 우선순위 지정, 웨이브 계획, 마이그레이션 메타데이터 수집을 담당합니다. 포트폴리오 워크스트림은 이러한 자산을 마이그레이션 워크스트림에 전달하고, 마이그레이션 워크스트림은 서버와 애플리케이션을 마이그레이션합니다.

WORM

Write Once, Read Many(WORM)를 참조하세요.

WQF

AWS Workload Qualification Framework를 참조하세요.

Write Once Read Many(WORM)

데이터를 한 번 쓰고 데이터가 삭제되거나 수정되지 않도록 하는 스토리지 모델입니다. 권한 있는 사용자는 필요한 만큼 여러 번 데이터를 읽을 수 있지만 데이터를 변경할 수는 없습니다. 이 데이터 스토리지 인프라는 변경 불가능한 항목으로 간주됩니다.

# Z

제로데이 익스플로잇

[제로데이 취약성](#)을 악용하는 공격(일반적으로 맬웨어)입니다.

제로데이 취약성

프로덕션 시스템의 명백한 결함 또는 취약성입니다. 위협 행위자는 이러한 유형의 취약성을 사용하여 시스템을 공격할 수 있습니다. 개발자는 공격의 결과로 취약성을 인지하는 경우가 많습니다.

제로샷 프롬프팅

태스크를 수행하기 위해 [LLM](#)에 명령을 제공하지만 안내에 도움이 되는 예제(샷)는 제공하지 않습니다. LLM은 사전 훈련된 지식을 사용하여 태스크를 처리해야 합니다. 제로샷 프롬프팅의 효과는 태스크의 복잡성과 프롬프트의 품질에 따라 달라집니다. [퓨샷 프롬프팅](#)도 참조하세요.

좀비 애플리케이션

평균 CPU 및 메모리 사용량이 5% 미만인 애플리케이션입니다. 마이그레이션 프로젝트에서는 이러한 애플리케이션을 사용 중지하는 것이 일반적입니다.

기계 번역으로 제공되는 번역입니다. 제공된 번역과 원본 영어의 내용이 상충하는 경우에는 영어 버전이 우선합니다.