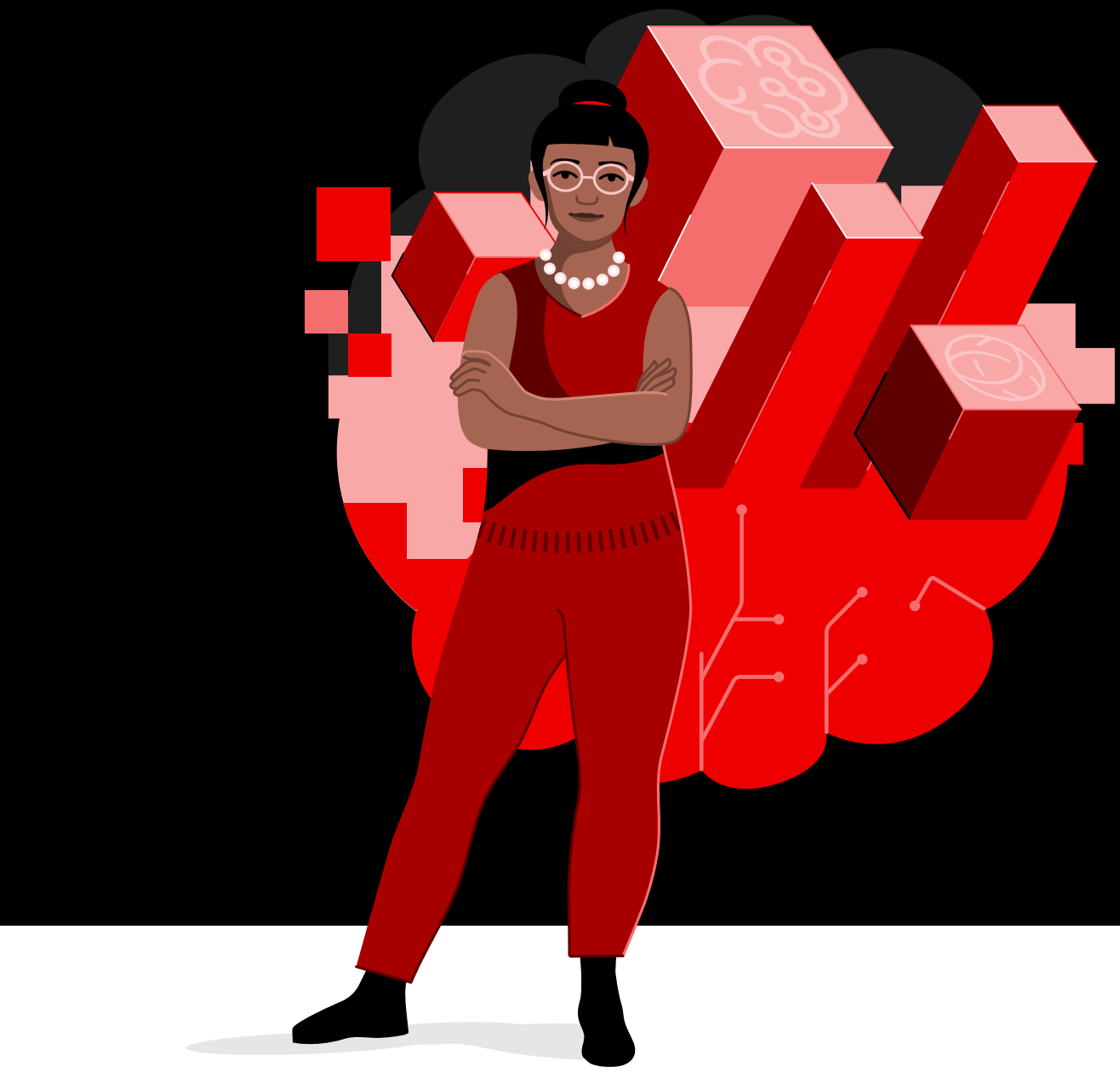
Red Hat



생성형 AI를 위한 기반 구축

시 주요 고려 사항

목차

비즈니스 혁신을 향한 새로운 가능성 알아보기

생성형 AI(인공지능)는 빠르게 변화하는 시장에서 혁신적인 제품을 만들고 프로세스를 최적화하며 경쟁 우위를 확보하고자 하는 조직을 위한 강력한 툴입니다. 딥러닝과 신경망의 발전을 바탕으로 데이터 처리는 물론 새롭고 독창적인 콘텐츠를 생성함으로써 예측형 AI의 기능을 뛰어넘습니다. 생성형 AI는 사람과 머신의 협업을 재구성하고 문제 해결에 대한 새로운 접근 방식을 장려하며 산업 전반에서 상당한 비즈니스 이익을 제공합니다.

전 세계적으로 기업과 조직은 생성형 AI 기술을 사용하여 새롭고 혁신적인 애플리케이션을 구축하고 있습니다. 실제로 39%는 현재 생성형 AI 기술에 투자하고 있고 37%는 잠재적 활용 사례를 탐색하고 있습니다.[1] 다음은 오늘날 생성형 AI의 수많은 활용 사례 중 일부입니다.

- **복잡한 시나리오에 대한 예측 생성.** 생성형 AI는 과거 데이터를 분석하고 패턴을 식별하며 정확한 예측을 개발하여 전략적 계획 수립과 위험 관리에 도움을 줄 수 있습니다.
- **개인화된 마케팅 개발.** 생성형 AI는 데이터를 분석하여 고객의 선호 사항과 행동을 파악함으로써 참여율과 전환율을 극대화하는 개인화된 마케팅 자료(이메일, 광고, 프로모션 등)를 생성할 수 있습니다.
- **고객 서비스 자동화 및 개인화.** 생성형 AI는 지능형 챗봇과 가상 어시스턴트의 기반으로, 고객 문의와 상호 작용에 자동으로 응답함으로써 개인화되고 효율적인 고객 서비스를 제공할 수 있습니다.

1 IDC Web Conference Proceeding. 'Unlocking Business Success with Generative AI'. Document #US50789223. 2023년 6월.

생성형 AI로 인한 새로운 고려 사항

생성형 AI의 장점과 단점이 계속 드러나고 있지만, 대부분의 기업과 조직은 현재 이러한 신기술에 투자하기를 원합니다. 생성형 AI와 관련된 문제를 파악하면 기업이 명확한 윤리 지침과 개발 체계를 확립하고, 정부 및 업계 규정을 준수하며, 잠재적 문제를 탐지하여 해결하는 데 도움이 될 수 있습니다.

- **데이터 프라이버시.** 생성형 AI 모델을 민감한 데이터나 개인 데이터를 사용해 훈련하면 개인 프라이버시 보호 문제가 발생하여 프라이버시에 대한 우려가 제기됩니다.
- **데이터 소유권.** 독점 모델(또는 독점 데이터를 사용하여 사전 훈련한 모델)을 사용하면 소송으로 이어질 가능성이 있는 데이터 소유권 문제가 발생합니다.
- **편견 및 공정성.** 생성형 AI 툴에서 제공하는 응답은 위험한 고정 관념과 혐오 발언 등 인간의 편향을 반영하는 것으로 나타났습니다.
- **윤리적 사용.** 생성형 AI 모델은 개인정보 침해 및 잘못된 정보 캠페인과 같은 악의적인 활동에 사용되는 합성 콘텐츠와 딥페이크를 생성할 수 있습니다.
- **설명 가능성 및 해석 가능성.** 생성형 AI 툴의 투명성이 부족하면 모델 출력을 해석, 이해, 설명하기 어렵고 그 결과 잘못되었거나 조작된 정보에 대한 책임 부족으로 이어집니다.
- **의도하지 않은 결과.** 생성형 AI의 자율적 특성은 의도하지 않은 결과를 초래하여 사람과 조직에 실질적인 해를 끼칠 수 있습니다.
- **규제 관련 과제.** 생성형 AI 기술이 빠르게 발전함에 따라 규제 체계가 그 속도를 따라가지 못하여 책임감 있고 윤리적인 사용을 보장하는 지침을 만들고 시행하기 어려울 수 있습니다.
- **에너지 소비.** AI 모델 훈련은 컴퓨팅 리소스를 많이 사용하고 에너지 수요가 많아 환경에 미치는 영향과 지속 가능성에 관한 우려가 제기되고 있습니다.

이 e-book에서는 생성형 AI 이니셔티브를 지원하기 위해 신뢰할 수 있는 인프라 기반 구축과 관련된 주요 고려 사항을 검토합니다.

생성형 AI를 위한 준비

IDC는 'Unlocking Business Success with Generative AI'(생성형 AI를 통한 비즈니스 성공 실현)에서 조직이 생성형 AI 이니셔티브를 위해 준비할 수 있도록 해당 조치를 권장합니다.[2]

- 비즈니스 요구 사항을 충족하는 우선 활용 사례를 위한 **애자일 실험 환경을 조성합니다.**
- 악의적인 행동을 억제하는 책임 있는 사용을 위한 **기업 정책을 개발합니다.**
- 선제적 변경 관리에 참여하고 생성형 AI가 **인력에 미치는 영향을 평가합니다.**
- AI 인프라를 **위해 신뢰할 수 있는 기술 벤더** 및 서비스 공급자와 협력합니다.
- 채용, 교육 또는 전문 서비스 지원을 통해 **적합한 엔지니어링 기술을 확보합니다.**

2 IDC 웹 컨퍼런스 진행. 'Unlocking Business Success with Generative AI'. Document #US50789223. 2023년 6월.

생성형 AI를 위한 기반 구축 시 고려 사항

생성형 AI 이니셔티브를 위해 선택하는 기술 기반은 손쉬운 채택과 전반적인 성공에 큰 영향을 미칠 수 있습니다. 이 장에서는 생성형 AI 기반과 관련된 주요 고려 사항을 설명합니다.

고려 사항 1: 입증된 툴셋을 통한 빌드

생성형 AI 모델을 기반으로 애플리케이션을 개발하는 일은 복잡한 태스크일 수 있습니다. 올바른 툴셋(오픈소스 프로젝트 및 상용 솔루션 기반의 언어, 프레임워크, 런타임 포함)을 사용하면 모델 튜닝 속도를 높이고 애플리케이션 개발과 배포를 간소화할 수 있습니다.

혁신적인 AI 솔루션을 빠르고 효율적으로 개발하려면 선호하는 툴셋을 제공하는 AI 기반을 선택하세요. 대화형 인터페이스를 통해 탐색적 데이터 사이언스, 교육, 튜닝이 지원되어 협업을 단순화할 수 있습니다. 사전 통합된 툴셋과 셀프 서비스 기능은 IT 운영을 간소화하면서 환경 전반에서 이식성과 일관성을 유지하는 데 도움이 됩니다.

고려 사항 2: 신속한 모델 미세 튜닝

생성형 AI 모델을 훈련하는 프로세스에는 비용과 시간이 많이 소요되기 때문에 대부분의 기업과 조직은 범용 데이터를 기반으로 사전 훈련된 기반 모델을 사용하여 AI 솔루션을 구축합니다. 그런 다음 데이터 과학자가 다양한 영역별 데이터를 사용하여 전문적인 태스크를 수행하도록 기반 모델을 조정합니다. 그러나 미세 튜닝에는 여전히 컴퓨팅 리소스가 많이 사용되며 강력한 프로세서와 분산형 하이브리드 클라우드 인프라가 필요합니다.

분산형 워크로드 관리 기능과 오케스트레이션 기능을 갖추고 하이브리드 클라우드 환경 전반에서 모델 크기, 데이터 용량, 기간에 상관없이 훈련 실행을 배포하는 AI 플랫폼을 확인해 보세요. 온사이트 데이터센터의 기반 모델을 미세 튜닝하는 옵션을 사용하면 제한된 모델에 대한 기술 및 규제 요구 사항 컴플라이언스가 간소화됩니다. 배치 훈련 기능을 사용하면 미세 튜닝 워크로드를 선점하며 리소스를 더 쉽게 공유하고 관리할 수 있습니다.

모델 미세 튜닝에 대한 대안

연구원들은 기반 모델을 더 빠르고 효율적으로 튜닝하는 방법을 연구하고 있습니다. **RAG(검색 증강 생성, Retrieval-augmented generation)**는 생성형 AI 모델에 가장 정확한 최신 정보를 제공하기 위해 외부 소스(예: 내부 데이터베이스, 기업 인트라넷, 인터넷)에서 팩트를 검색하는 AI 프레임워크입니다.

프롬프트 튜닝에서는 AI 모델에 원하는 의사 결정으로 모델을 유도하는 단서 또는 프론트엔드 프롬프트(추가 단어 또는 AI 생성 숫자 포함)를 제공하므로 제한된 데이터만을 사용하는 조직에서 특정 태스크에 맞게 기반 모델을 조정할 수 있습니다.

고려 사항 3: 효율적인 모델 서빙

IT 운영팀에게는 생성형 AI 솔루션을 통해 뛰어난 사용자 경험을 제공하는 일이 어려울 수 있습니다. 다양한 애플리케이션 수요로 인해 확장 가능한 인프라와 자동화된 관리가 필요합니다. 효율적인 모델 배포를 위해서는 성능을 모니터링하고 신속하게 이전 버전으로 되돌릴 수 있는 기능이 필요합니다. 또한 AI 솔루션은 방대한 양의 데이터를 처리하기 때문에 환경 전반에서 엄격한 보안 표준을 적용해야 합니다.

온사이트 인프라와 퍼블릭 클라우드 리소스, 엣지 장치를 비롯하여 하이브리드 클라우드 전반에서 생성형 AI 모델과 애플리케이션을 배포하고 확장할 수 있는 플랫폼을 고려해 보세요. 온사이트 또는 격리된 환경에서 생성형 AI 모델을 서빙하는 옵션을 사용하면 공개적으로 제공되는 모델을 재훈련하는 데 독점 데이터가 사용되지 않도록 할 수 있습니다. 또한 카나리아 롤아웃 및 설명 가능성 툴을 지원하여 모델 응답의 일관성과 신뢰성 개선에 도움이 됩니다.

고려 사항 4: 라이프사이클 관리 자동화

CI/CD(지속적 통합/지속적 제공) 파이프라인은 생성형 AI 솔루션을 자동으로 배포하고 관리할 수 있습니다. 신속하고 단계적인 변경을 통해 모델과 애플리케이션을 재훈련하고 업데이트함으로써 개발 속도를 높이고 모델의 성능을 개선할 수 있습니다. 그러나 AI 파이프라인은 표준 CI/CD 워크플로우보다 훨씬 더 복잡합니다. 대부분 데이터 추출, 훈련, 미세 튜닝, 검증, 재훈련과 같은 단계가 추가로 포함되기 때문입니다.

Tekton 및 Jenkins와 같은 CI/CD 툴을 기반으로 AI 파이프라인을 생성하여 기존 DevOps 워크플로우에 통합할 수 있는 기반을 선택하세요. 생성형 AI 모델을 빠르고 효율적으로 개발, 훈련, 모니터링, 재훈련할 수 있습니다. ArgoCD와 같은 **GitOps** 지속적 제공(CD) 툴을 사용하면 복잡한 AI 솔루션 배포를 코드로 정의하고 자동화하여 일관된 모델과 애플리케이션을 제공할 수 있습니다.

생성형 AI를 위한 컨테이너

컨테이너 및 **쿠버네티스** 기술은 애자일 배포, 관리, 확장성을 제공하여 생성형 AI 솔루션의 클라우드 네이티브 개발을 가속화합니다. 온사이트 데이터센터, 퍼블릭 클라우드, 엣지 장치 전반에서 온디맨드 방식으로 환경을 프로비저닝합니다. 물리 인프라와 가상 인프라에서 컨테이너 인스턴스를 자동으로 생성, 배포, 확장, 관리합니다. 또한 오픈소스 공급업체와 상용 공급업체로 구성된 강력한 에코시스템에서 제공하는 구성 요소와 데이터 저장소를 생성형 AI 솔루션에 통합합니다. **AI용 컨테이너의 이점을 자세히 알아보세요.**

고려 사항 5: 일관된 모델 모니터링

생성형 AI는 사람과 기업에 실질적이고 중대한 영향을 미칠 수 있습니다. 모델 동작을 추적하면 의사 결정과 타당성을 분석하고 성능 저하를 식별하여 문제가 있는 동작을 즉시 보고할 수 있습니다. 이러한 정보를 기반으로 한 효과적인 모델 거버넌스를 통해 모델이 프로덕션 환경에서 편견이 없고 공정하며 올바른 정보를 사용하여 대응하도록 할 수 있습니다.

중앙집중식 모니터링 기능을 갖추어 생성형 AI 모델을 조사하고 유지 관리 및 수정하는 데 도움이 되는 편향 및 데이터 변동 지표, 이상 탐지, 지점별 설명 가능성을 제공하는 AI 기반을 살펴보세요. 프로덕션 환경에서 지속적인 자동 모니터링을 수행하면 기업의 모델 거버넌스 표준 컴플라이언스가 향상됩니다. 또한 사용자 친화적인 툴 인터페이스와 사람이 읽을 수 있는 비기술적 리포트는 모델을 책임감 있게 사용하고 유지 관리하도록 장려합니다.

생성형 AI 모델의 주요 개념

- **편견**은 특정 집단을 선호하거나 고정관념에 부합하는 반응을 생성하는 것을 포함하여 생성된 출력의 공정성, 포용성, 윤리에 영향을 미치는 모델 동작의 패턴을 나타냅니다.
- **데이터 변동**은 훈련 데이터의 통계적 속성이 시간에 따라 변할 때 발생하며, 이로 인해 모델 성능이 저하되고 응답의 정확도와 관련성이 떨어집니다.
- **이상 탐지**는 훈련 중 확인된 예시와 다르거나 일반적이지 않은 모델 동작을 확인하고 보고하는 프로세스입니다.
- **지점별 설명 가능성**은 모델이 특정 출력을 생성하는 이유를 이해하여 투명성이 중요한 애플리케이션에 가시성을 제공하는 기능입니다.

고려 사항 6: 파트너 에코시스템 활용

생성형 AI 솔루션은 혁신적인 사용자 경험을 구현하기 위해 다양한 통합 구성 요소가 필요합니다. 신뢰할 수 있는 벤더로 구성된 협업 에코시스템에서 제공하는 기술을 올바르게 조합하면 애플리케이션 개발 속도를 높이고 편향 및 데이터 변동 과제를 해결하여 솔루션 전반에서 일관되고 안정적인 성능을 보장할 수 있습니다.

광범위하고 인증된 파트너 에코시스템을 통해 생성형 AI 모델과 애플리케이션을 개발하고 배포하는 데 필요한 완벽한 솔루션을 제공하는 플랫폼 벤더를 찾으세요. 데이터 통합 및 준비부터 모델 훈련 및 서빙에 이르기까지 광범위한 구성 요소를 통해 AI 솔루션을 더 빠르고 효율적으로 개발하고 배포할 수 있습니다. 상호 운용성이 입증된 인증된 솔루션을 선택하면 IT 지원 요청을 줄이고 생산성을 높일 수 있습니다.

고려 사항 7: 플랫폼 전문가와의 협력

생성형 AI 솔루션을 효과적으로 배포하고 관리하려면 전문적인 지식과 경험이 필요합니다. 확장성 요구 사항, 안정성 문제, 기존 시스템과의 통합으로 인해 프로덕션 배포가 복잡해질 수 있습니다. 컴퓨팅 리소스를 비효율적으로 사용하면 불필요한 비용이 발생할 수 있습니다. 또한 보안 표준, 개인 정보 취급 방침, AI 규제 프레임워크를 준수하지 않으면 의도하지 않은 결과를 초래할 수 있습니다.

생성형 AI 솔루션 구축에 필요한 포괄적인 지원과 지침을 제공하는 전문가 팀이 있는 벤더를 선택하세요. 예를 들어 전담 엔지니어는 툴과 리소스, 지식을 통해 전체 플랫폼을 지원함으로써 AI 프로젝트의 속도를 높일 수 있습니다. 전문 컨설턴트는 배포 문제를 해결하고, 인프라 효율성을 최적화하며, AI 솔루션 전반의 상호 운용성을 보장할 수 있습니다. 또한 전문 교육 서비스를 통해 새로운 생성형 AI 프로젝트를 더 빠르게 시작하는 데 필요한 지식과 전문 지식을 얻을 수 있습니다.

협업이 필요한 생성형 AI

성공적인 생성형 AI 프로젝트의 핵심은 다양한 역량을 갖춘 팀을 구성하는 것입니다.[3]

- **비즈니스 리더**는 솔루션을 사용하거나 솔루션의 영향을 받는 사람들을 대표합니다.
- **AI 전문가**는 생성형 AI 모델을 튜닝, 유지 관리, 업데이트합니다.
- **데이터 과학자**는 정확하고 편향되지 않은 훈련 데이터를 전처리하여 모델에 제공합니다.
- **윤리 및 컴플라이언스 담당자**는 생성형 AI 이니셔티브가 규정을 준수하는지 확인합니다.
- **IT 운영 전문가**는 솔루션을 기존 인프라에 통합하고 보안 정책을 적용합니다.

3 Kearney. 'Standing up tiger teams to tackle generative AI complexity', 2023년 11월.

유연한 개방형 파운데이션을 통한 빠른 혁신

Red Hat은 완벽한 기술 포트폴리오, 입증된 전문성, 전략적 파트너십을 제공하여 생성형 AI 목표 실현을 지원합니다. 생성형 AI 모델과 애플리케이션을 개발하고 배포하는 데 필요한 기반은 물론 신속한 도입을 위한 서비스와 교육도 제공합니다.

Red Hat® OpenShift®는 클라우드 네이티브 혁신을 위한 엔터프라이즈 수준의 통합 애플리케이션 플랫폼입니다. 온디맨드 컴퓨팅 리소스, 하드웨어 가속 지원, 온사이트와 퍼블릭 클라우드 및 엣지 환경 전반의 일관성을 통해 팀이 성공하는 데 필요한 속도와 유연성을 제공합니다. Red Hat OpenShift를 사용하면 데이터 과학자, 데이터 엔지니어, 개발자를 위한 셀프 서비스 플랫폼을 생성하여 지능형 애플리케이션을 신속하게 개발할 수 있습니다. 또한 협업 기능 덕분에 팀은 컨테이너화된 모델링 결과를 생성하고 이 결과를 동료 및 개발자와 동일한 방식으로 공유할 수 있습니다.

Red Hat OpenShift AI는 Red Hat OpenShift를 기반으로 구축되며, 모델과 애플리케이션을 구축, 훈련, 미세 튜닝, 배포, 모니터링하는 포괄적인 플랫폼을 제공할 뿐만 아니라 최신 생성형 AI 솔루션의 워크로드 및 성능 요구 사항을 충족합니다. 팀은 NVIDIA, Intel, Starburst, Anaconda, IBM, Run:ai, Pachyderm과 같은 파트너의 주요 인증 제품을 통합하는 일관된 협업 환경을 통해 실험 단계에서 프로덕션으로 신속하게 이동할 수 있습니다. Red Hat OpenShift AI는 Red Hat의 기술 에코시스템과 함께 하이브리드 클라우드 전반에서 혁신적인 생성형 AI 솔루션을 더 신속하게 개발하고 배포하는 데 필요한 구성 요소와 기능을 제공합니다.

IBM watsonx.ai AI studio는 지능형 애플리케이션에 필요한 생성형 AI 기능과 함께 다양한 모델 및 배포 옵션을 제공합니다. 워크로드가 있는 모든 곳에 모델(오픈소스, 타사, IBM 개발 기반 모델 포함)을 배포하여 AI 솔루션의 성능과 효율성을 높이세요. 또한 엔터프라이즈 관련 데이터를 기반으로 훈련된 **IBM 개발 기반 모델**에서는 생성형 AI 솔루션이 해당 비즈니스 영역의 미묘한 차이를 이해하여 경쟁 우위를 제공합니다.

IBM watsonx Code Assistant가 통합된 Red Hat Ansible® Lightspeed는 팀이 더욱 효율적으로 자동화 콘텐츠를 제작, 채택, 유지 관리하도록 설계된 생성형 AI 서비스입니다. IBM watsonx Code Assistant와 연결된 Red Hat Ansible Lightspeed는 자연어 프롬프트를 사용하여 자동화 아이디어를 Ansible 코드로 전환할 수 있도록 지원합니다. 이 서비스를 사용하면 생산성을 높이고 조직 전반에서 자동화에 대한 접근성을 높일 수 있습니다.

생성형 AI를 시작할 준비가 되셨나요?

생성형 AI는 독창적인 콘텐츠를 생성하고 애플리케이션 및 기술과 상호 작용하는 방식을 바꾸는 강력한 툴입니다.

Red Hat은 기술과 전문성, 파트너십을 통해 팀이 투명성과 제어 권한을 가지고 AI 애플리케이션과 ML 모델을 구축하고 배포할 수 있는 공통 기반을 제공합니다. 실제로 다른 오픈소스 소프트웨어의 유용성을 개선하는 데에도 Red Hat의 AI 툴과 플랫폼을 사용합니다. Red Hat의 파트너 통합을 통해 사용자는 Red Hat OpenShift AI와 같은 오픈소스 플랫폼과 연동되도록 구축한 신뢰할 수 있는 AI 툴 에코시스템을 활용할 수 있습니다.

Red Hat OpenShift AI에 대해 자세히 알아보고 체험판을 무료로 사용해 보세요.

지금 바로 Red Hat Consulting 시작하기

Red Hat 전문가와 협력해 AI/ML 프로젝트를 신속하게 시작하세요. Red Hat은 조직에서 AI/ML을 더욱 신속하게 도입할 수 있도록 컨설팅과 교육 서비스를 제공합니다.

- AI/ML 서비스 자세히 알아보기: red.ht/aiml-consulting
- 무료 디스커버리 세션을 예약해 보세요. redhat.com/ko/services/consulting

1057715_0324_KVM

Red Hat