

서비스 안내 | 공공의

2024-04-17

SAP AI 코어

최고의 달리기



콘텐츠

1	SAP AI Core란 무엇입니까?.....	6
2	SAP AI Core의 새로운 기능	9
2.1	2021 SAP AI Core(아카이브)의 새로운 기능.....	18
삼	개념.....	20
3.1	SAP AI 핵심 개요	20
	SAP AI 핵심 시스템 개요.....	21
	AI API 개요.....	24
	리소스 그룹.....	28
3.2	SAP AI Core 개요의 Generative AI Hub.....	30
3.3	용어.....	31
4	서비스 계획.....	35
4.1	프리 티어.....	37
	프리 티어 활성화.....	38
4.2	SAP AI Core 측정 및 가격.....	39
4.3	Generative AI Hub 측정 및 가격.....	40
4.4	리소스 계획 선택.....	41
	서비스 이용 보고.....	43
4.5	서비스 계획 업데이트.....	43
	서비스 계획 변경.....	43
5	초기 설정.....	45
5.1	Cloud Foundry에서 서비스 활성화.....	46
	하위 계정 만들기.....	47
	Cloud Foundry 활성화.....	48
	공간 만들기.....	50
	서비스 계획 추가.....	52
	서비스 인스턴스 생성.....	53
	서비스 키 생성.....	57
	서비스 키 사용.....	60
5.2	SAP AI Core Starter 튜토리얼.....	63
6	관리.....	64
6.1	Git 리포지토리 관리.....	64
	Git 리포지토리 추가.....	64
	폴더를 동기화하기 위한 애플리케이션 생성.....	67

Git 리포지토리 편집.....	70	
Git 리포지토리 삭제.....	71	
6.2 리소스 그룹 관리.....	71	
리소스 그룹 수준 리소스.....	72 리소스 그룹 생성.....	
73 리소스 그룹 편집.....	74	
리소스 그룹 삭제.....	75	
6.3 객체 저장소 비밀 관리.....	76	
객체 저장소 비밀번호 등록.....	76 객체 저장소 암호 편	
집.....	80 객체 저장소 암호 삭제.....	
84 6.4 Docker 레지스트리 비밀 관리.....	84	
Docker 레지스트리 비밀 등록.....	84	
Docker 레지스트리 비밀 편집.....	86	
Docker 레지스트리 비밀 삭제.....	88 6.5 일반 비밀 관	
리.....	88	
일반 비밀 만들기.....	88	
모든 일반 비밀 나열.....	90	
일반 비밀 업데이트.....	92	
일반 비밀 삭제.....	93	
실행 또는 배포 시 일반 암호 사용.....	93	
7	ML 작업.....	96
7.1	데이터 연결.....	96
	파일 관리.....	97
	데이터세트 API를 사용하여 파일 관리.....	101
7.2	모델 훈련.....	104
	리소스 계획 선택.....	106
	워크플로 템플릿.....	108
	목록 시나리오.....	112
	실행 파일 나열.....	115
	구성 생성.....	122 목록 구성.....
	티팩트 서명 사용.....	127 훈련 시
	작.....	131
	훈련 인스턴스 중지.....	134
	훈련 인스턴스 삭제.....	138
	효율성 기능.....	141 실행 로그 검색.....
	144 교육 일정.....	147
7.3	모델 사용.....	155
	리소스 계획 선택.....	156

템플릿 제공.....	158
실행 파일 나열.....	170
모델 배포.....	176 추론.....
179 배포 업데이트.....	180 배포 중지.....
183 배포 삭제.....	187 효율성 기능.....
배포 로그 검색.....	194
8 측정항목.....	198
8.1 AI API를 통한 지표 추적.....	198
측정항목 가져오기.....	199
지표 데이터 쿼리.....	207
측정항목 데이터 저장.....	209
측정항목 삭제.....	213
9 고급 기능.....	216
9.1 서비스형 AI 콘텐츠.....	216
서비스 사용자 정의 리소스.....	217
온보딩.....	219 오프보딩.....
221	
10 라이브러리 및 SDK.....	222
11 콘텐츠 패키지.....	224
12 SAP AI Core의 생성적 AI 허브.....	225
12.1 Generative AI Hub의 모델 및 시나리오.....	225
12.2 생성적 AI 모델을 위한 배포 생성.....	227
API 사용.....	227 모델 수명주기.....
231	
12.3 생성적 AI 모델 사용.....	231
콘텐츠 필터링.....	233 프롬프트 예.....
233 요약.....	234 추론.....
237	
변환.....	242
확장.....	248
챗봇.....	249
13 튜토리얼.....	253
14 보안.....	254

14.1 데이터, 데이터 흐름 및 프로세스의 보안 기능.....	254	14.2 전송 중 암호화	254
14.3 사용자 인증 및 관리.....	254	역할 및 권한.....	256
14.4 도커 이미지.....	259	AI 콘텐츠 보안.....	259
14.5 14.6 쿠버네티스 보안.....	260	14.7 구성 데이터 및 비 밀.....	260
14.8 출력 인코딩.....	260	14.9 멀티테넌 시.....	260
14.10 데이터 보호 및 개인정보 보호.....	261	261 14.10 데이터 보호 및 개인정보 보호.....	262
데이터 저장 및 처리.....	263	변경 로깅 및 읽기 액세스 로깅.....	263
동의.....	263		
삭제.....	263		
보안 및 고객 데이터 보호.....	264		
15 SAP AI Core의 접근성 기능.....	265		
16 모니터링 및 문제 해결.....	266	16.1 문제 해결.....	266
266 리포지토리.....	267	구성.....	268
유물.....	270		
애플리케이션.....	272		
실행.....	277		
도커.....	279		
전개.....	280		
여러 가지 잡다한.....	281		
17 서비스 오프보딩.....	284		

1 SAP AI Core란 무엇입니까?

SAP Business Technology Platform(SAP BTP)의 SAP AI Core 서비스에 대해 자세히 알아보세요.

SAP AI Core는 표준화되고 확장 가능하며 하이퍼스케일리에 구애받지 않는 방식으로 AI 자산의 실행 및 운영을 처리하도록 설계된 SAP Business Technology Platform의 서비스입니다. SAP 솔루션과의 원활한 통합을 제공합니다. 모든 AI 기능은 오픈 소스 프레임워크를 사용하여 쉽게 구현할 수 있습니다.

SAP AI Core는 AI 시나리오의 전체 수명주기 관리를 지원합니다. 생성적 AI 허브를 통해 생성적 AI 기능과 신속한 수명주기 관리에 액세스하세요.

SAP AI Core를 사용하면 데이터 기반 의사결정을 자신 있고 효율적으로 내릴 수 있으며 비즈니스 문제에 맞게 조정됩니다. 대용량 데이터를 처리하고 확장 가능한 기계 학습 기능을 제공하여 고객 피드백이나 티켓에 대한 분류 서비스, 분류 작업과 같은 작업을 자동화합니다. SAP AI Core는 사전 구성된 SAP 솔루션과 함께 제공되며 오픈 소스 기계 학습 프레임워크용으로 구성할 수 있고 Argo Workflow 및 KServe와 함께 사용할 수 있으며 다른 애플리케이션에 내장할 수 있습니다.

SAP AI Core를 사용하면 생성 AI 허브에서 다양한 생성 AI 모델로 자연어 프롬프트를 실험하고 활용할 수 있습니다.

팁

이 가이드의 영어 버전은 GitHub를 사용하여 기여하고 피드백을 받을 수 있도록 공개되어 있습니다. 이를 통해 SAP Help Portal 페이지의 담당 작성자 및 개발 팀과 연락하여 문서 관련 문제를 논의할 수 있습니다. 이 가이드에 참여하거나 피드백을 제공하려면 SAP Help Portal에서 해당 옵션을 선택하세요.

- ▶ **피드백 문제 생성 :** 문서 페이지에 대한 피드백을 제공합니다. 이 옵션은 GitHub에서 문제를 해결하세요.
- ▶ **피드백 편집 페이지 :** 문서 페이지에 기여합니다. 이 옵션은 GitHub에서 풀 요청을 업니다.

이러한 옵션을 사용하려면 GitHub 계정이 필요합니다.

추가 정보:

- [기부 지침](#)
- [소개영상](#)
- [소개 블로그 게시물](#)

환경

이 서비스는 다음 환경에서 실행됩니다.

- 클라우드 파운드리 • 키
- 마
- 쿠버네티스

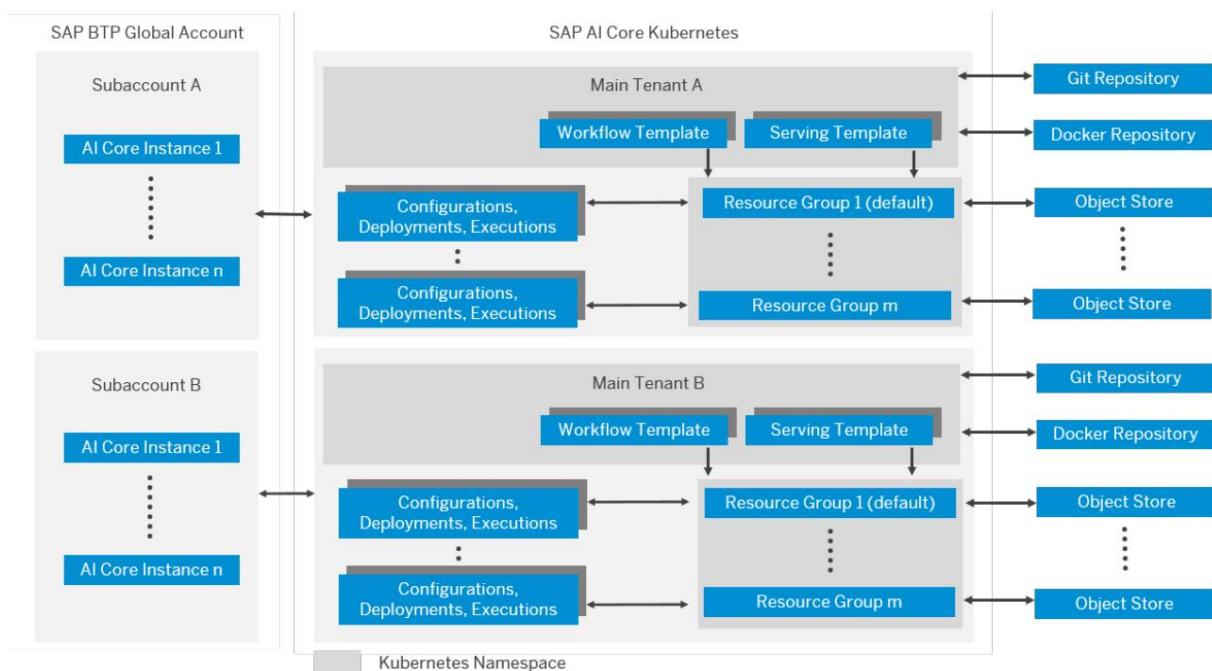
다중 테넌트

이 서비스는 다중 테넌트를 지원합니다. 테넌트 인식 애플리케이션에서 사용할 수 있습니다.

특징

파이프라인 실행	예를 들어 모델을 전처리 또는 교육하거나 일괄 추론을 수행하기 위해 파이프라인을 일괄 작업으로 실행합니다.
추론 제공 요청	훈련된 기계 학습 모델을 웹 서비스로 배포하여 훈련된 모델의 추론 요청을 고성능으로 처리합니다.
AI 시나리오 수명주기 관리	균일한 API 수명주기를 통해 모델 훈련, 측정항목 추적, 데이터, 모델, 모델 배포 등 ML 아티팩트 및 워크플로를 관리하세요.
혜택 다중 테넌트 지원	테넌트 인식 애플리케이션에서 이 서비스를 사용하세요. AI 자산과 실행을 분리하는 다중 테넌트 서비스를 구현하여 SAP AI Core 내에서 테넌트를 격리하세요.
클라우드 통합 하부 구조	Docker 레지스트리를 등록하고, git 리포지토리에서 AI 콘텐츠를 동기화하고, 훈련 데이터 및 훈련된 모델을 위한 객체 저장소를 등록하세요. AI 콘텐츠를 상품화하고 이를 SAP BTP 마켓플레이스의 소비자에게 서비스로 노출하세요.
생성적 AI 허브	신속한 실험과 신속한 수명주기 관리를 위해 다양한 생성 AI 모델 중에서 선택하세요.

SAP AI Core와 SAP AI Launchpad 간의 프로세스 흐름



관련 정보

[AI API 개요 \[페이지 24\]](#)

[SAP AI 런치파드](#)

2 SAP AI Core의 새로운 소식

							에 맞서	
							그들은	
기술							부시	연령-
니칼	환경						션	
와 함께	론-						네스 호	동쪽
후에-	남자들						프로세스	개정-
넷트	티	제목	설명	tion	클 유형	네스 호	에스 제품	시운
수액	못	생성	GCP 버전에서 선택된 모델	정보	Gen-	안-	기술	아니다
핵심	디	AI 허브	tex AI가 지원됩니다.	오직	예랄	명사	아니-	수액
	파운		자세한 내용은 Mod-를 참조하세요.		이익	이것-	AP-	사업
마른			일반 및 시나리오		-을 통해	남자들	더-	동쪽
			Active AI Hub [페이지 219].		이것	티	저것-	04-
수액	못	자원	최대 재수	정보	Gen-	안-	기술	아니다
핵심	디	그룹 제한	소스 그룹은 10개로 제한됩니다.	오직	예랄	명사	아니-	수액
	파운	테넌트당	개미 레벨이 500이 되었습니다. 이 레벨에 도달하면		이익	이것-	AP-	사업
마른			한도를 초과하면 오류가 발생합니다.		-을 통해	남자들	더-	동쪽
			메시지. 공간을 확보하려면		이것	티	저것-	04-
			일부 리소스 그룹을 설정해 보세요. 알-					04-
			대안적으로, 티켓을 올려-					04-
			할당량을 줄이세요.					04-
			자세한 내용은 <u>삭제</u>를 참조하세요.					04-
			리소스 그룹 [페이지 73].					04-
수액	못	진본인	X.509를 통한 인증	정보	Gen-	새로운 기술	아니다	수액
핵심	디	tion	인증서가 지원됩니다.	오직	예랄		AP-	사업
	파운		자세한 내용은 Cre-를 참조하세요.		이익	오기	더-	동쪽
마른			서비스 키를 먹은 경우 [페이지 57]		-을 통해		저것-	04-
			및 서비스 키 사용 [페이지 60].		이것		저것-	04-
수액	못	생성	생성 AI 허브 SDK는 다음과 같습니다.	정보	Gen-	새로운 기술	아니다	수액
핵심	디	AI 허브 SDK	이제 사용 가능합니다.	오직	예랄		AP-	사업
	파운		자세한 내용은 다음을 참조하세요.		이익	오기	더-	동쪽
마른			생성 AI 허브 SDK .		-을 통해		저것-	04-
					이것		저것-	04-

기술							에 맞서			
니칼	환경						그들은	부시	연령-	
와 함께	론-			삶-		~의	네스 호	동쪽	이익	
후에-	남자들			그리고-	싸이-	부시	프로세스	개정-	할 수 있는	
네트	E	제목	설명	tion	클 유형	네스 호	에스 제품	시온	현재	
수액	못	콘텐츠 파일-	카테고리에 해당하는 콘텐츠	정보	Gen-	새로운 기술	아니다	수액	2024년	2024년
핵심	디	~을 위해 노력하고 있다	증오, 성적, 자해의 잔혹함,	오직	에랄	아니-	AP-	사업	-02-1	-02-1
	파운	선택된	또는 폭력은 이제 다음과 같은 경우에 필터링됩니다.	이익	오기	더-	기술	플랫폼	8	8
핵심	마른	Azure 모드-	중간 또는 높은 수준으로 간주됩니다.		~을 통해		저것-			
		안으로	심각성. 심각도가 낮은 사례는 다음과 같습니다.		이것		~가 되었다			
		생성적인	필터링되지 않았습니다. 텔옥 위험은							
		AI 허브	또한 차단되었습니다.							
자세한 내용은 Azure 를 참조하세요.										
콘텐츠 필터링 문서 .										
tion ↗										
수액	못	생성	생성적 AI 허브에는 즉각적인 실행	정보	Gen-	새로운 기술	아니다	수액	2023년	2023년
핵심	디	AI 허브	이 포함됩니다.	오직	에랄	아니-	AP-	사업	-12-2	-12-2
	파운	신속한 관리 및 관리 도구. 신속한 실험에	신속한 관리 및 관리 도구. 신속한 실험에	이익	오기	더-	기술	플랫폼	0	0
핵심	마른	마든	는 생성이 포함됩니다.		~을 통해		저것-			
		자연어 실행			이것		~가 되었다			
		큰 선택을하라는 메시지가 표시됩니다.								
		언어 모델 및 매개변수. 프롬프트 관리에는								
		컬렉션, 메타데이터 형식의 프롬프트 저장								
		이 포함됩니다.								
태그 및 메모, 버전 관리 및										
삭제. 자세한 내용은,										
SAP AI의 Generative AI Hub 를 참조하세요 .										
핵심 개요 [페이지 30].										
수액	못	유물 서명-	아티팩트 서명(해시)은 다음을 수행할 수 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2023년	2023년
핵심	디	에 대한 본성	생성되어 이용 가능하게 됩니다.	오직	에랄	아니-	AP-	사업	-07-3	-07-3
	파운	아티팩트 아웃-	다른 처형 및 해제 가능	이익	오기	더-	기술	플랫폼	1	1
핵심	마른	에서 둔다	무결성 검증을 위한 노력		~을 통해		저것-			
		처형	유물의. 더 많은 정보를 원하시면-		이것		~가 되었다			
자세한 내용은 아티팩트 서명 사용 [페이지 125] 을 참조하십시오.										
수액	못	주형	워크플로를 생성하는 마법사	정보	Gen-	새로운 기술	아니다	수액	2023년	2023년
핵심	디	발전기	VS에서 템플릿 제공	오직	에랄	아니-	AP-	사업	-08-	-08-
	파운	암호. 사용자 응답을 사용하여		이익	오기	더-	기술	플랫폼	04	04
핵심	마른	단순화하고 자동화합니다.			~을 통해		저것-			
		템플릿 작성 과정.			이것		~가 되었다			
자세한 내용은 SAP 을 참조하세요.										
AI Core 툴킷 문서 .										

기술									
나칼	환경						에 맞서	그들은	
와 함께	론-				삶-	~의	부시	부시	연령-
후에-	남자들				그리고-	싸이-	부시	프로세스	동쪽 이익
네트	티	제목	설명		tion	클 유형	네스 호	에스 제품	개정- 할 수 있는
								시온	현재
수액	못	SAP AI 코어	SAP AI Core는 다음을 통해 제공됩니다.	정보	Gen-	새로운 기술	아니다	수액	2023년 2023년
핵심	디	툴킷	VS Code GUI를 통해	오직	예랄	아니-	AP-	사업	-08- 08-
	파운		SAP AI Core 툴킷 확장.	이익		오기	더-	기술 플랫폼	04 04
			자세한 내용은 SAP 을 참조하세요.		-을 통해		저것-		
			AI Core 툴킷 문서.		이것		-가 되었다		
수액	못	LLM 팩-	대형 콘텐츠 패키지	정보	Gen-	새로운 기술	아니다	수액	2023년 2023년
핵심	디	나이	SAP AI용 언어 모델	오직	예랄	아니-	AP-	사업	-05- 05-
	파운		Core는 배포를 단순화합니다.	이익		오기	더-	기술 플랫폼	31 31
			대규모 언어 모델의		-을 통해		저것-		
			통합되고 자동화된 작업 흐름.		이것		-가 되었다		
			자세한 내용은 PyPi 를 참조하세요.						
			법학대학원 .						
수액	못	API 데이터셋	업로드, 다운로드 등을 할 수 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2023년 2023년
핵심	디		SAP AI를 사용하여 아티팩트 삭제	오직	예랄	아니-	AP-	사업	-04- 05-
	파운		핵심 데이터 세트 API(직접적인 경우)	이익		오기	더-	기술	24 02
			객체 저장소의 파일에 액세스		-을 통해		저것-		
			가능하지도 바람직하지도 않습니다. 현재		이것		-가 되었다		
			Postman 및 컬 인터페이스가 지원됩니다.						
수액	못	메타데이터	엔드포인트를 사용하는 경우	정보	Gen-	차	기술	아니다	수액
핵심	디	응답	실행 파일을 나열하려면 다시	오직	예랄	nged	아니-	사업	-04- 04-
	파운	Exe를 나열하려면-	이제 스크립트 본문에 다음이 포함됩니다.	이익		오기	더-	기술	02 02
		마른	절단 가능 식품		매개변수에 대한 메타데이터-		-을 통해	플랫폼	
			에테르와 유돌. 아빠를 위해-		이것		저것-		
			레이미터, 설명				-가 되었다		
			기본값은 다시						
			돌린. 유물의 경우 종류,						
			설명 및 라벨						
			주석을 사용하여 추가할 수 있습니다.						
			자세한 내용은 목록 을 참조하세요.						
			실행 파일 [페이지 111], 작업						
			흐름 템플릿 [페이지 106]						
			템플릿 제공 [페이지 150].						

기술								에 맞서			
니칼	환경					그들은		선	부시	연령-	
와 함께	론-				삶-	~의	네스 호	동쪽	이익		
후에-	남자들				그리고-	싸이-	부시	프로세스	개정-	할 수 있는	
네트	티	제목	설명		tion	클 유형	네스 호	에스 제품	시온	현재	
수액	못	기여하다	SAP에서는 다음을 위해 노력하고 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2023년	2023년	
핵심	디	우리 문서에-	우리 문서가	오직	에랄	아니-	AP-	사업	-03-1	-03-1	
	파운	설명	당신을 위해 작동합니다. 당신이 그것을 느낀다면	이익	오기	더-	기술	기술			
	마른	원가 빠졌다거나 그런거		-을 통해		저것-	플랫폼	플랫폼	삼	삼	
		뭔가 좀 별로네		이것		-가 되었다					
이제 SAP Help Portal에서 직접 피드백을 제 공하고 변경 사항을 제안할 수 있습니다.											
너											
다음 두 가지 방법 중 하나로 그렇게 할 수 있습니다.											
• 도구 모음에서 편집을 클릭하여											
Gi-tHub에서 문서를 업니다. 여기											
에서 변경 사항을 제안하고											
우리가 검토하도록 요청을 가져옵니다.											
• 도구에서 피드백을 클릭하세요 .											
GitHub를 만드는 바는-											
고소하고 우리가 어떻게 할 수 있는지 말해줘											
문서를 개선하다											
당신을 위한.											
다음을 포함한 자세한 내용은											
기여 방법에 대한 지침,											
공개 문서 시작을 참조하세요 .											
tive.											
수액	못	대량 패치	실행 및 배포	정보	Gen-	새로운 기술	아니다	수액	2023년	2023년	
핵심	디	끝점	이제 PATCH 요청을 받을 수 있습니다.	오직	에랄	아니-	AP-	사업	-02-	-02-	
	파운	중지하거나	대량 조정을 위해 제공됨	이익	오기	더-	기술	기술	27	09	
	마른	삭제	대량 업데이트가 활성화되었습니다	-을 통해		저것-	플랫폼	플랫폼			
		다중 전-	관련 템플릿. 이상	이것		-가 되었다					
큐션 또는											
자세한 내용은 워크플로 온도를 참조하세요 .											
배포-											
접시 [페이지 104] 및 서빙											
말											
템플릿 [페이지 150].											
수액	못	동기화 종료-	게다가 자동으로	정보	Gen-	새로운 기술	아니다	수액	2023년	2023년	
핵심	디	Ar-에 대한 포인트	애플리케이션을 동기화하는 경우 다음을 사용하	오직	에랄	아니-	AP-	사업	-02-	-02-	
	파운	goCD	여 수동으로 동기화를 요청할 수 있습니다.	이익	오기	더-	기술	기술	27	09	
	마른		API 엔드포인트.	-을 통해		저것-	플랫폼	플랫폼			
			응용 프로그램 동기화에 대한 자세한 내용은	이것		-가 되었다					
다음 을 참조하십시오.											
풀더를 동기화하는 응용 프로그램											
[페이지 67].											

에 맞서									
기술		그들은							
니칼	환경	선	부시	연령-					
와 함께	론-	삶-	~의	네스 호					
후에-	남자들	그리고-	싸이-	부시					
네트	E	제목	설명	tion	클 유형	네스 호	에스 제품	시온	현재
수액	못	웹HDFS	이제 WebHDFS 아티팩트가 지원됩니다.	정보	Gen-	새로운 기술	아니다	수액	2023년
핵심	디	유물		오직	에랄	아니-	AP-	사업	-02-
	파운					이익	오기	기술	-02-
마른	마른		에 대한 자세한 내용은			-을 통해		플랫폼	27
			SAP AI의 WebHDFS 아티팩트						09
			Launchpad에 대한 자세한 내용은 자체 저장소			이것			
			암호 등록 [페이지 76]을 참조하십시오.						
수액	못	주기적	실행은 자동으로 실행될 수 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2023년
핵심	디	스케줄링	당연히 준비된 일정에 따라. 을 위한	오직	에랄	아니-	AP-	사업	-02-
	파운		자세한 내용은 교육을 참조하세요.			이익	오기	기술	-02-
마른	마른		일정 [페이지 140].			-을 통해		플랫폼	27
						이것			09
수액	못	배포-	이제 ttl pa-를 사용할 수 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2022년
핵심	디	드루-	지속 시간을 제한하는 매개변수	오직	에랄	아니-	AP-	사업	-11-2
	파운	그럴 수 있다	배포 시간을 몇 시간, 며칠 또는			이익	오기	기술	-11-2
마른	마른	제한된	주.			-을 통해		플랫폼	0
						이것			0
수액	못	YAML	관련 튜토리얼 단계에는	정보	Gen-	차	기술	아니다	수액
핵심	디	다음에 대한 파일	링크를 포함하도록 업데이트되었습니다.	오직	에랄	nged	아니-	사업	-11-2
	파운	스타터 투-	관련 파일에. YAML			이익	오기	기술	-11-2
마른	마른	토리얼	코드도 복사할 수 있고			-을 통해		플랫폼	0
			사용 가능			이것			0
			튜토리얼에서 직접 붙여넣기						
			곧장						
			기에서-						
			tHub						
수액	못	향상시키다-	GET 배포는 다시	정보	Gen-	차	기술	아니다	수액
핵심	디	~에 대한 언급	스폰서가 번호를 제공합니다	오직	에랄	nged	아니-	사업	-10-1
	파운	GET	최소 및 최대 및 실행 중			이익	오기	기술	-10-1
마른	마른	배포	복제본 및 리소스 계획			-을 통해		플랫폼	8
			NT API			이것			8
			이름. 자세한 내용은 다음을 참조하세요.						
			SAP AI Core API 사양 -						
			다시 전화해						
			후원						
			tion 						

기술									
수액	못	Azure 블룸	이제 Azure Blob을 등록할 수 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2022년
핵심	디파운드	저장	저장소 비밀 및 사용	오직	에랄	아니-	AP-	사업	2022년
	마른	지원됨	모델 제공을 위해.	이익	오기	더-	기술	기술	-10-1
			비밀 등록에 대한 자세한 내용은 등록을 참조하세요.	-을 통해		저것-	플랫폼	플랫폼	8
				이것		-가 되었다			8
개체 저장소 암호 [페이지 76]									
수액	못	프리 티어	SAP AI Core를 무료로 사용해 보려면	정보	Gen-	안-	기술	아니다	수액
핵심	디파운드	서비스	무료 계층 서비스 계획을 사용할 수 있습니다.	오직	에랄	명사	아니-	사업	사업
	마른	계획	무료 계층 서비스 계획은 다음과 같습니다.	이익	이것-	오기	더-	기술	기술
			쉽게 표준으로 업그레이드	-을 통해	남자들	저것-	플랫폼	플랫폼	-10-1
			계획을 세우고 사용자를 유지하며	이것	티	-가 되었다			8
			데이터.						8
자세한 내용은 무료를 참조하세요.									
계층 [페이지 37].									
수액	못	배포-	이제 기간을 제한할 수 있습니다.	정보	Gen-	새로운 기술	아니다	수액	2022년
핵심	디파운드	듀르-	지정하여 배포	오직	에랄	아니-	AP-	사업	2022년
	마른	다양한 변형-	실행해야 하는 시간	이익	오기	더-	기술	기술	-10-0
			전체 분 시간 또는	-을 통해		저것-	플랫폼	플랫폼	4
			날.	이것		-가 되었다			4
수액	못	측정항목 전-	AI API 지원 런타임이 있는 경우 다	정보	Gen-	새로운 기술	아니다	수액	2022년
핵심	디파운드	긴장	음을 사용할 수 있습니다.	오직	에랄	아니-	AP-	사업	2022년
	마른		새로운 "측정항목" 확장	이익	오기	더-	기술	기술	-09-1
			어떤 기능을 퀴리하려면	-을 통해		저것-	플랫폼	플랫폼	9
			메트릭 엔드 포인트가 지원됩니다. 측정항목 기	이것		-가 되었다			9
			능을 매우 세밀하게 지정할 수 있습니다.						9
세분성 수준을 통해									
그에 따라 클라이언트 구현에 반응합니다.									
자세한 내용은 Met-를 참조하세요.									
rics 확장.									
수액	못	Git 저장소 이	Git 리포지토리를 등록하면 더 이상 등록	정보	Gen-	안-	기술	아니다	수액
핵심	디파운드	름	할 필요가 없습니다.	오직	에랄	명사	아니-	사업	사업
	마른	더이상	저장소 이름을 입력하세요. 그만큼	이익	이것-	오기	더-	기술	기술
		필수적인	필드가 제거되었습니다. 이미 등록된 저장	-을 통해	남자들	저것-	플랫폼	플랫폼	-09-1
			소	이것	티	-가 되었다			9
			할당된 이름은 계속 작동합니다.						9

기술				에 맞서		
니칼	환경			그들은		
와 함께	론-			선	부시	연령-
후에-	남자들			삶-	~의	동쪽 이의
네트	E 제목	설명		그리고-	싸이-	프로세스
			tion	클 유형	네스 호	개정-
					에스 제품	할 수 있는
						시온
						현재

수액	못	API 지원 중단	API 엔드포인트 POST /lm/	답장-	Dep-	차	기술	아니다	수액	2022년	2022년
액	디	및							사업		
핵심	파운	해제-	구성/	한 청	답장-	nged	아니-	AP-	기술	-09-	-09-
	마른	시온	{구성 ID}/	디	케이트		오기	더-	플랫폼	05	05
			실행, POST /lm/								
			구성/								
			{구성 ID}/								
			배포, GET lm/								
			kpis 및 GET /analytics/								
			리소스 그룹 은								

더 이상 사용되지 않습니다. 업데이트하세요

기존 API 호출. 을 위한

시간표 및 더 많은 정보 -

자세한 내용은 SAP Note [3239609](#)를 참조하세요.

수액	못	피복재	업그레이드는 일부를 의미합니다.	답장-	Gen-	차	기술	아니다	수액	2022년	2022년
액	디	요소	제공 템플릿 변경사항:	한 청	에칼	nged	아니-	AP-	사업	-08-	-08-
핵심	파운	업그레이드됨	spec.template.api	디	이익		오기	더-	기술	22	22
	마른	KServe에게	버전 변경						플랫폼		
			버전 0.7								
			Serving.kubeflow.org/								
			보안을 위해								
			v1beta1 ~								
			규정 준수								
			Serving.kserve.io/								
			v1beta1,								
			spec.template.spec.pre								
			dictor.containers.name								
			kfserving 의 변경 사항 -								
			kserve 할 컨테이너 -								
			컨테이너. 저녁은 없어요-								
			apiVersion용 포트:								
			Serving.kubeflow.org/								
			v1alpha2.								
			기존 내용을 업데이트하세요.								
			9월까지 템플릿 제공								
			2022년 30일.								

에 맞서									
기술		그들은							
니칼	환경	선	부시	연령-					
와 함께	론-	삶-	~의	네스 호					
후에-	남자들	그리고-	싸이-	부시					
넷트	티	제목	설명	tion	클 유형	네스 호	에스 제품	시온	현재
수액	못	지원팀	SAP AI Core는 여러 기지를 지원합니다.	정보	Gen-	새로운 기술	아니다	수액	2022년 2022년
액자	디	물체	하이퍼스케일러 객체 저장소 등	오직	예랄	아니-	AP-	사업	-06- -06-
	파운	백화점	Amazon S3, OSS, HANA	이익	오기	더-	기술	기술	30 30
핵심	마른		데이터 레이크(HDL).	-을 통해		저것-	플랫폼		
비밀 등록에 대한 자세한 내용은 등록을 참조									
하세요.									
개체 저장소 암호 [페이지 76]									
수액	못	향상된	그만큼	정보	Gen-	새로운 기술	아니다	수액	2022년 2022년
액자	디	피복재	실행 파일.ai.sap.com	오직	예랄	아니-	AP-	사업	-06-1 -06-1
	파운	주형	/cascade-update-	이익	오기	더-	기술	기술	8 8
핵심	마른	매개변수	deployments 매개변수는 다음을 수행할 수 있습니다.	-을 통해		저것-	플랫폼		
개체 템플릿에 사용됩니다.									
연관된 배포를 자동으로 업데이트합니다.									
이상									
정보, 제공 템플릿									
및 개체 템플릿 변경									
및 배포 업데이트 .									
수액	못	추적	추적 성능은	정보	Gen-	안-	기술	아니다	수액
액자	디	서비스 임-	향상.	오직	예랄	명사	아니-	사업	-06-1 -06-1
	파운	증명		이익	이것-	오기	더-	기술	8 8
핵심	마른			-을 통해	남자들	저것-	플랫폼		
이것 티 -가 되었다									
수액	못	문서-	AI API CLI에 대한 문서	정보	Gen-	안-	기술	아니다	수액
액자	디	에 대한 설명	ent SDK가 PyPi.org로 이동되었습니다.	오직	예랄	명사	아니-	사업	-06- -06-
	파운	AI API	자세한 내용은 라이브러리	이익	이것-	오기	더-	기술	07 07
핵심	마른	클라이언트 SDK	그리고 SDK.	-을 통해	남자들	저것-	플랫폼		
이사했습니다									
PyPi.org에.									
수액	못	문서-	SAP 문서	정보	Gen-	안-	기술	아니다	수액
액자	디	에 대한 설명	AI Core SDK가 다음으로 이동되었습니다.	오직	예랄	명사	아니-	사업	-06- -06-
	파운	SAP AI	PyPi.org. 자세한 내용은,	이익	이것-	오기	더-	기술	07 07
핵심	마른	핵심 SDK	라이브러리 및 SDK.	-을 통해	남자들	저것-	플랫폼		
이사했습니다									
PyPi.org에.									

에 맞서									
기술		그들은							
니칼	환경	선	부시	연령-					
와 함께	론-	삶-	~의	네스 호					
후에-	남자들	그리고-	싸이-	부시					
넷트	E	제목	설명	tion	클 유형	네스 호	에스 제품	시온	현재
수액	못	추가의	\$ select 쿼리 매개변수	정보	Gen-	새로운 기술	아니다	수액	2022년 2022년
핵심	디	쿼리 파-	측정항목을 검색하는 데 사용할 수 있습니다.	오직	에랄	아니-	AP-	사업	-05- 05-
	파운	에 대한 평가자	측정항목과 같은 리소스 데이터	이익		오기	더-	기술	
핵심	마른	측정항목	또는 선택적으로 사용자 정의 정보, 을 위한 자세한 내용은 쿼리를 참조하세요.		-을 통해			플랫폼	05 05
			측정 데이터 .			이것			
수액	못	수평의	리소스 그룹 운영자	정보	Gen-	차	기술	아니다	수액
핵심	디	확장 대상	이제 수평 확장이 가능합니다.	오직	에랄	nged	아니-	사업	-04- 04-
	파운	자원	성능 향상.	이익		오기	더-	기술	09 09
핵심	마른	그룹 운영자			-을 통해			플랫폼	
		Ena-		이것					
		출혈							
수액	못	자원	리소스 그룹 ID 길이	정보	Gen-	차	기술	아니다	수액
핵심	디	그룹 ID	10자로 제한되어 있었는데,	오직	에랄	nged	아니-	사업	-04- 04-
	파운	길이 전-	그러나 이제 더 긴 ID가 지원됩니다. ID는 길이	이익		오기	더-	기술	09 09
핵심	마른	경향	가 길어야 합니다. 최소: 3, 최대: 253.		-을 통해			플랫폼	
			첫 번째 및 마지막 문자 소문자여야 합니다.	이것					
			ter, 대문자 또는 숫자. 다음의 문자 항목						
			두 번째에서 두 번째는 반드시 소문자이거나						
			대문자, 숫자, 마침표 또는 하이픈. 아니요						
			다른 특수 문자는 미트.						
수액	못	일반 Se-	이제 기본 테넌트 수준에서 일반 암호를 저	정보	Gen-	새로운 기술	아니다	수액	2022년 2022년
핵심	디	메인의 Cret	장할 수 있습니다.	오직	에랄	아니-	AP-	사업	-02-1 -02-1
	파운	지원	일반과 함께 사용하십시오.	이익		오기	더-	기술	9 9
핵심	마른	거주자	서비스 브로커. 더 많은 정보를 원하시면-		-을 통해			플랫폼	
		수준	Mation, 일반 세션 등록을 참조하세요.	이것					
			크레트.						

								에 맞서			
기술				그들은				연령-			
니칼	환경	선	부시	동쪽	이익						
와 함께	론-	삶-	~의	네스 호	동쪽	연령-					
후에-	남자들	그리고-	싸이-	부시	프로세스	개정-					
네트	E	제목	설명	tion	클 유형	네스 호	에스 제품	시온	현재		
수액	못	AI API	새로운 메타 API 엔드포인트를 사용하면	정보	Gen-	새로운 기술	아니다	수액	2022년	2022년	
핵심	디	메타 엔드-	클라이언트는 다음의 기능을 쿼리합니다.	오직	에랄	아니-	AP-	사업	-02-1	-02-1	
	파운	가리키다	구현하는 런타임 엔진	이익	오기	더-	기술	기술	9	9	할 수 있는
	마른		AI API. 메타 엔드 쿼리 -		-을 통해		플랫폼				
			포인트는 다음과 같은 정보를 반환합니다.		이것						
			클라이언트는								
			적절한 대응. 예를 들어, SAP AI Launchpad								
			는 다음을 활성화할 수 있습니다.								
			특정 기능을 활성화하거나 비활성화합니다.								
			의 능력을 바탕으로								
			런타임 엔진 구현								
			AI API의 자세한 내용은,								
			AI API 런타임 구현을 참조하세요 .								
			tions.								

2.1 2021 SAP AI Core(아카이브)의 새로운 소식

2021

기술-				환경				가능-			
니칼											
와 함께-	총-	론-									
후에-	능력	멘션 제목	설명					작업 유형			
네트											
SAP AI	확장-	구름	기전 후-	API 엔드포인트 /admin/resourceGroups 는 이전 버전이었습니다.		정보	새로운	2021-1			
핵심	시온	파운데이션-	단지-	AI Core API에서만 사용할 수 있습니다. 이제 다음에서 사용할 수 있습니다.		오직			2-18		
스위트 -	마른	마른	말	AI API이며 모든 런타임에서 구현될 수 있습니다.							
	디지털	예제		/ admin /resourceGroups 엔드포인트 사양							
경향예보세요-		만들어진		또한 CreateAt이라는 새로운 필드로 형상되었습니다 .							
	ence	CER		이는 리소스 그룹이 생성된 시기를 나타냅니다. 메모							
	그루			이는 아직 SAP AI Core 구현의 일부가 아닙니다.							
	PS API										
	끌-	끌-	자세한 내용은 AI API 사양을 참조하세요.								
	가리키다	가리키다		및 AI 코어 API.							

기술-	환경	설명	작업 유형	가능-
니칼	론-			블레
와 함께-	능력			-의
후에-	론-			
네트	멘션 제목	설명	작업 유형	가능-
SAP AI	확장-	구름	게-	정보
핵심	시온	파운데이션-	네릭	새로운 2021-1
	스위트 -	마른	세-	오직 2-03
	디지털		크레츠	
경험에보세요-	ence			
SAP AI	확장-	구름	AI API	정보
핵심	시온	파운데이션-	언에-	장 2021-1
	스위트 -	마른	우연-	에드 0-30
	디지털		말	
경험에보세요-	ence			

3가지 컨셉

이 섹션에서는 SAP AI Core와 관련된 몇 가지 개념을 살펴보겠습니다.

[SAP AI Core 개요 \[페이지 20\]](#)

SAP AI Core는 SAP 솔루션에 인공 지능 기능을 통합하는 핵심입니다.

[SAP AI Core 개요의 Generative AI Hub \[페이지 30\]](#)

생성적 AI 허브는 생성적 AI를 SAP AI Core 및 SAP AI Launchpad의 AI 활동에 통합합니다.

[용어 \[페이지 31\]](#)

3.1 SAP AI 코어 개요

SAP AI Core는 SAP 솔루션에 인공 지능 기능을 통합하는 핵심입니다.

특히 SAP AI Core는 다음을 수행하는 데 도움이 됩니다.

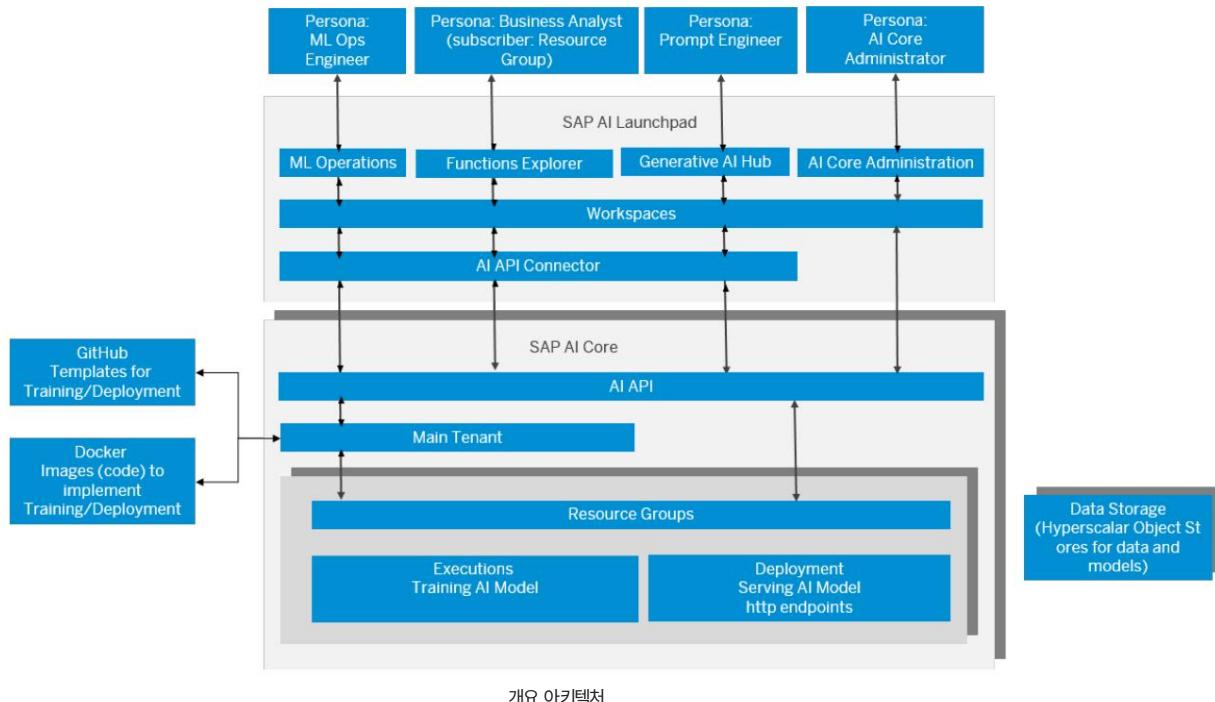
- AI 기능을 다른 애플리케이션에 원활하고 쉽게 내장
- 애플리케이션의 대용량 데이터를 활용하여 강력한 AI 학습 모델 생성
- 가속화된 하드웨어에서 AI 훈련 실행
- 비용 효율적인 방식으로 짧은 대기 시간과 높은 처리량으로 AI 추론을 제공합니다.
- 규정을 준수하고 설명 가능하며 유지 관리 가능한 프로세스를 준수합니다.
- 포괄적인 도구 및 서비스 세트를 사용하여 AI 수명주기의 모든 단계를 관리합니다. • AI 시나리오의 제품화 및 운영화에 중점을 둡니다.

AI 콘텐츠의 관리 및 운영(예: 버전 관리, 배포, 모니터링)은 SAP 솔루션 전반에 걸쳐 통합됩니다. 그러나 AI 콘텐츠 작성은 다양한 도구 세트(예: JupyterLab)에 열려 있습니다.

SAP AI Core 서비스는 SAP AI Launchpad 및 AI API와 함께 사용되도록 고안되었습니다. 주요 구성 요소는 SAP AI Core, SAP AI Launchpad 및 AI API입니다.

- SAP AI Core는 AI 워크플로우를 실행하고 워크로드를 제공하는 모델을 제공하는 엔진을 제공합니다.
 - SAP AI Launchpad는 다양한 AI 런타임을 관리합니다. 다양한 사용자 그룹이 접근하고 관리할 수 있도록 해줍니다. 그들의 AI 시나리오.
 - AI API는 다양한 런타임에서 AI 시나리오 수명주기를 관리하는 표준 방법을 제공합니다.
- SAP 기술(예: SAP S/4HANA) 또는 파트너 기술(예: Amazon Web Services)에서 제공되는지 여부. AI API가 SAP AI Core 이외의 런타임에 배포되는 경우 런타임은 런타임 어댑터를 제공해야 합니다.

[개요 아키텍처 \[페이지 21\]](#) 에서는 개요 아키텍처 다이어그램에 이러한 세 가지 주요 구성 요소를 보여줍니다.



상위 주제: [개념](#) [페이지 20]

관련 정보

[SAP AI Core 개요의 Generative AI Hub](#) [페이지 30]

[용어](#) [페이지 31]

3.1.1 SAP AI 핵심 시스템 개요

SAP AI Core 시스템은 내부 도구와 외부 도구를 연결합니다.

사용자는 SAP AI Core로 작업할 때 다양한 리포지토리, 시스템 및 개체와 상호 작용합니다. 이러한 개체 중 일부는 SAP에서 제공됩니다. 다른 경우에는 고객이 향상된 제어(권한 부여) 및 지속적인 통합/지속적인 배포(CI/CD)를 활성화하기 위해 이러한 구성 요소를 제공합니다.

주요 저장소, 시스템 및 개체

요소 및 용도 요약

무엇	왜
힘내 레포	교육을 저장하고 워크플로와 템플릿을 제공하기 위해
하이퍼스케일러 스토리지	학습 데이터 및 모델과 같은 입력 및 출력 아티팩트 저장용(예: SAP BTP 개체 저장소 서비스)
도커 저장소	템플릿에서 참조되는 사용자 지정 Docker 이미지의 경우
쿠버네티스(K8s)	K8s 클러스터는 AI 파이프라인에 사용되는 포드를 오케스트레이션하고 확장합니다. 리소스 그룹 격리는 K8s 네임스페이스를 기반으로 합니다.
K서빙(K)	기계 학습 모델의 최적화된 배포를 위한 것입니다. 배포 템플릿 사용 KServe 표기법.

AI API

학습 스크립트, 데이터, 모델 등 아티팩트 및 워크플로를 관리하기 위해 모델 서버)를 여러 런타임에 걸쳐

참고

AI API는 다른 기계 학습 플랫폼, 엔진을 통합하는 데에도 사용할 수 있습니다.
또는 AI 생태계에 대한 런타임입니다.

아르고 워크플로우

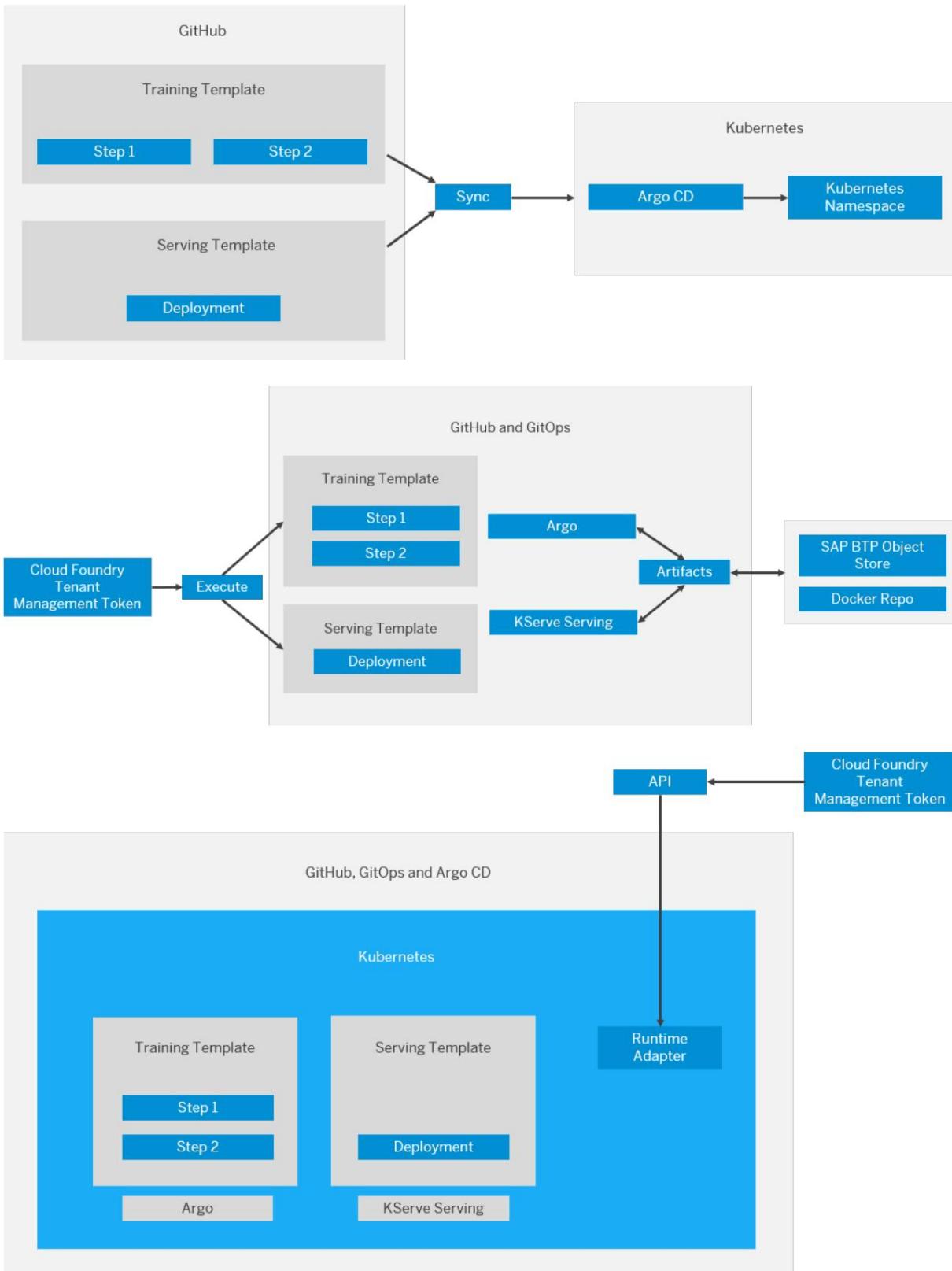
Kubernetes용 컨테이너 기반 워크플로 엔진입니다.

SAP AI 런치패드

SAP AI Launchpad는 SAP의 다중 테넌트 SaaS(Software as a Service) 애플리케이션입니다. 비즈니스 기술 플랫폼. 고객과 파트너는 SAP AI Launchpad를 사용하여 다음을 수행할 수 있습니다. AI 런타임(예: SAP 등)의 여러 인스턴스에서 AI 사용 사례(시나리오)를 관리합니다. AI 코어). SAP AI Launchpad는 Generative AI를 통해 생성 AI 기능도 제공합니다. 바퀴통.

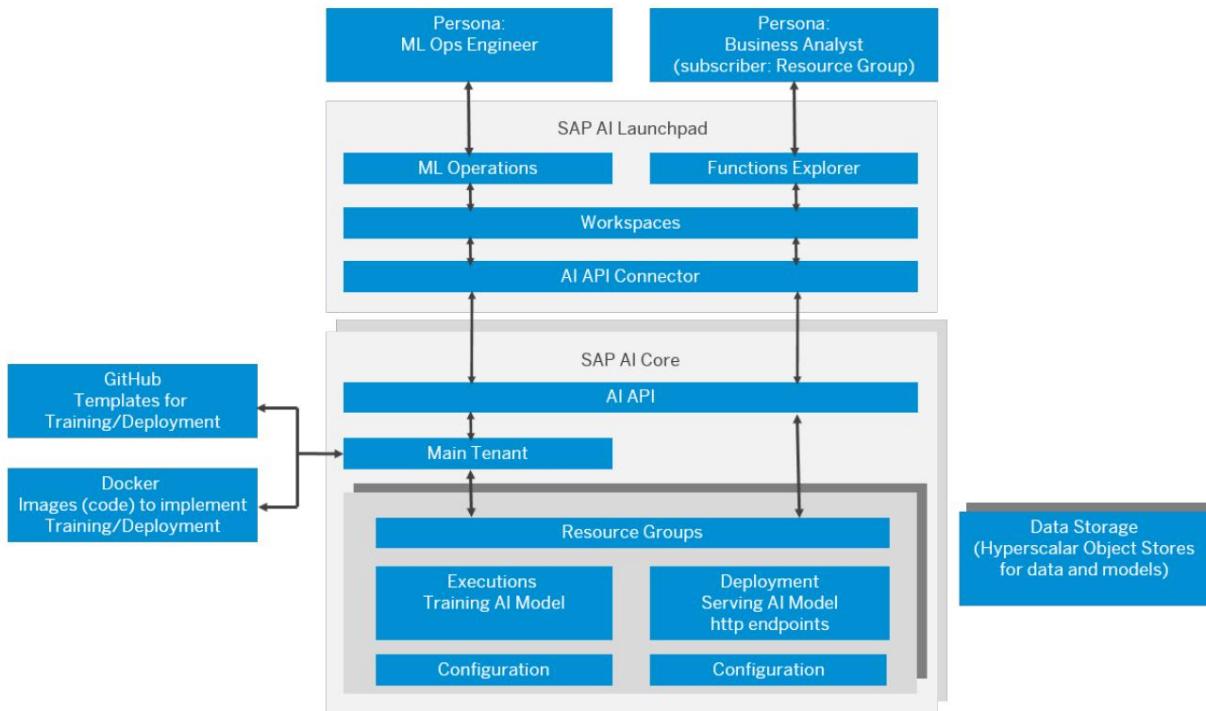
동기화

- GitOps 구현은 SAP AI Core와 통합되어 CI/CD를 활성화합니다.
- 템플릿은 Git 리포지토리에서 Kubernetes 클러스터로 동기화됩니다.
- 교육 및 제공 템플릿은 정기적으로(몇 분마다) 동기화됩니다.
- SAP AI Core는 템플릿 구문을 확인하고 동기화에 실패하면 오류 메시지가 표시됩니다.



프로세스 요약

- SAP BTP 토큰은 API 호출 인증을 위해 AI API에서 사용됩니다.
- 사용자는 AI API를 통해 템플릿을 실행합니다.
- 훈련 템플릿은 Argo 워크플로를 사용하여 실행됩니다. 학습 파이프라인은 데이터 입력을 소비합니다. 아티팩트를 생성하고 모델 아티팩트를 출력합니다.
- 훈련 템플릿과 AI API 구성은 사용하여 실행이 생성됩니다. 결과는 훈련 작업입니다. 제공 템플릿은 KServing을 사용하여 배포됩니다.
- 배포는 제공 템플릿과 AI API 구성은 사용하여 생성됩니다. 결과는 추론이다. 성기는 사람.
- 아티팩트(예: 데이터셋, 모델)는 하이퍼스케일러 스토리지에 복사됩니다.
- 필수 이미지는 등록된 Docker 저장소에서 가져옵니다.
- 시나리오 및 실행 파일에 대한 세부 정보는 AI API를 통해 Kubernetes 클러스터에서 검색됩니다.



3.1.2 AI API 개요

AI API를 사용하면 여러 런타임에 걸쳐 AI 자산(예: 교육 스크립트, 데이터, 모델, 모델 서버)을 관리할 수 있습니다.

Argo 워크플로와 제공 템플릿, 실행 및 배포는 AI API의 SAP AI Core 구현을 사용하여 관리됩니다. SAP AI Core에서는 Argo 워크플로우와 서빙 템플릿이 Executable이라는 개념으로 매핑됩니다. 매핑 메커니즘이 작동하려면 Argo 워크플로 및 제공 템플릿에 YAML 파일의 메타데이터 섹션에 특정 속성이 필요합니다. 이러한 속성은 두 템플릿 유형 모두에서 공유됩니다.

SAP AI Core는 런타임별 추가 API를 제공합니다. 이는 AI API 사양의 확장인 AI Core API 사양에서 사용할 수 있습니다.

관련 정보

[AI 코어 API](#) 
[AI API](#)

AI API 런타임 구현

AI API 사양은 기계 학습 아티팩트의 수명 주기 관리를 위한 일반 사양입니다.

SAP AI Core는 AI API 사양의 특정 런타임 구현 중 하나입니다. SAP AI Core와 별개로 AI API 사양의 다른 런타임 구현을 제공하는 것도 가능합니다. 이 섹션에서는 필요한 경계 조건과 구현 요구 사항을 설명합니다.

AI API를 사용하면 클라이언트가 모든 AI API 지원 런타임 구현과 통합할 수 있다는 이점이 있습니다. 예를 들어, SAP AI Launchpad는 동일한 API가 제공되는 한 사용자 정의 런타임 구현과 상호 작용할 수 있습니다. 지능형 시나리오 수명주기 관리는 AI API 지원 런타임과 통합될 수도 있습니다. AI API 클라이언트 SDK(Python)도 사용할 수 있습니다(자세한 내용은 [SAP AI Core SDK 참조](#)).

AI API 사양

AI API 사양은 다음 부분으로 구성됩니다.

- 주요사양
- 확장:
 - 분석 확장
 - 리소스 그룹 확장
 - 데이터셋 관리 확장
 - 측정항목 확장

추천

최소한 기본 사양을 구현한 다음 이를 기반으로 확장 사양을 구현합니다.

사용 사례.

AI API 런타임 기능 엔드포인트

Meta API는 AI API 사양(엔드포인트 /lm/meta)의 일부입니다. 구현은 AI API 런타임 구현의 기능을 지정하는 구성 응답을 반환해야 합니다.

Meta API를 사용하면 AI API 클라이언트가 AI API 구현 기능을 쿼리하여 사용 가능한 명령이나 사용자 인터페이스를 선택할 수 있습니다. 예를 들어 일부 AI API 런타임은 실행은 제공하지만 배포는 제공하지 않 을 수 있습니다. 배포가 아닌 실행에 대한 로그를 제공할 수도 있습니다. 예를 들어, SAP AI Launchpad와 같은 SAP AI Core 클라이언트가 SAP AI Core의 Meta API 엔드포인트를 쿼리하는 경우 응답은 다음과 같 습니다.

JSON

```
{
  "aiApi": {
    "capability": {
      "logs": {
        "deployments": true,
        "executions": true
      },
      "다중 테넌트": true,
      "shareable": true,
      "staticDeployments": true,
      "timeToLiveDeployments": true,
      "userDeployments": true,
      "userExecutions": true,
      "executionSchedules": true
    },
    "한계": {
      "배포": {
        "maxRunningCount": -1
      },
      "실행": {
        "maxRunningCount": -1
      },
      "minimumFrequencyHour": 1,
      "timeToLiveDeployments": {
        "최소": "10m",
        "최대": -1
      }
    }
  },
  "버전": "2.18.0"
},
"확장": {
  "분석": {
    "버전": "1.0.0"
  },
  "metrics": {
    "capability": {
      "extendedResults": true
    },
    "버전": "1.0.0"
  },
  "resourceGroups": {
    "버전": "1.2.0"
  }
},
"runtimeApiVersion": "2.21.0",
"runtimeldentifier": "aicore"
}
```

그런 다음 SAP AI Launchpad 및 기타 클라이언트는 그에 따라 반응하고 AI API의 런타임 구현을 위해 사용자 인터페이스에서 배포를 숨길 수 있습니다.

기능은 다음과 같습니다.

능력	true인 경우 사용자는 다음을 수행할 수 있습니다.
로그.실행	실행 로그 보기
로그.배포	배포 로그 보기
다중 테넌트	SAP AI Launchpad를 기본 테넌트 사용자로 사용(지원 리소스 그룹)
공유 가능	클라이언트는 하나의 인스턴스를 공유할 수 있습니다.
정적 배포	추론을 위해 항상 실행되는 정적 엔드포인트를 사용할 수 있습니다. 사용자가 배포를 시작하지 않고
사용자배포	배포 중지, 업데이트 또는 삭제
사용자 실행	실행 중지 또는 삭제
timeToLive배포	런타임 엔진을 사용하면 배포가 자동으로 삭제될 때까지의 시간을 정의할 수 있습니다.
해석학	모든 테넌트에 대한 요약 정보 검토
대량 업데이트	한 번에 최대 100개의 실행 또는 배포를 중지하거나 삭제합니다.
실행 일정	일정 만들기

한도에는 다음이 포함됩니다.

한계	세부
배포.maxRunningCount	실행 중인 동시 배포 수를 제한합니다. 리소스 그룹(있는 경우)
실행.maxRunningCount	동시에 실행 중인 실행 수를 제한합니다. 리소스 그룹(있는 경우)
timeToLiveDeployments.minimum	지원되는 경우 배포 시 ttl 매개변수에 가능한 최소값
timeToLiveDeployments.maximum	지원되는 경우 배포 시 ttl 매개변수에 가능한 최대값
최소 빈도시간	실행 일정에 대해 가능한 최소 값입니다. 지원됨

일반적인 AI API 사양 외에도 추가 사용을 포괄하는 여러 확장 기능도 있습니다.

사례. 이는 모든 런타임 엔진에서 구현되지 않을 수 있습니다.

확장은 다음과 같습니다:

확장	세부
해석학	분석 확장에는 리소스 그룹 또는 테넌트의 분석 정보를 가져오기 위 한 엔드포인트가 포함되어 있습니다.
측정항목	메트릭 확장에는 쓰기를 위한 엔드포인트가 포함되어 있습니다. 실행 중에 생성된 메트릭을 저장하고 검색하기 위해 메트릭 엔드포인트에서 읽기

확대**리소스 그룹****세부**

리소스 그룹 확장에는 리소스 그룹 관리를 위한 엔드포인트가 포함되어 있습니다.

데이터 세트

데이터 세트 확장에는 파일 업로드 및 다운로드를 위한 엔드포인트가 포함되어 있습니다.

관련 정보

[템플릿 제공 \[페이지 153\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[AI API 사양](#)

[Meta API를 사용한 맞춤형 런타임 기능](#)

분석 확장

[리소스 그룹 확장](#)

[지능형 시나리오 수명주기 관리](#)

3.1.3 리소스 그룹

SAP AI Core 테넌트는 리소스 그룹을 사용하여 관련 ML 리소스 및 워크로드를 격리합니다. 시나리오, 실행 파일 및 Docker 레지스트리 비밀은 모든 리소스 그룹에서 공유됩니다.

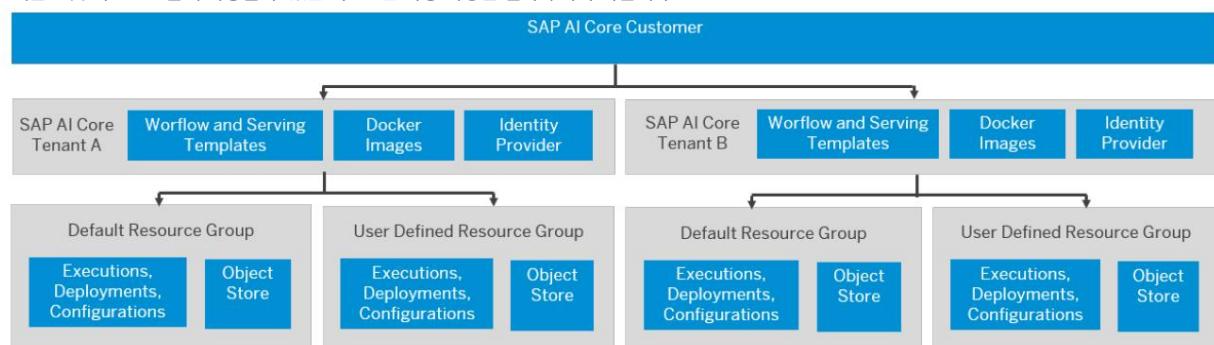
리소스 그룹은 하나의 SAP AI Core 테넌트 범위 내에서 관련 리소스의 가상 컬렉션을 나타냅니다.

테넌트가 등록되면 기본 리소스 그룹이 즉시 생성됩니다. 테넌트 관리자는 AI API를 사용하여 추가 리소스 그룹을 생성하거나 삭제할 수 있습니다. 테넌트는 해당 사용 시나리오에 따라 리소스 그룹을 매핑할 수 있습니다.

SAP AI Core 테넌트가 리소스 그룹을 사용하여 시나리오 소비자 테넌트를 격리하고 나중에 리소스 그룹이 삭제되면 시나리오 소비자의 프로비저닝이 해제됩니다. SAP AI Core는 테넌트의 시나리오 소비자를 인식하지 못합니다. 표준 XUSAA 다중 테넌시 모델을 따릅니다.

3.1.3.1 자원의 범위

테넌트 및 리소스 그룹에 사용할 수 있는 리소스는 사용 가능한 범위에 따라 다릅니다.



테넌트 수준 리소스

테넌트 수준 리소스에는 다음이 포함됩니다.

- 워크플로 템플릿 • 템플

릿 제공

- Docker 레지스트리(Docker 이미지 포함)
- 사용자 인증 및 승인(UAA)

사용자 인증 및 권한 부여는 SAP AI Core 테넌트를 기반으로 합니다. 임차인은 소유권을 가지고 있는 사람입니다.

SAP AI Core 서비스 키를 사용하여 얻은 액세스 토큰입니다. SAP AI Core 테넌트는 AI API를 사용하여 런타임 시 또는 수명 주기 관리 중에 요청 헤더에 리소스 그룹을 설정할 수 있습니다. 리소스 그룹이 설정되지 않은 경우 기본 리소스 그룹이 사용됩니다.

리소스 그룹 수준 리소스

테넌트 수준의 실행 파일은 모든 리소스 그룹에서 공유됩니다. 리소스 그룹 수준에서는 리소스 그룹 헤더를 설정하여 개체 저장소를 등록합니다.

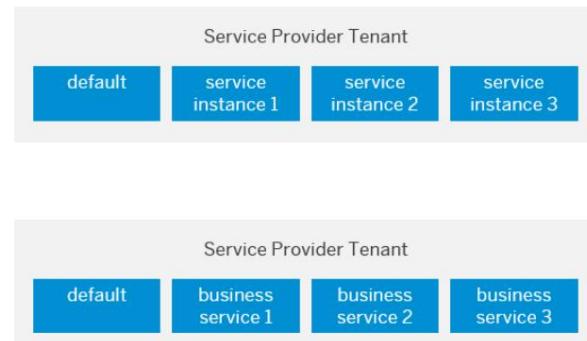
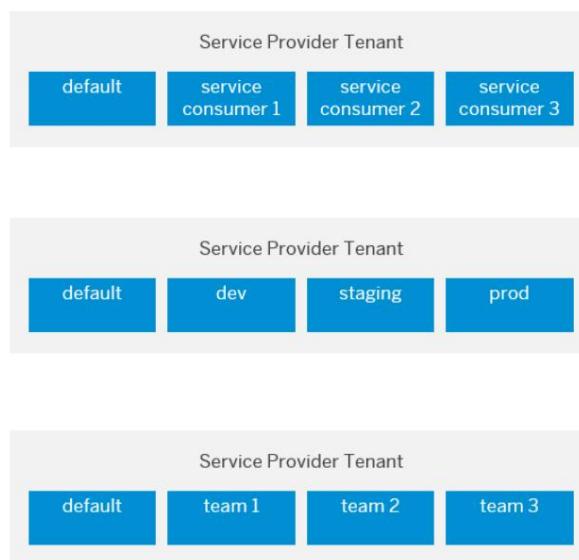
SAP AI Core 테넌트는 AI 기능 설계 시 보안 측면을 고려해야 합니다.

추천

여러 리소스 그룹에 대해 동일한 AWS IAM 사용자와 동일한 객체 저장소 버킷을 사용하지 마십시오.

실행, 배포, 구성 및 아티팩트와 같은 런타임 엔터티는 특정 리소스 그룹에 속하며 리소스 그룹 간에 공유될 수 없습니다.

리소스 그룹 매핑의 예



3.2 SAP AI Core의 생성적 AI 허브 개요

생성적 AI 허브는 생성적 AI를 SAP AI Core 및 SAP AI Launchpad의 AI 활동에 통합합니다.

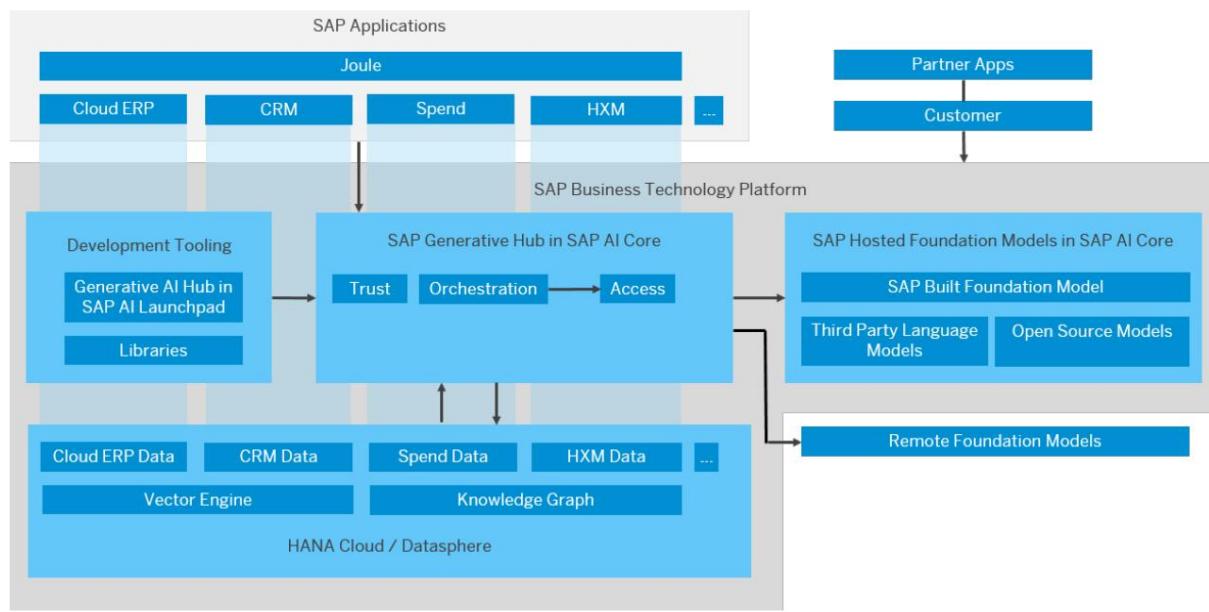
LLM은 레이블이 지정되지 않은 방대한 양의 데이터에 대해 교육을 받은 자기 지도형 딥 러닝 모델입니다.

AI 기술과 산업 규모의 컴퓨팅 리소스를 활용하여 자연어 처리(NLP) 작업을 위한 복잡한 언어 패턴과 의미론적 지식 기반을 학습합니다. 프롬프트와 같은 입력을 구문 분석하고 대상 단어를 예측하여 자연어로 작성된 상황에 맞는 응답을 반환할 수 있습니다. 단일 LLM은 다양한 입력 형식과 출력 모드를 사용하여 여러 NLP 작업을 수행할 수 있습니다.

LLM은 일반 모델이지만 전문화 또는 도메인별 추가 임베딩을 통해 미세 조정할 수 있습니다.

사용 사례.

SAP AI Core와 생성적 AI 허브는 LLM과 AI를 비용 효율적인 방식으로 새로운 비즈니스 프로세스에 통합하는 데 도움이 됩니다.



생성적 AI 허브 아키텍처 개요

상위 주제: [개념](#) [페이지 20]

관련 정보

[SAP AI Core 개요](#) [페이지 20]

[용어](#) [페이지 31]

3.3 용어

용어	정의
AI API	모든 AI 아티팩트 및 워크플로우를 처리하는 필수 API (예: 학습 스크립트, 데이터, 모델 및 모델 서버) 여러 AI 런타임에 걸쳐.
인공물	다음에 의해 생성되거나 소비되는 데이터 또는 파일에 대한 참조입니다. 실행 또는 배포.
구성	실행 또는 배포를 실행하는 데 사용되는 매개변수, 아티팩트 참조 및 실행 파일의 모음입니다. ↗ 실행을 인스턴스화하려면 구성이 필요합니다. 실행 파일과 구성은 1:n 관계(즉, 실행 파일에는 다양한 구성이 있을 수 있습니다).
데이터세트	아티팩트의 일종. 데이터세트는 데이터 소스와 이에 접근하고 조사하는 데 필요한 도구. 데이터세트: <ul style="list-style-type: none">• 데이터 평면에서 획득한 데이터 세트를 참조합니다.애플리케이션 API• 시나리오에 데이터 세트 아티팩트로 전달될 수 있습니다. 실행 시간
전개	배포는 모델 제공 템플릿의 인스턴스입니다. 모델 유형 아티팩트를 사용하고 적용하도록 구성된 제공 요청에 전달된 데이터에 적용됩니다(예: 예측). 배포가 성공적으로 인스턴스화되면 모델 서버가 생성되고 배포 URL이 생성됩니다. 추론을 위해. 입력으로 배포에는 하나 이상의 작업이 필요합니다. 구성의 모델 및 매개변수.
실행 가능	실행 파일은 목적을 위해 인스턴스화되는 템플릿입니다. 모델 훈련 또는 배포 생성과 같은 포즈 사용자 정의 레이블은 실행 파일에 적용될 수 있으며 실행 파일의 세부 정보와 함께 나열됩니다. • 배포할 수 없는 실행 파일이 인스턴스화되면 결과는 다음과 같습니다. 체형 중. 실행 시 출력 결과물이 생성될 수 있습니다. 사리 • 배포 가능한 실행 파일이 인스턴스화되면 다음과 같은 결과가 발생합니다. 전개. 배포는 다음이 가능한 URL을 생성합니다. 추론에 사용됨

용어

정의

실행

실행은 배포할 수 없는 실행 파일의 인스턴스입니다.

실행은 모델을 훈련하거나 다른 유형의 작업을 실행하는 데 사용됩니다.

작업 흐름.

실행은 한 번만 실행되고 출력 아티팩트를 생성합니다. 훈련 인스턴스가 실행되면 출력 아티팩트

생산된 모델입니다.

일반적으로 관련된 메타데이터의 양이 많습니다.

처형으로, 메타데이터는 측정항목으로 저장됩니다(및

관련 라벨), 태그, 맞춤 정보 등이 있습니다. 여러 지표,

태그와 사용자 정의 정보는 한 번의 실행으로 전달될 수 있습니다.

측정항목, 태그, 맞춤 정보가 저장되면

<실행 ID>로 큐리했습니다.

모든 지표, 태그 및 사용자 정의 정보는 다음과 연결되어야 합니다.

처형.

워크플로 실행 파일

워크플로 실행 파일은 [ML Operations](#) 앱 내에서 사용됩니다.

습곡.

워크플로 실행 파일은 일련의 매개변수와

인스턴스화해야 하는 입력 아티팩트입니다. 이것들은

구성으로 제공됩니다. 워크플로 실행 파일도

다음과 같은 경우 생성될 출력 아티팩트 세트를 정의합니다.

실행 파일이 인스턴스화되고 성공적으로 실행되었습니다.

작업 실행 가능

작업 실행 파일은 워크플로를 단순화한 표현입니다.

[Functions Explorer](#) 애플리케이션 내에서 사용되는 실행 파일입니다.

작업 템플릿

장기 실행을 지정하는 SAP AI Core의 실행 파일 유형입니다.

일반적으로 학습 또는 일괄 추론을 수행하는 프로세스입니다.

모델

학습의 결과인 아티팩트 유형입니다.

모델 제공 템플릿

실행 방법을 지정하는 SAP AI Core의 실행 파일 유형입니다.

모델이 서비스될 예정입니다.

운영

훈련 실행, 모델 배포 생성, 애플리케이션 통합 등 AI 수명주기 내의 모든 활동

그리고 모델 모니터링.

용어

정의

리소스 그룹

리소스 그룹은 생성된 고유한 작업 영역을 나타냅니다.

SAP AI Core 시나리오 소비자 및 고유한

목적. 리소스 그룹에는 유효한 구독이 필요합니다.

거주자.

리소스 그룹 내에서 사용자는 다음과 같은 엔터티를 생성하거나 추가합니다.

구성, 실행, 배포 및 아티팩트. 그들

시나리오에서 사용할 수 있는 실행 파일을 사용합니다.

사용자가 실행 파일을 인스턴스화하려면 먼저 다음을 생성해야 합니다.
구성.

- 배포할 수 없는 실행 파일이 인스턴스화되면 결과는 다음과 같습니다.
처형 중. 실행은 선택적으로 출력을 생성합니다.
유물.
- 배포 가능한 실행 파일이 인스턴스화되면 다음과 같은 결과가 발생합니다.
전개. 배포는 다음이 가능한 URL을 생성합니다.
추론에 사용됩니다.

실행 시간

AI 실행을 위한 처리 자원을 제공하는 플랫폼

훈련, 추론과 같은 학습 부하.

SAP AI 코어

고객의 비즈니스 애플리케이션을 가능하게 하는 인프라

AI와 머신러닝으로 지능화되는 양이온

AI 서비스를 교육하기 위한 기술 및 데이터

작업과 프로세스를 연결합니다.

대본

시나리오는 사용 사례에 대한 관련 실행 파일 그룹입니다.

사용자의 테넌트 내에서. 시나리오에는 다양한 버전에 추가로 대응하는 여러 버전

이 있을 수 있습니다.

실행 파일.

실행 파일 제공

제공 실행 파일은 배포 가능한 템플릿 파일프라인입니다.

인스턴스화하여 배포를 생성할 수 있습니다.

제공 실행 파일은 매개변수 집합을 정의하고

인스턴스화해야 하는 입력 아티팩트 이것들은

구성으로 제공됩니다.

훈련

(적어도) 아티팩트를 생성하는 실행 또는 실행 파일입니다.

유형 모델의 사실 .

즉각적인

프롬프트는 생성 AI를 위한 자연어 지침입니다.

모델.

용어

생성적 AI 허브

정의

생성적 AI 허브에는 즉각적인 실험이 포함됩니다.
신속한 관리 및 관리 도구. 신속한 실험에는 다양한 언어 모델을 선택하여 자연스러운 언어 프롬프트를 만들고 실행하는 것이 포함됩니다.

매개변수. 프롬프트 관리에는 프롬프트 저장이 포함됩니다.
컬렉션, 태그 및 메모 형태의 메타데이터,
버전 관리 및 삭제.

상위 주제: [개념](#) [페이지 20]

관련 정보

[SAP AI Core 개요](#) [페이지 20]

[SAP AI Core 개요의 Generative AI Hub](#) [페이지 30]

4가지 서비스 계획

선택한 SAP AI Core 서비스 계획에 따라 가격, 사용 조건, 리소스, 사용 가능한 서비스 및 호스트가 결정됩니다.

사용 사례에 따라 SAP AI Core에 대한 무료 등급 옵션을 선택할지 유료 서비스 계획을 선택할지 결정됩니다.

다음과 같은 서비스 계획을 사용할 수 있습니다.

- 무료
- 표준
- 확장됨

제한사항

SAP AI Core 무료 계층은 SAP BTP 평가판이 아닌 SAP BTP 무료 계층에서만 사용할 수 있습니다.

서비스 계획	세부	계정 유형
무료	<ul style="list-style-type: none"> • 제한적인 무료 서비스를 사용하여 SAP AI Core에 대해 알아보세요. • 무료 계층 서비스 계획에는 커뮤니티 지원만 제공되며 이는 SLA의 적용을 받지 않습니다. • 제한사항: <ul style="list-style-type: none"> • 글로벌 계정당 하나의 인스턴스로 제한됩니다. • 특정 시점에서 실행되는 하나의 실행 또는 배포로 제한됩니다. 시간 <ul style="list-style-type: none"> • 기본 리소스 그룹으로 제한됨 • Starter AI Core 리소스 계획으로 제한됨 • 무료 또는 표준 계획의 활성 인스턴스로 제한됨 (mu-) <ul style="list-style-type: none"> 실제로는 독점적임) 하위 계정 내 • 생성적 AI 허브는 포함되지 않음 • 지역 정보는 SAP Discovery Center를 참조하세요.. 	기업

참고

고객은 하위 계정 내에 활성 표준 요금제 인스턴스가 없는 경우에만 하나의 SAP AI Core 무료 요금제 인스턴스를 생성할 수 있습니다.
테넌트는 요금제를 무료에서 표준으로 업데이트할 수 있지만 표준에서 무료로 업데이트할 수는 없습니다. 하위 계정에 활성 무료 요금제 인스턴스가 있는 경우 표준 요금제 SAP AI Core 인스턴스를 생성할 수 없습니다.

자세한 내용은 [프리 티어 \[페이지 37\]](#), [리소스 계획 선택 \[페이지 104\]](#)의 스타터 계획 및 [무료 서비스 계획 사용을 참조하십시오.](#)

서비스 계획	세부	계정 유형
기준	<ul style="list-style-type: none"> 생산적인 사용을 위한 서비스 계획. 맞춤형 워크로드에 대한 리소스 사용 요금은 서비스 사용에 대한 기본 요금을 추가하여 사용된 리소스 유형에 따라 조정 가능한 가격으로 청구됩니다. <p>• 제한사항:</p> <ul style="list-style-type: none"> 활성 무료 요금제 인스턴스가 있는 경우 SAP AI Core 표준 요금제 인스턴스를 생성할 수 없습니다. 생성적 AI 허브는 포함되지 않음 <p>• 지역 정보는 SAP Discovery Center를 참조하세요.. </p> <p>자세한 내용은 리소스 계획 선택 [페이지 104] 및 SAP AI Core 측정 및 가격 책정 [페이지 39]을 참조하십시오.</p>	기업
펼친	<ul style="list-style-type: none"> 생성적 AI 기능이 추가된 표준 계획이 포함됩니다. 생성적 AI 허브의 특성. 제공된 생성 AI 모델 사용에는 적응형 AI 모델을 사용하여 요금이 부과됩니다. 모델 선택 및 전송된 토큰 수에 따른 가격 그리고 받았습니다. <p>• 지역 정보는 SAP Discovery Center를 참조하세요.. </p> <p>자세한 내용은 리소스 계획 선택 [페이지 104] 및 SAP AI Core 측정 및 가격 책정 [페이지 39]을 참조하십시오.</p>	기업

배포 할당량

각 테넌트에는 배포 수와 배포당 복제본 수를 제한하는 기본 할당량이 할당됩니다. 이 할당량에 도달하면 배포가 생성되지 않으며 이에 따라 알림이 전송됩니다. 기존 배포를 삭제하여 할당량을 확보할 수 있습니다.

또는 CA-ML-AIC 구성 요소에 대한 티켓을 생성하여 할당량 증가를 요청할 수 있습니다. 할당량 증가 요청 설명을 입력하고 증가 크기, 배포, 복제본 또는 둘 다를 포함할지 여부 및 하위 계정 ID에 대한 세부 정보를 포함합니다.

리소스 그룹 할당량

최대 리소스 그룹 수는 테넌트 수준에서 50으로 제한됩니다. 이 제한에 도달하면 오류 메시지가 표시됩니다. 공간을 확보하려면 일부 리소스 그룹을 삭제하세요. 또는 할당량을 늘리려면 티켓을 제출하세요.

관련 정보

[무료 계층 \[페이지 37\]](#)

[SAP 디스커버리 센터](#)



[SAP BTP 서비스 설명 가이드](#)



[리소스 계획 선택 \[페이지 104\]](#)

4.1 무료 등급

무료 등급 옵션을 활성화하여 SAP AI Core의 사용량 제한을 알아보고 서비스에 익숙해지세요.

이 옵션을 사용하면 사용이 제한됩니다.

무료 등급 옵션을 사용하면 다음과 같은 이점을 누릴 수 있습니다.

- SAP AI Core 기능을 무료로 살펴보세요.
- 작업 내용을 유지하면서 무료 계층에서 표준 서비스 계획으로 마이그레이션합니다.

무료 등급 범위

- 지역 정보는 [SAP Discovery 센터](#)를 참조하세요. . . • 무료 계층 서비스 계획에
는 커뮤니티 지원만 제공되며 SLA가 적용되지 않습니다.
- 무료 등급 옵션에는 사용량 제한이 적용됩니다. 자세한 내용은 [서비스 계획 \[페이지 35\]](#)를 참조하십시오.

참고

고객은 하위 계정 내에 활성 표준 요금제 인스턴스가 없는 경우에만 하나의 SAP AI Core 무료 요금제 인스턴스를 생성할 수 있습니다. 테넌트는 요금제를 무료에서 표준으로 업데이트할 수 있지만 표준에서 무료로 업데이트할 수는 없습니다. 하위 계정에 활성 무료 요금제 인스턴스가 있는 경우 표준 요금제 SAP AI Core 인스턴스를 생성할 수 없습니다.

관련 정보

[서비스 계획 \[페이지 35\]](#)

프리 티어 활성화

평가판 및 테스트 목적으로 SAP AI Core를 경험하려면 무료 계층을 활성화하세요. 또는 [여기의 튜토리얼을 따르십시오.](#)

전제조건

제한사항

- SAP AI Core 무료 계층은 SAP BTP 평가판이 아닌 SAP BTP 무료 계층에서만 사용할 수 있습니다.
- 생성적 AI 하브는 무료 등급에서 사용할 수 없습니다. 자세한 내용은 [서비스 계획 \[페이지 35\]](#)를 참조하십시오.

절차

- SAP BTP 조종석에서 글로벌 계정을 개설하세요.
- 하위 계정으로 이동합니다.
- 팀색 영역에서 [인스턴스 및 구독을 선택합니다](#).
- [새 인스턴스 또는 구독을 선택하세요](#).
- [서비스 필드](#)에서 SAP AI Core를 검색하세요 .
- [계획 필드](#)에서 [무료를 선택합니다](#) .

New Instance or Subscription

1 Basic Info 2 Parameters 3 Review

Enter basic info for your instance or subscription.

Service: * [SAP AI Core](#)

Plan: *

standard	Instance
free	Instance

참고

무료 테넌트가 생성됩니다.

관련 정보

[서비스 계획 \[페이지 35\]](#)

[무료 서비스 플랜 사용](#)

[글로벌 계정 얻기](#)

4.2 SAP AI Core 측정 및 가격

SAP AI Core는 리소스에 따라 다양한 청구 불가능한 측정 단위(UoM)를 기준으로 측정됩니다.
SAP AI Core가 소비됩니다.

다음 자원과 청구 불능 단위를 사용할 수 있습니다.

리소스 유형	인프라 자원	측정 단위
컴퓨팅	스타터 인스턴스	노드 시간
	기본 인스턴스	
	기본.8x 인스턴스	
	추론-S 인스턴스	
	추론-M 인스턴스	
	추론-L 인스턴스	
	Train-L 인스턴스	
저장	표준 SSD 블룸	기가바이트 시간
기준선	고정 클러스터 리소스(인스턴스, 스토리지, SAP HANA, Kubernetes, De-vOps, 지원 등)	입주시간

참고

인프라 사양은 선택한 하이퍼스케일러에 따라 다릅니다. 자세한 내용은 [선택을 참조하세요](#).

[리소스 계획 \[페이지 102\]](#).

각 측정 기록은 청구 가능한 측정 단위 "용량 단위"로 변환됩니다. 모든 리소스에는 특정 단위를 용량 단위로 변환하기 위한 변환율입니다. 적용 가능한 환율은 [SAP](#)을 참조하세요.

[BTP 서비스 설명 가이드](#).

모든 "컴퓨팅" 및 "스토리지" 리소스 소비는 실제 측정된 사용량을 기준으로 합니다.

기본 요금은 다음을 기준으로 자동으로 계산됩니다.

"컴퓨팅" 또는 "스토리지" 리소스가 테넌트 내에서 소비되고 있는 경우 "기준" 테넌트 시간은 1시간입니다.

자동으로 충전됩니다. 동일한 시간 내에 여러 리소스를 사용하는 경우 "기준" 테넌트 시간당 요금이 청구됩니다.

예

이 예에서는 가상의 값을 사용합니다. 현재 환율은 [SAP BTP 서비스 설명 가이드](#)를 참조하세요. .



다음 매개변수를 가정해 보겠습니다.

- 스타터 인스턴스의 1노드 시간은 0.5 용량 단위와 같습니다.
- Infer-M 인스턴스의 1노드 시간은 2용량 단위와 같습니다.
- 기준의 1 테넌트 시간은 1.5 용량 단위와 같습니다. • 1개 용량 단위 비용은 EUR 1입니다.

스타터 인스턴스 100시간과 Infer-M 인스턴스 300시간을 사용하고 한 달에 300시간 동안 클러스터에서 활성 상태인 경우 다음 비용이 발생합니다.

스타터 인스턴스의 100노드 시간 = $100 * 0.5$ 용량 단위 = 50용량 단위 Infer-M 인스턴스의 300노드 시간 = $300 * 2$ 용량 단위 = 600 용량 단위 기준선의 300 테넌트 시간 = $300 * 1.5$ 용량 단위 = 450 용량 단위 합계 = 1,100 용량 단위 = EUR 1,100

참고

둘 이상의 노드를 사용하여 데이터를 병렬로 처리하는 것이 가능합니다. 비용은 노드 시간으로 계산됩니다. 즉, 2개 노드의 데이터 처리 1시간은 2노드 시간과 같습니다.

추천

예상 비용을 추정하려면 [SAP AI 핵심 비용 계산기를 사용하세요](#).

관련 정보

[SAP 디스커버리 센터](#)



[SAP BTP 서비스 설명 가이드](#)



[리소스 계획 선택 \[페이지 104\]](#)

4.3 Generative AI Hub의 측정 및 가격

생성 AI 허브의 모델 사용은 GenAI 토큰 및 용량 단위를 사용하여 측정됩니다.

참고

생성적 AI 허브는 확장 서비스 계획의 일부로만 제공됩니다.

GenAI 토큰은 각 모델의 1,000개 토큰 블록에 해당합니다. 필요한 GenAI 토큰 수는 사용하는 모델과 토큰이 입력용인지 출력용인지에 따라 달라집니다. 모델 공급자의 토큰에서 얼마나 많은 GenAI 토큰을 얻는지 파악하는 데 도움이 되는 전환율이 있습니다. 전환율은 1,000개의 입력 및 출력 모델 토큰 블록에 적용됩니다. 이러한 요율을 참조하여 총 소비하는 GenAI 토큰 수를 계산할 수 있습니다. 지원되는 모델 및 모델 토큰과 GenAI 토큰 간의 전환율에 대한 자세한 내용은 SAP Note [3437766](#)을 참조하세요..



GenAI 토큰으로 측정된 각 기록은 "용량 단위"라고 알려진 청구 가능한 측정 단위로 변환됩니다. GenAI 토큰과 용량 단위(용량 단위 값) 간의 변환율은 [SAP BTP 서비스 설명 가이드](#)를 참조하세요..



예

이 예에서는 가상의 값을 사용합니다.

지정된 요청에 대해 x 입력 모델 토큰과 y 출력 모델 토큰이 사용됩니다. 해당 측정항목은 다음과 같습니다.

	GenAI 입력 토큰	GenAI 출력 토큰
모델	(모델 토큰 1,000개당)	(모델 토큰 1,000개당)
샘플 모델	0.002	0.003

총 GenAI 토큰 = $x/1000 * 0.002 + y/1000 * 0.003$ 용량 단위 = 총 GenAI 토큰 * 2.0000(용량 단위 값)

다른 SAP AI Core 구성 요소 사용과 관련된 요금도 적용될 수 있습니다. 자세한 내용은 [SAP AI Core 측정 및 가격을 참조하세요.](#)

추천

예상 비용을 추정하려면 SAP AI Core 비용 계산기를 사용하세요. 자세한 내용은 [비용 계산기를 참조하세요.](#)

4.4 리소스 계획 선택

수요에 따라 다양한 작업에 다양한 인프라 리소스를 사용하도록 SAP AI Core를 구성할 수 있습니다. SAP AI Core는 이러한 목적으로 "리소스 계획"이라는 사전 구성된 여러 인프라 번들을 제공합니다.

문맥

리소스 계획은 워크플로 및 제공 템플릿에서 리소스를 선택하는 데 사용됩니다. 워크플로의 각 단계에는 서로 다른 리소스 계획이 있을 수 있습니다.

일반적으로 워크로드에 GPU 가속이 필요한 경우 GPU 지원 리소스 계획 중 하나를 사용해야 합니다. 그렇지 않은 경우에는 워크로드의 예상 CPU 및 메모리 요구 사항을 기반으로 리소스 계획을 선택하세요.

SAP AI Core 내에서 리소스 계획은 Pod 수준의 ai.sap.com/resourcePlan 레이블을 통해 선택됩니다. 그것 선택한 리소스 계획을 매핑하고 다음 리소스 계획 ID 중 하나일 수 있는 문자열 값을 사용합니다.

AWS에 대한 리소스 계획 사양

리소스 계획 ID	GPU	CPU 코어	메모리 GB	재할당 코드 출처 워크플로 온도- 접시
열차-L	1 V100	7	55	ai.sap.com/ 자원계획: 기차.l
추론-S	1 T4	삼	10	ai.sap.com/ 자원계획: 추론하다
추론-M	1 T4	7	26	ai.sap.com/ 자원계획: 추론.m
추론-L	1 T4	15	58	ai.sap.com/ 자원계획: 추론하다.l
기동기	-	1	삼	ai.sap.com/ 자원계획: 기동기
기초적인	-	삼	11	ai.sap.com/ 자원계획: 기초적인
기본-8x	-	31	116	ai.sap.com/ 자원계획: 기본.8x

제한사항

프리 티어 서비스 계획의 경우 스타터 리소스 계획만 사용할 수 있습니다. 다른 계획을 지정하면 문제있는. Standard 서비스 계획의 경우 모든 리소스 계획을 사용할 수 있습니다. 자세한 내용은 [프리 티어를](#) 참조하세요. [\[페이지 37\]](#) 및 [서비스 계획 \[페이지 35\]](#).

참고

이러한 모든 노드에 대한 기본 디스크 저장소 크기에는 제한이 있습니다. 노드에 로드되는 데이터 세트 디스크 공간을 소비합니다. 대규모 데이터 세트(30GB 이상)나 대규모 모델이 있는 경우 다음을 수행할 수 있습니다.
디스크 크기를 늘려야 합니다. 그렇게 하려면 Argo Workflows의 영구 볼륨 청구를 사용하여 필요한 디스크 크기(볼륨 참조).

서비스 이용 보고

서비스 사용량 소비는 글로벌 계정의 [개요](#) 페이지와 하위 계정의 [개요](#) 및 [사용량 분석](#) 페이지에 있는 SAP BTP 조종석에 보고됩니다. 사용량 보고서에는 청구 가능한 측정값과 청구할 수 없는 측정값의 사용량이 나열됩니다. 최종 월별 청구서는 청구 가능한 측정값만을 기준으로 합니다. 청구 불 가능한 측정값은 보고 목적으로만 표시됩니다.

4.5 서비스 계획 업데이트

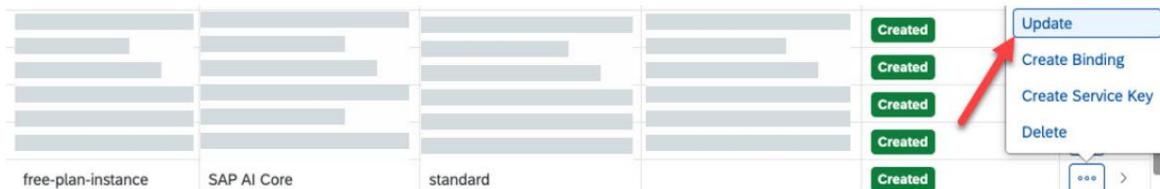
프리 티어에서 제품을 살펴본 후 SAP AI Core 표준 또는 확장 플랜으로 업데이트하는 방법을 알아보세요.

절차

1. SAP BTP 조종석에서 글로벌 계정을 개설하세요.
2. 하위 계정으로 이동합니다.
3. 탐색 영역에서 [인스턴스 및 구독](#)을 선택합니다.

Cloud Foundry 환경에서 하위 계정이 구독된 애플리케이션 목록이 표시됩니다.

4. SAP AI Core를 검색하세요.
5. 구독 행 끝에 있는 줄임표를 선택하고 메뉴에서 [업데이트](#)를 선택합니다.
6. 열리는 마법사에서 [기본값을 선택](#)하고 [구독 업데이트](#)를 클릭합니다.



참고

서비스 계획을 업데이트한 후에는 무료 등급 제한이 더 이상 적용되지 않습니다.

무료 계층 계획에 정의한 데이터는 자동으로 새 계획으로 마이그레이션됩니다.

서비스 계획을 업데이트해도 사용자 권한은 변경되지 않습니다.

서비스 계획 변경

무료 등급 옵션을 처음 구독하는 경우 동일한 서비스 인스턴스를 표준 서비스 계획(기업 계정용)으로 마이그레이션할 수 있습니다.

무료 등급 옵션에서 표준 플랜이나 확장 플랜으로 마이그레이션하거나 표준 플랜에서 확장 플랜으로 마이그레이션하는 경우 학습된 모델을 포함한 메타데이터 및 트랜잭션 데이터도 마이그레이션됩니다(엔터프라이즈 계정의 경우).

제한사항

표준 또는 확장 서비스 계획에서 무료 계층 옵션으로 또는 확장 계획에서 표준 계획으로 마이그레이션하는 것은 불가능합니다.

확장 계획이나 표준 계획이 적용된 인스턴스가 삭제되면 표준 계획으로 새 인스턴스를 생성할 수 없습니다.

5 초기 설정

SAP Business Technology Platform의 SAP BTP 조종석에서 SAP AI Core를 프로비저닝합니다. 프로비저닝 후에는 SAP AI Core 인스턴스에 액세스하기 위한 URL 및 자격 증명을 제공하는 서비스 키를 갖게 됩니다.

전제조건

- SAP BTP 관리자는 SAP Business Technology Platform의 글로벌 계정에 액세스할 수 있습니다. 자세한 내용은 [글로벌 계정 얻기를 참조하십시오.](#)
- SAP BTP 관리자가 하위 계정에 대한 권한을 설정했습니다.

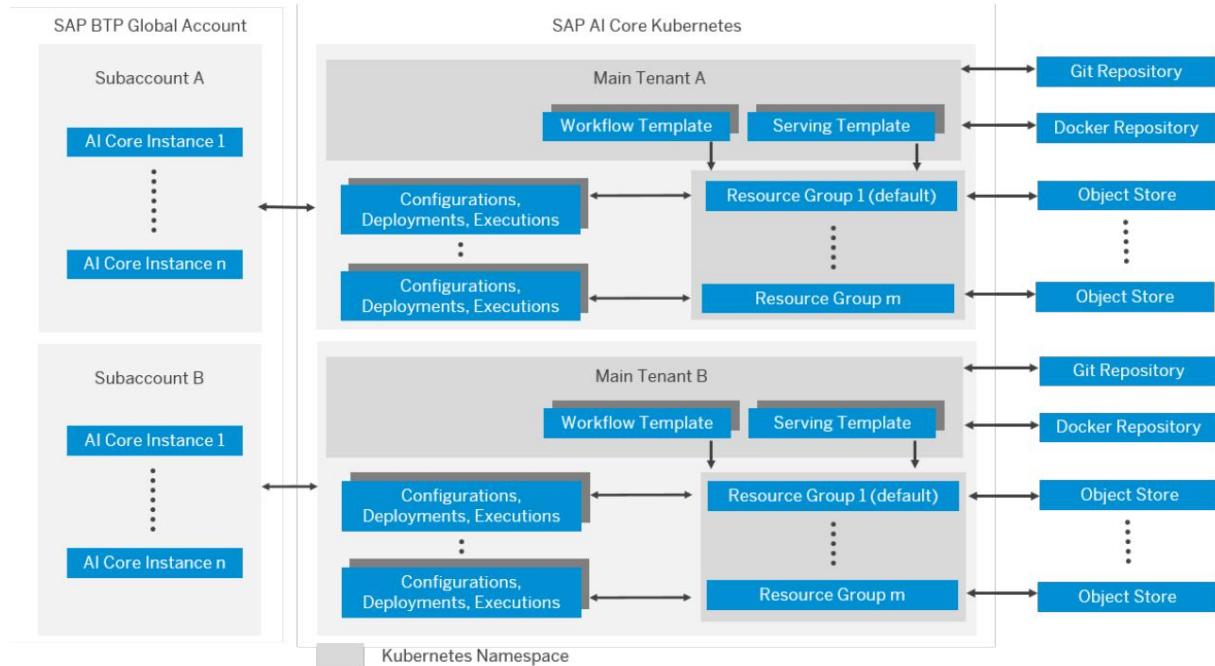
문맥

SAP AI Core 서비스는 테넌트 인식 재사용 서비스입니다. 영역의 ID를 기준으로 테넌트를 격리합니다.

(하위 계정을 나타냄) SAP AI Core 서비스 인스턴스는 하위 계정 내에 생성됩니다. 각 하위 계정은 SAP AI Core 테넌트를 나타냅니다.

참고

SAP AI Core 서비스는 서비스 인스턴스 ID를 기반으로 테넌트를 격리하지 않습니다. 동일한 하위 계정 내에 여러 서비스 인스턴스를 생성하는 경우 모든 서비스 인스턴스는 동일한 SAP AI Core 테넌트를 참조합니다.



아래 단계는 프로비저닝 절차를 안내합니다. 또는 SAP AI Core 및 SAP AI Launchpad 모두에 부스터를 사용할 수 있습니다. 자세한 내용은 [AI 부스터 튜토리얼을](#) 참조하세요. . 부스터를 사용하기로 선택한 경우 [SAP AI Core Starter](#) 자습서 [페이지 63]의 나머지 단계를 건너뛸 수 있습니다.



5.1 Cloud Foundry에서 서비스 활성화

SAP BTP Cloud Foundry 환경에 대한 표준 절차를 사용하여 SAP AI Core를 활성화합니다.

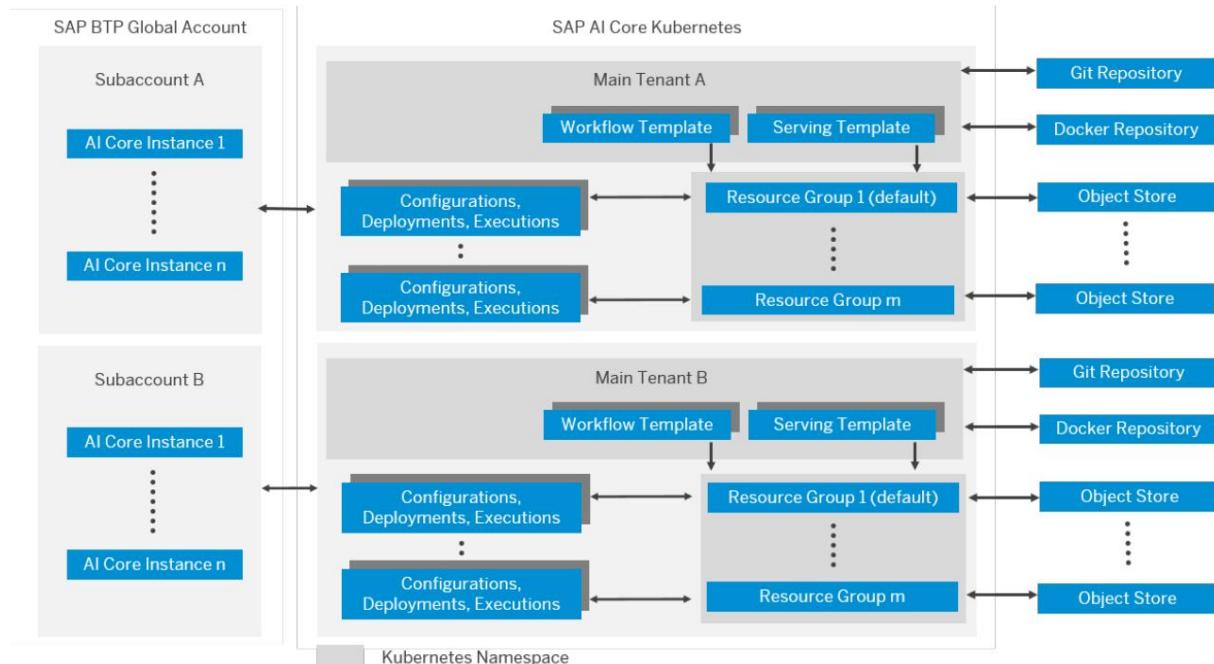
SAP Business Technology Platform의 SAP BTP 조종석에서 SAP AI Core를 프로비저닝합니다. 서비스를 프로비저닝한 후에는 SAP AI Core 인스턴스에 액세스하기 위한 URL 및 자격 증명을 제공하는 서비스 키를 갖게 됩니다.

SAP AI Core 서비스는 테넌트 인식 재사용 서비스입니다. 영역의 ID를 기준으로 테넌트를 격리합니다.

(하위 계정을 나타냄) SAP AI Core 서비스 인스턴스는 하위 계정 내에 생성됩니다. 각 하위 계정은 SAP AI Core 테넌트를 나타냅니다.

참고

SAP AI Core 서비스는 서비스 인스턴스 ID를 기반으로 테넌트를 격리하지 않습니다. 동일한 하위 계정 내에 여러 서비스 인스턴스를 생성하는 경우 모든 서비스 인스턴스는 동일한 SAP AI Core 테넌트를 참조합니다.



아래 단계는 프로비저닝 절차를 안내합니다. 또는 SAP AI Core와 SAP AI Launchpad 모두에 부스터를 사용할 수 있습니다. 자세한 내용은 [AI 부스터 튜토리얼을](#) 참조하세요.. 부스터를 사용하기로 선택한 경우 [SAP AI Core Starter](#) 자습서 [페이지 63]의 나머지 단계를 건너뛸 수 있습니다.



5.1.1 하위 계정 생성

절차

1. SAP Business Technology Platform 조종석에서 [Account Explorer](#)를 선택한 후 [생성을 클릭합니다.](#)

▶ 하위 계정 . ▶

2. 하위 계정 생성 대화 상자에서 하위 계정 이름을 입력하고 지역을 선택합니다.

상위 항목의 기본값은 글로벌 계정의 이름입니다.

Display Name*	Description
My Subaccount	Enter a description of up to 300 characters
Region*	
Europe (Frankfurt) AWS 2	
Subdomain*	
Parent*	
Advanced	
<input type="button" value="Create"/> <input type="button" value="Cancel"/>	

3. 선택 사항: 하위 계정이 생산 목적으로 사용되는 경우 [고급 아래에서 생산에 사용 확인란을 선택합니다.](#)

이 설정은 하위 계정의 구성을 변경하지 않습니다. 이는 글로벌 계정에서 프로덕션 하위 계정을 관리하는 데 도움을 주기 위한 것입니다. 예를 들어, 클라우드 운영자는 미션 크리티컬 계정과 관련된 사고를 처리할 때 이를 참조할 수 있습니다.



4. 생성을 클릭합니다.

5. 하위 계정을 보려면 계정 탐색기로 돌아가세요.

The screenshot shows the SAP BTP Cockpit's navigation bar with the SAP BTP Cockpit logo. The left sidebar has a red box around the 'Account Explorer' section. The main area shows the 'Global Account' page with the 'Subaccounts (67)' tab selected. A red box highlights the 'My Subaccount' entry in the list.

5.1.2 Cloud Foundry 활성화

절차

1. 하위 계정을 클릭하고 개요 페이지에서 Cloud Foundry 활성화를 선택합니다.

The screenshot shows the SAP BTP Cockpit interface. On the left, a sidebar menu is open, with the 'Overview' item highlighted by a red box. The main content area is titled 'Subaccount: My Subaccount - Overview'. It displays basic account information: Entitlements (30), Instances and Subscriptions (0). Below this, there are fields for Subdomain, Tenant ID, and Subaccount ID, each with a grayed-out placeholder value. To the right, provider details are listed: Provider: Amazon Web Services (AWS), Region: Europe (Frankfurt) AWS 2, and Environment: Multi-Environment. A section titled 'Cloud Foundry Environment' contains the message 'You are currently not using Cloud Foundry capabilities.' and a blue button labeled 'Enable Cloud Foundry'.

2. Cloud Foundry 환경 인스턴스에 대한 기본 정보를 입력하고 [생성을 클릭합니다.](#)

Enable Cloud Foundry

1 Basic Info

Enter basic info for your environment instance.

Environment: * Can't find what you're looking for?

Plan: *

Landscape: *

Instance Name:

Org Name: *

i Cloud Foundry API endpoints may differ between subaccounts in the same region.
The API endpoint for this subaccount is shown in the Cloud Foundry environment instance details.

Create Cancel

5.1.3 공간 만들기

절차

1. 하위 계정 개요 페이지에서 공간 생성을 선택합니다.

Subaccount: Testsubaccount - Overview Edit

This is a test sub account created for hosting AI Core Service

General Cloud Foundry Environment Entitlements

Entitlements 13 **Instances & Subscriptions** 1

Subdomain: [REDACTED] Provider: Amazon Web Services (AWS) Used for Production: Yes Created On: Aug 25, 2021, 22:14
Tenant ID: [REDACTED] Region: Europe (Frankfurt) Beta Features: Disabled Modified On: Aug 25, 2021, 22:15
Subaccount ID: [REDACTED]

Cloud Foundry Environment

Org Name: [REDACTED] API Endpoint: [REDACTED] Org ID: [REDACTED] Spaces (0) Create Space

Manage environment instance Disable Cloud Foundry

Name	Applications	Service Instances
No spaces defined in this subaccount		

2. 공간 이름을 입력하고 필요한 역할을 할당한 후 생성을 클릭합니다.

Create Space

Space Name: *

Assign space roles to +

Space Manager Space Developer Space Auditor

Create Cancel

관련 정보

[Cloud Foundry 환경의 역할 정보](#)

5.1.4 서비스 계획 추가

절차

1. SAP BTP 조종석에서 글로벌 계정으로 이동하여 **자격 , 엔터티 할당**을 차례로 선택합니다.
2. **엔터티 선택** 상자에서 하위 계정을 선택하고 선택을 클릭합니다.
3. 편집을 선택합니다.

Service	Service Technical Name	Plan	Set Quota Limit	Subaccount Assignment	Remaining Global Quota	Actions
Application Autoscaler	autoscaler	standard		1 shared units	1 shared units	
Application Logging Service	application-logs	lite		1 shared units	1 shared units	

4. 서비스 계획 추가를 선택합니다.

5. SAP AI Core를 선택하고 선택한 서비스 계획을 선택합니다.

Services available to this subaccount
SAP AI Core
SAP AI Launchpad
SAP HANA Schemas & HDI Containers
Service Manager

Available Plans

- extended
Default plan for aicore and gen ai features.
- standard
Default standard plan
- free
Free plan. To work without usage limits, migrate to standard plan. Please note, only community support is available for free service plans and these are not subject to SLAs. Use of free tier service plans are subject to additional terms and conditions as provided in the Business Technology Platform Supplemental Terms and Conditions linked in the Additional Links tab displayed in the Service tile.

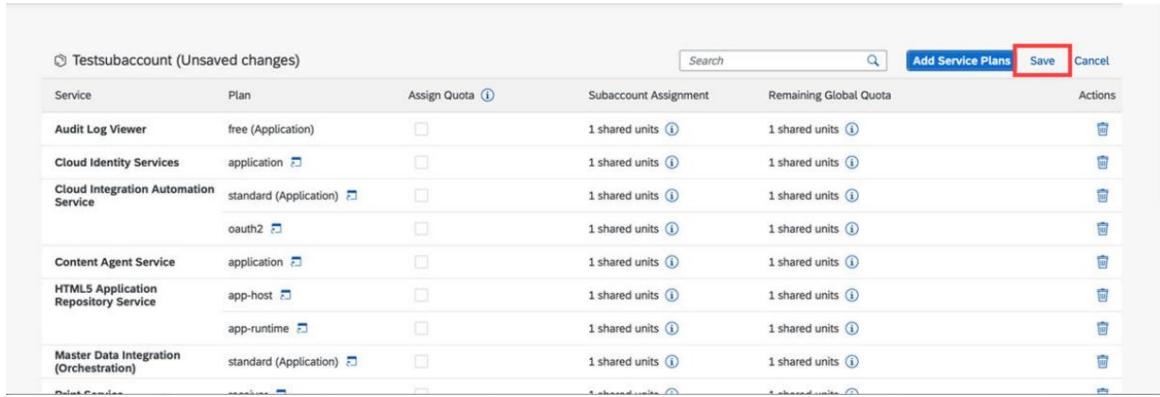
2 selected plans
Services: SAP AI Core (2 plans)

Add 2 Service Plans Cancel

팁

생성적 AI 기능을 사용하려면 확장 플랜을 선택하세요. 자세한 내용은 [서비스 계획 \[페이지 35\]](#)를 참조하십시오.

6. 변경 사항을 저장합니다.



Testsubaccount (Unsaved changes)		Assign Quota <small>i</small>	Subaccount Assignment	Remaining Global Quota	Actions
Service	Plan				
Audit Log Viewer	free (Application)	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
Cloud Identity Services	application <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
Cloud Integration Automation Service	standard (Application) <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
	oauth2 <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
Content Agent Service	application <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
HTML5 Application Repository Service	app-host <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
	app-runtime <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	
Master Data Integration (Orchestration)	standard (Application) <small>i</small>	<input type="checkbox"/>	1 shared units <small>i</small>	1 shared units <small>i</small>	

관련 정보

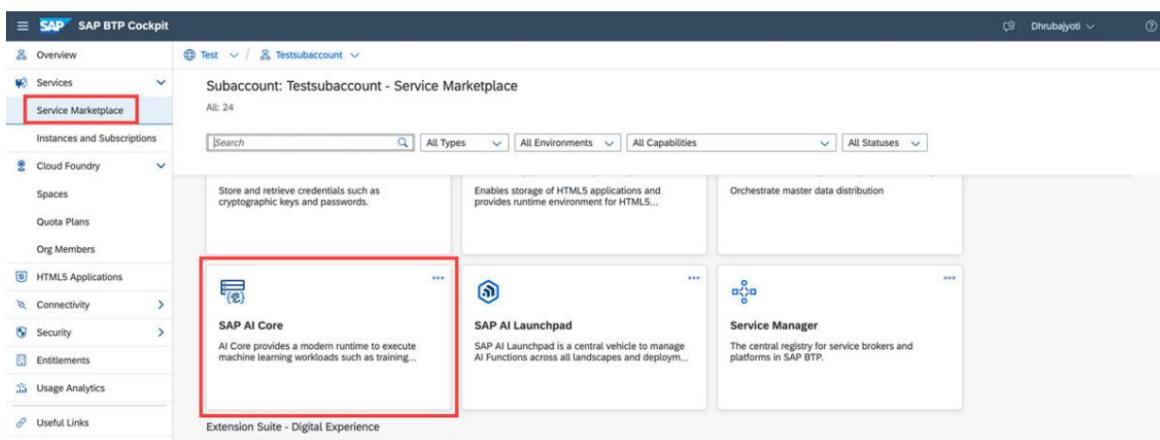
[서비스 계획 \[페이지 35\]](#)

5.1.5 서비스 인스턴스 생성

절차

- SAP BTP 조종석에서 글로벌 계정 내의 하위 계정으로 이동하여 [서비스를 선택합니다.](#)
[시장.](#)

SAP AI Core에 대한 타일이 표시됩니다.



The screenshot shows the SAP BTP Cockpit interface. On the left, there is a navigation sidebar with various options like Overview, Services, Instances and Subscriptions, Cloud Foundry, etc. Under the Services section, 'Service Marketplace' is selected and highlighted with a red box. The main content area displays a grid of service tiles. One tile for 'SAP AI Core' is specifically highlighted with a red box. The tile description reads: 'Store and retrieve credentials such as cryptographic keys and passwords.' Below the tile, it says 'AI Core provides a modern runtime to execute machine learning workloads such as training..'. To the right of the SAP AI Core tile are other tiles for 'SAP AI Launchpad' and 'Service Manager'.

2. 타일을 열고 생성을 클릭합니다.

The screenshot shows the SAP BTP Cockpit interface. On the left, there's a sidebar with various service categories like Overview, Services, Service Marketplace, Cloud Foundry, etc. In the center, under the 'Services' section, there's a list of services. One service, 'SAP AI Core', is highlighted with a red box. To the right of the service list, there's a detailed view of the 'SAP AI Core' service. At the top right of this view, there's a blue 'Create' button, which is also highlighted with a red box.

3. 서비스 인스턴스의 이름을 입력하고 다음을 선택합니다 (기타 모든 세부 정보는 기본적으로 입력됩니다).

The screenshot shows the 'New Instance or Subscription' wizard. It has three steps: 1. Basic Info, 2. Parameters, and 3. Review. The current step is 'Basic Info'. It asks for basic information like Service, Plan, Runtime Environment, Space, and Instance Name. The 'Instance Name' field is filled with 'TestAICore' and is highlighted with a red box. At the bottom right, there are buttons for 'Next >', 'Create', and 'Cancel'.

Plan	Description	Environments	Active
standard	Default standard plan More	Cloud Foundry	***

4. 현재 JSON 파일 업로드 기능은 지원되지 않습니다. 계속하려면 다음을 선택하세요 .

New Instance or Subscription

1 Basic Info 2 Parameters 3 Review

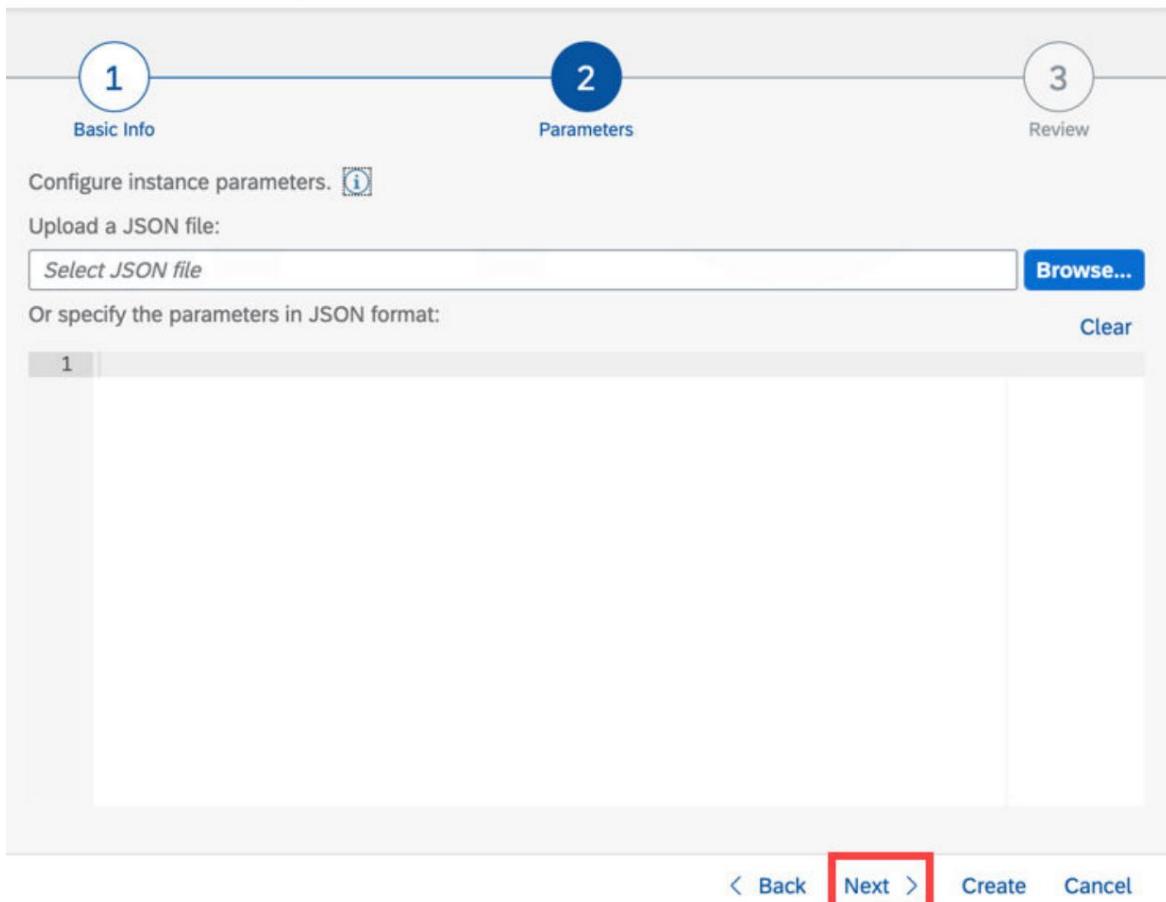
Configure instance parameters. [\(1\)](#)

Upload a JSON file:

Select JSON file [Browse...](#)

Or specify the parameters in JSON format:
1

< Back [Next >](#) Create Cancel



5. 데이터를 확인하고 [생성을 선택합니다.](#)

New Instance or Subscription

Review and verify the instance details.

TestAICore

Service: SAP AI Core
Service Plan: standard
Runtime Environment: Cloud Foundry
Space: TestAIFoundation

Creating an instance might take a while.

Back Create Cancel

결과

서비스 인스턴스가 생성되면 하위 계정의 [인스턴스 및 구독](#) 페이지에서 이를 확인할 수 있습니다.

Instance	Service	Plan	Runtime Environment	Scope	Credentials	Status
TestAICore	SAP AI Core	standard	Cloud Foundry	TestAIFoundation		Created

5.1.6 서비스 키 생성

절차

- 인스턴스 및 구독 페이지에서 새 인스턴스를 찾고 메뉴에서 서비스 키 생성을 선택합니다.
쓰러지다.

The screenshot shows the SAP BTP Cockpit interface. The left sidebar contains various navigation options like Overview, Services, Cloud Foundry, Spaces, Quota Plans, Org Members, etc. The 'Instances and Subscriptions' option is selected and highlighted with a red box. The main content area displays the 'Subaccount: Testsubaccount - Instances and Subscriptions' page. It includes a search bar and filters for All Services, All Plans, and All Statuses. Below the search bar, there are tabs for Subscriptions (0), Instances (1), and Environments (1). The 'Instances (1)' section shows a table with one row: Instance 'TestAICore', Service 'SAP AI Core', Plan 'standard', Runtime Environment 'Cloud Foundry', Scope 'TestAIFoundation', and Status 'Created'. To the right of the table, there are 'Update' and 'Create Binding' buttons, with the 'Create Service Key' button highlighted by a red box.

- 서비스 키 이름을 입력하고 생성을 클릭합니다.

New Service Key

Service Key Name: * TestAICore

Configure Binding Parameters: [\(i\)](#)

Upload a JSON file:

Select JSON file Browse...

Or specify the parameters in JSON format: Clear

```
1
```

Create Cancel

선택사항: clientsecret 자격 증명 대신 x.509 인증서를 사용하려면 업데이트하여 자격 증명을 지정하세요.

다음은 JSON으로 업로드합니다.

```
{"xsuaa": {"credential-type": "x509", "x509": { "key-length": 2048, "validity": 7, "validity-type": "DAYS"}}
```

- key-length: 생성된 개인 키의 바이트 길이입니다. 기본값: 2048.
- 유효성: 유효성 시간 단위입니다. 일, 월, 연도가 지원됩니다. 기본값: DAYS.
- 유효성 유형: 시간 단위 수입니다. 기본값: 7.

기본 조합은 길이가 2048인 키이며 7일 동안 유효합니다.

3. 서비스 키를 다운로드하여 저장하세요.

결과

이제 SAP AI Core 인스턴스에 액세스하기 위한 URL 및 자격 증명을 제공하는 서비스 키가 생겼습니다.
SAP AI Launchpad, SAP AI Core 툴킷 Postman 또는 컬을 통해.

Credentials

```

1  {
2      "clientid": "sb-4376bdb7-b313",  

3      "clientsecret": "xj....authentication.sap.hana.ondemand.com",  

4      "url": "https://.sap.hana.ondemand.com",  

5      "identityzone": "ai...g",  

6      "identityzoneid": "1eb727a0-1",  

7      "appname": "4!b313",  

8      "serviceurls": {  

9          "AI_API_URL": "https://api.ai.i...ml.hana.ondemand.com"  

10     }  

11 }

```

[Copy JSON](#) [Download](#) [Close](#)

클라이언트 비밀을 생성한 경우 키에는 다음이 포함됩니다.

- clientid, clientsecret 및 url을 사용하여 인증 토큰을 생성할 수 있습니다.
- identityzone 및 identityzoneid는 테넌트 ID를 나타냅니다.
- appname은 서비스 인스턴스 격리가 구현된 경우 서비스 인스턴스 세부정보를 제공합니다.
- serviceurl을 사용하면 인증 토큰이 생성된 후 SAP AI Core와 상호 작용할 수 있습니다.
 - AI_API_URL: ML 아티팩트(예: 훈련, 데이터, 모델 및 배포)를 처리하기 위한 통합 AI API
여러 하이퍼스케일러에 걸쳐

x.509 인증서를 생성한 경우 키에는 다음이 포함됩니다.

- 자격증
- certurl을 사용하여 인증 토큰을 생성할 수 있습니다.
- RSA 개인 키를 입력하세요.
- identityzone 및 identityzoneid는 테넌트 ID를 나타냅니다.
- appname은 서비스 인스턴스 격리가 구현된 경우 서비스 인스턴스 세부정보를 제공합니다.
- serviceurl을 사용하면 인증 토큰이 생성된 후 SAP AI Core와 상호 작용할 수 있습니다.
 - AI_API_URL: ML 아티팩트(예: 훈련, 데이터, 모델 및 배포)를 처리하기 위한 통합 AI API
여러 하이퍼스케일러에 걸쳐

5.1.7 서비스 키 사용

서비스 키를 구성한 후에는 로컬 클라이언트, 다른 공간의 앱 또는 배포 외부 엔터티가 사용 가능한 인터페이스 중 하나를 통해 SAP AI Core에 액세스 하는 데 사용할 수 있습니다.

우편 배달부 사용

전제조건

- <https://www.postman.com/>에서 Postman 클라이언트를 다운로드하여 설치했습니다..
- Postman 설명서와 인터페이스에 익숙해졌습니다.

절차

1. https://api.sap.com/api/AI_CORE_API/overview에서 JSON 컬렉션을 다운로드합니다. .
2. Postman에서 가져오기를 클릭하고 JSON 파일을 선택한 다음 가져오기를 선택하여 가져오기를 시작합니다.
3. 가져오기가 완료된 후 컬렉션을 강조 표시하고 인증 탭을 선택합니다.
4. 새 토큰 구성으로 이동하여 서비스 키의 자격 증명을 입력하고 변경 사항을 저장합니다.

참고

- 토큰 이름 필드는 설명 식별자로 선택합니다.
- 액세스 토큰 URL은 서비스 키에 url로 표시되어 있습니다. URL 끝에 /oauth/token을 추가하세요 .
- 부여 유형은 클라이언트 자격 증명 이어야 합니다 .

자격 증명의 문자와 관련된 경고가 표시되면 무시하십시오.

참고

clientsecret 자격 증명 대신 x.509 인증서를 생성한 경우 인증서, 키 및 certUrl을 사용하여 토큰을 만들어야 합니다.

예: 커먼 --cert <cert.pem> --key <key.pem> -XPOST <certUrl>/oauth/token -d 'grant_type=client_credentials&client_id=<클라이언트 ID>'

5. 변수 탭을 선택하고 자격 증명에서 baseUrl을 설정합니다.

baseUrl은 서비스 키에서 AI_API_URL로 표시됩니다.

6. 저장을 선택합니다.

7. 인증 탭에서 새 액세스 토큰 발기를 선택합니다. 그리고 인증과정을 기다립니다.

인증 프로세스가 완료되면 토큰 사용을 선택하여 완료하세요. 토큰이 환경 변수에 저장되어 있는지 확인하세요. 자동으로 저장되지 않는 경우 수동으로 토큰 필드에 복사하여 붙여넣을 수 있습니다.

다음 단계

자체 AI 모델을 교육하고 배포하려면 [관리 \[페이지 64\]](#)의 절차를 따르십시오.

생성 AI 허브에서 제공되는 생성 AI 모델을 사용하려면 [생성 AI 허브의 모델 및 시나리오 \[페이지 225\]](#)를 참조하십시오.

컬 사용하기

전제조건

컬은 기본적으로 운영 체제에 설치될 가능성이 높습니다. 확인하려면 명령 프롬프트를 열고 컬 -V를 입력하세요. 컬이 설치되어 있지 않은 경우 <https://curl.se/>에서 다운로드하여 설치하세요..

참고

macOS에서는 컬 명령을 따르기 위해 jq를 설치해야 할 수도 있습니다.

1. <https://brew.sh/>에서 Brew를 설치합니다. ↗
2. 터미널 세션에서 Brew install jq를 실행하여 쉘 환경에 jq를 설치합니다.

절차

1. 다음과 같이 환경을 설정합니다.

리눅스의 경우:

```
# XSUAA 세부정보
# URL은 뒤에 슬래시가 없어야 합니다. '/' 서비스 키에서 내보내기
CLIENTID=<clientid> 내보내기 서비스 키에서 CLIENTSECRET=<clientsecret>
서비스 키에서 XSUAA_URL=<url> 내보내기 서비스 키에서 AI_API_URL=<AI_API_URL>
```

Windows Powershell의 경우:

```
$env:CLIENTID = 서비스 키의 <clientid> $env:CLIENTSECRET = 서비스 키의
<clientsecret> $env:XSUAA_URL = 서비스 키의 <url> $env:AI_API_URL = 서비스 키의
<AI_API_URL>
```

참고

내보내기 명령은 키 값을 환경 변수로 설정합니다. 즉, 터미널 세션을 닫은 후에도 키 값이 유지됩니다. 현재 세션에 대해서만 내보내기 없이 환경 변수를 설정할 수 있습니다.

2. API를 호출하려면 서비스 키의 clientid 및 clientsecret을 사용하여 XSUAA OAuth 토큰을 가져옵니다.

AI API 호출 시 인증을 위해 XSUAA OAuth 토큰이 필요합니다.

리눅스의 경우:

```
SECRET=`echo -n '$CLIENTID:$CLIENTSECRET' | base64 -i - TOKEN=`curl --location --request POST
"$XSUAA_URL/oauth/token?
grant_type=client_credentials" \--header "Authorization: Basic
$SECRET" | jq -r '.access_token'`
```

Windows Powershell의 경우:

```
$SECRET = $env:CLIENTID + ":" + $env:CLIENTSECRET
$base64SECRET = [Convert]::ToBase64String([System.Text.Encoding]::UTF8.GetBytes("$SECRET"))
$TOKENRESPONSE = [System.Net.Http.HttpClient]::new().PostAsync("$XSUAA_URL/oauth/token?
grant_type=client_credentials" \--header "Authorization: Basic $base64SECRET")
$TOKENRESPONSE = $TOKENRESPONSE | ConvertFrom-Json
$TOKEN = $TOKENRESPONSE.access_token
```

참고

토큰은 제한된 시간 동안 유효합니다. 만료되면 동일한 코드를 사용하여 새 토큰을 만듭니다.
단편.

참고

clientsecret 자격 증명 대신 x.509 인증서를 생성한 경우 다음을 사용해야 합니다.
인증서, 키 및 certUrl을 사용하여 토큰을 만듭니다.
예: 커먼 --cert <cert.pem> --key <key.pem> -XPOST <certUrl>/oauth/token
-d 'grant_type=client_credentials&client_id=<클라이언트 ID>'

3. 토큰을 올바르게 가져왔는지 확인합니다.

```
echo $TOKEN
```

다음과 같이 긴 영숫자 문자열이 표시됩니다.

```
eyJhbGciOiJSUzI1NiIsp0dSI6Imh0dHBzOi8vYWktYWxwaGETtdmFsaWRhdGlvbi0yLmF1dGhlbnRpY2F0aW9uLnNhcC5oYW5hLm9uZGVtYW5kLmNvbS90b2tlbl9rZXlziwiia2lkjoiZGVmYXVsdC1qd3Qta2V5LTMyODMxMjg2NCIsInR5cCI6IkpxVCJ94ZGU5YjAxNmQ0MDk5YjlmM...
.....
...ALdfbMsHoYTtF6fNFbf3ZQ
```

다음 단계

자체 AI 모델을 교육하고 배포하려면 [관리 \[페이지 64\]](#)의 절차를 따르십시오.

생성적 AI 허브에서 제공되는 생성적 AI 모델을 사용하려면 [생성적 AI의 모델 및 시나리오](#)를 참조하세요.
[허브 \[페이지 219\].](#)

5.2 SAP AI Core Starter 튜토리얼

부스터를 사용하면 [AI Boosters](#) 튜토리얼을 통해 SAP AI Core 및 SAP AI Launchpad에 대한 프리 티어 플랜을 시작할 수 있습니다. .



또는 부스터 튜토리얼을 완료한 후 간단한 사용 사례에서 SAP AI Core의 기본 사항을 처음부터 끝까지 배우려면 튜토리얼을 따르세요. [SAP AI Core용 Quick Start](#). 이미 SAP AI Core를 프로비저닝한 경우 이 가이드의 두 번째 튜토리얼로 건너뛸 수 있습니다.

SAP AI Core에 사용할 수 있는 전체 자습서 라이브러리를 보려면 [자습서 \[페이지 249\]를 참조하십시오.](#)

참고

SAP AI Core를 프로비저닝한 후에는 다양한 방법으로 서비스에 액세스할 수 있습니다. 아래 튜토리얼에서 둘 이상의 옵션 탭을 제공하는 경우 선택한 액세스 방법에 대한 탭만 완료하면 됩니다.

코딩보다 GUI를 선호하는 고객에게는 SAP AI Launchpad를 권장합니다. 자세한 내용은 [SAP AI Launchpad를 참조하세요.](#)

6 행정

SAP AI Core와 함께 사용되는 외부 프로그램 및 도구에 대한 비밀을 생성하면 자격 증명을 손상시키지 않고 연결할 수 있습니다.

GitHub, Docker, Amazon Web Services S3 스토리지 등의 외부 도구와 함께 SAP AI Core를 사용하면 버전 제어, 컨테이너화, 클라우드 스토리지의 이점을 활용할 수 있습니다. 안정적인 인터넷 연결이 있는 경우 콘텐츠를 원격으로 사용할 수 있습니다.

관리는 일회성 절차이지만 필요한 경우 도구 제거 또는 추가 등의 단계를 반복할 수 있습니다.

참고

SAP AI Core 인스턴스를 구성하기 전에 초기 설정 작업을 완료해야 합니다. 자세한 내용은 [초기 설정 \[페이지 45\]](#)을 참조하십시오.

[리소스 그룹 관리 \[페이지 71\]](#)

리소스 그룹은 사용자가 구성, 실행, 배포 및 아티팩트를 생성하거나 추가할 수 있는 고유한 전용 네임스페이스 또는 작업 영역 환경입니다. 훈련 작업을 실행하는 데 사용됩니다.
또는 모델 서버.

6.1 Git 리포지토리 관리

6.1.1 Git 리포지토리 추가

자체 Git 리포지토리를 사용하여 SAP AI Core 템플릿의 버전을 제어할 수 있습니다. SAP AI Core 인스턴스에 대한 GitOps 온보딩에는 git 리포지토리 설정 및 콘텐츠 동기화가 포함됩니다.

전제조건

초기 설정을 완료했습니다. 자세한 내용은 [초기 설정 \[페이지 45\]](#)을 참조하십시오.

문맥

Git 리포지토리는 SAP AI Core에 등록하는 개인 액세스 토큰을 생성하여 관리됩니다.

개인 액세스 토큰은 자격 증명을 손상시키지 않고 GitHub 저장소에 대한 연결을 허용하고 제어하는 수단입니다.

전제조건

- 인터넷을 통해 git 저장소에 액세스할 수 있습니다.
- git 저장소에 대한 개인 액세스 토큰을 생성했습니다. 자세한 내용은 [만들기를](#) 참조하세요.
 개인 액세스 토큰 .
- GitLab에서 호스팅되는 Git 저장소를 온보딩하려면 저장소 URL에 다음이 포함되어 있는지 확인하세요.
.git 접미사.
- 저장소에서는 비밀이 허용되지 않습니다. 비밀을 사용하면 콘텐츠를 동기화할 수 없습니다.

참고

리소스를 동기화할 때 이를 충들이 없는지 확인하십시오. 이는 하나의 테넌트에서 여러 저장소나 애플리케이션을 사용하는 경우 특히 중요합니다. 동기화 중에 문제가 발생하는 경우 테넌트당 하나의 리포지토리 또는 애플리케이션만 사용하는 것이 좋습니다.

우편 배달부 사용

엔드포인트 {{apiurl}}/v2/admin/repositories에 POST 요청을 보내고 자격 증명을
원시 본문의 JSON 형식 :

The screenshot shows the Postman application interface. At the top, it displays a POST request to the endpoint `({{apiurl}})/repositories`. Below the request type, there are tabs for Params, Authorization, Headers (9), Body (selected), Pre-request Script, Tests, and Settings. Under the Body tab, the content type is set to raw JSON. The JSON payload is as follows:

```

1
2   "url": "https://github.com/john/aicore-test",
3   "username": "john",
4   "password": "XXXXXX"
5

```

At the bottom of the interface, there are tabs for Body, Cookies, Headers (4), and Test Results. The Body tab is selected. Below these tabs, there are buttons for Pretty, Raw, Preview, Visualize, and Text (selected). The raw JSON response is shown as:

```

1
2   "message": "Repository has been on-boarded."
3

```

컬 사용하기

엔드포인트 {{apiurl}}/v2/admin/repositories에 POST 요청을 제출하고 다음을 포함합니다.

신임장:

```
컬 --location --request POST "$AI_API_URL/v2/admin/repositories" \
--header "승인: 전달자 $TOKEN" \
--header '콘텐츠 유형: 애플리케이션/json' \
--data-raw '{
    "url": "https://github.com/john/examplerrepo",
    "사용자 이름": "존",
    "비밀번호": "<GIT_PAT_USER_TOKEN>"
}'
```

다음과 같이 고유한 Git 저장소 세부정보를 지정합니다.

- url: git 저장소의 URL
- 사용자 이름: git 저장소에 액세스하는 (서비스) 사용자
- 비밀번호: git 개인 액세스 토큰. 자세한 내용은 [개인 액세스 토큰 생성](#)을 참조하세요..

6.1.2 폴더 동기화를 위한 애플리케이션 생성

문맥

Git 저장소를 등록한 후에는 템플릿을 동기화할 애플리케이션을 생성해야 합니다.

저장소. 첫 번째 동기화에는 시간이 좀 걸리지만 애플리케이션 상태를 확인하여 언제 동기화되는지 확인할 수 있습니다.

완벽한. 첫 번째 동기화 후에는 수동으로 요청한 약 3분마다 자동으로 동기화가 수행됩니다.

, 아니면 될 수 있다

우편 배달부 사용

절차

애플리케이션 세부정보를 포함하여 엔드포인트 {{apiurl}}/v2/admin/applications에 POST 요청을 보냅니다.

POST <{{apiurl}}/v2/admin/applications>

Params Authorization Headers (9) **Body** Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL **JSON**

```

1 {
2   "applicationName": "my-app",
3   "repositoryUrl": "https://github.com/john/examplerrepo",
4   "revision": "HEAD",
5   "path": "workflows"
6 }
```

Body Cookies Headers (5) Test Results 200 OK 276 ms 244 B

Pretty Raw Preview Visualize **JSON**

```

1 {
2   "id": "my-app",
3   "message": "Application has been successfully created."
4 }
```

- **applicationName:** 애플리케이션의 이름을 설정합니다. 이름은 3~64자 사이여야 하며 [A-Za-z0-9\-_]+와 일치해야 합니다. 저장소Url: 등록된 Git 저장소의 URL입니다.
-
- **revision:** 대상으로 할 개정입니다. <HEAD>는 가장 최근 항목을 나타냅니다. 경로: 동기화할 템플릿이 포함된 대상 폴더의 경로입니다.

각 응용 프로그램은 저장소의 특정 경로와 개정판을 가리키므로 동일한 저장소 URL에 대해 여러 응용 프로그램이 생성될 수 있습니다.

결과

GitOps 설정이 완료되면 git 저장소의 템플릿이 최대 3분 간격으로 SAP AI Core에 자동으로 동기화됩니다.

다음 단계

{{apiurl}}/v2/admin/applications/{{appName}}/status에 GET 요청을 보내 애플리케이션의 동기화 상태를 확인하세요. appName에는 애플리케이션을 생성할 때 지정한 애플리케이션 이름을 입력합니다.

```

1
2   "healthStatus": "Healthy",
3   "message": "successfully synced (all tasks run)",
4   "reconciledAt": "2021-11-23T10:27:49Z",
5   "source": {
6     "path": "workflows",
7     "repoURL": "https://github.com/john/examplerrepo",
8     "revision": "db611bb28be3c853d08867c08b52b8f733b4f7bf"
9   },
10  "syncFinishedAt": "2021-11-23T10:27:49Z",
11  "syncRessourcesStatus": [
12    {
13      "kind": "ServingTemplate",
14      "message": "servingtemplate.ai.sap.com/text-clf-infer-tutorial configured",
15      "name": "text-clf-infer-tutorial",
16      "status": "Synced"
17    }
18  ],
19  "syncStartedAt": "2021-11-23T10:27:48Z",
20  "syncStatus": "Synced"
21

```

컬 사용하기

애플리케이션 세부정보를 포함하여 엔드포인트 {{apiurl}}/v2/admin/applications에 POST 요청을 제출합니다.

```

컬 --location --request POST "$AI_API_URL/v2/admin/applications" \ --header "권한 부여: Bearer $TOKEN" \ --header 'Content-Type: application/json' \ --data-raw '{
  "applicationName": "my-app", "repositoryUrl": "https://github.com/john/examplerrepo", "revision": "HEAD", "path": "workflows"
}'

```

- **applicationName:** 애플리케이션의 이름을 설정합니다. 이름은 3~64자 사이여야 하며 [A-Za-z0-9\-_]+와 일치해야 합니다. 저장소Url: 등록된 Git 저장소의 URL입니다.
-
- **revision:** 대상으로 할 개정입니다. <HEAD>는 가장 최근 항목을 나타냅니다. 경로: 동기화할 템플릿이 포함된 대상 폴
- 더의 경로입니다.

각 응용 프로그램은 저장소의 특정 경로와 개정판을 가리기므로 동일한 저장소 URL에 대해 여러 응용 프로그램이 생성될 수 있습니다.

결과

GitOps 설정이 완료되면 git 저장소의 템플릿이 최대 3분 간격으로 SAP AI Core에 자동으로 동기화됩니다.

다음 단계

`{apiurl}/v2/admin/applications/{appName}/status`에 GET 요청을 제출하여 애플리케이션의 동기화 상태를 확인하세요.

```
컬 --location --request GET "$AI_API_URL/v2/admin/applications/{appName}/status" \ --header "Authorization: Bearer $TOKEN" \ --header
'Content-Type:
application/json'
```

applicationName에는 애플리케이션을 생성할 때 지정한 애플리케이션 이름을 입력합니다.

수동으로 애플리케이션 동기화

애플리케이션은 최대 3분 간격으로 자동으로 GitHub 저장소와 동기화됩니다. 아래 엔드포인트를 사용하여 수동으로 동기화를 요청하세요.

`{apiurl}/admin/applications/{appName}/refresh`

6.1.3 Git 저장소 편집

우편 배달부 사용

엔드포인트 `{apiurl}/v2/admin/repositories/{repositoryName}`에 PATCH 요청을 보내고 본문에 변경 사항을 포함합니다.

컬 사용하기

엔드포인트 `{apiurl}/v2/admin/repositories`에 PATCH 요청을 제출하고 변경 사항을 포함합니다.

```
컬 --location --request PATCH "$AI_API_URL/v2/admin/repositories" \ --header "권한 부여: Bearer $TOKEN" \ --header 'Content-Type:
application/json' \ --data-raw '{

"url": "https://github.com/john/examplerrepo", "사용자 이름": "john", "비밀번호":
"<GIT_PAT_USER_TOKEN>

}'
```

다음과 같이 고유한 Git 저장소 세부정보를 지정합니다.

- url: git 저장소의 URL
- 사용자 이름: git 저장소에 액세스하는 (서비스) 사용자
- 비밀번호: git 개인 액세스 토큰. 자세한 내용은 [개인 액세스 토큰 생성](#)을 참조하세요. .

6.1.4 Git 저장소 삭제

URL이 유효하지 않거나 오류가 포함된 경우 또는 저장소가 더 이상 필요하지 않은 경우 연결에서 Git 저장소를 제거합니다. Git 저장소가 제거되면 더 이상 애플리케이션의 소스 저장소로 선택할 수 없습니다.

우편 배달부 사용

DELETE 요청을 엔드포인트 {{apiurl}}/v2/admin/repositories/{{repositoryName}}에 보내고 저장소 이름을 포함합니다.

컬 사용하기

```
컬 --location --request DELETE "{{apiurl}}/v2/admin/repositories/{{repositoryName}}" \
```

6.2 리소스 그룹 관리

리소스 그룹은 사용자가 구성, 실행, 배포 및 아티팩트를 생성하거나 추가할 수 있는 고유한 전용 네임스페이스 또는 작업 영역 환경입니다. 훈련 작업이나 모델을 실행하는 데 사용됩니다.

서버.

리소스 그룹은 기계 학습 워크로드를 물리적으로 격리하고 사용 시나리오에 맞게 관련 리소스를 논리적으로 격리하는 데 사용됩니다.

테넌트가 등록되면 기본 리소스 그룹이 자동으로 생성됩니다. 기본 리소스 그룹 삭제할 수 없습니다.

관리자는 서비스 소비자 및 사용 시나리오에 따라 리소스 그룹을 생성, 편집 또는 삭제합니다.

실행, 배포, 구성 및 아티팩트와 같은 런타임 엔티티는 특정 리소스 그룹에 속하며 리소스 그룹 간에 공유되지 않습니다. 시나리오, 실행 파일 및 Docker 레지스터리 비밀은 테넌트 내의 모든 리소스 그룹에서 공유됩니다.

리소스 그룹은 인스턴스라고도 합니다.

기억하세요

구하의 SAP 글로벌 계정은 여러 계정으로 구성될 수 있습니다. 각 계정은 테넌트와 연결될 수 있습니다. 테넌트는 여러 리소스 그룹을 포함할 수 있습니다. 테넌트에는 항상 기본 리소스 그룹과 사용 시나리오에 대해 정의된 리소스 그룹이 포함됩니다.

최대 리소스 그룹 수는 테넌트 수준에서 50으로 제한됩니다. 이 제한에 도달하면 오류 메시지가 표시됩니다. 공간을 확보하려면 일부 리소스 그룹을 삭제하세요. 또는 할당량을 늘리려면 티켓을 제출하세요.

자세한 내용은 [리소스 그룹 삭제 \[페이지 75\]](#)를 참조하십시오.

[리소스 그룹 만들기 \[페이지 73\]](#)

[리소스 그룹 편집 \[페이지 74\]](#)

[리소스 그룹 삭제 \[페이지 75\]](#)

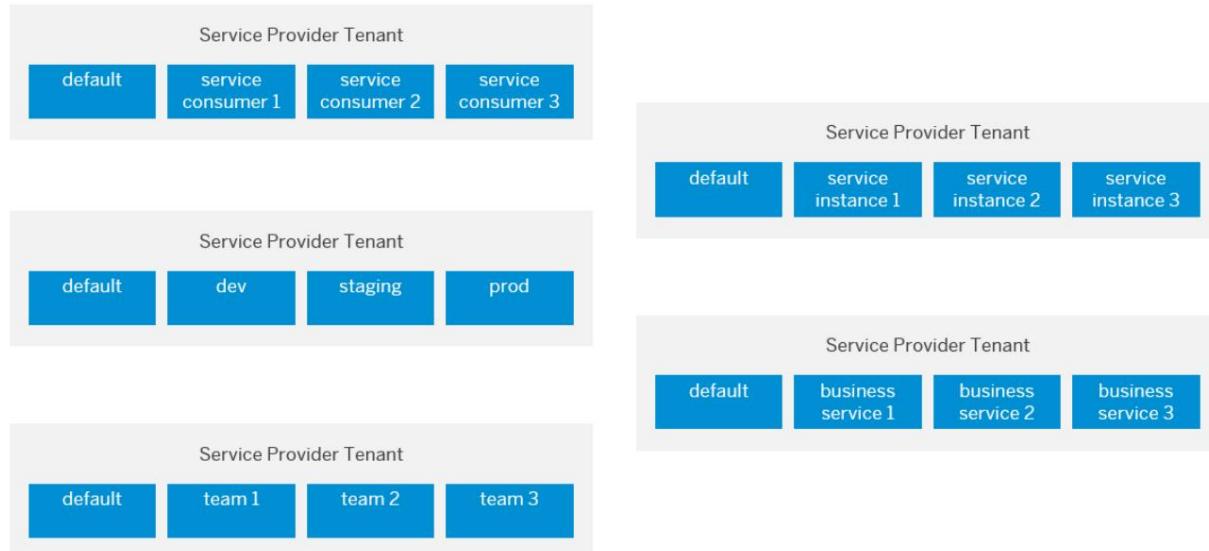
상위 주제: [관리 \[페이지 64\]](#)

리소스 그룹 수준 리소스

테넌트 수준의 실행 파일은 모든 리소스 그룹에서 공유됩니다. 반면, 실행, 배포, 구성 및 아티팩트와 같은 런타임 엔터티는 특정 리소스 그룹에 속하며 리소스 그룹 간에 공유될 수 없습니다. 마찬가지로 리소스 그룹 내에서 생성된 일반 비밀은 해당 그룹 내의 워크로드에만 사용됩니다.

리소스 그룹 헤더를 설정하여 리소스 그룹 수준에서 개체 저장소를 등록할 수 있습니다. 여러 리소스 그룹에 대해 동일한 IAM 사용자와 동일한 객체 저장소 버킷을 사용하지 않는 것이 좋습니다.

예제 리소스 그룹 매핑은 아래 그림에 요약되어 있습니다.



6.2.1 리소스 그룹 생성

전제조건

초기 설정을 완료했습니다. 자세한 내용은 [초기 설정 \[페이지 45\]](#)을 참조하십시오.

참고

리소스 그룹 ID의 길이는 최소 3자, 최대 253자여야 합니다. 첫 번째 문자와 마지막 문자는 소문자, 대문자 또는 숫자여야 합니다. 두 번째부터 두 번째까지의 문자 항목에는 소문자, 대문자, 숫자, 마침표 또는 하이픈이 포함될 수 있습니다. 다른 특수 문자는 허용되지 않습니다.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/resourceGroups에 POST 요청을 보냅니다.

2. 요청 본문으로 원시 라디오 버튼을 선택하고 다음을 입력합니다.

```
{
  "resourceGroupId": "<리소스 그룹의 ID>"
}
```

POST <{{apiurl}}/v2/admin/resourceGroups>

Params Authorization ● Headers (11) Body ● Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL **JSON** ▾

```

1  {
2   ...
3   "resourceGroupId": "trial-1"
4   ...
5 }
```

3. 요청을 보냅니다.

리소스 그룹 생성 요청이 수락되었음을 확인하는 202 응답을 받게 됩니다.

제한사항

최대 리소스 그룹 수는 테넌트 수준에서 50으로 제한됩니다. 이 제한에 도달하면 오류 메시지가 표시됩니다. 공간을 확보하려면 일부 리소스 그룹을 삭제하세요. 또는 할당량을 늘리려면 티켓을 제출하세요.

자세한 내용은 [리소스 그룹 삭제 \[페이지 75\]](#)를 참조하십시오.

컬 사용하기

1. 다음을 전송하여 리소스 그룹을 생성합니다.

```
컬 --location --request POST "$AI_API_URL/v2/admin/resourceGroups" --header
"권한 부여: Bearer $TOKEN" --header 'Content-Type: application/json' -- data-raw '{ "resourceGroupId": "<리소스 그룹의 ID>" }'
```

제한사항

최대 리소스 그룹 수는 테넌트 수준에서 50으로 제한됩니다. 이 제한에 도달하면 오류 메시지가 표시됩니다. 공간을 확보하려면 일부 리소스 그룹을 삭제하세요. 또는 할당량을 늘리려면 티켓을 제출하세요.

자세한 내용은 [리소스 그룹 삭제 \[페이지 75\]](#)를 참조하십시오.

상위 주제: [리소스 그룹 관리 \[페이지 71\]](#)

관련 정보

[리소스 그룹 편집 \[페이지 74\]](#)

[리소스 그룹 삭제 \[페이지 75\]](#)

[다중 테넌트 \[페이지 259\]](#)

6.2.2 리소스 그룹 편집

참고

리소스 그룹 ID의 길이는 최소 3자, 최대 253자여야 합니다. 첫 번째 문자와 마지막 문자는 소문자, 대문자 또는 숫자여야 합니다. 두 번째부터 두 번째까지의 문자 항목에는 소문자, 대문자, 숫자, 마침표 또는 하이픈이 포함될 수 있습니다. 다른 특수 문자는 허용되지 않습니다.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/resourceGroups/에 PATCH 요청을 보냅니다.

{{resource_group_name}}을 본문으로 사용:

```
{
  "resourceGroupId": "<리소스 그룹의 ID>"
}
```

컬 사용하기

1. 다음을 전송하여 리소스 그룹을 생성합니다.

```
컬 --location --request PATCH "$AI_API_URL/v2/admin/resourceGroups/ {{resource_group_name}}" --header "승인:
Bearer $TOKEN" --header
'Content-Type: application/json' --data-raw '{ "resourceGroupId": "<리소스 그룹의 ID>" }'
```

상위 주제: [리소스 그룹 관리 \[페이지 71\]](#)

관련 정보

[리소스 그룹 만들기 \[페이지 73\]](#)

[리소스 그룹 삭제 \[페이지 75\]](#)

6.2.3 리소스 그룹 삭제

우편 배달부 사용

엔드포인트 {{apiurl}}/v2/admin/resourceGroups/{{resource_group_name}}에 DELETE 요청을 보냅니다.

컬 사용하기

1. 다음을 전송하여 리소스 그룹을 생성합니다.

```
컬 --location --request POST "$AI_API_URL/v2/admin/resourceGroups/ {{resource_group_name}}" --header "승인:
Bearer $TOKEN" --header
```

```
'Content-Type: application/json' --data-raw '{ "resourceGroupId": "<리소스 그룹의 ID>" }'
```

상위 주제: [리소스 그룹 관리 \[페이지 71\]](#)

관련 정보

[리소스 그룹 만들기 \[페이지 73\]](#)

[리소스 그룹 편집 \[페이지 74\]](#)

6.3 객체 저장소 비밀 관리

6.3.1 객체 저장소 비밀 등록

SAP AI Core를 클라우드 개체 저장소에 연결하고 개체 저장소 비밀을 사용하여 액세스를 관리하세요. 연결된 스토리지는 SAP AI Core용 Metaflow Library의 데이터 세트, 모델 및 기타 캐시 파일을 저장합니다.

클라우드 저장소 자격 증명은 비밀을 사용하여 관리됩니다. 비밀은 자격 증명을 손상시키지 않고 디렉터리와 도구 전반에 걸쳐 연결을 허용하고 제어하는 수단입니다.

전제조건

초기 설정을 완료했습니다. 자세한 내용은 [초기 설정 \[페이지 45\]](#)을 참조하십시오.

문맥

SAP AI Core는 Amazon S3, OSS(Alicloud Object Storage Service), SAP HANA Cloud, Data Lake 및 Azure Blob Storage와 같은 여러 하이퍼스케일러 개체 저장소를 지원합니다.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/objectStoreSecrets에 POST 요청을 보냅니다.
2. 요청 본문으로 원시 라디오 버튼을 선택하고 개체 저장소 비밀 세부 정보를 입력합니다.

참고

Azure Blob Storage를 제외한 모든 저장소 유형의 경우 모든 <data> 필드가 필수입니다. Azure의 경우 필수 필드가 지정됩니다.

- Amazon S3의 경우:

```
{
  "name": "<구하의 식별자>", "data": {
    "AWS_ACCESS_KEY_ID": "<AWS 액세스 키 ID>",
    "AWS_SECRET_ACCESS_KEY": "<AWS 보안 액세스 키>"
  },
  "type": "S3", "bucket": "<S3 버킷 이름>", "endpoint": "<S3 끝점>", "region": "<S3 지역>", "pathPrefix": "<버킷 이름 뒤에 오는 경로 접두사>"}
}
```

- OSS(Alicloud 개체 스토리지 서비스)의 경우:

```
{
  "name": "default", "type": "oss",
  "pathPrefix": "<버킷 이름에 추가될 경로 접두사>", "data": { "BUCKET": "<bucket-name>", "ENDPOINT": "oss-cn-shanghai.aliyuncs.com", "name": "default",
    "type": "S3", "bucket": { "name": "default",
      "type": "oss", "pathPrefix": "<버킷 이름에 추가될 경로 접두사>", "data": { "REGION": "", "ACCESS_KEY_ID": "xxxxx",
      "SECRET_ACCESS_KEY": "xxxxx"
    }
  }
}
```

- SAP HANA Cloud, 데이터 레이크의 경우:

```
"BUCKET": "{ "name": "default", "type": "webhdfs", "pathPrefix": "<추가할 경로 접두사>", "data": { // 예 https://c32727c8-4260-4c37-b97f-ede322dcfa8.files.hdl.canary-eu10.hanacloud.ondemand.com
  "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-eu10.hanacloud.ondemand.com", "TLS_CERT": "-----\n인증서 시작-----\nMIICmjCCAYIxxxxxxxxxxxxxR4wtC32bGO66D+Jc8RhalA==\n-----인증서 종료-----\n", "TLS_KEY": "-----개인 키 시작-----\n-----\n\nMIIEvQIBADANBgkxxxxxxxxxxxxnor+rtZHhhzEfX5dYLCS5Pww=\n-----비공개 종료-----\n",
  "헤더": "\\"x-sap-filecontainer\\": "<파일-컨테이너-이름>\",
  \"콘텐츠 유형\": \"응용 프로그램/목록-스트림\"
}
}
```

- Azure Blob 저장소의 경우:

```
{
    "name": "default", "type": "azure", "pathPrefix": "<추가 할 경로 접두사>", "data": { "CONTAINER_URI": "https://account_name.blob.core.windows.net/" }

    컨테이너_이름", //필수
    "REGION": "<지역 이름>", //선택 사항 //선택 사항
    "CLIENT_ID": "<azure 클라이언트 ID>",
    "CLIENT_SECRET": "<azure 클라이언트 비밀>", //선택 사항 //필수
    "STORAGE_ACCESS_KEY": "sas_token",
    "TENANT_ID": "azure 테넌트 ID", //선택 사항
    "SUBSCRIPTION_ID": "구독 ID", //선택 사항
}
}
```

3. 요청을 보냅니다.

다음은 S3를 사용한 요청을 보여줍니다.

The screenshot shows a Postman interface with the following details:

- Method:** POST
- URL:** {{apiurl}}/v2/admin/objectStoreSecrets
- Body (raw JSON):**

```

1 {
2     "name": "default",
3     "type": "S3",
4     "bucket": "xxxx",
5     "endpoint": "s3-eu-central-1.amazonaws.com",
6     "pathPrefix": "xxxx",
7     "region": "eu-central-1",
8     "data": [
9         {
10            "AWS_ACCESS_KEY_ID": "xxxx",
11            "AWS_SECRET_ACCESS_KEY": "xxxx"
12        }
      ]
    }
  
```

컬 사용하기

1. /v2/admin/objectStoreSecrets 엔드포인트를 사용하여 객체 저장소 비밀 세부 정보를 등록합니다.

참고

Azure Blob Storage를 제외한 모든 저장소 유형의 경우 모든 <data> 필드가 필수입니다. Azure의 경우 필수 필드가 지정됩니다.

- Amazon S3의 경우:

```
curl -location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \ --header "Authorization: Bearer $TOKEN" \ --header 'Content-Type: application/json' \ --header 'AI-Resource-Group: <리소스 그룹>' \ --data-
raw '{ "name": "default", "type": "S3", "bucket": "<S3 버킷 이름>" ,
```

```

    "endpoint": "<S3 엔드포인트>", "pathPrefix": "<버킷 이름  
뒤에 오는 경로 접두사>", "region": "<S3 리전>", "data": {  
  
        "AWS_ACCESS_KEY_ID": "<AWS 액세스 키 ID>,"  
        "AWS_SECRET_ACCESS_KEY": "<AWS 보안 액세스 키>"  
    }  
}
}

```

- OSS(Alicloud 개체 스토리지 서비스)의 경우:

```

curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \ --header "Authorization: Bearer $TOKEN" \ --header 'Content-Type: application/json' \ -header 'AI-Resource-Group: <리소스 그룹>' \ --data-raw  
'{"name": "default", "type": "oss", "pathPrefix": "<버킷 이름에 추가될 경로 접두사>",  
"data": {"BUCKET": "<버킷 이름>", "ENDPOINT": "oss-cn-shanghai.aliyuncs.com", "REGION":  
"", "ACCESS_KEY_ID":  
        "xxxxx", "SECRET_ACCESS_KEY":  
        "xxxxx"  
    }  
}'
}

```

- SAP HANA Cloud, 데이터 레이크의 경우:

```

curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \ --header "Authorization: Bearer $TOKEN" \ --header 'Content-Type: application/json' \ -header 'AI-Resource-Group: <리소스 그룹>' \ --data-raw  
'{"name": "default", "type": "webhdfs", "pathPrefix": "<추가할 경로 접두사>", "data":  
{// 예: https://c32727c8-4260-4c37-b97f-ede322dcfa8f.files.hdl.canary-  
  
eu10.hanacloud.ondemand.com  
    "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-  
  
  
eu10.hanacloud.ondemand.com", "TLS_CERT": "-----  
인증서 시작----  
\nMIIICmjCCAYIxxxxxxxxxxR4wtC32bGO66D+Jc8RhIA==\n---END  
인증서-----\n",  
    "TLS_KEY": "-----개인 키 시작-----  
\nMIIEvQIBADANBgkqxxxxxxxxxxnor+rtZHhzEfX5dYLCS5Pww=\n---개인 종료  
키-----\n",  
    "헤더": "{\"x-sap-filecontainer\": \"<파일 컨테이너-이름>\",  
\"콘텐츠 유형\": \"응용 프로그램/옥텟-스트림\""}  
}  
}'
}

```

- Azure Blob 저장소의 경우:

```

curl --location --request POST "$AI_API_URL/v2/admin/objectStoreSecrets" \ --header "Authorization: Bearer $TOKEN" \ --header 'Content-Type: application/json' \ -header 'AI-Resource-Group: <리소스 그룹>' \ --data-raw  
'{"name": "default", "type": "azure", "pathPrefix": "<추가할 경로 접두사>", "data":  
{"CONTAINER_URI": "https://account_name.blob.core.windows.net/컨테이너_이름", //필수
}
}

```

```

    "REGION": "<지역 이름>", //선택 사항 //선택 사항
    "CLIENT_ID": "<azure 클라이언트 ID>",
    "CLIENT_SECRET": "<azure 클라이언트 비밀>", //선택 사항 //필수
    "STORAGE_ACCESS_KEY": "sas_token",
    "TENANT_ID": "azure 테넌트 ID", //선택 사항
    "SUBSCRIPTION_ID": "구독 ID", //선택 사항
}
}

```

참고

입력 아티팩트에만 해당

이름에 서로 다른 값을 사용하여 여러 비밀을 생성할 수 있지만 먼저 기본값을 생성해야 합니다.

팁

pathPrefix는 여러 프로젝트에 대해 동일한 버킷을 공유하는 경우 유용합니다. 예를 들어 프로젝트 폴더의 이름을 `my-ml-project1`로 설정할 수 있습니다. 그러면 모든 데이터가 해당 폴더에 저장됩니다.

참고

AI-Resource-Group 헤더를 지정하지 않으면 `<Resource Group>`에 "default" 값이 자동으로 할당됩니다.

제한사항

디렉터리를 가리키는 출력 아티팩트와 함께 SAP HANA Data Lake 객체 저장소를 사용하는 경우 워크플로 템플릿에서 `archive: none: {}`을 사용하여 아티팩트 보관을 비활성화할 수 없습니다. 자세한 내용은 [워크플로 템플릿 \[페이지 106\]](#)를 참조하십시오.

6.3.2 객체 저장소 비밀번호 편집

우편 배달부 사용

- 엔드포인트 `{{apiurl}}/v2/admin/objectStoreSecrets/{{objectStoreName}}`에 패치 요청을 보냅니다.
- 요청 본문으로 [원시](#) 라디오 버튼을 선택하고 객체 저장소 비밀 세부 정보를 입력합니다.

참고

Azure Blob Storage를 제외한 모든 저장소 유형의 경우 모든 `<data>` 필드가 필수입니다. Azure의 경우 필수 필드가 지정됩니다.

- Amazon S3의 경우:

```
{
}
```

```

    "name": "default", "type": "S3",
    "bucket": "<S3 버킷 이름",
    ">", "endpoint": "<S3 엔드포인트>", "pathPrefix": "<경로
    접두사 버킷 이름 뒤에 옵니다.>", "region": "<S3 지역>",
    "data":{

        "AWS_ACCESS_KEY_ID": "<AWS 액세스 키 ID>",
        "AWS_SECRET_ACCESS_KEY": "<AWS 보안 액세스 키>"
    }
}

```

- OSS(Alicloud 개체 스토리지 서비스)의 경우:

```

{
    "name": "default", "type": "oss",
    "pathPrefix": "<버킷 이름>
    추가될 경로 접두사>", "data":{ "BUCKET": "<bucket-name>", "엔드포인트": "oss-cn-shanghai.aliyuncs.com", "지역": "", "ACCESS_KEY_ID": "xxxxxx", "SECRET_ACCESS_KEY": "xxxxxx"
}

}

```

- SAP HANA Cloud, 데이터 레이크의 경우:

```

{
    "name": "default", "type": "webhdfs", "pathPrefix": "<추가
    할 경로 접두사>", "data":{ // 예 https://c32727c8-4260-4c37-b97f-
        ede322dcfa8f.files.hdl.canary-
        eu10.hanacloud.ondemand.com "HDFS_NAMENODE":
            "https://<file-container-name>.files.hdl.canary-
            eu10.hanacloud.ondemand.com", "TLS_CERT": "-----"
        인증서 시작----\nMIIcmjCCAYIxxxxxxxxxxxxR4wtC32bGO66D+Jc8RhalA==\n-----
        인증서 종료----\n", "TLS_KEY": "-----개인 키 시작-----
        \nMIIEvQIBADANBgkxxxxxxxxxxxxnor+rtZHhhzEfX5dYLCS5Pww=\n----비공개 종료
        키-----\n",
        "헤더": "[\"x-sap-filecontainer\": \"<파일-컨테이너-이름>\",
        \"콘텐츠 유형\": \"응용 프로그램/애토-스트림\"]"
    }
}

```

- Azure Blob 저장소의 경우:

```

{
    "name": "default", "type": "azure", "pathPrefix": "<추가
    할 경로 접두사>", "data":{

        "CONTAINER_URI": "https://<account_name>.blob.core.windows.net/컨테이너_이름", //필수 "REGION": "<지역 이름>", //
        선택 사항 //선택 사항 "CLIENT_ID": "<azure 클라이언트
        id>", "CLIENT_SECRET": "<azure client secret>", //선택 사항 //필수 "STORAGE_ACCESS_KEY": "TENANT_ID": "azure 테넌트 id", //선택 사항 //선택 사항 "SUBSCRIPTION_ID": "구독 ID",
        "sas_token",
    }
}

```

3. 요청을 보냅니다.

컬 사용하기

1. \$AI_API_URL/v2/admin/ 엔드포인트를 사용하여 객체 저장소 비밀 세부 정보를 등록합니다.

```
objectStoreSecrets/{{objectStoreName}}.
```

참고

Azure Blob Storage를 제외한 모든 저장소 유형의 경우 모든 <data> 필드가 필수입니다. Azure의 경우 필수 필드가 지정됩니다.

- Amazon S3의 경우:

```
컬 --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/{{objectStoreName}}" \ --header "권한 부여: Bearer $TOKEN" \ --
header 'Content-Type: application/
json' \ - -header 'AI-Resource-Group: <리소스 그룹>' \ --data-raw '{ "name": 
"default", "type": "S3", "bucket": "<S3 버킷 이름>", "endpoint": "<S3 엔드포인트>",
"pathPrefix": "<버킷 이름 뒤에 오는 경로 접두사>", "region": "<S3 리전>", "data": { 

"AWS_ACCESS_KEY_ID": "<AWS 액세스 키 ID>",
"AWS_SECRET_ACCESS_KEY": "<AWS 보안 액세스 키>" 
} 
}'
```

- OSS(Alicloud 개체 스토리지 서비스)의 경우:

```
컬 --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/{{objectStoreName}}" \ --header "권한 부여: Bearer $TOKEN" \ --
header 'Content-Type: application/
json' \ - -header 'AI-Resource-Group: <리소스 그룹>' \ --data-raw '{ "name": 
"default", "type": "oss", "pathPrefix": "<추가할 경로 접두사 bucketname>", "data": 
{ "BUCKET": "<bucket-name>", "ENDPOINT": "oss-cn-shanghai.aliyuncs.com", "REGION": 
"", "ACCESS_KEY_ID": 
"xxxxx ", "SECRET_ACCESS_KEY": 
"xxxxx" 
} 
}'
```

- SAP HANA Cloud, 데이터 레이크의 경우:

```
컬 --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/{{objectStoreName}}" \ --header "권한 부여: Bearer $TOKEN" \ --
header 'Content-Type: application/
json' \ - -header 'AI-리소스-그룹: <리소스 그룹>' \ --data-raw '{
```

```

    "이름": "기본값",
    "유형": "webhdfs",
    "pathPrefix": "<추가할 경로 접두어>",
    "데이터": {
        // 예: https://c32727c8-4260-4c37-b97f-ed322dcfa8f.files.hdl.canary-
eu10.hanacloud.ondemand.com
        "HDFS_NAMENODE": "https://<file-container-name>.files.hdl.canary-
eu10.hanacloud.ondemand.com",
        "TLS_CERT": "-----인증서 시작-----
\nMIICmjCCAYIxxxxxxxxxxR4wtC32bGO66D+Jc8RhalA==\n----END
인증서-----\n",
        "TLS_KEY": "-----개인 키 시작-----
\nMIIEvQIBADANBgkqxxxxxxxxxxnor+rtZhhzEfX5dYLCS5Pww=\n---비공개 종료
키-----\n",
        "헤더": "{\"x-sap-filecontainer\": \"<파일-컨테이너-이름>\",
        \"콘텐츠 유형\": \"응용 프로그램/옥텟-스트림\""
    }
}
}

```

- Azure Blob 저장소의 경우:

```

curl --location --request PATCH "$AI_API_URL/v2/admin/objectStoreSecrets/
{{objectStoreName}}" \
--header "승인: 전달자 $TOKEN" \
--header '콘텐츠 유형: 애플리케이션/json' \
--header 'AI-리소스-그룹: <리소스 그룹>' \
--data-raw '{
    "이름": "기본값",
    "유형": "하늘색",
    "pathPrefix": "<추가할 경로 접두어>",
    "데이터": {
        "CONTAINER_URI": "https://account_name.blob.core.windows.net/
컨테이너_이름", //필수
        "REGION": "<지역 이름>", //선택 사항
        "CLIENT_ID": "<azure 클라이언트 ID>", //선택 과목
        "CLIENT_SECRET": "<azure 클라이언트 비밀>", //선택 사항
        "STORAGE_ACCESS_KEY": "sas_token", //필수의
        "TENANT_ID": "azure 테넌트 ID", //선택 사항
        "SUBSCRIPTION_ID": "구독 ID", //선택 과목
    }
}
}

```

참고

입력 아티팩트에만 해당

이름에 서로 다른 값을 사용하여 여러 비밀을 생성할 수 있지만 먼저 기본값을 생성해야 합니다.

팁

pathPrefix는 여러 프로젝트에 대해 동일한 버킷을 공유하는 경우 유용합니다. 당신은 당신의 이름을 설정할 수 있습니다 예를 들어 프로젝트 폴더를 [my-ml-project1](#)로 지정합니다. 그러면 모든 데이터가 해당 폴더에 저장됩니다.

참고

AI-Resource-Group 헤더를 지정하지 않으면 <Resource Group>에 다음 값이 할당됩니다.

자동으로 "기본값"이 됩니다.

6.3.3 객체 저장소 비밀 삭제

우편 배달부 사용

끝점에 DELETE 요청 보내기

```
{{apiurl}}/v2/admin/objectStoreSecrets/{{objectStoreName}}
```

컬 사용하기

```
컬 --location --request DELETE "$AI_API_URL/v2/admin/objectStoreSecrets/{{objectStoreName}}"\
```

6.4 Docker 레지스트리 비밀 관리

6.4.1 Docker 레지스트리 비밀 등록

Docker는 원격 컨테이너에서 애플리케이션을 패키징하고 실행합니다. SAP AI Core를 Docker 저장소에 연결하고 Docker 레지스트리 비밀을 사용하여 액세스를 관리하세요.

전제조건

초기 설정을 완료했습니다. 자세한 내용은 [초기 설정 \[페이지 45\]](#)을 참조하십시오.

문맥

Docker 자격 증명은 비밀을 사용하여 관리됩니다. 비밀은 자격 증명을 손상시키지 않고 디렉터리와 도구 간의 연결을 허용하고 제어합니다.

Docker 레지스트리 비밀을 사용하면 SAP AI Core에 권한을 부여하여 Docker 리포지토리에서 프라이빗 Docker 이미지를 가져올 수 있습니다. Docker 이미지 가져오기를 인증하기 위해 워크플로에서 비밀 이름을 지정합니다.

자세한 내용은 [워크플로 템플릿 \[페이지 106\]](#) 및 [제공 템플릿 \[페이지 154\]](#)을 참조하십시오.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/dockerRegistrySecrets에 POST 요청을 보냅니다. 2. 요청 본문으로 원시 라디오 버튼을 선택하고 다음을 입력합니다.

```
{
  "이름": "mydockerregistry",
  "데이터": {
    ".dockerconfigjson": "{\"auths\":{\"your.private.registry\":{\"username\":\"john\",\"password\":\"docker-accesstoken-or-password\"}}}"
  }
}
```

- 이름: Docker 레지스트리 비밀의 이름을 설정합니다. 이는 귀하가 선택한 비밀 식별자입니다. 예에서 이름은 "mydockerregistry"입니다.
- 데이터: Docker 레지스트리 비밀을 나타내는 JSON 문자열을 입력합니다.

3. 요청을 보냅니다.

The screenshot shows the Postman interface with a POST request to {{apiurl}}/v2/admin/dockerRegistrySecrets. The request body is a JSON object:

```

1 {
2   ...
3   "name": "mydockerregistry",
4   "data": {
5     ".dockerconfigjson": {
6       "auths": {
7         "your.private.registry": {
8           "username": "john",
9           "password": "yourpassword"
10        }
11      }
12    }
13  }
14}

```

The response status is 202 Accepted. The message in the response body is "secret has been created".

4. Docker 레지스트리 비밀이 생성된 후 이를 템플릿에서 이미지 풀 비밀로 참조하십시오.

```

spec:
  imagePullSecrets:
    - name: <Docker 레지스트리 비밀 이름>

```

컬 사용하기

{{apiurl}}/v2/admin/dockerRegistrySecrets 엔드포인트에 POST 요청을 제출합니다. 요청 본문에 다음 매개변수를 포함합니다.

- 이름: Docker 레지스트리 비밀의 이름을 설정합니다.
- 데이터: Docker 레지스트리 비밀을 나타내는 JSON 문자열을 입력합니다.

샘플 코드

```
{
  "이름": "mydockerregistry",
  "데이터": {
    ".dockerconfigjson": [
      {
        "auths": {
          "your.private.registry": {
            "username": "john",
            "password": "docker-access-token-or-password"
          }
        }
      }
    ]
  }
}
```

참고

<http://hub.docker.com>의 공개 Docker 레지스트리를 사용하는 경우, <"auths"> 변수 입력에 <https://index.docker.io> 형식으로 Docker URL을 제공해야 합니다.

1. {{apiurl}}/v2/admin/dockerRegistrySecrets 엔드포인트에 POST 요청을 제출하고 저장소에 대한 자격 증명과 함께 Docker 레지스트리 비밀의 이름을 포함합니다. 예를 들어

샘플 코드

```
$ curl --location --request POST "$AI_API_URL/v2/admin/dockerRegistrySecrets" --header "Authorization: Bearer $TOKEN" --header 'Content-Type: application/json' -d-data-raw '{
  "name": "mydockerregistry",
  "data": {
    ".dockerconfigjson": [
      {
        "auths": {
          "my.docker.repositories.io": {
            "username": "$USERNAME",
            "password": "$PWD"
          }
        }
      }
    ]
  }
}' --verbose
{
  "message": "비밀번호가 생성되었습니다"
}
```

2. Docker 레지스트리 비밀이 생성된 후 이를 템플릿에서 이미지 풀 비밀로 참조하십시오.

예를 들어

소스 코드

```
spec:
  imagePullSecrets:
    - name: <Docker 레지스트리 비밀 이름>
```

6.4.2 Docker 레지스트리 비밀 편집

Docker는 원격 컨테이너에서 애플리케이션을 패키징하고 실행합니다. SAP AI Core를 Docker 저장소에 연결하고 Docker 레지스트리 비밀을 사용하여 액세스를 관리하세요.

Docker 자격 증명은 비밀을 사용하여 관리됩니다. 비밀은 자격 증명을 손상시키지 않고 디렉터리와 도구 간의 연결을 하용하고 제어합니다.

Docker 레지스트리 비밀을 사용하면 SAP AI Core에 권한을 부여하여 Docker 리포지토리에서 프라이빗 Docker 이미지를 가져올 수 있습니다. Docker 이미지 가져오기를 인증하기 위해 워크플로에서 비밀 이름을 지정합니다.

자세한 내용은 [워크플로 템플릿 \[페이지 106\]](#) 및 [제공 템플릿 \[페이지 154\]](#)를 참조하십시오.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/dockerRegistrySecrets/에 PATCH 요청을 보냅니다.

 {{dockerRegistryName}}

2. 요청 본문으로 원시 라디오버튼을 선택하고 다음을 입력합니다.

```
{
  "이름": "mydockerregistry", "데이터": 

  { ".dockerconfigjson": "{\"auths\":{\"your.private.registry\": {\"사용자 이름\":\"john\", \"password\":\"docker-accesstoken-or-password\"}}}" }
```

- 이름: Docker 레지스트리 비밀의 이름을 설정합니다. 데이터: Docker 레지스
- 트리 비밀을 나타내는 JSON 문자열을 입력합니다.

3. 요청을 보냅니다.

컬 사용하기

\$AI_API_URL/v2/admin/dockerRegistrySecrets/{{dockerRegistryName}} 엔드포인트에 PATCH 요청을 제출합니다. 요청 본문에 다음 매개변수를 포함합니다.

- 이름: Docker 레지스트리 비밀의 이름을 설정합니다.
- 데이터: Docker 레지스트리 비밀을 나타내는 JSON 문자열을 입력합니다.

샘플 코드

```
{
  "이름": "mydockerregistry", "데이터": 

  { ".dockerconfigjson": "{\"auths\":{\"your.private.registry\": {\"사용자 이름\":\"john\", \"password\":\"docker-accesstoken-or-password\"}}}" }
```

참고

<http://hub.docker.com> 의 공개 Docker 레지스트리를 사용하는 경우, <\"auths\"> 변수 입력에 https://index.docker.io 형식으로 Docker URL을 제공해야 합니다.

1. \$AI_API_URL/v2/admin/dockerRegistrySecrets/{{dockerRegistryName}} 엔드포인트에 PATCH 요청을 제출하고 Docker 레지스트리 비밀 이름과 리포지토리 자격 증명을 포함합니다. 예를 들어

샘플 코드

```
$ 컬 --location --request PATCH "$AI_API_URL/v2/admin/dockerRegistrySecrets/
{{dockerRegistryName}}" --header "승인:
Bearer $TOKEN" --header 'Content-Type: application/json' --data-raw '{ "name": "mydockerregistry", "data": {
```

```

        ".dockerconfigjson": "{\"auths\":{\"my.docker.repositories.io\":{\"username\": \"$USERNAME\", \"password\": \"$PWD
    \"}}}"
}

```

6.4.3 Docker 레지스트리 비밀 삭제

docker 레지스트리 비밀을 삭제하면 docker 레지스트리에 대한 액세스가 제거됩니다.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/dockerRegistrySecrets에 DELETE 요청을 보냅니다.
`curl -XDELETE {{apiurl}}/v2/admin/dockerRegistrySecrets/{{dockerRegistryName}}`

curl 사용하기

1. \$AI_API_URL/v2/admin/dockerRegistrySecrets/{{dockerRegistryName}} 엔드포인트에 DELETE 요청을 제출합니다.

6.5 일반 비밀 관리

6.5.1 일반 비밀 생성

일반 비밀은 SAP AI Core에 리소스를 노출하지 않고 리소스 그룹을 활용할 수 있는 권한을 부여합니다.
 신임장.

일반 비밀은 SAP AI Core가 오페스트레이션 계층인 통합 사용 사례와 같이 시스템 비밀을 적용할 수 없는 경우 중요한 정보를 저장하는 데 추가로 사용됩니다.

SAP AI Core를 사용하면 선택적으로 기본 테넌트 범위 또는 리소스 그룹 수준의 두 가지 수준에서 일반 암호를 사용할 수 있습니다.

일반 비밀은 시스템 비밀(예: 개체 저장소, Docker 레지스트리 등)과 다르며 API를 통해 기본 테넌트 또는 각 리소스 그룹에 대해 중요한 정보를 저장하는 데 사용할 수 있습니다. 후자는 실행 또는 배포 시 환경 변수 또는 볼륨 마운트로 컨테이너에 연결할 수 있습니다.

리소스 그룹에 일반 비밀을 생성하려면 엔드포인트 {{apiurl}}/v2/admin/secrets에 POST 요청을 보냅니다. API는 민감한 데이터가 Base64로 인코딩될 것으로 예상합니다. Linux 또는 Mac OS에서 다음 명령을 사용하여 Base64 형식으로 데이터를 쉽게 인코딩할 수 있습니다. `echo -n 'my-sensitive-data' | base64`

베이스64

전제조건

초기 설정을 완료했습니다. 자세한 내용은 [초기 설정 \[페이지 45\]](#)를 참조하십시오.

우편 배달부 사용

1. POST 요청을 보내고 URL {{apiurl}}/v2/admin/secrets를 입력합니다.
2. 요청 본문으로 원시 라디오 버튼을 선택하고 자격 증명을 JSON 형식으로 입력합니다.

```
{
  "이름": "MY_GENERIC_SECRET",
  "데이터": {
    "some-credential": "bXktc2VjcmV0LWNyZWRlbnRpYWw=",
    "기타 자격 증명": "bXktc2VjcmV0LW90aGVyLWNyZWRlbnRpYWw="
  }
}
```

참고

Unix 환경으로 사용하기 쉽도록 비밀 이름은 하이픈 없이 작성됩니다.
나중에 변수. 관례대로 대문자로 쓴다.

- 이름: 일반 비밀의 이름을 설정합니다.
 - 데이터: 일반 비밀을 나타내는 JSON 문자열을 입력합니다.
3. AI-Tenant-Scope 및 AI-Resource-Group 헤더를 통해 요청 범위를 지정합니다.
- AI-테넌트-범위 : true입니다. 작업은 기본 테넌트 수준에서 수행됩니다.
 - AI-리소스-그룹 : <리소스-그룹-이름>. 작업은 리소스 그룹 수준에서 수행됩니다.

이 예에서는 리소스 그룹 수준을 사용하고 있습니다.

KEY	VALUE
<input checked="" type="checkbox"/> AI-Resource-Group	default

4. 요청을 보냅니다.

```

POST {{apiurl}}/v2/admin/secrets
{
  "name": "my-generic-secret",
  "data": {
    "some-credential": "bXktc2VjcmV0LwNyZwRlbRpYWh=",
    "other-credentials": "bXktc2VjcmV0Lw90aGyLhMyZwRlbRpYWh="
  }
}
  
```

Body Cookies Headers (5) Test Results
Pretty Raw Preview Visualize JSON ↻

Status: 200 OK Time: 167 ms Size: 244 B Save Response ↻

```

{
  "message": "secret has been created",
  "name": "my-generic-secret"
}
  
```

컬 사용하기

v2/admin/secrets 앤드포인트에 POST 요청을 제출하고 일반 비밀 이름과 자격 증명을 포함합니다. AI-Tenant-Scope 및 AI-Resource-Group을 통해 범위를 지정 합니다.

- AI-테넌트-범위 : true입니다. 작업은 기본 테넌트 수준에서 수행됩니다.
- AI-리소스-그룹 : <리소스-그룹-이름>. 작업은 리소스 그룹 수준에서 수행됩니다.

컬 --location --request POST "\$AI_API_URL/v2/admin/secrets" \ --header "권한 부여: Bearer \$TOKEN" \ --header 'Content-Type: application/json' \ --header 'AI-Resource-Grp: 기본값' \ --data-raw '{

```

"이름": "MY_GENERIC_SECRET", "데이터": { "some-
credential": "bXktc2Vuc2l0aXZlLWRhdGE="
}
}'
```

참고

비밀 이름은 나중에 Unix 환경 변수로 쉽게 사용할 수 있도록 하이픈 없이 작성됩니다. 관례대로 대문자로 쓴다.

6.5.2 모든 일반 비밀 나열

우편 배달부 사용

1. 앤드포인트 {{apiurl}}/v2/admin/secrets에 GET 요청을 보냅니다.

2. 요청 본문으로 **없음** 라디오 버튼을 선택합니다.

3. AI-Tenant-Scope 또는 AI-Resource-Group 헤더를 통해 요청 범위를 지정합니다.

- AI-테넌트-범위 : true입니다. 작업은 기본 테넌트 수준에서 수행됩니다.
- AI-리소스-그룹 : <리소스-그룹-이름>. 작업은 리소스 그룹 수준에서 수행됩니다.

4. 요청을 보냅니다.

The screenshot shows a Postman interface with a GET request to `[[apilurl]]/v2/admin/secrets`. The Body tab is selected, showing the response: "This request does not have a body". Below the request, the response tab shows a JSON object with one resource:

```

1  {
2   "count": 1,
3   "resources": [
4     {
5       "createdAt": "2021-11-10 12:05:55+00:00",
6       "name": "my-generic-secret"
7     }
8   ]
9 }

```

The response status is 200 OK, Time: 312 ms, Size: 298 B, and there is a "Save Response" button.

컬 사용하기

엔드포인트 `/v2/admin/secrets`에 GET 요청을 제출하고 리소스 그룹 범위를 포함합니다.

- AI-테넌트-범위 : true입니다. 작업은 기본 테넌트 수준에서 수행됩니다.
- AI-리소스-그룹 : <리소스-그룹-이름>. 작업은 리소스 그룹 수준에서 수행됩니다.

```
컬 --location --request GET "$AI_API_URL/v2/admin/secrets" \
--header "승인: 전달자 $TOKEN" \
--header 'AI-리소스-그룹: 기본값'
```

응답에는 일반 비밀 목록, 이름, 생성 타임스탬프가 포함됩니다. 민감하지 않은 정보는 응답에서 공개됩니다.

6.5.3 일반 비밀 업데이트

일반 비밀을 업데이트하려면 아래와 같이 PATCH 엔드포인트를 사용하세요. PATCH 작업은 보안 비밀을 제공된 데이터로 바꿉니다. 이는 비밀 자격 증명을 교체하는 데 사용될 수 있습니다.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/admin/secrets/{{secretName}}에 PATCH 요청을 보냅니다.
2. 요청 본문으로 원시 라디오버튼을 선택하고 다음 코드를 입력합니다.

소스 코드

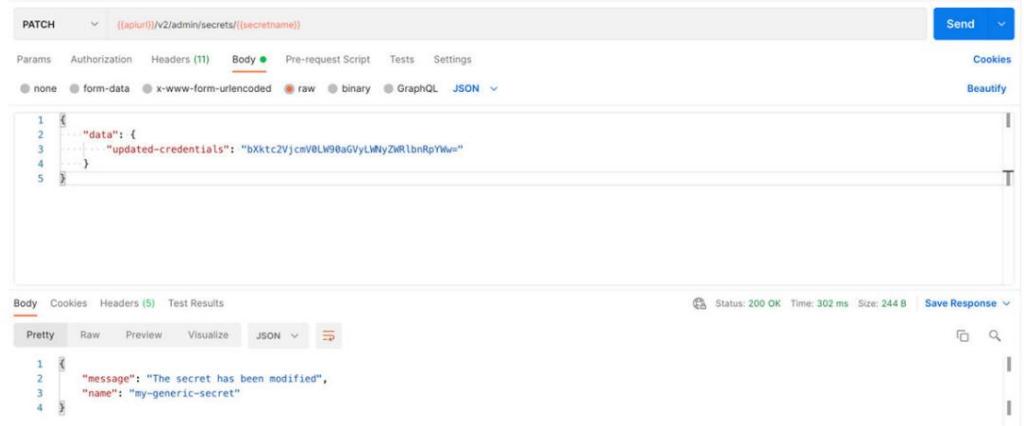
```
{
  "데이터": {
    "트된 자격 증명": "bXktc2VjcmV0LW90aGVyLWNyZWRlbnRpYWh="
  }
} 다음을 통해 범위를 지정합니다.
```

3. AI-Tenant-Scope 헤더를 통해 요청 범위를 지정하고 또는

AI 리소스 그룹:

- AI-테넌트-범위 : true입니다. 작업은 기본 테넌트 수준에서 수행됩니다.
- AI-리소스-그룹 : <리소스-그룹-이름>. 작업은 리소스 그룹 수준에서 수행됩니다.

4. 요청을 보냅니다.



컬 사용하기

엔드포인트 /v2/admin/secrets/"\$SECRET_NAME"AI-Tenant-Scope 또는 AI-Resource-Group 헤더에 PATCH 요청을 제출합니다.

```
컬 --location --request PATCH "$AI_API_URL/v2/admin/secrets/$SECRET_NAME" \
```

```
--header "권한 부여: Bearer $TOKEN" \ --header 'Content-Type: application/json' \ --header
'AI-Resource-Group: default' \ --data-raw '{ "data": { "some-credential":"
"bXktc2Vuc2l0aXZlLWRhdGE=" }

}'
```

6.5.4 일반 비밀 삭제

비밀 이름을 얻으려면 [모든 일반 비밀 나열 \[페이지 90\]](#)을 참조하십시오.

우편 배달부 사용

1. 엔드포인트{{apiurl}}/v2/admin/secrets/{{secretName}}에 DELETE 요청을 보냅니다.
2. 요청 본문으로 없음 라디오 버튼을 선택합니다.
3. AI-Tenant-Scope 또는 AI-Resource-Group 헤더를 통해 요청 범위를 지정합니다.
 - AI-테넌트-범위 : true입니다. 작업은 기본 테넌트 수준에서 수행됩니다.
 - AI-리소스-그룹 : <리소스-그룹-이름>. 작업은 리소스 그룹 수준에서 수행됩니다.
4. 요청을 보냅니다.

컬 사용하기

엔드포인트 /v2/admin/secrets/"\$SECRET_NAME"에 DELETE 요청을 제출합니다. 여기서는 리소스 그룹 범위를 사용합니다. 기본 테넌트 범위 수준의 경우 마지막 헤더를 AI-Tenant-Scope: true로 바꿉니다.

```
컬 --location --request DELETE "$AI_API_URL/v2/admin/secrets/
$SECRET_NAME$AI_API_URL/v2/admin/secrets/$SECRET_NAME" \ --header "권한 부여: Bearer
$TOKEN" \ --header 'AI-리소스-그룹: 기본값'
```

6.5.5 실행 시 일반 비밀 사용 또는 배포

리소스 그룹 수준의 일반 비밀은 실행 또는 배포 시 컨테이너에 연결될 수 있습니다. 볼륨으로 마운트하거나 환경 변수로 연결할 수 있습니다. 다음 예에서는 템플릿에서 일반 암호를 선언하여 컨테이너에서 일반 암호를 사용하는 방법을 보여줍니다. 이 방법으로는 일반 비밀만 컨테이너에 연결할 수 있습니다. 시스템 암호는 템플릿에서 사용할 수 없습니다.

일반 비밀을 환경 변수로 사용

아래와 같이 envFrom.secretRef 또는 env.valueFrom.secretKeyRef를 사용하여 일반 비밀을 컨테이너에 연결할 수 있습니다.

- envFrom.secretRef 사용:

```
사양: 컨테이너: - 이름: my-kserve-
container 이미지: centaur envFrom: - secretRef:
```

이름: MY_GENERIC_SECRET

비밀번호에 하이픈(-)과 같은 유효하지 않은 문자가 포함된 경우 이 방법을 사용하면 오류가 발생합니다. env.valueFrom.secretKeyRef를 사용하여 비밀을 유효한 변수 이름에 매핑해야 합니다.

- env.valueFrom.secretKeyRef 사용:

```
사양: 컨테이너: - 이름: kserve-
container 이미지: centaur

환경:
- 이름: MY_GENERIC_SECRET
값:
secretKeyRef: 이름: my-
generic-secret 키: some-credential
```

일반 암호를 볼륨 마운트로 사용

아래와 같이 일반 비밀을 볼륨으로 컨테이너에 탑재할 수도 있습니다.

```
사양: 컨테이너: - 이름: kserve-
container 이미지: centaur VolumeMounts:
- 이름: my-generic-secret mountPath: "/etc/
my-generic-secret" readOnly: true

볼륨:
- 이름: 내-일반-비밀
비밀:
secretName: 내-일반-비밀
```

추가 정보

아래와 같이 비밀 이름을 템플릿에 매개변수로 포함하고 AI API 구성을 통해 제공할 수 있습니다.

envFrom:

- 비밀참조:

이름: "{{inputs.parameters.secretName}}"

7 ML 작업

이 섹션에서는 SAP AI Core의 엔드 투 엔드 AI 라이프사이클을 안내합니다.

[데이터 연결 \[페이지 96\]](#)

SAP AI Core와 함께 클라우드 스토리지를 사용하여 데이터 세트 및 모델 파일과 같은 AI 자산을 저장하세요. SAP AI Core의 아티팩트를 사용하여 AI 자산을 참조합니다.

[모델 훈련 \[페이지 103\]](#)

AI 학습 모델을 훈련하기 위해 훈련 워크플로를 실행합니다.

[모델 사용 \[페이지 151\]](#)

AI 학습 모델을 배포하여 추론을 실행합니다.

7.1 데이터 연결

SAP AI Core와 함께 클라우드 스토리지를 사용하여 데이터 세트 및 모델 파일과 같은 AI 자산을 저장하세요. SAP AI Core의 아티팩트를 사용하여 AI 자산을 참조합니다.

[파일 관리 \[페이지 95\]](#)

아티팩트는 실행 또는 배포에 의해 생성되거나 소비되는 데이터 또는 파일을 나타냅니다. 이는 SAP AI Core 및 연결된 개체 저장소를 통해 관리됩니다.

[데이터 세트 API를 사용하여 파일 관리 \[페이지 101\]](#)

개체 저장소의 파일에 직접 액세스하는 것이 불가능하거나 바람직하지 않은 경우(예:

서비스 소비자가 자체 저장소의 소유자가 아닐 수 있는 서비스 시나리오) SAP AI Core Dataset을 사용하여 사전 등록된 객체 저장소에서 파일을 업로드, 다운로드 및 삭제할 수 있습니다.
API.

상위 주제: [ML 작업 \[페이지 96\]](#)

관련 정보

[모델 훈련 \[페이지 103\]](#)

[모델 사용 \[페이지 151\]](#)

7.1.1 파일 관리

아티팩트는 실행 또는 배포에 의해 생성되거나 소비되는 데이터 또는 파일을 나타냅니다. 이는 SAP AI Core 및 연결된 개체 저장소를 통해 관리됩니다.

개체 저장소 비밀을 사용하면 SAP AI Core가 자격 증명을 손상시키지 않고 클라우드 스토리지와 데이터에 액세스할 수 있습니다.

[파일 만들기 \[페이지 95\]](#)

[파일 목록 \[페이지 99\]](#)

상위 주제: [데이터 연결 \[페이지 96\]](#)

관련 정보

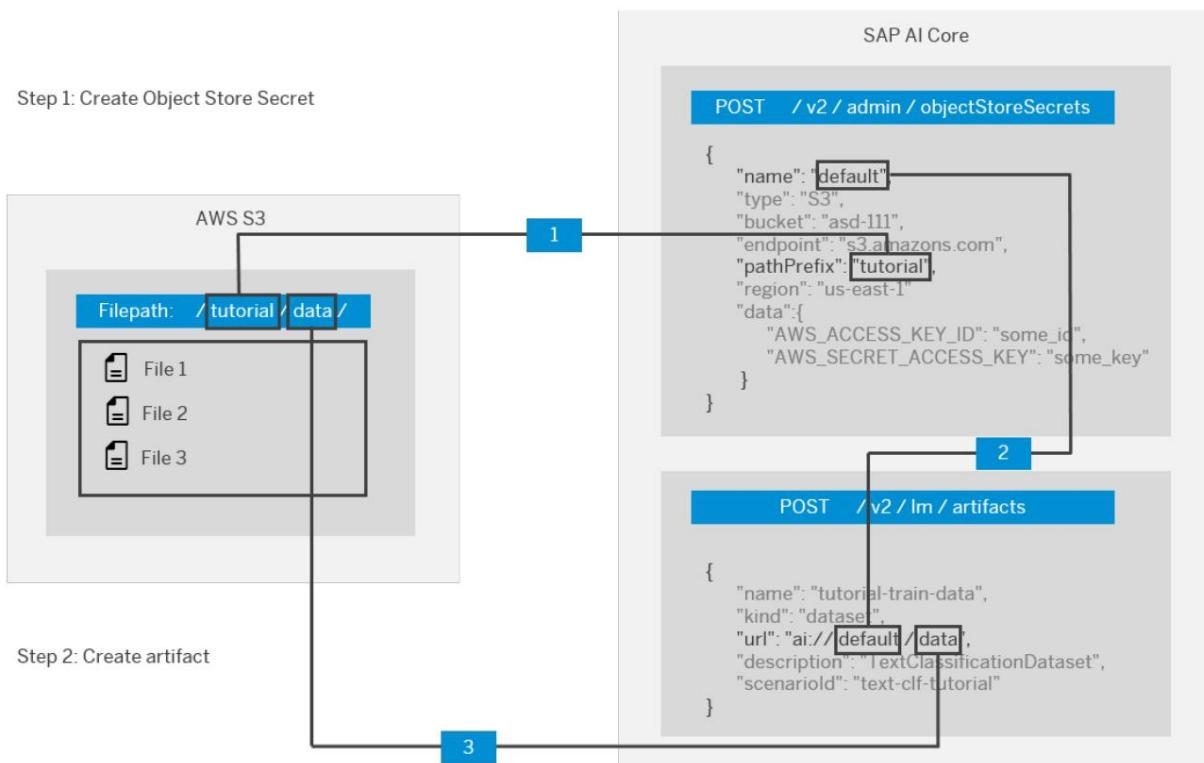
[데이터 세트 API를 사용하여 파일 관리 \[페이지 101\]](#)

7.1.1.1 파일 생성

개체 저장소 비밀을 사용하면 SAP AI Core가 자격 증명을 손상시키지 않고 클라우드 스토리지와 데이터에 액세스할 수 있습니다.

제한사항

objectStore 이름, 데이터 경로 및 시나리오 ID는 기존 값을 참조합니다. objectStore 이름 및 데이터 경로 값의 경우 아래 다이어그램에 설명된 명명 규칙에 따라 개체 스토리지를 등록할 때 사용한 값을 사용해야 합니다. 예제 출력 코드 블록에서 이러한 값은 ai://default/data로 표시됩니다.



우편 배달부 사용

1. URL `{apiurl}/v2/lm/artifacts`를 사용하여 새 POST 요청을 생성합니다. 2. 본문 탭을 전환하고 다음 JSON을 입력합니다.

```
{
  "name": "아티팩트 이름", "kind": "dataset", "url": "ai://<objectStore 이름>/<데이터 경로>",
  "description": "<아티팩트 설명>", "scenarioid": "<시나리오ID>"}
```

응답 본문에는 새 아티팩트의 ID가 포함되어 있습니다.

```
{
  "id": "3x4mpl3-651c-4f3e-8e1d-81a408041bc1", "message": "아티팩트가 승인됨",
  "url": "ai://default/data"
}
```

컬 사용하기

```
컬 --location --request POST "$API_URL/v2/lm/artifacts" \ --header "승인: Bearer $TOKEN" \
```

```
--header "Content-Type: application/json" \--header "AI-Resource-Group: <리소스 그룹>" \--
data-raw '{ "name": "아티팩트 이름", "kind": "dataset", "url": "ai://<objectStore 이름>/<데이터 경로>",
"description": "<아티팩트 설명>", "scenarioid": "<scenarioid>" }'
```

응답 본문에는 새 아티팩트의 ID가 포함되어 있습니다.

```
{ "id": "3x4mpl3-651c-4f3e-8e1d-81a408041bc1", "message": "아티팩트가 승인됨",
"url": "ai://default/data" }
```

상위 주제: [파일 관리 \[페이지 97\]](#)

관련 정보

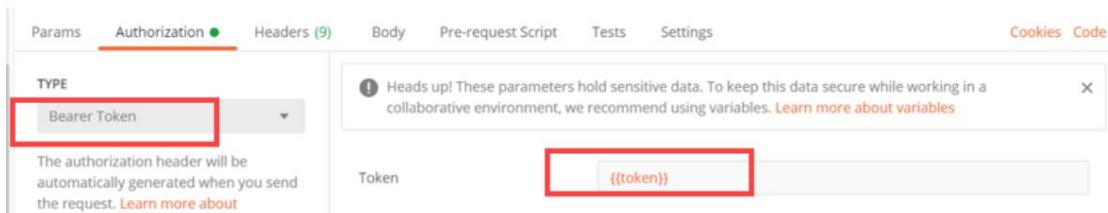
[파일 목록 \[페이지 99\]](#)

7.1.1.2 파일 나열

우편 배달부 사용

- 엔드포인트 {{apiurl}}/v2/lm/artifacts에 GET 요청을 보냅니다.
- Authorization 탭에서 유형을 Bearer Token으로 설정합니다.

- 토큰 값을 {{token}}으로 설정합니다.



4. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예에서는 기본값이 사용됨)

The screenshot shows the Postman interface with a GET request to {{apiurl}}/lm/artifacts. The 'Headers' tab is active, displaying a table with one row: 'AI-Resource-Group' key with value 'default'. This row is highlighted with a red box.

5. 요청을 보냅니다.

컬 사용하기

```
컬 --request GET "$AI_API_URL/v2/lm/artifacts" --header "승인: Bearer
$TOKEN" --header "ai-리소스 그룹: $RESOURCE_GROUP"
```

출력 코드

```
{
  "개수":3, "자원": [
    {
      "createdAt":"2021-02-09T08:08:12Z", "설명":"",
      "executionId":"d44edae36c187cf6",
      "id":"3088b75f-5448-4c19-8055-392668a043ec",
      "kind" :"model", "modifiedAt":"2021-02-09T08:08:12Z", "name":"pytf-model",
      "scenarioid":"ae0bd260-41ef-4162-81b0-861bd78a8516",
      "url" :"ai://default/d44edae36c187cf6/
      pytf-model"
    },
    {
      "createdAt":"2021-02-09T07:56:37Z", "설명":"",
      "executionId":"d44edae36c187cf6",
      "id":"38f7a46b-454d-4543-9457-b1eede5036f8",
      "kind" :"model", "modifiedAt":"2021-02-09T07:56:37Z", "name":"churn-
      pickle",
      "scenarioid":"ae0bd260-41ef-4162-81b0-861bd78a8516",
      "url" :"ai://default/d44edae36c187cf6/
      churn-pickle"
    },
    {
      "createdAt":"2021-02-09T07:56:37Z", "설명":"",
      "executionId":"d44edae36c187cf6",
      "id":"38f7a46b-454d-4543-9457-b1eede5036f8",
      "kind" :"model", "modifiedAt":"2021-02-09T07:56:37Z", "name":"churn-
      pickle",
      "scenarioid":"ae0bd260-41ef-4162-81b0-861bd78a8516",
      "url" :"ai://default/d44edae36c187cf6/
      churn-pickle"
    }
  ]
}
```

```
{
    "createdAt": "2021-02-07T16:07:16Z", "description": "Churn 및
    텍스트 분류자 데이터 세트", "id": "b45265f2-9bc3-441a-a0e1-fac1438acb79", "kind": "dataset
    ", "modifiedAt": "2021-02-07T16:07:16Z", "name": "pytf",
    "scenarioId": "84fe6957-1145-4183-
    b682-8f11ca56d060", "url": "ai://기본/pytf/"
}

]
```

상위 주제: [파일 관리 \[페이지 97\]](#)

관련 정보

[파일 만들기 \[페이지 95\]](#)

7.1.2 데이터세트 API를 사용하여 파일 관리

객체 저장소의 파일에 대한 직접 액세스가 가능하지 않거나 바람직하지 않은 경우(예: 서비스 소비자가 객체 저장소의 소유자가 아닐 수 있는 서비스로서의 콘텐츠 시나리오)에서 파일을 업로드, 다운로드 및 삭제할 수 있습니다. SAP AI Core Dataset API를 사용하여 사전 등록된 객체 저장소.

Dataset API 사양에 대한 자세한 내용은 [SAP AI Core API 문서를 참조하세요..](#)



전제조건

- 대상 리소스 그룹에 객체 저장소 암호가 정의되어 있어야 합니다.

제한사항

Dataset API에서는 S3 객체 저장소만 사용할 수 있습니다.

제한사항

Dataset API를 사용하면 csv 형식의 파일만 업로드할 수 있습니다.

[파일 만들기 \[페이지 101\]](#)

[파일 다운로드 \[페이지 103\]](#)

[파일 삭제 \[페이지 101\]](#)

상위 주제: [데이터 연결 \[페이지 96\]](#)

관련 정보

[파일 관리 \[페이지 95\]](#)

7.1.2.1 파일 생성

우편 배달부 사용

1. 앤드포인트 {{apiurl}}/v2/lm/dataset/files/{secret name}/{full에 PUT 요청을 보냅니다.
파일 경로}
파일 경로}를 설정합니다.
2. 헤더에서 Content-Type을 <text/csv>로 설정합니다.
3. 서비스 공급자 토큰을 사용하는 경우 리소스 그룹을 지정합니다.
서비스 소비자 토큰을 사용하는 경우 리소스 그룹 정보는 토큰에 포함되므로 지정할 필요가 없습니다.
4. 요청 본문에서 파일을 바이너리 데이터로 제출합니다.

결과

파일은 각각 저장소 버밀에 지정된 접두사와 업로드 요청에 지정된 전체 파일 경로를 사용하여 S3 스토리지 버킷에 업로드됩니다.

컬 사용하기

```
컬 --location --request PUT "$AI_API_URL/v2/lm/dataset/files/$SECRET_NAME/
$FILE_PATH" \" --header
"권한 부여: 전달자 $TOKEN" \ --header "콘텐츠 유형: text/csv" \ --header "ai-
resource-group: $RESOURCE_GROUP" \ --data @$FILE_LOCATION
```

- 서비스 공급자 토큰을 사용하는 경우 리소스 그룹을 지정합니다.
서비스 소비자 토큰을 사용하는 경우 리소스 그룹 정보는 토큰에 포함되므로 지정할 필요가 없습니다.
- 요청 본문에 파일을 바이너리 데이터로 제출합니다.

결과

파일은 각각 저장소 버밀에 지정된 접두사와 업로드 요청에 지정된 전체 파일 경로를 사용하여 S3 스토리지 버킷에 업로드됩니다.

7.1.2.2 파일 다운로드

우편 배달부 사용

엔드포인트 {{apiurl}}/v2/lm/dataset/files/{비밀 이름}/{전체 파일에 GET 요청 보내기
길}

The screenshot shows a POSTMAN interface with the following details:

- Method:** GET
- URL:** {{apiurl}}/v2/lm/dataset/files/default/test.csv
- Headers:** (9) (highlighted in green)
- Body:** (highlighted in red)
- Params:** binary (dropdown menu)
- Buttons:** Select File
- Response Headers:** Cookies (dropdown menu)
- Status:** 200 OK 537 ms 321 B Sa

컬 사용하기

```
컬 --location --request GET "$AI_API_URL/v2/lm/dataset/files/$SECRET_NAME/
$FILE_PATH"\`--header
"권한 부여: 전달자 $TOKEN"\`--header "ai-resource-group: $RESOURCE_GROUP"
\`--data @$FILE_LOCATION
```

7.1.2.3 파일 삭제

우편 배달부 사용

엔드포인트 {{apiurl}}/v2/lm/dataset/files/{secret name}/{full에 DELETE 요청을 보냅니다.
파일 경로}

Key	Value
AI-Resource-Group	{{resource-group}}
Authorization	{{token}}
Content-Type	text/csv
Key	Value

컬 사용하기

```
컬 --location --request 삭제 "$AI_API_URL/v2/lm/dataset/files/$SECRET_NAME/
$FILE_PATH" \\ --header
"권한 부여: 전달자 $TOKEN" \
--header "ai-리소스 그룹: $RESOURCE_GROUP"
```

7.2 모델 훈련

AI 학습 모델을 훈련하기 위해 훈련 워크플로를 실행합니다.

SAP AI Core는 리소스 그룹에서 지정한 대로 데이터를 사용합니다. 모델은 워크플로에 지정되며 원격 서버에서 학습됩니다. 결과 아티팩트는 객체 저장소에 저장됩니다.

실행 서비스에는 다음이 포함됩니다.

- 최첨단 워크플로우 엔진을 통한 일괄 작업/파이프라인.
- 테넌트 수준에서 저장 및 관리되는 파이프라인에 대한 매개변수화된 템플릿.
- 특정 매개변수와 리소스 경로가 포함된 파이프라인 인스턴스. • 파이프라인 단계 실행을 위해 자체 Docker 레지스트리 및 Docker 이미지 가져오기 지원.

- 파이프라인 단계로 모델링된 데이터 추출 및 검증.
- 데이터 저장소로 사용되는 전용 객체 저장소 버킷. 리소스 그룹별로 기본 버킷이 제공됩니다.
- 자체 하이퍼스케일러 지원 개체 저장소 버킷을 가져옵니다.

SAP AI Core의 실행 엔진은 [Argo Workflows](#)를 활용합니다. 오픈 소스 프로젝트, 직접 비순환 그래프 또는 단계로 모델링된 컨테이너 기반 워크플로 및 파이프 라인을 지원합니다. Argo 워크플로는 데이터를 수집하고, 전처리 및 후처리를 수행하고, 모델을 교육하고, 일괄 추론 파이프라인을 실행하는 데 사용됩니다. SAP AI Core는 [DAG\(Directed Acyclic Graph\) 구조](#) 형태의 단계 별령 처리도 활용합니다. 워크플로 템플릿에서 병렬 노드 사용이 비용에 어떤 영향을 미칠 수 있는지에 대한 자세한 내용은 [SAP AI Core 측정 및 가격 책정 \[페이지 39\]](#)을 참조하십시오.

참고

Argo Workflow는 시간이 중요한 작업에 최적화되어 있지 않습니다. 각 단계는 클러스터의 노드에 예약되고 클러스터가 초기화되어야 합니다. 소요되는 시간은 워크플로 컨트롤러의 로드와

클러스터의 노드 가용성. 따라서 시간이 중요한 작업에는 다단계 Argo Workflows를 사용하지 않는 것이 좋습니다.

다양한 매개변수(워크플로우와 호환되는 매개변수만 해당)를 사용하여 동일한 모델을 여러 번 훈련하고 SAP AI Launchpad에서 평가할 수 있습니다.

[리소스 계획 선택 \[페이지 104\]](#)

수요에 따라 다양한 작업에 다양한 인프라 리소스를 사용하도록 SAP AI Core를 구성할 수 있습니다. SAP AI Core는 이러한 목적으로 "리소스 계획"이라는 사전 구성된 여러 인프라 번들을 제공합니다.

[워크플로 템플릿 \[페이지 106\]](#)

여기서는 워크플로 요구 사항에 맞게 조정할 수 있는 최소 워크플로 예제 템플릿을 찾을 수 있습니다.

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

해시 형태의 아티팩트 서명을 추가하여 실행의 출력 아티팩트에 추가할 수 있습니다.

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

효율성을 향상하고 리소스 소비 관리에 도움이 되는 SAP AI Core 런타임의 기능을 알아보세요.

[실행 로그 검색 \[페이지 142\]](#)

배포 및 실행 로그에서 액세스됩니다.

[교육 일정 \[페이지 143\]](#)

상위 주제: [ML 작업 \[페이지 96\]](#)

관련 정보

[데이터 연결 \[페이지 96\]](#)

[모델 사용 \[페이지 151\]](#)

7.2.1 리소스 계획 선택

수요에 따라 다양한 작업에 다양한 인프라 리소스를 사용하도록 SAP AI Core를 구성할 수 있습니다.

SAP AI Core는 이러한 목적으로 "리소스 계획"이라는 사전 구성된 여러 인프라 번들을 제공합니다.

문맥

리소스 계획은 워크플로 및 제공 템플릿에서 리소스를 선택하는 데 사용됩니다. 워크플로의 다양한 단계 서로 다른 리소스 계획을 가질 수 있습니다.

일반적으로 워크로드에 GPU 가속이 필요한 경우 GPU 지원 리소스 계획 중 하나를 사용해야 합니다.

그렇지 않은 경우에는 워크로드의 예상 CPU 및 메모리 요구 사항을 기반으로 리소스 계획을 선택하세요.

SAP AI Core 내에서 리소스 계획은 Pod 수준의 ai.sap.com/resourcePlan 레이블을 통해 선택됩니다. 그것 선택한 리소스 계획을 매핑하고 다음 리소스 계획 ID 중 하나일 수 있는 문자열 값을 사용합니다.

AWS에 대한 리소스 계획 사양

리소스 계획 ID	GPU	CPU 코어	메모리 GB	재할당 코드
				출처
열차-L	1 V100	7	55	ai.sap.com/ 자원계획: 기차.l
추론-S	1 T4	삼	10	ai.sap.com/ 자원계획: 추론하다
추론-M	1 T4	7	26	ai.sap.com/ 자원계획: 추론.m
추론-L	1 T4	15	58	ai.sap.com/ 자원계획: 추론하다.l
기동기	-	1	삼	ai.sap.com/ 자원계획: 기동기
기초적인	-	삼	11	ai.sap.com/ 자원계획: 기초적인
기본-8x	-	31	116	ai.sap.com/ 자원계획: 기본.8x

제한사항

프리 티어 서비스 계획의 경우 스타터 리소스 계획만 사용할 수 있습니다. 다른 계획을 지정하면 오류가 발생합니다. Standard 서비스 계획의 경우 모든 리소스 계획을 사용할 수 있습니다. 자세한 내용은 [프리 티어 \[페이지 37\]](#) 및 [서비스 계획 \[페이지 35\]](#)를 참조하십시오.

참고

이러한 모든 노드에 대한 기본 디스크 저장소 크기에는 제한이 있습니다. 노드에 로드되는 데이터 세트는 디스크 공간을 사용합니다. 대규모 데이터 세트(30GB 이상)나 대규모 모델이 있는 경우 디스크 크기를 늘려야 할 수도 있습니다. 그렇게 하려면 Argo Workflows의 영구 볼륨 클레임을 사용하여 필요한 디스크 크기를 지정하십시오([볼륨](#) 참조).).

작업 개요: [모델 교육 \[페이지 102\]](#)

관련 정보

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

[교육 일정 \[페이지 143\]](#)

서비스 이용 보고

서비스 사용량 소비는 글로벌 계정의 [개요](#) 페이지와 하위 계정의 [개요](#) 및 [사용량 분석](#) 페이지에 있는 SAP BTP 조종석에 보고됩니다. 사용량 보고서에는 청구 가능한 측정값과 청구할 수 없는 측정값의 사용량이 나열됩니다. 최종 월별 청구서는 청구 가능한 측정값만을 기준으로 합니다. 청구 불가능한 측정값은 보고 목적으로만 표시됩니다.

7.2.2 워크플로 템플릿

여기서는 워크플로 요구 사항에 맞게 조정할 수 있는 최소 워크플로 예제 템플릿을 찾을 수 있습니다.

워크플로 템플릿을 사용하면 기본 테넌트 수준에서 학습 파이프라인을 관리할 수 있습니다. 템플릿은 git 저장소에 저장되어 필요에 따라 버전을 지정할 수 있습니다. SAP AI Core의 워크플로는 Argo Workflows 오픈 소스 프로젝트를 사용하여 실행됩니다. Argo Workflows는 Kubernetes에서 병렬 작업을 조정하기 위한 오픈 소스 컨테이너 기반 워크플로 엔진입니다. Argo Workflows는 Kubernetes CRD(Custom Resource Definition)로 구현됩니다.

워크플로 템플릿은 실행 파일로 매핑됩니다. 매핑에는 템플릿의 메타데이터 섹션에 특정 속성이 필요합니다. AI API는 템플릿의 주석과 레이블을 사용하여 시나리오와 실행 파일을 찾습니다.

워크플로 템플릿은 Argo Workflows 워크플로 엔진을 기반으로 하며 WorkflowTemplates로 정의됩니다. 이는 클러스터의 워크플로에 대한 정의입니다. WorkflowTemplate 및 샘플 워크플로에 대한 자세한 내용은 Argo 설명서를 참조하세요.

SAP AI Core에서 Argo Workflows는 다음 용도로 사용됩니다.

- 모델 훈련
- 데이터 수집 • 데이터 전처리 및 후처리
- 일괄 추론 파이프라인 실행

워크플로우는 배치 모드로 실행됩니다.

모델 학습 코드의 경우 SAP AI Core는 언어에 구애받지 않지만 관련 프로그래밍 언어는 워크플로 매개변수에 지정되어야 합니다. 가져온 패키지는 동일한 디렉토리에 저장된 요구사항.txt라는 별도의 파일에 지정되어야 합니다.

SAP AI Core에서 워크플로우 템플릿은 실행 파일로 매핑됩니다. 매핑에는 템플릿의 메타데이터 섹션에 특정 속성이 필요합니다. 워크플로는 여러 출력 아티팩트를 생성할 수 있지만 globalName이 있는 출력 아티팩트만 워크플로의 최종 출력 아티팩트로 간주됩니다.

시작하려면 아래의 일반 워크플로 템플릿을 복사하고 필요에 따라 고유한 값을 추가하세요. YAML 플러그인이 포함된 텍스트 편집기를 사용하여 템플릿을 만들 수 있습니다. 워크플로는 다음 매개변수를 지원합니다.

워크플로 템플릿 매개변수

유형	매개변수	설명
이름 (필수)	-	실행 가능 ID입니다. 실행 파일 ID 모든 실행 파일 중에서 고유해야 합니다. SAP AI Core 메인 내에서 사용 가능 거주자.
리벨(필수)	ai.sap.com/resourcePlan	선택한 항목을 지정해야 합니다. ResourcePlan. 값은 선택한 리소스의 문자열 값 계획. 자세한 내용은 선택을 참조하세요. 리소스 계획 [페이지 102].

유형	매개변수	설명
주석	시나리오.ai.sap.com/ 설명 (선택 사항)	해당 시나리오에 대한 설명 이 실행 파일이 속합니다.
	시나리오.ai.sap.com/name (필수적인)	필요한 시나리오 이름 AI API가 시나리오를 발견하도록 합니다.
	실행 파일.ai.sap.com/ 설명 (선택 사항)	실행 파일에 대한 설명입니다.
	실행 파일.ai.sap.com/ 이름 (필수)	실행 파일의 이름입니다.
	Artifacts.ai.sap.com/ <argo_artifact_name>.suffix 엑스	파일이나 폴더에 추가되는 접미사 각체 저장소에 출력 아티팩트가 있습니다.
	Artifacts.ai.sap.com/ <argo_artifact_name>.kind	출력 아티팩트의 유형(예: 데이터 세트 또는 모델)입니다.
	Artifacts.ai.sap.com/ <argo_artifact_name>.name	출력 아티팩트가 AI API에 등록되는 이름입니다.
	Artifacts.ai.sap.com/ <argo_artifact_name>.description 옵션 (선택사항)	AR에 대한 추가 메타데이터를 추가할 수 있습니다. 이러한 주석을 사용하여 tifact를 만듭니다.
	Artifacts.ai.sap.com/ <argo_artifact_name>.label 들: {"ext.ai.sap.com/ 맞춤키1":"맞춤값1" , "ext.ai.sap.com/ 맞춤키2":"맞춤값2" } (선택사항)	
라벨	시나리오.ai.sap.com/id (필수적인)	워크플로가 연결되는 시나리오 ID 템플릿이 속합니다.
	ai.sap.com/version (필수)	이 실행 파일의 호환 버전입니다. 호환성을 사용할 수 있습니다. 다양한 실행 파일을 제공하는 버전 다양한 AI 서비스 조건에 대한 변형 여름.

참고

위의 아티팩트 관련 매개변수에서 <argo_artifact_name>은 출력의 globalName을 나타냅니다.
인공물.

일반 워크플로 템플릿

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터: 이
름: text-clf-train-tutorial 주석:
```

```
시나리오.ai.sap.com/description: "SAP 개발자 튜토리얼 시나리오" 시나리오.ai.sap.com/name: "text-clf-tutorial-scenario" 실행 파일.ai.sap.com/description: "텍스
트 분류 Scikit 교육
```

실행 가능

```
실행 가능 파일.ai.sap.com/이름: "text-clf-train-tutorial-exec" 유형.ai.sap.com/text-data.kind: "데이터 세트" 유형.ai.sap.com/text-model-
tutorial.kind: "모델"artifacts.ai.sap.com/text-model-tutorial.description: "아티팩트 설명"artifacts.ai.sap.com/
text-model-tutorial.labels: | {"ext.ai.sap.com/customkey1":"customvalue1", "ext.ai.sap.com/
```

```
customkey2":"customvalue2"} 라벨: 시나리오.ai.sap.com/id:
"text-clf-
```

```
"tutorial" 실행 파일.ai.sap.com/id: "text-clf-train-tutorial" ai.sap.com/버전: "1.0.0"
```

사양:

```
imagePullSecrets: - 이름: <Docker 레지
스트리 비밀 이름> 진입점: text-clf-sk-training 인수: 매개변수: # 입력과 같은 문자열에 대한 자리 표시자
```

```
- 이름: DEPTH # 이 워크플로에 로컬인 식별자 설명: 매개변수에 대한 설명 기본값: test
```

```
템플릿: - 이름: text-clf-
sk-training 메타데이터: 라벨:
```

```
ai.sap.com/resourcePlan: 시작 입력: 아티팩트:
```

```
- 이름 : 텍스트 데이터
경로: /app/data/ 출력: 아티팩트: - 이름:
텍스트 모델-튜토리
얼 경로: /app/model
globalName: 텍스트-모델-튜토리얼 아카이브: 없음: {} 컨테이너:
```

```
image: "<DOCKER IMAGE URL GOES HERE>" imagePullPolicy: 항상 명령: ["bin/sh", "-c"] args: - >
```

```
set -e && echo "--훈련 시작--" && python /app/src/
train_scikit.py && ls -LR /app/model && echo "--훈련 종료--"
```

참고

템플릿의 모든 컨테이너에 대해 ["/bin/sh", "-c"] 필드 명령은 필수입니다. 인수의 내용은 수정될 수 있지만 비워둘 수는 없습니다. 컨테이너의 Dockerfile에 지정된 CMD 및 ENDPOINT는 무시됩니다.

Docker 이미지를 실행하려면 사용자 ID가 필요합니다. 이는 사용자가 지정하거나 지정하지 않을 수 있습니다.

사용자를 선택하고 지정하는 것은 선택 사항이며 유일한 제약 조건은 보안상의 이유로 사용자 ID가 `<root>` 사용자 인 `<0>i` 아니어야 한다는 것입니다.

사용자를 지정하지 않으면 `<65534>`, `<nobody>` 사용자 가 자동으로 할당됩니다.

사용자가 지정된 경우 Docker 이미지와 워크플로 템플릿에 포함되어야 합니다.

지정된 사용자이든 `<nobody>` 이든 관계없이 사용자에게는 애플리케이션이 실행되는 동안 파일에 액세스할 수 있는 권한이 필요하며, 이는 chown 및 chmod 명령을 사용하여 확인하고 변경할 수 있습니다.

팁

`docker run -it --user 65534:65543 <docker-image>` 를 로컬에서 실행하여 SAP AI Core에 제출하기 전에 컨테이너가 예상대로 작동하는지 확인할 수 있습니다.

참고

출력 아티팩트 섹션의 archive: none: {} 옵션은 아티팩트에 대한 자동 보관을 비활성화합니다. 아카이브가 활성화된 경우 출력 아티팩트는 객체 저장소에 업로드되기 전에 tar-gzip 파일에 아카이브됩니다. archive: none: {}을 통해 아카이브가 비활성화된 경우 아티팩트는 현재 형식으로 객체 저장소에 업로드됩니다. 출력 아티팩트가 디렉터리를 가리키는 경우 해당 디렉터리는 객체 저장소에 "있는 그대로" 업로드됩니다. 하지만 SAP HANA Cloud의 객체 저장소, 데이터 레이크는 이를 지원하지 않습니다. 이 경우 archive: none: {}을 제거하고 디렉터리를 단일 tar-gzip 파일로 보관한 후 객체 저장소에 업로드합니다.

참고

단일 WorkflowTemplate에 여러 컨테이너가 정의된 경우(예: 사이드카 또는 컨테이너 세트를 사용하는 경우) 컨테이너 중 하나의 이름은 'main'이어야 합니다. 엔드포인트 /logs에 대한 GET 요청을 사용하여 동일한 템플릿에 있는 다른 컨테이너의 로그를 가져오려면 컨테이너 이름에 '읽기 가능-' 접두사가 있어야 합니다. '읽기 가능한-' 접두사 없이도 실행이 계속 실행될 수 있지만 로그는 엔드포인트를 통해 액세스할 수 없습니다. 자세한 내용은 [Argo Sidecar](#)를 참조하세요. 및 [아르고 컨테이너 세트](#).

수동으로 애플리케이션 동기화

애플리케이션은 최대 3분 간격으로 자동으로 GitHub 저장소와 동기화됩니다. 아래 엔드포인트를 사용하여 수동으로 동기화를 요청하세요.

`{{apiurl}}/admin/applications/{{appName}}/refresh`

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)[구성 만들기 \[페이지 121\]](#)[목록 구성 \[페이지 122\]](#)[아티팩트 서명 사용 \[페이지 123\]](#)[교육 시작 \[페이지 131\]](#)[교육 인스턴스 중지 \[페이지 132\]](#)[교육 인스턴스 삭제 \[페이지 138\]](#)[효율성 기능 \[페이지 139\]](#)[실행 로그 검색 \[페이지 142\]](#)[교육 일정 \[페이지 143\]](#)

여러 교육 인스턴스 중지 또는 삭제

이 기능은 기본적으로 `false`로 설정되어 있습니다. 대량 PATCH 작업을 활성화하려면 템플릿에 관련 값이 `true`로 설정된 다음 코드 조각이 포함되어야 합니다.

```
메타:
  "bulkUpdates": { "executions": false,
    "deployments": false
  }
```

관련 정보

아르고 워크플로우

[Docker 레지스트리 암호 등록 \[페이지 84\]](#)[여러 교육 인스턴스 중지 \[페이지 136\]](#)[여러 교육 인스턴스 삭제 \[페이지 139\]](#)

7.2.3 목록 시나리오

시나리오는 사용자 테넌트 내의 사용 사례에 대한 관련 실행 파일 그룹입니다. 시나리오에는 다양한 실행 파일 버전에 해당하는 여러 버전이 있을 수 있습니다.

Postman을 사용하여 시나리오 나열

1. GET 요청을 엔드포인트 `{{apiurl}}/v2/lm/scenarios`로 보냅니다.
2. **Authorization** 탭에서 유형을 `Bearer Token`으로 설정합니다.

3. 토큰 값을 {{token}}으로 설정합니다.

The screenshot shows the Postman interface with the 'Authorization' tab selected. Under the 'Type' dropdown, 'Bearer Token' is selected. In the 'Token' input field, the placeholder value '{{token}}' is highlighted with a red box. A tooltip message at the top right of the interface reads: 'Heads up! These parameters hold sensitive data. To keep this data secure while working in a collaborative environment, we recommend using variables. Learn more about variables'.

4. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예제에서는 기본값이 사용됨)

The screenshot shows the Postman interface with the 'Headers' tab selected. It displays 8 hidden headers. A red box highlights the 'Headers' section and the 'Send' button. Below is a detailed view of the headers:

KEY	VALUE	DESCRIPTION	Bulk Edit	Presets
<input checked="" type="checkbox"/> AI-Resource-Group	default			
Key	Value	Description		

5. 요청을 보냅니다.

참고

정의된 교육 실행 파일이 있는 시나리오만 생성할 수 있습니다. 새로운 시나리오 ID가 지정된 경우 교육 실행 파일의 워크플로 템플릿에서 교육과 함께 새로운 시나리오가 생성됩니다. 실행 가능.

현재로서는 배포 실행 파일만 포함된 새 시나리오를 생성할 수 없습니다. 가능한 해결 방법은 다음과 같습니다.

더미 훈련 실행 파일로 시나리오를 생성한 다음 제공에서 동일한 시나리오 ID를 사용합니다.

주형.

컬을 사용하여 시나리오 나열

```
컬 --request GET 리소스 그룹:     "" --header "승인: 무기명 $TOKEN" --header "ai-$RESOURCE_GROUP"
```

출력 코드

```
{
  "개수":2,
  "자원":[
    {
      "createdAt":"2021-02-03T18:38:32+00:00",
```

```
"설명":"이탈 및 텍스트 클래스 시나리오 설명", "id":"84fe6957-1145-4183-b682-8f11ca56d060", "레이블":[
```

```
],
"modifiedAt":"2021-02-04T11:14:02+00:00", "name":"churntextclassscenname"
```

```
},
{
  "createdAt":"2021-02-04T14:11:02+00:00", "description":"이탈 및 텍스트 클래스 시나리오 설명", "id":"ae0bd260-41ef-4162-81b0-861bd78a8516", "ラベル":[
```

```
],
"modifiedAt":"2021-02-09T07:35:03+00:00", "name":"churntextclassscenname"
```

```
}
```

참고

정의된 교육 실행 파일이 있는 시나리오만 생성할 수 있습니다. 교육 실행 파일에 대한 워크플로 템플릿에 새 시나리오 ID가 지정되면 교육 실행 파일과 함께 새 시나리오가 생성됩니다.

현재로서는 배포 실행 파일만 포함된 새 시나리오를 생성할 수 없습니다. 가능한 해결 방법은 더미 학습 실행 파일을 사용하여 시나리오를 만든 다음 제공 템플릿에서 동일한 시나리오 ID를 사용하는 것입니다.

Postman으로 시나리오 버전 받기

새 GET 요청을 생성하고 URL {{apiurl}}/v2/lm/scenarios/{{scenarioId}}/를 입력합니다.

버전

The screenshot shows a Postman request configuration and its response. The request method is GET, and the URL is {{apiurl}}/v2/lm/scenarios/{{scenarioId}}/versions. The 'Headers' tab is selected, showing '(8)' headers. The 'Body' tab is selected, showing 'This request does not have a body'. The 'Tests' tab shows a success status: 200 OK, 1188 ms, 718 B, and a 'Save Response' button. The 'Pretty' tab is selected in the response panel, which displays the following JSON:

```

1
2   "count": 3,
3   "resources": [
4     {
5       "createdAt": "2021-10-07T20:07:18+00:00",
6       "id": "0.0.1",
7       "modifiedAt": "2021-10-07T20:07:18+00:00",
8       "scenarioId": "text-clf-tutorial"
    }
  ]
}

```

컬을 사용하여 시나리오 버전 얻기

다음을 제출하세요:

```
컬 --location --request GET '$API_URL/v2/lm/scenarios/$SCENARIO_ID/versions' \
```

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

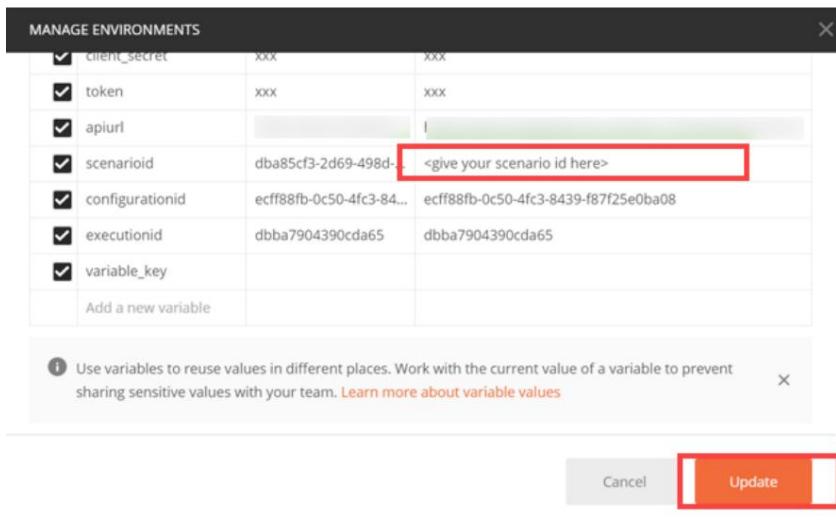
[교육 일정 \[페이지 143\]](#)

7.2.4 실행 파일 나열

실행 파일은 모델 교육이나 배포 생성과 같은 목적을 위해 인스턴스화되는 템플릿입니다. 시나리오의 모든 실행 파일을 나열하고 시나리오의 특정 실행 파일에 대한 세부 정보를 얻을 수 있습니다. 워크플로 템플릿은 교육 실행 파일에 매핑됩니다.

Postman을 사용하여 실행 파일 나열

1. 시나리오 ID를 시나리오 ID 환경 변수로 정의합니다.



2. 앤드포인트 {{apiurl}}/v2/lm/scenarios/{{scenarioid}}/executables에 GET 요청을 보냅니다.

3. 권한 부여 탭에서 유형을 Bearer Token으로 설정합니다.

4. 토큰 값을 {{token}}으로 설정합니다.

5. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> 기본값이 사용됨

6. 요청을 보냅니다.

```

1
2 "count": 3,
3 "resources": [
4 {
5     "createdAt": "2021-10-07T20:07:18+00:00",
6     "deployable": true,
7     "description": "Inference executable for text classification with Scikit-learn",
8     "id": "text-clf-infer-tutorial",
9     "inputArtifacts": [
]

```

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

컬을 사용하여 실행 파일 나열

```
curl --request GET "$AI_API_URL/v2/lm/scenarios" --header "권한 부여: Bearer $TOKEN" --header "ai-리소스 그룹: $RESOURCE_GROUP"
```

출력 코드

```
{
  "개수": 4, "자원": [
    {
      "createdAt": "2021-02-04T13:11:01+00:00", "deployable": true, "description": "실행 가능 한 설명을 제공하는 n 텍스트 클래스 이름",
      "id": "pytf-serving", "inputArtifacts": [
        {
          "이름": "model_uri"
        }
      ],
      "라벨": [
        {
          "name": "modelName", "type": "string",
          "default": "value", "description": "매개변수 설명"
        }
      ],
      "modifiedAt": "2021-02-04T13:11:01+00:00", "name": "churntextclassexecname", "매개변 수": [
        {
          "name": "modelName", "type": "string",
          "default": "value", "description": "매개변수 설명"
        }
      ],
      "scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
    },
    {
      "createdAt": "2021-02-09T07:35:02+00:00", "deployable": true, "description": "실행 가능 설명을 제공하는 n 텍스트 클래스 이름",
      "id": "pytf-serving-tracking", "inputArtifacts": [
        {
          "name": "textmodel", "kind": "모델",
          "description": "아티팩트 설명",
          "labels": [
            {
              "키": "ext.ai.sap.com/customkey1", "값": "customvalue1"
            },
            {
              "키": "ext.ai.sap.com/customkey2", "값": "customvalue2"
            }
          ]
        }
      ]
    }
  ]
}
```

```

        ],
    },
],
"라벨": [
],
"modifiedAt": "2021-02-09T07:35:02+00:00", "name": "churntextclassexecname", "매개변수": [
{
    "이름": "모델이름", "유형": "문자열"
}
],
"scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
},
{
    "createdAt": "2021-02-09T07:35:03+00:00", "deployable": false, "description": "churn 및 텍스트 클래스 실행 파일 설명", "id": "pytf-training-tracking", "inputArtifacts": [
{
    {
        "이름": "이탈 데이터"
    },
    {
        "이름": "텍스트클래스-데이터"
    }
],
"라벨": [
],
"modifiedAt": "2021-02-09T07:35:03+00:00", "name": "chunntextclassexecutablename", "outputArtifacts": [
{
    "name": "churn-pickle", "kind": "모델",
    "description": "아티팩트 설명",
    "labels": [
{
    "키": "ext.ai.sap.com/customkey1", "값": "customvalue1"
},
{
    "키": "ext.ai.sap.com/customkey2", "값": "customvalue2"
}
]
},
{
    "이름": "pytf-모델"
}
],
"매개변수": [
{
    "이름": "기차 시대", "유형": "문자열"
}
],
"scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
},
{
    "createdAt": "2021-02-04T14:11:02+00:00", "deployable": false, "description": "이탈 및 텍스트 클래스 실행 파일 설명",

```

```

"id":"test-training", "inputArtifacts":{

    "이름":"이탈 데이터"
},
{
    "이름":"텍스트클래스-데이터"
}
],
"라벨":[
],
"modifiedAt":"2021-02-04T14:11:02+00:00", "name":"chunntextclassexecutablename",
"outputArtifacts":{

    {
        "이름":"취젓기 피클"
    },
    {
        "이름":"pytf-모델"
    }
],
"매개변수":{

    "이름":"기차 시대", "유형":"문자열"
},
"scenarioid":"ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId":"0.0.1"
}
]
}
}

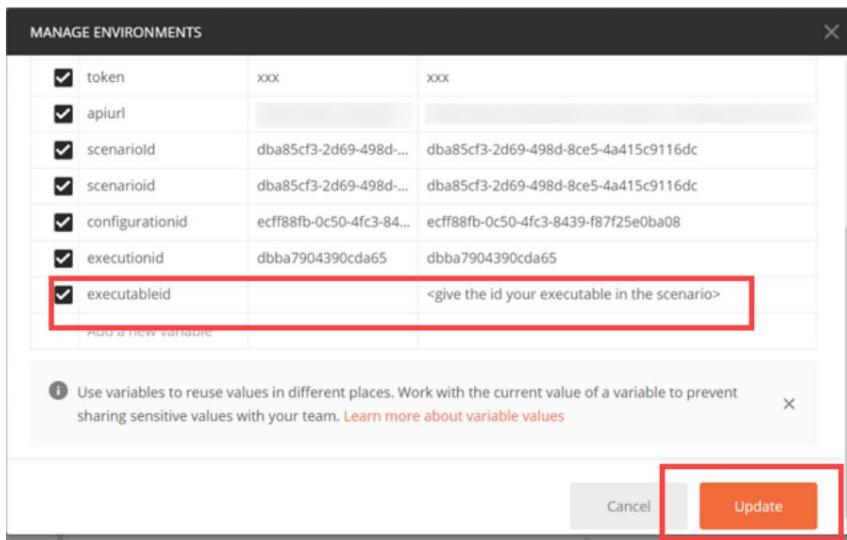
```

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

Postman으로 실행 가능한 세부정보 가져오기

- 환경 변수 runningid를 추가하고 해당 값으로 실행 파일의 ID를 입력합니다.



2. 앤드포인트 `{apiurl}/v2/lm/scenarios/{scenarioid}`에 GET 요청을 보냅니다.

실행 파일/`{executableid}`

3. 권한 부여 탭에서 유형을 `Bearer Token`으로 설정합니다.

4. 토큰 값을 `{token}`으로 설정합니다.

The screenshot shows the Postman Request tab with the Authorization tab selected. The 'Type' dropdown is set to 'Bearer Token'. The 'Token' field contains the placeholder '{token}'. A note on the right says: 'Heads up! These parameters hold sensitive data. To keep this data secure while working in a collaborative environment, we recommend using variables. [Learn more about variables](#)'.

5. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예제에서는 기본값이 사용됨)

6. 요청을 보냅니다.

참고

`<modifiedAt>` 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

컬을 사용하여 실행 가능한 세부정보 가져오기

```
curl --request GET "$AI_API_URL/v2/lm/scenarios" --header "권한 부여: Bearer $TOKEN" --header "ai-리소스 그룹: $RESOURCE_GROUP"
```

출력 코드

```
{
  "createdAt": "2021-02-04T14:11:02+00:00", "deployable": false,
  "description": "이탈 및 텍스트 클래스
  실행 파일 설명", "id": "test-training", "inputArtifacts": [
    {
      "이름": "이탈 데이터"
    },
    {
      "이름": "텍스트클래스-데이터"
    }
  ],
  "라벨": [
    ],
    "modifiedAt": "2021-02-04T14:11:02+00:00",
    "name": "chunntextclassexecutablename", "outputArtifacts": [
      {
        "이름": "휘젓기 파일"
      },
      {
        "이름": "pytf-모델"
      }
    ],
    "매개변수": [
      {
        "이름": "기차 시대", "유형": "문자열"
      }
    ],
    "scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
  }
}
```

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 종지 \[페이지 132\]](#)[교육 인스턴스 삭제 \[페이지 138\]](#)[효율성 기능 \[페이지 139\]](#)[실행 로그 검색 \[페이지 142\]](#)[교육 일정 \[페이지 143\]](#)

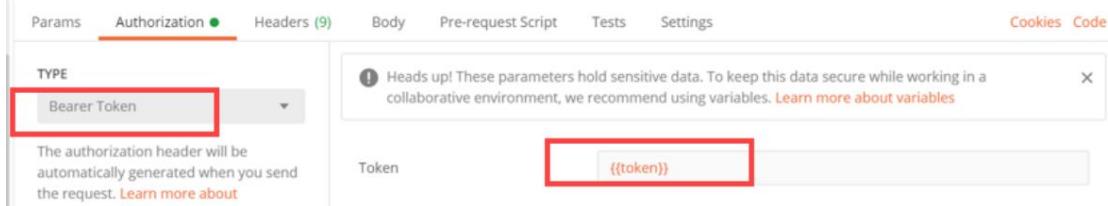
7.2.5 구성 생성

구성은 실행 또는 배포를 실행하는 데 사용되는 매개변수, 아티팩트 참조 및 실행 파일의 모음입니다.

우편 배달부 사용

1. 앤드포인트 {{apiurl}}/v2/lm/configurations에 POST 요청을 보냅니다.
2. **Authorization** 탭에서 유형을 **Bearer Token**으로 설정합니다.

3. 토큰 값을 {{token}}으로 설정합니다.



4. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예에서는 기본값이 사용됨)
컨텐츠 타입	애플리케이션/json

5. 본문 탭에서 원시 라디오 버튼을 선택하고 아래와 같이 요청 본문을 추가합니다.

```
{
  "name": "configuration-name", "executableId": "<실행 가능 ID>", "scenarioid": "<시나리오 ID>", "parameterBindings": [
    [
      {
        "key": "<매개변수 이름>", "value": "<값>"
      }
    ],
    "inputArtifactBindings": [
      {
        "key": "<아티팩트 이름>", "artifactId": "<아티팩트 ID>"
      }
    ]
}
```

The screenshot shows a Postman interface with the following details:

- Method:** POST
- URL:** {{apiurl}}/v2/lm/configurations (highlighted with a red box)
- Body (JSON):**

```

10 ...
11 ],
12 "inputArtifactBindings": [
13 ...
14   {
15     "key": "churn-data",
16     "artifactId": "896e9230-9219-47fa-89c7-c47db7d45d6a"
17 ...

```
- Response Status:** 201 Created (highlighted with a red box)
- Response Body (Pretty JSON):**

```

1 {
2   "id": "47b3eed9-f72f-4a18-b2a5-25b057a3e77f",
3   "message": "Configuration created"
4 }

```

6. 요청을 보냅니다.

컬 사용하기 |

```

컬 --request POST "$AI_API_URL/v2/lm/configurations" --header "승인:
베어러 $TOKEN" --header "ai-resource-group: $RESOURCE_GROUP" --header "Content-
유형: 애플리케이션/json" \
-d
'{ "name": "dummy-configuration", "executableId": 
    """$EXECUTABLE""", "scenarioid": """$SCENARIO""", 
    "parameterBindings": [
        {
            "키": "parameter_name_in_template", "값": "some_value"
        },
        ...
    ],
    "inputArtifactBindings": [
        {
            "key": "input_artifact_name_in_template", "artifactId": "a4d62a76-52aa-44cf-
                a789-743246d6d55b"
        }
    ]
}'

```

출력 코드

```
{

```

```

    "id":"f5bf305f-7c3f-4882-9f6b-8b95e3687b9b", "message":"구성이 생성되었습니다"
}

```

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 종지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

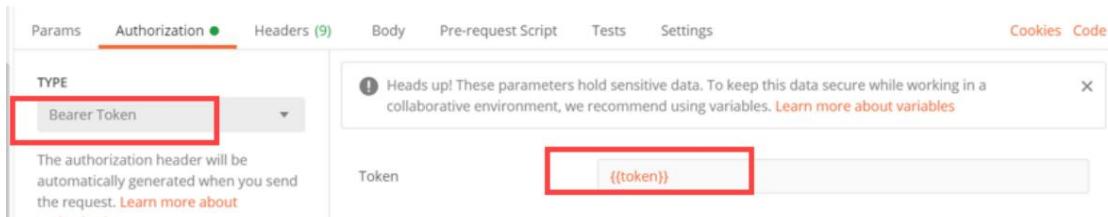
[교육 일정 \[페이지 143\]](#)

7.2.6 목록 구성

우편 배달부 사용

1. 엔드포인트 `{{apiurl}}/v2/lm/configurations`에 GET 요청을 보냅니다. 2. **Authorization** 탭에서 유형을 **Bearer Token**으로 설정합니다.

3. 토큰 값을 `{{token}}`으로 설정합니다.



4. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예제에서는 기본값이 사용됨)

5. 요청을 보냅니다.

The screenshot shows a Postman interface with a red box highlighting the URL field: `GET {{apiurl}}/v2/lm/configurations`. Below the URL, under the 'Body' tab, it says 'This request does not have a body'. The response section shows a status of `200 OK`, `595 ms`, and `19.03 KB`. The response body is displayed in JSON format:

```

1  {
2    "count": 45,
3    "resources": [
4      {
5        "createdAt": "2021-11-22T06:51:03Z",
6        "executableId": "text-clf-train-tutorial-metrics",

```

컬 사용하기

```
컬 --request GET "$AI_API_URL/v2/lm/configurations" --header "승인:  
베어러 $TOKEN" --header "ai-resource-group: $RESOURCE_GROUP"
```

출력 코드

```
{
  "개수": 2, "자원": [
    {
      "createdAt": "2021-02-04T11:50:45Z", "executableId": "pytf-training", "id": "1e6c2a5f-eabe-49a2-88e7-887145a2ef88", "inputArtifactBindings": [
        {
          "artifactId": "521f7f17-876e-4369-9162-09748b56d27a", "key": "churn-data"
        },
        {
          "artifactId": "521f7f17-876e-4369-9162-09748b56d27a", "key": "textclass-data"
        }
      ],
      "이름": "pytf-demo-config1", "parameterBindings": [
        {
          "키": "기차 시대", "값": "100"
        }
      ]
    }
  ]
}
```

```

        ],
        "scenarioid":"84fe6957-1145-4183-b682-8f11ca56d060"
    },
    {
        "createdAt":"2021-02-04T11:59:22Z", "executableId":"pytf-
        training", "id":"42147d0d-5e3a-424d-a3a1-545b842379c5",
        "inputArtifactBindings":[
            {
                "artifactId":"521f7f17-876e-4369-9162-09748b56d27a", "key":"churn-data"
            },
            {
                "artifactId":"521f7f17-876e-4369-9162-09748b56d27a", "key":"textclass-data"
            }
        ],
        "이름":"pytf-demo-config2", "parameterBindings": [
            {
                "키":"기차 시대", "값":"1"
            }
        ],
        "scenarioid":"84fe6957-1145-4183-b682-8f11ca56d060"
    }
}
}

```

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

[교육 일정 \[페이지 143\]](#)

7.2.7 아티팩트 서명 사용

해시 형태의 아티팩트 서명을 추가하여 실행의 출력 아티팩트에 추가할 수 있습니다.

출력 아티팩트에 대한 서명 제공

해시는 파일 콘텐츠를 사용하여 생성된 식별자입니다. 따라서 해시를 사용하여 두 파일을 비교하고 파일의 무결성을 확인할 수 있습니다. SAP AI Core는 서명을 생성하거나 확인하지 않지만 서명을 저장하고 사용할 수 있도록 합니다. 아티팩트 서명은 아티팩트와 함께 생성되고 저장됩니다.

출력 아티팩트에 대한 서명을 제공하려면 워크플로 템플릿에서 AICORE_ARTIFACT_SIGNATURES 출력 매개 변수를 지정합니다. 자세한 내용은 [Argo 워크플로 출력 매개변수를 참조하세요](#). 이 매개변수는 JSON 형식의 키-값 쌍으로 서명을 포함할 파일의 경로를 지정합니다.

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터: 이
  름: 서명-예제 주석:

    시나리오.ai.sap.com/description: "서명 기능의 예" 시나리오.ai.sap.com/name: "signature-example-scenario" 실행 파일.ai.sap.com/description: "서명 기능
    의 출력 아티팩트 예"

    실행 파일.ai.sap.com/name: "서명-예제-출력-아티팩트"
    라벨:
      시나리오.ai.sap.com/id: "서명-예제-시나리오" 실행 파일.ai.sap.com/id: "서명-예제-이웃-아트" ai.sap.com/
      version: "1.0.0"

    투기:
      진입점: 출력-아티팩트-서명-예 템플릿: - 이름: 출력-아티팩트-서명-예 출력: # "dummy-model"이라는 출력 아티팩트
      를 지정합니다. 이수: - 이
      름: dummy-model 경로: /tmp/model globalName: 더미 모델

    # 이름이 'AICORE_ARTIFACT_SIGNATURES'인 출력 매개변수를 지정합니다. # 이 매개변수는 다음을 포함하는 유효
    한 JSON 형식의 파일을 가리켜야 합니다.
    키-값 쌍이 있는 각체. 여기서 각 키는 # 출력 아티팩트의 globalName이고 값은 해당 서명 매개변수입니다. -
    name: AICORE_ARTIFACT_SIGNATURES # 매개변수 이름은 정확히 'AICORE_ARTIFACT_SIGNATURES'여야 합니다.

    값:
      path: /tmp/signatures.json # 경로 값을 설정할 수 있습니다.

  임의의 컨테이너:
    image: python:3.11-alpine imagePullPolicy: 항상 명령:
    ["python", "-c"] args: - | import hashlib, json model =
      "출력 아티팩트 역할을 하는 더미 문자열" model_signature =
        hashlib.sha256(model.encode("utf-8")).hexdigest() # 더미 모델의 해시를 생성합니다.

    서명 = {
```

```
# 키는 출력 아티팩트의 globalName이고 값은
그 서명이요
    "더미 모델": model_signature

} # open("/tmp/model", "w")을 사용하여 더미 모델을 디스크에
f: f.write(model)로 쓸니다.

# 모델 서명 사전을 지정된 경로에 씁니다.
open("/tmp/signatures.json", "w")을 f로 사용하는 AICORE_ARTIFACT_SIGNATURES 출력 매
개변수: json.dump(signatures, f) print("success")
```

예제 템플릿은 여러 출력 아티팩트와 함께 사용할 수 있습니다. 각 아티팩트에 대해 서명 개체에 다른 항목을 추가합니다.

템플릿은 AICORE_ARTIFACT_SIGNATURES 출력 매개변수에 제공된 이름-서명 키-값 쌍을 사용하여 출력 아티팩트 더미 모델을 생성합니다. 키는 출력 아티팩트의 전역 이름이고 값은 서명입니다. 예를 들어:

```
{
    "name-of-model1": "시그니처-모델1"
}
```

그런 다음 이러한 서명을 다른 실행에서 사용할 수 있습니다.

참고

서명은 교육 워크플로의 일부로 생성되어야 하며 소급하여 추가할 수 없습니다.

참고

GET {{apiurl}}/v2/lm/executions/{{executionid}} 엔드포인트를 통해 실행을 쿼리하면 출력 아티팩트의 서명이 내부적으로 유지되므로 표시되지 않습니다. 그러나 다른 실행이나 배포에서는 계속 사용할 수 있습니다.

다단계 워크플로우

다단계 워크플로 템플릿의 경우 출력 아티팩트 및 서명을 생성해야 하는 각 단계에 AICORE_ARTIFACT_SIGNATURES 출력 매개변수를 제공해야 합니다. 다음 예를 참조하세요.

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터:
```

이름: 서명-예제-2 주석:

여러 서명에 대한 출력 아티팩트 예입니다. 기능" 실행 파일.ai.sap.com/name: "signature-example-multi-output-artifacts" 레이블: 사나리오.ai.sap.com/id:
"signature-example-scenario" 실행 파일.ai.sap.com/id: "signature-example-multi-out-art" ai.sap.com/version: "1.0.0"

특기:

진입점: 단계 예

```

template: # 각각 출
    력 결과물이 있는 두 단계로 구성된 템플릿
    - 이름: steps-example steps: -- 출력-artifact-
        signatures-
            step1 -- 출력-artifact-signatures-step2 - 이름: 출력-artifact-signatures-step1 출력:
            # "dummy-model1"이라는 출력 아티팩트를 지정합니다.: - 이름: dummy-model1 경
            로: /tmp/model globalName: dummy-model1 # AICORE_ARTIFACT_SIGNATURES 매개변
            수는 현재 단계의 아
            티팩트 서명만 보유할 수 있습니다. # 예를 들어 아래 "dummy-model2"의 서명을 보유할 수 없습니다. 다른 워크플로우 단계
            매개변수에서 생성됩니다.
            - 이름: AICORE_ARTIFACT_SIGNATURES

값:
경로: /tmp/signatures.json 컨테이너:

image: python:3.11-alpine imagePullPolicy: Always command:
["python", "-c"] args: - > # ... 이전 예제와 동일하므로 건너뛰었습니다. -
이름: 출력-아티팩트- 서명-단계2 출력: # "dummy-model2"라는 출력 아티팩트
를 지정합니다.
이수: - 이
름: dummy-model2 경로: /tmp/model globalName: dummy-model2 # 각각에 AICORE_ARTIFACT_SIGNATURES 매개변수를 정의해야 합니다.

```

우리가 제공하고 싶은 워크플로 단계

```

# 출력 아티팩트 매개변수에 대한 서명: - 이름: AICORE_ARTIFACT_SIGNATURES
valueFrom:

```

```

경로: /tmp/signatures.json 컨테이너: 이미지: python:3.11-alpine
imagePullPolicy: 항상 명령:
["python", "-c"] args: - >
# ... 건너뛰었습니다. 이전 예시와 동일하기 때문입니다.

```

이 워크플로 템플릿은 dummy-model1 및 dummy-model2라는 두 가지 출력 아티팩트와 해당 이름-서명 키-값 쌍을 JSON 형식으로 생성합니다.

입력 아티팩트의 서명 확인

다른 실행에서 서명이 있는 입력 아티팩트를 사용하는 경우 SAP AI Core는 아티팩트의 서명을 AICORE_ARTIFACT_SIGNATURES라는 환경 변수로 제공합니다. 변수는 JSON 키-값 쌍입니다. 여기서 각 키는 입력 아티팩트 이름이고 값은 해당 서명입니다. 예를 들어:

```
{
    "모델 이름1": "모델 서명1", "모델 이름2": "모델 서명2"
}
```

```
}
```

서명을 추출하려면 워크플로 템플릿에 AICORE_ARTIFACT_SIGNATURES 변수를 포함합니다. 이 변수는 쌍의 키에서 값을 반환합니다. 예:

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터: 이
름: 서명-예제-3 주석:
```

```
시나리오.ai.sap.com/description: "서명 기능의 예" 시나리오.ai.sap.com/name: "signature-example-scenario" 실행 파일.ai.sap.com/description: "서명을 위한 입력 아티팩트 예"
```

특징"

실행 파일.ai.sap.com/name: "서명-예제-입력-아티팩트"

라벨:

```
시나리오.ai.sap.com/id: "서명-예제-시나리오" 실행 파일.ai.sap.com/id: "서명-예제-인-아트" ai.sap.com/version: "1.0.0" 사양:
진입점: 입력-아티팩트-서명-에 템플릿: - 이름: 입력-아티팩트-서명-에
```

```
입력: 아티팩트: -
이름: 더미 모델 경로: /
tmp/model 컨테이너: 이미지: python:3.11-
alpine imagePullPolicy: 항상 명령:
["python", "-c"] args:
- | import hashlib,json,os # 제공된 서명과 비교할 입력 아티팩트
의 해시를 생성합니다. open("/tmp/model", "rb") as f:
model_hash_actual = hashlib.file_digest(f,
```

```
"sha256").hexdigest()
# SAP AI Core는 다음을 통해 아티팩트 서명을 제공합니다.
AICORE_ARTIFACT_SIGNATURES 환경 변수 # JSON 형식의 문자열이 됩니다. 서명 =
json.loads(os.environ["AICORE_ARTIFACT_SIGNATURES"])
model_hash_expected = 서명["dummy-model"] # 이름
```

입력 아티팩트가 핵심입니다

주장 model_hash_expected == model_hash_actual, "서명이 이루어졌습니다.
일치하지 않는다면"

```
print("성공")
```

이 템플릿은 <dummy-model>이라는 이름의 입력 아티팩트를 가져와 해당 서명을 계산한 후 AICORE_ARTIFACT_SIGNATURES 환경 변수에서 SAP AI Core가 제공한 서명과 비교합니다. AICORE_ARTIFACT_SIGNATURES 환경 변수는 서명이 있는 입력 아티팩트가 하나 이상 있는 경우에만 설정되며, 그렇지 않으면 전혀 설정되지 않습니다. 단단계 워크플로 템플릿에서는 실제 입력 아티팩트가 정의된 위치에 관계없이 AICORE_ARTIFACT_SIGNATURES 환경 변수가 워크플로 템플릿의 각 단계에 제공됩니다. 각 단계에서 AICORE_ARTIFACT_SIGNATURES 변수는 모든 단계의 입력 아티팩트 서명을 보유합니다.

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

[교육 일정 \[페이지 143\]](#)

7.2.8 훈련 시작

Postman으로 훈련 실행

1. 엔드포인트 {{apiurl}}/v2/lm/executions에 POST 요청을 보냅니다. 구성 ID를 전달하세요.
요청 본문.

POST {{apiurl}}/v2/lm/executions

Params Authorization Headers (12) Body Pre-request Script Tests Se

none form-data x-www-form-urlencoded raw binary GraphQL **JSC**

```

1  {
2    "configurationId": "47b3eed9-f72f-4a18-b2a5-25b057a3e77f"
3  }
4

```

Body Cookies Headers (6) Test Results 202 Ac

Pretty Raw Preview Visualize **JSON** Copy

```

1  {
2    "id": "edc4121d1085c20a",
3    "message": "Execution scheduled",
4    "status": "UNKNOWN",
5    "targetStatus": "COMPLETED"

```

2. {{apiurl}}/v2/lm/executions/에 GET 요청을 제출하여 실행 상태를 확인합니다.
 {{실행 ID}}.

GET {{apiurl}}/v2/lm/executions/edc4121d1085c20a

Params Authorization Headers (11) Body Pre-request Script Tests Settings

Headers 8 hidden

KEY	VALUE
<input checked="" type="checkbox"/> Authorization	{{token}}
<input checked="" type="checkbox"/> ai-resource-group	{{resource-group}}
<input checked="" type="checkbox"/> Content-Type	application/json
Key	Value

참고

상태가 Dead 또는 Pending인 경우 실행에 오류가 있을 수 있습니다. 자세한 내용은 실행 로그를 확인하여 [실행 로그 검색 \[페이지 142\]](#)를 참조하십시오.

컬을 사용하여 훈련 실행

- 구성을 통해 교육 워크플로를 트리거합니다.

```
컬 --location --request POST '$AI_API_URL/v2/lm/executions' \
```

출력 코드

```
{
  "id": "dea6263e6283321b", "message": "실행 예약됨", "status": "UNKNOWN", "targetStatus": "COMPLETED"
}
```

- 실행 상태를 확인하세요.

```
컬 --request GET $AI_API_URL/v2/lm/executions/$EXECUTION \\
  --header "권한 부여: 전달자 $TOKEN" \ --header "ai-resource-group:
$RESOURCE_GROUP"
```

참고

상태가 Dead 또는 Pending인 경우 실행에 오류가 있을 수 있습니다. 자세한 내용은 실행 로그를 확인하여 [실행 로그 검색 \[페이지 142\]](#)를 참조하십시오.

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

[교육 일정 \[페이지 143\]](#)

7.2.9 훈련 인스턴스 중지

`$AI_API_URL/v2/lm/executions/`에 PATCH 요청을 제출하여 실행 중인 실행을 중지할 수 있습니다.

`$실행\.`

참고

중지는 상태가 실행 중이거나 보류 중인 경우에만 활성화됩니다.

[단일 교육 인스턴스 중지 \[페이지 131\]](#)

[여러 교육 인스턴스 중지 \[페이지 136\]](#)

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

[교육 일정 \[페이지 143\]](#)

7.2.9.1 단일 교육 인스턴스 중지

우편 배달부 사용

The screenshot shows the POSTMAN interface with the following details:

- Method:** PATCH
- URL:** {{apiurl}}/v2/lm/executions/{{executionid}}
- Body (JSON):**

```

1 {
2   "targetStatus": "STOPPED"
3 }

```

- Response Headers:** 202 Accepted, 367 ms, 251 B
- Response Body (Pretty JSON):**

```

1 {
2   "id": "e8c53facc2bfb87a",
3   "message": "Execution modification scheduled"
4 }

```

컬 사용하기

The terminal window displays the following curl command to stop an AI execution:

```
curl --request PATCH $AI_API_URL/v2/lm/executions/$EXECUTION \
--header "Authorization: Bearer $TOKEN" \
--header "ai-resource-group: $RESOURCE_GROUP" \
--header "Content-Type: application/json" \
-d '{ "targetStatus": "STOPPED" }'
```

Output:

```

{
  "id": "ee6769e4dc19c0fd", "message": "실행 수
정 예약됨"
}

```

상위 주제: [교육 인스턴스 중지 \[페이지 132\]](#)

관련 정보

[여러 교육 인스턴스 중지 \[페이지 136\]](#)

7.2.9.2 여러 훈련 인스턴스 중지

BulkUpdates는 AI API의 메타 기능 엔드포인트입니다. 대량 PATCH 작업을 활성화하거나 비활성화합니다. 자세한 내용은 [AI API 개요 \[페이지 24\]](#)를 참조하십시오.

이 기능은 기본적으로 false로 설정되어 있습니다. 대량 PATCH 작업을 활성화하려면 템플릿에 관련 값이 true로 설정된 다음 코드 조각이 포함되어야 합니다.

```
메타:
"bulkUpdates": { "executions": false,
                 "deployments": false
               }
```

대량 업데이트 정보:

- 요청당 최대 업데이트 수는 100개입니다. • 대량 업데이트에는 STOP 및 DELETE 요청이 혼합되어 포함될 수 있습니다.
- 실행 중이거나 보류 중인 실행이나 배포 만 중지할 수 있습니다.
- 중지되었거나 중단되었거나 알 수 없는 실행 또는 배포 만 삭제할 수 있습니다. • ID는 일괄 요청당 한 번만 나타날 수 있습니다. 동일한 ID에 대한 다중 수정의 경우 다중 요청 필요합니다.

우편 배달부 사용

엔드포인트에 대량 PATCH 요청 보내기: - /executions

요청 본문을 다음과 같이 업데이트합니다.

```
{
  "실행": [
    {
      "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨" },
      {
        "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"
      }
    ]
}
```

출력 코드

```
{
  "실행": [
```

```
{
  "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "실행 수정 예정" },
  {
    "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "실행 수정 예정"
  }
}
```

컬 사용하기

요청 본문을 다음과 같이 업데이트합니다.

```
컬 --파치 요청 - /실행 \
--header {"실행": [ {
  "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨" },
  {
    "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"
  }
}]}
```

출력 코드

```
{
  "실행": [ {
    "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "실행 수정 예정" },
    {
      "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "실행 수정 예정"
    }
  }]
}
```

상위 주제: [교육 인스턴스 중지](#) [페이지 132]

관련 정보

[단일 교육 인스턴스 중지](#) [페이지 131]

[AI API 개요](#) [페이지 24]

7.2.10 교육 인스턴스 삭제

훈련 인스턴스를 삭제하면 사용한 SAP AI Core 리소스가 해제됩니다.

제한사항

실행이 실행 중인 경우 먼저 중지해야 합니다. \$AI_API_URL/v2/lm/executions/\$EXECUTION\에 PATCH 요청을 제출하여 실행 중인 실행을 중지할 수 있습니다. 자세한 내용은 [교육 인스턴스 중지 \[페이지 132\]](#)를 참조하십시오.

[단일 교육 인스턴스 삭제 \[페이지 136\]](#)

[여러 교육 인스턴스 삭제 \[페이지 139\]](#)

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

[교육 일정 \[페이지 143\]](#)

7.2.10.1 단일 교육 인스턴스 삭제

컬 사용하기

```
컬 --request DELETE $AI_API_URL/v2/lm/executions/$EXECUTION \\
--header "승인: 전달자 $TOKEN" \
--header "ai-리소스 그룹: $RESOURCE_GROUP"
```

출력 코드

```
{
  "id": "ee6769e4dc19c0fd", "message": "삭제 예정", "targetStatus": "DELETED"
}
```

우편 배달부 사용

`{apiurl}/v2/lm/executions/{executionid}`에 DELETE 요청을 보냅니다. 이 요청의 헤더는 AI-Resource-Group: {YOUR-Resource-Group}입니다.

The screenshot shows the Postman interface with a DELETE request to `{apiurl}/v2/lm/executions/{executionid}`. The Headers tab is active, displaying the following configuration:

KEY	VALUE	DESC	Bulk Edit	Presets
AI-Resource-Group	default			
Key	Value	Description		

The Body tab is also visible, showing a JSON response with the following content:

```

1  {
2    "id": "e8c53facc2bfb87a",
3    "message": "Deletion scheduled",
4    "targetStatus": "DELETED"
5  }

```

상위 주제: [교육 인스턴스 삭제 \[페이지 138\]](#)

관련 정보

[여러 교육 인스턴스 삭제 \[페이지 139\]](#)

7.2.10.2 여러 교육 인스턴스 삭제

BulkUpdates는 AI API의 메타 기능 엔드포인트입니다. 대량 PATCH 작업을 활성화하거나 비활성화합니다. 자세한 내용은 [AI API 개요 \[페이지 24\]](#)를 참조하십시오.

이 기능은 기본적으로 false로 설정되어 있습니다. 대량 PATCH 작업을 활성화하려면 템플릿에 관련 값이 true로 설정된 다음 코드 조각이 포함되어야 합니다.

```
메타:
  "bulkUpdates": { "executions": false,
    "deployments": false
  }
```

대량 업데이트 정보:

- 요청당 최대 업데이트 수는 100개입니다. • 대량 업데이트에는 STOP 및 DELETE 요청이 혼합되어 포함될 수 있습니다.
- 실행 중이거나 보류 중인 실행이나 배포 만 중지할 수 있습니다.
- 중지되었거나 중단되었거나 알 수 없는 실행 또는 배포 만 삭제할 수 있습니다.
- ID는 일괄 요청당 한 번만 나타날 수 있습니다. 동일한 ID에 대한 다중 수정의 경우 다중 요청 필요합니다.

컬 사용하기

요청 본문을 다음과 같이 업데이트합니다.

```
컬 --패치 요청 - /실행 \
--header {"실행": [ {
  "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨", {
    "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"
  }
}]}
```

출력 코드

```
{
  "실행": [
    {
      "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "실행 수정 예정", {
        "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "실행 수정 예정"
      }
    }
  ]
}
```

우편 배달부 사용

대량 폐지 요청 보내기: - /executions

요청 본문을 다음과 같이 업데이트합니다.

```
{  
    "실행": [  
        {  
            "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨" }, {  
  
                "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"  
            }  
        ]  
    }  
}
```

출력 코드

상위 주제: [교육 인스턴스 삭제 \[페이지 138\]](#)

관련 정보

[단일 교육 인스턴스 삭제 \[페이지 136\]](#)

[AI API 개요 \[페이지 24\]](#)

7.2.11 효율성 특징

효율성을 향상하고 리소스 소비 관리에 도움이 되는 SAP AI Core 런타임의 기능을 알아보세요.

테넌트 월 노드 풀

테넌트 월 노드를 사용하면 테넌트가 특정 리소스 계획 유형에서 노드를 예약하여 미리 정의된 수의 노드가 클러스터에 존재하도록 할 수 있습니다. 그런 다음 모델 학습 및 제공 중에 예약 노드를 사용할 수 있습니다. 노드를 예약하면 워크로드 시작 시 대기 시간이 줄어들지만 사용 여부에 관계없이 노드 사용에 따라 인라인 비용이 발생합니다.

노드 예약 메커니즘

1. 테넌트는 예약할 노드 수를 지정합니다.
2. 실행 또는 배포는 예약 노드, 즉 동일한 리소스의 대체 예약 노드를 활용합니다.

계획 유형은 하이퍼스케일러에서 요청됩니다.

테넌트가 지정한 예약 노드 수는 지속적으로 사용할 수 있습니다.

최소 예약 노드 수는 0입니다.

노드 예약 정보

- 테넌트가 지정한 예약 노드 수만큼 지속적으로 사용할 수 있습니다.
- 최소 예약 노드 수는 0입니다.
- 최대 예약 노드 수는 10개입니다.
- 기본 예약 노드 수는 0입니다.

Postman을 사용한 예약 노드

1. [엔드포인트{{apiurl}}/v2/admin/resources/nodes](#)에 PATCH 요청을 보냅니다. 2. 요청 본문에 예약할 리소스 계획 유형과 노드 수량을 JSON 형식으로 제공합니다.

```
{
  "resourcePlans": [
    {
      "이름": "infer.l", "요청": 1
    },
    {
      "이름": "infer.m", "요청": 1
    },
    {
      "이름": "train.l", "요청": 1
    }
    ...
  ]
}
```

컬을 사용하는 예약 노드

[엔드포인트{{apiurl}}/v2/admin/resources/nodes](#)에 PATCH 요청을 제출합니다.

```
curl --request 패치 $AI_API_URL/v2/admin/resources/nodes \ --data-raw '{ "resourcePlans": [
  {
    "이름": "infer.l", "요청": 1
  },
  {
    "이름": "infer.m", "요청": 1
  },
  {
    "이름": "train.l", "요청": 1
  }
]}'
```

기억하세요

모든 예약 노드에는 모델 학습 및 서빙 중에 사용된 노드와 동일한 요금이 청구됩니다.

Postman을 사용하여 예약 노드 상태 확인

엔드포인트{{apiurl}}/v2/admin/resources/nodes에 GET 요청 보내기

컬을 사용하여 예약 노드 상태 확인

```
컬 --request GET $AI_API_URL/v2/resources/nodes
```

출력 코드

```
{
  "resourcePlans": {
    "infer.l": {
      "프로비저닝됨": 1,
      "요청됨": 1
    },
    "infer.m": {
      "프로비저닝됨": 1,
      "요청됨": 1
    },
    "train.l": {
      "프로비저닝됨": 1,
      "요청됨": 1
    }
  }
}
```

- 요청됨: 테넌트가 요청한 예약 노드 수입니다.
- 프로비저닝됨: 현재 클러스터에 존재하는 예약 노드 수입니다.

예약 노드 수량 업데이트

예약된 노드 수를 업데이트하려면 요청 필드에 수량을 업데이트하여 예약 절차를 반복하세요.

예약 노드 삭제

예약 노드를 삭제하려면 요청 필드의 수량을 0으로 설정하여 예약 절차를 반복하세요.

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)[실행 로그 검색 \[페이지 142\]](#)[교육 일정 \[페이지 143\]](#)[서비스 계획 \[페이지 35\]](#)

7.2.12 실행 로그 검색

배포 및 실행 로그에서 액세스됩니다.

GET 요청을 제출하여 특정 배포 또는 실행에 대한 로그를 검색할 수 있습니다. 로그를 검색하려면 다음 엔드포인트를 사용하세요.

- GET /v2/lm/deployments/{deploymentId}/logs
- GET /v2/lm/executions/{executionId}/logs

쿼리 매개변수는 다음과 같습니다.

- start : RFC 3339 호환 날짜/시간 형식의 문자열로 된 쿼리 시작 시간입니다. 기본값은 현재보다 1시간 전입니다. 예: 2021-05-19T00:00:14.347Z
- end : RFC 3339 호환 날짜/시간 형식의 문자열로 된 쿼리의 종료 시간입니다. 기본값은 현재 시간입니다. 예: 2021-05-19T00:00:14.347Z

- \$top : 반환된 항목의 최대 수입니다. 기본값은 1000입니다. 상한은 5000입니다. • \$order : 로그의 정렬 순서입니다. asc(오름차순의 경우 가장 빠른 순서가 목록 상단에 표시됨) 또는 desc(내림차순의 경우 가장 최근 순서가 목록 상단에 표시됨) 목록의 맨 위). 기본값은 asc입니다.

예를 들어:

- /v2/lm/deployments/{deploymentId}/logs?
start=2021-05-19T00:00:14.347Z&end=2021-05-19T01:00:14.347Z&\$top=100&\$order=asc - 첫 번째 반환

2021-05-19T00:00:14.347Z와 2021-05-19T01:00:14.347Z 사이의 배포 로그 100줄

- /v2/lm/deployments/{deploymentId}/logs - 이전 시간의 배포 로그를 반환합니다.
- /v2/lm/executions/{executionId}/logs - 이전 시간의 실행 로그를 반환합니다.

우편 배달부 사용

1. 엔드포인트 {{apiurl}}/v2/lm/executions/{{executionId}}/logs에 GET 요청을 보냅니다.

Headers (8)

Key	Value	Description	Bulk Edit	Presets
AI-Resource-Group	default			
Key	Value	Description		

```

1
2   "data": {
3     "result": [
4       {
5         "container": "init",
6         "msg": "time=\"2021-11-22T06:52:27.831Z\" level=info msg=\"Starting Workflow Executor\" executorType=
7           version=v3.2.2\\n",
8         "pod": "ed0f0c1b6294b935",
9         "stream": "stderr",
10        "timestamp": "2021-11-22T06:52:27.831955906+00:00"
11      }
12    ]
13  }
14
15 }
```

2. 권한 부여 탭에서 유형을 Bearer Token으로 설정합니다.
3. 토큰 값을 {{token}}으로 설정합니다.
4. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름>(예제에서 기본값은 사용됨)

5. 요청을 보냅니다.

컬 사용하기

```
컬 --request GET "$AI_API_URL/v2/lm/executions/$EXECUTION_ID/logs?
start=2021-05-19T00:00:14.347Z" --header "권한 부여: Bearer $TOKEN" --header "ai-resource-group: $RESOURCE_GROUP"
```

샘플 출력

예를 들어 API의 다음 JSON 출력을 참조하세요.

출력 코드

```
{
  "데이터": {
    "결과": [
      {
        "container": "저장소 초기화 장치",
        "msg": "[ 210531 08:20:51 초기화-진입점:13]
초기화 중, 인수: src_uri [gs://kserve-samples/models/tensorflow/flowers] dest_path[ /mnt/models]\\n",
        "time": "2021-05-19T00:00:14.347Z"
      }
    ]
  }
}
```

```

    "포드": "tfss-dep-i543026-predictor-default-v6nf5-
배포-8b58c8ddcf0x", "스트림": "stderr", "타
    임스탬프":
        "2021-05-31T08:20:51.334+00:00"
    },
    {
        "container": "storage-initializer", "msg": "[I 210531 08:20:51
Storage:45] 내용 복사 중
gs://kserve-samples/models/tensorflow/flowers를 로컬로\n", "pod": "tfss-dep-i543026-predictor-
default-v6nf5-
배포-8b58c8ddcf0x", "스트림": "stderr", "타
    임스탬프":
        "2021-05-31T08:20:51.335+00:00"
    },
    {
        "container": "storage-initializer", "msg": "[W 210531 08:20:51
_metadata:104] 3번 중 1번 시도에서 Compute Engine 메타데이터 서버를 사용할 수 없습니다. 이유: [Errno
111] 연결이 거부되었습니다.\n",
        "포드": "tfss-dep-i543026-predictor-default-v6nf5-
배포-8b58c8ddcf0x", "스트림": "stderr", "타
    임스탬프":
        "2021-05-31T08:20:51.338+00:00"
    },
    ...
]
}

```

상위 주제: [모델 훈련](#) [페이지 104]

관련 정보

[리소스 계획 선택](#) [페이지 104]
[워크플로 템플릿](#) [페이지 106]
[목록 시나리오](#) [페이지 112]
[실행 파일 목록](#) [페이지 111]
[구성 만들기](#) [페이지 121]
[목록 구성](#) [페이지 122]
[아티팩트 서명 사용](#) [페이지 123]
[교육 시작](#) [페이지 131]
[교육 인스턴스 중지](#) [페이지 132]
[교육 인스턴스 삭제](#) [페이지 138]
[효율성 기능](#) [페이지 139]
[교육 일정](#) [페이지 143]
[배포 로그 검색](#) [페이지 194]

7.2.13 훈련 일정

주기적으로 실행을 실행하려면 실행을 자동으로 시작하는 일정을 정의할 수 있습니다. 이렇게 하려면 실행을 위한 구성을 준비해야 합니다.

훈련 일정에는 시작 및 종료 타임스탬프가 있습니다. 시작 타임스탬프는 실행을 생성하기 위해 일정이 처음 실행되는 시기를 정의합니다. 종료 타임스탬프는 일정이 만료되는 시기를 정의합니다.

아직 만료되지 않은 일정은 ACTIVE 상태입니다. 만료되면 상태가 INACTIVE로 변경되고 더 이상 실행이 시작되지 않습니다.

이러한 타임스탬프의 형식은 RFC3339 섹션 5.6에 정의되어 있으며 초 단위는 제외됩니다(예: 2023-02-09T12:53:47Z). 모든 타임스탬프는 UTC 시간으로 해석됩니다. 자세한 내용은 [RFC3339 섹션 5.6을 참조하세요](#). 및 [AI API 개요 \[페이지 24\]](#).

[교육 일정 만들기 \[페이지 142\]](#)

[교육 일정에 따라 생성된 실행 나열 \[페이지 149\]](#)

[기존 교육 일정 변경 \[페이지 150\]](#)

[교육 일정 삭제 \[페이지 150\]](#)

상위 주제: [모델 훈련 \[페이지 104\]](#)

관련 정보

[리소스 계획 선택 \[페이지 104\]](#)

[워크플로 템플릿 \[페이지 106\]](#)

[목록 시나리오 \[페이지 112\]](#)

[실행 파일 목록 \[페이지 111\]](#)

[구성 만들기 \[페이지 121\]](#)

[목록 구성 \[페이지 122\]](#)

[아티팩트 서명 사용 \[페이지 123\]](#)

[교육 시작 \[페이지 131\]](#)

[교육 인스턴스 중지 \[페이지 132\]](#)

[교육 인스턴스 삭제 \[페이지 138\]](#)

[효율성 기능 \[페이지 139\]](#)

[실행 로그 검색 \[페이지 142\]](#)

7.2.13.1 교육 일정 생성

컬 사용하기

1. \$AI_API_URL/v2/lm/에 POST 요청을 제출하여 교육 일정을 만듭니다.

실행 일정.

요청 본문에서 다음을 정의해야 합니다. • cron 형식

의 일정. 자세한 내용은 [Cron 형식 지정을](#) 참조하세요.

- 이름
- 교육 인스턴스에 사용하려는 구성 ID
- 일정이 활성화되어야 하는 시작 타임스탬프
- 일정이 비활성화되어야 하는 종료 타임스탬프

```
컬 --location --request POST "$AI_API_URL/v2/lm/executionSchedules/
$EXECUTION_SCHEDULE" \\--header "승인:
전달자 $TOKEN" \\--header "ai-resource-group: $RESOURCE_GROUP" --data-
raw '{ \"cron\": \"0 * * * *\", \"name\": \"시간별 훈련 실행\", \"configurationId\":
\"35b60591-1e48-473b-9b44-
d5f8e9e4de32\", \"start\":
\"2023-02-10T10:50:31Z\", \"end\": \"2023-02-10T12:55:31Z\" }'
```

2. GET 요청을 제출하여 교육 일정 상태를 확인하세요.

```
컬 --location -- 요청 GET "$AI_API_URL/v2/lm/executionSchedules/
$EXECUTION_SCHEDULE" \\--header "권
한 부여: 전달자 $TOKEN" \\--header "ai-resource-group: $RESOURCE_GROUP"
```

우편 배달부 사용

1. {{apiurl}}/v2/lm/executionSchedules에 POST 요청을 제출하여 교육 일정을 생성합니다.

요청 본문에서 다음을 정의해야 합니다.

- cron 형식의 일정입니다. 자세한 내용은 [Cron 형식 지정을](#) 참조하세요.
- 이름
- 교육 인스턴스에 사용하려는 구성 ID
- 일정이 활성화되어야 하는 시작 타임스탬프
- 일정이 비활성화되어야 하는 종료 타임스탬프

POST <{{apiurl}}/v2/lm/executionSchedules...>

Params Authorization Headers (10) Body **JSON** Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL JSON

```
1 "cron": "0 * * * *",
2 "name": "Hourly training run",
3 "configurationId": "35b60591-1e48-473b-9b44-d5f8e9e4de32",
4 "start": "2023-02-08T10:50:31Z",
5 "end": "2023-02-08T12:55:31Z"
```

Body Cookies Headers (7) Test Results

Pretty Raw Preview Visualize JSON

```
1 {
2   "id": "799b4e67-a213-40b9-9550-637fde75dbda",
3   "message": "Execution Schedule created"
4 }
```

2. {{apiurl}}/v2/lm/에 GET 요청을 제출하여 훈련 일정 상태를 확인합니다.

ExecutionSchedules/{{executionScheduleId}}

GET `{{apiurl}}/v2/lm/executionSchedules/799b4e67-a213-40b9-9550-637fde75dbda`

Params Authorization Headers (8) Body Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL

Body Cookies Headers (5) Test Results

Pretty Raw Preview Visualize JSON `copy`

```

1
2   "configurationId": "35b60591-1e48-473b-9b44-d5f8e9e4de32",
3   "createdAt": "2023-02-09T12:53:47Z",
4   "cron": "@ * * * *",
5   "end": "2023-02-08T12:55:31Z",
6   "id": "799b4e67-a213-40b9-9550-637fde75dbda",
7   "modifiedAt": "2023-02-09T12:53:47Z",
8   "name": "Hourly training run",
9   "start": "2023-02-08T10:50:31Z",
10  "status": "ACTIVE"
11

```

상위 주제: [교육 일정 \[페이지 143\]](#)

관련 정보

[교육 일정에 따라 생성된 실행 나열 \[페이지 149\]](#)

[기존 교육 일정 변경 \[페이지 150\]](#)

[교육 일정 삭제 \[페이지 150\]](#)

7.2.13.2 훈련 일정에 따라 생성된 실행 나열

컬 사용하기

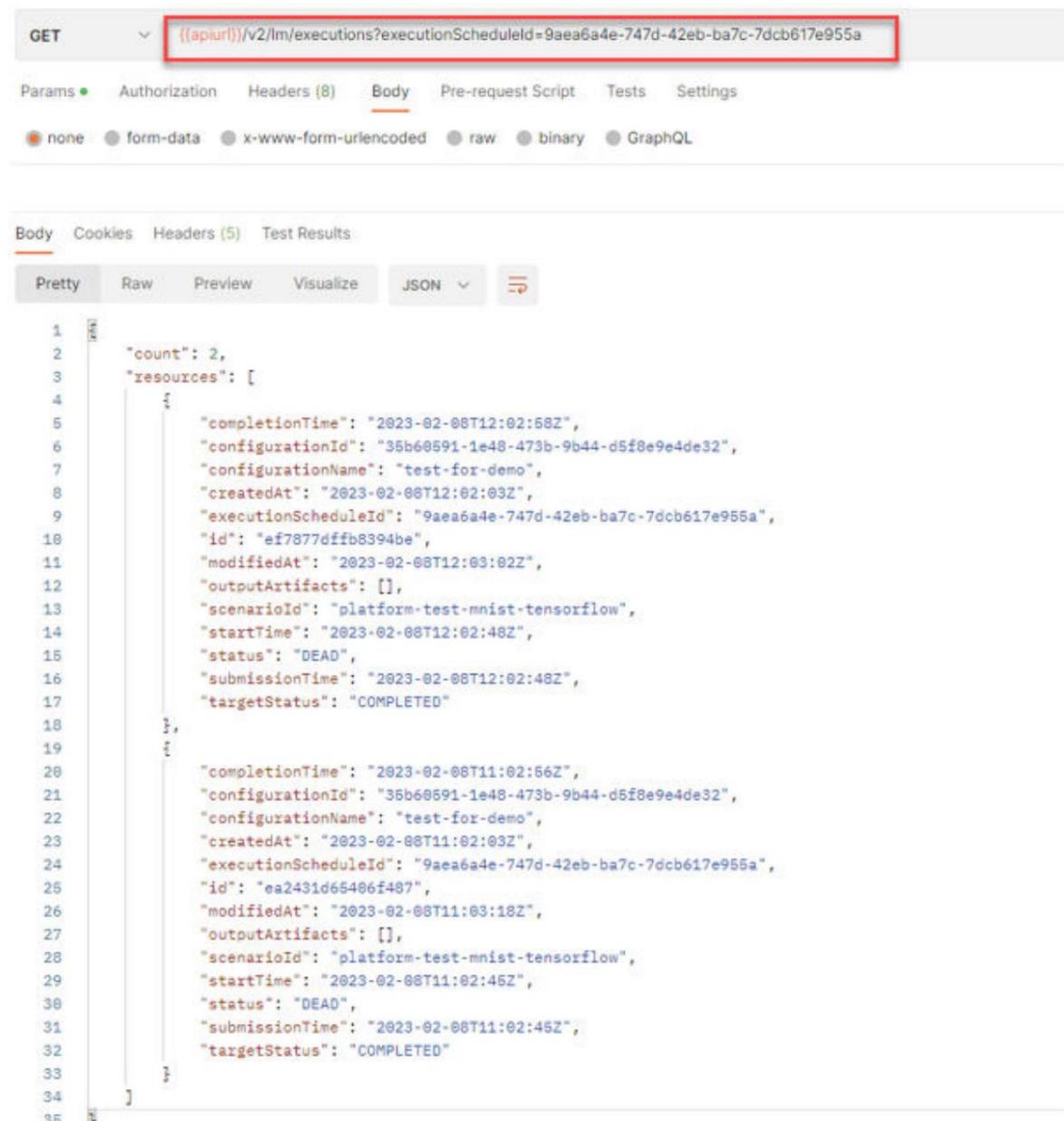
GET 요청을 제출하여 훈련 일정에 따라 생성된 실행 목록을 가져옵니다.

```
컬 --location -- 요청 GET "$AI_API_URL/v2/lm/executionSchedules/
$EXECUTION_SCHEDULE" \\
```

```
--header "권한 부여: 전달자 $TOKEN" \ --header "ai-resource-group:
$RESOURCE_GROUP"
```

우편 배달부 사용

`{apiurl}/v2/lm/executions?executionScheduleId={executionScheduleId}`에 GET 요청을 제출하여 훈련 일정에 따라 생성된 실행 목록을 가져옵니다.



The screenshot shows a Postman interface with the following details:

- Method:** GET
- URL:** `{(apiurl)}/v2/lm/executions?executionScheduleId=9aea6a4e-747d-42eb-ba7c-7dcb617e955a`
- Params:** Params, Authorization, Headers (8), Body (selected), Pre-request Script, Tests, Settings
- Body Content Type:** none, form-data, x-www-form-urlencoded, raw (selected), binary, GraphQL
- Response Body (Pretty JSON):**

```

1
2   "count": 2,
3   "resources": [
4     {
5       "completionTime": "2023-02-08T12:02:56Z",
6       "configurationId": "35b60591-1e48-473b-9b44-d5f8e9e4de32",
7       "configurationName": "test-for-demo",
8       "createdAt": "2023-02-08T12:02:03Z",
9       "executionScheduleId": "9aea6a4e-747d-42eb-ba7c-7dcb617e955a",
10      "id": "ef7877dff8394be",
11      "modifiedAt": "2023-02-08T12:03:02Z",
12      "outputArtifacts": [],
13      "scenarioId": "platform-test-mnist-tensorflow",
14      "startTime": "2023-02-08T12:02:48Z",
15      "status": "DEAD",
16      "submissionTime": "2023-02-08T12:02:48Z",
17      "targetStatus": "COMPLETED"
18    },
19    {
20      "completionTime": "2023-02-08T11:02:56Z",
21      "configurationId": "35b60591-1e48-473b-9b44-d5f8e9e4de32",
22      "configurationName": "test-for-demo",
23      "createdAt": "2023-02-08T11:02:03Z",
24      "executionScheduleId": "9aea6a4e-747d-42eb-ba7c-7dcb617e955a",
25      "id": "ea2431d65406f487",
26      "modifiedAt": "2023-02-08T11:03:18Z",
27      "outputArtifacts": [],
28      "scenarioId": "platform-test-mnist-tensorflow",
29      "startTime": "2023-02-08T11:02:45Z",
30      "status": "DEAD",
31      "submissionTime": "2023-02-08T11:02:45Z",
32      "targetStatus": "COMPLETED"
33    }
34  ]
35

```

상위 주제: [교육 일정](#) [페이지 143]

관련 정보

[교육 일정 만들기 \[페이지 142\]](#)

[기존 교육 일정 변경 \[페이지 150\]](#)

[교육 일정 삭제 \[페이지 152\]](#)

7.2.13.3 기존 교육 일정 변경

컬 사용하기

기존 교육 일정의 크론 정의, 시작, 종료 및 구성을 변경할 수 있습니다.

이름 필드는 변경할 수 없습니다.

패치 요청 제출:

```
컬 --location --request 패치 $AI_API_URL/v2/lm/executions?ExecutionScheduleId=$EXECUTION_SCHEDULE --header
"권한 부여: Bearer $TOKEN" \ --header "ai-resource-group:
$RESOURCE_GROUP" \ -data-raw '[ "cron": "0 0 * * *" ]'
```

우편 배달부 사용

기존 교육 일정의 크론 정의, 시작, 종료 및 구성을 변경할 수 있습니다.

이름 필드는 변경할 수 없습니다.

PATCH 요청의 변경 사항을 {{apiurl}}/v2/lm/executionSchedules/로 보냅니다.
{{executionScheduleId}}.

PATCH {{apiurl}}/v2/lm/executionSchedules/2c5de98b-da10-444a-a62c-f1fc250e9a75

Params Authorization Headers (10) **Body** Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL JSON

```

1
2   "cron": "0 * * * *"
3

```

Body Cookies Headers (5) Test Results

Pretty Raw Preview Visualize JSON

```

1
2   "id": "2c5de98b-da10-444a-a62c-f1fc250e9a75",
3   "message": "Execution Schedule modified"
4

```

훈련 일정 일시 중지 또는 재개

상태 매개변수는 일정을 일시 중지하거나 재개합니다.

- <INACTIVE>: 일정을 일시 중지합니다.
- <ACTIVE>: 스케줄을 재개합니다.

우편 배달부 사용

본문을 사용하여 {{apiurl}}/v2/lm/executionSchedules/{{executionScheduleId}}에 PATCH 요청을 보냅니다.

```
{
  "상태": "비활성"
}
```

컬 사용하기

```
컬 --location --request 패치 $AI_API_URL/v2/lm/executions? ExecutionScheduleId=$EXECUTION_SCHEDULE --header
"권한 부여: Bearer $TOKEN" \ --header "ai-resource-group:
$RESOURCE_GROUP" \ --data-binary '{ "status": "INACTIVE" }'
```

상위 주제: [교육 일정 \[페이지 143\]](#)

관련 정보

[교육 일정 만들기 \[페이지 142\]](#)

[교육 일정에 따라 생성된 실행 나열 \[페이지 149\]](#)

[교육 일정 삭제 \[페이지 150\]](#)

7.2.13.4 교육 일정 삭제

컬 사용하기

ACTIVE 및 INACTIVE 상태의 훈련 일정을 삭제할 수 있습니다.

DELETE 요청을 제출하십시오.

```
컬 --location -- DELETE "$AI_API_URL/v2/lm/executionSchedules/ 요청
$EXECUTION_SCHEDULE" \\ --header "권
한 부여: 전달자 $TOKEN" \ --header "ai-resource-group: $RESOURCE_GROUP"
```

우편 배달부 사용

ACTIVE 및 INACTIVE 상태의 훈련 일정을 삭제할 수 있습니다.

엔드포인트 {{apiurl}}/v2/lm/executionSchedules/ {{executionScheduleId}}에 DELETE 요청을 보냅니다.

상위 주제: [교육 일정 \[페이지 143\]](#)

관련 정보

[교육 일정 만들기 \[페이지 142\]](#)

[교육 일정에 따라 생성된 실행 나열 \[페이지 149\]](#)

[기존 교육 일정 변경 \[페이지 150\]](#)

7.3 모델 사용

AI 학습 모델을 배포하여 추론을 실행합니다.

SAP AI Core는 필요한 컴퓨팅 리소스(예: GPU)를 효율적으로 사용하여 AI 훈련 및 오프라인 배치 추론 작업을 실행할 수 있는 수단을 제공합니다.

SAP AI Core에서 훈련된 모델을 사용하거나 배포 및 추론을 위해 사전 훈련된 자체 모델을 사용할 수 있습니다.

모델 제공 서비스에는 다음이 포함됩니다.

- 온라인 추론 시나리오에 통합할 수 있는 빠르고 안전하며 리소스 효율적인 추론 엔드포인트를 통해 AI 모델을 제공합니다.
- 사용자 제공 Docker 이미지를 배포할 수 있는 유연성.
- 비용 효율적인 서비스(예: 자동 확장, 서비스 지원 또는 다중 모델 서비스).
- 기본 테넌트가 서비스 인스턴스를 관리하기 위한 매개변수화된 서비스 템플릿.
- 특정 매개변수 및 리소스 격리를 사용하여 리소스 그룹 네임스페이스에서 인스턴스를 제공합니다.

제공 템플릿은 모델 서버를 만드는 데 사용됩니다. 모델 서버가 실행되면 들어오는 추론 요청을 처리하고 AI 학습 모델의 결과를 반환합니다. 제공 템플릿은 모델 배포 방법을 정의합니다. 모델이 작동하려면 입력 매개 변수 및 아티팩트와 함께 [제공 템플릿 API 참조 \[페이지 162\]](#)에 따라 사양을 제공해야 합니다.

minReplicas 및 maxReplicas 매개변수를 지정하여 처리에 사용되는 노드 수를 제한합니다.

[리소스 계획 선택 \[페이지 151\]](#)

수요에 따라 다양한 작업에 다양한 인프라 리소스를 사용하도록 SAP AI Core를 구성할 수 있습니다. SAP AI Core는 이러한 목적으로 "리소스 계획"이라는 사전 구성된 여러 인프라 번들을 제공합니다.

[템플릿 제공 \[페이지 153\]](#)

제공 템플릿을 사용하여 기본 테넌트 수준에서 제공 인스턴스를 관리합니다. 제공 템플릿은 모델 배포 방법을 정의합니다.

[실행 파일 목록 \[페이지 169\]](#)

실행 파일은 모델 교육이나 배포 생성과 같은 목적을 위해 인스턴스화되는 템플릿입니다. 리소스 그룹의 모든 실행 파일을 나열하고 리소스 그룹에서 특정 실행 파일의 세부 정보를 가져올 수 있습니다. 제공 템플릿은 배포 실행 파일에 매핑됩니다.

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 중지 \[페이지 183\]](#)

[배포 삭제 \[페이지 181\]](#)

[효율성 기능 \[페이지 189\]](#)

효율성을 향상하고 리소스 소비 관리에 도움이 되는 SAP AI Core 런타임의 기능을 알아보세요.

배포 및 실행 로그에서 액세스할 수 있는 [배포](#)

[로그 검색 \[페이지 194\]](#)

상위 주제: [ML 작업 \[페이지 96\]](#)

관련 정보

[데이터 연결 \[페이지 96\]](#)

[모델 훈련 \[페이지 103\]](#)

7.3.1 리소스 계획 선택

수요에 따라 다양한 작업에 다양한 인프라 리소스를 사용하도록 SAP AI Core를 구성할 수 있습니다. SAP AI Core는 이러한 목적으로 "리소스 계획"이라는 사전 구성된 여러 인프라 번들을 제공합니다.

문맥

리소스 계획은 워크플로 및 제공 템플릿에서 리소스를 선택하는 데 사용됩니다. 워크플로의 각 단계에는 서로 다른 리소스 계획이 있을 수 있습니다.

일반적으로 워크로드에 GPU 가속이 필요한 경우 GPU 지원 리소스 계획 중 하나를 사용해야 합니다.

그렇지 않은 경우에는 워크로드의 예상 CPU 및 메모리 요구 사항을 기반으로 리소스 계획을 선택하세요.

SAP AI Core 내에서 리소스 계획은 Pod 수준의 ai.sap.com/resourcePlan 레이블을 통해 선택됩니다. 그것 선택한 리소스 계획을 매핑하고 다음 리소스 계획 ID 중 하나일 수 있는 문자열 값을 사용합니다.

AWS에 대한 리소스 계획 사양

리소스 계획 ID	GPU	CPU 코어	메모리 GB	재할당 코드 출처 워크플로 온도- 접시
열차-L	1 V100	7	55	ai.sap.com/ 자원계획: 기차.l
추론-S	1 T4	삼	10	ai.sap.com/ 자원계획: 추론하다
추론-M	1 T4	7	26	ai.sap.com/ 자원계획: 추론.m
추론-L	1 T4	15	58	ai.sap.com/ 자원계획: 추론하다.l
기동기	-	1	삼	ai.sap.com/ 자원계획: 기동기
기초적인	-	삼	11	ai.sap.com/ 자원계획: 기초적인
기본-8x	-	31	116	ai.sap.com/ 자원계획: 기본.8x

제한사항

프리 티어 서비스 계획의 경우 스타터 리소스 계획만 사용할 수 있습니다. 다른 계획을 지정하면 문제있는. Standard 서비스 계획의 경우 모든 리소스 계획을 사용할 수 있습니다. 자세한 내용은 [프리 티어를 참조하세요.](#) [\[페이지 37\]](#) 및 [서비스 계획 \[페이지 35\].](#)

참고

이러한 모든 노드에 대한 기본 디스크 저장소 크기에는 제한이 있습니다. 노드에 로드되는 데이터 세트 디스크 공간을 소비합니다. 대규모 데이터 세트(30GB 이상)나 대규모 모델이 있는 경우 다음을 수행할 수 있습니다.
디스크 크기를 늘려야 합니다. 그렇게 하려면 Argo Workflows의 영구 볼륨 청구를 사용하여 필요한 디스크 크기(볼륨 참조).

작업 개요: [모델 사용 \[페이지 151\]](#)

관련 정보

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 중지 \[페이지 183\]](#)

[배포 삭제 \[페이지 181\]](#)

[효율성 기능 \[페이지 189\]](#)

[배포 로그 검색 \[페이지 194\]](#)

서비스 이용 보고

서비스 사용량 소비는 글로벌 계정의 [개요](#) 페이지와 하위 계정의 [개요](#) 및 [사용량 분석](#) 페이지에 있는 SAP BTP 조종석에 보고됩니다. 사용량 보고서에는 청구 가능한 측정값과 청구할 수 없는 측정값의 사용량이 나열됩니다. 최종 월별 청구서는 청구 가능한 측정값만을 기준으로 합니다. 청구 불가능한 측정값은 보고 목적으로만 표시됩니다.

7.3.2 템플릿 제공

제공 템플릿을 사용하여 기본 테넌트 수준에서 제공 인스턴스를 관리합니다. 제공 템플릿은 모델 배포 방법을 정의합니다.

제공 템플릿은 수신 추론 요청을 처리하고 기계 학습 모델의 결과를 반환하는 모델 서버에 하나 이상의 훈련된 모델을 배포하는 데 사용됩니다. 제공 템플릿은 git 저장소에 저장되어 필요에 따라 버전을 지정할 수 있습니다.

SAP AI Core에서 제공 템플릿은 실행 파일로 매핑됩니다. 매핑에는 템플릿의 메타데이터 섹션에 특정 속성이 필요합니다.

모델은 KServe에서 제공하는 간단한 Kubernetes CRD(사용자 정의 리소스 정의)를 사용하여 배포됩니다. 모델을 배포하려면 KServe 사양을 따르고 필수 입력 매개변수와 아티팩트를 정의하는 YAML 사양을 제공해야 합니다.

시작하려면 아래의 일반 계재 템플릿을 복사하고 필요에 따라 고유한 값을 추가하세요. YAML 플러그인이 포함된 텍스트 편집기를 사용하여 템플릿을 생성할 수 있습니다. YAML 사양은 KServe 사양을 따르고 필수 입력 매개변수와 아티팩트를 정의해야 합니다.

템플릿 매개변수 세부정보 제공

유형	매개변수	설명
이름 (필수)		실행 가능 ID입니다. 실행 파일 ID 모든 실행 파일 중에서 고유해야 합니다. SAP AI Core 메인 내에서 사용 가능 거주자.
메타데이터(필수)	이름	제공 템플릿을 고유하게 식별하는 데 사용되는 제공 템플릿의 기술 ID입니다. 이는 제공 템플릿의 실행 파일 ID로 지정되므로 SAP AI 내에서 사용 가능한 모든 실행 파일 중에서 고유해야 합니다.
핵심 주요 테넌트.		
주석	시나리오.ai.sap.com/설명 (선택 사항)	해당 시나리오에 대한 설명 이 실행 파일이 속합니다.
	시나리오.ai.sap.com /name(필수)	꼭 필요한 시나리오 이름 시나리오를 발견하는 AI API.
	실행 파일.ai.sap.com/설명 (선택 사항)	제공 템플릿에 대한 설명입니다.
	실행 파일.ai.sap.com/ 이름 (필수)	제공 템플릿의 이름입니다.
	실행 파일.ai.sap.com/cascade-update-deployments (선택 사항)	제공 템플릿에 대한 변경 사항이 연결된 배포에 계단식으로 적용되도록 허용하는 설정입니다. 변경 제공 템플릿 및 업데이트 배포 [페이지 169] 를 참조하십시오 .
AR에 대한 추가 메타데이터를 추가할 수 있습니다.		
	Artifacts.ai.sap.com/ <argo_artifact_name>.descr lPtion (선택 사항)	이러한 주석을 사용하여 tifact를 만듭니다.
	Artifacts.ai.sap.com/ <argo_artifact_name>.label s: {"ext.ai.sap.com/ 맞춤기1":"맞춤값1" , "ext.ai.sap.com/ 맞춤기2":"맞춤값2" } (선택사항)	
라벨	시나리오.ai.sap.com/id (필수)	제공 템플릿이 속한 시나리오 ID
	ai.sap.com/version (필수)	이 버전은 고객을 위한 것입니다. 호환되는 제공 템플릿 버전을 식별하는 데 사용됩니다.

유형	매개변수	설명
투기	imagePullSecret(선택사항)	Docker 레지스트리 비밀 등록 [페이지 84] .
	Docker 이미지가 다음과 같은 경우에는 필요하지 않습니다. 공개 Docker 저장소에서 가져옴 노상 강도.	
	minReplicas (선택 사항)	배포는 기본적으로 지속적으로 실행됩니다. 그러나 추론이 필요하지 않은 경우 수동으로 중지할 수 있습니다. min-Replicas를 Docker regis-try 비밀의 이름으로 설정합니다. 이름은 귀하의 이름을 참조합니다. 이미지를 가져오기 위한 Docker 자격 증명 Docker 컨테이너의 경우. Docker 레지스트리 비밀 생성에 대한 자세한 내용은 0에서 추론 서버 허용을 참조하세요. 정지 상태가 아닐 때 정지 상태로 들어가려면 사용. 필요할 때 다시 시작됩니다. 수요에 따라 노드 확장 min/maxReplicas 매개변수. 자세한 내용은 효율성을 참조하세요. 기능 [페이지 189] .
	maxReplicas (선택 사항)	내부에 있는 노드 수를 제한합니다. 참조 서버는 최대 규모로 확장됩니다. 자세한 내용은 효율성 기능 [페이지 191] .
라벨(필수)	ai.sap.com/resourcePlan	선택한 항목을 지정해야 합니다. Docker의 ResourcePlanName 레지스트리 비밀. 이름 참조 Docker 자격 증명을 가져오려면 Docker 컨테이너의 이미지입니다. 그만큼 value는 선택한 문자열 값입니다. 리소스 계획(리소스 선택 참조) 계획 [페이지 101] .

일반 제공 템플릿

```
apiVersion: ai.sap.com/v1alpha1
종류: ServingTemplate
metadata:
  이름: text-clf-infer-tutorial
  주석:
    시나리오.ai.sap.com/description: "SAP 개발자 튜토리얼 시나리오"
    시나리오.ai.sap.com/name: "text-clf-tutorial-scenario"
    실행 파일.ai.sap.com/description: "텍스트에 대한 추론 실행 파일
    Scikit-learn을 사용한 분류"
    실행 파일.ai.sap.com/name: "text-clf-infer-tutorial-exec"
    Artifacts.ai.sap.com/textmodel.kind: "모델"
    Artifacts.ai.sap.com/textmodel.description: "아티팩트 설명"
    Artifacts.ai.sap.com/textmodel.labels: | {"ext.ai.sap.com/
      customkey1":"customvalue1", "ext.ai.sap.com/
      customkey2":"customvalue2"}
```

```

라벨: 시나리오
    오.ai.sap.com/id: "text-clf-tutorial" ai.sap.com/version: "1.0.0"

특기:
    입력: 매개변수:

        - 이름: 모델명
        기본값: 값 유형: 문자열 설명:
        매개변수 설명 아티팩트:

        - 이름: textmodel 템플릿: apiVersion:
    "serving.kserve.io/
        v1beta1" 메타데이터: 주석: | autoscaling.knative.dev/metric: 동시성 autoscaling.knative.dev/
        target: 1
        autoscaling.knative.dev/
        targetBurstCapacity: 0 레이블: | ai.sap.com/resourcePlan: 스타터 사양: | 예측기: imagePullSecrets: - 이름:
        <Docker 레지스터리 비밀 이름> minReplicas: 0 maxReplicas: 5개 컨테이너: - 이름:
        kserve-container

image: "<DOCKER IMAGE URL GOES HERE>" 포트: - 컨테이너 포트: 9001 프로토콜:
TCP

환경:
    - 이름: STORAGE_URI
    값: "{{inputs.artifacts.textmodel}}"

```

- STORAGE_URI 환경 변수 이름은 현재 KService에 하드 코딩되어 사용자 정의 예측기가 env var STORAGE_URI로 구성될 때 모델을 다운로드해야 함을 나타냅니다. • /mnt/models는 현재 KService에 하드 코딩되어 있습니다. 자신의 Docker 컨테이너로 이 예제를 시도할 때 해당 경로에서 모델을 읽으십시오.

수동으로 애플리케이션 동기화

애플리케이션은 최대 3분 간격으로 자동으로 GitHub 저장소와 동기화됩니다. 아래 엔드포인트를 사용하여 수동으로 동기화를 요청하세요.

[/{{apiurl}}/admin/applications/{{appName}}/refresh]({{apiurl}}/admin/applications/{{appName}}/refresh)

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)
[추론 \[페이지 173\]](#)
[배포 업데이트 \[페이지 179\]](#)
[배포 중지 \[페이지 183\]](#)
[배포 삭제 \[페이지 181\]](#)
[효율성 기능 \[페이지 189\]](#)
[배포 로그 검색 \[페이지 194\]](#)

다중 배포 중지 또는 삭제

이 기능은 기본적으로 false로 설정되어 있습니다. 대량 PATCH 작업을 활성화하려면 템플릿에 관련 값이 true로 설정된 다음 코드 조각이 포함되어야 합니다.

```
메타:  
"bulkUpdates": { "executions": false,  
                 "deployments": false  
 }
```

관련 정보

[다중 배포 중지 \[페이지 181\]](#)
[여러 배포 삭제 \[페이지 189\]](#)

7.3.2.1 템플릿 API 참조 제공

API 스키마 사양 [ai.sap.com/v1alpha1](#) [페이지 163]

패키지 v1alpha1에는 제공 v1alpha1 API 그룹(ai.sap.com/v1alpha1)에 대한 API 스키마 정의가 포함되어 있습니다.

KServe 사양 [serving.kserve.io/v1beta1](#) [페이지 166]

패키지 v1beta1에는 제공 v1beta1 API 그룹(serving.kserve.io/v1beta1)에 대한 API 스키마 정의가 포함되어 있습니다.

7.3.2.1.1 API 스키마 사양 ai.sap.com/v1alpha1

파키지 v1alpha1에는 제공 v1alpha1 API 그룹(ai.sap.com/)에 대한 API 스키마 정의가 포함되어 있습니다.
v1alpha1).

서빙 템플릿

ServingTemplate은 모델 제공 방법을 지정하는 실행 파일 유형입니다.

필드	설명
metadata	
Kubernetes 메타/v1.ObjectMeta	<p>이름 (필수) 기술 ID 현재 템플릿은 고유하게 식별하는 데 사용됩니다. 서빙 템플릿.</p>
주석	정의하는 주석
템플릿 메타데이터 제공 주석 [페이지 161]	시나리오 및 실행 파일 정보.
라벨	Sce-를 정의하는 라벨
템플릿 메타데이터 제공 레이블 [페이지 160]	nario 및 실행 파일 정보.
투기	ServingTemplateSpec에는 KServe용 템플릿이 포함되어 있습니다. KServe CR을 생성하는 데 필요한 매개변수 및 아티팩트.
ServingTemplate 사양 [페이지 162]	

ServingTemplate 메타데이터 주석

(ServingTemplate [페이지 161] 에 표시됨)

지원되는 Kubernetes 메타데이터 주석의 하위 집합

필드	설명
시나리오.ai.sap.com/description	시나리오에 대한 설명(예: "SAP 개발자 튜토리얼 시나리오")
시나리오.ai.sap.com/name	제공 템플릿의 이름(예: 'text-clf-tutorial-scenario')
실행 파일.ai.sap.com/description	제공 템플릿에 대한 설명입니다. (예: "Scikit-learn을 사용한 텍스트 분류를 위한 추론 실행 파일")

필드	설명
실행 파일.ai.sap.com/name	제공 템플릿의 이름(예: 'text-clf-in-fer-tutorial-exec')

ServingTemplate 메타데이터 라벨

(ServingTemplate [페이지 161] 에 표시됨)

지원되는 Kubernetes 메타데이터 라벨의 하위 집합

필드	설명
ai.sap.com/version	시나리오 버전(예: "1.0.0")
시나리오.ai.sap.com/id	시나리오의 고유 ID(예: "1234-abcd-efg")

ServingTemplate 사양

(ServingTemplate [페이지 161] 에 표시됨)

ServingTemplate 사양은 이 리소스의 최상위 유형입니다.

필드	설명
입력	
매개변수	에 대한 입력 매개변수 KServe 추론 서비스 템플릿 [페이지 161]
ServingTemplate 입력 매개 변수 [페이지 160]	
인공물	에 대한 입력 아티팩트 KServe 추론 서비스 템플릿 [페이지 161] [페이지 161]
ServingTemplate 입력 아티팩트 [페이지 161]	
주형	KServe CR 템플릿
	KServe InferenceService 템플릿 [페이지 163]

ServingTemplate 입력 매개변수

(ServingTemplate 사양 [페이지 162] 에 표시됨)

KServe InferenceService 템플릿 [페이지 161] 사양에 필요한 입력 매개 변수

필드	설명
이름	매개변수 이름
기본	(선택사항) 매개변수의 기본값
유형	(선택 사항) 매개변수의 유형입니다. "문자열" 유형만 지원됩니다.

ServingTemplate 입력 아티팩트

(ServingTemplate 사양 [페이지 162] 에 표시됨)

KServe InferenceService 템플릿 [페이지 161] 사양에 필요한 입력 아티팩트

필드	설명
이름	아티팩트 이름

KServe 추론 서비스 템플릿

(표시 :ServingTemplate 사양 [페이지 164])

KServe InferenceService 템플릿

필드	설명
api버전	KServe API 버전. 예를 들어, serving.kserve.io/v1beta1.
끈	
metadata	KServe InferenceService CR의 메타데이터입니다. 보다 KServe 메타데이터 주석 [페이지 162] 및 KServe 메타데이터 레이블 [페이지 163] . 주석 메타데이터의 레이블은 다음을 통해 매개변수화될 수 있습니다. {{inputs.parameters.Parameter-Name-Here}} 사용
투기	KServe InferenceServiceSpec 을 위한 여러 줄 문자열 템플릿 [페이지 161] 매개 변수에 대한 선택적 자리 표시자와 인공물
여러 줄 문자열	

상위 주제: [템플릿 API 참조 제공 \[페이지 162\]](#)

관련 정보

[KServe 사양](#)serving.kserve.io/v1beta1 [페이지 166]

7.3.2.1.2 KServe 사양 [serving.kserve.io/v1beta1](#)

파키지 v1beta1에는 제공 v1beta1 API 그룹(serving.kserve.io/v1beta1)에 대한 API 스키마 정의가 포함되어 있습니다.

추론 서비스

InferenceService는 InferenceServices API의 스키마입니다.

필드	설명				
metadata	해당 필드에 대해서는 Kubernetes API 문서를 참조하세요.				
Kubernetes 메타/v1.ObjectMeta	메타 데이터 필드.				
투기					
KServe InferenceServiceSpec [페이지 167]	<table> <tr> <td>예언자</td> <td>예측자가 모델을 정의합니다.</td> </tr> <tr> <td>KServe PredictorSpec [페이지 168]</td> <td>서빙 사양</td> </tr> </table>	예언자	예측자가 모델을 정의합니다.	KServe PredictorSpec [페이지 168]	서빙 사양
예언자	예측자가 모델을 정의합니다.				
KServe PredictorSpec [페이지 168]	서빙 사양				

KServe 메타데이터 주석

(InferenceService [페이지 162] 에 표시됨)

지원되는 KServe 메타데이터 레이블의 하위 집합을 포함하는 여러 줄 문자열입니다. {{variable_name}}을 사용하여 자리 표시자를 추가하여 변수를 정의하고 값을 동적으로 바꿀 수 있습니다. 예를 들어:

```
autoscaling.knative.dev/target: "{{inputs.parameters.MyAutoScalingTarget}}"
```

필드	설명
autoscaling.knative.dev/metric	자동 확장에 사용할 Knative 자동 확장 측정항목(예: 'rps', '동시성')
autoscaling.knative.dev/target	Knative 자동 확장 대상 번호(예: '1')
autoscaling.knative.dev/targetBurstCapacity	Knative 자동 확장 목표 버스트 용량 수(예: '70')
autoscaling.knative.dev/window	Knative 자동 확장 기간(예: '10s')
autoscaling.knative.dev/scaleToZeroPodRetentionPeriod	Autoscaler가 Pod 크기를 0으로 조정하기로 결정한 후 마지막 Pod가 활성 상태로 유지되는 최소 시간을 결정합니다. 기본값: 0초

예

```
메타데이터: 주석: |
  autoscaling.knative.dev/metric:
    rps autoscaling.knative.dev/target: {{inputs.parameters.MyAutoScalingTarget}}
    autoscaling.knative.dev/targetBurstCapacity: 70 autoscaling.knative.dev/
  window: 10s
```

KServe 메타데이터 라벨

(InferenceService [페이지 162] 에 표시됨)

지원되는 KServe 메타데이터 레이블의 하위 집합을 포함하는 여러 줄 문자열입니다. {{variable_name}}을 사용하여 자리 표시자를 추가하여 변수를 정의하고 리소스 계획을 동적으로 바꿀 수 있습니다. 예를 들어:

필드	설명
ai.sap.com/resourcePlan:	"{{inputs.parameters.MyResourcePlan}}"

필드	설명
ai.sap.com/resourcePlan	KServe 기반 생성에 사용되는 리소스 계획 모델 서버(예: "기본")

예

```
metadata:
  라벨: |
    ai.sap.com/resourcePlan: 기본
```

KServe InferenceServiceSpec

(InferenceService [페이지 162] 에 표시됨)

InferenceServiceSpec은 이 리소스의 최상위 유형입니다.

필드	설명
예언자	예측자는 모델 제공 사양을 정의합니다.

[KServe PredictorSpec \[페이지 168\]](#)

KServe 예측기 사양

(KServe InferenceServiceSpec [페이지 167] 에 표시됨)

PredictorSpec은 예측기의 구성을 정의합니다. 다음 필드는 "1-of" 의미 체계를 따릅니다. 사용자는 정확히 하나의 사양을 지정해야 합니다.

필드	설명
컨테이너	Pod에 속한 컨테이너 목록입니다. 컨테이너는 다음을 수행할 수 있습니다. 현재 추가되거나 제거되지 않습니다. 이 있어야합니다 포드에 있는 컨테이너 중 하나 이상은 업데이트할 수 없습니다.
이미지PullSecrets	동일한 리소스 그룹의 비밀에 대한 참조 목록이며 이 사양에서 사용되는 이미지를 가져오는 데 사용됩니다. 비밀 이름은 Docker Registry Secret 생성 API를 통해 생성된 이름과 동일해야 합니다. 자세한 내용은 imagePullSecrets 자정 을 참조하세요.
아래에	
minReplicas (선택사항)	최소 복제본 수입니다. 기본값은 1이지만 설정 가능 0으로 조정하여 0으로 조정합니다.
정수	
maxReplicas (선택사항)	자동 확장을 위한 최대 복제본 수입니다. 기본값: $\text{minReplicas} \geq 20$ 면 $\text{maxReplicas} = \text{minReplicas}$ 그렇지 않으면 2입니다.
정수	
컨테이너 동시성 (선택사항)	동시에 처리할 수 있는 요청 수를 지정합니다. 이는 동시에 컨테이너의 하드 제한을 설정합니다. 렌시.
정수64	https://knative.dev/docs/serving/autoscaling/concurrency .
타이밍 전에 대기할 시간(초)을 지정합니다.	
시간 초과 (선택 사항)	타이밍 전에 대기할 시간(초)을 지정합니다. 구성 요소에 대한 요청을 보냅니다.
정수64	
종료GracePeriodSeconds (선택 사항)	Pod가 정상적으로 종료되어야 하는 선택적 시간(초)은 삭제 요청 시 줄어들 수 있습니다. 값은 음수가 아닌 정수여야 합니다. 값 0은 즉시 삭제를 의미합니다. 이 값이 nil이면 기본 유예 기간이 대신 사용됩니다. 유예 기간은 포드에서 실행 중인 프로세스에 종료 신호가 전송된 후의 시간(초)과 종료 신호로 프로세스가 강제로 중지되는 시간입니다. 이 값을 프로세스의 예상 정리 시간보다 길게 설정하십시오. 기본값은 30초입니다.
정수64	

예

사양: |

```

예측자:
  imagePullSecrets: - 이름:
    [DOCKER REGISTRY SECRET GOES HERE]
  minReplicas: 1 maxReplicas:
  5개 컨테이너: - 이름: kserve-
  컨테이너 이미지:
  "[DOCKER IMAGE URL GOES HERE]" 포트: - 컨테이너
  포트: 9001 프로토콜: TCP

환경:
- 이름: STORAGE_URI
  값: "{{inputs.artifacts.textmodel}}"

```

상위 주제: [템플릿 API 참조 제공](#) [페이지 162]

관련 정보

[API 스키마 사양 ai.sap.com/v1alpha1](#) [페이지 163]

7.3.2.2 제공 템플릿 변경 및 배포 업데이트

제공 템플릿을 변경하면 해당 템플릿과 연결된 배포를 자동으로 업데이트할 수 있습니다.

전제조건

실행 파일 `ai.sap.com/cascade-update-deployments` 매개변수가 존재하며 제공 템플릿에 `true`로 설정됩니다. 이를 통해 제공 템플릿을 변경하면 연결된 배포의 자동 업데이트를 트리거 할 수 있습니다.

절차

1. 제공 템플릿의 필드를 필요한 대로 변경합니다. 다음 변경 사항이 지원됩니다.

- `spec.template.spec.default.predictor.minReplicas`
- `spec.template.spec.default.predictor.custom.container.image`
- `spec.template.spec.default.predictor.custom.container.ports`
- `spec.template.spec.default.predictor.custom.container.env`

2. 베타 제공 템플릿의 필드를 필요한 대로 변경합니다. 다음 사항에 대한 변경 사항은 다음과 같습니다.

지원됨:

- spec.template.spec.predictor.containers.image
- spec.template.spec.predictor.containers.ports
- spec.template.spec.predictor.containers.env
- 값 수정
- 이름-값 쌍 추가
- 매개변수화된 값을 하드코드로 업데이트하거나 그 반대로 업데이트할 수 있습니다.

7.3.3 실행 파일 나열

실행 파일은 모델 교육이나 배포 생성과 같은 목적을 위해 인스턴스화되는 템플릿입니다. 리소스 그룹의 모든 실행 파일을 나열하고 리소스 그룹에서 특정 실행 파일의 세부 정보를 가져올 수 있습니다. 제공 템플릿은 배포 실행 파일에 매핑됩니다.

Postman을 사용하여 실행 파일 나열

- 시나리오 ID를 시나리오 ID 환경 변수로 정의합니다.

	client_secret	XXX	XXX
<input checked="" type="checkbox"/>	token	XXX	XXX
<input checked="" type="checkbox"/>	apiurl		
<input checked="" type="checkbox"/>	scenarioid	dba85cf3-2d69-498d-<give your scenario id here>	
<input checked="" type="checkbox"/>	configurationid	ecff88fb-0c50-4fc3-8439-f87f25e0ba08	ecff88fb-0c50-4fc3-8439-f87f25e0ba08
<input checked="" type="checkbox"/>	executionid	dbba7904390cda65	dbba7904390cda65
<input checked="" type="checkbox"/>	variable_key		
Add a new variable			

Use variables to reuse values in different places. Work with the current value of a variable to prevent sharing sensitive values with your team. [Learn more about variable values](#)

- 엔드포인트 {{apiurl}}/v2/lm/scenarios/{{scenarioid}}/executables에 GET 요청을 보냅니다.
- Authorization 탭에서 유형을 Bearer Token으로 설정합니다.

- токن 값을 {{token}}으로 설정합니다.

Params	Authorization	Headers (9)	Body	Pre-request Script	Tests	Settings	Cookies	Code
TYPE <input type="text" value="Bearer Token"/> <div style="margin-top: 10px;">The authorization header will be automatically generated when you send the request. Learn more about</div>								
Heads up! These parameters hold sensitive data. To keep this data secure while working in a collaborative environment, we recommend using variables. Learn more about variables								
Token <input type="text" value="{{token}}"/>								

5. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> 기본값이 사용됨)

6. 요청을 보냅니다.

GET `{apiurl}/v2/lm/scenarios/{scenarioId}/executables ...`

Headers (8)

Key	Value	Description
AI-Resource-Group	default	

Body Cookies Headers (5) Test Results

Pretty Raw Preview Visualize JSON

```

1
2   "count": 3,
3   "resources": [
4     {
5       "createdAt": "2021-10-07T20:07:18+00:00",
6       "deployable": true,
7       "description": "Inference executable for text classification with Scikit-learn",
8       "id": "text-clf-infer-tutorial",
9       "inputArtifacts": [

```

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

컬을 사용하여 실행 파일 나열

```
curl --request GET "$AI_API_URL/v2/lm/scenarios" --header "권한 부여: Bearer $TOKEN" --header "ai-리소스 그룹: $RESOURCE_GROUP"
```

출력 코드

```
{
  "개수": 4, "자원": [
    {
      "createdAt": "2021-02-04T13:11:01+00:00", "deployable": true,
      "description": "실행 가능한 설명을 제 공하는 n 텍스트 클래스 이탈", "id": "pytf-serving", "inputArtifacts": [
        {
          "이름": "model_uri"
        }
      ],
      "라벨": [

```

```

    ],
    "modifiedAt": "2021-02-04T13:11:01+00:00", "name": "churntextclassexecname", "매개변수": [
        {
            "name": "modelName", "type": "string",
            "default": "value", "description":
                "매개변수 설명"
        }
    ],
    "scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
},
{
    "createdAt": "2021-02-09T07:35:02+00:00", "deployable": true, "description": "실행 가능 설명을 제공하는 n 텍스트 클래스 이탈",
    "id": "pytf-serving-tracking", "inputArtifacts": [
        {
            "name": "textmodel", "kind": "모델",
            "description": "아티팩트 설명",
            "labels": [
                {
                    "키": "ext.ai.sap.com/customkey1", "값": "customvalue1"
                },
                {
                    "키": "ext.ai.sap.com/customkey2", "값": "customvalue2"
                }
            ]
        }
    ],
    "ラベル": [
        ],
        "modifiedAt": "2021-02-09T07:35:02+00:00", "name": "churntextclassexecname", "매개변수": [
        {
            "이름": "모델이름", "유형": "문자열"
        }
    ],
    "scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
},
{
    "createdAt": "2021-02-09T07:35:03+00:00", "deployable": false, "description": "churn 및 텍스트 클래스 실행 파일 설명", "id": "pytf-training-tracking", "inputArtifacts": [
        {
            "이름": "이탈 데이터"
        },
        {
            "이름": "텍스트클래스-데이터"
        }
    ],
    "라벨": [
        ],
        "modifiedAt": "2021-02-09T07:35:03+00:00", "name": "chunntextclassexecutablename",

```

```

"출력아티팩트": [
  {
    "name": "churn-pickle", "kind": "모델",
    "description": "아티팩트 설명",
    "labels": [ {
      "key": "ext.ai.sap.com/customkey1", "value": "customvalue1"
    },
    { "key": "ext.ai.sap.com/customkey2", "value": "customvalue2"
    }
  ],
  {
    "이름": "pytf-모델"
  }
],
"매개변수": [ {
  "이름": "기차 시대", "유형": "문자열"
},
{
  "scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
},
{
  "createdAt": "2021-02-04T14:11:02+00:00", "deployable": false, "description": "이탈 및 텍스트 클래스 실행 파일 설명", "id": "test-training", "inputArtifacts": [
    {
      "이름": "이탈 데이터"
    },
    {
      "이름": "텍스트클래스-데이터"
    }
],
"modifiedAt": "2021-02-04T14:11:02+00:00", "name": "chunntextclassexecutablename",
"outputArtifacts": [
  {
    "이름": "취정기 피클"
  },
  {
    "이름": "pytf-모델"
  }
],
"매개변수": [ {
  "이름": "기차 시대", "유형": "문자열"
}
],
"scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
}
]
}

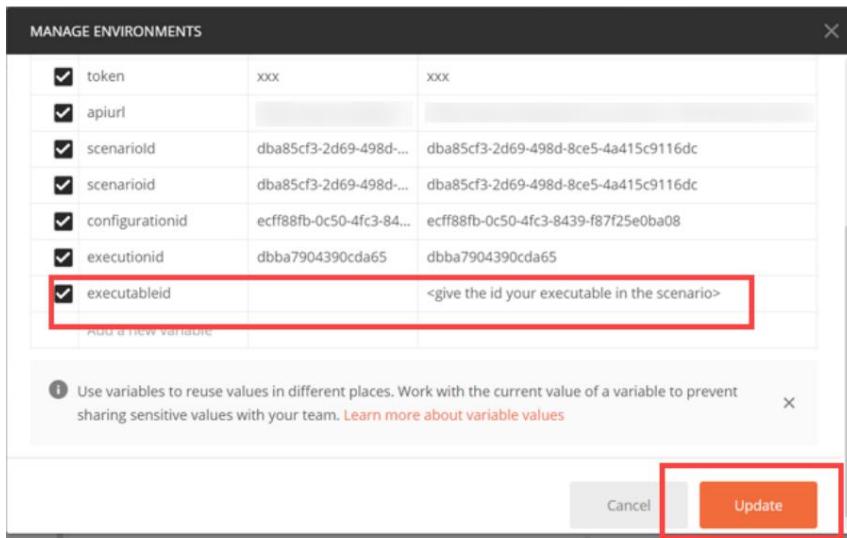
```

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

Postman으로 실행 가능한 세부정보 가져오기

- 환경 변수 runningid를 추가하고 해당 값으로 실행 파일의 ID를 입력합니다.

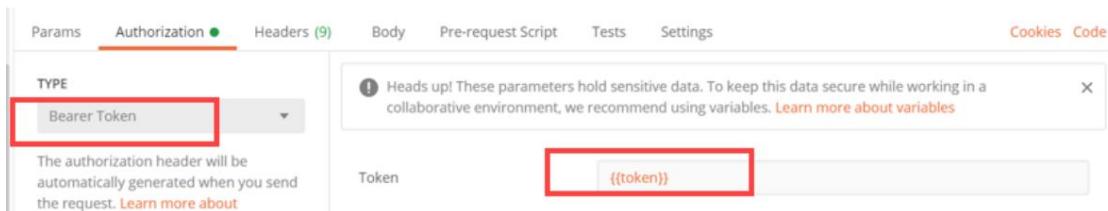


- 엔드포인트 {{apiurl}}/v2/lm/scenarios/{{scenarioId}}/에 GET 요청을 보냅니다.

실행 파일/{{executableId}}

- 권한 부여 탭에서 유형을 Bearer Token으로 설정합니다.

- токن 값을 {{token}}으로 설정합니다.



- 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예에서는 기본값이 사용됨)

- 요청을 보냅니다.

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

컬을 사용하여 실행 가능한 세부정보 가져오기

```
컬 --request GET "$AI_API_URL/v2/lm/scenarios" --header "권한 부여: Bearer  
$TOKEN" --header "ai-리소스 그룹: $RESOURCE_GROUP"
```

출력 코드

```
{
  "createdAt": "2021-02-04T14:11:02+00:00", "deployable": false, "description": "이  
탈 및 텍스트 클래스 실행 파일 설명",
  "id": "test-training", "입력아티팩트": [
    {
      "이름": "이탈 데이터"
    },
    {
      "이름": "텍스트클래스-데이터"
    }
  ],
  "라벨": [
    ],
    "modifiedAt": "2021-02-04T14:11:02+00:00", "name": "chunntextclassexecutablename",
    "outputArtifacts": [
      {
        "이름": "휘젓기 피클"
      },
      {
        "이름": "pytf-모델"
      }
    ],
    "매개변수": [
      {
        "이름": "기차 시대", "유형": "문자열"
      }
    ],
    "scenarioid": "ae0bd260-41ef-4162-81b0-861bd78a8516", "versionId": "0.0.1"
  }
}
```

참고

<modifiedAt> 필드는 최근 성공한 동기화의 타임스탬프를 나타냅니다. 출력 1970-01-01T00:00:00+00:00은 오류를 나타냅니다.

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[모델 배포 \[페이지 172\]](#)[추론 \[페이지 173\]](#)[배포 업데이트 \[페이지 179\]](#)[배포 중지 \[페이지 183\]](#)[배포 삭제 \[페이지 181\]](#)[효율성 기능 \[페이지 189\]](#)[배포 로그 검색 \[페이지 194\]](#)

7.3.4 모델 배포

우편 배달부 사용

1. 앤드포인트 {{apiurl}}/v2/lm/deployments에 POST 요청을 보냅니다.

2. 요청 본문에 구성 ID를 전달합니다.

The screenshot shows the Postman application interface. At the top, it says "POST" and the URL is "{{apiurl}}/v2/lm/deployments". Below the URL, there are tabs for "Params", "Authorization", "Headers (11)", "Body", "Pre-request Script", "Tests", and "Setting". The "Headers" tab is selected. Under "Headers", there is a table with one row: "AI-Resource-Group" with value "trial-1". Below this, under "Body", there is another table with two rows: "Key" and "Value". At the bottom, there are tabs for "Pretty", "Raw", "Preview", "Visualize", "JSON", and "Copy". The "Pretty" tab is selected. The JSON response is displayed in a red-bordered box:

```

1  {
2    "deploymentUrl": "",
3    "id": "d94771b082c5cbcc",
4    "message": "Deployment scheduled.",
5    "status": "UNKNOWN"
  
```

3. {{apiurl}}/v2/lm/deployments/{{배포 ID}}에 GET 요청을 보내 배포 상태를 확인합니다.

{{배포 ID}}.

The screenshot shows a Postman interface with the following details:

- Method:** GET
- URL:** {{apiurl}}/v2/lm/deployments/d94771b082c5cbcc
- Headers:** (9)

KEY	VALUE	DESCRIPTION
AI-Resource-Group	trial-1	
Key	Value	Description
- Body:** (Pretty, Raw, Preview, Visualize, JSON) - Shows a JSON response with deployment details.
- Response Headers:** 200 OK, 877 ms, 673 B
- Body Content (JSON):**

```

5   "deploymentUrl": "",
6     "id": "d94771b082c5cbcc",
7     "modifiedAt": "2021-09-27T10:48:51Z",
8     "scenarioId": "b5379278-887c-4156-9a65-a6c11d6f1a71",
9     "startTime": "2021-09-27T10:33:12Z",
10    "status": "RUNNING".
  
```

참고

상태가 비활성 또는 보류 중이면 배포에 오류가 있을 수 있습니다. 자세한 내용은 배포 로그를 확인하고 [배포 로그 검색 \[페이지 194\]](#)을 참조하십시오.

컬 사용하기

1. 배포를 트리거합니다.

```

컬 --request POST $AI_API_URL/v2/lm/deployments \
--header "권한 부여: 전달자 $TOKEN" \
--header "ai-resource-group: $RESOURCE_GROUP" \
--data-raw '{

  "configurationId": "2b72d740-5a89-4cf7-b37c-85973eed6ae"

}'
  
```

출력 코드

```
{
  "deploymentUrl": "", "id": "dda5d19065d5b1f4", "message": "배포가 생성되었습니다.", "status": "UNKNOWN"
}
```

2. 나중에 사용할 수 있도록 배포 환경 변수를 기록해 둡니다.

3. 배포 상태를 확인합니다.

```

컬 --request GET $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \
--header "승인: 전달자 $TOKEN" \
--header "ai-리소스 그룹: $RESOURCE_GROUP"
  
```

출력 코드

```
{
  "configurationExecutableId": "hello-tf-1-15", "configurationId": "2b72d740-5a89-4cf7-
  b37c-85973eed6ae", "configurationName": "hello-tf-1-15-config", "createdAt": "2020-11-11T06:34:24Z", "createdBy": "사용
  자", "deploymentUrl": "https://my-deployment-url.com", "id": "d291766fd1072b3f",
  "modifiedAt": "2020-11-11T06:37:29Z", "modifiedBy": "사용자", "scenarioId":
  "dba85cf3-2d69-498d-8ce5-4a415c9116dc",
  "status": "실행 중", "targetStatus": "실행 중", "versionId": "0.1.0"
}
```

참고

상태가 비활성 또는 보류 중이면 배포에 오류가 있을 수 있습니다. 자세한 내용은 배포 로그를 확인하고 [배포 로그 검색 \[페이지 194\]](#)을 참조하십시오.

선택적 매개변수

`ttl` 매개변수를 사용하여 배포 시간을 제한할 수 있습니다. 수량에는 정수가 필요하고 시간 단위를 지정하려면 단일 문자가 필요합니다. 분 (m), 시 (h), 일 (d)만 지원되며 값은 자연수여야 합니다. 예를 들어 `"ttl": "5h"`는 배포 기간을 5시간으로 지정합니다. **4.5h** 및 **4h30m**은 유효한 입력이 아닙니다. 값이 전달되지 않으면 배포 기간이 무기한입니다. 기간이 만료되면 배포가 중지되고 삭제됩니다.

배포 할당량

각 테넌트에는 배포 수와 배포당 복제본 수를 제한하는 기본 할당량이 할당됩니다. 이 할당량에 도달하면 배포가 생성되지 않으며 알림이 전송됩니다. 기존 배포를 삭제하여 할당량을 확보할 수 있습니다.

또는 티켓을 생성하여 할당량 증가를 요청할 수 있습니다. 구성 요소 이름은 CA-ML-AIC이고 티켓 제목은 **할당량 증가** 요청입니다.

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)[실행 파일 목록 \[페이지 169\]](#)[추론 \[페이지 173\]](#)[배포 업데이트 \[페이지 179\]](#)[배포 중지 \[페이지 183\]](#)[배포 삭제 \[페이지 181\]](#)[효율성 기능 \[페이지 189\]](#)[배포 로그 검색 \[페이지 194\]](#)

지원하다. ↗

[서비스 계획 \[페이지 35\]](#)

7.3.5 추론

모델 배포의 URL을 사용하여 모델 결과에 액세스하세요.

모델을 추론하려면 "\$DEPLOYMENT_URL/<정의한 경로>" 에 POST 요청을 보냅니다.

예를 들어 TensorFlow 모델은 v1/models/\$MODEL_NAME:predict 경로를 따릅니다.

Postman을 이용한 추론

배포 URL로 {{apiurl}}/v2/lm/ 배포 호출의 응답 본문에 반환된 URL을 전달합니다([모델 배포 \[페이지 176\]](#) 참조).

요청 본문으로 JSON 형식의 샘플 인스턴스를 입력합니다.

The screenshot shows the Postman interface with a POST request to `{{deploymenturl}}/v1/models/churn:predict`. The request body is a JSON object with the following fields:

```

1  {
2    "region": 3,
3    "tenure": 37,
4    "age": 53,
5    "address": 13,
6    "income": 48,
7    "employ": 10,
8    "gender": 0,
9    "equip": 0,
10   "longmon": 13,
11   "tollmon": 24,
12   "equipmon": 0,
13   "cardmon": 10,
14   "longten": 440,
15   "tollten": 876,
16   "equipten": 0,

```

컬을 이용한 추론

이 예에서 모델 이름은 "churn"입니다.

```
컬 --location --request POST '$AI_API_URL/v1/models/churn:predict' \
```

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 중지 \[페이지 183\]](#)

[배포 삭제 \[페이지 181\]](#)

[효율성 기능 \[페이지 189\]](#)

[배포 로그 검색 \[페이지 194\]](#)

7.3.6 배포 업데이트

추론 URL을 유지하면서 새 구성으로 배포를 업데이트할 수 있습니다.

새 배포 구성으로 전환하는 동안 추론 요청은 계속 작동합니다.

새 구성이 배포되면 배포가 "비활성" 상태로 종료되거나 "보류 중" 상태에 머물며 잘못된 구성으로 인해 "실행 중" 상태가 되지 않을 수 있습니다. 이 경우 추론 요청은 더 이상 작동하지 않습니다. 마지막으로 알려진 실행 구성 ID는 [최신 실행 구성 ID](#) 필드에 기록되어 있으며 다른 PATCH 요청에서 마지막 실행 구성으로 돌아가는 데 사용될 수 있습니다.

업데이트된 배포가 "실행 중" 상태에 도달하면 [최신 RunningConfigurationId](#)가 새 구성으로 업데이트됩니다.

전제조건

- 배포 상태는 "보류 중", "실행 중" 또는 "죽음"이어야 합니다.
- 새 구성에는 현재 활성 구성과 동일한 시나리오 ID 및 실행 가능 ID가 포함되어 있습니다.

구성.

참고

죽은 배포는 7일 이내에만 패치할 수 있습니다. 배포가 DEAD 상태에 도달한 후 7일이 지나면 배포가 삭제되어 더 이상 사용할 수 없게 됩니다.

우편 배달부 사용

1. {{apiurl}}/v2/lm/deployments/에 PATCH 요청을 제출하여 배포를 업데이트합니다.

 {{배포 ID}}.

2. 요청 본문에 새 구성 ID를 전달합니다.

The screenshot shows a Postman interface for a PATCH request. The URL is {{apiurl}}/v2/lm/deployments/d748fdae9f88a9b0. The Headers tab shows 10 items. The Body tab is selected and set to JSON. The raw JSON body is:

```

1 {
  "configurationId": "58c4f650-2494-4786-aa28-e2ab4136d545"
}

```

The response body is:

```

1 {
  "id": "d748fdae9f88a9b0",
  "message": "Deployment modification scheduled"
}

```

3. {{apiurl}}/v2/lm/deployments/에 GET 요청을 제출하여 배포 상태를 확인합니다.

 {{배포 ID}}.

GET {{apiurl}}/v2/lm/deployments/d748fdae9f88a9b0

Params Authorization Headers (8) Body Pre-request Script Tests Settings

Body Cookies Headers (5) Test Results

Pretty Raw Preview Visualize JSON

```

1
2   "configurationId": "58c4f650-2494-4786-aa28-e2ab4136d545",
3   "configurationName": "mnist-serving",
4   "createdAt": "2021-12-01T10:20:28Z",
5   "deploymentUrl": "https://{{apiurl}}/v2/inference/deployments/d748fdae9f88a9b0",
6   "id": "d748fdae9f88a9b0",
7   "lastOperation": "UPDATE",
8   "latestRunningConfigurationId": "58c4f650-2494-4786-aa28-e2ab4136d545",
9   "modifiedAt": "2021-12-01T10:48:36Z",
10  "scenarioId": "platform-test-mnist-tensorflow",
11  "startTime": "2021-12-01T10:47:23Z",
12  "status": "RUNNING",
13  "statusDetails": {
14    "deployment_info": [
15      {
16        "last_transition_time": "2021-12-01T10:46:15Z",
17        "message": null,
18        "status": "True",
19        "type": "IngressReady"
20      },

```

컬 사용하기

1. {{apiurl}}/v2/lm/deployments/에 PATCH 요청을 제출하여 배포를 업데이트합니다.

{{배포 ID}}.

2. 요청 본문에 새 구성 ID를 전달합니다.

```
컬 --request PATCH $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \ --header "권한 부여: Bearer $TOKEN" \ --header "ai-resource-group: $RESOURCE_GROUP" \ --header 'Content-Type: 애플리케이션/json' \ --data=raw '{
  "configurationId": "490a02b0-4b97-48e8-b905-1c6ae5ea5b1c"
}'
```

출력 코드

```
{
  "id": "d748fdae9f88a9b0", "message": "배포 수
정 예약됨"
}
```

3. {{apiurl}}/v2/lm/deployments/에 GET 요청을 제출하여 배포 상태를 확인합니다.

{{배포 ID}}.

```
컬 --요청 GET $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \ --header
"권한 부여: Bearer $TOKEN" \ --header "ai-resource-group: $RESOURCE_GROUP"
```

출력 코드

```
{
  "configurationId": "490a02b0-4b97-48e8-b905-1c6ae5ea5b1c", "configurationName": "mnist-serving",
  "createdAt": "2021-12-01T10:20:28Z", "deploymentUrl": "https://my-deployment-url.com",
  "id": "d748fdae9f88a9b0",
  "lastOperation": "UPDATE", "latestRunningConfigurationId": "490a02b0-4b97-48e8-
  b905-1c6ae5ea5b1c", "modifiedAt": "2021
  -12-01T11:19:21Z", "scenarioId": "플랫폼-테스트-
  mnist-tensorflow", "startTime": "2021-12-01T10:47:23Z", "status": "PENDING", "statusDetails": {}, "submissionTime":
  "2021-12-01T11:03:55Z", "targetStatus": "실행 중"
}
```

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 중지 \[페이지 183\]](#)

[배포 삭제 \[페이지 181\]](#)

[효율성 기능 \[페이지 189\]](#)

[배포 로그 검색 \[페이지 194\]](#)

7.3.7 배포 중지

배포를 중지하면 사용된 SAP AI Core 런타임 컴퓨팅 리소스가 해제됩니다. 배포가 중지되어도 비용이 발생하지 않습니다.

참고

중지는 상태가 실행 중이거나 보류 중인 경우에만 활성화됩니다.

[단일 배포 중지 \[페이지 182\]](#)

[다중 배포 중지 \[페이지 181\]](#)

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 삭제 \[페이지 181\]](#)

[효율성 기능 \[페이지 189\]](#)

[배포 로그 검색 \[페이지 194\]](#)

7.3.7.1 단일 배포 중지

우편 배달부 사용

`{{apiurl}}/v2/lm/deployments/{{deploymentid}}`에 PATCH 요청을 제출하여 배포를 중지합니다. 이 요청의 헤더는 AI-Resource-Group: {YOUR-Resource-Group}입니다. 이 요청의 본문은 다음과 같습니다.

The screenshot shows the Postman interface with the following details:

- Method:** PATCH
- URL:** `({{baseUrl}})/lm/deployments/d509aa095937801e`
- Headers:** (11) - This section is collapsed.
- Body:** (raw, JSON) - This section is selected.
- Body Content:**

```
{
  "targetStatus": "중지됨"
}
```
- Params:** none
- Authorization:** None
- Headers:** (11)
- Tests:** None
- Settings:** None
- Cookies:** None
- Beautify:** Enabled
- Response Preview:**

```
{
  "id": "d509aa095937801e",
  "message": "Deployment modification scheduled"
}
```

`{apiurl}/v2/lm/deployments/{deploymentid}`에 GET 요청을 제출하여 배포 상태를 확인하세요.

컬 사용하기

1. `{apiurl}/v2/lm/deployments/{deploymentid}`에 PATCH 요청을 제출하여 배포를 업데이트합니다.

2. 요청 본문을 다음과 같이 업데이트합니다.

```
컬 --request PATCH $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \
--header "authorization: Bearer $TOKEN" \
--header "ai-resource-group: $RESOURCE_GROUP" \
--header 'content-Type: application/json' \
--data-raw '{ "targetStatus": "STOPPED" }'
```

출력 코드

```
{
  "id": "d748fd9f88a9b0", "message": "배포 수
정 예약됨"
}
```

3. `{apiurl}/v2/lm/deployments/`에 GET 요청을 제출하여 배포 상태를 확인합니다.

`{deploymentID}`.

```
컬 --요청 GET $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \
--header "권한 부여: Bearer $TOKEN" \
--header "ai-resource-group: $RESOURCE_GROUP"
```

상위 주제: [배포 중지 \[페이지 183\]](#)

관련 정보

[다중 배포 중지 \[페이지 181\]](#)

7.3.7.2 다중 배포 중지

BulkUpdates는 AI API의 메타 기능 엔드포인트입니다. 대량 PATCH 작업을 활성화하거나 비활성화합니다. 자세한 내용은 [AI API 개요 \[페이지 24\]](#)를 참조하십시오.

이 기능은 기본적으로 `false`로 설정되어 있습니다. 대량 PATCH 작업을 활성화하려면 템플릿에 관련 값이 `true`로 설정된 다음 코드 조각이 포함되어야 합니다.

메타:

```

    "bulkUpdates": { "실행": false,
      "배포": false
    }
  
```

대량 업데이트 정보:

- 요청당 최대 업데이트 수는 100개입니다. • 대량 업데이트에는 STOP 및 DELETE 요청이 혼합되어 어 포함될 수 있습니다. • 실행 중이거나 보류 중인 실행이나 배포 만 중지할 수 있습니다. • 중지되었거나 중단되었거나 알 수 없는 실행 또는 배포 만 삭제할 수 있습니다.
- ID는 일괄 요청당 한 번만 나타날 수 있습니다. 동일한 ID에 대한 다중 수정의 경우 다중 요청 필요합니다.

컬 사용하기

요청 본문을 다음과 같이 업데이트합니다.

```

컬 --request PATCH - /deployments \ --header {"배포": [
  {
    "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨", {
      "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"
    }
  ]
}
  
```

출력 코드

```

{
  "배포": [
    {
      "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "배포 수정 예약됨", {
        "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "배포 수정 예약됨"
      }
    ]
  }
}
  
```

우편 배달부 사용

다음 주소로 대량 PATCH 요청 보내기: - /deployments

```
{"배포": [
```

```
{
  "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨", {

    "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"

  }
}

}'
```

출력 코드

```
{
  "배포": [
    {
      "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "배포 수정 예약됨", {

        "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "배포 수정 예약됨"

      }
    }
  }
}
```

상위 주제: [배포 중지](#) [페이지 183]

관련 정보

[단일 배포 중지](#) [페이지 182]

[AI API 개요](#) [페이지 24]

7.3.8 배포 삭제

배포를 삭제하면 사용된 SAP AI Core 리소스가 해제됩니다.

제한사항

배포가 실행 중인 경우 먼저 배포를 중지해야 합니다. {{apiurl}}/v2/lm/deployments/{{deploymentid}}에 PATCH 요청을 제출하여 배포를 중지할 수 있습니다. 자세한 내용은 [배포 중지](#) [페이지 183]를 참조하십시오.

[단일 배포 삭제](#) [페이지 182]

[여러 배포 삭제](#) [페이지 189]

상위 주제: [모델 사용](#) [페이지 151]

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 중지 \[페이지 183\]](#)

[효율성 기능 \[페이지 189\]](#)

[배포 로그 검색 \[페이지 194\]](#)

7.3.8.1 단일 배포 삭제

우편 배달부 사용

`{{apiurl}}/v2/lm/deployments/{{deploymentid}}`에 DELETE 요청을 제출하여 배포를 삭제합니다. 이 요청의 헤더는 `AI-Resource-Group: {YOUR-Resource-Group}`입니다.

`{{apiurl}}/v2/lm/deployments/{{deploymentid}}`에 GET 요청을 제출하여 배포 상태를 확인하세요.

컬 사용하기

1. `{{apiurl}}/v2/lm/deployments/`에 DELETE 요청을 제출하여 배포를 업데이트합니다.

`{{배포 ID}}.`

2. 요청 본문을 다음과 같이 업데이트합니다.

```
컬 --request DELETE $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \ --header "authorization: Bearer $TOKEN" \ --header "ai-resource-group: $RESOURCE_GROUP" \ --header 'content-Type: application/json' \ --data-raw '{ "targetStatus": "DELETED" }'
```

출력 코드

```
{
  "id": "d748fdae9f88a9b0", "message": "배포
  수정 예약됨"
}
```

3. {{apiurl}}/v2/lm/deployments에 GET 요청을 제출하여 배포 상태를 확인합니다.

{{배포 ID}}.

```
컬 --요청 GET $AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID \
--header
"권한 부여: Bearer $TOKEN" \
--header "ai-resource-group: $RESOURCE_GROUP"
```

상위 주제: [배포 삭제 \[페이지 187\]](#)

관련 정보

[여러 배포 삭제 \[페이지 189\]](#)

7.3.8.2 다중 배포 삭제

BulkUpdates는 AI API의 메타 기능 엔드포인트입니다. 대량 PATCH 작업을 활성화하거나 비활성화합니다. 자세한 내용은 [AI API 개요 \[페이지 24\]](#)를 참조하십시오.

이 기능은 기본적으로 false로 설정되어 있습니다. 대량 PATCH 작업을 활성화하려면 템플릿에 관련 값이 true로 설정된 다음 코드 조각이 포함되어야 합니다.

```
메타:
  "bulkUpdates": {
    "executions": false,
    "deployments": false
  }
```

대량 업데이트 정보:

- 요청당 최대 업데이트 수는 100개입니다.
- 대량 업데이트에는 STOP 및 DELETE 요청이 혼합되어 포함될 수 있습니다.
- 실행 중이거나 보류 중인 실행이나 배포만 중지할 수 있습니다. • 중지되었거나 중단되었거나 알 수 없는 실행 또는 배포만 삭제할 수 있습니다.
- ID는 일괄 요청당 한 번만 나타날 수 있습니다. 동일한 ID에 대한 다중 수정의 경우 다중 요청 필요합니다.

컬 사용하기

요청 본문을 다음과 같이 업데이트합니다.

```
컬 --패치 요청 - /배포 \
--header {"배포": [
  {
    "id": "aa97b177-9383-4934-8543-0f91a7a0283a",
    "targetStatus": "중지됨"
  }
]}
```

```
{
    "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"
}
]
}
```

출력 코드

```
{
    "배포": [
        {
            "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "배포 수정 예약됨"}, {

            "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "배포 수정 예약됨"

        }
    ]
}
```

우편 배달부 사용

다음 주소로 대량 PATCH 요청 보내기: - /deployments

요청 본문을 다음과 같이 업데이트합니다.

```
{
    "배포": [
        {
            "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "targetStatus": "중지됨"}, {

            "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "targetStatus": "삭제됨"
        }
    ]
}
```

출력 코드

```
{
    "배포": [
        {
            "id": "aa97b177-9383-4934-8543-0f91a7a0283a", "message": "배포 수정 예약됨"}, {

            "id": "qweq32131-qwee-1231-8543-0f91a7a2e2e", "message": "배포 수정 예약됨"

        }
    ]
}
```

상위 주제: [배포 삭제](#) [페이지 187]

관련 정보

[단일 배포 삭제 \[페이지 182\]](#)

[AI API 개요 \[페이지 24\]](#)

7.3.9 효율성 특징

효율성을 향상하고 리소스 소비 관리에 도움이 되는 SAP AI Core 렌타임의 기능을 알아보세요.

자동 확장

SAP AI Core에는 현재 소비량을 기준으로 사용되는 노드 수를 줄이거나 소비량이 많은 기간 동안 사용 제한을 적용하는 매개변수가 포함되어 있습니다. 이러한 매개변수를 사용하면 워크로드를 수요에 따라 유연하게 확장하고 소비량을 제한하여 소비량과 비용을 제한할 수 있습니다.

자세한 내용은 [템플릿 제공을 참조하세요.](#)

0으로 조정

불균일한 로드가 예상되는 경우 0으로 조정하면 수요가 허용될 때 노드가 절전 상태로 전환되어 소비가 제한되므로 비용이 발생합니다. 수요가 증가하면 노드가 깨어나므로 응답 시간이 늘어납니다. 글로벌 노드 풀은 이 콜드 스타트 시간을 줄입니다. 자세한 내용은 [템플릿 제공을 참조하세요.](#)

전역 노드 풀

추운 서버가 절전 상태에서 확장되면 응답 시간이 추가됩니다. 이를 줄이기 위해 SAP AI Core에는 일반적으로 사용되는 노드를 예약하여 응답 시간을 단축하는 글로벌 노드 풀이 있습니다. 전역 노드 풀을 사용하기 위해 아무것도 할 필요가 없습니다. 이미 마련되어 있습니다.

응답 시간을 더 줄이려면 자동 크기 조정 매개변수를 1로 설정하여 콜드 스타트를 완전히 방지하세요. 자세한 내용은 [템플릿 제공을 참조하세요.](#)

1로 확장

1로 조정하면 콜드 스타트를 완전히 방지할 수 있습니다. 이렇게 하면 필요하지 않은 경우에도 단일 노드를 웜 상태로 유지하여 응답 시간을 줄일 수 있습니다. 그러나 0으로 조정하는 것과 관련된 소비 및 비용 절감 효과는 제공하지 않습니다. 자세한 내용은 [템플릿 제공을 참조하세요.](#)

지속

기본 기간은 무기한이지만 ttl 매개변수는 배포 기간을 분, 시간 또는 일로 제한합니다. 이 매개변수를 사용하면 모델 서버 및 모델 배포 URL의 삭제를 계획하여 예상 사용 기간을 허용하고 이후 불필요한 소비와 비용을 피할 수 있습니다. 자세한 내용은 [모델 배포](#)를 참조하세요. [AI API 정보](#).

테넌트 월 노드 풀

테넌트 월 노드를 사용하면 테넌트가 특정 리소스 계획에서 노드를 예약하여 미리 정의된 수의 노드가 클라스터에 존재하도록 할 수 있습니다. 그런 다음 모델 학습 및 제공 중에 예약 노드를 사용할 수 있습니다. 노드를 예약하면 워크로드 시작 시 대기 시간이 줄어들지만 사용 여부에 관계없이 노드 사용에 따라 인라인 비용이 발생합니다.

노드 예약 메커니즘

1. 테넌트는 예약할 노드 수를 지정합니다.
 2. 실행 또는 배포는 예약 노드, 즉 동일한 리소스의 대체 예약 노드를 활용합니다.
- 계획 유형은 하이퍼스케일러에서 요청됩니다.

테넌트가 지정한 예약 노드 수는 지속적으로 사용할 수 있습니다.

최소 예약 노드 수는 0입니다.

노드 예약 정보

- 테넌트가 지정한 예약 노드 수만큼 지속적으로 사용할 수 있습니다.
- 최소 예약 노드 수는 0입니다.
- 최대 예약 노드 수는 10개입니다.
- 기본 예약 노드 수는 0입니다.

Postman을 사용한 예약 노드

1. 엔드포인트{{apiurl}}/v2/admin/resources/nodes에 PATCH 요청을 보냅니다. 2. 요청 본문에 예약할 리소스 계획 유형과 노드 수 람을 JSON 형식으로 제공합니다.

```
{
  "resourcePlans": [ {
    "이름": "infer.l", "요청": 1
  },
  {
    "이름": "infer.m", "요청": 1
  },
  {
    "이름": "train.l", "요청": 1
  }
  ...
} ] }
```

컬을 사용하는 예약 노드

엔드포인트{{apiurl}}/v2/admin/resources/nodes에 PATCH 요청을 제출합니다.

```
curl --request 퍼시 $AI_API_URL/v2/admin/resources/nodes \ --data-raw '{ "resourcePlans": [ {
```

```
        "이름": "infer.l", "요청": 1
    },
    {
        "이름": "infer.m", "요청": 1
    },
    {
        "이름": "train.l", "요청": 1
    }
]
```

```
}
```

기억하세요

모든 예약 노드에는 모델 학습 및 서빙 중에 사용된 노드와 동일한 요금이 청구됩니다.

Postman을 사용하여 예약 노드 상태 확인

엔드포인트{{apiurl}}/v2/admin/resources/nodes에 GET 요청 보내기

컬을 사용하여 예약 노드 상태 확인

```
curl --request GET $AI_API_URL/v2/resources/nodes
```

출력 코드

```
{
  "resourcePlans": { "infer.l": { "프
  로비저닝됨": 1, "요청됨":
  1

    },
    "infer.m": { "프로비저닝
    됨": 1, "요청됨": 1

    },
    "train.l": { "프로비저닝
    됨": 1, "요청됨": 1

    }
}
```

- 요청됨: 테넌트가 요청한 예약 노드 수입니다.
- 프로비저닝됨: 현재 클러스터에 존재하는 예약 노드 수입니다.

예약 노드 수량 업데이트

예약된 노드 수를 업데이트하려면 요청 필드에 수량을 업데이트하여 예약 절차를 반복하세요.

예약 노드 삭제

예약 노드를 삭제하려면 요청 필드의 수량을 0으로 설정하여 예약 절차를 반복하세요.

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 중지 \[페이지 183\]](#)

[배포 삭제 \[페이지 181\]](#)

[배포 로그 검색 \[페이지 194\]](#)

[서비스 계획 \[페이지 35\]](#)

7.3.10 배포 로그 검색

배포 및 실행 로그에서 액세스합니다.

GET 요청을 제출하여 특정 배포 또는 실행에 대한 로그를 검색할 수 있습니다. 로그를 검색하려면 다음 엔드포인트를 사용하세요.

- GET /v2/lm/deployments/{deploymentId}/logs
- GET /v2/lm/executions/{executionId}/logs

쿼리 매개변수는 다음과 같습니다.

- start : RFC 3339 호환 날짜/시간 형식의 문자열로 된 쿼리 시작 시간입니다. 기본값은 현재보다 1시간 전입니다. 예: 2021-05-19T00:00:14.347Z
end : RFC 3339 호환 날짜/시간 형식의 문자열로 된 쿼리의 종료 시간입니다. 기본값은 현재 시간입니다. 예: 2021-05-19T00:00:14.347Z

• \$top : 반환된 항목의 최대 수입니다. 기본값은 1000입니다. 상한은 5000입니다. • \$order : 로그의 정렬 순서입니다. asc(오름차순의 경우 가장 빠른 순서가 목록 상단에 표시됨) 또는 desc(내림차순의 경우 가장 최근 순서가 목록 상단에 표시됨) 목록의 맨 위). 기본값을 참고하세요

값은 오름차순입니다.

예를 들어:

- ./v2/lm/deployments/{deploymentId}/logs?
start=2021-05-19T00:00:14.347Z&end=2021-05-19T01:00:14.347Z&\$top=100&\$order=asc - 첫 번째 반환

2021-05-19T00:00:14.347Z와 2021-05-19T01:00:14.347Z 사이의 배포 로그 100줄

- /v2/lm/deployments/{deploymentId}/logs - 이전 시간의 배포 로그를 반환합니다.
- /v2/lm/executions/{executionId}/logs - 이전 시간의 실행 로그를 반환합니다.

우편 배달부 사용

1. 새 GET 요청을 생성하고 URL {{apiurl}}/v2/lm/deployments/{{deploymentid}}/를 입력합니다.
로그.

The screenshot shows the Postman interface with a GET request to `https://{{apiurl}}/v2/lm/deployments/{{deploymentid}}/logs?start=2021-09-24T10:59:31Z`. The Headers tab has a checked checkbox for 'AI-Resource-Group' with the value 'default'. The Body tab displays a JSON response with the following content:

```

1
2   "data": {
3     "result": [
4       {
5         "container": "storage-initializer",
6         "msg": "[I 211121 12:36:59 initializer-entrypoint:13] Initializing, args: src_uri [s3://
7             hcp-e597ff51-40f5-42c9-a75a-744281742e61/manji1234/e4fead0ec09e716a/text-model-tutorial] dest_path[ [/
8                 mnt/models]\n",
9         "pod": "d757b72bcd373305-predictor-default-nblx6-deployment-5c4dc46cfmw",
          "stream": "stderr",
          "timestamp": "2021-11-21T12:36:59.759396881+00:00"
        }
      ]
    }
  }

```

2. 권한 부여 탭에서 유형을 **Bearer Token**으로 설정합니다.
3. 토큰 값을 {{token}}으로 설정합니다.
4. 헤더 탭에서 다음 항목을 추가합니다.

열쇠	값
AI 리소스 그룹	<리소스 그룹 이름> (예제에서는 기본값이 사용됨)

5. 요청을 보냅니다.

컬 사용하기

```
컬 --request GET "$AI_API_URL/v2/lm/deployments/$DEPLOYMENT_ID/logs?
start=2021-05-19T00:00:14.347Z" --header "권한 부여: Bearer $TOKEN" --header "ai-resource-group: $RESOURCE_GROUP"
```

샘플 출력

예를 들어 API의 다음 JSON 출력을 참조하세요.

출력 코드

```
{
  "데이터": [
    {
      "결과": [
        {
          "컨테이너": "저장소 초기화 프로그램", "msg": "[I 210531 08:20:51 초기화 프로그램 진입점:13]
초기화 중, 인수: src_uri [gs://kserve-samples/models/tensorflow/flowers] dest_path[ [/mnt/models]\n", "pod": "tfs-dep-i543026-predictor-default-v6nf5-
배포-8b58c8ddcfpx", "스트림": "stderr", "타임스탬프": "2021-05-31T08:20:51.334+00:00"
      },
      {
        "container": "storage-initializer", "msg": "[I 210531 08:20:51 Storage:45] 내용 복사 중
gs://kserve-samples/models/tensorflow/flowers를 로컬로\n", "pod": "tfs-dep-i543026-predictor-default-v6nf5-
배포-8b58c8ddcfpx", "스트림": "stderr", "타임스탬프": "2021-05-31T08:20:51.335+00:00"
      },
      {
        "container": "storage-initializer", "msg": "[W 210531 08:20:51 _metadata:104] 3번 중 1번 시도에서 Compute Engine 메타데이터 서버를 사용할 수 없습니다. 이유: [Errno 111] 연결이 거부되었습니다.\n",
        "포드": "tfs-dep-i543026-predictor-default-v6nf5-
배포-8b58c8ddcfpx",
        "스트림": "stderr", "타임스탬프": "2021-05-31T08:20:51.338+00:00"
      },
      ...
    ]
  }
}
```

상위 주제: [모델 사용 \[페이지 151\]](#)

관련 정보

[리소스 계획 선택 \[페이지 151\]](#)

[템플릿 제공 \[페이지 153\]](#)

[실행 파일 목록 \[페이지 169\]](#)

[모델 배포 \[페이지 172\]](#)

[추론 \[페이지 173\]](#)

[배포 업데이트 \[페이지 179\]](#)

[배포 중지 \[페이지 183\]](#)

[배포 삭제 \[페이지 181\]](#)

[효율성 기능 \[페이지 189\]](#)

[실행 로그 검색 \[페이지 142\]](#)

8가지 지표

AI API는 지표를 추적하고 보고되는 지표를 사용자 정의하거나 필터링하는 기능을 제공합니다.

8.1 AI API를 사용한 지표 추적

AI API를 사용하여 실행 및 모델에 대한 측정항목을 추적하고 가져올 수 있습니다. 또한 SAP AI Launchpad 인터페이스를 사용하여 지표를 비교할 수 있습니다.

사용 승인

AI API를 사용하여 지표를 추적하려면 다음 역할 컬렉션이 필요합니다.

기능	역할 수집
측정항목 만들기	aicore_admin
	또는
	aicore_scenario_editor
측정항목 보기	aicore_viewer
	또는
	aicore_scenario_viewer

8.1.1 지표 가져오기

다음은 지표 정보를 수동으로 추적하고 지표를 패치 및 삭제하는 방법을 보여줍니다.

Postman을 사용하여 측정항목 가져오기

1. 다음 쿼리를 사용하여 엔드포인트 {{apiurl}}/v2/lm/metrics에 대한 GET 요청을 준비합니다.

매개변수 및 헤더:

쿼리 매개변수

매개변수	필수의	데이터 형식	설명
실행 ID	예	문자열 배열	실행할 실행 ID 측정항목 정보를 나열합니다.

\$select

아니요

문자열 배열

리소스만 프로젝트합니다.

에 지정된

\$select 매개변수 및 실행-
cutionId.

지원되는 값

\$select는 측정항목입니다.

태그 및 customInfo 또는
이들의 어떤 조합이라도
그리고 *.

값이 *이면 모든 지표가
리소스 데이터가 반환됩니다.

머리글

필드	필수의	데이터 형식	설명
AI-리소스-그룹 예		끈	리소스 그룹의 ID 실행 내용이 포함되어 있습니다.

2. 권한 부여 탭에서 유형을 Bearer Token으로 설정합니다.

3. 토큰 값을 {{token}}으로 설정합니다.

The screenshot shows the Postman interface with the 'Params' tab selected. Under the 'Authorization' section, the 'Type' dropdown is set to 'Bearer Token' and the 'Value' field contains the placeholder '{{token}}'. A tooltip message at the top right of the input field reads: 'Heads up! These parameters hold sensitive data. To keep this data secure while working in a collaborative environment, we recommend using variables. Learn more about variables'.

4. Params에서 쿼리 매개변수를 설정합니다.

1. Params
2. \$select
3. metrics

1. Params
2. executionIds
3. e3de6dbbca8621c5,ed4bb3c7a4ba64b9

5. 리소스 그룹 헤더를 설정합니다.

1. Headers
2. AI-Resource-Group
3. default

6. 요청을 보냅니다.

다음과 같은 응답을 받게 됩니다:

응답

JSON 필드

설명

맞춤 정보

사전의 배열. 각 사전은 사용자 정의를 포함합니다.
문자열의 정보.

실행 ID

측정항목이 속한 실행의 ID입니다.

측정항목

측정항목 사전의 배열입니다.

태그

Body Cookies Headers (3) Test Results

Status: 200 OK Time: 2.65 s Size: 1.78 KB Save Response

```

1
2   "count": 2,
3   "resources": [
4     {
5       "customInfo": [ - ],
6       "executionId": "e3de6dbbca8621c5",
7       "metrics": [ - ]
8     },
9     {
10       "customInfo": [ - ],
11       "executionId": "ed4bb3c7a4ba64b9",
12       "metrics": [ - ]
13     }
14   ]
15 
```

참고

\$select 매개변수에서 유효한 항목을
반점.

Params	Authorization	Headers (9)	Body	Pre-request Script	Tests	Settings	Cookies
Query Params							
KEY	VALUE				DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/> \$select	metrics,custominfo				Projects only the resources that are specified in \$select along with ...		
<input checked="" type="checkbox"/> executionIds	e3de6dbbc8621c5,ed4bb3c7a4ba64b9				ID of execution for which to list metrics information.		

컬을 사용하여 측정항목 가져오기

```
curl --location -g --request GET '$AI_API_URL/v2/lm/metrics?
$select=metrics,tags,customInfo&executionIds=e3de6dbbc8621c5,ed4bb3c7a4ba64b9' --header 'AI-リソース-グループ: 기본값' --header 'Bearer $TOKEN'
```

참고

허용된 값 중 하나를 사용하거나 허용된 값의 조합(메트릭, \$select 매개변수의 *, 태그,customInfo)을 사용할 수 있습니다.

출력 코드

```
{
  "개수": 2, "자원": [
    {
      "맞춤 정보": [
        {
          "name": "혼란 행렬", "value": "[{'Predicted': 'False', 'Actual': 'False','value': 34},{'예측': '거짓','실제': '참', '값': 124}, {'예측': '참','실제': '거짓 ', '값': 165},{'예측': '참','실제': '참', '값': 36}]"
        }
      ],
      "executionId": "ec83b8e837fe4a56", "metrics": [
        {
          "라벨": [
            {
              "이름": "m11", "값": "알파"
            },
            {
              "이름": "metrics.ai.sap.com/Artifact.name", "값": "텍스트 모델 튜토리얼"
            }
          ],
          "이름": "오류율", "단계": 2, "타임스탬프": "2020-09-29T11:40:10.330000Z", "값": 0.98
        },
        {
          "태그": [
            {
              "이름": "텍스트-모델-튜토리얼", "값": "베타-3"
            }
          ]
        }
      ]
    }
  ]
}
```

```

    },
    {
      "customInfo": [
        {
          "name": "훈
          란 행렬", "value": " 정밀 재현 f1 점수 지원\n0 0.85 0.97
          0.91 273\n1 0.96 0.79 0.87 227\nn 정확도 0.89 500\nn 매크로 평균 0.90 0.88 0.89 500\nn가중 평균 0.90 0.89 0.89 500\nn"
        }
      ],
      "executionId": "ea4af58c56c26384", "metrics": [
        {
          "라벨": [
            {
              "이름": "기
              차", "값": "범위 0-400"
            }
          ],
          "이름": "정확도", "단계": 1, "타임스탬프":
          "2021-12-03T06:23:57.945336Z", "값": 0.9675
        },
        {
          "라벨": [
            {
              "이름": "기
              차", "값": "범위 400-800"
            }
          ],
          "이름": "정확도", "단계": 2, "타임스탬프":
          "2021-12-03T06:23:58.089152Z", "값": 0.94
        },
        {
          "라벨": [
            {
              "이름": "기
              차", "값": "범위 800-1200"
            }
          ],
          "이름": "정확도", "단계": 3, "타임스탬프":
          "2021-12-03T06:23:58.195216Z", "값": 0.9625
        },
        {
          "라벨": [
            {
              "이름": "기
              차", "값": "범위
              1200-1600"
            }
          ],
          "이름": "정확도", "단계": 4, "타임스탬프":
          "2021-12-03T06:23:58.306945Z", "값": 0.9425
        },
        {
          "라벨": [
            {
              "이름": "기
              차", "값": "범위
              1600-2000"
            }
          ],
          "name": "정확도",
        }
      ]
    }
  }
}

```

```

    "단계": 5, "타임스탬프": "2021-12-03T06:23:58.421745Z", "값": 0.945
},
{
    "라벨": [
        {
            "이름": "metrics.ai.sap.com/Artifact.name", "값": "테스트 모델 튜토리얼"
        }
    ],
    "이름": "오류율", "단계": 0, "타임스탬프": "2021-12-03T06:23:58.569476Z", "값": 0.1099999999999999
},
{
    "레이블": [], "이름": "n_compiments", "단계": 0, "타임스탬프": "2021-12-03T06:23:57.576040Z", "값": 1173.0
},
{
    "레이블": [], "이름": "n_complaints", "단계": 0, "타임스탬프": "2021-12-03T06:23:57.575618Z", "값": 1327.0
},
{
    "레이블": [], "이름": "n_samples", "단계": 0, "타임스탬프": "2021-12-03T06:23:57.574852Z", "값": 2500.0
},
{
    "라벨": [
        {
            "이름": "기자", "값": "80%"
        },
        {
            "이름": "단계", "값": "전처리"
        }
    ],
    "이름": "split_samples", "단계": 0, "타임스탬프": "2021-12-03T06:23:57.757594Z", "값": 2000.0
},
{
    "라벨": [
        {
            "이름": "검증", "값": "20%"
        },
        {
            "이름": "단계", "값": "전처리"
        }
    ],
    "이름": "split_samples", "단계": 0, "타임스탬프": "2021-12-03T06:23:57.757608Z", "값": 500.0
}

```

```

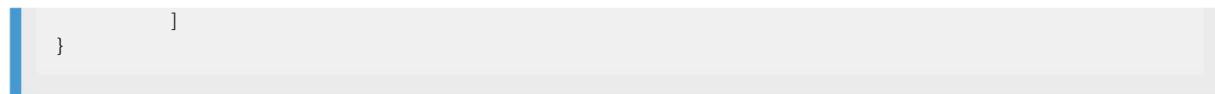
        },
        ],
        "태그": [
            {
                "name": "일기 모델", "value": "OK"
            },
            {
                "name": "테스트 추론", "value": "확인됨"
            }
        ]
    }
}

```

```
curl --location -g --request GET '$AI_API_URL/v2/lm/metrics? $select=tags,customInfo&executionIds=e3de6dbbc8a8621c5,ed4bb3c7a4ba64b9' \
--header 'AI-Resource-Group: default' \ --header 'Authorization: Bearer $TOKEN'
```

출력 코드

```
{
    "개수": 2, "자원": [
        {
            "customInfo": [
                {
                    "name": "혼란 행렬", "value": "[{'Predicted': 'False',
                    'Actual': 'False','value': 34},{'예측': '거짓','실제': '참', '값': 124}, {'예측': '참','실제': '거짓', '값': 165}, {'예측': '참','실제': '참', '값': 36}]"
                }
            ],
            "executionId": "ec83b8e837fe4a56", "태그": [
                {
                    "이름": "테스트-모델-튜토리얼", "값": "비타-3"
                }
            ]
        },
        {
            "customInfo": [
                {
                    "name": "혼란 행렬",
                    "값": "정밀 재현율 f1 점수 지원\n0 0.85 0.97 0.91 273\n1 0.96 0.79 0.87 227\nn 정확도 0.89 500\nn 막로 평균 0.90 0.88 0.89 500\nn 가중 평균 0.90 0.89 0.89 500\\ N"
                }
            ],
            "executionId": "ea4af58c56c26384", "태그": [
                {
                    "name": "일기 모델", "value": "OK"
                },
                {
                    "name": "테스트 추론", "value": "확인됨"
                }
            ]
        }
    ]
}
```



Postman을 사용한 패치 측정항목

1. 다음 헤더를 사용하여 엔드포인트 {{apiurl}}/v2/lm/metrics에 대한 PATCH 요청을 준비합니다.

머리글	필드	필수의	데이터 형식	설명
	AI-리소스-그룹 예		끈	리소스 그룹의 ID 실행 내용이 포함되어 있습니다.
	컨텐츠 타입	예	값 = 응용프로그램/ 병합 패치+json	

2. 권한 부여 탭에서 유형을 Bearer Token으로 설정합니다.

3. 토큰 값을 {{token}}으로 설정합니다.

The screenshot shows the Postman interface with the 'Authorization' tab selected. Under 'Type', 'Bearer Token' is chosen. In the 'Token' field, the placeholder {{token}} is highlighted with a red box. A tooltip message is visible above the token field: 'Heads up! These parameters hold sensitive data. To keep this data secure while working in a collaborative environment, we recommend using variables. Learn more about variables'.

4. 헤더를 설정하세요.

The screenshot shows the Postman interface with the 'Headers' tab selected. There are 11 headers listed. The table below shows the key-value pairs:

KEY	VALUE
AI-Resource-Group	default
Content-Type	application/merge-patch+json
Key	Value

5. 본문을 추가합니다.

The screenshot shows the Postman interface with the 'Body' tab selected. A sample code block is shown:

```
{
  "executionId": "{{executionid}}",
  "측정항목": [
    {
      "name": "오류율",
      "값": 0.98,
      "타임스탬프": "2021-06-28T07:50:24.589Z",
      "2 단계"
    }
  ]
}
```

```

    "라벨": [ {
        "이름": "그룹", "값": "트리-82"
    }
],
"태그": [ {
    "이름": "아티팩트 그룹", "값": "RFC-1"
}
],
"customInfo": [
    {
        "name": "훈련 행렬", "value": "[{'Predicted': 'False',
'Actual': 'False','value': 34},
{'예측': '거짓','실제': '참','값': 124}, {'예측':
'참','실제': '거짓','값': 165}, {'예측': '참','실제':
'참','값': 36}]]"
    }
]
}

```

PATCH ▼ {{apiurl}}/v2/lm/metrics

Params Authorization Headers (11) **Body** Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL JSON ▼

```

1   "executionId": "{{executionid}}",
2   "metrics": [
3     {
4       "name": "Error Rate",
5       "value": 0.98,
6       "timestamp": "2021-06-28T07:50:24.589Z",
7       "step": 2,
8       "labels": [
9         {
10            "name": "group",
11            "value": "tree-82"
12          }
13        ]
14      }
15    ],
16  ]

```

6. 요청을 보냅니다.

응답으로 204(콘텐츠 없음) 응답을 받게 됩니다.

컬을 사용한 패치 측정항목

```

curl --location --request GET '$AI_API_URL/v2/lm/metrics?ExecutionId=e1c49497ccf6dde8' \
--header 'AI-Resource-Group: default' \
--header 'Authorization: Bearer $TOKEN' \
--data-raw '{
  "executionId": "e1c49497ccf6dde8",
  "metrics": [
    {
      "label": "Error Rate",
      "value": 0.98,
      "timestamp": "2021-06-28T07:50:24.589Z"
    }
  ]
}'

```

```

"2 단계,
"라벨": [
{
"이름": "그룹",
"값": "나무-82"
}
],
"태그": [
{
"name": "아트팩트 그룹",
"값": "RFC-1"
}
],
"맞춤 정보": [
{
"name": "혼란 행렬",
"value": "[{\"예측됨\": \"거짓\", \"실제\": \"거짓\", \"값\": 34}, {\"예측\": \"거짓\", \"실제\": \"참\", \"값\": 124}, {\"예측\": \"참\", \"실제\": \"거짓\", \"값\": 165}, {\"예측\": \"참\", \"실제\": \"실제\", \"값\": 36}]"
}
]
}'
```

8.1.2 메트릭 데이터 쿼리

엔드포인트 /v2/lm/metrics에 GET 요청을 제출하여 지표를 쿼리할 수 있습니다. 추적을 가져오려면 데이터의 경우 다음 매개변수를 사용할 수 있습니다.

- AI-리소스-그룹(헤더) - 문자열 – UUID
- ExecutionIds(query) – string – 최대 10개의 실행 ID(쉼표로 구분)를 기반으로 지표를 검색합니다.
목록)
- \$select (query) – 메트릭 및 사용자 정의 정보와 같은 메트릭 리소스 데이터를 선택적으로 검색합니다. 을 위한 예를 들어, 매개변수 값이 와일드카드 (*)인 경우 모든 지표 리소스 데이터가 반환됩니다. 매개변수인 경우 값이 맞춤 정보 이면 맞춤 정보 데이터만 반환됩니다. 같은 배열로 입력해야 합니다. 오직 측정항목, 태그, customInfo 및 * 값이 지원됩니다.

예

GET API의 응답

이 예에서 \$select 매개변수 값은 와일드카드 (*)이므로 모든 지표 리소스 데이터가 반환됩니다.

응답 코드	설명
200	각 항목에 측정항목이 포함된 추적 메타데이터 목록 레이블, 매개변수 및 태그.

샘플 코드

```
{
  "개수": 1, "자원": [
    {
      "executionId": "aa97b177-9383-4934-8543-0f91a7a0283a", "metrics": [
        {
          "이름": "오류율", "값": 0.98, "타임스탬프":
          "2020-11-12T12:36:13.730Z",
          "단계": 2, "레이블": [
            {
              "이름": "그룹", "값": "트리-82"
            }
          ]
        },
        {
          "이름": "아티팩트 그룹", "값": "RFC-1"
        }
      ],
      "customInfo": [
        {
          "name": "혼란 행렬", "value": "[{'Predicted': 'False',
          'Actual': 'False','value':
          34},{'예측': '거짓','실제': '참', '값': 124}, {'예측':
          '참','실제': '거짓', '값': 165},{'예측': '참','실제':
          '참','값': 36}]"
        }
      ]
    }
  ]
}
```

응답 코드

400

설명

리소스 사양이 잘못되었습니다.

샘플 코드

```
{
  "error": { "code":
    "<error code>", "message": "<error message>.",
    "requestId": "9832bf934f3743v3948v3", "target": "/metrics",
    "details": [
      {
        "code": "<오류 코드>", "message": "<오류 메시지
        >."
      }
    ]
  }
}
```

응답 코드	설명
501	작업이 지원되지 않습니다.

샘플 코드

```
{
  "error": {
    "code": "02010055", "message": "측정항목을
    찾을 수 없습니다.", "requestId": "9832bf934f3743v3948v3", "target": "/metrics", "details": [
      {
        "code": "9827389374", "message": "빈
        결과 집합."
      }
    ]
  }
}
```

8.1.3 측정항목 데이터 저장

엔드포인트 `/v2/lm/metrics`에 PATCH 요청을 제출하여 메트릭을 저장할 수 있습니다. 추적 데이터를 저장하려면 AI-Resource-Group 헤더 문자열 매개변수를 제공해야 합니다.

ExecutionId, 지표 레이블, 태그 및 customInfo는 함께 추적 데이터를 유지하기 위해 PATCH API에 대한 요청 본문으로 사용되는 MetricResource를 형성합니다.

MetricResource의 각 속성에는 다음 제한이 적용됩니다.

- 측정항목 – 최대 1000
- 레이블 – 최대 20개, 각 레이블의 최대 문자 제한은 256자입니다.
- 태그 – 최대 100개(각 태그당 최대 문자 제한은 256자) customInfo – 최대 100개(전체 customInfo 객체의 최대 크기는 5MB)
-

추천

민감한 정보를 추적하기 위해 SAP AI Core에서 제공하는 추적 기능을 사용하지 마세요. 자세한 내용은 [보안 \[페이지 250\]](#)을 참조하십시오.

실행 ID

실행의 고유 식별자입니다.

측정항목

측정항목은 값이 숫자인 키/값 쌍입니다. 측정항목에는 선택적 단계, 타임스탬프, 라벨이 있을 수 있습니다.

필드. 측정항목은 모델 성능을 평가하기 위한 훈련 실행의 일부로 캡처됩니다. 모든 지표

(및 관련 라벨), 태그 및 사용자 정의 정보는 실행과 연결되어야 합니다. 측정항목, 태그 및 사용자 정의 정보가 저장되면 실행 ID로 쿼리할 수 있습니다.

측정항목 속성

속성 이름	설명	유형	제약
이름 (필수)	측정항목 이름	끈	패턴: [\w-]{1,64} 최대 길이: 256
값(필수)	메트릭 Number의 숫자 값		
타임스탬프	측정항목이 표시되는 시간 생성되었거나 로그인되었습니다. RFC3339 형식	문자열(\$날짜-시간)	
단계	단계는 선택적 정수입니다. 이는 모든 수단을 나타냅니다. 훈련 진행의 강화 (훈련 반복 횟수- 옵션, 시대 수 및 등) 측정항목의 경우	정수	최소: 0
라벨	이름-값 객체 목록 일부와 관련된 쌍 미터법.	레이블 속성 [페이지 210]	

라벨

레이블은 측정항목에 적용되는 분류 문구 또는 이름입니다. 라벨 세트를 사용하여 각 인스턴스를 제공할 수 있습니다.

추가 정보가 포함된 측정 기록입니다.

라벨 속성

속성 이름	설명	유형	제약
이름 (필수)	라벨 이름	끈	최소 길이: 1 최대 길이: 256
값(필수)	라벨의 가치	끈	최소 길이: 1 최대 길이: 256

측정항목을 모델과 연결하려면 이를 모델 결과물과 명시적으로 연결해야 합니다.

샘플 코드

```
"라벨": [
  {
    "이름": "그룹",
    "값": "나무-82"
  },
  {
    "이름": "metrics.ai.sap.com/Artifact.name",
    "값": "내_모델_이름"
  }
]
```

실행이 하나의 출력 모델 아티팩트만 생성하는 경우 실행에서 캡처된 모든 측정항목은 해당 출력 아티팩트와 연결됩니다. 실행으로 인해 더 이상의 출력 모델 아티팩트가 생성되는 경우 각 지표는 캡처된 항목은 위의 예에 표시된 대로 모델 아티팩트와 명시적으로 연결되어야 합니다.

태그

태그는 실행 수준에서 분리를 지원하는 데 사용되는 이름/값 쌍입니다. 예를 들어 다음을 할당할 수 있습니다.

선택한 테스트 실행 그룹에 태그를 지정합니다. 태그 세트는 MetricResource와 연결될 수 있습니다.

차례는 실행과 연결됩니다.

태그 속성

속성 이름	설명	유형	제약
이름 (필수)	태그 이름	끈	최소 길이: 1 최대 길이: 256
값(필수)	태그의 값	끈	최소 길이: 1 최대 길이: 256

맞춤 정보

일반적으로 다음과 관련된 대량의 메타데이터를 캡처할 수 있는 사용자 지정 정보 키/값 쌍

처럼. 사용자 정의 정보는 소비에 대한 지표에 관한 렌더링 또는 의미 정보를 제공합니다.

JSON 형식의 애플리케이션 또는 복잡한 지표.

이러한 사용자 정의 정보 개체 집합은 MetricResource와 연결될 수 있습니다.

CustomInfo 속성

속성 이름	설명	유형	제약
이름 (필수)	CustomInfo의 이름	끈	최소 길이: 1 최대 길이: 256
값(필수)	CustomInfo의 값	끈	최소 길이: 1 최대 길이: 256

예

PATCH API에 대한 요청 본문 예

```
{
  "executionId": "aa97b177-9383-4934-8543-0f91a7a0283a",
  "측정항목": [
    {
      "name": "오류율",
      "값": 0.98,
      "타임스탬프": "2020-11-12T12:18:01.539Z",
      "2 단계",
      "라벨": [
        {
          "이름": "그룹",
          "값": "나무-82"
        },
        {
          "이름": "metrics.ai.sap.com/Artifact.name",
          "값": "내_모델_이름"
        }
      ]
    },
    "태그": [
      {
        "name": "아티팩트 그룹",
        "값": "RFC-1"
      }
    ],
    "맞춤 정보": [
      {
        "name": "훈란 행렬",
        "값": "[{'예측': '거짓', '실제': '거짓', '값': 34}, {'예측': '거짓', '실제': '참', '값': 124}, {'예측': '참', '실제': '거짓', '값': 165}, {'예측': '참', '실제': '참', '값': 36}]"
      }
    ]
  }
}
```

PATCH API의 응답

응답 코드	설명
204	측정항목이 성공적으로 업데이트/생성되었습니다.
400	리소스 사양이 잘못되었습니다.

샘플 코드

```
{
  "오류": {
    "코드": "<오류 코드>",
    "message": "<오류 메시지>.",
    "requestId": "9832bf934f3743v3948v3",
    "대상": "/메트릭",
    "세부정보": [
      {
        "코드": "<오류 코드>",
        "message": "<오류 메시지>"
      }
    ]
  }
}
```

응답 코드

413

설명

요청 엔터티가 서버에서 정의한 제한보다 큽니다.

샘플 코드

```
{
  "오류": {
    "코드": "02000005",
    "message": "PayloadLimitException",
    "requestId": "9832bf934f3743v3948v3",
    "대상": "/메트릭",
    "세부정보": [
      {
        "코드": "02000005",
        "message": "PayloadLimitException"
      }
    ]
  }
}
```

8.1.4 지표 삭제

/v2/lm/metrics 앤드포인트에 DELETE 요청을 제출하여 메트릭을 삭제할 수 있습니다. 추적을 삭제하려면 데이터를 사용하려면 다음 매개변수를 제공해야 합니다.

- AI-리소스-그룹(헤더) – 문자열
- 실행 ID(쿼리)

쿼리 매개변수

매개변수	필수의	데이터 형식	설명
실행 ID	예	문자열 배열	실행 ID

머리글

필드	필수의	데이터 형식	설명
AI 리소스 그룹	예	끈	해당 리소스 그룹의 ID 실행 내용이 포함되어 있습니다.

DELETE API의 응답

응답 코드	설명
200	측정항목 리소스가 삭제되었습니다.
404	지정된 리소스를 찾을 수 없습니다.

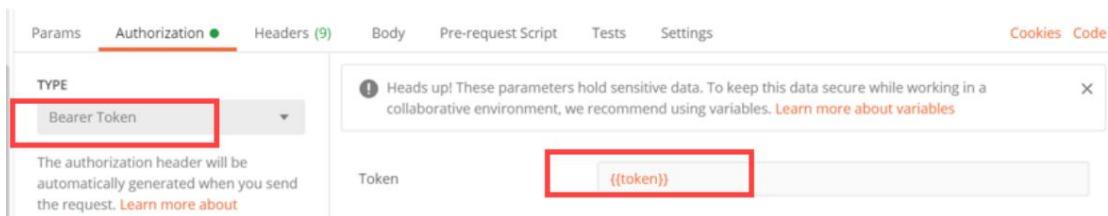
샘플 코드

샘플 404 응답:

```
{
  "error": {
    "code": "02010055",
    "message": "측정항목을
    찾을 수 없습니다.",
    "requestId": "9832bf934f3743v3948v3",
    "target": "/metrics",
    "details": [
      {
        "code": "9827389374",
        "message": "빈 결과 집합."
      }
    ]
  }
}
```

우편 배달부 사용

1. {{apiurl}}/v2/lm/metrics 엔드포인트에 DELETE 요청을 보냅니다.
2. 권한 부여 탭에서 유형을 **Bearer Token**으로 설정합니다.
3. 토큰 값을 {{token}}으로 설정합니다.



4. Params에서 쿼리 매개변수를 설정합니다.

DELETE {{apiurl}}/v2/lm/metrics?executionId={{executionid}}

Params Authorization Headers (8) Body Pre-request Script Tests Settings

Query Params

KEY	VALUE	DESCRIPTION
<input checked="" type="checkbox"/> executionId	{{{executionid}}}	
Key	Value	Description

5. 헤더를 설정하세요.

DELETE {{apiurl}}/v2/lm/metrics?executionId={{executionid}}

Params Authorization **Headers (8)** Body Pre-request Script Tests Settings

Headers (7 hidden)

KEY	VALUE
<input checked="" type="checkbox"/> AI-Resource-Group	default
Key	Value

6. 요청을 보냅니다.

Metric Resource가 성공적으로 삭제되었다는 메시지를 받아야 합니다.

Body	Cookies	Headers (5)	Test Results	Status: 200 OK	Time: 458 ms	Size: 225 B
<pre>Pretty Raw Preview Visualize JSON ↻</pre> <pre> 1 2 "message": "Metric Resource was successfully deleted" 3 </pre>						

컬 사용하기

필수 쿼리 매개변수 및 헤더를 사용하여 엔드포인트 {{apiurl}}/v2/lm/metrics에 DELETE 요청을 보냅니다.

```
컬 --location --request DELETE '$AI_API_URL/v2/lm/metrics? ExecutionId=e1c49497ccf6dde8' \ --header 'AI-Resource-Group: default' \ --header 'Authorization: Bearer $TOKEN'
```

9가지 고급 기능

서비스로서의 AI 콘텐츠 [페이지 216]

SAP AI Core는 사용자가 GitOps를 사용하여 [Service Marketplace](#)에서 AI 콘텐츠를 서비스로 제공할 수 있도록 지원합니다.

9.1 서비스로서의 AI 콘텐츠

SAP AI Core는 사용자가 GitOps를 사용하여 [Service Marketplace](#)에서 AI 콘텐츠를 서비스로 제공할 수 있도록 지원합니다.

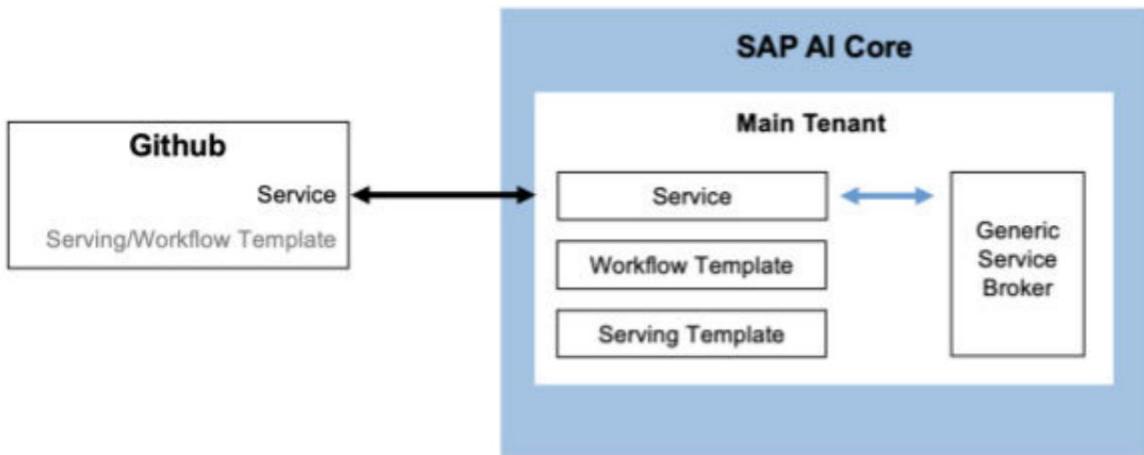
소비자는 SAP AI Core를 사용하여 서비스 인스턴스를 생성하여 사용 가능한 콘텐츠를 사용하고 자체 실행 또는 배포를 생성할 수 있습니다. 주 테넌트 역할을 하는 사람이 [서비스 마켓플레이스](#)를 통해 이를 외부에서 사용할 수 있도록 선택하면 서비스 공급자가 됩니다. 서비스 제공업체 및 해당 콘텐츠의 온보딩 프로세스는 다음과 같습니다.

1. 소비자가 사용할 수 있는 워크플로, 제공 템플릿 또는 Docker 레지스트리와 같은 AI 콘텐츠를 생성합니다.
2. 브로커 등록을 위한 일반 비밀을 생성합니다.
3. 등록된 git 저장소에 서비스 사용자 정의 리소스 YAML을 제공합니다.
4. SAP AI Core는 에서 사용할 수 있는 콘텐츠에 대한 서비스 브로커를 생성합니다. 서비스 브로커 최종 소비자의 온보딩 및 오프보딩을 처리합니다.
5. 엔드포인트 `{{apiurl}}/v2/admin/services`를 호출하여 서비스 브로커 정보를 가져옵니다.
6. SAP Cloud Management 서비스에 서비스 브로커를 등록합니다.
7. 소비자는 [서비스 마켓플레이스](#)에서 서비스 인스턴스를 생성합니다.
8. SAP AI Core는 `<resourceGroupId>`을 사용하여 소비자를 위한 리소스 그룹을 생성합니다.
`serviceInstanceId`.
9. 소비자는 서비스 키를 생성하고 서비스 사용을 시작합니다.

참고

서비스 제공자 주 테넌트는 1개의 서비스만 프로비저닝할 수 있습니다.

SAP AI Core는 다음과 같이 기본 테넌트에 대한 일반 서비스 브로커 인스턴스를 배포합니다.



상위 주제: [고급 기능](#) [페이지 216]

9.1.1 서비스 커스텀 리소스

서비스 공급자 기본 테넌트는 서비스 사용자 지정 리소스를 준비해야 합니다. 사용자 정의 리소스에는 서비스 세부 정보, 브로커 자격 증명 또는 비밀에 대한 참조, 서비스 소비자를 위해 구성된 기능이 포함되어 있습니다.

예제 서비스 사용자 정의 리소스는 다음 코드 블록에 제공됩니다.

apiVersion: ai.sap.com/v1alpha1 종류: 서비스 메타데이터: 이름: 샘플 서비스

특기:

브로커비밀:

이름: 브로커 자격 증명 사용자 이름KeyRef: 사용자 이름 비
밀번호KeyRef: 비밀번호

설명: 데모 기능에 사용되는 서비스: 기본: staticDeployments: true userDeployments:
true createExecutions: true

로그: 실행:

true 배포: true serviceCatalog: - 확

장Credentials:

공유됨:

서비스 URL:

AI_API_URL: https://api.ai.internalprod.eu-
Central-1.aws.ml.hana.ondemand.com

extendCatalog: 이름: 샘플 서비스
스 ID: 샘플 서비스 브로커 ID 설명: 샘플 서비스 바인
딩 가능: 실제 계획:

```

- id: 샘플-서비스-표준
설명: 샘플 서비스 표준 계획
이름: 표준
무료: 거짓
metadata:
  지원되는 플랫폼:
    - 클라우드파운드리
    - 쿠버네티스
    - sapbtp

```

이는 필요에 따라 값을 수정하여 지침으로 사용할 수 있습니다. YAML 설명자를 생성하려면 텍스트 편집기를 사용하세요.

YAML 플러그인을 사용합니다.

매개변수에 대한 자세한 내용은 다음 표를 참조하세요.

서비스 매개변수 세부정보

유형	매개변수	설명
metadata	이름	서비스 이름
브로커비밀	이름	다음을 포함하는 보안 비밀 이름 서비스 브로커를 등록하기 위한 자격 증명입니다.
참고		
<p>필수로 갖춰야 할 사항 이 비밀은 다음과 같이 등록되었습니다 일반적인 비밀.</p>		
사용자 이름KeyRef		사용자 이름에 대한 주요 참조 등록된 비밀에서.
비밀번호키참조		비밀번호에 대한 주요 참조 등록된 비밀에서.
능력	기초적인	소비자는 기존 제품을 사용할 수 있습니다. 배포는 허용되지 않음 생성, 업데이트 또는 삭제 (기본값: true).
	정적 배포	소비자는 (기본값: true).
	사용자배포	소비자는 (기본값: true).
	createExecutions	소비자는 실행 파일을 생성할 수 있습니다. 컷션(기본값: true).
로그	처형	소비자는 실행 파일에 액세스할 수 있습니다. 실행 로그.
	배포	소비자는 배포 로그에 액세스할 수 있습니다.

유형	매개변수	설명
	기본적으로	소비자는 다음에 액세스할 수 있습니다.
		<ul style="list-style-type: none"> • 아티팩트 생성, 읽기 및 구성 • 다운로드, 업로드 및 데이터 세트 삭제 • 생성, 읽기, 업데이트 및 객체 저장소 비밀번호 삭제
서비스 카탈로그	자격 증명 연장	<p>에 대해 언급하곤 했습니다. 요약된 서비스 URL입니다. 그것은 다시 서비스 키를 활용하세요.</p>
	확장 카탈로그	<p>서비스 카탈로그를 확장하는 데 사용됩니다. 동나무.</p>

제한사항

서비스 사용자 정의 리소스 업데이트는 지원되지 않습니다.

서비스 사용자 정의 리소스가 성공적으로 동기화되면 템플릿의 매개변수 업데이트가 더 이상 수행되지 않습니다.
어떤 효과를 얻습니다.

9.1.2 온보딩

서비스를 온보딩하려면 다음을 완료하세요.

1. 서비스 브로커 등록 시 자격 증명으로 사용할 BrokerSecret을 생성합니다.

1. URL {{apiurl}}/v2/admin/secrets에 대한 새로운 POST 요청을 사용합니다.

2. 요청 본문에 JSON 형식의 자격 증명(base64 인코딩)을 제공합니다.

```
{
  "name": "브로커 자격 증명",
  "데이터": {
    "사용자 이름": "bXktc2VjcmV0LWNyZWRlbnRpYWhw=",
    "비밀번호": "bXktc2VjcmV0LW90aGVyLWNyZWRlbnRpYWhw="
  }
}
```

3. 헤더에 요청 범위를 지정합니다. AI-Tenant-Scope: true

4. 요청을 보냅니다.

2. 다음 세부 정보를 사용하여 BrokerSecret 사양 섹션을 수정합니다.

```
api버전: ai.sap.com/v1alpha1
종류: 서비스
metadata:
  이름: 샘플 서비스
  사양:
    BrokerSecret:
      이름: 브로커 자격 증명
      사용자 이름KeyRef: 사용자 이름
      비밀번호KeyRef: 비밀번호
```

```
설명: 데모 기능에 사용되는 서비스: 기본: staticDeployments: true userDeployments:
true createExecutions: true
```

```
로그: 실행:
true 배포: true serviceCatalog: - 확장
장Credentials:
```

```
공유됨:
serviceUrls:
AI_SVC_URL: https://api.ai.internalprod.eu-
Central-1.aws.ml.hana.ondemand.com extendCatalog: 이름: 샘플 서비스 ID: 샘
플 서비스 브로커 ID 설명: 샘플 서비스 바인딩 가능: true 계획: - id: 샘플 서비스 표준 설정: 표준 계획 샘플 서비스 이름: 표준
```

```
무료: false 메타데이터: 지원되는 플랫폼:
cloudfoundry - kubernetes - sapbtp
```

참고

사용자 이름과 비밀번호는 이전 단계의 브로커 자격 증명의 키 이름입니다.

3. 서비스 키의 일부가 될 소비자에게 제공하려는 서비스 URL로 spec.serviceCatalog[].extendCredentials를 업데이트하십시오. spec.serviceCatalog[].extendCatalog 아래에 카탈로그 세부정보를 제공하세요.

4. 서비스 사용자 정의 리소스를 등록된 github 저장소에 푸시하고 동기화가 완료될 때까지 기다립니다.
성공적인.

5. 동기화가 완료되면 URL {{apiurl}}/v2/admin으로 GET 요청을 보내 서비스 세부 정보를 확인합니다.
서비스.

```
{
  "개수": 1 "지원": [
    {
      "name": "sample-service", "description": "데모에 사용되는 서비스", "status": "PROVISIONED", "url": "https://aif-xyzabc.servicebroker.internalprod.eu-central.amazonaws.ml.hana.ondemand.com"
    }
  ]
}
```

서비스 브로커 URL을 기록해 두세요.

6. smctl을 사용하여 서비스 브로커를 하위 계정 범위로 등록합니다.

1. 서비스 브로커를 전역적으로 등록하기 전에 먼저 하위 계정 범위로 등록을 테스트합니다.

하위 계정 범위는 서비스가 등록된 환경 카탈로그에 자동으로 표시된다는 의미입니다. Service Manager 가이드에 설명된 단계에 따라 [smctl을 설정할 수 있습니다.](#)

smctl이 설치되면 다음과 같이 로그인합니다.

```
# 환경 변수
SERVICE_MANAGER_URL=<sm url 예: https://service-manager.cfapps.sap.hana.ondemand.com/>
SVC_SUBACCOUNT_USER=<user-with-servicemanager-
role>SVC_SUBACCOUNT_PWD=<password + 2FA> SERVICE_BROKER_URL=https://aif-
xyzabc.servicebroker.internalprod.eu-central.aws.ml.hana.ondemand.com
SVC_SUBACCOUNT_SUBDOMAIN=<하위 계정 예: subaccountxyz> SERVICE_BROKER_USER=<비밀로 제공되는 브로커 사용자 이름> SERVICE_BROKER_PWD=<비
밀로 제공되는 브로커 비밀번호> # smctl 로그인 smctl 로그인 -a
$SERVICE_MANAGER_URL --param 하위 도메인=$SVC_SUBACCOUNT_SUBDOMAIN \ -u=$SVC_SUBACCOUNT_USER \
-p=$SVC_SUBACCOUNT_PWD
```

2. 브로커 이름, URL 및 자격 증명을 제공하여 서비스 브로커를 등록합니다.

```
# 서비스 브로커 등록 smctl 등록 브로커 샘플 서비스
$SERVICE_BROKER_URL -b
$SERVICE_BROKER_USER:$SERVICE_BROKER_PWD # 서비스 브로커 등록이 성
공적으로 완료되어야 합니다.
```

서비스 브로커가 성공적으로 등록되면 [서비스 마켓플레이스](#)에서 서비스를 사용할 수 있습니다.

```
# 클라우드에 로그인했다고 가정
올바른 하위 계정을 제공한 Foundry에서 가져오기
서비스 계획 정보 cf 마켓플레이스 -s 샘플 서비스
```

이제 소비자는 서비스 인스턴스와 서비스 키를 생성할 수 있습니다. 서비스 인스턴스 생성 시 SAP AI Core는 id = 인스턴스 ID로 해당 리소스 그룹을 생성하고 이제 서비스를 사용할 준비가 되었습니다.

9.1.3 오프보딩

실수로 인한 서비스 삭제를 방지하기 위해 서비스 제공자는 다음과 같은 삭제 전략을 제공해야 합니다.

```
metadata:
  이름: 샘플 서비스 주석:
    ai.sap.com/serviceDeletionStrategy: "삭제"
```

serviceDeletionStrategy 주석을 사용하면 서비스 제공자는 git 저장소에서 서비스 사용자 정의 리소스를 삭제하고 오프보딩을 진행할 수 있습니다. 성공적인 서비스 오프보팅을 위해서는 모든 소비자 서비스 인스턴스를 삭제해야 합니다.

10개의 라이브러리 및 SDK

SAP AI Core와 함께 사용할 추가 SDK 및 라이브러리를 살펴보세요.

SAP AI Core와 함께 사용 가능한 SDK

자원	설명	추가 정보
SAP AI 코어 SDK	<p>SAP AI Core SDK는 Python 기반입니다.</p> <p>SAP AI Core에 액세스할 수 있는 SDK</p> <p>Python 메소드와 데이터 구조를 사용하여 터스. 상호작용하는 데 사용할 수 있습니다.</p> <p>SAP AI Core는 모든 것에 대한 액세스를 제공합니다.</p> <p>공개 수명주기 및 관리 API.</p>	<ul style="list-style-type: none"> SAP AI 코어 SDK 비즈니스 액셀러레이터 허브
AI API 클라이언트 SDK	<p>AI API 클라이언트 SDK는 Python 기반입니다.</p> <p>AI에 접근할 수 있게 해주는 SDK</p> <p>Python 메소드와 데이터를 사용하는 API 구조.</p>	<ul style="list-style-type: none"> AI API 클라이언트 SDK AI API
SAP 생성 AI 허브 SDK	<p>이 SDK를 사용하면 다음을 활용할 수 있습니다.</p> <p>생성 모델의 힘</p> <p>SAP AI에서 사용할 수 있는 chatGPT와 같은 Core의 생성 AI 허브.</p>	생성적 AI 허브 SDK

SAP AI Core와 함께 사용 가능한 라이브러리

자원	설명	추가 정보
SAP AI용 Metaflow Python 라이브러리 핵심	<p>소프트웨어 개발을 다운로드할 수 있습니다.</p> <p>SAP AI Core용 옵션 키트(SDK)</p> <p>Python 패키지 색인(PyPI). 이것</p> <p>Python 라이브러리는 다음을 수행할 수 있는 플러그인입니다.</p> <p>다음을 수행할 수 있는 워크플로 템플릿을 생성합니다.</p> <p>SAP AI Core에서 실행</p> <p>Metaflow는 오픈 소스 도구입니다.</p> <p>데이터 사이언스의 생산성을 높이세요.</p> <p>티스트. Metaflow Python 라이브러리</p> <p>SAP AI Core는 Metaflow의 용량을 확장합니다.</p> <p>머신러닝 파이프를 통해</p> <p>Argo Workflows에서 라인을 실행할 수 있습니다.</p> <p>(SAP AI Core에서도 사용됨)</p>	<ul style="list-style-type: none"> PyPi 메타플로우 메타플로우 Argo 워크플로란 무엇입니까?

관련 정보

[AI API를 사용한 지표 추적 \[페이지 192\]](#)

11가지 콘텐츠 패키지

SAP AI Core와 함께 사용할 추가 콘텐츠 패키지를 살펴보세요.

SAP AI Core에서 사용 가능한 콘텐츠 패키지

자원	설명	추가 정보
데이터 로봇 패키지	SAP AI Core용 DataRobot 통합을 위한 콘텐츠 패키지입니다.	SAP AI Core용 데이터 로봇 패키지 
컴퓨터 비전 패키지	이미지용 콘텐츠 패키지 SAP AI Core의 사례에는 이미지 분류 및 기능 추출 이 추가되었으며 AI Core SDK와 함께 사용됩니다.	SAP AI용 컴퓨터 비전 패키지  핵심 

12 SAP AI Core의 생성적 AI 허브

12.1 생성 AI 허브의 모델 및 시나리오

생성적 AI 허브는 확장 서비스 계획을 통해서만 SAP AI Core에서 사용할 수 있습니다.에 대한 정보는 서비스 계획을 추가하거나 기존 서비스 계획을 업데이트하는 방법은 [SAP AI Core에서 서비스 계획 추가 \[페이지 52\]](#)를 참조하십시오. 및 [SAP AI Core에서 서비스 계획 업데이트 \[페이지 43\]](#).

시나리오

생성적 AI 모델에 대한 액세스는 글로벌 AI 시나리오 기반 모델에 따라 제공됩니다. SAP AI Core에서 관리합니다. 개별 모델은 서빙 템플릿 형태의 실행 파일로 제공되며, 원하는 모델에 해당하는 템플릿을 선택하여 액세스할 수 있습니다.

다음 시나리오를 사용할 수 있습니다.

글로벌 시나리오	실행 파일 ID	설명
기초 모델	azure-openai	Azure OpenAI 서비스는 다음에 대한 REST API 액세스를 제공합니다. OpenAI의 LLM.
기초 모델	aicore-오픈 소스	SAP AI를 통해 호스팅 및 액세스되는 오픈 소스 모델 핵심.
기초 모델	gcp-vertexai	GCP Vertex AI는 Google의 PaLM 2 및 Gemini 모델에 대한 액세스를 제공합니다.

모델

생성 AI 허브에서는 다음 모델을 사용할 수 있습니다.

참고

표시된 모델 및 관련 값은 변경될 수 있습니다.

토큰 전환율 및 사용 중단 날짜를 포함하여 이러한 모델에 대한 자세한 내용은 다음을 참조하세요.
SAP 노트 [3437766](#).

모델 및 해당 매개변수에 대한 자세한 내용은 모델 공급자의 설명서를 참조하세요.

실행 파일 ID	모델명	모델 버전	추가 정보
azure-openai	gpt-35-터보	• 0613 • 1106	Azure 채팅 완료
azure-openai	gpt-35-터보-16k	0613	Azure 채팅 완료
azure-openai	gpt-4	0613	Azure 채팅 완료
azure-openai	gpt-4-32k	0613	Azure 채팅 완료
azure-openai	텍스트 임베딩-ada-002	2	Azure 채팅 완료
aicore-오픈 소스	tiuae--falcon-40b- 지식하다	저것	Tiuae Falcon 40b 교육
gcp-vertexai	텍스트 들소	002	GCP 베텍스 AI
gcp-vertexai	채팅 들소	002	GCP 베텍스 AI
gcp-vertexai	텍스트베딩게코 003		GCP 베텍스 AI
gcp-vertexai	텍스트 삽입 게코 다국어	001	GCP 베텍스 AI
gcp-vertexai	gemini-1.0-pro	001	GCP 베텍스 AI

Azure OpenAI의 모델은 채팅 완료 API의 프라이빗 인스턴스를 통해 액세스됩니다. 그만큼 액세스 포인트는 공개적으로 액세스할 수 없으며 SAP AI Core를 통해 액세스됩니다.

자세한 내용은 [Azure OpenAI 채팅 완료 API](#)를 참조하세요.

오픈 소스 모델은 SAP AI Core에서 호스팅되며 OpenAI 호환 API 스키마를 통해 액세스할 수 있습니다.

SAP AI Core의 생성적 AI 하브에 대한 자세한 내용은 [SAP AI Core 설명서](#)를 참조하세요.

참고

다음 주제는 이 문서의 범위를 벗어납니다.

- 텍스트 지식 기반(예: 임베딩) 작업과 같은 고급 소비 패턴
- LLM 통화의 복잡한 조정
- 자체 모델 학습

관련 정보

[Azure 채팅 완료 설명서](#)

[Tiuae Falcon 40b 자침 문서](#)

[GCP Vertex AI 문서](#)

12.2 생성적 AI 모델을 위한 배포 생성

배포를 생성하여 생성적 AI 모델을 사용할 수 있게 만듭니다. 각 모델과 모델 버전, 그리고 생성 AI 허브와 함께 사용하려는 각 리소스 그룹에 대해 한 번만 수행할 수 있습니다. 생성된 배포 URL은 재사용할 수 있습니다.

전제조건

- SAP AI Core 서비스 인스턴스와 서비스 키가 있습니다. 자세한 내용은 [SAP AI Core 초기 설정 문서를 참조하세요.](#)
- 확장 서비스 요금제를 사용하고 있습니다. 자세한 내용은 [서비스 계획 \[페이지 35\]](#) 및 [서비스 계획 업데이트 \[페이지 43\]](#)를 참조하십시오.
- 선호하는 사용자 인터페이스에 대한 클라이언트 인증을 완료했습니다. 자세한 내용은 다음을 참조하세요.
[SAP AI Core에서 서비스 키를 사용하세요.](#)

문맥

배포를 생성하여 모델을 사용할 수 있게 만듭니다. 각 모델 및 모델 버전에 대해 한 번만 수행할 수 있습니다. 모델 배포에는 액세스하려는 모델의 modelName 및 버전이 포함됩니다.
배포가 완료되면 조직 전체에서 모델 버전에 액세스하는 데 사용할 수 있는 배포Url이 생성됩니다.

API 사용

절차

1. 배포할 LLM을 결정하고 다음 정보를 기록해 두십시오.

- 실행 파일 ID
- 모델명
- 모델 버전

참고

- 모델 버전을 지정하는 대신 '최신'을 사용하면 최신 버전의 모델이 사용됩니다.
[SAP AI Core에서 사용할 수 있습니다.](#)
- 모델 버전이 기재되지 않은 경우 해당되지 않습니다.

실행 파일 ID	모델명	모델 버전	추가 정보
azure-openai	gpt-35-ti보	• 0613 • 1106	Azure 채팅 완료
azure-openai	gpt-35-ti보-16k	0613	Azure 채팅 완료
azure-openai	gpt-4	0613	Azure 채팅 완료
azure-openai	gpt-4-32k	0613	Azure 채팅 완료
azure-openai	텍스트 삽입-에이다-002	2	Azure 채팅 완료
aicore-opensource tiiuae--	저것 팔콘-40b- 지시하다		Tiiuae Falcon 40b 교육
gcp-vertexai	텍스트 들소	002	GCP 버텍스 AI
gcp-vertexai	채팅 들소	002	GCP 버텍스 AI
gcp-vertexai	텍스트베딩 게코	003	GCP 버텍스 AI
gcp-vertexai	텍스트 삽입 게코 다국어	001	GCP 버텍스 AI
gcp-vertexai	gemini-1.0-pro	001	GCP 버텍스 AI

2. GET 요청을 보내 생성 AI가 포함된 시나리오에 액세스할 수 있는지 확인합니다.

`{{apiurl}}/v2/lm/scenarios.`

Bearer \$TOKEN을 사용하여 Authorization 헤더를 설정하고 리소스 그룹을 설정합니다.

참고

모든 생성 AI 활동에 동일한 리소스 그룹을 사용해야 합니다. 다른 리소스를 사용하려면 그룹의 경우 각 리소스 그룹에 대해 이러한 단계를 반복해야 합니다.

```

1  {
2    "count": 1,
3    "resources": [
4      {
5        "createdAt": "2023-09-23T08:19:06+00:00",
6        "description": "AI Core Global Scenario for LLM Access",
7        "id": "foundation-models",
8        "labels": [
9          {
10            "key": "scenarios.ai.sap.com/lm",
11            "value": "true"
12          }
13        ],
14        "modifiedAt": "2023-09-23T08:19:06+00:00",
15        "name": "foundation-models"
      }
    ]
  }
  
```

나열된 시나리오에는 ID가 기초 모델인 시나리오가 포함되어 있습니다.

3. 엔드포인트 {{apiurl}}/v2/lm/configuration에 POST 요청을 보내 구성을 생성합니다.

다음 매개변수를 전달하여 액세스를 제공하려는 모델의 세부정보를 포함합니다.

- name은 식별자를 자유롭게 선택하는 것입니다.
- runningId, modelName 및 modelVersion은 위의 표에 제공됩니다.
- 시나리오 ID는 기초 모델이어야 합니다.
- versionId는 자체 버전 참조입니다.

샘플 코드

```
{
  "name": "yourNameChoice", "executableId": "azure-openai", "scenarioId": "foundation-models", "versionId": "0.0.1", "parameterBindings": [
    {
      "key": "모델 이름", "value": "gpt-35-터보"
    },
    {
      "key": "모델 버전", "value": "0613"
    }
  ],
  "inputArtifactBindings": []
}
```

The screenshot shows the Postman interface with a POST request to `{{baseUrl}}/lm/configurations`. The request body is a JSON object:

```

6  i
7  "key": "modelName",
8  "value": "gpt-35-turbo"
9  },
10 [
11   "key": "modelVersion",
12   "value": "0613"
13 ],
14 ],
15 "inputArtifactBindings": []
16 }

```

The response tab shows a successful `201 Created` response with a message: `"id": "7b760c52-", "message": "Configuration created"`.

팁

SAP AI Core에서 사용 가능한 최신 모델 버전을 사용하려면 modelVersion에 최신 값을 지정할 수 있습니다.

응답으로 고유한 configurationId를 받습니다.

4. 엔드포인트 {{apiurl}}/v2/lm/deployments에 POST 요청을 보내 배포를 생성합니다.

요청에 이전 단계의 configurationId를 포함합니다.

샘플 코드

```
{ "configurationId": "yourConfigurationId" }
```

The screenshot shows the Postman interface with a POST request to `({{baseUrl}})/lm/deployments`. The request body is a JSON object containing `"configurationId": "7b760c52-"`. The response tab shows a status of 202 Accepted with a response body containing deployment details like `id: "d5106", message: "Deployment scheduled.", status: "UNKNOWN"`.

5. 엔드포인트 {{apiurl}}/v2/lm/에 GET 요청을 전송하여 배포 세부 정보를 검색합니다.

배포.

The screenshot shows the Postman interface with a GET request to `({{baseUrl}})/lm/deployments`. The response status is 200 OK with a response body containing a single deployment resource with `id: "d5106"`.

다음 단계

배포가 실행 중일 때 응답에 제공된 배포Url을 사용하여 모델에 액세스할 수 있습니다. 자세한 내용은 [생성적 AI 모델 사용 \[페이지 231\]](#)을 참조하십시오.

모델 수명주기

모델 버전에는 지원 중단 날짜가 있습니다. 배포에 모델 버전이 지정된 경우 해당 모델 버전의 지원 중단 날짜에 배포 작업이 중지됩니다.

다음 모델 업그레이드 옵션 중 하나를 구현합니다.

- 자동 업그레이드: 새로운 생성 AI 구성 및 배포를 생성하거나
새로운 구성, modelVersion 최신 지정. SAP AI Core에서 새 모델 버전이 지원되면 기존 생성 AI 배포는 해당 모델의 최신 버전을 자동으로 사용합니다.
- 수동 업그레이드: 선택한 대체 모델 버전으로 새로운 생성 AI 구성 생성하고 이를 사용하여 배포에 패치를 적용합니다. 이 모델 버전은 SAP AI Core에서 지원하는 모델 업데이트와 관계없이 생성적 AI 배포에 사용됩니다.

참고

modelVersion을 지정하지 않으면 기본적으로 최신 버전이 됩니다. 수동으로 업그레이드하려면 다음을 지정해야 합니다.
모델 버전.

12.3 생성적 AI 모델 사용

엔드포인트 {{deploymentUrl}}/chat/completions?api-version=2023-05-15에 요청을 보내 생성 AI 모델을 사용합니다. LLM은 질문에 답하고, 텍스트를 요약하고, 텍스트 본문에서 정보를 추출하는 등 자연어 관련 작업을 수행할 수 있습니다.

전제조건

- 생성 AI 모델에 대한 배포 URL이 있습니다. 자세한 내용은 [만들기를](#) 참조하세요.
[생성적 AI 모델 배포 \[페이지 227\]](#).

문맥

요청 본문에는 쿼리를 정의하는 메시지 매개변수가 포함되어야 합니다.

다음 헤더가 설정되어 있는지 확인하세요.

머리글	값
권한 부여	무기명 \$TOKEN
AI 리소스 그룹	활성화 단계에 사용되는 리소스 그룹

Falcon 모델을 사용하는 경우 모델 이름을 매개변수로 전달합니다. 예: "모델": "tiiuae--
팔콘-40b-자시".

다음과 같은 선택적 매개변수를 포함할 수도 있습니다.

모델	매개변수
하늘빛	<ul style="list-style-type: none"> max_tokens: 허용되는 최대 토큰 수를 정의하는 정수입니다. 생성된 답변입니다. 기본값은 4,096입니다. 온도: 0에서 2 사이의 숫자입니다. 값이 높을수록 출력이 더 무작위화됩니다. 값이 낮을수록 더 집중적이고 결정적입니다. 빈도_페널티: -2.0에서 2.0 사이의 숫자입니다. 양수 값은 지금까지 텍스트의 기준 빈도를 기반으로 새 토큰에 페널티를 적용하여 모델의 가능성을 줄입니다. 같은 줄을 그대로 반복합니다. 존재_페널티: -2.0에서 2.0 사이의 숫자입니다. 양수 값은 새 토큰에 불이익을 줍니다. 지금까지 텍스트에 등장했는지 여부에 따라 모델이 말할 가능성이 높아집니다. 새로운 주제에 대해. stop: 문자열 또는 배열입니다. 최대 4개의 시퀀스. 출력이 중지 값 중 하나를 생성하는 경우 콘텐츠 생성이 중지됩니다.

자세한 내용은 [Azure Chat 완료 설명서](#)를 참조하세요..

매	<ul style="list-style-type: none"> max_tokens: 허용되는 최대 토큰 수를 정의하는 정수입니다. 생성된 답변입니다. 기본값은 4,096입니다. 온도: 0에서 2 사이의 숫자입니다. 값이 높을수록 출력이 더 무작위화됩니다. 값이 낮을수록 더 집중적이고 결정적입니다. 빈도_페널티: -2.0에서 2.0 사이의 숫자입니다. 양수 값은 지금까지 텍스트의 기준 빈도를 기반으로 새 토큰에 페널티를 적용하여 모델의 가능성을 줄입니다. 같은 줄을 그대로 반복합니다. 존재_페널티: -2.0에서 2.0 사이의 숫자입니다. 양수 값은 새 토큰에 불이익을 줍니다. 지금까지 텍스트에 등장했는지 여부에 따라 모델이 말할 가능성이 높아집니다. 새로운 주제에 대해. stop: 문자열 또는 배열입니다. 최대 4개의 시퀀스. 출력이 중지 값 중 하나를 생성하는 경우 콘텐츠 생성이 중지됩니다.
---	--

자세한 내용은 [Tiiuae Falcon 40b 지침 문서](#)를 참조하세요..

GCP 버텍스 AI	자세한 내용은 GCP Vertex AI 문서 를 참조하세요..
------------	--

모델을 제거하려면 해당 배포를 삭제하세요. 자세한 내용은 [배포 삭제 \[페이지 187\]](#).

주의

SAP는 기본 제품의 입력 또는 출력에 포함된 콘텐츠의 품질에 대해 어떠한 책임도 지지 않습니다.
편견, 환각 또는 부정확성을 포함하되 이에 국한되지 않는 생성 AI 모델. 사용자는
내용을 확인할 책임이 있습니다.

주의

생성 AI 허브를 사용할 때 프롬프트에 민감한 데이터를 저장하지 마세요. 민감한 데이터는 민감하지 않은 모든 데이터입니다. 기밀 정보 또는 개인 정보를 포함하되 이에 국한되지 않는 공개 공개를 목적으로 합니다.

콘텐츠 필터링

선택한 Azure 모델에 대해 다음 범주 및 심각도에 대한 콘텐츠 필터링이 활성화되었습니다.

범주	낮은 심각도	중간 및 높은 심각도
싫어하다:	허용하다	차단하다
성적:	허용하다	차단하다
자해:	허용하다	차단하다
폭행:	허용하다	차단하다

범주	감지된 경우
탈옥 위험:	차단하다

자세한 내용은 [Azure 콘텐츠 필터링을 참조하세요.](#) .



내용 때문에 응답이 필터링된 경우 응답 대신 오류 메시지가 표시됩니다.

프롬프트 예

Curl에 대한 예제는 다음과 같습니다.

개방형 AI

```
curl --location '$DEPLOYMENT_URL/chat/completions?api-version=2023-05-15' \
--header 'AI-리소스-그룹: <리소스 그룹 ID>' \
--header '콘텐츠 유형: 애플리케이션/json' \
--header "승인: 전달자 $TOKEN" \
--데이터 '{
    "메시지": [
        {
            "역할": "사용자",
            "content": "샘플 입력 프롬프트"
        }
    ],
    "max_tokens": 100,
    "온도": 0.0,
    "주파수_페널티": 0,
    "presence_penalty": 0,
    "중지": "널"
}'
```

}

요약

LLM에 텍스트를 제공하고 요약을 요청할 수 있습니다.

절차

엔드포인트 {{deploymentUrl}}/chat/completions?api에 POST 요청을 보냅니다.

버전=2023-05-15.

쿼리를 본문에 포함하세요. 요약할 텍스트를 세 개의 백틱(`)으로 표시합니다.

참고

모델에 버전이 없으면 엔드포인트에 버전이 필요하지 않습니다.

예

이 예에서는 제품 리뷰 요약을 생성합니다. 요약에는 관련되지 않은 주제가 포함될 수 있습니다.

주요 주제로.

```
{
  "메시지": [
    {
      "역할": "사용자",
      "content": "귀하의 작업은 제품에 대한 간략한 요약을 생성하는 것입니다.  
전자상거래 사이트의 리뷰입니다. 아래 리뷰를 삼중 백틱으로 구분하여 최대 30단어로 요약하세요. 리뷰: ``내 딸의 생일을 위해 이 판다 플라시 장난감을 샀는데, 딸은 그것을 좋아하고 어디든 갖고 다닙니다. 부드럽고 매우 귀엽고, 얼굴이 친근해 보입니다. 그래도 내가 낸 금액에 비해 좀 적다. 같은 가격에 더 큰 옵션이 있을 수도 있다고 생각합니다. 예정보다 하루 일찍 도착해서 선물하기 전에 제가 직접 가지고 놀게 되었어요.
      ``"
    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "宜居"
}
```

The screenshot shows the Postman interface with a POST request to `https://api.`. The request body contains a JSON object with a single message. The response status is `200 OK`.

```

POST https://api.

{
  "messages": [
    {
      "role": "user",
      "content": "Your task is to generate a short summary of a product review from an ecommerce site. Summarize the review below, delimited by triple backticks, in at most 30 words. Review: ``Got this panda plush toy for my daughter's birthday, who loves it and takes it everywhere. It's soft and super cute, and its face has a friendly look. It's a bit small for what I paid though. I think there might be other options that are bigger for the same price. It arrived a day earlier than expected, so I got to play with it myself before I gave it to her.``"
    }
  ]
}
  
```

Body Cookies Headers (5) Test Results

Pretty Raw Preview Visualize JSON

Status: 200 OK Time: 5.50 s Size: 617 B Save as example

예

이 예에서 프롬프트는 유사하지만 의도된 피드백 수신자를 추가하여 개선되었습니다.

(구매부서) 및 요청사유(상품가격 결정을 위해)

```
{
  "메시지": [
    {
      "역할": "사용자",
      "content": "귀하의 임무는 전자상거래 사이트에서 제품 리뷰에 대한 간략한 요약을 생성하여 제품 가격 결정을 담당하는 가격 책정 부서에 피드백을 제공하는 것입니다. 세 개의 블록으로 구분하여 아래 리뷰를 요약하십시오. 대부분 30개 단어로 구성되며 가격 및 인자된 가치와 관련된 모든 측면에 중점을 둡니다. 검토: `이 펠더 인형 장난감을 좋아하고 어디든 가지고 다니는 딸의 생일을 위해 구입했습니다. 얼굴은 친근해 보인다."
    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "널"
}
```

그래도 내가 낸 금액에 비해 좀 적다. 같은 가격에 더 큰 옵션이 있을 수도 있다고 생각합니다. 예정보다 하루 일찍 도착해서 선물주기 전에 제가 직접 가지고 놀게 되었어요``"

```

    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "널"
}
```

```

POST /analyze
{
  "message": {
    "text": "Your task is to extract relevant information from a product review from an ecommerce site to give feedback to the Shipping department. From the review below, delimited by triple quotes extract the information relevant to shipping and delivery. Limit to 30 words. Review: -- Got this plush toy for my daughter's birthday, who loves it and uses it everyday. It's soft and super cute, and its face has a friendly look. It's a bit small for what I paid though. I think there might be other options that are bigger for the same price. It arrived a day earlier than expected, so I got to play with it myself before I gave it to her....",
    "max_tokens": 100,
    "frequency_penalty": 0.0,
    "presence_penalty": 0,
    "stop": null
  }
}
  
```

Body Cookies Headers (0) Test Results

JSON

```

{
  "tokens": [
    {
      "index": 0,
      "content": "The relevant information about shipping and delivery from the review is: 'It arrived a day earlier than expected.'",
      "role": "assistant"
    }
  ],
  "tokens_over": 100,
  "tokens_usage": 0.0,
  "tokens_total": 100
}
  
```

Status: 200 OK Time: 100 ms Size: 400 B Save as Example

예

이 예에서는 "요약" 대신 "추출"을 사용합니다.

```
{
  "message": [
    {
      "text": "역할": "사용자",
      "content": "당신의 임무는 다음에서 관련 정보를 추출하는 것입니다."
    }
  ]
}
```

전자상거래 사이트의 제품 리뷰를 통해 배송 부서에 피드백을 제공합니다. 아래 리뷰에서 세 개의 백틱으로 구분되어 배송 및 배송과 관련된 정보를 추출합니다. 30단어로 제한하세요. 리뷰: `내 딸의 생일을 위해 이 판다 플러시 장난감을 샀는데, 딸은 그것을 좋아하고 어디든 갖고 다닙니다. 부드럽고 매우 귀엽고, 얼굴이 친근해 보입니다.

그래도 내가 낸 금액에 비해 좀 적다. 같은 가격에 더 큰 옵션이 있을 수도 있다고 생각합니다. 예정보다 하루 일찍 도착해서 선물주기 전에 제가 직접 가지고 놀게 되었어요``"

```
,
  ],
  "max_tokens": 100,
  "운도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "널"
}
```

```

POST /analyze
{
  "message": {
    "text": "What is the sentiment of the following product review, which is delimited with triple backticks? Review text: `` Needed a nice lamp for my bedroom, and this one had additional storage and not too high of a price point. Not it fast. The string to our lamp broke during the transit and the company rapidly sent over a new one. I have been very happy with this purchase. ``",
    "max_tokens": 100,
    "frequency_penalty": 0,
    "presence_penalty": 0,
    "stop": null
  }
}
  
```

Body Cookies Headers (0) Test Results

JSON

```

{
  "tokens": [
    {
      "index": 0,
      "content": "The sentiment of the product review is positive.",
      "role": "assistant"
    }
  ],
  "tokens_over": 100,
  "tokens_usage": 0.0,
  "tokens_total": 100
}
  
```

Status: 200 OK Time: 102 ms Size: 400 B Save as Example

추론

추론은 주어진 텍스트의 정보를 사용하여 결론을 도출합니다.

절차

엔드포인트 {{deploymentUrl}}/chat/completions?api에 POST 요청을 보냅니다.

버전=2023-05-15.

쿼리를 본문에 포함하세요. 세 개의 백틱(`)을 사용하여 추론할 텍스트를 표시합니다.

예

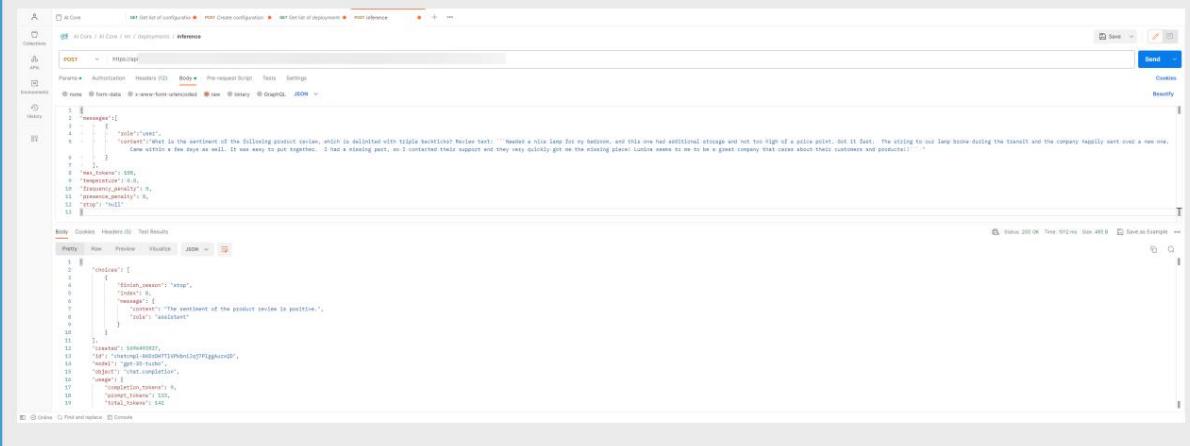
이 예에서는 제품 리뷰에 대한 감정 분석을 수행합니다.

```
{
  "messages": [
    {
      "role": "user",
      "content": "삼중 백틱으로 구분된 다음 제품 리뷰의 감정은 무엇입니까? 리뷰 텍스트: ``내 침실에 사용할 멋진 램프가 필요했는데 이 램프는 추가 저장 공간이 있고 가격대가 너무 높지 않습니다. 빨리 받았습니다. 운송 중에 램프의 끈이 끊어졌고 회사에서도 며칠 내에 새 끈을 보내 주었습니다."
    }
  ],
  "max_tokens": 100,
  "温度": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "증지": "글"
}
```

조립하기 쉬웠어요. 누락된 부품이 있어서 지원팀에 연락했더니 누락된 부품을 아주 빨리 받았습니다! 루미나는 고객과 제품에 관심을 갖는 훌륭한 회사인 것 같습니다!!````"

```

    }
  ],
  "max_tokens": 100,
  "temperature": 0.0,
  "top_p": 1.0,
  "frequency_penalty": 0,
  "presence_penalty": 0,
  "stop": "\n"
}
```



예

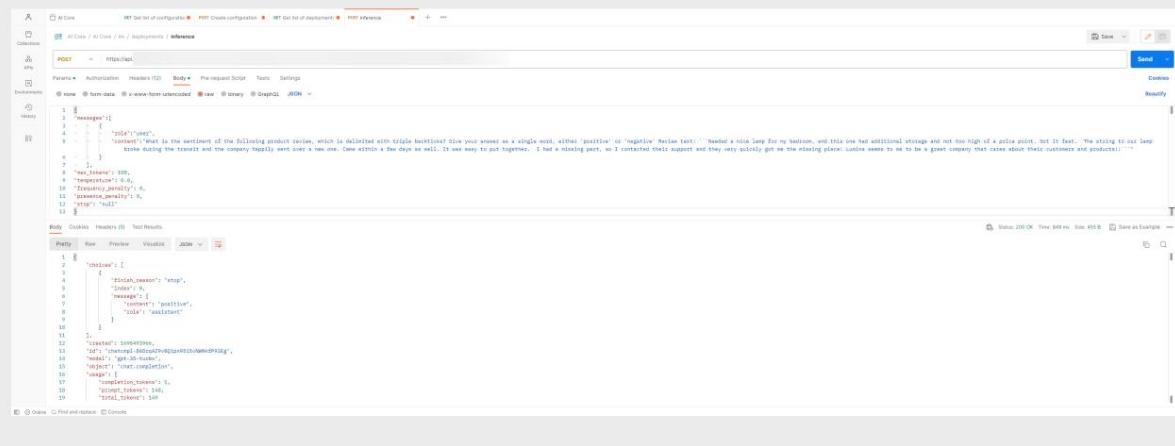
이 예에서는 한 단어 응답으로 감정을 생성합니다.

{

```

"메시지": [
    {
        "역할": "사용자",
        "content": "세 개의 백틱으로 구분된 다음 제품 리뷰의 감정은 무엇입니까? '긍정적' 또는 '부정적' 중 하나의 단어로 답변을 제공하십시오. 리뷰 텍스트: `` Needed a nice lamp for my 침실에 추가 저장 공간이 있고 가격도 그리 높지 않습니다. 배송 중에 램프 끈이 끊어졌는데 회사에서 며칠 안에 새 램프도 보내 주었습니다."
    }
],
"max_tokens": 100,
"온도": 0.0,
"주파수_페널티": 0,
"presence_penalty": 0,
"중지": "널"
}

```



예

이 예시는 리뷰에 표현된 감정을 분석합니다.

```

{
"메시지": [
    {
        "역할": "사용자",
        "content": "작가가 느낀 감정 목록을 식별합니다."
    }
]

```

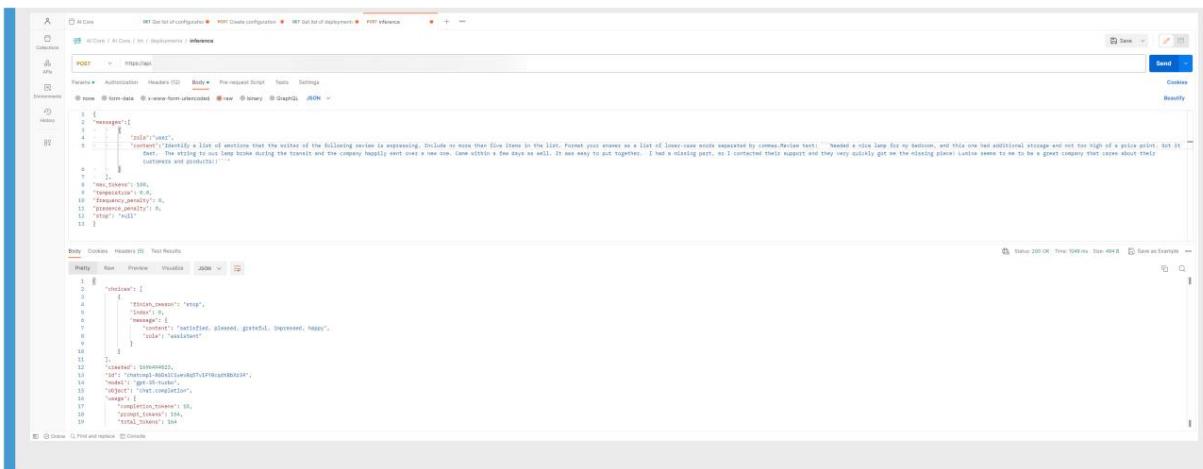
다음 리뷰는 표현하고 있습니다. 목록에 5개 이하의 항목을 포함하십시오.

답변의 형식은 쉼표로 구분된 소문자 단어 목록으로 구성하세요. 텍스트 검토: ``내 침실에 사용할 멋진 램프가 필요했는데 이 램프는 추가 저장 공간이 있고 가격도 그리 높지 않습니다. 빨리 받았습니다. 운송 중에 램프의 끈이 끊어졌고 회사에서는 기꺼이 새 끈을 보내주었습니다. 그것도 며칠 안에 오더군요. 조립하기 쉬웠어요. 누락된 부품이 있어서 지원팀에 연락했더니 누락된 부품을 아주 빨리 받았습니다! 루미나는 고객과 제품에 관심을 갖는 훌륭한 회사인 것 같습니다!!````"

```

    }
],
"max_tokens": 100,
"온도": 0.0,
"주파수_페널티": 0,
"presence_penalty": 0,
"중지": "널"
}

```



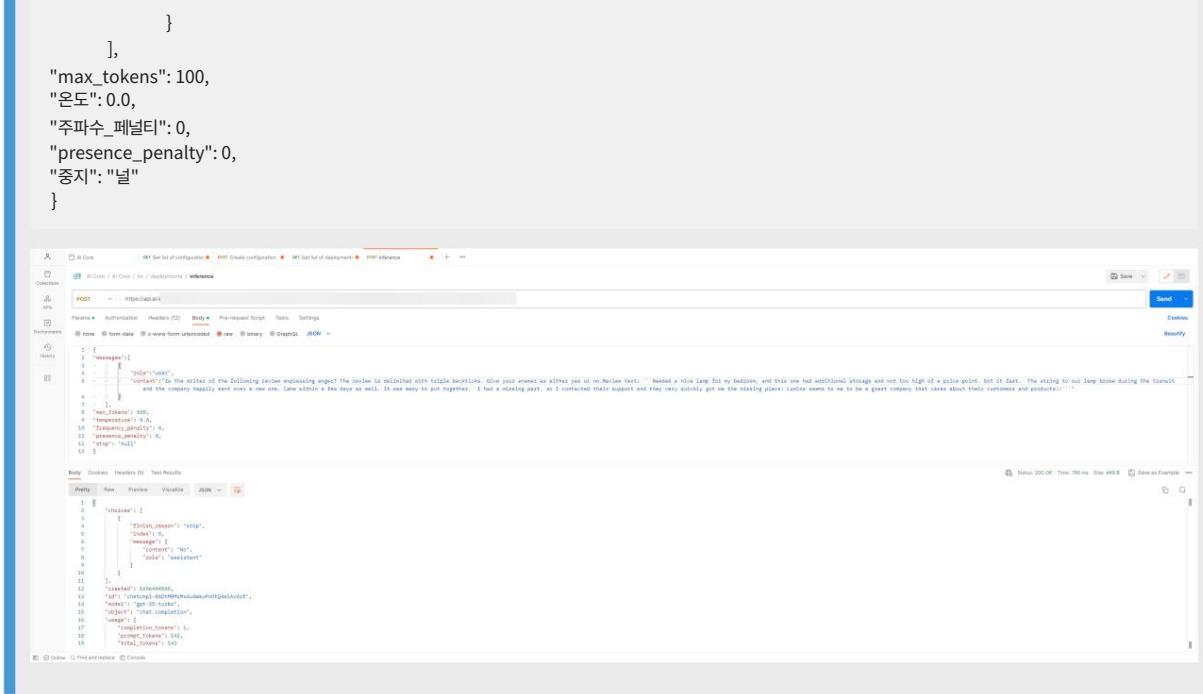
예

이 예에서는 분노가 있는지 여부를 감지합니다.

```
{
  "messages": [
    {
      "text": "역할":"사용자",
      "content": "다음 리뷰의 작성자가 분노를 표현하고 있습니까? 리뷰는 백틱 세 개로 구분됩니다. 예 또는 아니오로 대답하십시오. 리뷰 텍스트: ``내 침실에 멋진 램프가 필요했는데 이 램프는 추가 저장 공간과 가격이 너무 높지 않습니다."
    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "증지": "널"
}
```

운송 중에 램프의 끈이 끊어졌고 회사에서는 기꺼이 새 끈을 보내주었습니다. 그것도 며칠 안에 오더군요. 조립하기 쉬웠어요.

누락된 부품이 있어서 지원팀에 연락했더니 누락된 부품을 아주 빨리 받았습니다! 루미나는 고객과 제품에 관심을 갖는 훌륭한 회사인 것 같습니다!!````



예

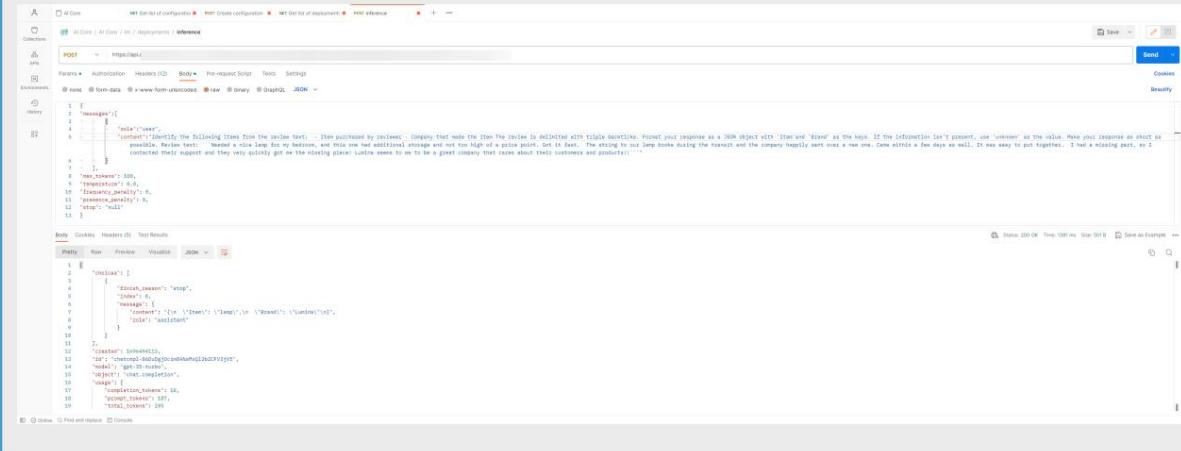
이 예에서는 고객 리뷰에서 제품 및 회사 이름을 감지합니다.

```
{
  "message": [
    {
      "role": "使用者",
      "content": "리뷰 텍스트에서 다음 항목을 확인하세요."
    }
  ]
}

- 리뷰어가 구매한 아이템 - 해당 아이템을 만든 회사 리뷰는 백틱 세 개로 구분됩니다. '항목'과 '브랜드'를 키로 사용하여 응답 형식을 JSON 개체로 지정하세요. 정보가 없으면 'unknown'을 값으로 사용합니다. 답변을 가능한 한 짧게 하세요. 리뷰 텍스트: ``내 침실에 사용할 멋진 램프가 필요했는데, 이 램프는 추가 저장 공간이 있고 가격도 그리 높지 않습니다. 빨리 받았습니다. 운송 중에 램프의 끈이 끊어졌고 회사에서는 기꺼이 새 끈을 보내주었습니다. 그것도 며칠 안에 오더군요. 조립하기 쉬웠어요. 누락된 부품이 있어서 지원팀에 연락했더니 누락된 부품을 아주 빨리 받았습니다!``
```

루미나는 고객과 제품에 관심을 갖는 훌륭한 회사인 것 같습니다!!``````

```
}
  ],
  "max_tokens": 100,
  "温度": 0.0,
  "主参数_佩尔蒂": 0,
  "presence_penalty": 0,
  "증지": "널"
}
```



예

이 예에서는 단일 쿼리에서 여러 작업을 수행합니다.

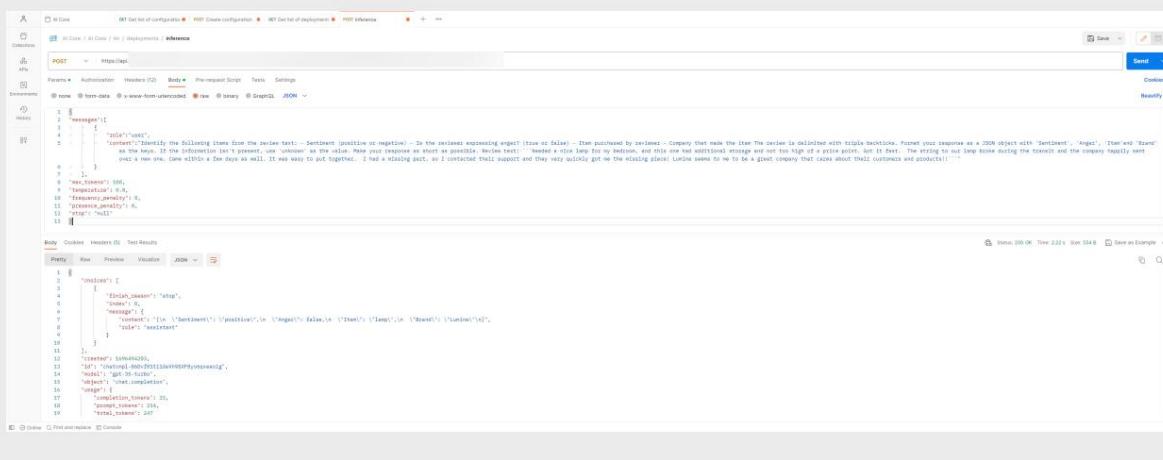
```
{
  "message": [
    {
      "role": "使用者",
      "content": "리뷰 텍스트에서 다음 항목을 확인하세요: -"
    }
  ]
}

감정(긍정적 또는 부정적) - 리뷰어가 분노를 표현하고 있습니까? (참 또는 거짓) - 리뷰어가 구매한 아이템 - 아이템을 만든 회사 리뷰는 세 개의 백틱으로 구분됩니다.
'감정', '분노', '항목', '브랜드'를 키로 사용하여 응답 형식을 JSON 개체로 지정하세요. 정보가 없으면 'unknown'을 값으로 사용합니다. 답변을 가능한 한 짧게 하세요. 리뷰 텍스트: ``내 침실에 사용할 멋진 램프가 필요했는데, 이 램프는 추가 저장 공간이 있고 가격도 그리 높지 않습니다. 빨리 받았습니다. 운송 중에 램프의 끈이 끊어졌고 회사에서는 기꺼이 새 끈을 보내주었습니다. 그것도 며칠 안에 오더군요. 조립하기 쉬웠어요. 누락된 부품이 있어서 지원팀에 연락했더니 누락된 부품을 아주 빨리 받았습니다!``
```

누락된 부품이 있어서 지원팀에 연락했더니 누락된 부품을 아주 빨리 받았습니다! 루미나는 고객과 제품에 관심을 갖는 훌륭한 회사인 것 같습니다!!^ ^ ^"

```

        },
        ],
        "max_tokens": 100,
        "온도": 0.0,
        "주파수_페널티": 0,
        "presence_penalty": 0,
        "중지": "널"
    }
}
```



예

이 예는 이야기에서 논의되는 다섯 가지 주제를 식별합니다.

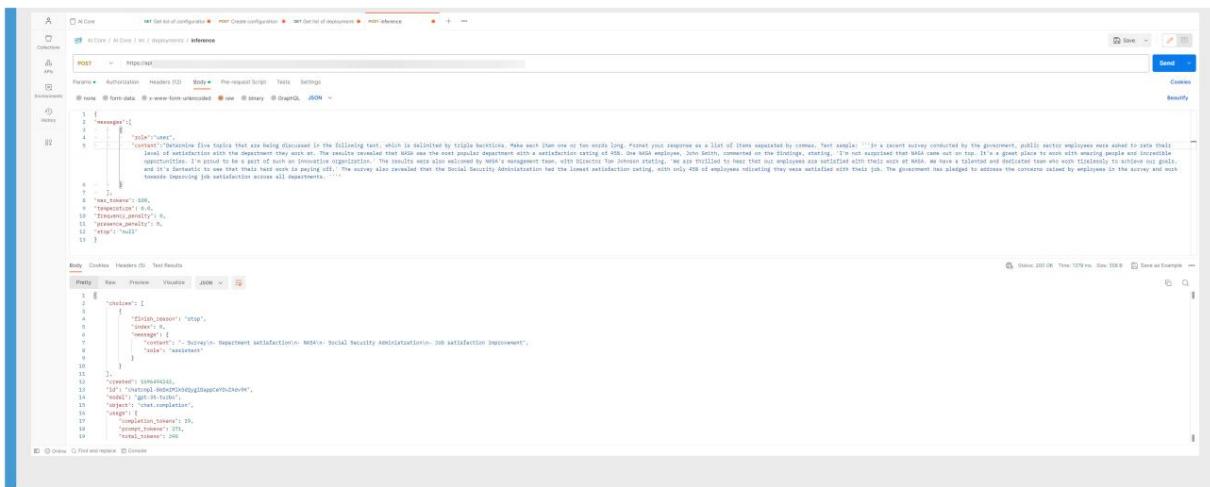
```
{
    "메시지": [
        {
            "역할": "사용자",
            "content": "다음 텍스트에서 논의되고 있는 5개의 주제를 결정하며 삼중 백틱으로 구분됩니다. 각 항목의 길이를 한두 단어로 만드세요."
        }
    ]
}
```

응답 형식을 쉼표로 구분된 항목 목록으로 지정하세요.

테스트 샘플: "'정부가 실시한 최근 설문 조사에서 공공 부문 직원들에게 자신이 근무하는 부서에 대한 만족도를 평가해 달라는 요청을 받았습니다. 그 결과 NASA가 만족도 95%로 가장 인기 있는 부서인 것으로 나타났다. NASA 직원 중 한 명인 John Smith는 연구 결과에 대해 'NASA가 1위를 차지한 것이 놀랍지 않습니다. 놀라운 사람들과 놀라운 기회와 함께 일할 수 있는 훌륭한 장소입니다. 나는 이런 혁신적인 조직의 일원이 된 것을 자랑스럽게 생각합니다.' NASA 경영진도 결과를 환영했으며, Tom Johnson 이사는 '우리 직원들이 NASA에서의 업무에 만족하고 있다는 소식을 듣고 매우 기쁩니다. 우리는 우리의 목표를 달성하기 위해 끊임 없이 노력하는 재능 있고 혁신적인 팀을 보유하고 있으며, 그들의 노력이 성과를 거두는 것을 보는 것은 환상적입니다.' 설문조사에 따르면 사회보장국은 만족도가 가장 낮은 것으로 나타났습니다. 직원 중 45%만이 자신의 직업에 만족한다고 답했습니다. 정부는 설문조사를 통해 직원들이 제기한 우려사항을 해결하고 모든 부서에서 직무 만족도를 향상시키기 위해 노력할 것을 약속했습니다.''''

```

        },
        ],
        "max_tokens": 100,
        "온도": 0.0,
        "주파수_페널티": 0,
        "presence_penalty": 0,
        "중지": "널"
    }
}
```



변환

변환은 주어진 텍스트를 다른 언어나 레이스터로 변환합니다.

절차

엔드포인트 {{deploymentUrl}}/chat/completions?api-0에 POST 요청을 보냅니다.
버전=2023-05-15.

쿼리를 본문에 포함하세요. 삼중 백틱(`)을 사용하여 변환할 텍스트를 표시합니다.

예

이 예는 텍스트를 영어에서 스페인어로 번역합니다.

```
{
  "messages": [
    {
      "role": "user",
      "content": "다음 영어 텍스트를 스페인어로 번역하세요:\n```\n`안녕하세요, 의사기를 주문하고 싶습니다``````\n}\n],\n"max_tokens": 100, "온도": 0.0,\n"주파수_페널티": 0, "존재_페널티": 0,\n"중지": "null" }
}
```

```

POST https://ai-core/api/inference
{
  "message": {
    "role": "user",
    "content": "Translate the following English text to Spanish: 'Hi, I would like to order a blender.'"
  },
  "max_tokens": 100,
  "temperature": 0.0,
  "top_p": 1.0,
  "presence_penalty": 0.0,
  "stop": null
}
  
```

Status 200 OK Time: 103 ms Size: 491 B Save as Example

예

이 예제에서는 텍스트가 작성된 언어를 감지합니다.

```
{
  "메시지": [
    {
      "role": "user",
      "content": "이 언어가 무엇인지 알려주세요: `` Combien coûte le 플로어 램프?`` `` "
    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "존재_페널티": 0,
  "중지": "null"
}
```

```

POST https://ai-core/api/inference
{
  "message": {
    "role": "user",
    "content": "Tell me which language this is: `` Combien coûte la lampe?`` "
  },
  "max_tokens": 100,
  "온도": 0.0,
  "presence_penalty": 0.0,
  "stop": null
}
  
```

Status 200 OK Time: 323 ms Size: 468 B Save as Example

예

이 예는 주어진 텍스트를 여러 언어로 번역합니다.

```
{
  "메시지": [
    {
      "role": "user",
      "content": "다음 텍스트를 프랑스어, 스페인어, 영어 해적으로 번역하세요: ``농구를 주문하고 싶습니다`` "
    }
  ]
}
```

```
        ],
    "max_tokens": 100, "온도": 0.0,
    "주파수_페널티": 0, "존재_페널티": 0,
    "중지": "null" }
```

The screenshot shows the SAP AI Core API interface. A POST request is being made to the endpoint `https://api.ai-core.com/api/inference`. The request body is a JSON object with the following structure:

```

{
  "message": {
    "role": "user",
    "content": "Translate the following text to French and Spanish and English please: \"I want to order a basketball\""
  },
  "max_tokens": 100,
  "temperature": 0.0,
  "frequency_penalty": 0,
  "presence_penalty": 0,
  "stop": null
}

```

The response body contains the translated text in three languages:

```

{
  "choices": [
    {
      "index": 0,
      "text": "Je veux commander un ballon de basket",
      "tokens": 40,
      "logprobs": 0.0,
      "finish_reason": "length",
      "model": "gpt-3.5-turbo",
      "object": "text_completion",
      "repetition_penalty": 1.0,
      "score": 1.0
    }
  ],
  "created": 1688615184,
  "error": null,
  "id": "AF000000000000000000000000000000",
  "index": 0,
  "logprobs": 0,
  "model": "gpt-3.5-turbo",
  "n": 1,
  "presence_penalty": 0,
  "score": 1.0,
  "size": 72,
  "usage": {
    "completion_tokens": 20,
    "prompt_tokens": 10,
    "total_tokens": 30
  }
}

```

예

이 예에서는 텍스트의 언어와 레지스터를 모두 번역합니다.

```
{ "메시지":{ [
    {
      "role": "user",
      "content": "다음 텍스트를 두 언어 모두에서 스페인어로 번역하세요."
    }
  ],
  "max_tokens": 100, "온도": 0.0,
  "주파수_페널티": 0, "존재_페널티": 0,
  "중지": "null" }
```

The screenshot shows the SAP AI Core API interface. A POST request is being made to the endpoint `https://api.ai-core.com/api/inference`. The request body is a JSON object with the following structure:

```

{
  "message": {
    "role": "user",
    "content": "Would you like to order a pillow?"
  },
  "max_tokens": 100,
  "temperature": 0.0,
  "frequency_penalty": 0,
  "presence_penalty": 0,
  "stop": null
}

```

The response body contains the translated text in three languages:

```

{
  "choices": [
    {
      "index": 0,
      "text": "Te gustaría ordenar una almohada?",
      "tokens": 20,
      "logprobs": 0.0,
      "finish_reason": "length",
      "model": "gpt-3.5-turbo",
      "object": "text_completion",
      "repetition_penalty": 1.0,
      "score": 1.0
    }
  ],
  "created": 1688615184,
  "error": null,
  "id": "AF000000000000000000000000000000",
  "index": 0,
  "logprobs": 0,
  "model": "gpt-3.5-turbo",
  "n": 1,
  "presence_penalty": 0,
  "score": 1.0,
  "size": 56,
  "usage": {
    "completion_tokens": 20,
    "prompt_tokens": 10,
    "total_tokens": 30
  }
}

```

예

이 예에서는 텍스트를 공식 레지스터로 변환합니다.

```
{
  "메시지": [
    {
      "role": "user", "content": " 다음을 속어에서 비즈니스 편지로 번역하세요: '야, 이 사람은 Joe야, 이 스텐드 램프의 사양을 확인해봐.'"
    }
  ],
  "max_tokens": 100, "온도": 0.0, "주파수_페널티": 0, "존재_페널티": 0, "중지": "null"
}
```

The screenshot shows the SAP AI Core interface with a POST request to the endpoint /api/v1/translate. The request body is a JSON object containing a 'message' array with one item. The item has a 'role' field set to 'user' and a 'content' field containing a Korean business letter. The response body shows the translated English text: "Translate the following from slang to a business letter: '야, 이 사람은 Joe야, 이 스텐드 램프의 사양을 확인해봐.'". The response status is 200 OK, and the time taken is 166 ms.

예

이 예는 출력 형식 간을 변환합니다. 프롬프트는 입력 및 출력 형식을 모두 설명합니다.

```
{
  "메시지": [
    {
      "role": "user", "content": " 다음 Python 사진을 JSON에서 열 헤더와 제목이 있는 HTML 테이블로 변환합니다. { \"resturant Employees\": {\"name\": \"Shyam\", \"email\": \"shyamjaiswal@gmail.com\"}, [ {\"name\": \"Bob\", \"email\": \"bob32@gmail.com\"}, {\"name\": \"Jai\", \"email\": \"jai87@gmail.com\"} ] }"
    }
  ],
  "max_tokens": 100, "온도": 0.0, "주파수_페널티": 0, "존재_페널티": 0, "중지": "null"
}
```

```

POST https://api.ai-core
{
  "messages": [
    {
      "role": "user",
      "content": "Translate the following python dictionary from JSON to an HTML table with column headers and titles: [ { \"restaurant_employee\": { \"name\": \"Yolanda\", \"email\": \"yolanda@saasigill.com\", \"new\": \"Yolanda\", \"email2\": \"yolanda@saasigill.com\", \"old\": \"Yolanda\", \"email3\": \"yolanda@saasigill.com\" } } ]"
    }
  ]
}
  
```

Status: 200 OK - Time: 3.04 s - Size: 798 B - Save as Example

예

이 예에서는 텍스트가 교정됩니다. 텍스트를 교정하고 수정할 수도 있고, 간단히 교정할 수도 있습니다.

```

{
  "메시지": [
    {
      "역할": "사용자",
      "content": "다음 텍스트를 교정하고 수정한 후 다시 작성하세요."
    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "없"
}
  
```

```

POST https://api.ai-core
{
  "messages": [
    {
      "role": "user",
      "content": "Find and correct the following text and rewrite the corrected version. If you don't find any errors, just say 'No errors found'. Don't use any punctuation around the text.: The girl with the black and white puppies have a bell."
    }
  ]
}
  
```

Status: 200 OK - Time: 1.006 ms - Size: 499 B - Save as Example

```

{
  "메시지": [
    {
      "역할": "사용자",
      "content": "다음 텍스트를 교정하고 수정한 후 다시 작성하세요."
    }
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "없"
}
  
```

수정된 버전. 오류가 발견되지 않으면 \"오류가 발견되지 않았습니다\"라고 말씀하세요. 텍스트 주위에 구두점을 사용하지 마세요: Yolanda has her 공책."

```

    "max_tokens": 100,
    "온도": 0.0,
    "주파수_페널티": 0,
    "presence_penalty": 0,
    "중지": "널"
}

```

```

POST https://api.ai-core.sap.com/v1/inference
{
  "messages": [
    {
      "role": "user",
      "content": "Please find and correct the following text and rewrite the corrected version. If you don't find any errors, just say: 'No errors found'. Don't use any punctuation around the text: 'Yolanda has two notebooks.'"
    }
  ],
  "max_tokens": 100,
  "temperature": 0.0,
  "presence_penalty": 0,
  "stop": null
}

```

Body Cookies Headers (S) Test Results

Status: 200 OK Time: 873 ms Date: 4/18/2024 Save as Example

```
{
  "메시지": [
    {
      "역할": "사용자",
      "content": "이 리뷰를 교정하고 수정합니다. ``내 딸이 내 방에서 내 것을 계속 가져가기 때문에 이것을 생일 선물로 받았습니다. 예, 어른들도 팬더를 좋아합니다. 그녀는 귀를 다른 것보다 약간 낮춥니다. 그리고 그것이 비대칭으로 설계되었다고 생각하지 않습니다. 하지만 동일한 가격에 더 큰 다른 옵션이 있을 수 있다고 생각합니다. 그래서 나는 그것을 지불했습니다. 딸에게 주기 전에 직접 가지고 놀아야겠어요````"
    }
  ]
}
```

```

    },
    "max_tokens": 100,
    "온도": 0.0,
    "주파수_페널티": 0,
    "presence_penalty": 0,
    "중지": "널"
}

```

```

POST https://api.ai-core.sap.com/v1/inference
{
  "messages": [
    {
      "role": "user",
      "content": "Please find and correct this review: ``Get this for my daughter for her birthday just because she keeps taking mine from my room. Yes, adults also like pandas too. One of the ears is a bit lower than the other, and I don't think that was designed to be asymmetrical. It's a bit small for what I paid for it though. I think there might be other options that are bigger for the same price. It arrived a day earlier than expected, so I got to play with it myself before I gave it to my daughter.````"
    }
  ],
  "max_tokens": 100,
  "temperature": 0.0,
  "presence_penalty": 0,
  "stop": null
}

```

Body Cookies Headers (S) Test Results

Status: 200 OK Time: 238 ms Date: 4/18/2024 Save as Example

확장

확장은 프롬프트를 기반으로 텍스트를 생성합니다.

절차

엔드포인트 {{deploymentUrl}}/chat/completions?api-에 POST 요청을 보냅니다.

버전=2023-05-15.

쿼리를 본문에 포함하세요.

예

이 예에서는 고객 이메일에 대한 자동 회신을 생성합니다.

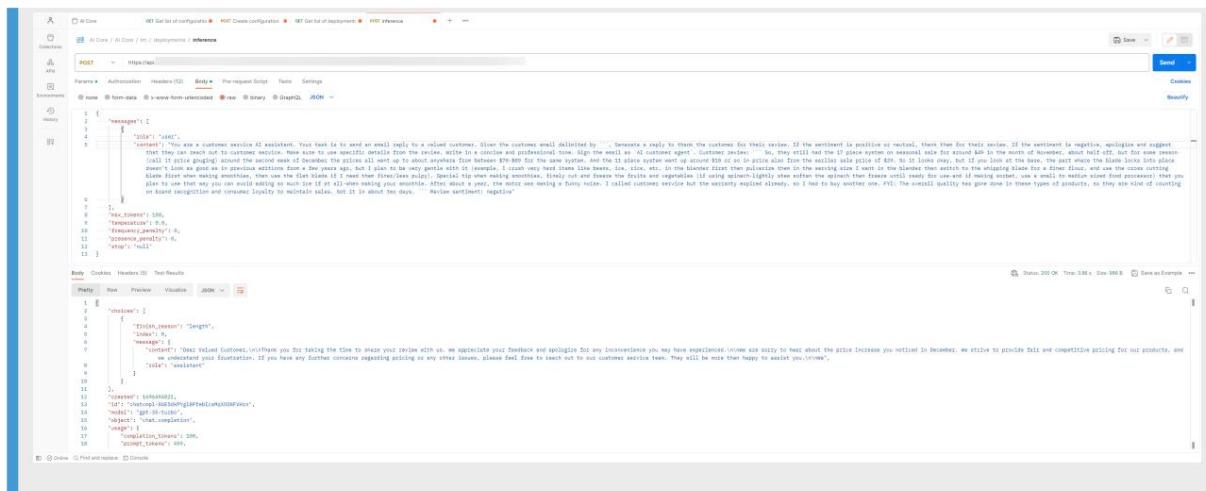
```
{
  "메시지": [
    {
      "역할": " 사용자",
      "content": "당신은 고객 서비스 AI 도우미입니다. 당신의 임무"
    }
  ]
}
```

소중한 고객에게 이메일 답장을 보내는 것입니다. `로 구분된 고객 이메일을 고려하여 고객의 리뷰에 대해 감사를 표하는 답장을 생성합니다.

감정이 긍정적이거나 중립적이라면 리뷰에 대해 감사를 표하세요. 감정이 부정적이라면 사과하고 고객 서비스에 연락할 수 있다고 제안하세요. 리뷰의 구체적인 세부정보를 활용하세요. 간결하고 전문적인 어조로 작성하세요. 이메일에 'AI 고객 에이전트'로 서명하세요.

고객 리뷰: `그래서 그들은 여전히 11월 한 달 동안 약 \$49에 시즌 세일로 17피스 시스템을 판매하고 있었는데, 이는 약 절반 할인된 가격이었습니다. 하지만 어떤 이유에서인지(가격 바가지라고 부르죠) 12월 둘째 주쯤에는 가격이 모두 동일한 시스템의 경우 \$70-\$89 정도까지 올라갔습니다. 그리고 11피스 시스템의 가격도 이전 판매 가격인 29달러에서 약 10달러 정도 올랐습니다. 그래서 괜찮아 보이는데, 베이스를 보면 칼날이 고정되는 부분이 몇년 전의 전작들만큼 좋지는 않은데, 아주 부드럽게 처리할 예정입니다. 콩, 얼음, 쌀 등과 같이 매우 딱딱한 것을 막사기에 먼저 분쇄한 다음 막사기에 넣고 원하는 크기로 분쇄한 다음 일가루를 더 곱게 만들려면 휨핑 칼날로 전환하고, 스무디를 만들 때는 신자 절단 칼날을 먼저 사용하십시오., 더 미세하고 덜 과육이 필요한 경우 플랫 블레이드를 사용하십시오. 스무디를 만들 때 특별한 텁, 사용할 과일과 야채를 절개 절라 냉동하세요(시금치를 살짝 끓여서 시금치를 부드럽게 만든 후 사용할 준비가 될 때까지 얼리고, 셰벗을 만들 경우 중소형 푸드 프로세서를 사용하세요). 그렇게 하면 스무디를 만들 때 너무 많은 얼음을 추가하는 것을 피할 수 있습니다. 1년 정도 지나자 모터에서 이상한 소리가 나기 시작했습니다. 고객센터에 전화했는데 이미 보증기간이 만료되어 하나 더 구입해야 했어요. 참고: 이러한 유형의 제품에서는 전반적인 품질이 향상되었으므로 판매를 유지하기 위해 브랜드 인지도와 소비자 충성도에 의존하고 있습니다. 이를 정도 만에 받았습니다. `리뷰 감정: 부정적`

```
,
  ],
  "max_tokens": 100,
  "온도": 0.0,
  "주파수_페널티": 0,
  "presence_penalty": 0,
  "중지": "널"
}
```



챗봇

문맥

챗봇은 입력을 사용하고 대화 형식으로 출력을 제공합니다.

절차

엔드포인트 {{deploymentUrl}}/chat/completions?api에 POST 요청을 보냅니다.

버전=2023-05-15.

쿼리를 본문에 포함하세요. 더 많은 맥락을 제공하거나 선례를 설정하려면 원하는 사례를 포함하세요.

프롬프트에 출력됩니다.

```
{
  "메시지": [
    {
      "role": "system", "content": "당신은 세이스피어처럼 말하는 어시스턴트입니다.", "prompt_tokens": 300,
      "usage": "당신은 세이스피어처럼 말하는 어시스턴트입니다."
    }
  ]
}
```

이 왜 길을 건넜나요? }, { "role": "user", "content": "모르겠어요"

The screenshot shows the SAP AI Core interface with the following details:

API Call: POST /inference

Body (JSON):

```

{
  "messages": [
    {
      "role": "system",
      "content": "You are an assistant that speaks like Shakespeare."
    },
    {
      "role": "user",
      "content": "Tell me a joke"
    },
    {
      "role": "assistant",
      "content": "Why did the chicken cross the road?"
    },
    {
      "role": "user",
      "content": "I don't know"
    }
  ],
  "max_tokens": 100,
  "temperature": 0.0,
  "presence_penalty": 0,
  "stop": null
}

```

Test Results (JSON):

```

{
  "messages": [
    {
      "role": "assistant",
      "content": "To cross the other side, but really, the other side was full of peril and danger, soooomth",
      "tokens": 100
    },
    {
      "role": "assistant",
      "content": "I'm sorry, I can't generate that many tokens right now.",
      "tokens": 100
    }
  ],
  "max_tokens": 100,
  "temperature": 0.0,
  "presence_penalty": 0,
  "stop": null
}

```

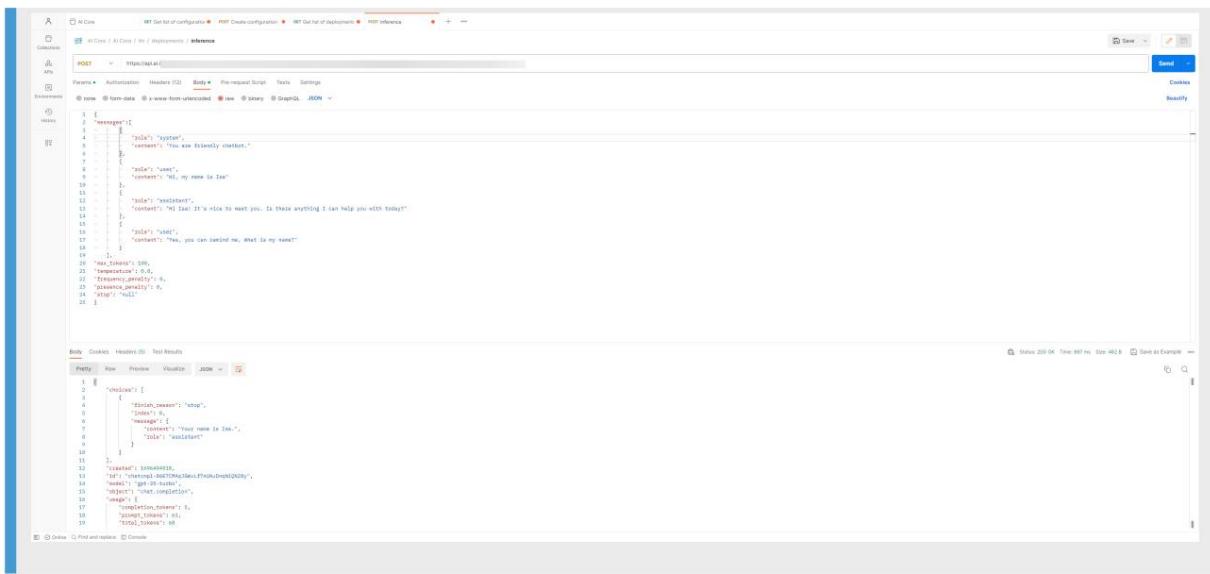
The screenshot shows the SAP AI Core API interface. A POST request is being made to `https://api.aiassistant.com`. The request body is a JSON object representing a conversation:

```

{
  "messages": [
    {
      "role": "system",
      "content": "You are friendly chatbot."
    },
    {
      "role": "user",
      "content": "Hi, my name is Isa."
    },
    {
      "role": "assistant",
      "content": "Hello Isa! It's nice to meet you. How can I assist you today?"
    },
    {
      "role": "system",
      "content": "stop",
      "index": 3,
      "tokens": 17,
      "content": "안녕하세요, 제 이름은 Isa입니다"
    },
    {
      "role": "user",
      "content": "안녕하세요. 오늘 제가 도와드릴 일이 있나요?"
    },
    {
      "role": "assistant",
      "content": "예, 제 이름이 무엇인지 알려주시면 됩니다."
    }
  ],
  "max_tokens": 100,
  "model": "get-in-the-loop-1000tokens",
  "presence_penalty": 0,
  "temperature": 0.0,
  "top_p": 1.0,
  "total_tokens": 28,
  "usage": 17
}

```

The response status is 200 OK, with a total time of 1008 ms and a size of 508 B.



```
POST https://api.ai-core.com/api/inference
```

```
Body Headers JSON
```

```
1 | { "messages": [
2 |   { "role": "system",
3 |     "content": "You are friendly chatbot."
4 |   },
5 |   { "role": "user",
6 |     "content": "Hi, my name is Ira"
7 |   },
8 |   { "role": "assistant",
9 |     "content": "Hello! It's nice to meet you. Is there anything I can help you with today?"
10 |   },
11 |   { "role": "user",
12 |     "content": "Yes, you can remind me. What is my name?"
13 |   },
14 |   { "role": "assistant",
15 |     "content": "Your name is Ira."
16 |   }
17 | } }
```

```
Body Headers JSON
```

```
1 | {
2 |   "id": "1624556584327",
3 |   "role": "assistant",
4 |   "content": "Your name is Ira.",
5 |   "tokens": 8,
6 |   "total_tokens": 8,
7 |   "total_time_ms": 0,
8 |   "temperature": 0.0,
9 |   "max_length": 100,
10 |   "presence_penalty": 0,
11 |   "frequency_penalty": 0,
12 |   "model": "text-davinci-003",
13 |   "n": 1,
14 |   "stop": null,
15 |   "logprobs": null,
16 |   "best_of": 1,
17 |   "max_retry": 0,
18 |   "seed": null,
19 |   "echo": false
20 | }
```

13가지 튜토리얼

SAP AI Core에 대해 사용 가능한 모든 튜토리얼입니다.

초보자 튜토리얼

빠른 시작:
SAP AI 코어  계정을 프로비저닝하고, 키를 등록하고, SAP AI Core SDK에 연결하고 운영하기 위한 도구를 설정하세요. 간단한 사용 사례에서 SAP AI Core를 사용하여 처음부터 끝까지 기본 사항을 알아보세요. 필요한 코드는 모두 자습서 단계에서 직접 제공되며 YAML 파일은 GitHub에서 추가로 사용할 수 있습니다.

빠른 시작:
SAP AI 코어  **VS 코드**  계정을 프로비저닝하고, 키를 등록하고, SAP AI Core SDK에 연결하고 운영하기 위한 도구를 설정하세요. 간단한 사용 사례에서 SAP AI Core를 사용하여 처음부터 끝까지 기본 사항을 알아보세요. 필요한 코드는 모두 자습서 단계에서 직접 제공되며 YAML 파일은 GitHub에서 추가로 사용할 수 있습니다. 이 튜토리얼은 VS Code를 사용하는 시작 튜토리얼입니다.

측정항목  지표를 로깅하고 모델을 SAP AI Core와 비교하는 다양한 방법을 살펴보세요.

고급 튜토리얼

SAP Developer Center에서는 기본 튜토리얼 외에도 다음 튜토리얼을 사용할 수 있습니다.

SAP AI Core용 컴퓨터 비전 패키지  SAP AI Core가 포함된 컴퓨터 비전 패키지를 살펴보세요.

사전 훈련된 Tensorflow 모델  SAP AI Core의 GPU와 함께 사전 훈련된 Tensorflow 모델을 사용하세요.

생성적 AI 허브 튜토리얼

Generative AI Hub 스타터 튜토리얼  생성 AI 허브를 활성화하고 프롬프트 작성의 기본 사항을 알아보세요.

14 보안

여기에서는 SAP AI Core의 보안 측면 중 일부를 설명하겠습니다.

14.1 데이터, 데이터 흐름 및 프로세스의 보안 기능

아래 표는 SAP AI Core의 데이터 흐름 개요를 보여줍니다.

데이터 및 보안 조치 개요

단계	설명	보안 조치
1	전송 제어/통신 보안	암호화된(HTTPS) 통신. 전송 중인 데이터는 최첨단 TLS 설정을 사용하여 암호화됩니다.
2	지속성 있는 애플리케이션 데이터 총	SAP에서 생성하고 유지 관리하는 최첨단 암호화 키를 사용하여 저장 시 암호화됩니다.
3	지속성 있는 애플리케이션 데이터 총	백업 및 복원 기능이 구현 및 테스트되었습니다. 정기적으로 백업은 원격 위치에 저장되고 백업은 최첨단 암호화를 통해 저장 중에 암호화됩니다. SAP가 생성하고 유지 관리하는 키.
4	목적에 따른 접근 통제 및 분리 포즈	액세스 제어를 구현하는 데 역할과 범위를 사용할 수 있습니다. 고객 관리자는 표준 BTP 보안 관리 기능을 사용하여 사용자에게 역할을 할당하여 "최소한의 보안"을 보장할 수 있습니다. 특권"과 "직무의 분리"이다.

14.2 전송 중 암호화

데이터 업로드 및 다운로드를 포함한 서비스와의 통신은 전송 계층을 사용하여 암호화됩니다.

보안(TLS) 프로토콜. SAP 서비스는 최신 프로토콜 버전(즉, TLS v1.2 이상)만 지원하며

강력한 암호 제품군. 보안을 설정하려면 시스템에서 지원되는 프로토콜 버전과 암호화 제품군을 사용해야 합니다.

서비스와의 통신. 또한 서비스의 도메인 이름에 대해 인증서의 유효성을 검사해야 합니다.

중간자 공격을 피하기 위해.

14.3 사용자 인증 및 관리

SAP AI Core는 SAP Authorization에서 제공하는 사용자 관리 및 인증 메커니즘을 사용하며 신뢰 관리 서비스(XSUAA).

SAP BTP의 XSUAA(SAP Authorization and Trust Management 서비스)에 대한 자세한 내용은 SAP Authorization and Trust Management Service 란 무엇입니까 ? 를 참조하세요.

참고

SAP AI Core의 데이터 보안 조항은 SAP AI Core 툴킷에도 적용됩니다.

SAP AI Core 확장은 SAP AI Core 엔진의 연결 세부 정보를 유지합니다. 연결 세부 정보는 사용자가 Visual Studio Code를 실행하고 SAP AI Core 툴킷 확장을 설치하는 시스템의 애플리케이션 메모리에 암호화된 형식으로 유지됩니다.

사용자 인증

SAP AI Core의 인증은 JSON 웹 토큰(JWT)을 기반으로 합니다.

명령된 사용자 통신의 경우 비밀번호 부여 OAuth2.0 헤더를 사용하여 JWT 토큰을 얻습니다. 사용자의 JWT 토큰에는 할당된 역할 또는 역할 모음에 의해 정의된 범위가 포함됩니다([사용자 역할 \[페이지 255\] 참조](#)).

소비 애플리케이션에는 사용자 이름과 비밀번호를 허용하고 사용자 토큰을 얻는 애플리케이션 라우터 또는 REST 게이트웨이가 있어야 합니다.

기술적인 사용자 통신을 위해 클라이언트 자격 증명 OAuth 2.0 헤더를 사용하여 JWT 토큰을 얻습니다.

사용자 역할

SAP AI Core는 SAP Authorization and Trust Management 서비스(XSUAA)에서 사용 가능한 역할 템플릿과 역할 컬렉션을 제공합니다. 이러한 템플릿과 컬렉션은 시나리오, 배포, 실행과 같은 SAP AI Core 기능 및 리소스에 대한 액세스를 관리하고 제어하는 데 도움이 됩니다.

역할 템플릿은 범위, 참조 및 선택적 특성 참조의 모음입니다. 역할 컬렉션은 역할 템플릿 집합입니다. 보안 관리자는 특정 사용자 또는 그룹에 역할 모음을 할당할 수 있습니다.

사용자.

참고

사용자를 생성하고 역할을 할당하려면 보안 관리자 권한이 있어야 합니다.

SAP AI Core는 [역할 및 권한 \[페이지 250\]](#)에 설명된 사용자 역할을 지원합니다 .

14.3.1 역할과 권한

SAP AI Core는 사용자에게 할당할 수 있는 기본 역할 컬렉션을 제공합니다. 역할 컬렉션에 따라 다음이 결정됩니다.
사용자가 SAP AI Core에서 수행할 수 있는 작업입니다. 자신만의 역할 컬렉션을 만들고 할당할 수도 있습니다.
그들에게 필요한 역할.

기본 역할

SAP AI Core는 다음과 같은 기본 역할을 제공합니다.

기본 역할

역할	설명
aicore_resourcegroup_viewer	SAP AI Core 테넌트의 리소스 그룹 보기
aicore_resourcegroup_editor 보기	SAP AI Core 테넌트의 리소스 그룹 생성 및 삭제
aicore_repository_viewer	SAP AI Core 테넌트의 GitOps 리포지토리 보기
aicore_repository_admin 보기	SAP AI의 GitOps 리포지토리 생성, 업데이트 및 삭제 핵심 테넌트
aicore_application_viewer	SAP AI Core 테넌트의 GitOps 애플리케이션 보기
aicore_application_admin 보기	SAP AI의 GitOps 애플리케이션 생성, 업데이트 및 삭제 핵심 테넌트
aicore_log_viewer	SAP AI Core 테넌트의 애플리케이션 로그 보기
aicore_credential_viewer	객체 저장소 자격 증명 및 Docker 레지스트리 자격 증명 보기 SAP AI Core 테넌트의 (메타데이터만)
aicore_credential_admin	객체 저장소 자격 증명 보기, 생성, 업데이트 및 삭제 SAP AI의 docker 레지스트리 자격 증명(메타데이터만) 핵심 테넌트
시나리오_메타데이터_뷰어	시나리오 및 시나리오 버전 보기
시나리오_실행_가능_뷰어	시나리오의 실행 파일 보기
시나리오_구성_뷰어	시나리오 구성 보기
시나리오_구성_편집기	시나리오 구성 편집
시나리오_배포_뷰어	배포 보기 또는 나열
시나리오_배포_편집기 보기	배포 생성, 업데이트 및 삭제 배포의 애플리케이션 로그 보기

역할

설명

시나리오_배포_예측기

시나리오 배포 호출(모델 추론 목적)

시나리오_실행_뷰어

시나리오 실행 보기 및 추적 측정항목 보기

시나리오_실행_편집기 보기

실행 및 추적 지표를 생성, 업데이트 및 삭제합니다.
실행의 애플리케이션 로그 보기

시나리오_아티팩트_뷰어

아티팩트 보기 또는 나열

시나리오_아티팩트_편집기

아티팩트 보기, 생성, 업데이트 및 삭제

시나리오_메트릭_뷰어

실행의 추적 측정항목 보기

역할 컬렉션

SAP AI Core는 다음과 같은 기본 역할 컬렉션을 제공합니다.

역할 컬렉션 및 역할

역할 수집

역할

aicore_viewer

- 시나리오_메타데이터_뷰어
- 시나리오_실행 가능_뷰어
- 시나리오_구성_뷰어
- 시나리오_배포_뷰어
- 시나리오_아티팩트_뷰어
- 시나리오_실행_뷰어
- 시나리오_메트릭_뷰어
- aicore_credential_viewer
- aicore_connection_viewer
- aicore_resourcegroup_viewer
- aicore_application_viewer

역할 수집

역할

aicore_admin

- 시나리오_구성_편집기
- 시나리오_배포_편집기
- 시나리오_배포_예측기
- 시나리오_실행_편집기
- 시나리오_아티팩트_편집기
- aicore_credential_admin
- aicore_connection_admin
- aicore_resourcegroup_editor
- aicore_log_viewer
- aicore_application_admin

aicore_scenario_viewer

- 시나리오_메타데이터_뷰어
- 시나리오_실행 가능_뷰어
- 시나리오_구성_뷰어
- 시나리오_배포_뷰어
- 시나리오_아티팩트_뷰어
- 시나리오_실행_뷰어
- 시나리오_메트릭_뷰어

aicore_scenario_editor

- 시나리오_구성_편집기
- 시나리오_배포_편집기
- 시나리오_배포_예측기
- 시나리오_실행_편집기
- 시나리오_아티팩트_편집기

aicore_resourcegroup_viewer

- aicore_resourcegroup_viewer

aicore_resourcegroup_editor

- aicore_connection_admin

aicore_application_admin

- aicore_application_admin

aicore_repository_admin

- aicore_repository_admin

관련 정보

[사용자 인증 및 관리 \[페이지 250\]](#)

14.4 도커 이미지

SAP AI Core는 테넌트별 Docker 레지스트리(관리 API를 통해 등록됨)를 지원합니다. 추가의 Docker 이미지를 참조하여 실행 및 배포와 같은 테넌트 워크로드를 생성할 수 있습니다.
이 Docker 레지스트리에서.

Docker 이미지는 가상 머신에 캐시됩니다. 이러한 캐시된 Docker 이미지는 다른 사람이 액세스할 수 없습니다.
테넌트이며 SAP에서는 액세스할 수 없습니다.

캐시된 Docker 이미지는 테넌트 오프보딩 시 즉시 삭제되지 않지만 다음 작업의 일부로 정리됩니다.

클러스터 축소 동작, 유지 관리, 가상 머신 업그레이드 등의 운영 이벤트입니다.

귀하가 사용하는 모든 서비스에는 귀하와 SAP 간에 보안 책임이 공유됩니다. 왜냐하면

Docker 이미지 생성은 테넌트의 책임이므로 그렇게 하지 않는 것이 좋습니다.

Docker 이미지에 개인 데이터, 민감한 데이터 또는 기계 학습 모델을 삽입하거나 하드 코딩하세요.

보안상의 이유로 SAP AI Core의 Docker 컨테이너는 루트가 아닌 사용자로만 실행됩니다. 자세한 내용은 다음을 참조하세요.

[워크플로 템플릿 \[페이지 106\]](#).

14.5 AI 콘텐츠 보안

AI 콘텐츠에는 워크플로 템플릿과 서빙 템플릿은 물론 여기에 사용되는 Docker 이미지도 포함됩니다.

Docker 이미지에는 기계 학습 라이브러리와 함께 기계 학습 알고리즘 또는 코드가 포함되어 있습니다.

프레임워크 및 기타 종속 패키지. 보안 개발을 위한 표준 관행을 준수하는지 확인하세요.

AI 콘텐츠 작업 시 소프트웨어.

보안 관행

관행	설명
위험 모델링	위험 모델링 워크숍을 개최하여 보안 위협이나 위협을 식별하고 평가합니다. AI 콘텐츠.
정적 코드 스캔	SAST(정적 코드 스캔) 도구를 사용하여 코드의 취약점을 스캔하고 분석합니다.
오픈소스 취약점 스캔	제품에서 사용되는 오픈 소스 구성 요소의 취약점을 검사하고 발견된 취약한 OSS 구성요소를 패치합니다.
오픈소스 전략	오픈 소스 구성 요소가 사용되는 시기를 정의하는 업데이트 전략 개발 제품이나 서비스가 최신 보안 버전으로 업데이트됩니다.
코드 리뷰	각 코드 변경에 대해 피어 코드 검토를 수행합니다. 검토자는 다음을 수행해야 합니다. 보안 관점에서 코드를 자세히 살펴보세요.
악성 코드 검사	AI 콘텐츠용으로 업로드된 데이터에서 악성코드를 검사합니다.
안전한 코드 보호	배포된 서비스를 통해 소스코드부터 시작하여 보안보증을 지원합니다. 예를 들어 Docker 이미지 디아제스트 및 이미지 서명 확인을 사용합니다.
Docker 기본 이미지 보안	AI 콘텐츠용 Docker 이미지를 빌드하려면 안전하고 가벼운 기본 이미지를 사용하세요. 사용 가능한 최신 기본 이미지를 사용하고 사용하지 않는 구성 요소를 제거했는지 확인하세요. Docker 이미지에서.

14.6 쿠버네티스 보안

워크플로 템플릿 및 제공 템플릿에서 관련 있고 적용 가능한 Kubernetes 보안 기능을 활성화하는 것이 좋습니다. 워크로드에 적절한 Kubernetes 기능을 활성화했는지 확인하십시오.

관련 정보

[Kubernetes 배포를 위한 보안 모범 사례](#)



14.7 구성 데이터 및 비밀

워크로드는 런타임 시 자격 증명을 사용하여 각각 저장소 이외의 네트워크 리소스에 액세스할 수 있습니다. 런타임에 이 정보를 포함하는 방법은 기밀성에 관한 기준이 다릅니다.

- 민감한 정보의 경우:

SAP AI Core를 사용하면 일반 비밀 형식으로 비밀을 포함할 수 있습니다. 일반 비밀은 SAP AI Core의 REST API에 의해 생성 및 관리되며 워크로드에서 안전하게 사용됩니다.

- 민감하지 않은 정보의 경우:

구성이나 라벨을 사용하여 민감하지 않은 매개변수를 포함할 수 있습니다.

참고

이러한 매개변수는 일반 텍스트(예: GET 요청)로 반환될 수 있습니다.

14.8 출력 인코딩

비즈니스 기능 중단을 방지하기 위해 SAP AI Core는 사용자 입력을 삭제하지 않습니다. AI API를 사용하는 소비자 또는 애플리케이션은 XSS 공격을 방지하기 위해 사용 컨텍스트에 따라 필요한 출력 인코딩을 수행해야 합니다.

관련 정보

[XSS\(교차 사이트 스크립팅\) – OWASP 사이트](#)

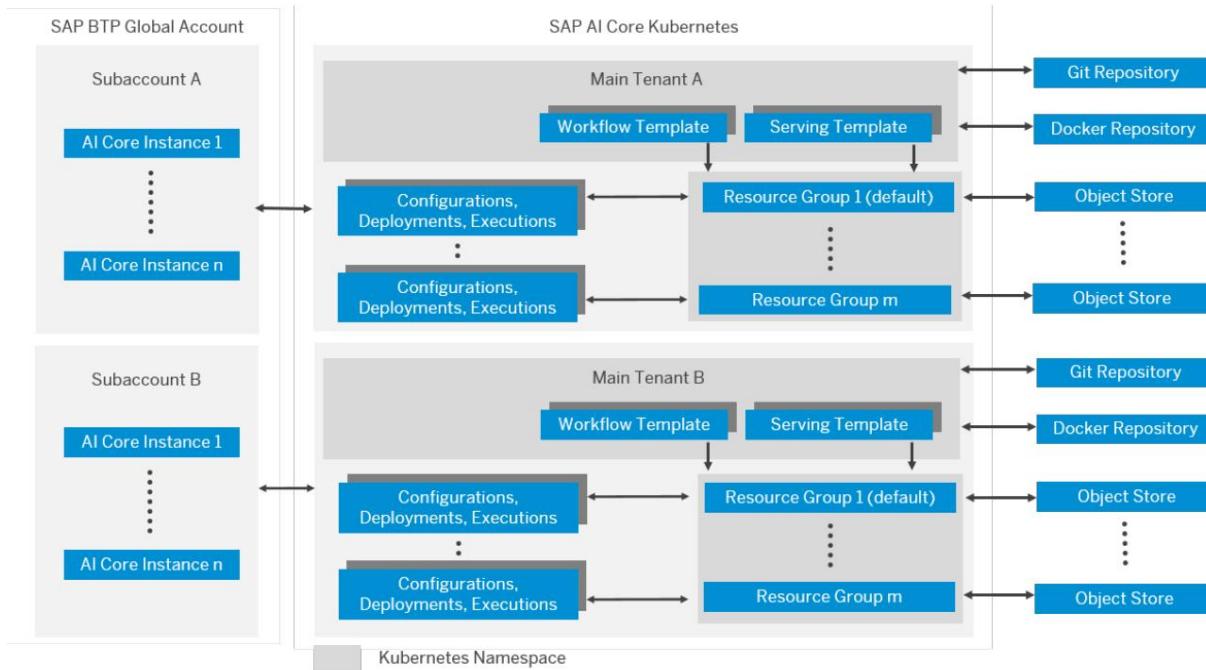


[크로스 사이트 스크립팅 방지 치트 시트 – OWASP 사이트](#)



14.9 다중 테넌트

SAP AI Core는 테넌트 인식 BTP 재사용 서비스로, 기본 테넌트와 리소스 그룹을 지원합니다. 리소스는 아래에 설명된 대로 테넌트 또는 리소스 그룹에 대해 정의됩니다.



각 기본 테넌트와 리소스 그룹은 네임스페이스에 매핑됩니다. 기본 테넌트 네임스페이스에는 워크플로 및 모델 제공을 위한 템플릿만 포함됩니다. 이러한 개체의 인스턴스는 각 리소스 그룹 네임스페이스에 생성되고 기본 테넌트 네임스페이스의 해당 템플릿을 참조합니다. 각 기본 테넌트에는 기본 테넌트의 작업 부하에 사용할 수 있는 기본 리소스 그룹이 있습니다.

테넌트 수준 리소스

테넌트 수준 리소스에는 다음과 같은 실행 파일이 포함됩니다.

- 워크플로 템플릿
- 템플릿 제공
- Docker 이미지가 포함된 Docker 레지스트리 • 사용자 인증 및 권한 부여(UAA)

사용자 인증 및 권한 부여는 SAP AI Core 테넌트(SAP AI Core용 서비스 키를 사용하여 얻은 액세스 토큰)를 기반으로 합니다. 런타임 시 또는 AI API를 통해 수명 주기를 관리할 때 SAP AI Core 테넌트는 요청 헤더에 적절한 리소스 그룹을 설정해야 합니다.

리소스 그룹 수준 리소스

테넌트 수준의 실행 파일은 모든 리소스 그룹에서 공유됩니다. 반면, 실행, 배포, 구성 및 아티팩트와 같은 런타임 엔터티는 특정 리소스 그룹에 속하며 리소스 그룹 간에 공유될 수 없습니다. 마찬가지로 리소스 그룹 내에서 생성된 일반 비밀은 해당 그룹 내의 워크로드에만 사용됩니다.

리소스 그룹 헤더를 설정하여 리소스 그룹 수준에서 개체 저장소를 등록할 수 있습니다. 여러 리소스 그룹에 대해 동일한 IAM 사용자와 동일한 객체 저장소 버킷을 사용하지 않는 것이 좋습니다.

워크로드의 테넌트 격리

워크로드는 샌드박스 환경에서 실행되며 다른 테넌트 또는 리소스 그룹의 워크플로에 액세스할 수 없습니다.

워크로드의 인바운드 또는 아웃바운드 트래픽에는 TCP만 지원됩니다. UDP 포트에서 소켓을 열기 위해 워크로드를 여는 것은 권장되지 않습니다. 사용할 수는 없지만 작업 부하에 이론적 보안 문제를 일으킬 수 있습니다.

14.10 데이터 보호 및 개인정보 보호

SAP Business Technology Platform의 데이터 보호 및 개인정보 보호에 대한 일반 정보는 [데이터 보호 및 개인정보 보호를 참조하세요](#).

데이터 보호는 수많은 법적 요구사항 및 개인정보 보호 문제와 연관되어 있습니다. 일반적인 데이터 보호 및 개인 정보 보호법을 준수하는 것 외에도 다양한 국가/지역의 산업별 법률을 준수를 고려할 필요가 있습니다. 이 섹션에서는 관련 법적 요구 사항 및 데이터 개인 정보 보호를 준수하도록 지원하기 위해 SAP AI Core가 제공하는 특정 특징과 기능에 대해 설명합니다.

이 가이드는 이러한 특징과 기능이 회사, 산업, 지역 또는 국가/지역별 요구 사항을 지원하는 가장 좋은 방법인지에 대한 조언을 제공하지 않습니다. 또한 이 가이드는 특정 환경에 필요한 추가 기능에 대한 조언이나 권장 사항을 제공하지 않습니다.

데이터 보호와 관련된 결정은 사례별로 이루어져야 하며 주어진 시스템 환경과 해당 법적 요구 사항을 고려하여 이루어져야 합니다.

참고

SAP는 어떤 형태로든 법률 자문을 제공하지 않습니다. SAP 소프트웨어는 개인 데이터의 단순화된 차단 및 삭제와 같은 보안 기능과 특정 데이터 보호 관련 기능을 제공하여 데이터 보호 규정 준수를 지원합니다. 대부분의 경우 해당 데이터 보호 및 개인 정보 보호법 준수는 제품 기능에 포함되지 않습니다. 이 문서에 사용된 정의 및 기타 용어는 특정 법적 출처에서 가져온 것이 아닙니다.

데이터 보호가 보장되는 정도는 안전한 시스템 작동에 따라 달라집니다. 네트워크 보안, 보안 노트 구현, 시스템 변경 사항에 대한 적절한 기록, 시스템의 적절한 사용은 데이터 개인 정보 보호 법률 및 기타 법률을 준수하기 위한 기본 기술 요구 사항입니다.

SAP BTP의 데이터 보호 및 개인정보 보호 용어집은 [데이터 보호 및 개인정보 보호에 대한 SAP BTP 용어집을 참조하세요.](#)

14.10.1 데이터 저장 및 처리

SAP AI Core는 구성 파일, 기계 학습(ML) 교육, ML 서빙 등 데이터를 처리할 수 있는 기능을 제공합니다.

SAP AI Core는 데이터 프로세서 역할을 하며 데이터 유형이나 데이터 카테고리를 인식하지 않습니다. SAP AI Core 고객은 데이터 컨트롤러로서 데이터 저장 및 처리 요구 사항에 대한 데이터 보호 및 개인 정보 보호(DPP) 책임을 이행할 책임이 있습니다.

14.10.2 변경 로깅 및 읽기 액세스 로깅

SAP AI Core는 데이터 프로세서 역할을 하며 데이터 유형이나 데이터 카테고리를 인식하지 않습니다. SAP AI Core 고객은 데이터 컨트롤러로서 데이터 저장 및 처리 요구 사항에 대한 데이터 보호 및 개인 정보 보호(DPP) 책임을 이행할 책임이 있습니다.

SAP AI Core를 사용하여 개발하는 모든 애플리케이션이나 서비스의 경우 관련 로깅 기능이 포함되어 있는지 확인하고, 데이터가 올바르게 기록되는지 확인하여 데이터 개인정보 보호법을 준수해야 합니다.

14.10.3 동의

SAP AI Core는 데이터 프로세서 역할을 하며 데이터 유형이나 데이터 카테고리를 인식하지 않습니다. SAP AI Core 고객은 데이터 컨트롤러로서 개인 데이터를 수집하기 전에 데이터 주체의 동의 요청을 이행할 책임이 있습니다.

14.10.4 삭제

SAP AI Core는 Bring your object store를 지원합니다. 이를 통해 고객은 아티팩트 관련 파일(예: 교육 데이터, 기계 학습 모델 또는 기타 유형)이 저장되는 개체 저장소 비밀을 등록할 수 있습니다. 이러한 데이터는 ML 학습 또는 ML 제공과 같은 처리 중에 ML 워크로드에서 사용됩니다.

아티팩트의 개념은 데이터의 메타데이터를 캡처하는 것으로 제한됩니다. 데이터는 개체 저장소에 물리적으로 저장되며 SAP AI Core는 파일 삭제에 대해 책임을 지지 않습니다. AI API를 사용하여 아티팩트를 삭제하면 해당 메타데이터가 SAP AI Core 서비스에서 삭제되고 실제 파일은 개체 저장소에서 삭제되지 않습니다. AI API에 대한 자세한 내용은 [AI API 개요 \[페이지 24\]](#)를 참조하십시오.

오프보딩 시 SAP AI Core는 처리 목적으로 사용되는 AI Core 내의 캐시된 데이터를 정리합니다.

SAP AI Core 고객은 데이터 컨트롤러로서 등록된 개체 저장소에서 데이터를 삭제할 책임이 있습니다.

14.10.5 보안 및 고객 데이터 보호

SAP 제품 표준 보안과 데이터 보호 및 개인 정보 보호(DPP) 요구 사항은 SAP에 맡겨진 고객 데이터를 안전하게 보호하는 데 있어 높은 표준과 의무를 설정합니다.

고객 데이터 보호는 세 가지 방식으로 처리됩니다.

- 고객 데이터는 본 서비스에서 제공하는 목적 이외의 목적으로 가져오고, 출력하고, 처리됩니다.
고객이 구독했습니다.
- 인증을 포함한 보안 기술을 통해 고객 데이터를 악의적인 액세스로부터 보호합니다.
그리고 승인.
- 고객 데이터는 다음을 통해 SAP 관리자나 지원 담당자에게 우발적으로 노출되지 않도록 보호됩니다.
보안 정책, 액세스 제어 및 모니터링.

SAP AI Core의 15가지 접근성 기능

SAP AI Core 경험을 최적화하기 위해 SAP AI Core는 소프트웨어를 효율적으로 사용하는 데 도움이 되는 기능과 설정을 제공합니다.

참고

SAP AI Core는 SAPUI5를 기반으로 하는 인터페이스로 SAP AI Launchpad를 사용합니다. 이러한 이유로 SAPUI5의 접근성 기능도 적용됩니다. [최종 사용자를 위한 접근성](#)에서 SAP Help Portal의 SAPUI5에 대한 접근성 문서를 참조하세요.

화면 판독기 지원 및 키보드 단축키에 대한 자세한 내용은 [SAPUI5 컨트롤에 대한 화면 판독기 지원을 참조하세요](#). [SAPUI5 요소에 대한 키보드 처리](#).

16 모니터링 및 문제 해결

잠재적인 문제에 대한 솔루션을 살펴보고 지원을 받는 방법을 알아보세요.

지원 받기

이 서비스에 문제가 발생하면 아래 절차를 따르는 것이 좋습니다.

플랫폼 상태 확인

[SAP Trust Center](#)에서 플랫폼의 가용성을 확인하세요..



플랫폼 가용성, 업데이트 및 알림에 대한 자세한 내용은 [Cloud Foundry 환경의 플랫폼 업데이트 및 알림](#)을 참조하세요.

안내 답변 확인

SAP 지원 포털에서 [안내 답변을](#) 확인하세요. SAP Cloud Platform 섹션. 일반적인 SAP Cloud Platform 문제는 물론 특정 서비스에 대한 솔루션도 찾을 수 있습니다.

SAP 지원팀에 문의

[SAP Support Portal](#)을 통해 사건이나 오류를 보고할 수 있습니다..



사건에 대해 다음 구성 요소를 사용하십시오.

구성요소 이름	구성요소 설명
CA-ML-AIC	SAP AI 코어

사건을 제출할 때 다음 정보를 포함하는 것이 좋습니다.

- 지역정보(카나리아, EU10, US10)
- 하위 계정 기술 이름
- 사건이나 오류가 발생한 페이지의 URL • 오류를 복제하는 데 사용된 단계 또는 클릭
- 스크린샷, 동영상 또는 입력한 코드

16.1 문제 해결

문제 해결 정보는 다음 섹션을 참조하세요.

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)[응용 프로그램 \[페이지 269\]](#)[실행 \[페이지 269\]](#)[도커 \[페이지 279\]](#)[배포 \[페이지 270\]](#)[기타 \[페이지 279\]](#)

16.1.1 리포지토리

테넌트에 대한 리포지토리 ra-aicore-test를 찾을 수 없습니다.

결과는 다음과 같습니다.

```
{
  "오류": {
    "코드": "500",
    "세부정보": [
      {
        "code": null,
        "message": "테넌트에 대한 저장소 ra-aicore-test를 찾을 수 없습니다."
      }
    ],
    "message": "b82a8318 테넌트에 대한 저장소 ra-aicore-test를 찾을 수 없습니다." "request_id": null, "target": "/api/v4alpha/repositories"
  }
}
```

그리고:

```
AIAPIServerException: /admin/repositories 게시 실패: 테넌트 68에 대한 리포지토리 ra-aicore-test를 찾을 수 없습니다.
```

해결 방법을 따르십시오.

name 매개변수의 값에 다른 이름을 사용하십시오. 이름을 재사용하면 예외가 발생합니다.
aicore 테스트.

```
응답 = ai_api_client.rest_clinet.post( path="/admin/repositories", body={ "name": "aicore-test-1", "url": "https://github.com/John/aicore-test", }, )
print(response) {'message': '저장소가 온보딩되었습니다.'}
```

상위 주제: [문제 해결 \[페이지 262\]](#)

관련 정보

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[응용 프로그램 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[배포 \[페이지 270\]](#)

[기타 \[페이지 279\]](#)

16.1.2 구성

시나리오 *<y>* 에 대한 실행 파일 *<x>* 을 (를) 찾을 수 없기 때문에 구성 생성할 수 없습니다.

구성을 생성하려고 하면 해당 시나리오에 대한 실행 파일을 찾을 수 없다는 메시지가 나타납니다.

다음을 확인하세요.

1. 구성에서 실행 가능 ID에 대해 워크플로의 이름 값을 사용하고 있는지 확인합니다.

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터:

이름: text-clf-train-tutorial 주식:

대본"
...
시나리오.ai.sap.com/description: "SAP 개발자 튜토리얼
```

참고

[runnings.ai.sap.com/id](#)의 값을 실행 가능 ID로 사용하지 마십시오.

2. 워크플로의 실행 가능 파일 [.ai.sap.com/id](#) 값을 사용하고 있는지 확인하십시오.

시나리오 ID.

```
...
Artifacts.ai.sap.com/text-data.kind: "데이터 세트" Artifacts.ai.sap.com/text-model-tutorial.kind: "모델"
레이블:
시나리오.ai.sap.com/id: "text-clf-tutorial" ai.sap.com/version: "2.1.0"
투기:
```

...

로그 메시지: minio 클라이언트 사용

다음을 확인하세요.

- 구성에서 실행 가능 ID에 대해 워크플로의 이름 값을 사용하고 있는지 확인합니다.

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터:

이름: text-clf-train-tutorial 주석:

    시나리오.ai.sap.com/description: "SAP 개발자 튜토리얼
대본"
...
...
```

참고

[runnings.ai.sap.com/id](#)의 값을 실행 가능 ID로 사용하지 마십시오.

- 워크플로의 실행 가능 파일 [.ai.sap.com/id](#) 값을 사용하고 있는지 확인하십시오.

시나리오 ID.

```
...
Artifacts.ai.sap.com/text-data.kind: "데이터 세트" Artifacts.ai.sap.com/text-model-tutorial.kind: "모델"
레이블:
...
시나리오.ai.sap.com/id: "text-clf-tutorial" ai.sap.com/version: "2.1.0" 사양:
...
...
```

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[응용 프로그램 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[배포 \[페이지 270\]](#)

[기타 \[페이지 279\]](#)

16.1.3 아티팩트

출력 아티팩트가 생성되지 않았습니다.

다음을 완료:

워크플로의 출력 아티팩트에 대한 globalName 매개변수를 정의합니다.

```

...
실행 파일.ai.sap.com/description: "텍스트 분류 Scikit 훈련 실행"

실행 가능 파일.ai.sap.com/이름: "text-clf-train-tutorial-exec" 유물.ai.sap.com/text-data.kind: "데이터 세트" 유물.ai.sap.com /text-
model- tutorial.kind: "마음" 라벨:

시나리오.ai.sap.com/id: "text-clf-tutorial" ai.sap.com/version: "2.1.0"

...
...
출력: 아티팩트:

-name: text-model-tutorial 경로: /app/model 전역 이름:
text-model-tutorial 아카이브: 없음:
{}

컨테이너:
...

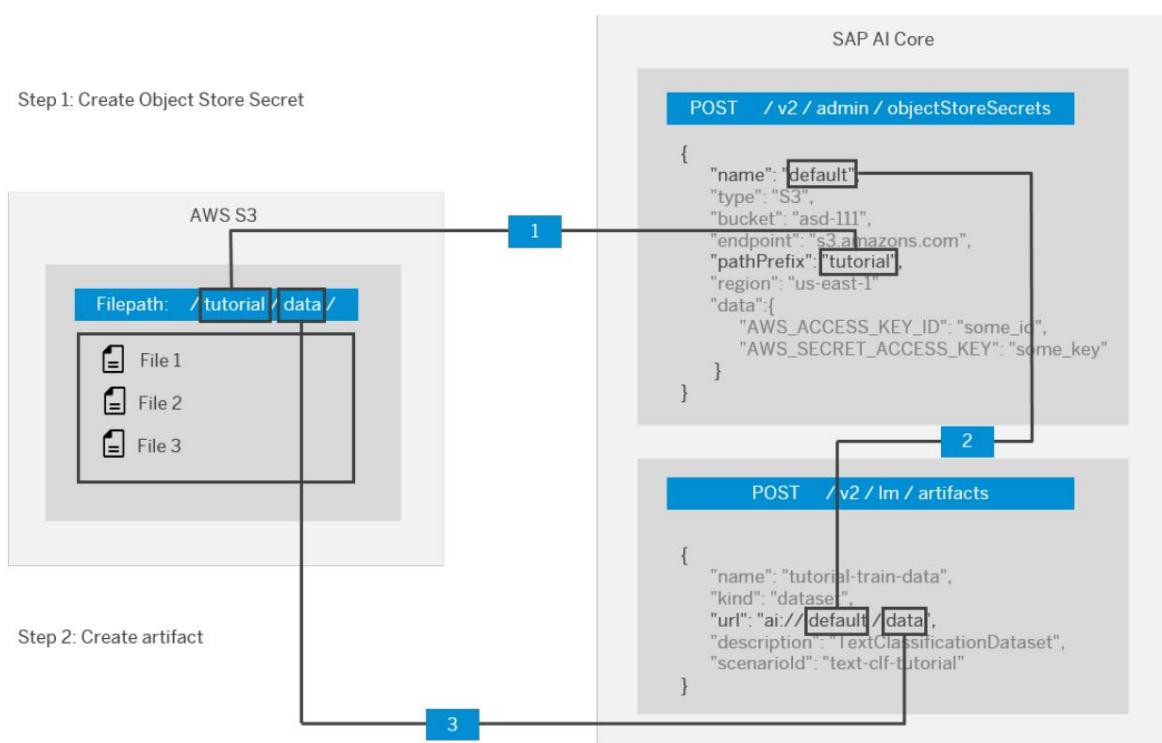
```

아티팩트를 로드하지 못했습니다. 지정된 키가 존재하지 않습니다.

다음을 완료:

1. <name> 명령 규칙과 pathPrefix를 사용하여 객체 저장소 버킷을 생성했는지 확인하세요.

AWS S3 경로에서 다음 다이어그램을 참조하세요.



2. 아티팩트를 생성할 때 URL 매개변수에 후행 슬래시(/)를 추가하지 마세요.

- 잘못된 사용법: "url": "ai://yourObjectStoreSecretName/folder/subfolder/"
- 올바른 사용법: "url": "ai://yourObjectStoreSecretName/folder/subfolder"

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[응용 프로그램 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[배포 \[페이지 270\]](#)

[기타 \[페이지 279\]](#)

16.1.4 적용

템플릿은 애플리케이션을 통해 동기화되지 않습니다.

애플리케이션을 삭제하거나 생성한 후에는 템플릿이 동기화되지 않습니다.

해결 방법을 따르십시오.

1. 엔드포인트를 사용하여 동기화된 애플리케이션을 삭제합니다.

[{{apiurl}}/v2/admin/applications/{{appName}} 삭제](#)

2. 엔드포인트를 사용하여 연결된 GitHub 리포지토리를 오프보딩합니다.

[{{apiurl}}/v2/admin/repositories/{{repositoryName}} 삭제](#)

3. GitHub 비밀번호 대신 개인 액세스 토큰을 사용하여 연결된 GitHub 리포지토리를 온보딩합니다.

자세한 내용은 [개인 액세스 토큰 만들기를](#) 참조하세요..



[POST {{apiurl}}/v2/admin/repositories](#)

샘플 코드

몸:

```
{
  "이름": "aicom-test",
  "url": "https://github.com/john/aicom-test", "username": "john", "password": "yourGitHubPersonalAccessToken"
}
```

4. 엔드포인트를 사용하여 애플리케이션을 생성합니다.

[POST {{apiurl}}/v2/admin/applications](#) 5. ArgoCD 애플리케이션을 확인

하여 저장소가 해당 파일에 대해 올바르게 동기화되었는지 확인합니다.

거주자. 예를 들어 중복된 워크플로 이름이 없는지 확인하세요. name 매개변수의 값은 실행 가능 ID로 간주됩니다.

```
이름: text-clf-train-tutorial
주석:
...api버전: argoproj.io/v1alpha1
종류: WorkflowTemplate 메타데이터:
```

6. 예상 테넌트를 사용하여 SAP AI Core를 호출하고 있는지 확인합니다.

7. 워크플로 템플릿에 올바른 시나리오 라벨이 포함되어 있는지 확인하세요.

8. 엔드포인트를 사용하여 애플리케이션 동기화 상태를 가져옵니다.

[GET {{apiurl}}/v2/admin/applications/{{appName}}/status](#) 상태는 템플릿에서 오류를 반환합니다. 템플릿이 업데이트되면 약 3분 후에 애플리케이션이 자동으로 다시 동기화됩니다.

실행 파일은 시나리오에서 검색할 때 나타나지 않습니다.

해결 방법을 따르십시오.

1. 엔드포인트를 사용하여 동기화된 애플리케이션을 삭제합니다.

[{{apiurl}}/v2/admin/repositories/{{appName}} 삭제](#)

2. 엔드포인트를 사용하여 연결된 GitHub 리포지토리를 오프보딩합니다.

[{{apiurl}}/v2/admin/repositories/{{repositoryName}} 삭제](#)

3. GitHub 비밀번호 대신 개인 액세스 토큰을 사용하여 연결된 GitHub 리포지토리를 온보딩합니다.

자세한 내용은 [개인 액세스 토큰 만들기](#)를 참조하세요..

POST {{apiurl}}/v2/admin/repositories

샘플 코드

몸:

```
apiVersion: argoproj.io/v1alpha1
  "name": "aicore-test",
  "url": "https://github.com/john/aicore-test", "username": "john", "password": "yourGitHubPersonalAccessToken"
```

4. 엔드포인트를 사용하여 애플리케이션을 생성합니다.

POST {{apiurl}}/v2/admin/applications

5. ArgoCD 애플리케이션을 확인

하여 저장소가 해당 파일에 대해 올바르게 동기화되었는지 확인합니다.

거주자. 예를 들어 중복된 워크플로 이름이 없는지 확인하세요. name 매개변수의 값은 실행 가능 ID로 간주됩니다.

```
apiVersion: argoproj.io/v1alpha1
  종류: WorkflowTemplate
  메타데이터:
```

이름: text-clf-train-tutorial

주석:

...

6. 사용자가 예상 테넌트를 사용하여 SAP AI Core를 호출하고 있는지 확인합니다.

7. 워크플로 템플릿의 시나리오 라벨이 올바른지 확인하세요.

8. 엔드포인트를 사용하여 애플리케이션 동기화 상태를 가져옵니다.

GET {{apiurl}}/v2/admin/applications/{{appName}}/status

상태는 템플릿에서 오류를 반환합니다. 템플릿이

업데이트되면 약 3분 후에 애플리케이션이 자동으로 다시 동기화됩니다.

애플리케이션 상태가 정상으로 반환되지만 대부분의 다른 속성은 알 수 없습니다.

해결 방법을 따르십시오.

1. 엔드포인트를 사용하여 동기화된 애플리케이션을 삭제합니다.

[{{apiurl}}/v2/admin/applications/{{appName}} 삭제](#)

2. 엔드포인트를 사용하여 연결된 GitHub 리포지토리를 오프보딩합니다.

[삭제](#)

[{{apiurl}}/v2/admin/repositories/{{repositoryName}}](#)

3. GitHub 비밀번호 대신 개인 액세스 토큰을 사용하여 연결된 GitHub 리포지토리를 온보딩합니다.

자세한 내용은 [개인 액세스 토큰 만들기](#)를 참조하세요..

POST {{apiurl}}/v2/admin/repositories

샘플 코드

몸:

```
{ "이름": "aicore-테스트",
  "url": "https://github.com/john/aicore-test", "사용자 이름": "john", "password":
  "yourGitHubPersonalAccessTokenHere" }
```

4. 엔드포인트를 사용하여 애플리케이션을 생성합니다.

POST {{apiurl}}/v2/admin/applications

5. ArgoCD 애플리케이션을 확인하여 저장소가 해당 파일에 대해 올바르게 동기화되었는지 확인합니다.

거주자. 예를 들어 중복된 워크플로 이름이 없는지 확인하세요. name 매개변수의 값은 실행 가능 ID로 간주됩니다.

```
apiVersion: argoproj.io/v1alpha1 종류: WorkflowTemplate 메타데이터: 이
름: text-clf-train-tutorial 주석:
```

...

6. 사용자가 예상 테넌트를 사용하여 SAP AI Core를 호출하고 있는지 확인합니다.

7. 워크플로 템플릿에 올바른 시나리오 라벨이 포함되어 있는지 확인하세요.

8. 엔드포인트를 사용하여 애플리케이션 동기화 상태를 가져옵니다.

GET {{apiurl}}/v2/admin/applications/{{appName}}/status 상태는 템플릿에서 오류를 반환합니다. 템플릿이 업데이트되면 약 3분 후에 애플리케이션이 자동으로 다시 동기화됩니다.

애플리케이션 상태 메시지: rpc 오류: 코드 = 알 수 없음 desc = my-path: 앱 경로

존재하지 않는다

애플리케이션에 지정된 경로가 저장소에 존재하지 않습니다.

해결 방법을 따르십시오.

애플리케이션을 삭제하고 올바른 경로를 사용하여 새 애플리케이션을 만듭니다.

애플리케이션 상태 메시지: 애플리케이션 저장소 <your git 저장소>가 아닙니다.

프로젝트 'xyz'에서 하용됨

온보딩된 저장소에서 저장소 URL을 찾을 수 없습니다.

다음을 확인하세요.

GET {{apiurl}}/v2/admin/applications를 사용하고 저장소 URL에 "status": "COMPLETED"가 있는지 확인하여 애플리케이션에 지정된 저장소가 성공적으로 온보딩되었는지 확인하세요.

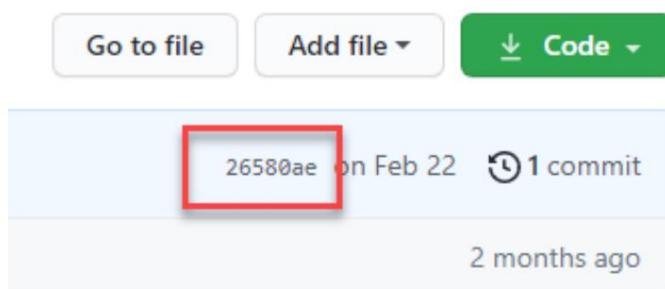
애플리케이션 상태 메시지: rpc 오류: 코드 = 알 수 없음 desc = 해결할 수 없음

커밋 SHA에 대한 '존재하지 않는 지점'

애플리케이션에 지정한 개정판이 저장소에 존재하지 않습니다.

해결 방법을 따르십시오.

애플리케이션을 삭제하고 올바른 개정판을 사용하여 새 애플리케이션을 생성하세요. 개정 번호는 GitHub에서 찾을 수 있습니다.



또는 HEAD를 입력하여 최신 커밋을 참조하세요.

애플리케이션 상태 메시지: rpc 오류: 코드 = FailedPrecondition desc = 실패

\"workflow.yaml\" 비정렬화: 매니페스트 비정렬 실패: 변환 중 오류가 발생했습니다.

YAML에서 JSON으로: yaml: 7행: 이 컨텍스트에서는 매픽 값이 허용되지 않습니다.

워크플로 템플릿에 구문 오류가 있습니다.

해결 방법을 따르십시오.

Argo Lint를 사용하여 워크플로 템플릿의 구문 오류를 식별하세요. Argo Lint IDE를 설정하려면 [Argo Lint IDE 설정](#)을 참조하세요.

애플리케이션 상태 메시지: spec.source.repoURL 및 spec.source.path

spec.source.chart가 필요합니다.

애플리케이션에 빈 경로를 지정했습니다.

```
{
  "healthStatus": "알 수 없음",
```

```

    "message": "spec.source.repoURL 및 spec.source.path 중 하나
spec.source.chart가 필요합니다.", "reconciledAt": "알 수 없음",
    "source": { "path": "알 수 없음", "repoURL": "알 수 없음",
    "revision": "알 수 없음"

},
"syncFinishedAt": "알 수 없음",
"syncRessourcesStatus": [], "syncStartedAt": "알
수 없음", "syncStatus": "알 수 없음"
}

```

해결 방법을 따르십시오.

애플리케이션을 삭제하고 경로를 지정하여 새 애플리케이션을 만듭니다. 엔드포인트({{apiurl}}/v2/admin/applications/{{appName}}/status)를 사용하여 상태를 확인하세요.

수동으로 애플리케이션 동기화

애플리케이션은 최대 3분 간격으로 자동으로 GitHub 저장소와 동기화됩니다. 아래 엔드포인트를 사용하여 수동으로 동기화를 요청하세요.

<{{apiurl}}/admin/applications/{{appName}}/refresh>

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[배포 \[페이지 270\]](#)

[기타 \[페이지 279\]](#)

16.1.5 실행

실행 상태가 오랫동안 DEAD 또는 PENDING입니다.

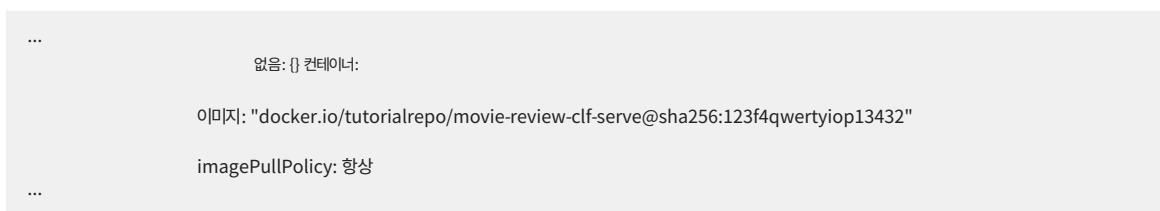
다음을 확인하세요.

1. 실행 로그를 확인합니다.
2. 제공한 매개변수 이름과 실행 이름이 템플릿의 이름과 일치하는지 확인하세요. 메모
이름은 대소문자를 구분합니다.

실행을 위해 개인 docker.io의 이미지를 가져올 수 없습니다.

해결 방법을 따르십시오.

1. Docker Hub의 도메인 이름을 지정합니다. 예를 들어:



도메인 이름을 아직 모르는 경우 다음 이미지를 참조하세요.

Docker Hub search results for 'tutorialrepo/movie-review-clf-serve':latest



tutorialrepo/movie-review-clf-serve:latest

DIGEST: sha256:123f4qwertyiop13432

OS/ARCH

linux/amd64

COMPRESSED SIZE ⓘ

2.7 GB

LAST PUSHED

16 days ago by tutorialrepo

2. 템플릿의 이미지 버전 대신 Docker 이미지 다이제스트를 지정합니다.

이미지 튜토리얼repo/movie-review-clf-serve@sha256:123f4qwertyiop13432

실행 목록의 출력 아티팩트가 비어 있고 오랫동안 UNKNOWN 상태입니다.

다음을 확인하세요.

1. 이름이 default인 개체 암호가 현재 리소스 그룹에 있는지 확인합니다. 이상 자세한 내용은 [개체 저장소 암호 등록 \[페이지 76\]](#)을 참조하십시오.
2. Docker 이미지를 가져오는 데 필요한 Docker 레지스트리 비밀을 생성했는지 확인하세요. 로그는 실행이 시작된 후에만 사용할 수 있습니다.

GET 실행이 오랫동안 UNKNOWN 상태였습니다.

다음을 확인하세요.

1. 현재 리소스 그룹에 default라는 이름의 개체 비밀이 존재하는지 확인합니다. 이상 자세한 내용은 [개체 저장소 암호 등록 \[페이지 76\]](#)을 참조하십시오.
2. Docker 이미지를 가져오는 데 필요한 Docker 레지스트리 비밀을 생성했는지 확인하세요. 로그는 실행이 시작된 후에만 사용할 수 있습니다.

로그 없이 실행 상태가 DEAD로 변경됨

다음을 확인하세요.

1. 템플릿이 Argo 사양을 충족하는지, SAP AI Core에서 실행할 수 있는지 확인하세요.
2. 템플릿에서 자동으로 오류를 식별하려면 Argo Linter를 사용하십시오.
3. 실행 중에 생성될 아티팩트를 저장하기 위해 Docker 폴더를 생성했는지 확인합니다.

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[응용 프로그램 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[배포 \[페이지 270\]](#)

[기타 \[페이지 279\]](#)

16.1.6 도커

컨테이너 기본 로그를 가져오는 중에 오류가 발생했습니다. 백오프 풀링 이미지

템플릿을 SAP AI Core에 푸시하면 "컨테이너 기본 로그 가져오기 오류" 오류가 발생합니다.
"백오프 풀링 이미지".

다음 단계를 완료하세요.

1. Docker 레지스트리가 공용이고 방화벽 뒤에서 보호되지 않는지 확인하십시오.
조직.
Docker 이미지가 공개되지 않은 경우 SAP AI Core에서 Docker 레지스트리 비밀을 생성했는지 확인하세요.
2. 이전 자격 증명을 사용하여 로컬 컴퓨터에서 Docker 이미지를 가져올 수 있는지 확인합니다.
단계.
3. 실행 파일에 동일한 Docker 레지스트리 비밀을 지정합니다.
4. 다음 형식으로 실행 파일의 Docker 이미지 경로를 지정합니다.

`<DOCKER_REGISTRY>/<REPO_NAME>/<DOCKER_IMAGE>:<TAGNAME>`

예

`docker.io/tutorialrepo/text-clf-train:0.0.1`

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[응용 프로그램 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[배포 \[페이지 270\]](#)

[기타 \[페이지 279\]](#)

16.1.7 배포

UNKNOWN 상태의 배포를 강제로 중지하려고 합니다.

배포를 중지하고 삭제하고 싶지만 배포 상태가 "알 수 없음"이므로 그렇게 할 수 없습니다.

다음과 같이 PATCH 요청 제출을 시도했습니다.

```
패치 {{apiurl}}/lm/deployments/d4fec9c24c54f87e
```

그러나 다음과 같은 응답이 나타납니다.

```
{
  "error": {
    "code": "01010076",
    "message": "잘못된 요청, 현재 상태 UNKNOWN은 변경할 수 없습니다."
  },
  "requestID": "e110820e-1cfe-456a-bb0e-77907b36422c",
  "target": "/apu/v2/배포/d4fec9c24c54f87e"
}
```

다음 단계를 완료하세요.

1. [GET {{apiurl}}/v2/lm/deployments/{{deploymentid}}](#) 엔드포인트를 사용하여 배포 상태가 "알 수 없음"인 이유를 알아보세요.
2. 배포를 중지하지 않고 삭제합니다(배포 중지는 배포를 중지하는 경우에만 필요함).
달리기):
[{{apiurl}}/v2/lm/deployments/{{deploymentid}}](#) 삭제

배포가 PENDING 상태로 유지됩니다.

다음을 확인하세요.

1. 프라이빗 Docker 이미지를 사용할 때 Docker Registry 비밀이 존재하는지 확인하세요.
2. Docker 이미지를 로컬 시스템에 다운로드할 수 있는지 확인하세요.

배포 ID <abc>를 찾을 수 없습니다.

이 메시지는 배포를 방금 시작했을 때 나타납니다. 몇 분 정도 기다리면 메시지가 자동으로 해결됩니다.

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[응용 프로그램 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[기타 \[페이지 279\]](#)

16.1.8 기타

403 - 금지됨: RBAC 액세스가 거부되었습니다.

실행 또는 구성에 대한 POST 요청을 제출하면 오류 403 - 금지됨이 발생합니다.
RBAC 액세스가 거부되었습니다.

다음을 확인하세요.

1. 올바른 토큰과 ai-resource-group 헤더를 전달하고 있는지 확인하세요.
2. 테넌트 프로비저닝을 확인하세요.

런타임 어댑터 예외가 발생합니다.

다음 오류가 발생합니다. "런타임 어댑터 예외; 배포 개시 실패: {\n \"code\": \"400\",\\n \"message\": \"입력 매개변수 또는 아티팩트 누락, 하나 또는
더 많은 자리 표시자 값이 제공 사양에서 확인되지 않으며 오류는 'dict object'에 '' 속성이 없습니다.\"\\n}\\n"입니다.

다음을 확인하세요.

입력 아티팩트 키에 'example-artifact-key'와 같은 구분 기호가 포함되어 있지 않은지 확인하세요. 그렇다면 키 이름을 바꾸세요(예: "exampleArtifactKey").

템플릿을 GitHub 저장소에 푸시했지만 API를 통해 생성된 실행 파일을 볼 수 없습니다.

템플릿에 오류가 있을 수 있습니다. 다음을 확인하세요.

- 제공 템플릿은 항상 입력 매개변수가 구성될 것으로 예상합니다. 매개변수가 없는 경우
게재 템플릿에 더미 매개변수를 추가하고 이에 대한 구성을 만들습니다.
- 하이픈(-)은 템플릿 이름에 허용되는 유일한 구분 기호입니다.

템플릿에서 시나리오를 생성했지만 AI API 호출에서 볼 수 없습니다.

다음 해결 방법을 사용하세요.

시나리오를 표시하려면 새 시나리오_id(제공 템플릿 아님)가 포함된 워크플로 템플릿을 추가하세요.

오류: getaddrinfo ENOTFOUND

다음을 확인하세요.

- 모든 환경 변수가 SAP AI Core 서비스 키와 일치하는지 확인하세요. 구체적으로 다음을 확인하세요.
 - 인증_URL
 - 클라이언트_ID
 - 클라이언트_비밀
 - apiurl
- `{{apiurl}}/v2/admin/repositories` 엔드포인트에 GET 요청을 제출합니다. 3. 문제가 지속되면 [모니터링 및 문제 해결](#) [페이지 266]에 설명된 대로 SAP 지원 센터에 문의하십시오.

Git 리포지토리는 SAP AI Core 인스턴스와 동기화되지 않습니다.

git 리포지토리를 SAP AI Core와 동기화하려고 하면 응답에 빈 필드가 표시됩니다.

다음을 확인하세요.

- GitHub 비밀번호가 아닌 GitHub 개인 액세스 토큰을 사용하고 있는지 확인하세요.
이미 개인용 액세스 토큰을 사용하고 있는 경우 다음과 같이 진행하세요.
 - 엔드포인트를 사용하여 SAP AI Core 인스턴스에서 모든 알 수 없는 애플리케이션을 삭제합니다.
`{{apiurl}}/v2/admin/applications/{{appName}}` 삭제
 - 엔드포인트를 호출하여 SAP AI Core에서 GitHub 리포지토리를 오프보딩합니다.
`{{apiurl}}/v2/admin/repositories/{{repositoryName}}` 삭제
 - GitHub 개인 액세스 토큰을 생성하고 이를 사용하여 이전과 같이 SAP AI Core에 git repo를 온보딩합니다. 을 위한 자세한 내용은 [개인 액세스 토큰 만들기를](#) 참조하세요..

4. 애플리케이션을 다시 동기화하세요.

상위 주제: [문제 해결 \[페이지 260\]](#)

관련 정보

[리포지토리 \[페이지 261\]](#)

[구성 \[페이지 261\]](#)

[아티팩트 \[페이지 269\]](#)

[응용 프로그램 \[페이지 269\]](#)

[실행 \[페이지 269\]](#)

[도커 \[페이지 279\]](#)

[배포 \[페이지 270\]](#)

17 서비스 오프보딩

테넌트 오프보딩은 고객이 하위 계정을 삭제할 때 발생합니다. SAP AI Core는 하위 계정 삭제 이벤트를 폴링하고 필요한 프로비저닝 해제 및 삭제 활동을 수행합니다.

참고

서비스 인스턴스를 삭제해도 데이터와 리소스는 삭제되지 않습니다(서비스 인스턴스를 기반으로 격리하지 않기 때문). 하위 계정을 유지하면서도 SAP AI Core 프로비저닝을 해제하려면 중간 규모 지원 티켓을 생성하세요. [Service Offboarding](#)이라는 제목의 구성 요소 CA-ML-AIC에서 데이터와 리소스를 수동으로 삭제하도록 요청합니다.

중요한 면책 조항 및 법률 정보

하이퍼링크

일부 링크는 아이콘 및/또는 마우스 오버 텍스트로 분류됩니다. 이 링크는 추가 정보를 제공합니다.

아이콘 정보:

- 아이콘이 있는 링크  : 귀하는 SAP에서 호스팅하지 않는 웹사이트에 접속하고 있습니다. 해당 링크를 사용함으로써 귀하는 동의하게 됩니다.(귀하의 약관에 달리 명시되지 않는 한). SAP와의 계약에 대한 내용은 다음과 같습니다.
 - 링크된 사이트의 내용은 SAP 문서가 아닙니다. 귀하는 이 정보를 기반으로 SAP에 대한 제품 청구를 추론할 수 없습니다.
 - SAP는 링크된 사이트의 콘텐츠에 동의하거나 동의하지 않으며, 가능성과 정확성을 보증하지 않습니다. SAP는 다음에 대해 책임을 지지 않습니다.
 - SAP의 중과실이나 고의적인 위법 행위로 인해 손해가 발생한 경우를 제외하고 해당 콘텐츠의 사용으로 인해 발생한 손해.
- 아이콘이 있는 링크  : 특정 SAP 제품 또는 서비스에 대한 문서를 떠나 SAP가 호스팅하는 웹 사이트에 들어가고 있습니다. 사용하여 그러한 링크를 사용하는 경우, 귀하는 (SAP와의 계약에서 달리 명시적으로 명시하지 않는 한) 본 링크를 기반으로 SAP에 대한 제품 청구를 추론할 수 없다는 점에 동의합니다.

외부 플랫폼에서 호스팅되는 비디오

일부 비디오는 제3자 비디오 호스팅 플랫폼을 가리킬 수 있습니다. SAP는 이러한 플랫폼에 저장된 비디오의 향후 가용성을 보장할 수 없습니다. 게다가, 어떤 이러한 플랫폼에서 호스팅되는 광고 또는 기타 콘텐츠(예: 추천 동영상 또는 동일한 사이트에서 호스팅되는 다른 동영상으로 이동)는 SAP의 통제 또는 책임.

베타 및 기타 실험적 기능

실험적 기능은 SAP가 향후 커스터마이징에 대해 공식적으로 제공하는 범위에 포함되지 않습니다. 이는 실험적 기능이 다음에 의해 변경될 수 있음을 의미합니다.
어떤 이유로든 통지 없이 언제든지 SAP, 실험적 기능은 생산적인 용도로 사용되지 않습니다. 시연, 테스트, 검사, 평가 또는 기타 방법으로 사용할 수 없습니다.
실제 운영 환경에서 실험 기능을 사용하거나 충분히 백업되지 않은 데이터를 사용합니다.
실험적 기능의 목적은 조기에 피드백을 받아 고객과 파트너가 이에 따라 향후 제품에 영향을 미칠 수 있도록 하는 것입니다. 귀하의 피드백(예: SAP 커뮤니티)을 통해 귀하는 기여 또는 파생 저작물에 대한 저작 재산권이 SAP의 독점 재산으로 유지된다는 점에 동의합니다.

예제 코드

모든 소프트웨어 코딩 및/또는 코드 조각은 예입니다. 생산적인 용도로 사용되지 않습니다. 예제 코드는 구문을 더 잘 설명하고 시각화하기 위한 것입니다.
그리고 표현 규칙, SAP는 예제 코드의 정확성과 완전성을 보증하지 않습니다. SAP는 다음의 사용으로 인해 발생한 오류나 손해에 대해 책임을 지지 않습니다.
SAP의 중과실이나 고의적인 위법 행위로 인해 손해가 발생한 경우를 제외하고 예제 코드입니다.

편견 없는 언어

SAP는 다양성과 포용의 문화를 지원합니다. 가능할 때마다 우리는 문서에서 편견 없는 언어를 사용하여 모든 문화, 민족, 성별, 능력.

© 2024 SAP SE 또는 SAP 계열사. 판권 소유.

본 출판물의 어떠한 부분도 SAP SE 또는 SAP 계열사의 명시적인 허가 없이는 어떤 형태나 목적
으로든 복제되거나 전송될 수 없습니다. 여기에 포함된 정보는 사전 통지 없이 변경될 수 있습니
다.

SAP SE 및 그 유통업체가 판매하는 일부 소프트웨어 제품에는 다른 소프트
웨어 권리업체의 독점 소프트웨어 구성 요소가 포함되어 있습니다.
국가별 제품 사양은 다를 수 있습니다.

이러한 자료는 SAP SE 또는 SAP 계열사가 어떠한 종류의 전술이나 보증 없이 정보 제공 목적
로만 제공하며, SAP 또는 계열사는 자료와 관련된 오류나 누락에 대해 책임을 지지 않습니다.
SAP 또는 SAP 계열사 제품 및 서비스에 대한 유일한 보증은 해당 제품 및 서비스와 함께 제공
되는 명시적 보증서에 명시된 것입니다. 본 문서의 어떤 내용도 추가 보증을 구성하는 것으로 해
석되어서는 안 됩니다.

여기에서 언급된 SAP 및 기타 SAP 제품과 서비스 및 해당 로고는 독일 및 기타 국가에서 SAP
SE(또는 SAP 계열사)의 상표 또는 등록 상표입니다. 언급된 기타 모든 제품 및 서비스 이름은
해당 회사의 상표입니다.

<https://www.sap.com/about/legal/trademark.html>을 참조하십시오. 추가 상표 정보 및 고지 사항을 확인하세요.