GenAI 설명서를 위한 BlueXP 워크로드 팩토리

GenAI

NetApp
June 30, 2025

목차

GenAI 설명서를 위한 BlueXP 워크로드 팩토리

릴리스 정보

GenAI를 위한 BlueXP 워크로드 팩토리의 새로운 기능

워크로드 팩토리의 Generative AI 워크로드 기능의 새로운 기능에 대해 알아보십시오.

2025년 6월 29일

일반 **NFS/SMB** 파일 시스템에 호스팅된 데이터 소스 지원

이제 일반 SMB 또는 NFS 공유에서 데이터 소스를 추가할 수 있습니다. 이를 통해 Amazon FSx for NetApp ONTAP 이외의 파일 시스템에서 호스팅되는 볼륨에 저장된 파일을 포함할 수 있습니다.

"지식 기반에 데이터 소스 추가"

"커넥터에 데이터 소스 추가"

2025년 6월 3일

추적기를 사용하여 작업을 모니터링하고 추적할 수 있습니다

GenAI에서 Tracker 모니터링 기능을 사용할 수 있습니다. Tracker를 사용하면 보류 중, 진행 중 및 완료된 작업의 진행 상황과 상태를 모니터링 및 추적하고, 작업 및 하위 작업의 세부 정보를 검토하고, 문제나 오류를 진단하고, 실패한 작업의 매개변수를 편집하고, 실패한 작업을 다시 시도할 수 있습니다.

"BlueXP 워크로드 팩토리의 Tracker를 사용하여 워크로드 작업을 모니터링합니다."

지식 기반에 대한 재순위 모델을 선택하세요

이제 지식 기반에 사용할 특정 리랭커 모델을 선택하여 리랭킹된 쿼리 결과의 관련성을 높일 수 있습니다. GenAI는 Cohere Rerank 및 Amazon Rerank 모델을 지원합니다.

"GenAI 기술 자료를 만듭니다"

2025년 5월 4일

Amazon Q Business용 **NetApp** 커넥터 지원

GenAI의 이번 릴리스에서는 Amazon Q Business용 NetApp Connector에 대한 지원이 도입되어 Amazon Q Business용 커넥터를 생성할 수 있습니다. Amazon Bedrock용 GenAI 지식 기반을 구축하는 것보다 초기 구성이 적은 Amazon Q Business AI Assistant를 빠르고 쉽게 활용할 수 있습니다.

"Amazon Q Business용 NetApp 커넥터 만들기"

향상된 채팅 모델 지원

GenAI는 이제 기술 자료에 대해 다음과 같은 추가 채팅 모델을 지원합니다.

- "미스트랄 AI 모델"

- "Amazon Titan 텍스트 모델"
- "Meta Llama 모델"
- "Jamba 1.5 모델"
- "COHERE 명령 모델"
- "Deepseek 모델"

GenAI는 Amazon Bedrock이 지원하는 각 공급자의 모델을 지원합니다. "아마존 Bedrock에서 기반 모델을 지원했습니다"

"GenAI 기술 자료를 만듭니다"

사용 권한 용어가 업데이트되었습니다

워크로드 팩토리 사용자 인터페이스와 문서에서는 이제 읽기 권한을 나타내는 데 "읽기 전용"을 사용하고 자동화 권한을 나타내는 데 "읽기/쓰기"를 사용합니다. == 2025년 3월 2일

임베디드 챗봇 개선 사항

이제 질문 및 응답을 클립보드에 직접 복사하고 채팅 창의 크기를 조정하며 제목을 변경할 수 있습니다. 또한 채팅 응답에는 이제 표를 포함할 수 있으며, 이 테이블도 복사할 수 있습니다.

"GenAI 기술 자료를 테스트합니다"

채팅 응답 인용 지원

채팅 응답에는 응답을 생성하는 데 사용된 파일 및 데이터 청크가 나열된 인용이 포함됩니다.

"GenAI 기술 자료를 테스트합니다"

향상된 파일 형식 지원

이번 GenAI 릴리스는 향상된 파일 지원을 제공합니다.

- 채팅 모델은 향상된 CSV 지원을 제공합니다. 이렇게 하면 CSV 파일에서 데이터를 쿼리할 때 보다 유용한 응답을 사용할 수 있습니다.
- GenAI는 이제 데이터 소스에서 Apache Parquet 파일을 수집할 수 있습니다.
- GenAI는 이제 이미지가 포함된 Microsoft Word DOCX 파일의 수집을 지원합니다. DOCX 문서에 포함된 이미지가 스캔되고 포함된 이미지의 텍스트 통찰력이 기술 자료 쿼리에 대한 응답에 포함됩니다.

"지원되는 데이터 소스 파일 형식"

2025년 2월 2일

Amazon Nova Foundation 모델 지원

GenAI는 이제 Amazon Nova 기반 모델을 지원합니다. Amazon Nova Micro, Amazon Nova Lite 및 Amazon Nova Pro가 지원됩니다.

"GenAI 요구 사항"

데이터 원본에 대한 파일 형식 필터링

GenAI는 이제 데이터 소스를 추가할 때 데이터 소스 스캔에 포함할 특정 파일 유형을 선택할 수 있도록 지원합니다.

"기술 문서에 데이터 원본을 추가합니다"

데이터 원본에 대한 파일 수정 날짜 필터링

GenAI는 이제 데이터 소스를 추가할 때 수정 날짜별로 데이터 소스 스캔에 포함할 파일 필터링을 지원합니다. 포함된 파일의 수정 날짜 범위를 선택할 수 있습니다.

"기술 문서에 데이터 원본을 추가합니다"

이미지 파일 지원 및 **PDF** 파일 지원 향상

GenAI는 이제 이미지와 그래프 설명뿐만 아니라 문서 텍스트로부터 얻은 통찰력으로 지식 기반 쿼리에 대한 응답성을 향상시켜 보다 풍부하고 높은 품질의 답변을 제공합니다. GenAI는 이제 PDF 파일 내에서 이미지 파일 및 이미지를 스캔할 수 있습니다(다중 모드 파일 지원이라고도 함). 이미지나 PDF 파일을 스캔하도록 선택하면 이미지의 텍스트(PDF 문서에 포함된 이미지 포함)가 데이터 원본으로 스캔되고 스캔의 통찰력은 기술 자료 쿼리에 대한 응답에 포함됩니다.

"기술 문서에 데이터 원본을 추가합니다"

하이브리드 검색 및 리랭크 지원

GenAI는 이제 하이브리드 검색을 사용하여 검색 결과의 관련성과 정확성을 크게 향상시킬 수 있습니다. 하이브리드 검색은 기존의 키워드 기반 검색의 강점과 고급 밀도 벡터 기반 의미 검색 기법을 결합합니다. 표준 키워드 검색 결과는 유사한 일치 및 언어적 뉘앙스로 보강되어 관련성을 향상시킵니다. 그런 다음 GenAI는 COHERE Rerank 및 Amazon Rerank와 같은 고급 재순위 모델을 사용하여 이러한 결과를 더욱 구체화하고 가장 관련성이 높은 결과를 반환합니다. 이 기능은 새로 작성된 기술 자료에 사용할 수 있습니다.

"GenAI를 위한 BlueXP 워크로드 팩토리에 대해 알아보십시오"

2025년 1월 5일

사용자 지정 스냅샷 이름입니다

이제 임시 스냅샷에 대한 스냅샷 이름을 제공할 수 있습니다.

"스냅샷으로 기술 자료 보호"

사용자 지정 **AI** 엔진 인스턴스 이름

이제 구축 중에 AI 엔진 인스턴스에 사용자 지정 이름을 지정할 수 있습니다.

"GenAI 인프라를 구축합니다"

손상되거나 누락된 **GenAI** 인프라를 재구축합니다

AI 엔진 인스턴스가 손상되거나 삭제된 경우 워크로드 공장이 자동으로 리빌드하도록 할 수 있습니다. 워크로드 공장에서 재구축이 완료된 후 지식 베이스를 인프라에 자동으로 다시 연결하여 사용할 수 있습니다.

"문제 해결"

2024년 12월 1일

스냅샷에서 지식 베이스를 복제합니다

GenAI용 BlueXP 워크로드 팩토리에서는 이제 스냅샷으로부터 기술 자료 클론 복제를 지원합니다. 이를 통해 기술 자료를 빠르게 복구하고 기존 데이터 소스를 사용하여 새로운 기술 자료를 작성할 수 있으며 데이터 복구 및 개발에 도움이 됩니다.

"기술 문서를 복제합니다"

온프레미스 **ONTAP** 클러스터 검색 및 복제

온프레미스 ONTAP 클러스터 데이터를 FSx for ONTAP 파일 시스템으로 검색하고 복제하여 AI 지식 기반을 보강하는 데 사용할 수 있습니다. 모든 사내 검색 및 복제 워크플로는 스토리지 인벤토리의 새로운 * 온-프레미스 ONTAP * 탭에서 사용할 수 있습니다.

"사내 ONTAP 클러스터를 검색합니다"

2024년 11월 3일

데이터 가드레일을 사용하여 개인 식별 정보를 마스킹합니다

Generative AI 워크로드에는 BlueXP 분류를 기반으로 하는 데이터 가드레일 기능이 도입되었습니다. 데이터 가드레일 기능은 개인 식별 정보(PII)를 식별 및 마스킹하여 규정 준수를 유지하고 중요한 조직 데이터의 보안을 강화합니다.

"GenAI 기술 자료를 만듭니다"

"BlueXP 분류에 대해 알아보십시오"

2024년 9월 29일

기술 자료 볼륨에 대한 스냅샷 및 복원 지원

이제 기술 자료의 시점 복사본을 만들어 3세대 AI 워크로드 데이터를 보호할 수 있습니다. 이렇게 하면 실수로 데이터가 손실되거나 기술 문서 설정이 변경되는 것을 방지할 수 있습니다. 언제든지 이전 버전의 Knowledge Base 볼륨을 복원할 수 있습니다.

"기술 자료 볼륨의 스냅샷을 생성합니다"

"기술 자료 볼륨의 스냅샷을 복원합니다"

예약된 스캔을 일시 중지합니다

이제 예약된 데이터 원본 스캔을 일시 중지할 수 있습니다. 기본적으로 Generative AI 워크로드는 매일 각 데이터 소스를 스캔하여 각 기술 자료에 새로운 데이터를 수집합니다. 테스트 중 또는 스냅샷 복원 중에 최신 변경 사항을 수집하지 않으려면 예약된 스캔을 일시 중지하고 언제든지 다시 시작할 수 있습니다.

"기술 자료 관리"

이제 기술 자료에 지원되는 데이터 보호 볼륨입니다

이제 기술 자료 볼륨을 선택할 때 NetApp SnapMirror 복제 관계에 포함되는 데이터 보호 볼륨을 선택할 수 있습니다. 따라서 SnapMirror 복제로 이미 보호되는 볼륨에 대한 기술 자료를 저장할 수 있습니다.

"기술 자료에 통합할 데이터 소스를 식별합니다"

2024년 9월 1일

추가 청킹 전략

생성 AI 워크로드는 이제 데이터 소스에 대해 다중 문장 청킹 및 중복 기반 청킹을 지원합니다.

각 기술 자료 전용 볼륨

이제 Generative AI 워크로드가 새로운 각 기술 자료에 대해 전용 Amazon FSx for NetApp ONTAP 볼륨을 생성하여 각 기술 자료에 대한 개별 스냅샷 정책을 지원하고 장애 및 데이터 중독에 대한 보호 기능을 향상합니다.

2024년 8월 4일

Amazon CloudWatch Logs 통합

이제 생성 가능한 AI 워크로드가 Amazon CloudWatch Logs에 통합되어 생성 가능한 AI 워크로드 로그 파일을 모니터링할 수 있습니다.

챗봇 애플리케이션 예

NetApp 워크로드 팩토리 GenAI 샘플 애플리케이션을 사용하면 웹 기반 챗봇 애플리케이션에서 직접 상호 작용하여 게시된 NetApp 워크로드 팩토리 기술 자료에서 인증 및 검색을 테스트할 수 있습니다.

2024년 7월 7일

GenAI를 위한 워크로드 팩토리 최초 릴리즈

초기 릴리스에는 조직의 데이터를 포함하여 사용자 지정된 기술 자료를 개발할 수 있는 기능이 포함되어 있습니다. 사용자를 위한 챗봇 애플리케이션에서 기술 자료에 액세스할 수 있습니다. 이 기능을 통해 조직별 질문에 대한 정확하고 적절한 응답을 제공함으로써 모든 사용자의 만족도와 생산성을 향상시킬 수 있습니다.

GenAI를 위한 BlueXP 워크로드 팩토리에 대해 알아보십시오

GenAI를 위한 BlueXP 워크로드 팩토리에 대해 알아보십시오

GenAI용 BlueXP 워크로드 팩토리를 사용하면 Amazon FSx for NetApp ONTAP 파일 시스템을 GenAI 기반 모델과 통합할 수 있습니다. 이 솔루션은 AI 데이터 세트를 위한 풍부한 보호, 보안 및 비용 최적화 기능을 갖춘 고성능 스토리지를 제공합니다.

GenAI의 BlueXP 워크로드 팩토리는 무엇입니까?

GenAI용 BlueXP 워크로드 팩토리를 사용하면 생성 AI 애플리케이션이 있는 Amazon FSx for NetApp ONTAP에서 엔터프라이즈 데이터 소스를 사용할 수 있습니다. 검색 증강 생성(RAG)을 활용하면 Amazon Bedrock 또는 Amazon Q Business를 통해 사용 가능한 기반 모델에 데이터 소스를 신속하게 연결하여 가상 도우미, Q&A 챗봇, 문서 요약, 콘텐츠 작성 등과 같은 생성 AI 기반 애플리케이션을 개발할 수 있습니다

조직 데이터와 함께 Generative AI를 사용하면 모델이 훈련된 공용 데이터에 기반한 모델의 인텔리전스에만 의존하지 않고 자체 지식과 전문 지식을 활용할 수 있습니다. RAG를 사용하여 모델을 사용자 정의하면 조직별 질문에 정확하고 적절한 답변을 할 수 있어 Generative AI를 사용하여 애플리케이션 사용자의 생산성과 효율성이 향상됩니다.

조직의 데이터에 맞춘 GenAI 애플리케이션을 개발하면 자신의 지식과 전문성을 활용할 수 있습니다. 이 사용자 지정 기능을 통해 조직별 질문에 대한 정확하고 적절한 응답을 보장함으로써 모든 사용자의 만족도와 생산성을 향상시킬 수 있습니다.

"기술 문서를 작성합니다"GenAI는 데이터 소스에서 데이터를 수집하고 벡터화된 결과를 데이터베이스에 저장하며 수집된 데이터를 사용하여 쿼리에 응답하는 방법을 완벽하게 제어할 수 있습니다. 이 방법을 사용하려면 초기 구성이 더 필요하지만 다른 결과에 대해 다른 채팅 모델을 선택할 수 있습니다. "Amazon Q Business용 NetApp 커넥터 정의 "사용자의 경우 데이터 원본의 데이터는 Amazon Q Business에서 수집되어 인덱스에 저장됩니다. 이 방법을 사용하면 초기 구성이 덜 필요하지만 결과를 제어할 수 없습니다.

작업 부하 공장에 대한 자세한 내용은 을 "워크로드 공장 개요"참조하십시오.

GenAI를 사용하여 생성 AI 애플리케이션을 생성할 때의 이점

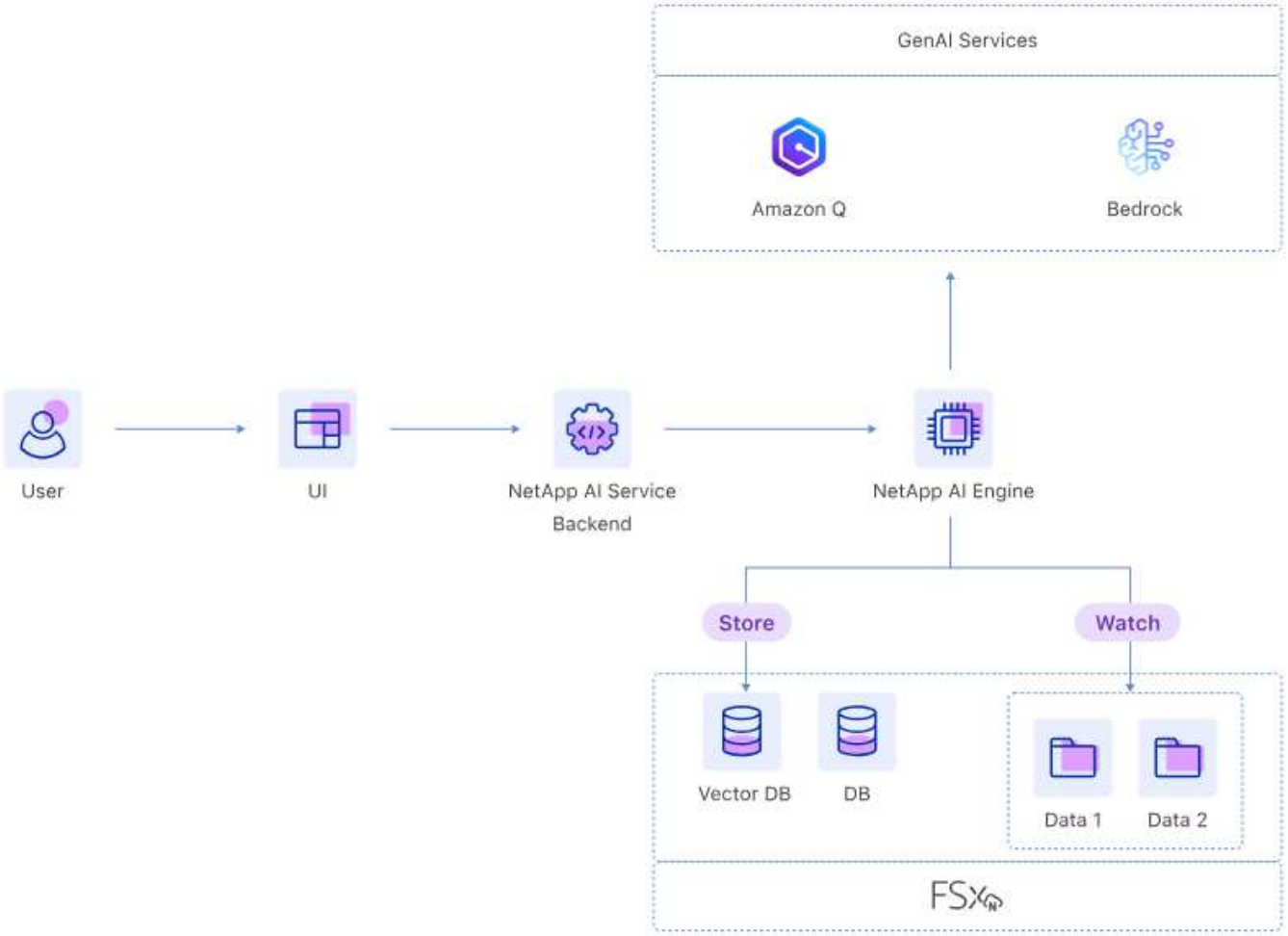
GenAI용 BlueXP 워크로드 팩토리에서는 검색 증강 생성(RAG)을 사용하여 생성 AI 애플리케이션을 구축하는 데 필요한 인프라 배포 프로세스를 간소화합니다. 특히 GenAI는 다음과 같은 이점을 제공합니다.

- 데이터 인프라, 기초 및 언어 모델에 대한 심층적인 지식이 없어도 IT 관리자와 개발자는 GenAI에서 제공하는 자동화를 활용하여 애플리케이션 개발을 가속화할 수 있습니다. 데이터 관리자와 개발자는 생성 AI 애플리케이션에서 사용할 조직의 비정형 데이터를 포함하는 엔터프라이즈 기술 자료를 쉽고 빠르게 생성할 수 있습니다.
- 데이터 보안 및 개인 정보 보호를 위해 기술 자료에 포함된 파일에 사용자 권한을 보존하여 보안을 강화합니다. 챗봇과 같은 애플리케이션은 인증된 사용자에게만 사용자가 액세스할 수 있는 데이터를 기반으로 해답을 제공하기 위해 개발될 수 있습니다.
- 조직 데이터가 외부적으로 노출되지 않는 AWS 고객 계정 내에서 엔터프라이즈 데이터를 안전하게 보호합니다.
- LangChain과 같은 오픈 소스 프레임워크를 사용하여 Q&A 챗봇과 같은 GenAI 애플리케이션의 개발을 가속화하여 GenAI API를 활용하여 지식 기반 및 커넥터를 프로비저닝 및 관리하고, 기술 자료와 채팅하고, 채팅 기록을 저장 및 검색합니다.

- FSx for NetApp ONTAP 파일 시스템에 세대별 AI 데이터 인프라를 구축하고 고가용성, 로컬 데이터 보호 및 복구를 위한 스냅샷, 재해 복구를 위한 SnapMirror, 데이터 인프라 백업을 위한 SnapVault와 같은 ONTAP 기능을 활용하여 데이터 보호 및 가용성 환경을 개선합니다.
- 데이터 중복제거, 압축 및 컴팩션, 데이터 계층화, 씬 프로비저닝과 같은 ONTAP 데이터 효율성 기능을 활용하여 생성 AI 데이터 인프라에 대한 전체 스토리지 비용을 절감합니다.
- GenAI에서 제공하는 하이브리드 검색 및 재순위 기능으로 데이터에서 고품질 결과를 얻을 수 있습니다. 하이브리드 검색과 재순위를 결합하면 검색 결과의 관련성이 크게 향상됩니다. 이러한 기능은 Amazon AWS를 통해 제공되며 지역에 따라 다릅니다.

GenAI의 작동 원리

GenAI는 조직의 개인 데이터를 사용하여 모델의 인텔리전스를 보완하여(훈련된 데이터를 기반으로) 조직 내 사용자가 묻는 질문에 대한 맞춤형 답변을 제공합니다. 먼저 RAG 프레임워크에 필요한 인프라를 배포한 다음 지식 기반을 구축하거나 Amazon Bedrock 또는 Amazon Q Business를 통해 사용할 수 있는 조직의 데이터 소스 및 기반 모델을 사용하여 커넥터를 정의한 다음 애플리케이션(예: Q&A 챗봇)을 기술 자료 또는 커넥터에 연결합니다.



GenAI용 BlueXP 워크로드 공장이 생성 AI 애플리케이션을 구축하는 데 어떤 도움이 되는지 알아보십시오

GenAI는 다음과 같은 방식으로 RAG를 사용하여 생성 AI 애플리케이션을 구축하는 데 도움이 됩니다.

- 검색-증강(RAG) 프레임워크에 필요한 인프라를 구축하고 FSx for ONTAP 파일 시스템 및 Amazon Bedrock 또는

Amazon Q Business에서 데이터 소스를 사용합니다. 인프라에는 데이터 관리를 위한 NetApp GenAI Engine 인스턴스, 내장 벡터 데이터베이스(LanceDB) 및 벡터 데이터베이스용 FSx for ONTAP 파일 시스템의 스토리지가 포함됩니다.

- 데이터 원본을 포함하고 사용자 쿼리에 대한 응답을 검색하기 위해 Amazon Bedrock 또는 Amazon Q Business를 통해 사용할 수 있는 임베디드 및 언어 모델에 데이터 원본을 연결할 수 있습니다. 데이터 소스는 모델 및 해당 구성과 함께 FSx for ONTAP 기술 자료로 제공됩니다.
- 소스 데이터를 기술 자료 또는 커넥터에 수집하여 소스 파일을 FSx for ONTAP 파일 시스템에 포함하고 SMB 공유 내에 파일에 대한 파일 권한은 저장합니다.
- 기술 문서의 내용을 기반으로 대화 시작 질문을 자동으로 작성합니다.
- 데이터 관리자가 기술 자료를 사용하여 채팅을 테스트할 수 있는 채팅 시뮬레이터를 제공합니다.
- 간단한 커넥터 인터페이스를 제공하므로 GenAI를 Amazon Q Business와 연결하고 이 AI 도우미의 기능을 빠르고 쉽게 활용할 수 있습니다.

워크로드 팩토리 사용을 위한 툴

다음 툴과 함께 BlueXP 워크로드 팩토리를 사용할 수 있습니다.

- * 워크로드 팩토리 콘솔 * : 워크로드 팩토리 콘솔은 애플리케이션과 프로젝트에 대한 시각적이고 포괄적인 뷰를 제공합니다.
- * BlueXP 콘솔 *: BlueXP 콘솔은 하이브리드 인터페이스 환경을 제공하므로 다른 BlueXP 서비스와 함께 BlueXP 워크로드 팩토리를 사용할 수 있습니다.
- * 질문하기 *: 질문하기 AI 도우미를 사용하여 워크로드 팩토리 웹 UI를 벗어나지 않고도 워크로드 공장에 대해 자세히 알아보십시오. 워크로드 기본 도움말 메뉴에서 질문하기 에 액세스합니다.
- * CloudShell CLI *: 워크로드 팩토리에는 단일 브라우저 기반 CLI를 통해 모든 계정에서 AWS 및 NetApp 환경을 관리하고 운영할 수 있는 CloudShell CLI가 포함되어 있습니다. 워크로드 팩토리 콘솔의 상단 표시줄에서 CloudShell에 액세스합니다.
- * REST API *: 워크로드 팩토리 REST API를 사용하여 FSx for ONTAP 파일 시스템 및 기타 AWS 리소스를 배포하고 관리합니다.
- * CloudFormation *: AWS CloudFormation 코드를 사용하여 AWS 계정의 CloudFormation 스택에서 AWS 및 타사 리소스를 모델링, 프로비저닝 및 관리하기 위해 워크로드 팩토리 콘솔에 정의한 작업을 수행합니다.
- * Terraform BlueXP 워크로드 팩토리 공급자 *: Terraform을 사용하여 워크로드 팩토리 콘솔에서 생성된 인프라 워크플로우를 구축하고 관리하십시오.

비용

워크로드 팩토리의 GenAI 기능을 사용하는 데 비용이 들지 않습니다.

그러나 생성 AI 인프라를 지원하기 위해 구축하는 AWS 리소스에 대한 비용을 지불해야 합니다. 예를 들어, Amazon Bedrock 또는 Amazon Q Business, FSx for ONTAP 파일 시스템 및 스토리지 용량, GenAI 엔진 EC2 인스턴스의 비용을 지불합니다.

텍스트 정보에 대한 이미지 스캔과 같은 일부 다중 모드 작업은 더 많은 리소스를 사용할 수 있으므로 비용이 더 많이 듭니다. 기술 자료 설정을 변경하는 등의 일부 구성 작업은 데이터 원본을 다시 스캔할 수 있으며 데이터 원본 스캔에 비용이 더 많이 들 수 있습니다.

라이센싱

워크로드 공장의 AI 기능을 사용하기 위해 NetApp에 특별한 라이센스가 필요하지 않습니다.

NetApp GenAI 엔진의 구성 요소

GenAI 인프라를 구축하면 워크로드 팩토리에서는 GenAI 엔진용 EC2 인스턴스를 생성합니다. 또한 이 인스턴스에 대한 IAM 역할, 보안 그룹 및 개인 엔드포인트도 생성합니다. 워크로드 팩토리에서 AWS 환경에서 생성하는 이러한 구성 요소에 대한 자세한 정보를 이해하기를 원할 수 있습니다.

EC2 인스턴스 유형입니다

m5.large를 참조하십시오

IAM 역할

GenAI 엔진 인스턴스에서는 데이터 집합을 Amazon Bedrock의 내장 모델에 전송하고 NetApp AI Service 백엔드와 통신할 수 있는 권한이 필요합니다. IAM 역할에는 다음 권한이 포함됩니다.

IAM 역할 권한

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Action": [
        "ssm:DescribeDocument",
        "ssm:DescribeAssociation",
        "ssm:GetDeployablePatchSnapshotForInstance",
        "ssm:GetManifest",
        "ssm:ListInstanceAssociations",
        "ssm:ListAssociations",
        "ssm:PutInventory",
        "ssm:PutComplianceItems",
        "ssm:PutConfigurePackageResult",
        "ssm:UpdateAssociationStatus",
        "ssm:UpdateInstanceAssociationStatus",
        "ssm:UpdateInstanceInformation",
        "ssmmessages:CreateControlChannel",
        "ssmmessages:CreateDataChannel",
        "ssmmessages:OpenControlChannel",
        "ssmmessages:OpenDataChannel"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ssm:GetParameter"
      ],
      "Resource": "arn:aws:ssm:*:*:parameter/netapp/wlmai/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "fsx:DescribeVolumes",
        "fsx:DescribeStorageVirtualMachines",
        "fsx:DescribeFileSystems"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
```

```
        "fsx:TagResource",
        "fsx:ListTagsForResource"
      ],
      "Resource": [
        "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
        "arn:aws:fsx:*:*:volume/*/*"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "fsx:CreateVolume"
      ],
      "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:storage-virtual-machine/*/*"
      ],
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai: <ai-engine-id>:kbId": "*"
        }
      },
      "Action": "fsx:DeleteVolume",
      "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
      ],
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai: <ai-engine-
id>:qConnectorId": "*"
        }
      },
      "Action": "fsx:DeleteVolume",
      "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
      ],
      "Effect": "Allow"
    },
```

```
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai: <ai-engine-id>": "*"
        }
      },
      "Action": "fsx:UntagResource",
      "Resource": "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai: <ai-engine-id>:kbId": "*"
        }
      },
      "Action": "fsx:UntagResource",
      "Resource": "arn:aws:fsx:*:*:volume/*/*",
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai: <ai-engine-
id>:qConnectorId": "*"
        }
      },
      "Action": "fsx:UntagResource",
      "Resource": "arn:aws:fsx:*:*:volume/*/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "bedrock:InvokeModel",
        "bedrock:Rerank",
        "bedrock:GetFoundationModel",
        "bedrock:GetInferenceProfile"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ec2messages:GetMessages",
        "ec2messages:GetEndpoint",
        "ec2messages:AcknowledgeMessage",
```

```
        "ec2messages:DeleteMessage",
        "ec2messages:FailMessage",
        "ec2messages:SendReply"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "qbusiness:ListWebExperiences",
        "qbusiness:GetApplication",
        "qbusiness:CreateDataSource",
        "qbusiness:DeleteDataSource",
        "qbusiness:ListIndices",
        "qbusiness:StartDataSourceSyncJob",
        "qbusiness:StopDataSourceSyncJob",
        "qbusiness:ListDataSourceSyncJobs",
        "qbusiness:BatchPutDocument",
        "qbusiness:BatchDeleteDocument"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogGroups"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogStreams",
        "logs:PutLogEvents",
        "logs:CreateLogStream",
        "logs:CreateLogGroup"
      ],
      "Resource": [
        "arn:aws:logs:*:*:log-group:/netapp/wlmai/*:log-stream:*",
        "arn:aws:logs:*:*:log-group:/netapp/wlmai/*"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "s3:GetObject",
```

```
            "s3:PutObject"
        ],
        "Resource": "*",
        "Effect": "Allow"
      },
      {
        "Action": [
          "kms:Decrypt",
          "kms:GenerateDataKey"
        ],
        "Resource": "*",
        "Effect": "Allow"
      }
    ]
}
```

보안 그룹

아웃바운드 규칙은 모든 트래픽에 대해 열려 있지만 인바운드 규칙은 완전히 닫힙니다.

전용 끝점

타겟 VPC에 아직 해당 VPC가 없는 경우 워크로드 팩토리는 GenAI 엔진 EC2 인스턴스에 대한 프라이빗 엔드포인트를 생성하여 다음 AWS 서비스와 통신할 수 있도록 합니다.

- 아마존 Bedrock
  - 베드록
  - Bedrock - 런타임
  - Bedrock-에이전트-런타임
- Amazon Elastic Container Registry(ECR)
  - API를 참조하십시오
  - Docker 를 참조하십시오
- AWS 시스템 관리자(SSM)
  - SSM을 참조하십시오
  - ec2messages를 참조하십시오
  - ssmessages(sms 메시지
- NetApp ONTAP용 Amazon FSx
- Amazon CloudWatch 를 참조하십시오

GenAI를 사용하여 Amazon Bedrock에 대한 지식 기반을 구축하십시오

시작하십시오

GenAI 지식 기반을 빠르게 시작합니다

Amazon FSx for NetApp ONTAP 파일 시스템에 있는 조직의 데이터를 사용하여 기술 자료 또는 Amazon Q Business 커넥터를 만들기 시작합니다. 챗봇과 같은 애플리케이션은 이 기술 자료 또는 커넥터에 액세스하여 최종 사용자에게 조직 중심의 응답을 제공합니다.

워크로드 공장 에 로그인합니다

를 사용하여 "콘솔 환경" 로그인해야 "워크로드 팩토리에 계정을 설정합니다" 합니다.

GenAI 요구사항을 충족하도록 환경을 설정합니다

AWS 인프라, 배포되어 검색된 FSx for ONTAP 파일 시스템, 기술 자료 또는 커넥터에 통합하려는 데이터 소스 목록, Amazon Bedrock AI 서비스 또는 Amazon Q Business 애플리케이션에 대한 액세스 등을 구축하려면 AWS 자격 증명이 필요합니다.

"GenAI 요구 사항에 대해 자세히 알아보십시오"..

데이터 소스가 포함된 **FSx for ONTAP** 파일 시스템을 식별합니다

기술 자료에 통합할 데이터 소스는 단일 FSx for ONTAP 파일 시스템 또는 여러 FSx for ONTAP 파일 시스템에 배치할 수 있습니다. 이러한 시스템이 서로 다른 VPC에 있는 경우 동일한 네트워크 내에서 액세스할 수 있거나 VPC를 피어링하여 AI 엔진과 동일한 지역 및 AWS 계정을 사용해야 합니다.

"데이터 소스를 식별하는 방법에 대해 알아봅니다"..

GenAI 인프라를 구축합니다

인프라 배포 마법사를 시작하여 AWS 환경에 GenAI 인프라를 구축합니다. 이 프로세스는 NetApp GenAI 엔진용 EC2 인스턴스와 NetApp AI Engine 데이터베이스를 포함할 FSx for ONTAP 파일 시스템의 볼륨을 배포합니다. 볼륨은 기술 문서에서 사용하는 벡터 데이터베이스를 저장하는 데 사용됩니다.

"기술 자료 인프라를 구축하는 방법에 대해 알아보십시오"..

다음 단계

이제 지식 기반을 구축하여 최종 사용자에게 조직 중심의 응답을 제공할 수 있습니다.

GenAI 기술 자료 요구 사항

기술 자료를 구축하기 전에 워크로드 팩토리 및 AWS가 올바르게 설정되었는지 확인하십시오.

여기에는 AWS 로그인 자격 증명, 기술 자료에 통합하려는 데이터 소스가 포함된 배포된 FSx for ONTAP 파일 시스템, Amazon Bedrock AI 서비스에 대한 액세스 등이 포함됩니다.

기본 **GenAI** 요구 사항

GenAI는 시작하기 전에 환경에 필요한 일반적인 요구 사항을 충족해야 합니다.

워크로드 공장 로그인 및 계정

를 사용하여 "콘솔 환경" 로그인해야 "워크로드 팩토리에 계정을 설정합니다" 합니다.

AWS 자격 증명 및 권한

워크로드 팩토리에 AWS 자격 증명을 읽기/쓰기 권한으로 추가해야 합니다. 즉, GenAI에 대해 워크로드 팩토리를 읽기/쓰기 모드로 사용해야 합니다.

_Basic_mode 및 _Read-only_mode 권한은 현재 지원되지 않습니다.

자격 증명을 설정할 때 아래 표시된 권한을 선택하면 FSx for ONTAP 파일 시스템을 관리하고 기술 자료 및 챗봇에 필요한 GenAI EC2 인스턴스 및 기타 AWS 리소스를 배포 및 관리할 수 있는 모든 권한을 얻을 수 있습니다.

"워크로드 팩토리에 AWS 자격 증명을 추가하는 방법에 대해 알아보십시오"

GenAI 기술 자료 요구 사항

기술 자료를 사용할 계획이라면 환경이 다음 요구 사항을 충족하는지 확인하십시오.

아마존 **Bedrock**

Amazon Bedrock은 기반 모델을 사용할 수 있으며 생성 가능한 AI 애플리케이션을 구축하는 기능을 제공합니다.

GenAI의 BlueXP 워크로드 공장을 시작하기 전에 Amazon Bedrock을 설정해야 합니다. GenAI 배포는 Amazon Bedrock이 지원되는 AWS 지역에 있어야 합니다.

- "AWS 설명서: Amazon Bedrock을 설정합니다"
- "AWS 설명서: Amazon Bedrock에 대한 지식 기반 지원 지역 및 모델"

GenAI는 검색 결과의 관련성을 개선하기 위해 검색 결과의 순위를 기본적으로 재지정합니다. 최상의 결과를 얻으려면 아마존 Bedrock 기반 모델 구성에 COHERE Rerank 또는 Amazon Rerank와 같은 재순위 모델에 대한 액세스가 포함되어 있어야 합니다(해당 지역에서 사용 가능한 경우).

모델 임베드

기술 문서를 만들기 전에 사용하려는 포함 모델을 활성화해야 합니다. 지원되는 임베드 모델은 다음과 같습니다.

- Titan Embeddings G1 - 텍스트
- Titan Embedding Text v2
- 타이탄 다중 모드 포함 G1
- 영어 포함
- 다국어 포함

"Amazon Titan에 대해 자세히 알아보십시오"

채팅 모델

기술 자료를 생성하기 전에 사용하려는 기본 채팅 모델을 활성화해야 합니다. 모델 지원은 AWS 지역에 따라 다르므로, 기술 자료를 배포할 계획이 있는 지역에서 사용할 수 있는 모델을 확인하려면 을 참조하십시오 "AWS 설명서를 참조하십시오" .

GenAI는 Anthropic, Amazon, Mistral AI, Meta, Jamba 및 COHERE의 다양한 모델을 지원합니다.

아마존 Bedrock에서 이러한 모델을 사용하는 방법에 대해 자세히 알아보십시오.

- "아마존 Bedrock에 있는 Anthropic's Claude"
- "아마존 Bedrock 콘솔에서 Amazon Nova를 시작하십시오"
- "미스트랄 AI 모델"
- "Amazon Titan 텍스트 모델"
- "Meta Llama 모델"
- "Jamba 모델"
- "COHERE 명령 모델"

FSx for ONTAP 파일 시스템

FSx for ONTAP 파일 시스템이 하나 이상 필요합니다.

- NetApp GenAI 엔진에서 기술 자료에 사용되는 벡터 데이터베이스를 저장하기 위해 하나의 파일 시스템을 사용(또는 없는 경우 생성)합니다.

  이 FSx for ONTAP 파일 시스템은 FlexVol 볼륨을 사용해야 합니다. FlexGroup 볼륨은 지원되지 않습니다.

- 하나 이상의 파일 시스템에는 기술 자료에 통합할 데이터 원본이 포함됩니다.

  하나의 FSx for ONTAP 파일 시스템을 두 가지 용도로 사용하거나 여러 FSx for ONTAP 파일 시스템을 사용할 수 있습니다.

- AWS FSx for ONTAP 파일 시스템이 상주하는 AWS 지역, VPC 및 서브넷을 알아야 합니다. 파일 시스템은 Amazon Bedrock이 활성화된 AWS 지역에 있어야 합니다.
- 이 배포의 일부인 AWS 리소스에 적용할 태그 키/값 쌍을 고려해야 합니다(선택 사항).
- NetApp AI 엔진 인스턴스에 안전하게 연결할 수 있는 키 쌍 정보를 알아야 합니다.

"FSx for ONTAP 파일 시스템을 구축 및 관리하는 방법에 관해 알아보십시오"

기술 문서나 커넥터에 추가할 데이터 소스 식별

기술 자료에 통합할 FSx for ONTAP 파일 시스템에 있는 문서(데이터 소스)를 식별하거나 생성합니다. 이러한 데이터 원본을 사용하면 기술 문서에서 조직과 관련된 데이터를 기반으로 사용자 쿼리에 대한 정확하고 개인화된 답변을 제공할 수 있습니다.

최대 데이터 원본 수입니다

지원되는 최대 데이터 원본 수는 10개입니다.

데이터 소스의 위치입니다

데이터 소스는 Amazon FSx for NetApp ONTAP 파일 시스템의 SMB 공유 또는 NFS 내보내기의 단일 볼륨 또는 볼륨 내의 폴더에 저장될 수 있습니다. 데이터 소스는 NetApp SnapMirror 데이터 보호 관계에 있는 Amazon FSx for NetApp ONTAP 볼륨에 저장할 수도 있습니다.

볼륨이나 폴더 내에서 개별 문서를 선택할 수 없으므로 데이터 원본이 포함된 각 볼륨이나 폴더에 기술 문서와 통합되지 않아야 하는 관련 문서가 포함되어 있지 않은지 확인해야 합니다.

각 기술 자료에 여러 데이터 소스를 추가할 수 있지만, 모두 AWS 계정에서 액세스할 수 있는 FSx for ONTAP 파일 시스템에 상주해야 합니다.

각 데이터 원본의 최대 파일 크기는 50MB입니다.

지원되는 프로토콜

Knowledge Base는 NFS 또는 SMB/CIFS 프로토콜을 사용하는 볼륨의 데이터를 지원합니다. SMB 프로토콜을 사용하여 저장된 파일을 선택할 때 기술 문서에서 해당 볼륨의 파일에 액세스할 수 있도록 Active Directory 정보를 입력해야 합니다. 여기에는 Active Directory 도메인, IP 주소, 사용자 이름 및 암호가 포함됩니다.

SMB를 통해 액세스되는 공유(파일 또는 디렉토리)에 데이터 소스를 저장하는 경우, 해당 공유에 액세스할 수 있는 권한이 있는 챗봇 사용자 또는 그룹만 데이터에 액세스할 수 있습니다. 이 "권한 인식 기능"이 활성화되면 AI 시스템은 auth0의 사용자 이메일을 SMB 공유에서 파일을 보거나 사용할 수 있는 사용자와 비교합니다. 챗봇은 포함된 파일에 대한 사용자 권한을 기반으로 해답을 제공합니다.

예를 들어, 기술 자료에 10개의 파일(데이터 소스)을 통합했고 이 중 2개의 파일이 제한된 정보를 포함하는 인적 리소스 파일인 경우, 이 두 파일에 액세스할 수 있도록 인증된 챗봇 사용자만 이러한 파일의 데이터를 포함하는 응답을 받게 됩니다.

지원되는 데이터 소스 파일 형식

현재 워크로드 공장 GenAI 기술 자료에서 지원되는 데이터 소스 파일 형식은 다음과 같습니다.

파일 형식	연장
아파치 Parquetfootnote: disclaimer [구조화된 데이터 파일을 지식베이스에 인제스트할 때 데이터 가드레일 기능은 지원되지 않는다.]	마루
쉼표로 구분된 값 파일:면책 사항 [ ]	.csv입니다
Graphics Interchange Format(그래픽 교환 형식)	gif
JPEG를 선택합니다	.jpg or.jpeg 을 참조하십시오
JSON 및 JSONP각주: 면책 사항 [ ]	제이슨
마크다운	진료 부서
Microsoft Word를	.doc 또는 .docx
일반 텍스트	.txt입니다
휴대용 문서 형식	PDF로 이동합니다
휴대용 네트워크 그래픽	png입니다
WebP 이미지	웹

GenAI 인프라를 구축합니다

조직을 위해 FSx for ONTAP 지식 베이스, 커넥터 및 애플리케이션을 구축하기 전에 환경에 RAG 프레임워크용 GenAI 인프라를 구축해야 합니다. 기본 인프라 구성요소는 Amazon Bedrock 서비스, NetApp GenAI 엔진의 가상 머신 인스턴스 및 FSx for ONTAP 파일 시스템입니다.

구축된 인프라는 여러 지식 베이스, 챗봇, 커넥터를 지원할 수 있으므로 일반적으로 이 작업은 한 번만 수행하면 됩니다.

인프라 세부 정보

GenAI 배포는 Amazon Bedrock이 지원되는 AWS 지역에 있어야 합니다. "지원되는 지역 목록을 봅니다"

인프라는 다음과 같은 구성 요소로 이루어집니다.

아마존 **Bedrock** 서비스

Amazon Bedrock은 단일 API를 통해 선도적인 AI 회사의 기반 모델(FMS)을 사용할 수 있는 완전 관리형 서비스입니다. 또한 안전한 생성 AI 애플리케이션을 구축하는 데 필요한 기능을 제공합니다.

"아마존 Bedrock에 대해 자세히 알아보십시오"

아마존 **Q** 비즈니스

Amazon Q는 Amazon Bedrock을 기반으로 구축되어 질문에 답하고 데이터 소스의 정보를 기반으로 콘텐츠를 생성하는 데 사용할 수 있는 완전 관리형 생성 AI 도우미를 제공합니다.

"아마존 Q 비즈니스에 대해 자세히 알아보십시오"

NetApp GenAI 엔진용 가상 머신입니다

이 프로세스 중에 NetApp GenAI 엔진이 배포됩니다. 데이터 소스에서 데이터를 수집한 다음 해당 데이터를 벡터 데이터베이스에 쓸 수 있는 처리 능력을 제공합니다.

FSx for ONTAP 파일 시스템

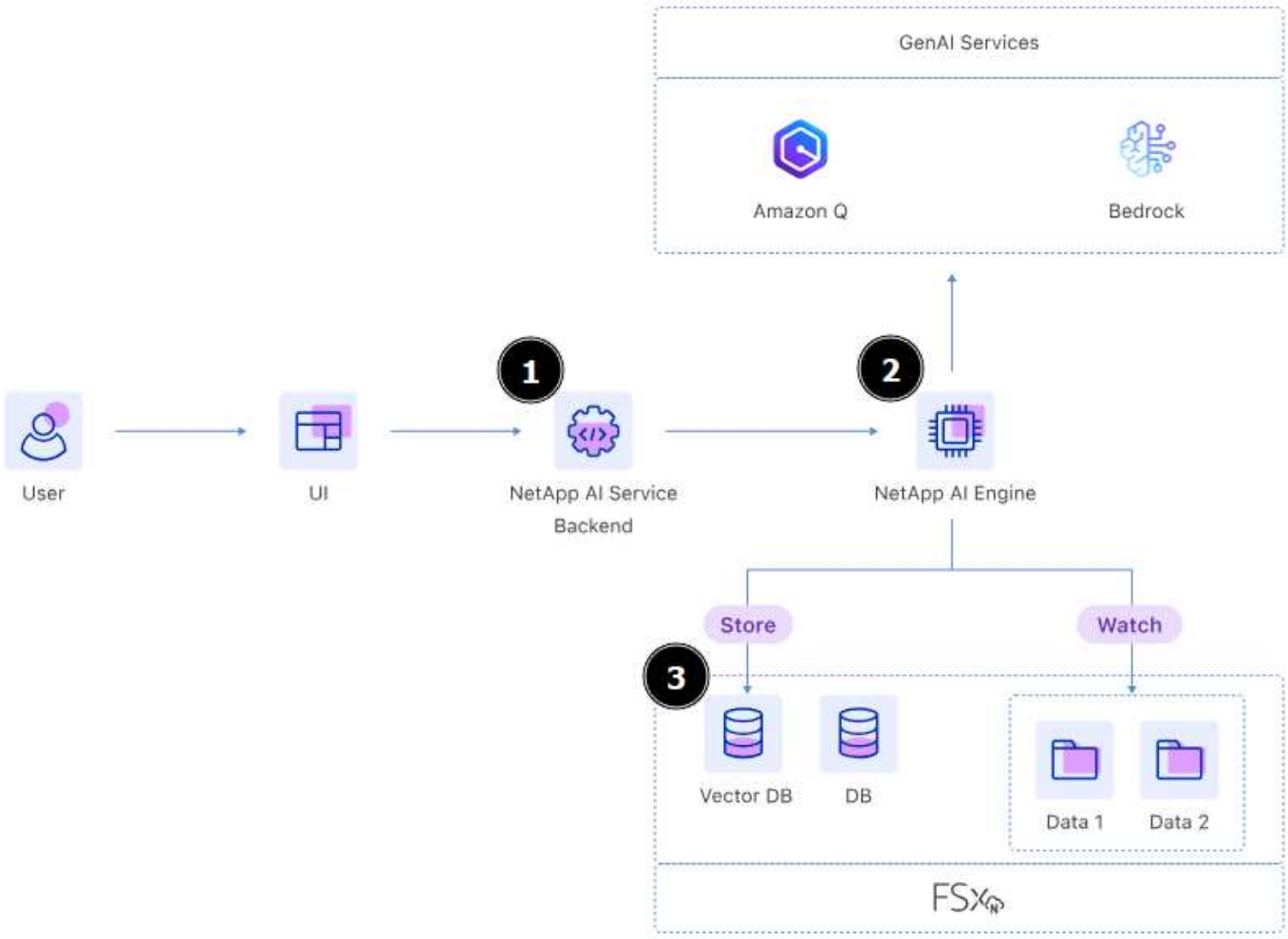
FSx for ONTAP 파일 시스템은 GenAI 시스템을 위한 스토리지를 제공합니다.

데이터 소스를 기반으로 기본 모델에 의해 생성된 데이터를 저장하는 벡터 데이터베이스를 포함하는 단일 볼륨이 배포됩니다.

기술 자료에 통합할 데이터 소스는 동일한 FSx for ONTAP 파일 시스템 또는 다른 시스템에 상주할 수 있습니다.

NetApp GenAI 엔진은 이 두 볼륨을 모두 모니터링하고 상호 작용합니다.

다음 이미지는 GenAI 인프라를 보여 줍니다. 이 절차를 수행하는 동안 번호가 1, 2, 3인 구성 요소가 전개됩니다. 배포를 시작하기 전에 다른 요소가 있어야 합니다.



GenAI 인프라를 구축합니다

AWS 자격 증명을 입력하고 FSx for ONTAP 파일 시스템을 선택하여 검색 증강 생성(RAG) 인프라를 배포해야 합니다.

시작하기 전에

이 절차를 시작하기 전에 사용자의 환경이 지식 베이스 또는 커넥터의 요구 사항을 충족하는지 확인하십시오.

- "기술 자료 요구 사항"
- "커넥터 요구 사항"

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. AI 워크로드 타일에서 * 배포 및 관리 * 를 선택합니다.
3. 인프라 다이어그램을 검토하고 * Next * 를 선택합니다.
4. AWS 설정 * 섹션의 항목을 완료합니다.
   a. * AWS 자격 증명 *: AWS 리소스 배포 권한을 제공하는 AWS 자격 증명을 선택하거나 추가합니다.
   b. * 위치 *: AWS 지역, VPC 및 서브넷을 선택합니다.

      GenAI 배포는 Amazon Bedrock이 활성화된 AWS 지역에 있어야 합니다. "지원되는 지역 목록을 봅니다"

5. 인프라 설정 * 섹션의 항목을 완료합니다.
   a. * 태그 * : 이 배포의 일부인 모든 AWS 리소스에 적용할 태그 키/값 쌍을 입력하십시오. 이러한 태그는 AWS 관리 콘솔 및 워크로드 공장 내의 인프라 정보 영역에 표시되며 워크로드 공장 리소스를 추적하는 데 도움이 됩니다.
6. Connectivity * 섹션을 완료합니다.
   a. * 키 쌍 * : NetApp GenAI 엔진 인스턴스에 안전하게 연결할 수 있는 키 쌍을 선택하십시오.
7. AI 엔진 * 섹션을 완료하십시오.
   a. * 인스턴스 이름 *: 필요에 따라 * 인스턴스 이름 정의 * 를 선택하고 AI 엔진 인스턴스의 사용자 정의 이름을 입력합니다. 인스턴스 이름은 AWS 관리 콘솔 및 워크로드 공장 내의 인프라 정보 영역에 표시되며, 작업 부하 공장 리소스를 추적하는 데 도움이 됩니다.
8. 배포 * 를 선택하여 배포를 시작합니다.

자격 증명 오류로 인해 배포가 실패하는 경우 오류 메시지 내에서 하이퍼링크를 선택하여 오류 세부 정보를 확인할 수 있습니다. 누락 또는 차단된 권한 목록과 GenAI 워크로드를 배포하기 위해 필요한 권한 목록을 확인할 수 있습니다.

결과

워크로드 팩토리가 챗봇 인프라 구축을 시작합니다. 이 프로세스는 최대 10분 정도 소요될 수 있습니다.

배포 프로세스 중에 다음 항목이 설정됩니다.

- 네트워크는 전용 끝점과 함께 설정됩니다.
- IAM 역할, 인스턴스 프로필 및 보안 그룹이 생성됩니다.
- GenAI 엔진의 가상 머신 인스턴스가 배포됩니다.
- Amazon Bedrock은 접두사가 있는 로그 그룹을 사용하여 Amazon CloudWatch 로그에 로그를 보내도록 구성되어 `/aws/bedrock/`있습니다.
- GenAI 엔진은 이름이 지정된 로그 그룹을 사용하여 Amazon CloudWatch 로그에 로그를 전송하도록 구성되어 `/netapp/wlmai/<tenancyAccountId>/randomId`있습니다. 여기서 는 현재 사용자에 대한 입니다. `<tenancyAccountID>` "BlueXP 계정 ID입니다"

GenAI 기술 자료를 만듭니다

AI 인프라를 구축하고 FSx for ONTAP 데이터 저장소의 기술 자료에 통합할 데이터 소스를 식별한 후에는 워크로드 팩토리를 사용하여 기술 자료를 구축할 준비가 된 것입니다. 이 단계의 일환으로 AI 특성을 정의하고 대화를 시작하는 방안을 마련합니다.

계속하기 전에 귀사의 환경이 기술 자료에 대한 을 충족하는지 "요구 사항"확인하십시오.

이 작업에 대해

Knowledge Base에는 두 가지 데이터 통합 Modality(_public mode_와_Enterprise mode_가 있습니다.

공개 모드

조직의 데이터 원본을 통합하지 않고도 기술 문서를 사용할 수 있습니다. 이 경우 기술 문서에 통합된 응용 프로그램은 인터넷에 공개된 정보의 결과만 제공합니다. 이를 _public mode_integration 이라고 합니다.

엔터프라이즈 모드

대부분의 경우 조직의 데이터 원본을 기술 자료에 통합할 수 있습니다. 이를 엔터프라이즈 모드_통합이라고 하며, 이는 기업으로부터 지식을 제공하기 때문입니다.

조직의 데이터 원본에는 PII(개인 식별 정보)가 포함될 수 있습니다. 이러한 중요한 정보를 보호하기 위해 기술 자료를 만들고 구성할 때 _data guardrails_를 활성화할 수 있습니다. BlueXP 분류를 기반으로 하는 데이터 가드레일은 PII를 식별하고 마스킹하므로 액세스 및 복구가 불가능합니다.

"BlueXP 분류에 대해 알아보십시오"..

GenAI용 BlueXP 워크로드 공장은 민감한 개인 정보를 마스킹하지 않습니다(SPii). 이 데이터 유형에 대한 자세한 내용은 을 "중요한 개인 데이터의 유형"참조하십시오.

데이터 가드레일은 언제든지 활성화 또는 비활성화할 수 있습니다. 데이터 가드레일 사용을 전환하면 워크로드 공장에서 전체 기술 자료를 처음부터 검사하여 비용이 발생합니다.

기술 문서를 만들고 구성합니다

기술 자료에서는 지식 기반을 생성하는 데 사용할 Bedrock AI 모델 및 내장 형식과 같은 특성을 정의합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. AI 워크로드 타일에서 * 배포 및 관리 * 를 선택합니다.
3. Knowledge Base & Connectors 탭에서 * Create New * 드롭다운을 선택하고 * Bedrock용 * NetApp GenAI Knowledge Base * 를 선택합니다.
4. 기술 문서 정의 페이지에서 기술 문서 설정을 구성합니다.
   a. * 이름 *: 기술 문서에 사용할 이름을 입력하십시오.
   b. **Description**: 기술 문서에 대한 자세한 설명을 입력합니다.
   c. * 포함 모델 * : 포함 모델은 귀하의 데이터가 기술 자료를 위한 벡터 임베딩으로 변환되는 방법을 정의합니다. 워크로드 팩토리에서는 다음 모델을 지원합니다.

      - Titan Embeddings G1 - 텍스트
      - Titan Embedding Text v2
      - 타이탄 다중 모드 포함 G1
      - 영어 포함
      - 다국어 포함

      아마존 Bedrock에서 임베딩 모델을 이미 활성화했어야 합니다.

      "Amazon Titan에 대해 자세히 알아보십시오"

   d. * 채팅 모델 * : 아마존 Bedrock에 통합된 다양한 채팅 모델 중에서 선택하십시오. 아마존 Bedrock에서 채팅 모델을 이미 활성화했어야 합니다.
   e. 재순위: 쿼리 결과의 관련성과 품질을 향상시킬 수 있는 재순위 기능을 활성화 또는 비활성화합니다. 재순위 기능에 사용할 표준 채팅 모델 또는 특수 재순위 모델을 선택하세요. 재순위 모델 옵션은 해당 지역에서 사용

가능한 경우에만 표시됩니다.

f. * 데이터 가드레일 * : 데이터 가드레일을 활성화 또는 비활성화할지 선택합니다. "BlueXP 분류를 기반으로 하는 데이터 가드레일에 대해 알아보십시오"..

데이터 가드레일을 사용하려면 다음 사전 요구 사항을 충족해야 합니다.

- BlueXP 분류와 통신하려면 서비스 계정이 필요합니다. 서비스 계정을 생성하려면 BlueXP Tenancy 계정에 _Organization admin_역할이 있어야 합니다. 조직 관리자 역할을 가진 구성원은 BlueXP 의 모든 작업을 완료할 수 있습니다. "BlueXP 에서 구성원에 역할을 추가하는 방법에 대해 알아봅니다"
- AI 엔진이 에 액세스할 수 있어야 "BlueXP API 끝점입니다"합니다.
- 에 설명된 대로 다음을 수행해야 "BlueXP 분류 문서"합니다.

A. BlueXP Connector를 생성합니다

B. 환경이 사전 요구 사항을 충족할 수 있는지 확인합니다

C. BlueXP 분류를 배포합니다

CSV, JSON, JSONP 또는 Parquet과 같은 구조화된 데이터 파일을 수집할 때는 데이터 가드레일 기능이 지원되지 않습니다.

g. * Conversation Starters *: 이 기술 자료를 사용하는 챗봇과 상호 작용하는 사용자에게 표시되는 최대 4개의 대화 시작 프롬프트를 제공할지 여부를 선택합니다. 이 설정을 사용하는 것이 좋습니다.

대화 시작점을 활성화하면 기본적으로 "자동 모드"가 선택됩니다. "수동 모드"는 기술 문서에 데이터 원본을 추가한 후에만 활성화할 수 있습니다. "기술 문서 설정을 수정하는 방법에 대해 알아봅니다"..

h. * FSx for ONTAP 파일 시스템 *: 새로운 기술 자료를 정의하면 워크로드 공장에서 이를 저장할 새로운 Amazon FSx for NetApp ONTAP 볼륨을 생성합니다. 새 볼륨을 생성할 기존 파일 시스템 이름과 SVM(스토리지 VM이라고도 함)을 선택합니다.

i. * Snapshot policy *: 워크로드 공장 저장소 인벤토리에 정의된 기존 정책 목록에서 스냅샷 정책을 선택합니다. 선택한 스냅샷 정책에 따라 기술 자료의 반복적인 스냅샷이 자동으로 생성됩니다.

필요한 스냅샷 정책이 없는 경우 "스냅샷 정책을 생성합니다" 볼륨이 포함된 스토리지 VM에서 수행할 수 있습니다.

5. GenAI에 기술 문서를 추가하려면 * 기술 문서 만들기 * 를 선택하십시오.

기술 문서가 작성되는 동안 진행률 표시기가 나타납니다.

기술 문서를 만든 후에는 데이터 원본을 새 기술 문서에 추가하거나 데이터 원본을 추가하지 않고 프로세스를 종료할 수 있습니다. 지금 * 데이터 원본 추가 * 를 선택하고 하나 이상의 데이터 원본을 추가하는 것이 좋습니다.

기술 문서에 데이터 원본을 추가합니다

하나 이상의 데이터 원본을 추가하여 조직의 데이터로 기술 문서를 채울 수 있습니다.

이 작업에 대해

지원되는 최대 데이터 원본 수는 10개입니다.

단계

1. *데이터 소스 추가*를 선택한 후 추가하려는 데이터 소스 유형을 선택합니다.
    - ONTAP 파일 시스템용 FSx 추가(기존 ONTAP 볼륨용 FSx의 파일 사용)
    - 파일 시스템 추가(일반 SMB 또는 NFS 공유의 파일 사용)

ONTAP 파일 시스템에 **FSx** 추가

1. * 파일 시스템 선택 *: 데이터 소스 파일이 있는 FSx for ONTAP 파일 시스템을 선택하고 * 다음 * 을 선택합니다.
2. * 볼륨 선택 *: 데이터 원본 파일이 있는 볼륨을 선택하고 * 다음 * 을 선택합니다.

   SMB 프로토콜을 사용하여 저장된 파일을 선택할 때 도메인, IP 주소, 사용자 이름 및 암호를 포함한 Active Directory 정보를 입력해야 합니다.

3. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.
4. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.

   ◦ * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.

   ▪ * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.

   ▪ * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

   높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   ◦ * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.

   ▪ 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

   이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

   이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든 PII가 GenAI 데이터베이스에 마스킹됩니다.

   이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

   ◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

   수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

5. 선택한 데이터 원본이 SMB 프로토콜을 사용하는 볼륨에 있을 때만 사용할 수 있는 * 권한 인식 * 섹션에서 권한 인식 응답을 활성화하거나 비활성화할 수 있습니다.

- 사용: 이 기술 자료에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.
- * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.

6. 이 데이터 소스를 기술 문서에 추가하려면 * 추가 * 를 선택하십시오.

일반 **NFS** 파일 시스템 추가

1. 파일 시스템 선택: 데이터 소스 파일이 있는 파일 시스템 호스트의 IP 주소 또는 FQDN을 입력하고, 네트워크 공유에 대한 NFS 프로토콜을 선택하고 *다음*을 선택합니다.
2. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.

> 경우에 따라 NFS 내보내기 이름을 직접 입력하고 *디렉터리 검색*을 선택하여 사용 가능한 디렉터리를 표시해야 할 수도 있습니다. 내보내기 전체 또는 내보내기에서 특정 폴더만 선택할 수 있습니다.

3. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.
   - * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.
     - * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.
     - * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

     > 높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   - * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.
     - 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

       이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

     이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든 PII가 GenAI 데이터베이스에 마스킹됩니다.

     > 이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

   - 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

4. *데이터 소스 추가*를 선택하여 이 데이터 소스를 지식 기반에 추가하세요.

일반 **SMB** 파일 시스템 추가

1. 파일 시스템 선택:
   a. 데이터 소스 파일이 있는 파일 시스템 호스트의 IP 주소나 FQDN을 입력하세요.
   b. 네트워크 공유에 SMB 프로토콜을 선택합니다.
   c. 도메인, IP 주소, 사용자 이름, 비밀번호 등 Active Directory 정보를 입력합니다.
   d. 다음 * 을 선택합니다.
2. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.

   경우에 따라 SMB 공유 이름을 직접 입력하고 *디렉터리 검색*을 선택하여 사용 가능한 디렉터리를 표시해야 할 수도 있습니다. 전체 공유를 선택하거나 공유에서 특정 폴더만 선택할 수 있습니다.

3. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.
   - * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.
     - * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.
     - * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

       높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   - 권한 인식: 권한 인식 응답을 활성화하거나 비활성화합니다.
     - 사용: 이 기술 자료에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.
     - * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.
   - * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.
     - 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

       이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

   이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든

PII가 GenAI 데이터베이스에 마스킹됩니다.

> 이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

> 수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

4. *데이터 소스 추가*를 선택하여 이 데이터 소스를 지식 기반에 추가하세요.

결과

데이터 원본이 기술 문서에 포함되기 시작합니다. 데이터 원본이 완전히 포함되면 상태가 "포함"에서 "포함"으로 변경됩니다.

기술 자료에 단일 데이터 소스를 추가한 후 챗봇 시뮬레이터 창에서 로컬로 테스트하고 필요에 따라 변경한 후 사용자가 챗봇을 사용할 수 있도록 할 수 있습니다. 또한 동일한 단계에 따라 기술 문서에 데이터 원본을 추가할 수도 있습니다.

GenAI 기술 자료를 테스트합니다

기술 자료를 만든 후에는 챗봇 시뮬레이터를 사용하여 로컬에서 테스트하고 필요에 따라 변경한 후 챗봇 애플리케이션을 통해 사용자에게 기술 자료를 제공할 수 있습니다.

이 작업에 대해

기술 자료를 테스트하여 예상대로 작동하는지 확인하고, 이 기술 자료의 챗봇 사용자가 기본적으로 사용할 수 있도록 대화 시작을 사용자 지정할 수 있습니다. 챗봇 시뮬레이터는 기술 자료에 포함된 모든 데이터 소스에 대해 실행됩니다.

챗봇 시뮬레이터에서 내장된 데이터 소스와 채팅을 통해 기술 자료를 테스트할 수 있습니다. 기술 자료를 로컬로 테스트할 때 GenAI 벡터 데이터베이스에는 상호 작용이나 통찰력이 캡처되지 않습니다.

사용자의 애플리케이션에 기술 자료를 배포하기 전에 워크로드 공장 내에서 대부분의 테스트를 수행합니다. 데이터 원본이나 챗봇 작업을 변경해야 하는 경우 기술 자료를 게시하기 전에 지금 변경해야 합니다.

> 챗봇 시뮬레이터 창의 크기를 조정하고 제목을 변경하고 질문 및 응답을 클립보드에 복사할 수 있습니다.

챗봇을 테스트하기 위해 수행해야 할 작업은 다음과 같습니다.

- 조직과 관련된 많은 질문을 입력하여 답변이 예상과 같은지 확인합니다.
- 챗봇 애플리케이션의 사용자가 기본적으로 사용할 수 있도록 하려는 대화 시작자를 사용자 지정합니다.
- 챗봇 답변 하단에 제공된 특성 콘텐츠에 올바른 참조가 포함되어 있는지 확인합니다.

단계

1. Knowledge Base 인벤토리 페이지에서 테스트할 기술 문서를 선택합니다.

오른쪽 창에 챗봇 시뮬레이터가 나타납니다. 정의된 경우 기존 대화 시작도 표시됩니다.

2. 챗봇 입력 필드에 프롬프트 또는 질문을 입력하고 을 선택하여 챗봇이 조직 지식에 어떻게 반응하는지 확인합니다 .

- 응답 아래의 * Sources * 목록을 확장하면 답변을 생성하는 데 사용되는 소스를 확인할 수 있습니다. 여기에 답변을 생성하는 데 사용되는 파일 목록이 나와 있습니다. 파일 이름 위로 마우스를 가져가면 각 파일 및 볼륨 경로에서 사용되는 데이터 청크를 보고 각 파일에 복사할 수 있습니다.
- 응답에 테이블이 포함된 경우 각 열의 데이터를 정렬하고 각 테이블을 클립보드에 복사할 수 있습니다.

3. 기술 문서에서 보다 집중적인 답변을 제공할 수 있도록 데이터 원본을 업데이트해야 하는 경우 지금 해당 내용을 변경한 다음 기술 문서를 다시 테스트합니다.

GenAI 기술 문서에 대한 외부 인증을 활성화합니다

지식 기반 인증을 활성화하여 API 엔드포인트를 사용하여 지식 기반을 챗봇 애플리케이션과 통합할 때 토큰 검증 및 ACL이 필요하도록 합니다. 인증을 활성화할 때 챗봇 클라이언트의 기술 자료에 대한 API 요청에 사용할 JSON Web Token의 설정을 구성합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 인증을 활성화할 기술 문서를 선택합니다.
3. 를 선택하고 * 기술 자료 관리 * 를 선택합니다.
4. 조치 * 메뉴를 선택하고 * 인증 설정 관리 * 를 선택합니다.
5. 인증 설정:

   a. 인증 설정 활성화 * 를 선택합니다.

   b. 필요한 정보를 제공하십시오. 예제가 제공되지만 인증 공급자로부터 이러한 필드의 값을 얻어야 합니다.

      - * 알고리즘 *: 인증 공급자가 사용하는 서명 알고리즘입니다.
      - * Audience * (선택 사항): 토큰의 의도된 수신자를 포함하는 문자열(때로는 URL)입니다.
      - * 발급자 *: 토큰을 발급한 공급자를 식별하는 문자열입니다.

        예를 들어 Amazon Cognito는 다음과 같은 형식의 발급사 문자열을 사용합니다.

        ```
        https://cognito-idp-<region>.amazonaws.com/<UserPoolID>
        ```

        여기서 `<region>` 는 사용자 풀이 포함된 AWS 영역이며 `<UserPoolID>` 은 사용자 풀 ID입니다. 다음 명령을 사용하여 사용자 풀 ID를 검색할 수 있습니다.

        ```
        aws cognito-idp list-user-pools --max-results=60 --output=table
        ```

- * JWKS URI *: 이 토큰의 서명을 확인하는 데 필요한 공개 키를 제공하는 URI 문자열입니다.

  예를 들어 Amazon Cognito는 JWKS URI 문자열을 다음과 같은 형식으로 사용합니다.

```
https://cognito-idp.<region>.amazonaws.com/<userPoolId>/.well-known/jwks.json
```

+
여기서 `<region>` 는 사용자 풀이 포함된 AWS 영역이며 `<UserPoolID>` 은 사용자 풀 ID입니다. 다음 명령을 사용하여 사용자 풀 ID를 검색할 수 있습니다.

```
aws cognito-idp list-user-pools --max-results=60 --output=table
```

6. 저장 * 을 선택합니다.

결과

기술 자료에 대한 인증이 활성화되었으며, API 엔드포인트를 사용하여 기술 자료와 상호 작용하고 기술 자료를 챗봇 애플리케이션과 통합할 수 있습니다.

GenAI 기술 자료를 게시하고 고유한 엔드포인트를 봅니다

기술 자료를 로컬로 구축 및 테스트한 후에는 사용자가 기술 자료를 쿼리할 수 있도록 챗봇 애플리케이션과 통합할 수 있도록 기술 자료를 게시할 수 있습니다.

이 작업에 대해

기술 문서를 게시하면 채팅 응용 프로그램에서 사용할 수 있습니다. 게시 작업은 워크로드 팩토리 API를 트리거하여 고유한 엔드포인트를 생성하고 게시합니다. 게시한 후에는 채팅 응용 프로그램에 대한 기술 자료에 액세스할 수 있으며 API 엔드포인트를 통합할 수 있습니다.

게시한 각 기술 문서에는 고유한 끝점이 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 게시하려는 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드 및 포함된 모든 데이터 원본의 목록이 표시됩니다.

4. 작업 * 메뉴를 선택하고 * 게시 * 를 선택합니다.

   워크로드 공장에서 기술 자료를 게시합니다. 기술 문서에 대한 세부 정보 페이지에서 상태가 * 게시되지 않음 * 에서 * 게시됨 * 으로 변경됩니다

   이제 기술 문서의 고유 끝점에 대한 세부 정보를 얻을 수 있습니다.

5. 게시됨 상태 옆에 있는 * 보기 * 를 선택합니다.

워크로드 팩토리 API를 사용하여 기술 자료에 액세스하는 방법에 대한 세부 정보가 표시됩니다.

6. 게시된 정보 보기 * 대화 상자에서 기술 자료를 응용 프로그램과 통합하는 데 사용할 수 있는 API 끝점을 복사합니다.

   API 엔드포인트에 대해 자세히 알아보려면 로 "API 설명서" 이동하여 * AI > 외부 * 를 선택하십시오.

   이러한 끝점을 사용하려면 인증 공급자로부터 사용자 토큰을 얻어야 합니다.

결과

이제 기술 자료를 챗봇 애플리케이션과 통합하는 데 사용할 수 있는 게시된 기술 자료 및 고유한 엔드포인트가 생겼습니다.

GenAI 외부 예제 챗봇 애플리케이션을 사용합니다

기술 자료를 구성, 활성화 및 게시한 후 외부 애플리케이션 개발자는 NetApp에서 제공하는 오픈 소스 예제 챗봇 애플리케이션을 구성 및 실행하여 기술 자료와 상호 작용하고 워크로드 팩토리 API를 사용하여 자체 생성 AI 애플리케이션을 생성하는 방법을 배울 수 있습니다.

단계

1. "기술 문서를 작성합니다"..
2. "인증을 활성화합니다" 사용자가 만든 기술 문서에 대한 정보를 제공합니다.

   이렇게 하면 기술 문서에서 API 요청을 인증할 수 있으며 API 엔드포인트를 사용할 때 토큰 유효성 검사와 ACL이 필요합니다.

   > ⓘ 이 기술 문서와 통합되는 외부 채팅 응용 프로그램은 기술 문서의 인증 설정에서 구성한 것과 동일한 인증 공급자(발급사)를 사용해야 합니다.

3. "기술 문서를 게시합니다" 외부 애플리케이션에 대한 API 액세스를 활성화합니다.

   기술 자료가 게시되면 API 끝점에 외부에서 액세스할 수 있으며, 기술 자료를 외부 채팅 애플리케이션(예: 챗봇 애플리케이션)과 통합할 수 있습니다.

4. 에서 예제 챗봇 애플리케이션 패키지를 "GitHub를 참조하십시오"다운로드합니다.
5. 패키지에 포함된 README 파일의 지침에 따라 챗봇 애플리케이션을 설치하고 실행합니다.
6. 로 "http://localhost:9091" 이동하여 응용 프로그램에 로그인합니다.

   챗봇 애플리케이션 예가 나타납니다.

자세한 정보

"워크로드 팩토리 API 설명서"

RAG 기반 GenAI 애플리케이션을 생성합니다

지식 기반을 구축하고 챗봇을 테스트하면 사용자가 챗봇을 쿼리할 수 있는 애플리케이션을 설정할 수 있습니다.

"FSx for ONTAP에서 RAG 기반 AI 애플리케이션을 생성하는 방법에 대해 알아보십시오"

GenAI를 통해 앞으로 할 수 있는 일

이제 엔터프라이즈 데이터를 사용하여 기술 자료를 만들고 사용자를 위해 배포했으므로 ONTAP용 FSx 파일 시스템을 비롯한 기술 자료, 데이터 소스 및 RAG 인프라를 관리할 수 있습니다.

기술 자료 구성 요소를 관리하기 위해 수행할 수 있는 작업은 다음과 같습니다.

- 데이터 원본의 콘텐츠를 업데이트하거나 새 데이터 원본을 추가하고 이러한 변경 내용을 기술 자료 및 챗봇과 동기화합니다.
- 청킹 전략 및 사용 권한 인식(SMB 파일 액세스용)을 비롯한 데이터 소스 설정을 관리합니다.
- 채팅 모델 및 대화 시작자를 포함한 기술 자료 설정을 관리합니다.
- 기술 문서 게시를 취소하거나 변경한 후 다시 게시합니다.
- FSx for ONTAP 파일 시스템에서 중요한 데이터를 백업 및 보호하여 기술 자료 데이터 및 기타 인프라 구성요소를 항상 사용할 수 있도록 합니다.

FSx for ONTAP 파일 시스템 관리에 대한 자세한 내용은 로 "Amazon FSx for NetApp ONTAP 설명서의 워크로드 팩토리" 이동하여 사용할 수 있는 백업 및 보호 기능을 확인하십시오.

GenAI를 사용하여 Amazon Q Business용 커넥터를 만듭니다

시작하십시오

GenAI 커넥터를 위한 빠른 시작

Amazon FSx for NetApp ONTAP 파일 시스템에 있는 조직의 데이터를 사용하여 Amazon Q Business용 NetApp Connector를 만들어 보세요. 커넥터를 생성한 후 최종 사용자는 Amazon Q Business Assistant에 액세스하여 질문에 대한 조직 중심의 응답을 받을 수 있습니다.

워크로드 공장 에 로그인합니다

를 사용하여 "콘솔 환경" 로그인해야 "워크로드 팩토리에 계정을 설정합니다" 합니다.

GenAI 요구사항을 충족하도록 환경을 설정합니다

AWS 인프라, 배포되고 검색된 FSx for ONTAP 파일 시스템, 커넥터에 통합하려는 데이터 소스 목록, Amazon Q Business 애플리케이션에 대한 액세스 등을 배포하려면 AWS 자격 증명이 필요합니다.

"GenAI 요구 사항에 대해 자세히 알아보십시오"..

데이터 소스가 포함된 **FSx for ONTAP** 파일 시스템을 식별합니다

커넥터에 통합할 데이터 원본은 단일 FSx for ONTAP 파일 시스템 또는 여러 FSx for ONTAP 파일 시스템에 있을 수 있습니다. 이러한 시스템이 서로 다른 VPC에 있는 경우 동일한 네트워크 내에서 액세스할 수 있거나 VPC를 피어링하여 AI 엔진과 동일한 지역 및 AWS 계정을 사용해야 합니다.

"데이터 소스를 식별하는 방법에 대해 알아봅니다"..

GenAI 인프라를 구축합니다

인프라 배포 마법사를 시작하여 AWS 환경에 GenAI 인프라를 구축합니다. 이 프로세스는 NetApp GenAI 엔진용 EC2 인스턴스와 NetApp AI Engine 데이터베이스를 포함할 FSx for ONTAP 파일 시스템의 볼륨을 배포합니다. 볼륨은 커넥터에 대한 정보를 저장하는 데 사용됩니다.

"GenAI 인프라를 구축하는 방법에 대해 알아보십시오"..

다음 단계

이제 Amazon Q Business용 커넥터를 만들어 최종 사용자에게 조직 중심의 응답을 제공할 수 있습니다.

GenAI 커넥터 요구 사항

Amazon Q Business용 NetApp 커넥터를 생성하기 전에 워크로드 팩토리와 AWS가 올바르게 설정되었는지 확인하세요.

기본 **GenAI** 요구 사항

GenAI는 시작하기 전에 환경에 필요한 일반적인 요구 사항을 충족해야 합니다.

워크로드 공장 로그인 및 계정

를 사용하여 "콘솔 환경" 로그인해야 "워크로드 팩토리에 계정을 설정합니다" 합니다.

AWS 자격 증명 및 권한

워크로드 팩토리에 AWS 자격 증명을 읽기/쓰기 권한으로 추가해야 합니다. 즉, GenAI에 대해 워크로드 팩토리를 읽기/쓰기 모드로 사용해야 합니다.

현재 기본 모드와 읽기 전용 모드 권한은 지원되지 않습니다.

자격 증명을 설정할 때 아래 표시된 권한을 선택하면 FSx for ONTAP 파일 시스템을 관리하고 기술 자료 및 챗봇에 필요한 GenAI EC2 인스턴스 및 기타 AWS 리소스를 배포 및 관리할 수 있는 모든 권한을 얻을 수 있습니다.

"워크로드 팩토리에 AWS 자격 증명을 추가하는 방법에 대해 알아보십시오"

Amazon Q Business용 **NetApp** 커넥터 요구 사항

환경이 아마존 Q 비즈니스 커넥터에 대한 다음과 같은 특정 요구 사항을 충족하는지 확인하십시오.

Amazon Q Business 응용 프로그램

아마존 Q 비즈니스 응용 프로그램을 생성하거나 기존 응용 프로그램을 사용해야 합니다.

- 애플리케이션이 AWS 지역 중 하나에 존재하는지 확인합니다.
- 응용 프로그램에 대해 이 "인덱스를 만들었습니다" 있는지 확인합니다.
- 응용 프로그램이 오류 상태가 아닌지 확인합니다.

FSx for ONTAP 파일 시스템

FSx for ONTAP 파일 시스템이 하나 이상 필요합니다.

- 커넥터에 대한 정보를 저장하기 위해 NetApp GenAI 엔진에서 하나의 파일 시스템을 사용(또는 없는 경우 생성)합니다.

  이 FSx for ONTAP 파일 시스템은 FlexVol 볼륨을 사용해야 합니다. FlexGroup 볼륨은 지원되지 않습니다.

- 하나 이상의 파일 시스템에 커넥터에 추가할 데이터 원본이 포함됩니다.

  하나의 FSx for ONTAP 파일 시스템을 두 가지 용도로 사용하거나 여러 FSx for ONTAP 파일 시스템을 사용할 수 있습니다.

- AWS FSx for ONTAP 파일 시스템이 상주하는 AWS 지역, VPC 및 서브넷을 알아야 합니다.
- 이 배포의 일부인 AWS 리소스에 적용할 태그 키/값 쌍을 고려해야 합니다(선택 사항).
- NetApp AI 엔진 인스턴스에 안전하게 연결할 수 있는 키 쌍 정보를 알아야 합니다.

"FSx for ONTAP 파일 시스템을 구축 및 관리하는 방법에 관해 알아보십시오"

커넥터에 추가할 데이터 소스 식별

커넥터에 통합할 FSx for ONTAP 파일 시스템에 있는 문서(데이터 소스)를 식별하거나 생성합니다. 이러한 데이터 소스를 통해 Amazon Q Business는 조직과 관련된 데이터를 기반으로 사용자 쿼리에 정확하고 맞춤화된 답변을 제공할 수 있습니다.

최대 데이터 원본 수입니다

지원되는 최대 데이터 원본 수는 10개입니다.

데이터 소스의 위치입니다

데이터 소스는 Amazon FSx for NetApp ONTAP 파일 시스템의 SMB 공유 또는 NFS 내보내기의 단일 볼륨 또는 볼륨 내의 폴더에 저장될 수 있습니다. 데이터 소스는 NetApp SnapMirror 데이터 보호 관계에 있는 Amazon FSx for NetApp ONTAP 볼륨에 저장할 수도 있습니다.

볼륨이나 폴더 내에서 개별 문서를 선택할 수 없으므로 데이터 원본이 포함된 각 볼륨이나 폴더에 기술 문서와 통합되지 않아야 하는 관련 문서가 포함되어 있지 않은지 확인해야 합니다.

각 커넥터에 여러 데이터 소스를 추가할 수 있지만, 모두 AWS 계정에서 액세스할 수 있는 FSx for ONTAP 파일 시스템에 상주해야 합니다.

각 데이터 원본의 최대 파일 크기는 50MB입니다.

지원되는 프로토콜

커넥터는 NFS 또는 SMB/CIFS 프로토콜을 사용하는 볼륨의 데이터를 지원합니다. SMB 프로토콜을 사용하여 저장된 파일을 선택할 때 커넥터가 해당 볼륨의 파일에 액세스할 수 있도록 Active Directory 정보를 입력해야 합니다. 여기에는 Active Directory 도메인, IP 주소, 사용자 이름 및 암호가 포함됩니다.

SMB를 통해 액세스되는 공유(파일 또는 디렉토리)에 데이터 소스를 저장하는 경우, 해당 공유에 액세스할 수 있는 권한이 있는 챗봇 사용자 또는 그룹만 데이터에 액세스할 수 있습니다. 이 "권한 인식 기능"이 활성화되면 AI 시스템은 auth0의 사용자 이메일을 SMB 공유에서 파일을 보거나 사용할 수 있는 사용자와 비교합니다. 챗봇은 포함된 파일에 대한 사용자 권한을 기반으로 해답을 제공합니다.

예를 들어, 10개의 파일(데이터 소스)을 커넥터에 통합하고 2개의 파일이 제한된 정보를 포함하는 인적 리소스 파일인 경우, 이 두 파일에 액세스할 수 있도록 인증된 챗봇 사용자만 이러한 파일의 데이터를 포함하는 챗봇으로부터 응답을 받게 됩니다.

> Amazon Q Business 커넥터에 데이터 원본을 추가하면 데이터 원본 파일에 사용자 권한만 적용됩니다. 그룹 권한이 적용되지 않습니다.

> 데이터 원본의 파일에 텍스트가 없는 경우(예: 텍스트가 없는 이미지) Amazon Q Business는 해당 파일을 인덱싱하지 않고 텍스트가 없음을 나타내는 항목을 Amazon CloudWatch 로그에 기록합니다.

지원되는 데이터 소스 파일 형식

현재 Amazon Q Business용 NetApp Connector에서 지원되는 데이터 소스 파일 형식은 다음과 같습니다.

파일 형식	연장
쉼표로 구분된 값 파일입니다	.csv입니다
JSON 및 JSONP	제이슨
마크다운	진료 부서
Microsoft Word를	.docx입니다
일반 텍스트	.txt입니다
휴대용 문서 형식	PDF로 이동합니다
Microsoft PowerPoint를 클릭합니다	.ppt 또는 .pptx
하이퍼텍스트 마크업 언어	.html을 참조하십시오
확장 가능한 마크업 언어	XML
XSLT를 참조하십시오	.XSLT입니다
Microsoft Excel을 선택합니다	xls입니다
서식 있는 텍스트	.RTF를 클릭합니다

GenAI 인프라를 구축합니다

조직을 위해 FSx for ONTAP 지식 베이스, 커넥터 및 애플리케이션을 구축하기 전에 환경에 RAG 프레임워크용 GenAI 인프라를 구축해야 합니다. 기본 인프라 구성요소는 Amazon Bedrock 서비스, NetApp GenAI 엔진의 가상 머신 인스턴스 및 FSx for ONTAP 파일 시스템입니다.

구축된 인프라는 여러 지식 베이스, 챗봇, 커넥터를 지원할 수 있으므로 일반적으로 이 작업은 한 번만 수행하면 됩니다.

인프라 세부 정보

GenAI 배포는 Amazon Bedrock이 지원되는 AWS 지역에 있어야 합니다. "지원되는 지역 목록을 봅니다"

인프라는 다음과 같은 구성 요소로 이루어집니다.

아마존 **Bedrock** 서비스

Amazon Bedrock은 단일 API를 통해 선도적인 AI 회사의 기반 모델(FMS)을 사용할 수 있는 완전 관리형 서비스입니다. 또한 안전한 생성 AI 애플리케이션을 구축하는 데 필요한 기능을 제공합니다.

"아마존 Bedrock에 대해 자세히 알아보십시오"

아마존 **Q** 비즈니스

Amazon Q는 Amazon Bedrock을 기반으로 구축되어 질문에 답하고 데이터 소스의 정보를 기반으로 콘텐츠를 생성하는 데 사용할 수 있는 완전 관리형 생성 AI 도우미를 제공합니다.

"아마존 Q 비즈니스에 대해 자세히 알아보십시오"

NetApp GenAI 엔진용 가상 머신입니다

이 프로세스 중에 NetApp GenAI 엔진이 배포됩니다. 데이터 소스에서 데이터를 수집한 다음 해당 데이터를 벡터 데이터베이스에 쓸 수 있는 처리 능력을 제공합니다.

FSx for ONTAP 파일 시스템

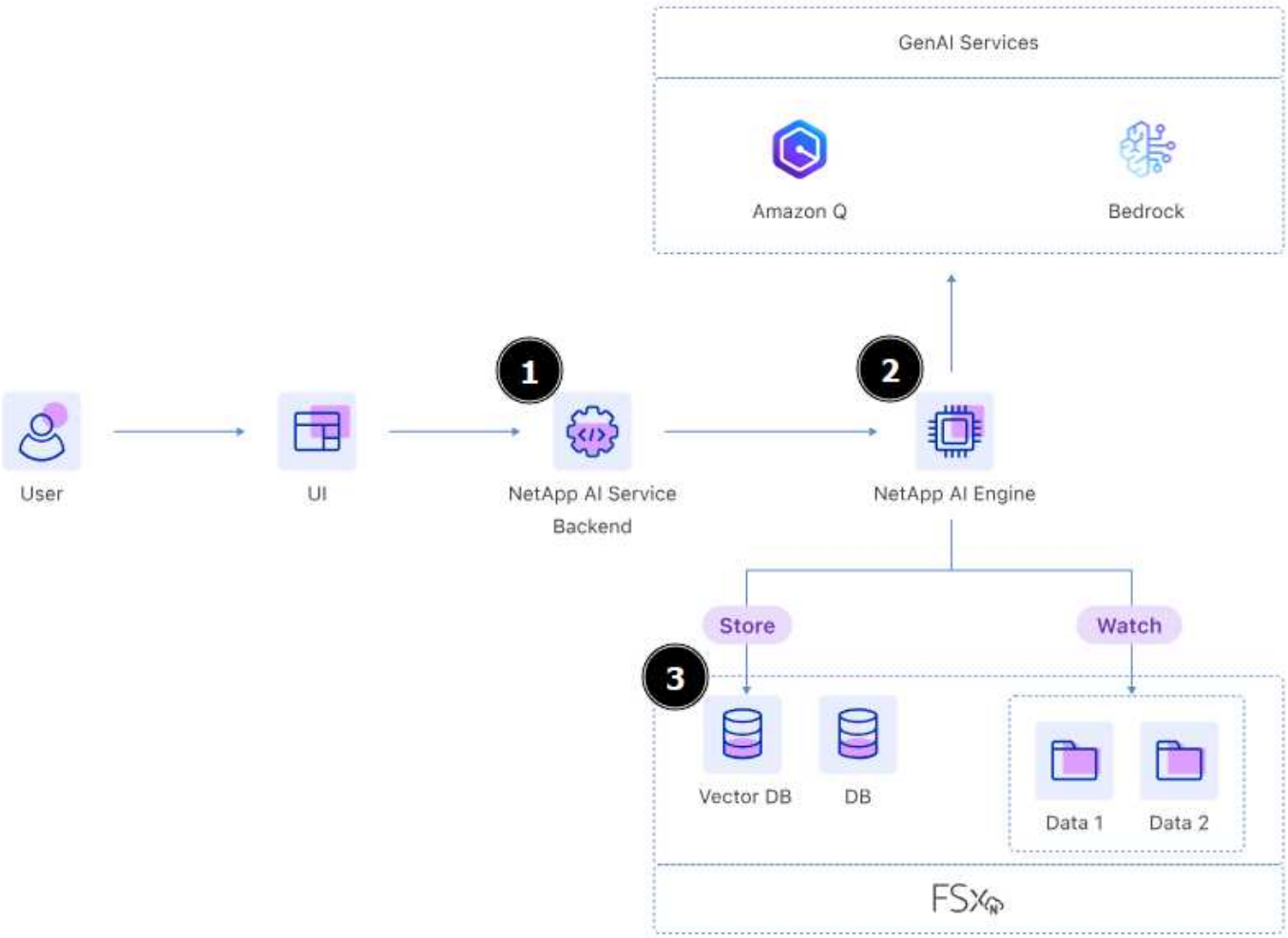
FSx for ONTAP 파일 시스템은 GenAI 시스템을 위한 스토리지를 제공합니다.

데이터 소스를 기반으로 기본 모델에 의해 생성된 데이터를 저장하는 벡터 데이터베이스를 포함하는 단일 볼륨이 배포됩니다.

기술 자료에 통합할 데이터 소스는 동일한 FSx for ONTAP 파일 시스템 또는 다른 시스템에 상주할 수 있습니다.

NetApp GenAI 엔진은 이 두 볼륨을 모두 모니터링하고 상호 작용합니다.

다음 이미지는 GenAI 인프라를 보여 줍니다. 이 절차를 수행하는 동안 번호가 1, 2, 3인 구성 요소가 전개됩니다. 배포를 시작하기 전에 다른 요소가 있어야 합니다.



GenAI 인프라를 구축합니다

AWS 자격 증명을 입력하고 FSx for ONTAP 파일 시스템을 선택하여 검색 증강 생성(RAG) 인프라를 배포해야 합니다.

시작하기 전에

이 절차를 시작하기 전에 사용자의 환경이 지식 베이스 또는 커넥터의 요구 사항을 충족하는지 확인하십시오.

- "기술 자료 요구 사항"
- "커넥터 요구 사항"

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. AI 워크로드 타일에서 * 배포 및 관리 * 를 선택합니다.
3. 인프라 다이어그램을 검토하고 * Next * 를 선택합니다.
4. AWS 설정 * 섹션의 항목을 완료합니다.
   a. * AWS 자격 증명 *: AWS 리소스 배포 권한을 제공하는 AWS 자격 증명을 선택하거나 추가합니다.
   b. * 위치 *: AWS 지역, VPC 및 서브넷을 선택합니다.

      GenAI 배포는 Amazon Bedrock이 활성화된 AWS 지역에 있어야 합니다. "지원되는 지역 목록을 봅니다"
5. 인프라 설정 * 섹션의 항목을 완료합니다.
   a. * 태그 * : 이 배포의 일부인 모든 AWS 리소스에 적용할 태그 키/값 쌍을 입력하십시오. 이러한 태그는 AWS 관리 콘솔 및 워크로드 공장 내의 인프라 정보 영역에 표시되며 워크로드 공장 리소스를 추적하는 데 도움이 됩니다.
6. Connectivity * 섹션을 완료합니다.
   a. * 키 쌍 * : NetApp GenAI 엔진 인스턴스에 안전하게 연결할 수 있는 키 쌍을 선택하십시오.
7. AI 엔진 * 섹션을 완료하십시오.
   a. * 인스턴스 이름 *: 필요에 따라 * 인스턴스 이름 정의 * 를 선택하고 AI 엔진 인스턴스의 사용자 정의 이름을 입력합니다. 인스턴스 이름은 AWS 관리 콘솔 및 워크로드 공장 내의 인프라 정보 영역에 표시되며, 작업 부하 공장 리소스를 추적하는 데 도움이 됩니다.
8. 배포 * 를 선택하여 배포를 시작합니다.

자격 증명 오류로 인해 배포가 실패하는 경우 오류 메시지 내에서 하이퍼링크를 선택하여 오류 세부 정보를 확인할 수 있습니다. 누락 또는 차단된 권한 목록과 GenAI 워크로드를 배포하기 위해 필요한 권한 목록을 확인할 수 있습니다.

결과

워크로드 팩토리가 챗봇 인프라 구축을 시작합니다. 이 프로세스는 최대 10분 정도 소요될 수 있습니다.

배포 프로세스 중에 다음 항목이 설정됩니다.

- 네트워크는 전용 끝점과 함께 설정됩니다.
- IAM 역할, 인스턴스 프로필 및 보안 그룹이 생성됩니다.
- GenAI 엔진의 가상 머신 인스턴스가 배포됩니다.
- Amazon Bedrock은 접두사가 있는 로그 그룹을 사용하여 Amazon CloudWatch 로그에 로그를 보내도록 구성되어 `/aws/bedrock/`있습니다.
- GenAI 엔진은 이름이 지정된 로그 그룹을 사용하여 Amazon CloudWatch 로그에 로그를 전송하도록 구성되어 `/netapp/wlmai/<tenancyAccountId>/randomId`있습니다. 여기서 는 현재 사용자에 대한 입니다. `<tenancyAccountID>` "BlueXP 계정 ID입니다"

Amazon Q Business용 NetApp 커넥터 만들기

AI 인프라를 배포하고 FSx for ONTAP 데이터 저장소에서 사용할 데이터 소스를 식별한 후 Amazon Q Business용 NetApp 커넥터를 정의할 준비가 되었습니다.

계속하기 전에 사용자의 환경이 아마존 Q 비즈니스를 위한 을 충족하는지 "요구 사항"확인하십시오.

이 작업에 대해

조직의 데이터 원본에는 PII(개인 식별 정보)가 포함될 수 있습니다. 이 중요한 정보를 보호하기 위해 커넥터를 정의할 때 _data guardrails_를 활성화할 수 있습니다. BlueXP 분류를 기반으로 하는 데이터 가드레일은 PII를 식별하고 마스킹하므로 액세스 및 복구가 불가능합니다.

"BlueXP 분류에 대해 알아보십시오"..

GenAI용 BlueXP 워크로드 공장은 민감한 개인 정보(SPII)를 마스킹하지 않습니다. 이 데이터 유형에 대한 자세한 내용은 을 "중요한 개인 데이터의 유형"참조하십시오.

데이터 가드레일은 언제든지 활성화 또는 비활성화할 수 있습니다. 데이터 가드레일 사용을 전환하면 워크로드 공장 출하 시 전체 데이터 소스를 처음부터 검사하여 비용이 발생할 수 있습니다.

커넥터를 정의합니다

Amazon Q Business용 NetApp 커넥터를 생성하세요. 이 커넥터는 GenAI와 Amazon Q Business 간의 API 및 데이터 소스 통신을 지원합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. AI 워크로드 타일에서 * 배포 및 관리 * 를 선택합니다.
3. Knowledge Base & Connectors 탭에서 * Create New * 드롭다운을 선택하고 * Amazon Q Business Connector * 를 선택합니다.
4. 커넥터 정의 페이지에서 커넥터 설정을 구성합니다.
   a. * 이름 *: 연결선에 사용할 이름을 입력합니다.
   b. * 설명 *: 커넥터에 대한 자세한 설명을 입력합니다.
   c. * Amazon Q *: 통합하려는 Amazon Q Business 인스턴스의 지역 및 응용 프로그램 이름입니다.
   d. * 데이터 가드레일 * : 데이터 가드레일을 활성화 또는 비활성화할지 선택합니다. "BlueXP 분류를 기반으로 하는 데이터 가드레일에 대해 알아보십시오"..

      데이터 가드레일을 사용하려면 다음 사전 요구 사항을 충족해야 합니다.

      - BlueXP 분류와 통신하려면 서비스 계정이 필요합니다. 서비스 계정을 생성하려면 BlueXP Tenancy 계정에 _Organization admin_역할이 있어야 합니다. 조직 관리자 역할을 가진 구성원은 BlueXP 의 모든 작업을 완료할 수 있습니다. "BlueXP 에서 구성원에 역할을 추가하는 방법에 대해 알아봅니다"
      - AI 엔진이 에 액세스할 수 있어야 "BlueXP API 끝점입니다"합니다.
      - 에 설명된 대로 다음을 수행해야 "BlueXP 분류 문서"합니다.

A. BlueXP Connector를 생성합니다

B. 환경이 사전 요구 사항을 충족할 수 있는지 확인합니다

C. BlueXP 분류를 배포합니다

> 데이터 가드레일 기능을 활성화하면 GenAI는 일반 텍스트(포함된 이미지 또는 미디어 텍스트 제외)만 수집하고 개인 또는 중요 데이터를 마스킹하여 .txt, .md, .csv, .docx 및 .pdf 파일을 처리합니다. 다른 모든 파일 형식은 개인 또는 중요 데이터를 마스킹하지 않고 정상적으로 처리됩니다.

e. **FSx for ONTAP** 파일 시스템: Amazon Q Business에 대한 새로운 NetApp 커넥터를 정의하면 워크로드 팩토리는 커넥터 정보를 저장하기 위해 새로운 Amazon FSx for NetApp ONTAP 볼륨을 생성합니다. 새 볼륨을 생성할 기존 파일 시스템 및 SVM(스토리지 VM이라고도 함)을 선택합니다.

f. * Snapshot policy *: 워크로드 공장 저장소 인벤토리에 정의된 기존 정책 목록에서 스냅샷 정책을 선택합니다. GenAI는 선택한 스냅샷 정책에 따라 커넥터 정보를 주파수 단위로 저장하는 볼륨의 반복 스냅샷을 자동으로 생성합니다.

필요한 스냅샷 정책이 없는 경우 "스냅샷 정책을 생성합니다" 볼륨이 포함된 스토리지 VM에서 수행할 수 있습니다.

5. Amazon Q Business를 GenAI와 통합하려면 * Create Connector * 를 선택하십시오.

커넥터가 만들어지는 동안 진행 표시기가 나타납니다.

커넥터가 만들어지면 데이터 원본을 커넥터에 추가하여 Amazon Q Business에서 데이터를 수집하여 인덱스에 추가할 수 있습니다. 지금 * 데이터 원본 추가 * 를 선택하고 하나 이상의 데이터 원본을 추가하는 것이 좋습니다.

데이터 원본을 커넥터에 추가합니다

하나 이상의 데이터 원본을 추가하여 Amazon Q Business 인덱스를 조직의 데이터로 채울 수 있습니다.

이 작업에 대해

- 지원되는 최대 데이터 원본 수는 10개입니다.
- 아마존 Q 비즈니스 지수의 특정 서비스 제한 사항은 을 "아마존 Q 비즈니스 문서" 참조하십시오.

단계

1. 데이터 소스 추가*를 선택하면 *파일 시스템 선택 페이지가 나타납니다.
2. * 파일 시스템 선택 *: 데이터 소스 파일이 있는 FSx for ONTAP 파일 시스템을 선택하고 * 다음 * 을 선택합니다.
3. * 볼륨 선택 *: 데이터 원본 파일이 있는 볼륨을 선택하고 * 다음 * 을 선택합니다.

SMB 프로토콜을 사용하여 저장된 파일을 선택할 때 도메인, IP 주소, 사용자 이름 및 암호를 포함한 Active Directory 정보를 입력해야 합니다.

4. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.
5. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.
   - * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.

- 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.
- 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

> ⓘ 수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

6. 선택한 데이터 원본이 SMB 프로토콜을 사용하는 볼륨에 있을 때만 사용할 수 있는 * 권한 인식 * 섹션에서 권한 인식 응답을 활성화하거나 비활성화할 수 있습니다.
   - * 활성화됨 *: 이 커넥터에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.
   - * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.

> ⓘ Active Directory 그룹 권한은 Amazon Q Business 커넥터 데이터 원본에 대해 지원되지 않습니다.

7. 이 데이터 소스를 아마존 Q 비즈니스 커넥터에 추가하려면 * 추가 * 를 선택하십시오.

결과

데이터 원본은 Amazon Q Business 인덱스에 포함됩니다. 데이터 원본이 완전히 포함되면 상태가 "포함"에서 "포함 "으로 변경됩니다.

커넥터에 단일 데이터 소스를 추가한 후 Amazon Q Business 챗봇 환경에서 테스트하고 필요한 사항을 변경한 후 사용자에게 서비스를 제공할 수 있습니다. 또한 같은 단계를 수행하여 데이터 원본을 커넥터에 추가할 수도 있습니다.

관리 및 모니터링

GenAI 인프라를 관리합니다

구축된 GenAI RAG 인프라에 대한 세부 정보를 보거나 더 이상 필요하지 않을 경우 챗봇 인프라를 제거할 수 있습니다.

인프라에 대한 정보를 봅니다

챗봇 인프라에 대한 정보를 볼 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 워크로드 팩토리 탐색 메뉴에서 * AI * 를 선택하십시오.
3. 인프라 * 탭을 선택합니다.
4. 다음 구성 요소에 대한 세부 정보가 포함된 인프라에 대한 정보를 봅니다.
    - AWS 설정
    - 인프라 설정
    - 있습니다
    - 벡터 데이터베이스

인프라를 제거합니다

하나 이상의 챗봇에 배포한 챗봇 인프라가 더 이상 필요하지 않다면 워크로드 팩토리에서 제거할 수 있습니다.

이 인프라에 배포된 모든 챗봇이 비활성화되고 모든 채팅 기록이 삭제됩니다.

이 작업을 수행하면 워크로드 팩토리에서 AI 인프라에 대한 링크만 제거되며 AWS에서 모든 구성요소가 제거되는 것은 아닙니다. AWS에서 다음 인프라 구성 요소를 수동으로 삭제해야 합니다.

- VM 인스턴스
- 전용 끝점
- AI 데이터베이스가 포함된 FSx for ONTAP 파일 시스템의 볼륨입니다
- IAM 역할
- 정책에 동의하게 됩니다
- 보안 그룹입니다

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 워크로드 팩토리 탐색 메뉴에서 * AI * 를 선택하십시오.
3. 인프라 * 탭을 선택합니다.

4. 를 ••• 선택하고 * 챗봇 인프라 제거 * 를 선택합니다.
5. 인프라를 삭제할 것인지 확인하고 * 제거 * 를 선택합니다.

결과

챗봇 인프라 구성 요소가 워크로드 팩토리에서 제거되었습니다.

GenAI 지식 기반 관리

기술 문서를 만든 후에는 기술 문서 세부 정보를 보거나 기술 문서를 수정하거나 추가 데이터 원본을 통합하거나 기술 문서를 삭제할 수 있습니다.

기술 문서에 대한 정보를 봅니다

통합된 기술 자료 및 데이터 원본에 대한 설정에 대한 정보를 볼 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 워크로드 팩토리 탐색 메뉴에서 * AI * 를 선택하십시오.
3. 보려는 기술 문서를 선택합니다.

   정의된 경우 현재 사용 중인 대화 시작점이 오른쪽 창에 표시됩니다.

4. 기술 문서 세부 정보를 보려면 * 기술 자료 관리 * 를 선택하고 ••• 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

   작업 * 메뉴를 사용하면 필요한 경우 기술 자료를 관리할 수 있습니다.

기술 문서를 편집합니다

일부 설정을 변경하여 기술 자료를 업데이트하거나 데이터 원본을 추가 또는 제거할 수 있습니다.

기술 문서에서 데이터 원본을 추가, 수정 또는 제거할 때마다 데이터 원본이 기술 문서에 다시 인덱싱되도록 데이터 원본을 동기화해야 합니다. 동기화는 증분 방식이므로 Amazon Bedrock은 마지막 동기화 이후 추가, 수정 또는 삭제된 FSx for ONTAP 볼륨의 개체만 처리합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 업데이트할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

4. 조치 * 메뉴를 선택하고 * 기술 문서 편집 * 을 선택합니다.
5. 지식 기반 편집 페이지에서는 지식 기반 이름, 설명, 내장 모델, 채팅 모델, 기능 활성화를 변경할 수 있으며, 대화 시작자를 자동으로 생성할지 수동으로 생성할지 선택할 수 있으며, 지식 기반이 포함된 볼륨에 사용되는 스냅샷

정책을 선택할 수 있습니다.

대화를 시작하는 데 수동 모드를 사용하는 경우 여기에서 대화 시작도 변경할 수 있습니다.

> 임베딩, 비용 등 모든 기술 자료 검사 기술 문서를 만든 후 데이터 가드레일을 활성화하면 기술 문서가 다시 검색되어 비용이 발생합니다. 마찬가지로 채팅 모델을 변경하면 GenAI가 관련 데이터 소스를 다시 스캔합니다(비용 발생).

6. 변경한 후 * 저장 * 을 선택합니다.

스냅샷으로 기술 자료 보호

기술 자료 볼륨의 스냅샷을 생성하고 복원하여 기술 자료 데이터를 보호할 수 있습니다. 언제든지 스냅샷에서 복원하여 이전 버전의 기술 자료로 되돌릴 수 있습니다.

스냅샷은 백업보다 더 빠르고 스토리지 효율적이며, 다른 보호 정책을 사용하여 각 기술 자료를 보호할 수 있습니다. 스냅샷이 유용한 시나리오는 다음과 같습니다.

- 우발적인 데이터 손실 또는 손상
- 기술 문서에 수집되는 잘못된 데이터를 복구하는 중입니다
- 다양한 데이터 소스 또는 청킹 전략을 테스트하고 테스트가 완료되면 신속하게 되돌릴 수 있습니다

기술 자료 볼륨의 스냅샷을 생성합니다

기술 자료 볼륨의 수동 스냅샷을 생성하여 기술 문서의 상태를 저장할 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 보호할 기술 자료를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

4. Actions * 메뉴를 선택하고 * Snapshot > Create new snapshot * 을 선택합니다.
5. 필요에 따라 * 스냅샷 이름 정의 * 를 선택하고 스냅샷에 대한 사용자 정의 이름을 입력합니다.

   사용자 지정 이름을 정의하면 나중에 복원해야 할 경우 스냅샷의 내용을 더 잘 확인할 수 있습니다.

6. Create * 를 선택합니다.

   기술 자료의 스냅샷이 생성됩니다.

기술 자료 볼륨의 스냅샷을 복원합니다

언제든지 기술 자료 볼륨의 수동 또는 예약 스냅샷을 복원할 수 있습니다.

볼륨에 저장된 데이터베이스가 손상되었거나 삭제된 경우 Generative AI 워크로드 UI를 사용하여 스냅샷을 복원할 수 없습니다. 이 문제를 해결하려면 볼륨이 호스팅되는 ONTAP 클러스터에서 을 사용하여 스냅샷을 복원할 수 "ONTAP CLI 를 참조하십시오" 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 복원할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

4. Actions * 메뉴를 선택하고 * Snapshot > Restore snapshot * 을 선택합니다.

   스냅샷 선택 대화 상자가 나타나고 이 기술 자료에 대해 생성된 스냅샷 목록을 볼 수 있습니다.

5. (선택 사항) 스냅샷이 복원된 후 예약 및 현재 실행 중인 데이터 소스 스캔을 계속하려면 * Pause running and scheduled scans after the snapshot Restore * (스냅숏 복원 후 실행 및 예약된 스캔 일시 중지) 옵션을 선택 취소합니다.

   이 옵션은 기본적으로 활성화되어 있으므로 기술 문서가 부분적으로 복원된 상태일 때 검사가 수행되지 않거나 검사가 새로 복원된 기술 문서를 이전 데이터로 업데이트하지 않습니다.

6. 목록에서 복구할 스냅샷을 선택합니다.
7. Restore * 를 선택합니다.

기술 문서를 복제합니다

기술 자료 스냅샷에서 새로운 기술 자료를 생성할 수 있습니다. 이 기능은 원본 기술 문서가 손상되었거나 손실된 경우에 유용합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 복원할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

4. Actions * 메뉴를 선택하고 * Snapshot > Clone Knowledge Base * 를 선택합니다.

   클론 대화 상자가 나타납니다.

5. 스냅숏이 클론 생성된 후 예약 및 현재 실행 중인 데이터 원본 스캔을 계속하려면 * 스냅숏 클론 생성 후 실행 및 예약된 검사 일시 중지 * 옵션을 선택 취소합니다.

   이 옵션은 기본적으로 활성화되어 있으므로 기술 문서가 부분적으로 복원된 상태일 때 검사가 수행되지 않거나 검사가 새로 복원된 기술 문서를 이전 데이터로 업데이트하지 않습니다.

6. 목록에서 복제할 스냅샷을 선택합니다.

7. Continue * 를 선택합니다.
8. 새 기술 문서의 이름을 입력합니다.
9. 새 기술 자료에 사용할 파일 시스템 SVM 및 볼륨 이름을 선택합니다.
10. 클론 * 을 선택합니다.

기술 문서에 데이터 원본을 추가합니다

추가 데이터 원본을 기술 문서에 포함시켜 추가 조직 데이터로 채울 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 데이터 소스를 추가할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 데이터 원본 추가 * 를 선택합니다.
4. 추가하려는 데이터 소스 유형을 선택하세요.
   - ONTAP 파일 시스템용 FSx 추가(기존 ONTAP 볼륨용 FSx의 파일 사용)
   - 파일 시스템 추가(일반 SMB 또는 NFS 공유의 파일 사용)

ONTAP 파일 시스템에 **FSx** 추가

1. * 파일 시스템 선택 *: 데이터 소스 파일이 있는 FSx for ONTAP 파일 시스템을 선택하고 * 다음 * 을 선택합니다.
2. * 볼륨 선택 *: 데이터 원본 파일이 있는 볼륨을 선택하고 * 다음 * 을 선택합니다.

   SMB 프로토콜을 사용하여 저장된 파일을 선택할 때 도메인, IP 주소, 사용자 이름 및 암호를 포함한 Active Directory 정보를 입력해야 합니다.

3. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.
4. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.

   ◦ * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.

   ▪ * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.

   ▪ * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

   높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   ◦ * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.

   ▪ 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

   이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

   이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든 PII가 GenAI 데이터베이스에 마스킹됩니다.

   이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

   ◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

   수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

5. 선택한 데이터 원본이 SMB 프로토콜을 사용하는 볼륨에 있을 때만 사용할 수 있는 * 권한 인식 * 섹션에서 권한 인식 응답을 활성화하거나 비활성화할 수 있습니다.

- 사용: 이 기술 자료에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.
- * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.

6. 이 데이터 소스를 기술 문서에 추가하려면 * 추가 * 를 선택하십시오.

일반 **NFS** 파일 시스템 추가

1. 파일 시스템 선택: 데이터 소스 파일이 있는 파일 시스템 호스트의 IP 주소 또는 FQDN을 입력하고, 네트워크 공유에 대한 NFS 프로토콜을 선택하고 *다음*을 선택합니다.
2. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.

> 경우에 따라 NFS 내보내기 이름을 직접 입력하고 *디렉터리 검색*을 선택하여 사용 가능한 디렉터리를 표시해야 할 수도 있습니다. 내보내기 전체 또는 내보내기에서 특정 폴더만 선택할 수 있습니다.

3. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.
   - * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.
     - * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.
     - * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

     > 높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   - * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.
     - 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

       이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

     이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든 PII가 GenAI 데이터베이스에 마스킹됩니다.

     > 이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

   - 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

4. *데이터 소스 추가*를 선택하여 이 데이터 소스를 지식 기반에 추가하세요.

일반 **SMB** 파일 시스템 추가

1. 파일 시스템 선택:
   a. 데이터 소스 파일이 있는 파일 시스템 호스트의 IP 주소나 FQDN을 입력하세요.
   b. 네트워크 공유에 SMB 프로토콜을 선택합니다.
   c. 도메인, IP 주소, 사용자 이름, 비밀번호 등 Active Directory 정보를 입력합니다.
   d. 다음 * 을 선택합니다.
2. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.

   경우에 따라 SMB 공유 이름을 직접 입력하고 *디렉터리 검색*을 선택하여 사용 가능한 디렉터리를 표시해야 할 수도 있습니다. 전체 공유를 선택하거나 공유에서 특정 폴더만 선택할 수 있습니다.

3. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.
   - * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.
     - * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.
     - * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

       높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   - 권한 인식: 권한 인식 응답을 활성화하거나 비활성화합니다.
     - 사용: 이 기술 자료에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.
     - * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.
   - * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.
     - 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

       이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

   이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든

PII가 GenAI 데이터베이스에 마스킹됩니다.

> 이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

> 수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

4. *데이터 소스 추가*를 선택하여 이 데이터 소스를 지식 기반에 추가하세요.

결과

데이터 원본은 기술 자료에 통합됩니다.

데이터 원본을 기술 문서와 동기화합니다

데이터 소스는 하루에 한 번 관련 기술 자료와 자동으로 동기화되므로 데이터 소스 변경 사항이 챗봇에 반영됩니다. 데이터 원본을 변경하고 데이터를 즉시 동기화하려는 경우 필요 시 동기화를 수행할 수 있습니다.

동기화는 증분 동기화이므로 Amazon Bedrock은 마지막 동기화 이후 추가, 수정 또는 삭제된 데이터 원본의 객체만 처리합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 동기화할 기술 자료를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.
4. 조치 * 메뉴를 선택하고 * 지금 스캔 * 을 선택합니다.

   데이터 원본을 스캔한다는 메시지와 검사가 완료되면 최종 메시지가 표시됩니다.

결과

기술 자료는 첨부된 데이터 원본과 동기화되며 활성 챗봇은 데이터 원본의 최신 정보를 사용하기 시작합니다.

예약된 동기화를 일시 중지하거나 다시 시작합니다

데이터 원본의 다음 동기화(스캔)를 일시 중지하거나 다시 시작하려면 언제든지 다시 시작할 수 있습니다. 데이터 원본을 변경하고 변경 기간 동안 동기화를 실행하지 않으려면 다음 예약된 동기화를 일시 중지해야 할 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base & Connectors 탭에서 스캔을 일시 중지하거나 다시 시작할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

4. Actions * 메뉴를 선택하고 * Scan > Pause Scheduled Scan * 또는 * Scan > Resume Scheduled Scan * 을 선택합니다.

   다음 예약된 스캔이 일시 중지되었거나 다시 시작되었다는 메시지가 표시됩니다.

기술 문서를 생성하기 전에 채팅 모델을 평가합니다

기술 자료를 작성하기 전에 사용 가능한 기본 채팅 모델을 평가할 수 있으므로 구현에 가장 적합한 모델을 확인할 수 있습니다. 모델 지원은 AWS 지역에 따라 다르므로, 기술 자료를 배포할 계획이 있는 지역에서 사용할 수 있는 모델을 확인하려면 을 참조하십시오 "이 AWS 설명서 페이지" .

> 이 기능은 Knowledge Base가 생성되지 않은 경우(Knowledge Base 인벤토리 페이지에 Knowledge Base가 없는 경우에만) 사용할 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 Chatbot 페이지 오른쪽에 채팅 모델을 선택하는 옵션이 표시됩니다.
3. 목록에서 채팅 모델을 선택하고 프롬프트 영역에 질문 집합을 입력하여 챗봇이 어떻게 응답하는지 확인합니다.
4. 여러 모델을 시도하여 구현에 가장 적합한 모델을 확인하십시오.

결과

기술 문서를 작성할 때 이 채팅 모델을 사용하십시오.

기술 문서 게시를 취소합니다

챗봇 애플리케이션과 통합될 수 있도록 기술 자료를 게시한 후, 챗봇 애플리케이션이 기술 자료에 액세스하지 못하도록 하려면 게시를 취소할 수 있습니다.

기술 문서의 게시를 취소하면 모든 채팅 응용 프로그램이 작동하지 않습니다. 기술 자료에 액세스할 수 있는 고유 API 끝점이 비활성화됩니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 게시를 취소할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드 및 포함된 모든 데이터 원본의 목록이 표시됩니다.

4. 작업 * 메뉴를 선택하고 * 게시 취소 * 를 선택합니다.

결과

기술 문서가 비활성화되고 챗봇 애플리케이션에서 더 이상 액세스할 수 없습니다.

기술 문서를 삭제합니다

기술 문서가 더 이상 필요하지 않은 경우 삭제할 수 있습니다. 기술 문서를 삭제하면 작업 부하 공장에서 제거되고 기술 문서가 포함된 볼륨이 삭제됩니다. 기술 자료를 사용하는 애플리케이션이나 챗봇이 작동하지 않습니다. 지식 기반

삭제는 되돌릴 수 없습니다.

기술 문서를 삭제할 때 기술 문서와 관련된 모든 상담원과의 연결을 해제하여 기술 문서와 연결된 모든 리소스를 완전히 삭제해야 합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 삭제할 기술 문서를 선택합니다.
3. 를 ••• 선택하고 * 기술 자료 관리 * 를 선택합니다.
4. 조치 * 메뉴를 선택하고 * 기술 문서 삭제 * 를 선택합니다.
5. 기술 자료 삭제 대화 상자에서 삭제할 내용을 확인하고 * 삭제 * 를 선택합니다.

결과

기술 문서가 작업 부하 공장에서 제거되고 관련 볼륨이 삭제됩니다.

Amazon Q Business 커넥터를 관리합니다

Amazon Q Business용 커넥터를 만든 후에는 커넥터 세부 정보를 보거나 커넥터를 수정하거나 추가 데이터 원본을 통합하거나 커넥터를 삭제할 수 있습니다.

커넥터에 대한 정보를 봅니다

통합된 커넥터 및 데이터 원본에 대한 설정 정보를 볼 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 워크로드 팩토리 탐색 메뉴에서 * AI * 를 선택하십시오.
3. 보려는 연결선을 선택합니다.
4. 커넥터 세부 정보를 보려면 * 커넥터 관리 * 를 선택하고 ••• 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

   조치 * 메뉴를 사용하면 커넥터를 원하는 경우 관리할 수 있습니다.

연결선을 편집합니다

일부 설정을 변경하여 연결선을 업데이트하거나 데이터 원본을 추가 또는 제거할 수 있습니다.

커넥터에서 데이터 원본을 추가, 수정 또는 제거할 때마다 GenAI는 다시 인덱싱되도록 데이터 원본 정보를 Amazon Q Business로 전송해야 합니다. 동기화는 증분 방식이므로 Amazon Q Business는 마지막 동기화 이후 추가, 수정 또는 삭제된 FSx for ONTAP 볼륨의 객체만 처리합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base & Connectors 인벤토리 페이지에서 업데이트할 커넥터를 선택합니다.

3. 을 선택하고 * 커넥터 관리 * 를 선택합니다.

   이 페이지에는 게시된 상태, 데이터 원본의 포함 상태, 포함 모드, 포함된 모든 데이터 원본의 목록 등이 표시됩니다.

4. 작업 * 메뉴를 선택하고 * 커넥터 편집 * 을 선택합니다.
5. 커넥터 편집 페이지에서 커넥터가 포함된 볼륨에 사용되는 커넥터 이름, 설명, 포함 모델, 데이터 가드레일 사용 및 스냅샷 정책을 변경할 수 있습니다.

   임베딩을 포함한 모든 데이터 소스 스캔에는 비용이 소요됩니다. 커넥터를 만든 후 데이터 가드레일을 활성화하면 데이터 소스가 다시 검사되어 비용이 발생합니다.

6. 변경한 후 * 저장 * 을 선택합니다.

커넥터에 데이터 원본을 추가합니다

커넥터에 추가 데이터 원본을 포함시켜 추가 조직 데이터로 채울 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base & Connectors 인벤토리 페이지에서 데이터 원본을 추가할 커넥터를 선택합니다.
3. 를 선택하고 * 데이터 원본 추가 * 를 선택합니다.
4. 추가하려는 데이터 소스 유형을 선택하세요.
   - ONTAP 파일 시스템용 FSx 추가(기존 ONTAP 볼륨용 FSx의 파일 사용)
   - 파일 시스템 추가(일반 SMB 또는 NFS 공유의 파일 사용)

ONTAP 파일 시스템에 **FSx** 추가

1. * 파일 시스템 선택 *: 데이터 소스 파일이 있는 FSx for ONTAP 파일 시스템을 선택하고 * 다음 * 을 선택합니다.
2. * 볼륨 선택 *: 데이터 원본 파일이 있는 볼륨을 선택하고 * 다음 * 을 선택합니다.

   SMB 프로토콜을 사용하여 저장된 파일을 선택할 때 도메인, IP 주소, 사용자 이름 및 암호를 포함한 Active Directory 정보를 입력해야 합니다.

3. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.
4. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.

   ◦ * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.

   ▪ * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.

   ▪ * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

   높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   ◦ * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.

   ▪ 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

   이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

   이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든 PII가 GenAI 데이터베이스에 마스킹됩니다.

   이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

   ◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

   수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

5. 선택한 데이터 원본이 SMB 프로토콜을 사용하는 볼륨에 있을 때만 사용할 수 있는 * 권한 인식 * 섹션에서 권한 인식 응답을 활성화하거나 비활성화할 수 있습니다.

◦ 사용: 이 기술 자료에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.

◦ * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.

6. 이 데이터 소스를 기술 문서에 추가하려면 * 추가 * 를 선택하십시오.

일반 **NFS** 파일 시스템 추가

1. 파일 시스템 선택: 데이터 소스 파일이 있는 파일 시스템 호스트의 IP 주소 또는 FQDN을 입력하고, 네트워크 공유에 대한 NFS 프로토콜을 선택하고 *다음*을 선택합니다.
2. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.

> 경우에 따라 NFS 내보내기 이름을 직접 입력하고 *디렉터리 검색*을 선택하여 사용 가능한 디렉터리를 표시해야 할 수도 있습니다. 내보내기 전체 또는 내보내기에서 특정 폴더만 선택할 수 있습니다.

3. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.

◦ * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.

▪ * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.

▪ * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

> 높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

◦ * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.

▪ 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든 PII가 GenAI 데이터베이스에 마스킹됩니다.

> 이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

4. *데이터 소스 추가*를 선택하여 이 데이터 소스를 지식 기반에 추가하세요.

일반 **SMB** 파일 시스템 추가

1. 파일 시스템 선택:
   a. 데이터 소스 파일이 있는 파일 시스템 호스트의 IP 주소나 FQDN을 입력하세요.
   b. 네트워크 공유에 SMB 프로토콜을 선택합니다.
   c. 도메인, IP 주소, 사용자 이름, 비밀번호 등 Active Directory 정보를 입력합니다.
   d. 다음 * 을 선택합니다.
2. * 데이터 소스 선택 *: 파일을 저장한 위치를 기준으로 데이터 소스 위치를 선택합니다. 전체 볼륨일 수도 있고 볼륨의 특정 폴더 또는 하위 폴더일 수도 있고 * 다음 * 을 선택합니다.

   경우에 따라 SMB 공유 이름을 직접 입력하고 *디렉터리 검색*을 선택하여 사용 가능한 디렉터리를 표시해야 할 수도 있습니다. 전체 공유를 선택하거나 공유에서 특정 폴더만 선택할 수 있습니다.

3. * 구성 *: 데이터 소스가 파일에서 정보를 수집하는 방법과 검색에 포함할 파일을 구성합니다.
   - * 데이터 소스 정의 *: * 청크 전략 * 섹션에서 데이터 소스가 기술 문서에 통합될 때 GenAI 엔진이 데이터 소스 컨텐츠를 청크로 분할하는 방법을 정의합니다. 다음 전략 중 하나를 선택할 수 있습니다.
     - * 다중 문장 청킹 *: 데이터 소스의 정보를 문장 정의 청크로 정리합니다. 각 청크를 구성하는 문장의 수(최대 100개)를 선택할 수 있습니다.
     - * 오버랩 기반 청크 *: 데이터 소스의 정보를 인접 청크와 겹칠 수 있는 문자 정의 청크로 구성합니다. 각 청크의 크기를 문자 단위로 선택하고 각 청크가 인접한 청크와 겹치는 정도를 선택할 수 있습니다. 청크 크기는 50자에서 3000자 사이이고 겹치는 비율은 1 ~ 99%로 구성할 수 있습니다.

       높은 중복 비율을 선택하면 검색 정확도가 약간 개선되어 저장소 요구 사항이 크게 증가할 수 있습니다.

   - 권한 인식: 권한 인식 응답을 활성화하거나 비활성화합니다.
     - 사용: 이 기술 자료에 액세스하는 챗봇 사용자는 액세스 권한이 있는 데이터 원본에서 쿼리에 대한 응답만 받습니다.
     - * 사용 안 함 * : 챗봇 사용자는 모든 통합 데이터 소스의 콘텐츠를 사용하여 응답을 받습니다.
   - * 파일 필터링 *: 검색에 포함할 파일을 구성합니다.
     - 파일 형식 지원 * 섹션에서 모든 파일 형식을 포함하거나 데이터 원본 검색에 포함할 개별 파일 형식을 선택합니다.

       이미지 또는 PDF 파일을 포함하는 경우 GenAI용 BlueXP 워크로드 공장에서 이미지(PDF 문서의 이미지 포함)의 텍스트를 구문 분석하므로 비용이 더 많이 듭니다.

   이미지의 텍스트 데이터를 포함할 경우, 스캔된 텍스트 데이터가 사용자 환경에서 AWS로 전송되기 때문에 GenAI는 이미지에서 PII(개인 식별 정보)를 마스킹할 수 없습니다. 그러나 데이터가 저장되면 모든

PII가 GenAI 데이터베이스에 마스킹됩니다.

> 이미지 파일을 스캔에 포함할지 여부는 기술 자료 채팅 모델과 관련이 있습니다. 스캔에 이미지 파일을 포함할 경우 채팅 모델은 이미지를 지원해야 합니다. 여기에서 이미지 파일 형식을 선택하면 기술 문서를 이미지 파일을 지원하지 않는 채팅 모델로 전환할 수 없습니다.

◦ 파일 수정 시간 필터 * 섹션에서 수정 시간에 따라 파일 포함을 활성화 또는 비활성화하도록 선택합니다. 수정 시간 필터링을 사용하는 경우 목록에서 날짜 범위를 선택합니다.

> 수정 날짜 범위를 기준으로 파일을 포함하는 경우 날짜 범위가 충족되지 않으면(지정한 날짜 범위 내에서 파일이 수정되지 않음) 파일이 정기 검색에서 제외되고 데이터 원본에 이러한 파일이 포함되지 않습니다.

4. *데이터 소스 추가*를 선택하여 이 데이터 소스를 지식 기반에 추가하세요.

결과

데이터 원본이 커넥터에 통합되어 있습니다.

데이터 원본을 커넥터와 동기화합니다

데이터 원본은 하루에 한 번 연결된 커넥터와 자동으로 동기화되므로 데이터 원본 변경 내용이 Amazon Q Business에 반영됩니다. 데이터 원본을 변경하고 데이터를 즉시 동기화(검사)하려는 경우 필요 시 동기화를 수행할 수 있습니다.

동기화는 증분 동기화이므로 Amazon Q Business는 마지막 동기화 이후 추가, 수정 또는 삭제된 데이터 원본의 객체만 처리합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 지식 기반 및 커넥터 탭에서 동기화할 커넥터를 선택합니다.
3. 을 선택하고 * 커넥터 관리 * 를 선택합니다.
4. 조치 * 메뉴를 선택하고 * 지금 스캔 * 을 선택합니다.

   데이터 원본을 스캔한다는 메시지와 검사가 완료되면 최종 메시지가 표시됩니다.

결과

커넥터는 첨부된 데이터 원본과 동기화되며 Amazon Q Business는 데이터 원본의 최신 정보를 사용하기 시작합니다.

예약된 동기화를 일시 중지하거나 다시 시작합니다

데이터 원본의 다음 동기화(스캔)를 일시 중지하거나 다시 시작하려면 언제든지 다시 시작할 수 있습니다. 데이터 원본을 변경하고 변경 기간 동안 동기화를 실행하지 않으려면 다음 예약된 동기화를 일시 중지해야 할 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 커넥터 인벤토리 페이지에서 스캔을 일시 중지하거나 다시 시작할 커넥터를 선택합니다.
3. 을 선택하고 * 커넥터 관리 * 를 선택합니다.

4. Actions * 메뉴를 선택하고 * Scan > Pause Scheduled Scan * 또는 * Scan > Resume Scheduled Scan * 을 선택합니다.

   다음 예약된 스캔이 일시 중지되었거나 다시 시작되었다는 메시지가 표시됩니다.

연결선을 삭제합니다

연결선이 더 이상 필요하지 않으면 삭제할 수 있습니다. 커넥터를 삭제하면 작업 부하 공장에서 제거되고 커넥터가 포함된 볼륨이 삭제됩니다. 커넥터 삭제는 되돌릴 수 없습니다.

커넥터를 삭제할 때 커넥터와 관련된 모든 에이전트에서 커넥터를 연결 해제하여 커넥터와 연결된 모든 리소스를 완전히 삭제해야 합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base & Connectors 인벤토리 페이지에서 삭제할 커넥터를 선택합니다.
3. 을 선택하고 * 커넥터 관리 * 를 선택합니다.
4. 조치 * 메뉴를 선택하고 * 커넥터 삭제 * 를 선택합니다.
5. 커넥터 삭제 대화 상자에서 삭제할 커넥터를 확인하고 * 삭제 * 를 선택합니다.

결과

커넥터가 작업 부하 공장에서 제거되고 관련 볼륨이 삭제됩니다.

GenAI 데이터 소스를 관리합니다

FSx for ONTAP 파일 시스템에서 데이터 소스를 사용하여 기술 자료 또는 커넥터를 생성한 후에는 데이터 소스 세부 정보를 보거나, 데이터 소스 콘텐츠를 업데이트 또는 변경하거나, 데이터 소스 설정을 편집하거나, 데이터 소스를 삭제할 수 있습니다.

데이터 원본에 대한 정보를 봅니다

데이터 원본의 내용에 대한 정보를 볼 수 있으며 기술 문서 또는 커넥터를 사용하여 포함 상태를 볼 수 있습니다. 데이터 원본은 기술 문서 또는 커넥터와 연결되어 있으므로 데이터 원본 세부 정보를 보려면 먼저 기술 문서나 커넥터를 선택해야 합니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경" 로그인합니다.
2. 워크로드 팩토리 탐색 메뉴에서 * AI * 를 선택하십시오.
3. 데이터 원본이 있는 기술 문서 또는 커넥터를 선택한 다음 * 지식 기반 관리 * 또는 * 커넥터 관리 * 를 선택하고 선택합니다.

   페이지의 아래쪽에는 연결된 데이터 원본이 나열됩니다.

4. 을 선택하여 각 행을 확장하여 FSx for ONTAP 파일 시스템, 볼륨 및 데이터 소스가 있는 경로와 같은 각 데이터 소스에 대한 자세한 정보를 확인합니다.

또한 포함 정보 및 해당 데이터 원본이 현재 기술 문서나 커넥터에 포함되어 있는지 여부도 나열합니다.

데이터 원본 설정을 편집합니다

기술 자료 또는 연결선과 통합된 데이터 원본에 대한 정보를 편집할 수 있습니다. 대부분의 정보는 데이터 원본을 추가한 후에 수정되지만 일부 구성(예: 청킹 정의 또는 권한 인식)을 변경할 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. Knowledge Base 인벤토리 페이지에서 데이터 원본이 있는 기술 문서를 선택한 다음 * Manage Knowledge Base * 를 선택하고 ••• 선택합니다.

   페이지 아래쪽에는 이 기술 자료에 포함된 데이터 원본이 나열됩니다.
3. 편집할 데이터 원본의 행에서 * 데이터 원본 편집 * 을 선택하고 ••• 선택합니다.
4. 데이터 원본 편집 페이지에서 ⌄청크 정의에 대한 행을 확장하도록 선택합니다.
5. 청킹 전략 및 구성, 사용 권한 인식(SMB 볼륨의 경우)에 대한 설정을 업데이트하고 * 저장 * 을 선택합니다.

결과

데이터 소스 설정이 업데이트되고 AI 시스템이 데이터 소스를 동기화하여 기술 자료에 다시 인덱싱되도록 합니다.

기존 데이터 원본의 내용을 업데이트합니다

언제든지 데이터 원본의 내용을 변경하여 조직 데이터를 추가하거나 업데이트할 수 있습니다. 이 데이터 원본이 기술 문서에서 활발하게 사용되는 경우 데이터 원본이 기술 문서에 다시 인덱싱되도록 데이터 원본을 동기화해야 합니다. 동기화는 증분 방식이므로 Amazon Bedrock은 마지막 동기화 이후 추가, 수정 또는 삭제된 FSx for ONTAP 볼륨의 개체만 처리합니다.

데이터 소스는 하루에 한 번 기술 자료와 자동으로 동기화되므로 데이터 소스 변경 사항이 챗봇에 반영됩니다. 데이터 원본을 변경하고 데이터를 즉시 동기화하려는 경우 할 수 "필요 시 동기화를 수행합니다"있습니다.

데이터 원본을 삭제합니다

데이터 원본이 더 이상 기술 문서에 포함되지 않아도 되는 경우 삭제할 수 있습니다.

단계

1. 중 하나를 사용하여 워크로드 팩토리에 "콘솔 환경"로그인합니다.
2. 기술 문서 인벤토리 페이지에서 데이터 원본이 있는 기술 문서를 선택한 다음 * 기술 자료 관리 * 를 선택하고 ••• 선택합니다.

   페이지 아래쪽에는 이 기술 자료에 포함된 데이터 원본이 나열됩니다.
3. 삭제할 데이터 원본의 행에서 * 데이터 원본 삭제 * 를 선택하고 ••• 선택합니다.
4. 데이터 원본 삭제 대화 상자에서 삭제할지 확인하고 * 확인 * 을 선택합니다.

결과

데이터 원본이 기술 자료에서 제거되고 AI 시스템은 기술 자료에서 이 데이터 원본에 대한 인덱싱된 정보를 제거합니다. 해당 데이터 소스의 정보는 기술 자료를 사용하는 챗봇에 더 이상 사용할 수 없습니다.

BlueXP 워크로드 팩토리의 Tracker를 사용하여 워크로드 작업을 모니터링합니다.

BlueXP 워크로드 공장에서 Tracker를 사용하여 작업 부하 작업의 실행을 모니터링 및 추적하고 작업 진행 상황을 모니터링합니다.

이 작업에 대해

워크로드 팩토리에는 작업 부하 작업의 진행 상황과 상태를 모니터링 및 추적하고 작업 작업 및 하위 작업에 대한 세부 정보를 검토하며 문제나 장애를 진단할 수 있는 모니터링 기능인 Tracker가 있습니다.

Tracker에서 몇 가지 작업을 사용할 수 있습니다. 시간 프레임(최근 24시간, 7일, 14일 또는 30일), 작업 부하, 상태 및 사용자별로 작업을 필터링하고 검색 기능을 사용하여 작업을 찾고 작업 테이블을 CSV 파일로 다운로드할 수 있습니다. 언제든지 Tracker를 새로 고치고 실패한 작업을 재시도하거나 실패한 작업에 대한 매개 변수를 편집하고 작업을 다시 시도할 수 있습니다.

추적기는 작업에 따라 두 가지 수준의 모니터링을 지원합니다. 파일 시스템 배포와 같은 각 작업은 작업 설명, 상태, 시작 시간, 작업 기간, 사용자, 영역, 프록시 리소스, 작업 ID 및 모든 관련 하위 작업을 표시합니다. API 응답을 보고 작업 중에 발생한 상황을 파악할 수 있습니다.

예를 포함한 추적기 작업 수준

- 레벨 1(작업): 파일 시스템 구축을 추적합니다.
- 레벨 2(하위 작업): 파일 시스템 구축과 관련된 하위 작업을 추적합니다.

작업 상태

Tracker의 작업 상태는 다음과 같습니다_in progress_,_success_및_failed_입니다.

작동 주파수

작업 빈도는 작업 유형 및 작업 일정을 기반으로 합니다.

이벤트 보존

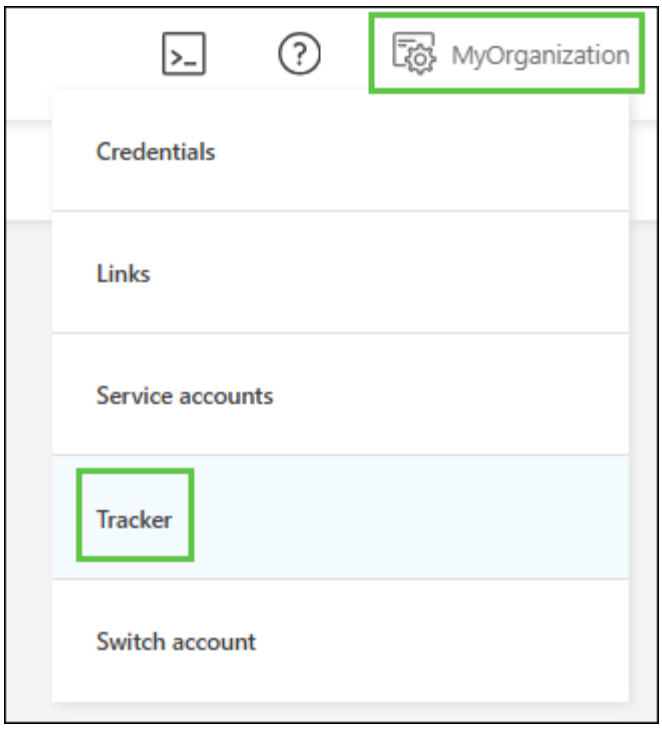
이벤트는 사용자 인터페이스에서 30일 동안 유지됩니다.

작업을 추적하고 모니터링합니다

추적기를 사용하여 BlueXP 에서 작업을 추적하고 모니터링합니다.

단계

1. 중 하나를 사용하여 "콘솔 환경"로그인합니다.
2. 작업 부하에서 계정 설정 메뉴를 선택한 다음 * Tracker * 를 선택합니다.



3. Tracker(추적기) 탭에서 필터를 사용하거나 검색을 사용하여 작업 결과를 좁힙니다. 작업 보고서를 다운로드할 수도 있습니다.

API 요청을 봅니다

Tracker의 작업에 대한 코드상자에서 API 요청을 봅니다.

단계

1. Tracker에서 작업을 선택합니다.
2. 3점 메뉴를 선택한 다음 * API 요청 보기 * 를 선택합니다.

실패한 작업을 다시 시도하십시오

Tracker에서 실패한 작업을 재시도하십시오. 실패한 작업의 오류 메시지를 복사할 수도 있습니다.

실패한 작업을 최대 10회까지 재시도할 수 있습니다.

단계

1. Tracker에서 실패한 작업을 선택합니다.
2. 세 개의 점 메뉴를 선택한 다음 * 재시도 * 를 선택하십시오.

결과

작업이 다시 시작됩니다.

실패한 작업을 편집한 후 다시 시도하십시오

실패한 작업의 매개 변수를 편집하고 Tracker 외부에서 작업을 재시도하십시오.

단계

1. Tracker에서 실패한 작업을 선택합니다.
2. 세 점 메뉴를 선택한 다음 * 편집 을 선택하고 다시 시도하십시오 *.

   매개 변수를 편집하고 작업을 다시 시도할 수 있는 작업 페이지로 리디렉션됩니다.

결과

작업이 다시 시작됩니다. Tracker로 이동하여 작업 상태를 확인합니다.

지식 및 지원

GenAI를 위한 BlueXP 워크로드 팩토리 지원에 등록하십시오

BlueXP 워크로드 팩토리 및 스토리지 솔루션 및 서비스와 관련된 기술 지원을 받으려면 지원 등록을 해야 합니다. 워크로드 공장 과는 별도의 웹 기반 콘솔인 BlueXP 콘솔에서 지원을 등록해야 합니다.

지원을 등록한다고 해서 클라우드 공급자 파일 서비스에 대한 NetApp 지원이 활성화되지 않습니다. 클라우드 공급자 파일 서비스, 인프라 또는 서비스를 사용하는 솔루션과 관련된 기술 지원은 해당 제품에 대한 워크로드 공장 설명서의 "도움말 얻기"를 참조하십시오.

"ONTAP용 Amazon FSx"

지원 등록 개요

계정 ID 지원 가입 등록(BlueXP의 지원 리소스 페이지에 있는 20자리 960xxxxxxxx 일련 번호)은 단일 지원 구독 ID로 사용됩니다. 각 BlueXP 계정 수준 지원 구독을 등록해야 합니다.

등록하면 지원 티켓 열기 및 자동 사례 생성과 같은 기능을 사용할 수 있습니다. 아래 설명과 같이 BlueXP에 NetApp Support 사이트(NSS) 계정을 추가하여 등록을 완료합니다.

NetApp 지원에 계정을 등록합니다

지원을 등록하고 지원 권한을 활성화하려면 계정 사용자 한 명이 NetApp Support 사이트 계정을 BlueXP 로그인과 연결해야 합니다. NetApp 지원에 등록하는 방법은 NetApp Support 사이트(NSS) 계정이 이미 있는지 여부에 따라 다릅니다.

NSS 계정이 있는 기존 고객

NSS 계정을 가지고 있는 NetApp 고객은 BlueXP를 통해 지원을 받기 위해 등록하기만 하면 됩니다.

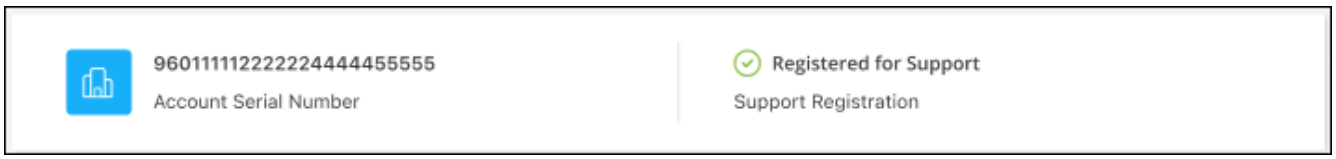
단계

1. 워크로드 팩토리 콘솔의 오른쪽 위에서 * Help > Support * 를 선택합니다.

   이 옵션을 선택하면 BlueXP 콘솔에 새 브라우저 탭이 열리고 지원 대시보드가 로드됩니다.

2. BlueXP 콘솔의 오른쪽 상단에서 설정 아이콘을 선택하고 * 자격 증명 * 을 선택합니다.
3. 사용자 자격 증명 * 을 선택합니다.
4. NSS 자격 증명 추가 * 를 선택하고 NetApp Support 사이트(NSS) 인증 프롬프트를 따릅니다.
5. 등록 프로세스가 성공적으로 완료되었는지 확인하려면 도움말 아이콘을 선택하고 * 지원 * 을 선택합니다.

   리소스 * 페이지에 계정이 지원을 위해 등록되었다는 내용이 표시됩니다.



다른 BlueXP 사용자는 BlueXP 로그인과 NetApp Support 사이트 계정을 연결하지 않은 경우 동일한 지원 등록 상태를 볼 수 없습니다. 그러나 BlueXP 계정이 지원을 위해 등록되지 않은 것은 아닙니다. 계정에 있는 한 사용자가 이 단계를 따랐을 경우 계정이 등록되었습니다.

기존 고객이지만 **NSS** 계정은 없습니다

기존 사용권 및 제품 번호가 _NO_NSS인 기존 NetApp 고객인 경우 NSS 계정을 만들어 BlueXP 로그인과 연결해야 합니다.

단계

1. 를 완료하여 NetApp 지원 사이트 계정을 만듭니다 "NetApp Support 사이트 사용자 등록 양식"
   a. 적절한 사용자 레벨(일반적으로 * NetApp 고객/최종 사용자 *)을 선택해야 합니다.
   b. 위에 사용된 BlueXP 계정 일련 번호(960xxxx)를 일련 번호 필드에 복사해야 합니다. 이렇게 하면 계정 처리 속도가 빨라집니다.
2. 아래의 단계를 완료하여 새 NSS 계정을 BlueXP 로그인에 연결합니다 NSS 계정이 있는 기존 고객.

NetApp이 처음이었습니다

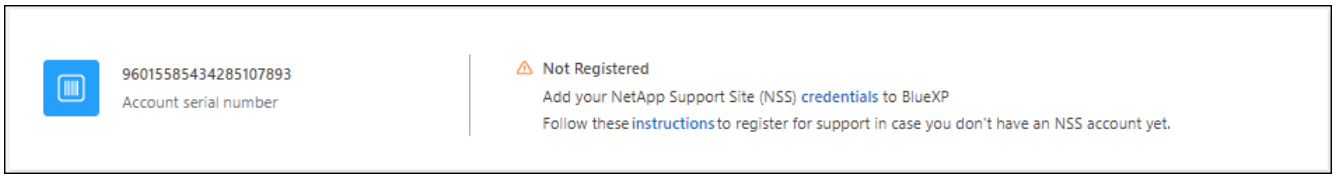
NetApp의 새로운 브랜드이고 NSS 계정이 없는 경우 아래의 각 단계를 수행하십시오.

단계

1. 워크로드 팩토리 콘솔의 오른쪽 위에서 * Help > Support * 를 선택합니다.

   이 옵션을 선택하면 BlueXP 콘솔에 새 브라우저 탭이 열리고 지원 대시보드가 로드됩니다.

2. 지원 리소스 페이지에서 계정 ID 일련 번호를 찾습니다.



3. 로 "NetApp의 지원 등록 사이트" 이동하여 * 등록된 NetApp 고객이 아님 * 을 선택합니다.
4. 필수 필드(빨간색 별표가 있는 필드)를 입력합니다.
5. [제품 라인] * 필드에서 * Cloud Manager * 를 선택한 다음 해당 청구 공급자를 선택합니다.
6. 위의 2단계에서 계정의 일련 번호를 복사하고 보안 검색을 완료한 다음 NetApp의 글로벌 데이터 개인 정보 보호 정책을 읽는지 확인합니다.

   이 보안 트랜잭션을 완료하기 위해 제공된 사서함으로 즉시 이메일이 전송됩니다. 몇 분 내에 확인 이메일이 도착하지 않으면 스팸 폴더를 확인해야 합니다.

7. 이메일 내에서 작업을 확인합니다.

확인 시 NetApp에 요청이 제출되고 NetApp Support 사이트 계정을 만들 것을 권장합니다.

8. 를 완료하여 NetApp 지원 사이트 계정을 만듭니다 "NetApp Support 사이트 사용자 등록 양식"
   a. 적절한 사용자 레벨(일반적으로 * NetApp 고객/최종 사용자 *)을 선택해야 합니다.
   b. 일련 번호 필드에 위에서 사용된 계정 일련 번호(960xxxx)를 복사해 주십시오. 이렇게 하면 계정 처리 속도가 빨라집니다.

작업을 마친 후

이 과정에서 NetApp이 연락을 드릴 것입니다. 신규 사용자를 위한 일회성 온보딩 연습입니다.

NetApp Support 사이트 계정이 있는 경우 아래의 단계를 완료하여 계정을 BlueXP 로그인에 연결합니다 NSS 계정이 있는 기존 고객.

GenAI 문제 해결

발생할 수 있는 몇 가지 일반적인 문제를 해결하는 방법에 대해 알아봅니다.

일반적인 문제 및 해결 방법

이러한 문제 중 하나가 있는 경우 해결 방법 열에 있는 단계를 사용하여 문제를 해결할 수 있습니다.

영역	문제	원인	해결 방법
구축	볼륨이 이미 있기 때문에 배포가 실패합니다.	GenAI용 BlueXP 워크로드 팩토리에는 구축 프로세스 중에 새 볼륨을 생성해야 하지만, 지정한 이름을 사용하는 볼륨이 이미 존재합니다.	새 볼륨에 사용할 고유한 이름을 지정하고 다시 배포하십시오.
구축	GenAI용 BlueXP 워크로드 팩토리에서 볼륨을 마운트할 수 없기 때문에 구축이 실패합니다.	FSx for NetApp ONTAP에 필요한 하나 이상의 인바운드 포트가 닫히거나 필터링되었습니다.	다음 인바운드 포트를 엽니다.

| 프로토콜 | 포트 | 목적

| 모든 ICMP | 모두 | 인스턴스에 Ping을 수행 중입니다

| HTTPS | 443 | Connector에서 fsxadmin 관리 LIF로 액세스하여 FSx로 API 호출을 전송합니다

| SSH를 클릭합니다 | 22 | 클러스터 관리 LIF 또는 노드 관리 LIF의 IP 주소에 SSH를 액세스할 수 있습니다

| TCP | 111 | NFS에 대한 원격 프로시저 호출

| TCP | 139 | CIFS에 대한 NetBIOS 서비스 세션입니다

| TCP | 161-162 을 참조하십시오 | 단순한 네트워크 관리 프로토콜

| TCP | 445 | Microsoft SMB/CIFS over TCP 및 NetBIOS 프레임

| TCP | 635 | NFS 마운트

| TCP | 749 | Kerberos

| TCP | 2049 | NFS 서버 데몬

| TCP | 3260 | iSCSI 데이터 LIF를 통한 iSCSI 액세스

| TCP | 4045 | NFS 잠금 데몬

| TCP | 4046 | NFS에 대한 네트워크 상태 모니터

| TCP | 10000 | NDMP를 사용한 백업

| TCP | 11104 | SnapMirror에 대한 인터클러스터 통신 세션의 관리

| TCP | 11105 | 인터클러스터 LIF를 사용하여 SnapMirror 데이터 전송

| UDP입니다 | 111 | NFS에 대한 원격 프로시저 호출

| UDP입니다 | 161-162 을 참조하십시오 | 단순한 네트워크 관리 프로토콜

| UDP입니다 | 635 | NFS 마운트

| UDP입니다 | 2049 | NFS 서버 데몬

| UDP입니다 | 4045 | NFS 잠금 데몬

| UDP입니다 | 4046 | NFS에 대한 네트워크 상태 모니터

| UDP입니다 | 4049 | NFS rquotad 프로토콜

유지 관리	AI 엔진을 시작하지 못하고 * Knowledge Base * 페이지에 "AI engine instance error" 오류가 표시됩니다.	AI 엔진 인스턴스가 손상되었거나 존재하지 않습니다.	Rebuild * 버튼을 선택합니다. GenAI를 위한 BlueXP 워크로드 공장은 인프라를 리빌드하고 리빌드 진행률을 표시합니다. 완료되면 지식 베이스가 재구축된 인프라에 다시 연결되고 지식 베이스 목록이 표시됩니다.
유지 관리	AI 엔진이 시작되지 않고 * Knowledge Base * 페이지에 "GenAI 엔진 인스턴스가 중지되었습니다" 오류가 표시됩니다.	AI 엔진 인스턴스가 실행되고 있지 않습니다.	AWS Management Console 또는 AWS CLI를 사용하여 AI 엔진 인스턴스를 시작합니다.

유지 관리	AI 엔진이 시작되지 않고 * Knowledge Base * 페이지에 "GenAI 엔진 서버가 응답하지 않습니다 "라는 오류가 표시됩니다.	AI 엔진 인스턴스가 응답하지 않습니다.	다음 복구 단계를 사용합니다. 단계 1. GenAI 엔진 인스턴스에 대한 SSH 액세스를 사용하도록 GenAI 엔진 인스턴스 보안 그룹을 수정합니다. 2. SSH를 사용하여 인스턴스에 로그인합니다. 3. 다음 명령을 실행합니다. `docker-compose up`

유지 관리	BlueXP 워크로드 팩터에서 GenAI를 사용하는 백엔드 Docker 인스턴스를 시작하지 못했습니다.	볼륨이 삭제되고 EC2 인스턴스가 다시 시작되었습니다.	다음 복구 단계를 사용합니다.
			
			단계
			1. FSx for NetApp ONTAP에서 새 볼륨을 생성합니다. 예를 들어, 볼륨 이름은 `netapp_ai` 이고 볼륨 경로는 일 수 `/netapp_ai` 있습니다.
			2. Amazon EC2 인스턴스에 대한 SSH.
			3. 볼륨 나열:
			`docker volume list`
			4. 이전 볼륨 제거:
			`docker volume rm ec2-user_persistent_folder`
			5. `docker-compose.yml`텍스트 편집기를 사용하여 파일을 엽니다.
			6. `volumes`섹션에서 디바이스 경로를 새 볼륨 경로로 변경합니다. 예를 들면 다음과 같습니다.

유지 관리	BlueXP 워크로드 팩터에서 GenAI를 사용하는 백엔드 Docker 인스턴스를 시작하지 못했습니다.	루트 볼륨이 삭제되었습니다.	이름과 경로를 사용하여 볼륨을 생성한 다음 Amazon EC2에서 백엔드 Docker 인스턴스를 다시 시작합니다.
유지 관리	BlueXP 워크로드 팩터에서 GenAI를 사용하는 백엔드 Docker 인스턴스를 시작하지 못했습니다.	루트 볼륨이 삭제되었습니다.	이름과 경로를 사용하여 볼륨을 생성한 다음 Amazon EC2에서 백엔드 Docker 인스턴스를 다시 시작합니다.

GenAI용 BlueXP 워크로드 팩토리에 도움을 받으십시오

NetApp은 BlueXP 워크로드 팩토리 및 클라우드 서비스를 다양한 방식으로 지원합니다. 기술 자료(KB) 기사 및 커뮤니티 포럼과 같은 광범위한 무료 셀프 지원 옵션이 24x7 제공됩니다. 지원 등록에는 웹 티켓팅을 통한 원격 기술 지원이 포함됩니다.

FSx for ONTAP에 대한 지원을 받으십시오

FSx for ONTAP, 인프라 또는 서비스를 사용하는 솔루션과 관련된 기술 지원은 해당 제품의 워크로드 공장 설명서의 "도움말 얻기"를 참조하십시오.

"ONTAP용 Amazon FSx"

Workload Factory 및 해당 스토리지 솔루션 및 서비스에 대한 기술 지원을 받으려면 아래에 설명된 지원 옵션을 사용하십시오.

자체 지원 옵션을 사용합니다

이 옵션은 하루 24시간, 주 7일 동안 무료로 사용할 수 있습니다.

- 문서화

  현재 보고 있는 작업 부하 공장 문서입니다.

- "기술 자료"

  워크로드 팩토리 기술 문서를 검색하여 문제 해결에 유용한 문서를 찾습니다.

- "커뮤니티"

  워크로드 공장 커뮤니티에 참여하여 진행 중인 토론을 따르거나 새 토론을 만드십시오.

NetApp Support로 케이스 생성

위의 자체 지원 옵션 외에도 NetApp 지원 전문가와 협력하여 지원을 활성화한 이후의 모든 문제를 해결할 수 있습니다.

시작하기 전에

케이스 생성 * 기능을 사용하려면 먼저 지원을 등록해야 합니다. NetApp 지원 사이트 자격 증명을 워크로드 공장

```
o:
"addr=svm-
0df66b96a890
d8a72.\
fs-
0c57008aac4
12bc3.\

fsx.us-east-
1.amazonaws.
com,nobrl,s
oft,rw"

device:
':/netapp_ai
' # Path to
new volume
```

로그인에 연결하십시오. "지원 등록 방법을 알아보십시오"..

단계

1. 워크로드 팩토리 콘솔의 오른쪽 위에서 * Help > Support * 를 선택합니다.

   이 옵션을 선택하면 BlueXP 콘솔에 새 브라우저 탭이 열리고 지원 대시보드가 로드됩니다.

2. 리소스 * 페이지의 기술 지원 아래에서 사용 가능한 옵션 중 하나를 선택합니다.

   a. 전화로 통화하려면 * 전화 * 를 선택하십시오. 전화를 걸 수 있는 전화 번호가 나열된 netapp.com 페이지로 연결됩니다.

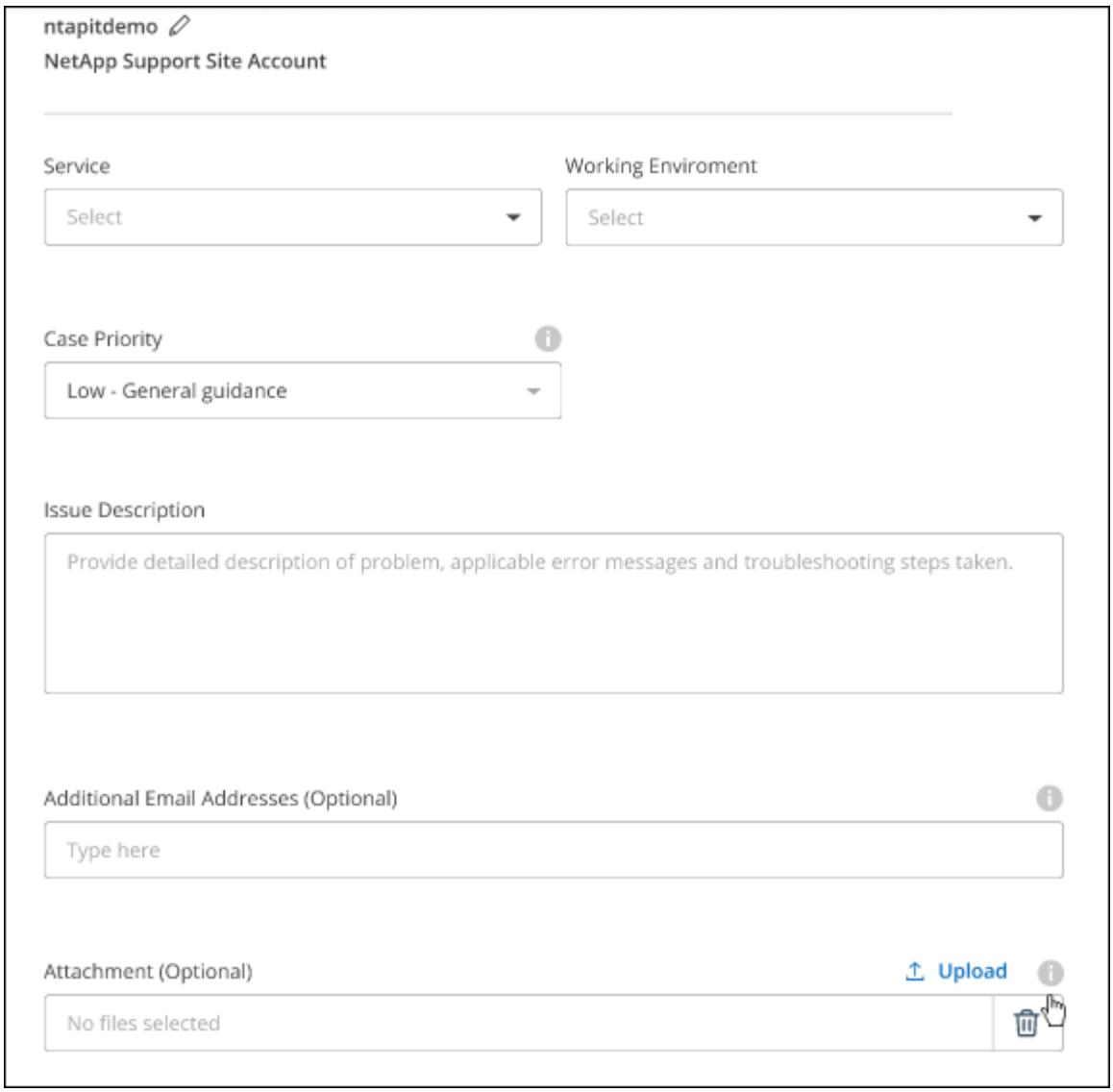
   b. NetApp 지원 전문가와 함께 티켓을 열려면 * 케이스 생성 * 을 선택하십시오.

      - * 서비스 *: * 워크로드 팩토리 * 를 선택합니다.
      - * 케이스 우선 순위 *: 케이스의 우선 순위를 선택합니다. 우선 순위는 낮음, 중간, 높음 또는 긴급입니다.

        이러한 우선 순위에 대한 자세한 내용을 보려면 필드 이름 옆에 있는 정보 아이콘 위로 마우스를 가져갑니다.

      - * 문제 설명 * : 해당 오류 메시지 또는 수행한 문제 해결 단계를 포함하여 문제에 대한 자세한 설명을 제공합니다.
      - * 추가 이메일 주소 *: 다른 사람에게 이 문제를 알고자 할 경우 추가 이메일 주소를 입력하십시오.
      - * 첨부 파일(선택 사항) *: 한 번에 하나씩 최대 5개의 첨부 파일을 업로드합니다.

        첨부 파일은 파일당 25MB로 제한됩니다. txt, log, pdf, jpg/jpeg, rtf, DOC/docx, xls/xlsx 및 CSV.



작업을 마친 후

지원 케이스 번호와 함께 팝업이 나타납니다. NetApp 지원 전문가가 귀사의 사례를 검토하고 곧 다시 연결해 드릴 것입니다.

지원 케이스 기록을 보려면 * 설정 > 일정 * 을 선택하고 "지원 케이스 생성"이라는 작업을 찾을 수 있습니다. 맨 오른쪽에 있는 버튼을 사용하면 작업을 확장하여 세부 정보를 볼 수 있습니다.

케이스를 생성하려고 할 때 다음과 같은 오류 메시지가 나타날 수 있습니다.

"선택한 서비스에 대해 케이스를 생성할 권한이 없습니다."

이 오류는 NSS 계정과 연결된 레코드 회사가 BlueXP 계정 일련 번호(960xxxx) 또는 작동 환경 일련 번호 다음 옵션 중 하나를 사용하여 지원을 요청할 수 있습니다.

- 제품 내 채팅을 사용합니다
- 에서 비기술적 케이스를 제출하십시오 https://mysupport.netapp.com/site/help

지원 사례 관리(Preview)

BlueXP에서 직접 활성 및 해결된 지원 사례를 보고 관리할 수 있습니다. NSS 계정 및 회사와 관련된 케이스를 관리할 수 있습니다.

케이스 관리를 미리 보기로 사용할 수 있습니다. NetApp은 이 경험을 개선하고 다음 릴리즈에서 향상된 기능을 추가할 계획입니다. 제품 내 채팅을 사용하여 피드백을 보내주십시오.

다음 사항에 유의하십시오.

- 페이지 상단의 케이스 관리 대시보드에서는 두 가지 보기를 제공합니다.
  - 왼쪽 보기에는 사용자가 제공한 NSS 계정으로 지난 3개월 동안 개설된 총 케이스가 표시됩니다.
  - 오른쪽 보기에는 사용자 NSS 계정을 기준으로 회사 수준에서 지난 3개월 동안 개설된 총 사례가 표시됩니다.

  테이블의 결과에는 선택한 보기와 관련된 사례가 반영됩니다.
- 관심 있는 열을 추가 또는 제거할 수 있으며 우선 순위 및 상태 등의 열 내용을 필터링할 수 있습니다. 다른 열은 정렬 기능만 제공합니다.

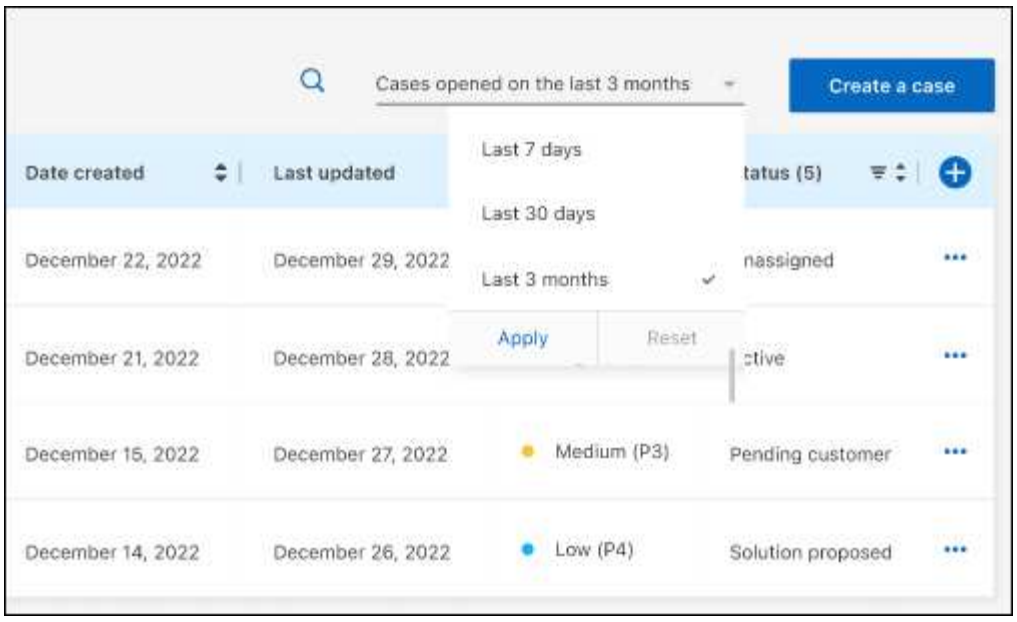
  자세한 내용은 아래 단계를 참조하십시오.
- 케이스 수준별로 케이스 메모를 업데이트하거나 아직 종결 또는 미결 종결 상태가 아닌 케이스를 종결할 수 있습니다.

단계

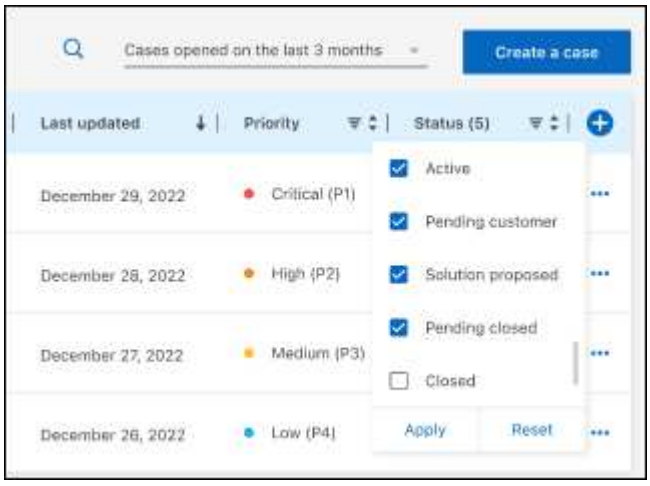
1. 워크로드 팩토리 콘솔의 오른쪽 위에서 * Help > Support * 를 선택합니다.

   이 옵션을 선택하면 BlueXP 콘솔에 새 브라우저 탭이 열리고 지원 대시보드가 로드됩니다.
2. Case Management * 를 선택하고 메시지가 표시되면 NSS 계정을 BlueXP에 추가합니다.

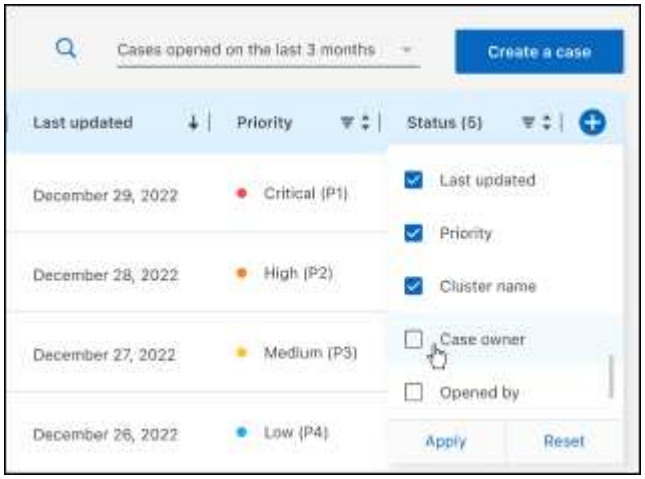
   케이스 관리 * 페이지에는 BlueXP 사용자 계정과 연결된 NSS 계정과 관련된 미해결 케이스가 표시됩니다. NSS 관리 * 페이지 상단에 나타나는 것과 동일한 NSS 계정입니다.
3. 필요한 경우 테이블에 표시되는 정보를 수정합니다.
   - 조직의 케이스 * 에서 * 보기 * 를 선택하여 회사와 관련된 모든 케이스를 봅니다.
   - 정확한 날짜 범위를 선택하거나 다른 기간을 선택하여 날짜 범위를 수정합니다.



◦ 열의 내용을 필터링합니다.



◦ 표시할 열을 선택한 다음 선택하여 테이블에 표시되는 열을 변경합니다 .



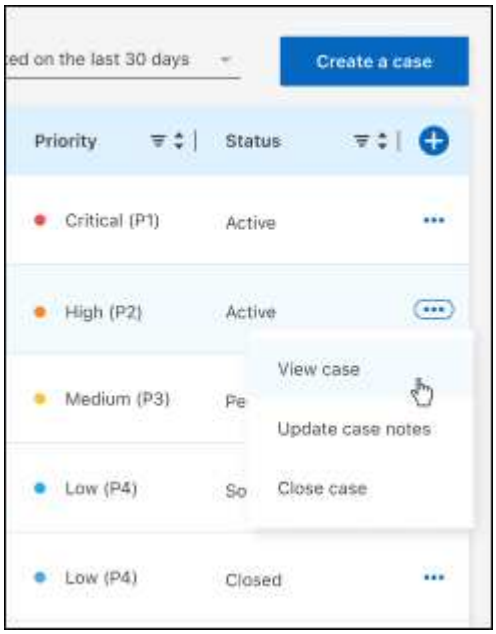
4. 사용 가능한 옵션 중 하나를 선택하고 선택하여 기존 케이스를 관리합니다.

◦ * 사례 보기 *: 특정 케이스에 대한 전체 세부 정보를 봅니다.

◦ * 케이스 메모 업데이트 *: 문제에 대한 추가 세부 정보를 제공하거나 * 파일 업로드 * 를 선택하여 최대 5개의 파일을 첨부할 수 있습니다.

첨부 파일은 파일당 25MB로 제한됩니다. txt, log, pdf, jpg/jpeg, rtf, DOC/docx, xls/xlsx 및 CSV.

◦ * 케이스 종료 *: 케이스를 종료하는 이유에 대한 세부 정보를 제공하고 * 케이스 닫기 * 를 선택합니다.



GenAI 법적 고지 사항을 위한 BlueXP 워크로드 공장

법적 고지 사항은 저작권 선언, 상표, 특허 등에 대한 액세스를 제공합니다.

저작권

"https://www.netapp.com/company/legal/copyright/"

상표

NetApp, NetApp 로고, NetApp 상표 페이지에 나열된 마크는 NetApp Inc.의 상표입니다. 기타 회사 및 제품 이름은 해당 소유자의 상표일 수 있습니다.

"https://www.netapp.com/company/legal/trademarks/"

특허

NetApp 소유 특허 목록은 다음 사이트에서 확인할 수 있습니다.

https://www.netapp.com/pdf.html?item=/media/11887-patentspage.pdf

개인 정보 보호 정책

"https://www.netapp.com/company/legal/privacy-policy/"

오픈 소스

통지 파일은 NetApp 소프트웨어에 사용된 타사의 저작권 및 라이센스에 대한 정보를 제공합니다.

"BlueXP 워크로드 공장"