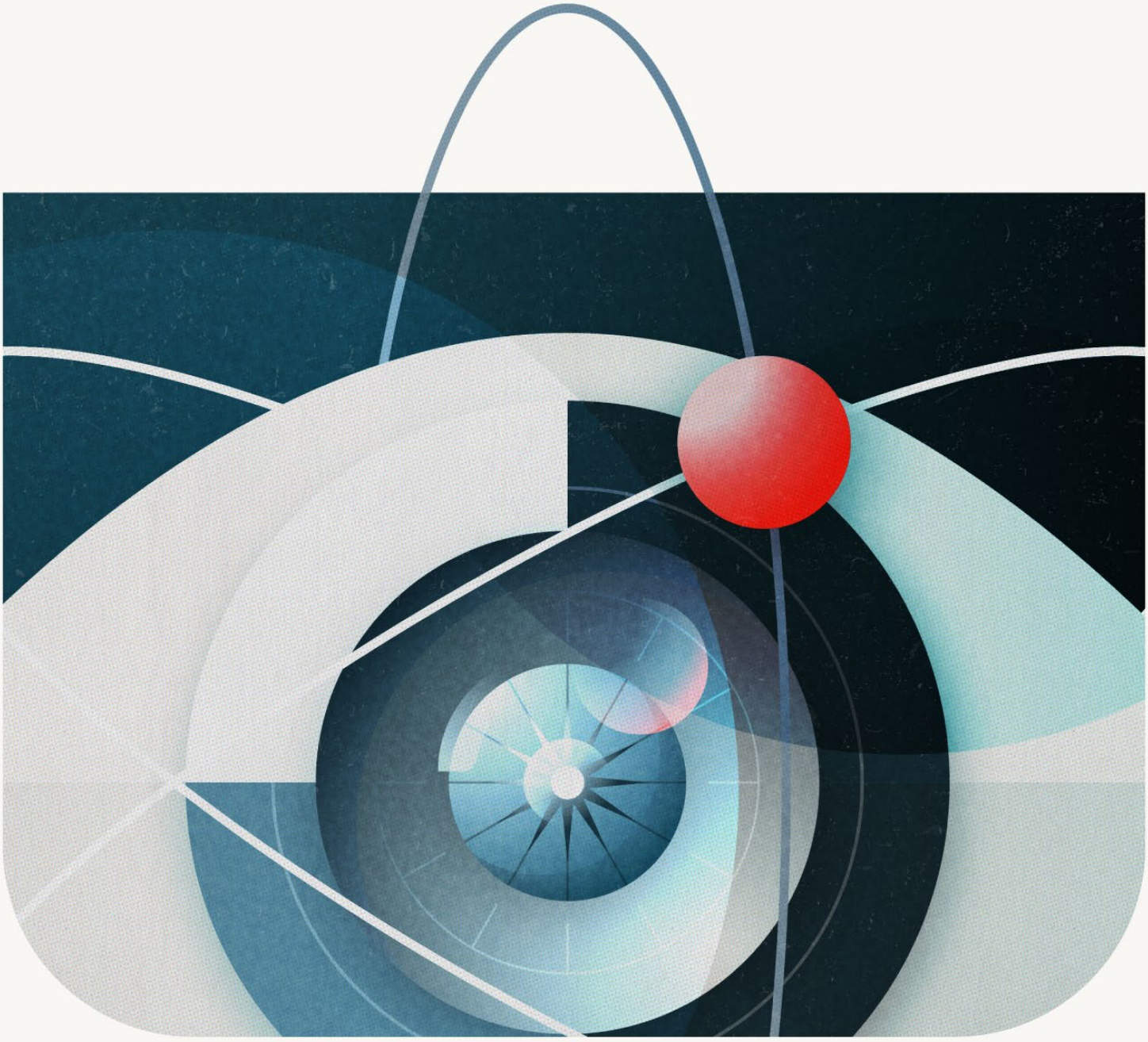
databricks

데이터 + AI 현황

데이터 인텔리전스
및 LLM 도입을 위한
무한경쟁



모든 조직은 데이터와 AI의 보편화를 서두르고 있습니다

소개

생성형 AI(GenAI)는 혁신, 창의성 및 생산성의 새로운 시대를 열어가고 있습니다. GenAI가 주류로 진입한 지 18개월 만에 전 세계 기업들이 조직을 혁신하기 위해 GenAI에 투자하고 있습니다. 기업들은 사용자에게 고품질의 GenAI 경험을 제공하는 데 있어 데이터가 핵심이라는 사실을 깨닫고 있습니다. 현대의 비즈니스 리더들은 다음과 같은 시급한 질문에 대한 답을 찾아야 합니다. *GenAI 경험을 제공하는 가장 빠르고 효과적인 방법은 무엇일까요?*

사일로화된 데이터 및 AI 플랫폼에서는 자연어를 사용해 데이터 관련 질문을 하든, 데이터로 지능형 앱을 구축하든, GenAI 프로젝트를 가속화하기 어렵습니다. Databricks는 Data Intelligence Platform이 조직 전반을 급속하게 보편화할 것이라고 믿습니다. 이 새로운 범주의 데이터 플랫폼은 GenAI를 사용하여 데이터를 더 쉽게 보호하고 활용할 뿐만 아니라 데이터에서 가치를 창출하기 위한 기술적 장벽을 낮춥니다. Databricks의 고객들은 이미 AI 도입을 가속화하고 있습니다.

데이터 및 AI 현황 *보고서에서는 조직이 데이터 및 AI 이니셔티브의 우선순위를 어떻게 정하고 있는지에 대한 스냅샷을 제공합니다.* 공유된 인사이트는 현재 Fortune 500대 기업 중 300개 이상을 포함해 Databricks Data Intelligence Platform을 사용 중인 10,000개 고객사의 데이터를 바탕으로 합니다. 가장 혁신적인 조직들이 머신 러닝을 성공적으로 활용하고, GenAI를 채택하여 진화하는 거버넌스 요구 사항에 대응하는 방법을 알아보세요.

이 보고서는 기업이 진화하는 엔터프라이즈 AI 시대에 효과적인 데이터 전략을 개발하는 데 활용할 수 있도록 작성되었습니다.

주요 결과

올해 11배 더 많은 AI 모델이 프로덕션에 활용되고 있습니다.

수년 동안 AI 실험에 매달린 기업들이 이제 1년 전보다 훨씬 더 많은 모델을 실제 환경에 배포하고 있습니다.

평균적으로 조직은 3배 더 효율적으로 모델을 프로덕션에 활용하고 있습니다.

자연어 처리는 가장 많이 사용되고 가장 빠르게 성장하는 머신 러닝 분야였습니다.

GenAI를 활용하는 기업의 70%가 기본 모델을 보강하기 위해 툴과 벡터 데이터베이스를 사용합니다.

통합된 지 1년도 채 되지 않아 LangChain은 가장 널리 사용되는 데이터 및 AI 제품 중 하나가 되었습니다.

다수의 기업들이 검색 증강 생성(Retrieval Augmented Generation, RAG)을 통해 비공개 데이터로 LLM을 맞춤 구성하는 데 집중하고 있습니다.

RAG에는 벡터 데이터베이스가 필요하며, 벡터 데이터베이스의 사용량은 전년 대비 377% 증가했습니다. (사용량에는 오픈 소스 및 비공개 LLM이 모두 포함됨)

LLM을 사용하는 기업의 76%가 오픈 소스를 선택하며, 종종 독점 모델과 함께 사용합니다.

많은 기업이 비용, 성능, 레이턴시 측면의 장단점을 고려하여 더 작은 오픈 소스 모델을 선택합니다.

Meta Llama 3는 출시 4주 만에 전체 오픈 소스 모델 사용량의 39%를 차지했습니다.

규제가 엄격한 산업에서 의외로 GenAI를 빠르게 도입했습니다. GPU 사용의 선두주자인 금융 서비스는 6개월 동안 88% 성장률을 기록하며, 가장 빠르게 움직이고 있습니다.

방법론:
Databricks가 이 보고서를 작성한 방법

2024년 데이터 및 AI 현황 보고서는 Databricks 고객이 Databricks Data Intelligence Platform과 광범위한 통합 에코시스템을 사용하는 방식을 기준으로 수집한 후 완전히 집계한 익명화된 데이터를 바탕으로 작성되었습니다.

이 보고서는 머신 러닝의 추세, GenAI의 채택, 통합 및 사용 사례에 초점을 맞춥니다. 이 보고서에서 언급된 고객은 모든 주요 산업을 대표하며, 스타트업부터 세계 최대 기업까지 그 규모가 다양합니다. 달리 명시되지 않은 경우, 이 보고서는 2023년 2월 1일부터 2024년 3월 31일까지의 데이터를 제시하고 분석하며 사용량은 고객 수 기준으로 측정됩니다. 가능한 경우 시간 경과에 따른 성장 추세를 보여주기 위해 전년 대비 증감률을 제공합니다.

머신 러닝

프로덕션에 AI 활용

ML 모델을 프로덕션에 활용하기 위해 경쟁하는 조직

올해 우리는 AI가 실험 단계에서 프로덕션 애플리케이션으로 전환되는 것을 목격하고 있습니다. 머신 러닝(ML)이 발전함에 따라 기업들은 ML 모델 수명주기의 두 가지 뚜렷한 부분을 탐색하는 방법을 학습하고 있습니다. 조직은 먼저 다양한 알고리즘과 초매개변수를 실험적으로 테스트하여 최적의 모델을 찾아내는 과정을 통해 ML 모델을 만든 다음, 이러한 모델을 프로덕션에 적용합니다. 이 과정에서 팀은 실험 단계를 최대한 시간 효율적으로 진행하는 동시에 엄격하게 테스트한 모델만 프로덕션에 활용해야 한다는 두 가지 상반된 목표를 갖게 됩니다.

이전에 프로덕션 환경에 모델을 배포할 때는 서로 다른 데이터 및 AI 플랫폼, 복잡한 배포 워크플로, 거버넌스를 위한 액세스 제어 부족, 모니터링 불가능 등 많은 어려움에 직면했습니다. 이 보고서의 데이터를 통해 기업들이 Data Intelligence Platform을 도입하여 이러한 과제를 어떻게 극복하는지 확인할 수 있습니다.

프로덕션 ML을 가속화하는 기업들

Databricks에서 개발한 오픈 소스 MLOps 플랫폼인 MLflow의 데이터는 고객이 모델을 로깅(실험을 의미)하는 빈도와 모델을 등록(프로덕션을 의미)하는 빈도를 보여줍니다.

이를 통해 더 많은 실험이 이루어지고 있을 뿐만 아니라 기업들이 훨씬 더 효율적으로 ML을 프로덕션에 활용하고 있음을 알 수 있습니다.

실험 모델 대 등록 모델의 비율

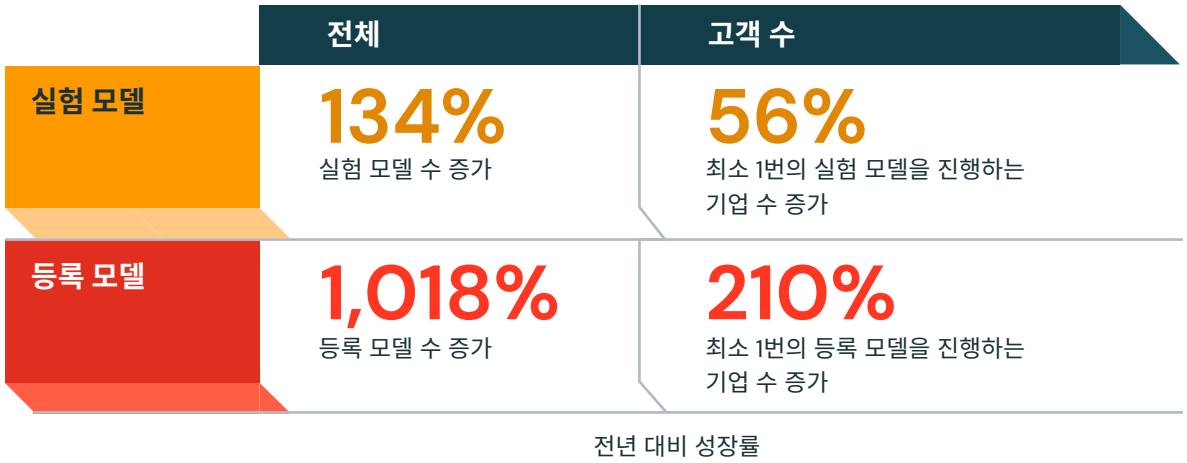
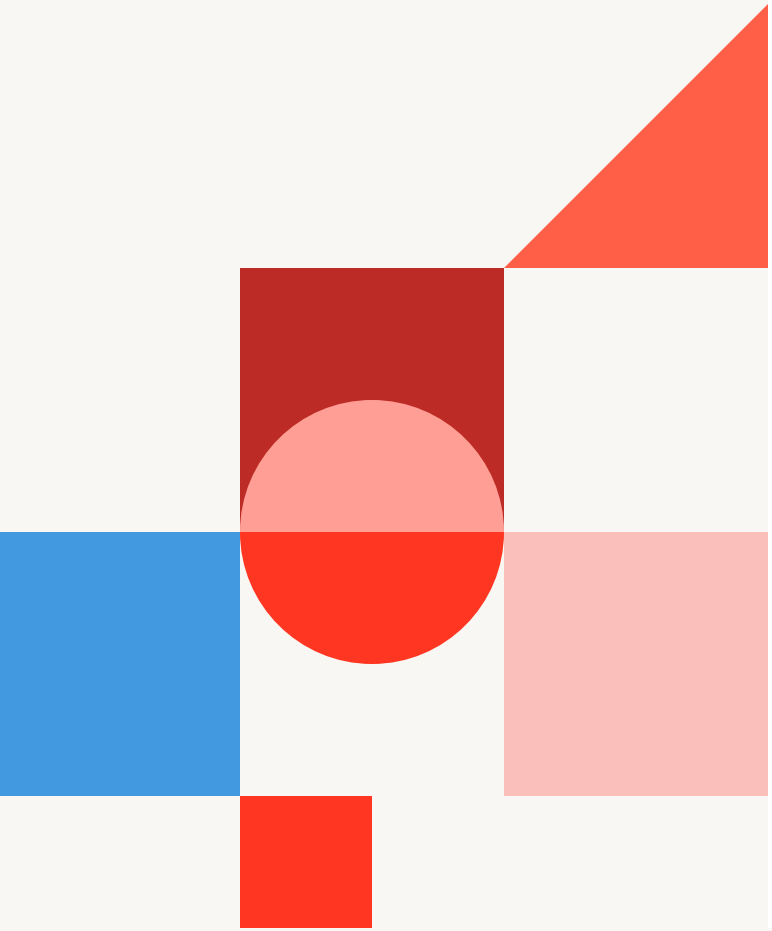


그림 1:
모델 등록의 증가율이 진행 중인 실험의 증가율을 훨씬 앞질렀으며, 이는 기업들이 실험에서 프로덕션으로 전환하고 있음을 의미합니다.

올해 11배 더 많은 AI 모델이 프로덕션에 활용되고 있습니다



모델 수는 측정 가능한 범위에서 크게 증가했습니다.

ML에 투자하는 기업의 수가 급증하고 있습니다

데이터에 따르면 1년 전에 비해 56% 더 많은 기업이 모델 실험을 진행하고 있으며, 210% 더 많은 기업이 모델을 등록하고 있는 것으로 나타났습니다. 이는 작년에 실험에 집중했던 많은 기업들이 이제 프로덕션 단계로 전환했음을 의미합니다.

기업 전반에서 ML 모델의 수가 증가하고 있습니다

수년 동안 실험에 집중해 온 조직들이 이제 프로덕션에 돌입하고 있습니다. 올해에는 1,018% 더 많은 모델이 등록되어 134% 증가한 실험 모델을 훨씬 앞질렀습니다. 이러한 추세는 회사 차원에서도 확인됩니다. 평균적으로 올해 조직은 261% 더 많은 모델을 등록하고 50% 더 많은 모델을 실험하고 있습니다.

핵심 내용

ML은 기업이 혁신을 실현하고 차별화하는 방법의 핵심입니다. 그리고 기업들이 계속해서 자신감을 쌓아감에 따라 이러한 추세는 앞으로도 계속될 것으로 예상됩니다. GenAI의 새로운 분야는 아직 테스트 단계에 있지만, 기업의 관심을 끌고 있습니다.

3배 더 효율적으로 모델을 프로덕션에 활용

ML 효율성은 시간, 비용, 리소스로 측정할 수 있는 실질적인 가치를 지니고 있습니다. 모델 개발과 실험도 중요하지만, 궁극적으로는 이러한 모델을 실제 사용 사례에 배포하여 비즈니스 가치를 창출해야 합니다.

이에 진행 상황을 평가하기 위해 전체 고객에 대한 실험 모델과 등록 모델의 비율을 살펴보았습니다. 2023년 2월 실험 모델과 등록 모델의 비율은 16대 1이었습니다. 즉, 실험 모델 16개당 하나의 모델이 프로덕션용으로 등록된다는 의미입니다. 하지만 데이터 범위의 막바지에는 실험 모델과 등록 모델의 비율이 5:1로 급격히 감소했는데, 이는 프로덕션용으로 등록되는 모델이 3배 증가했음을 의미합니다.

결론적으로 기업들은 실제 가치를 제공하지 않는 실험 모델에 투입되는 리소스를 줄이면서 모델을 프로덕션에 등록하는 효율을 훨씬 더 높이고 있습니다.

업계 수준의 효율성

산업마다 데이터세트, 전략적 목표, 위험 프로파일이 다릅니다. 따라서 ML 실험과 프로덕션의 혼합을 포함하여 ML 접근 방식에도 차이가 있을 것으로 예상할 수 있습니다.

이러한 추세를 더 잘 이해하기 위해 6개의 주요 산업을 분석했습니다.

산업별 실험 모델과 등록 모델의 비율

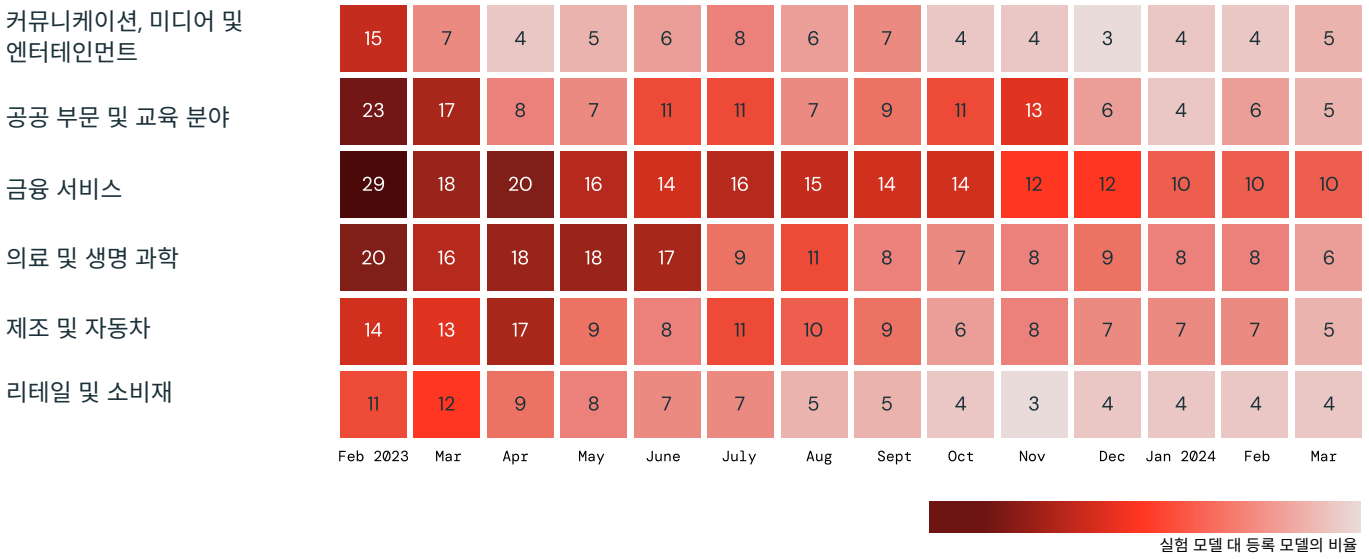


그림 2: 2024년 2월 1일부터 3월 31일 사이에 실험 모델 대 등록 모델의 비율은 꾸준히 감소했으며, 이는 기업들이 더 많은 실험 모델을 프로덕션에 배포했음을 나타냅니다.

참고: 모델 레지스트리 API와 추적의 변경으로 인해 올해의 데이터는 작년의 실험 모델 및 등록 모델의 차트와 상관관계가 없습니다.

가장 효율적인 산업인 소매업은 전체 모델의 25%를 프로덕션에 투입합니다.

리테일 및 소비재는 실험 모델 4개당 프로덕션 모델 1개의 비율로, 주요 산업 중 가장 높은 효율성을 달성했습니다. MIT Technical Review Insights 보고서에 요약된 바와 같이, 리테일 및 소비재는 경쟁 압력과 소비자의 기대치로 인해 오래 전부터 초기 AI 도입을 주도해 왔습니다.

효율성 향상:
금융 서비스는 3배 더 효율적으로 모델을 프로덕션에 활용하고 있습니다.

금융 서비스 부문은 가장 급격한 효율성 향상을 달성했습니다.

금융 서비스는 테스트가 가장 많이 이루어지는 산업입니다. 2023년 초에 금융 서비스는 가장 많은 테스트를 수행하여 29개의 실험 모델당 1개의 모델만 배포했습니다. 하지만 거의 3배 가까이 효율성이 향상되어 10대 1의 비율로 2024년 3월을 마감했습니다. 규제가 엄격한 산업에서 프로덕션 ML의 위험이 더 높기 때문에 긴 테스트 주기는 필수입니다.

올해 더 많은 회사가 더 많은 모델을 프로덕션에 활용할 수 있게 된 이유는 무엇일까요 한 가지 요인은 Data Intelligence Platform의 가용성일 가능성이 높은데, 이는 ML 수명주기 전반에 걸쳐 실무자에게 표준화된 개방형 환경을 제공합니다. 기업은 데이터 거버넌스, 개인정보 보호 및 보안을 보장하면서 데이터 준비 및 모델 교육부터 실시간 서비스, 모니터링에 이르는 각 단계를 하나의 플랫폼에서 실행할 수 있습니다. 이렇게 하면 결과의 품질이 향상되고 프로덕션 준비 상태를 지원할 수 있습니다.

머신 러닝 애플리케이션

NLP의 폭발적인 성장

NLP는 2년 연속 최고의 데이터 사이언스 및 ML 애플리케이션으로 입지를 굳혔습니다.

비정형 데이터는 산업과 지역에 관계없이 어디에나 존재하므로 자연어 처리(NLP) 기술을 통해 의미를 도출하는 것이 필수적입니다. GenAI는 NLP의 주요 사용 사례입니다.

다음 차트는 ML 발전과 AI의 최전선에 있으며 가장 인기 있는 프로그래밍 언어 중 하나로 꾸준히 순위를 유지하고 있는 Python 라이브러리에 초점을 맞추고 있습니다. 데이터에 따르면 조직 내에서 사용되는 상위 5개의 데이터 사이언스 및 ML 애플리케이션(DS/ML)을 결정하기 위해 특수 Python 라이브러리의 사용량을 집계했습니다.

산업별 상위 DS/ML 애플리케이션

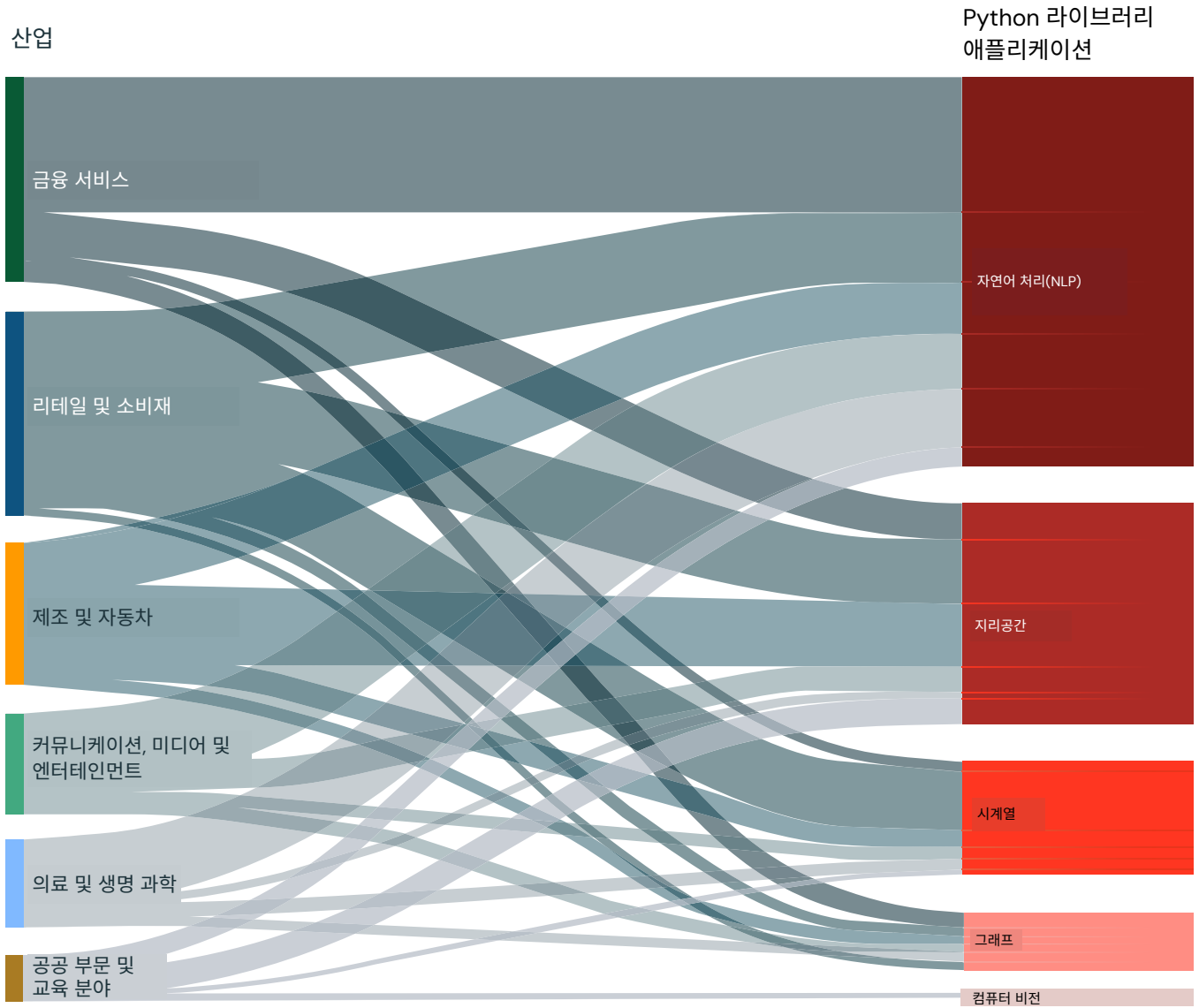


그림 3:
NLP는 가장 일반적으로 사용되는 Python 라이브러리 애플리케이션으로, 모든 주요 산업에서 많이 사용됩니다.

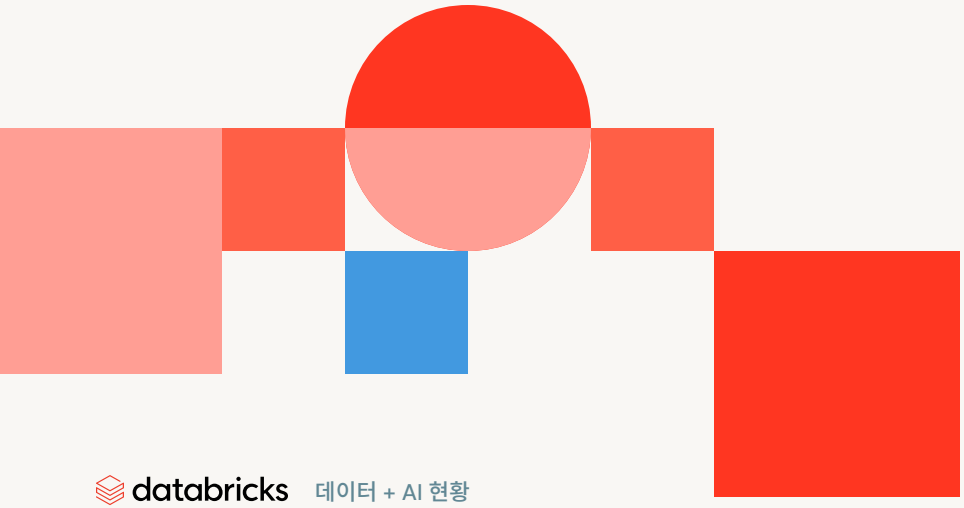
참고: 이 차트는 각 범주에서 ML 라이브러리를 사용하는 고유한 노트북 수를 반영합니다. 데이터 준비 및 모델링을 위한 도구에 사용된 라이브러리는 포함되지 않습니다.

데이터에 따르면, 2년 연속으로 가장 많이 사용되는 DS/ML 애플리케이션은 NLP이며 사용된 특수 Python 라이브러리의 50%가 NLP와 연관되어 있습니다.

또한 데이터 팀은 지리공간 및 시계열 애플리케이션도 앞다투어 활용하고 있습니다. 위치 기반 분석을 통해 사용자 경험을 맞춤화하는 데 자주 사용되는 지리공간 라이브러리는 두 번째로 많이 사용 되는 사용 사례로, Python 라이브러리 사용량의 30%를 차지합니다.

NLP 채택률이 가장 높은 의료 및 생명과학 분야

주요 산업 중 의료 및 생명과학 분야는 NLP 전용 Python 라이브러리 사용 비율이 69%로 가장 높습니다. Arcadia가 HIMSS(Healthcare Information and Management Systems Society)와 함께 실시한 설문조사에 따르면, 의료 산업은 전 세계 데이터 양의 30%를 생성하며 다른 어떤 산업보다 빠르게 성장하고 있습니다. NLP는 임상 연구 분석을 지원하고, 신약 개발 및 출시를 가속화하며, 영업 및 마케팅의 상업적 효과를 높일 수 있습니다.

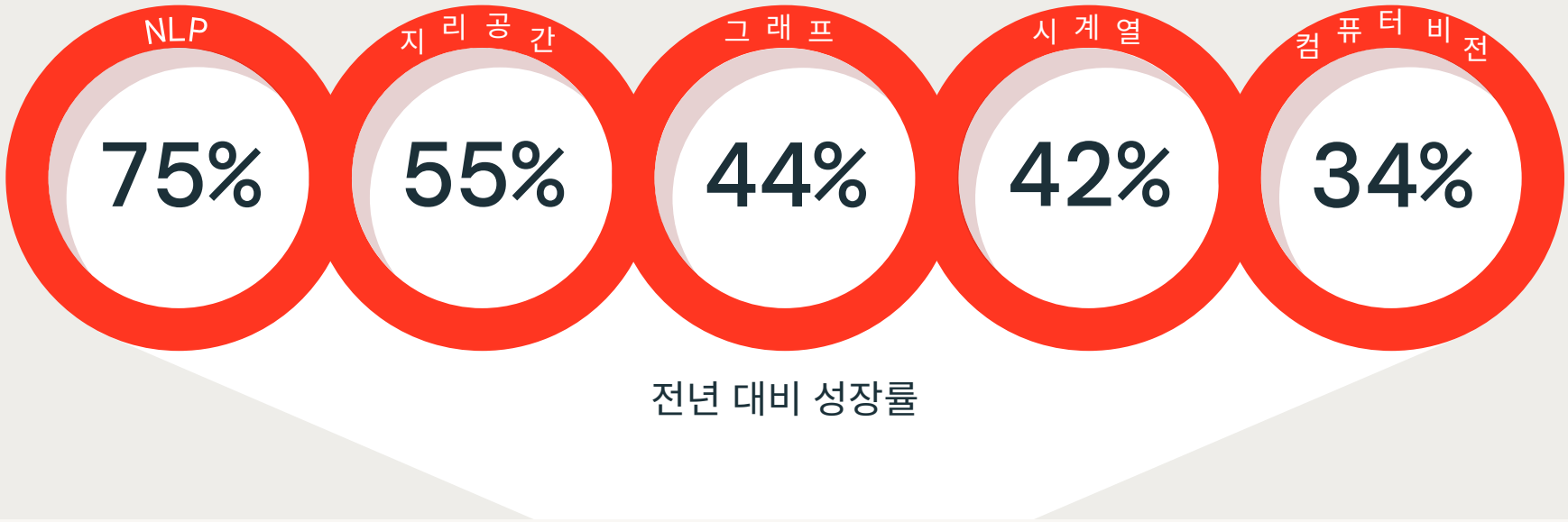


사용된 특수 Python 라이브러리의 50%가 NLP와 연관되어 있음

가장 널리 사용되는 DS/ML 애플리케이션인 NLP는 발전 속도를 늦추지 않습니다.

AI 기반 애플리케이션이 증가함에 따라 산업 전반에 걸쳐 NLP 솔루션에 대한 수요가 증가하고 있습니다. NLP는 Python 라이브러리 사용량의 대부분을 차지하고 있지만, 전년 대비 75%로 전체 애플리케이션 중 가장 높은 성장률을 보이고 있습니다.

빠르게 성장하는 DS/ML 애플리케이션



산업별 빠르게 성장하는 DS/ML 애플리케이션

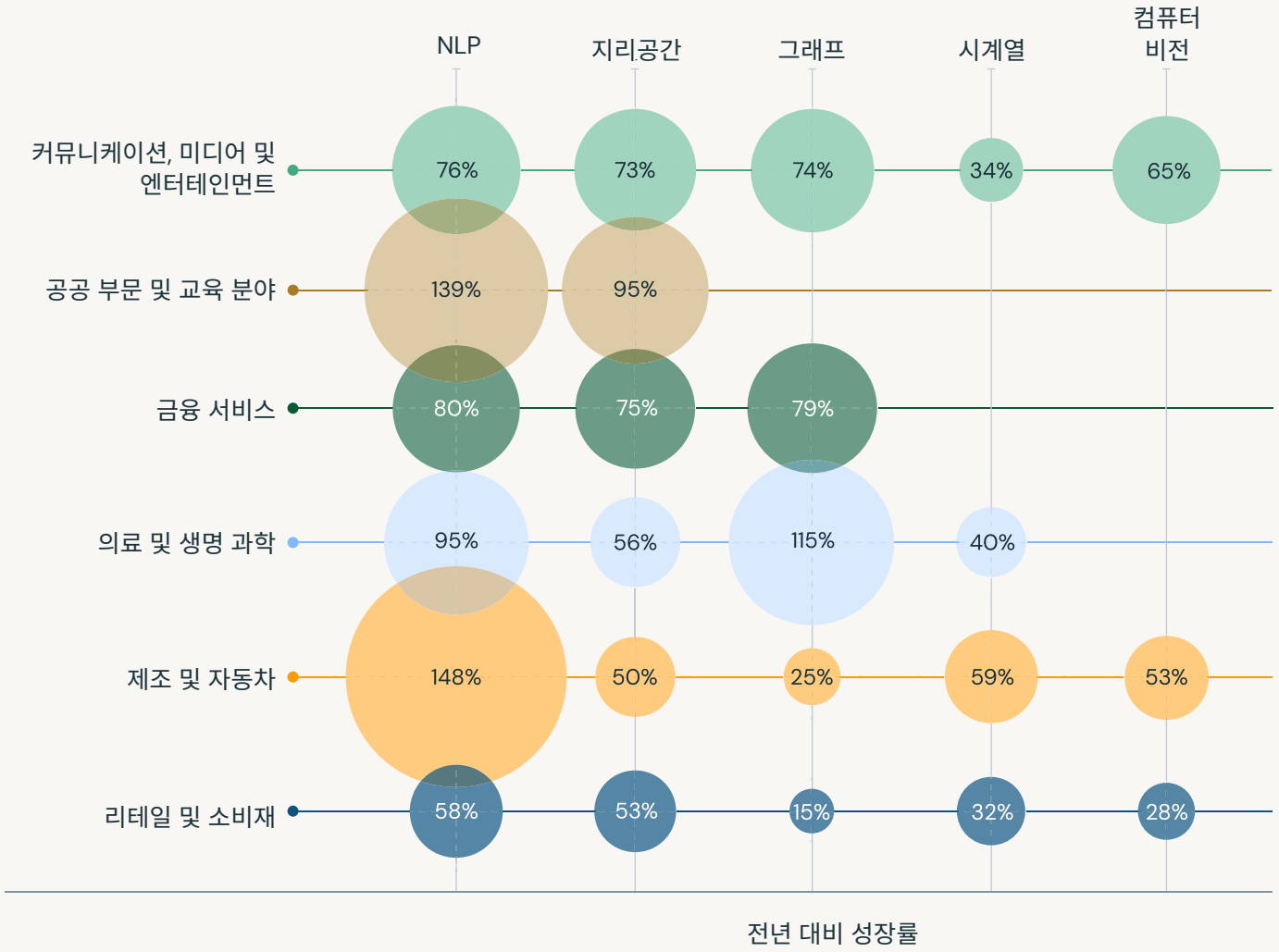


그림 4:
NLP는 애플리케이션 중 가장 높은 성장세를 보이고 있습니다. 가장 큰 성장률은 제조 및 자동차 분야의 NLP 사용 증가로, 전년 대비 148% 증가했습니다.

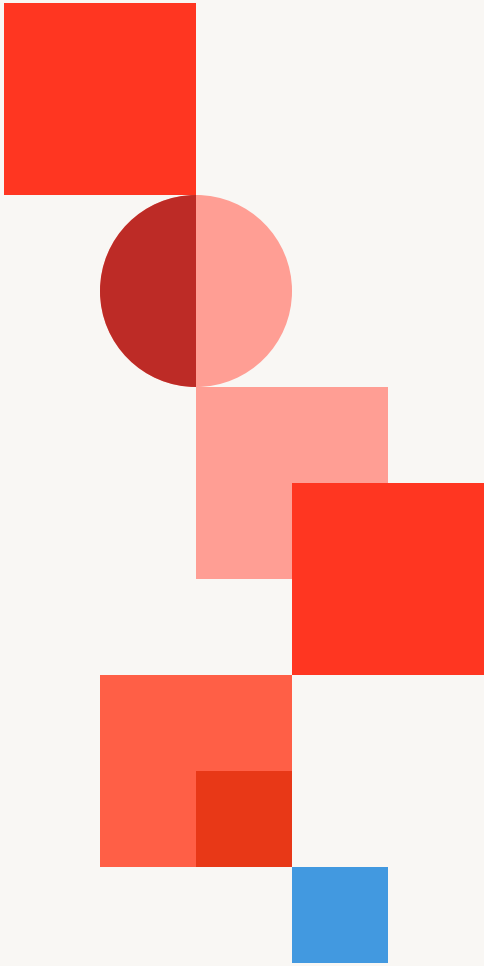
모든 산업에서 막대한 투자를 쏟고 있는 NLP

주요 산업 중에서 제조 및 자동차 분야는 NLP 사용이 전년 대비 148% 증가하여 가장 큰 증가세를 보였습니다. 고객의 피드백 분석부터 품질 관리 모니터링, 챗봇 지원에 이르기까지 모든 작업을 지원하는 NLP를 통해 기업은 운영 효율성을 개선할 수 있습니다. 공공 부문 및 교육 분야의 지난 1년간 NLP 성장률은 전년 대비 139%로 그 뒤를 바짝 쫓고 있습니다.

산불부터 조류 독감까지, ML 성장과 맞물린 시사 문제

지리공간은 6개 산업 분야 모두에서 크게 성장한 또 다른 애플리케이션입니다. 위치 기반 데이터에서 패턴, 추세, 상관관계를 찾는 기업이 점점 더 늘어나고 있습니다. 공공 부문 및 교육 분야의 지리공간 성장률이 높은 것은 재난 관리 및 비상 대응 계획과 관련이 있을 수 있습니다.

모든 애플리케이션과 산업에서 세 번째로 높은 성장률을 보인 것은 의료 및 생명과학 분야의 시계열 라이브러리 도입으로, 전년 대비 115%를 기록했습니다. 시계열은 환자 위험 예측, 공급 예측 및 신약 개발을 지원합니다. 2023년 NIH에서 실시한 리뷰 보고서에 따르면 시계열 분석을 통해 데이터에서 직접 추정함으로써 새로운 팬데믹에 대한 정확한 단기 예측을 쉽고 짧은 시간 내에 수행할 수 있습니다." [1]



1 역학에 시계열 분석 응용: 코로나19 팬데믹 기간 동안의 문헌 검토 및 경험(영문), 2023년 10월 16일.

최신 데이터 및 AI 스택

GenAI로 진화

GENAI의 다음 단계를 보여주는 최고의 데이터 및 AI 제품

데이터 리더들은 항상 AI 전략을 실현할 수 있는 최고의 도구를 찾고 있습니다. 이를 위해 Databricks Data Intelligence Platform에서 가장 널리 채택된 상위 10가지 데이터 및 AI 제품을 소개합니다. 여기에는 DS/ML, 데이터 거버넌스 및 보안, 오케스트레이션, 데이터 통합 및 데이터 소스 제품이 포함됩니다.

인기 제품 10개 중 9개는 오픈 소스입니다. 조직은 독점적인 장벽과 제한을 피하면서 더 유연한 제품을 선택하고 있습니다. 보고서의 뒷부분에서 보다 자세히 설명하겠지만 오픈 LLM의 인기도 점점 높아지고 있습니다.

상위 10개 데이터 및 AI 제품

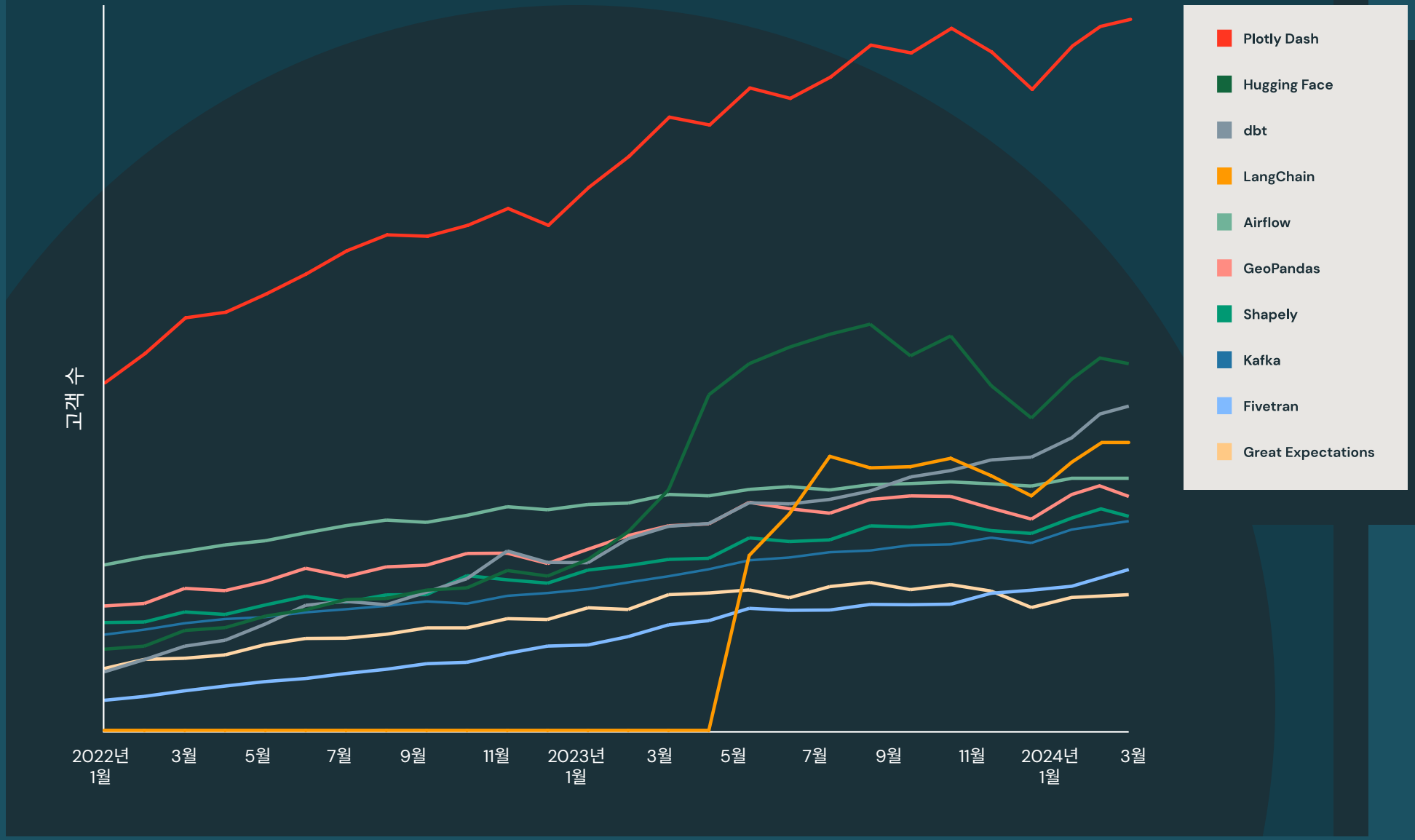


그림 5: 상위 10개 데이터 및 AI 제품에는 DS/ML, 데이터 거버넌스 및 보안, 오케스트레이션, 데이터 통합 및 데이터 소스 제품이 포함됩니다.

최상위 위치를 유지하고 있는 PLOTLY DASH

Plotly Dash는 데이터 사이언티스트들이 데이터 애플리케이션을 쉽게 구축, 확장 및 배포할 수 있게 해주는 로우코드 플랫폼입니다. Dash와 같은 제품은 기업들이 역동적인 비즈니스 요구에 부응하기 위해 애플리케이션을 보다 빠르고 쉽게 제공할 수 있도록 지원합니다. 2년이 넘는 기간 동안 Dash는 1위 자리를 지켜왔으며, 이는 데이터 사이언티스트들이 프로덕션급 데이터 및 AI 애플리케이션을 개발해야 한다는 압박을 점점 더 크게 느끼고 있다는 것을 말해줍니다.

2위로 뛰어오른 HUGGING FACE 트랜스포머

Hugging Face 트랜스포머는 1년 전 4위에서 2위로 올라서며 고객들이 가장 많이 사용하는 제품 중 두 번째로 높은 순위를 차지했습니다. 많은 기업들이 오픈 소스 플랫폼의 사전 학습된 트랜스포머 모델을 엔터프라이즈 데이터와 함께 사용하여 파운데이션 모델을 구축하고 세부 조정합니다. 이는 RAG 애플리케이션에서 볼 수 있는 성장 추세를 뒷받침합니다.

통합 후 몇 달 만에 최고의 제품이 된 LANGCHAIN

독점 LLM을 활용하고 구축하기 위한 오픈 소스 툴체인인 LangChain은 지난 봄에 상위권에 진입했으며, 통합 후 1년도 되지 않아 4위로 올라섰습니다. 기업이 자체적으로 최신 LLM 애플리케이션을 구축하고 특수한 트랜스포머 관련 Python 라이브러리를 사용하여 모델을 훈련할 때, LangChain을 사용하면 다른 시스템과의 신속한 인터페이스 또는 통합을 개발할 수 있습니다.

고품질의 데이터 세트를 구축하기 위해 지속적으로 투자하는 기업

상위 10위 안에 데이터 통합 제품 3개가 포함된 것은 기업들이 신뢰할 수 있는 데이터 세트 구축에 주력하고 있음을 나타냅니다. dbt(데이터 변환), Fivetran(데이터 파이프라인 자동화), Great Expectations(데이터 품질) 모두 꾸준한 성장세를 보이고 있습니다. 특히, dbt는 작년에 두 계단 상승했습니다.

라이징 스타

John Snow Labs는 의료 및 생명과학 조직이 AI 프로젝트를 구축, 배포 및 운영할 수 있도록 지원하는 AI 및 NLP 회사입니다. John Snow Labs는 고급 NLP, ML 모델 및 GenAI를 사용하여 질병 진단, 신약 개발 및 환자 치료의 정확성을 높이는 데 중요한 역할을 하고 있습니다.

의료 및 생명과학 분야에서 주로 사용되지만, 데이터 및 AI 제품 중에서는 15위에 랭크되어 있다는 점에서 John Snow Labs는 그 가치를 인정받을 만합니다. 널리 사용되는 Spark NLP 라이브러리는 텍스트 분류, 엔티티 인식, 감정 분석 등 다양한 NLP 작업을 지원하므로, 금융 서비스와 같은 다른 업종에서도 유용합니다.

벡터 데이터베이스

LLM 도입을 서두르는 기업

LLM은 언어 이해 및 생성 기능을 통해 다양한 비즈니스 사용 사례를 지원합니다. 그러나 특히 기업 환경에서는 LLM만으로는 한계가 있습니다. LLM은 신뢰할 수 없는 정보 출처일 수 있으며 할루시네이션(hallucination)이라고 하는 잘못된 정보를 제공하기 쉽습니다. 독립 실행형 LLM은 근본적으로 특정 조직의 도메인 지식과 요구 사항에 맞춰져 있지 않습니다.

데이터에 따르면 더 많은 기업이 독립 실행형 LLM에 의존하는 대신 RAG로 전환하고 있는 것으로 나타났습니다. RAG를 통해 조직은 자체 데이터를 사용하여 LLM을 보다 효과적으로 맞춤 구성하고 고품질의 GenAI 앱을 제공할 수 있습니다. LLM에 관련 정보를 추가로 제공하면 모델이 더 정확한 답변을 제공하고 오답 가능성을 줄일 수 있습니다.

기업의 GenAI를 주도하는 RAG

작년에는 5개월 만에 1,310% 성장한 LLM Python 라이브러리 차트를 통해 SaaS LLM의 뜨거운 성장세를 확인할 수 있었습니다. GPT-4와 같은 SaaS LLM은 대규모 텍스트 데이터세트로 학습되며 2년도 채 되지 않아 주류가 되었습니다.

올해 벡터 데이터베이스를 도입하여 높은 성장세를 보여주고 있습니다. 전체 벡터 데이터베이스 카테고리는 전년 대비 377% 성장했으며, Databricks Vector Search의 공개 미리 보기가 출시된 이후에는 186% 성장했습니다.[2]

RAG의 정의

검색 증강 생성(Retrieval Augmented Generation, RAG)은 질문이나 작업과 관련된 데이터와 문서를 찾아서 LLM이 보다 정확한 답변을 제공할 수 있도록 컨텍스트로 제공하는 GenAI 애플리케이션 패턴입니다.

벡터 데이터베이스와 RAG의 연동 방식

벡터 데이터베이스는 주로 비정형 데이터의 표현을 생성합니다. 이는 쿼리의 키워드와의 유사성을 기반으로 문서나 레코드를 찾기 위해 RAG 애플리케이션에서 정보 검색을 할 때 유용합니다.

RAG 애플리케이션은 기성 제품에 비해 많은 장점이 있습니다. RAG는 모델을 미세 조정하거나 사전 학습하는 데 비용과 시간을 들이지 않고도 독점적인 실시간 데이터를 LLM에 통합하는 인기 있는 방법으로 빠르게 부상했습니다.

벡터 데이터베이스의 기하급수적인 성장은 기업들이 자체 데이터를 LLM과 통합하기 위해 더 많은 RAG 애플리케이션을 구축하고 있음을 시사합니다.

2 Databricks Vector Search는 2023년 12월 7일에 공개 미리 보기로 출시되었습니다.

LLM PYTHON 라이브러리 사용

LLM 정의

트랜스포머 학습: 트랜스포머 학습: 트랜스포머 모델 (예: Hugging Face 트랜스포머)

SaaS LLMs: API 기반 LLM에 액세스하기 위한 라이브러리 (예: OpenAI)

LLM 도구: 독점 LLM 활용 및 구축을 위한 툴체인 (예: LangChain)

벡터 데이터베이스
벡터/KNN 인덱스(Pinecone 및 Databricks Vector Search)

그림 6: Databricks Vector Search의 공개 미리 보기 출시 이후 전체 벡터 데이터베이스 카테고리는 186% 성장했는데, 이는 다른 어떤 LLM Python 라이브러리보다 훨씬 높은 수치입니다.

참고: 고객은 한 카테고리에서 두 개 이상의 도구를 사용할 수 있으므로 두 번 이상 계산될 수 있습니다.
사용량은 외부 벡터 데이터베이스 서비스에 연결하는 패키지 사용과 당사 플랫폼의 API 호출을 통해 측정됩니다.
추세선은 계절적 하락을 해결하기 위해 12월 18일부터 1월 1일 사이의 평균값입니다.

점점 더 정교하게 LLM을 구축하는 기업들

작년에는 고객들이 기성 모델을 통해 LLM에 뛰어들었습니다. SaaS LLM을 사용하는 고객 수는 전년 대비 178% 성장했습니다. 하지만 기업들은 LLM을 더 세밀하게 제어하고 필요에 맞는 도구를 구축하기 시작했습니다.

벡터 데이터베이스, LLM 도구 및 트랜스포머 관련 라이브러리의 지속적인 성장은 많은 데이터 팀들이 구매 대신 구축을 선택하고 있음을 보여줍니다. 독점적인 LLM을 구축하고 활용하기 위해 LangChain과 같은 LLM 도구에 투자하는 기업들이 점점 더 많아지고 있습니다. Hugging Face와 같은 트랜스포머 관련 라이브러리는 LLM 훈련에 사용되며, 여전히 고객 수 기준으로 가장 많이 채택되고 있습니다. 이러한 라이브러리의 사용은 전년 대비 36% 증가했습니다. 이러한 추세선을 종합하면 오픈 소스 LLM이 더욱 정교하게 채택되고 있음을 알 수 있습니다.

벡터 데이터베이스를 사용하는 고객 수는 전년 대비 377% 증가

소규모 오픈 소스 모델을 선호하는 기업

오픈 소스 LLM 모델의 사용

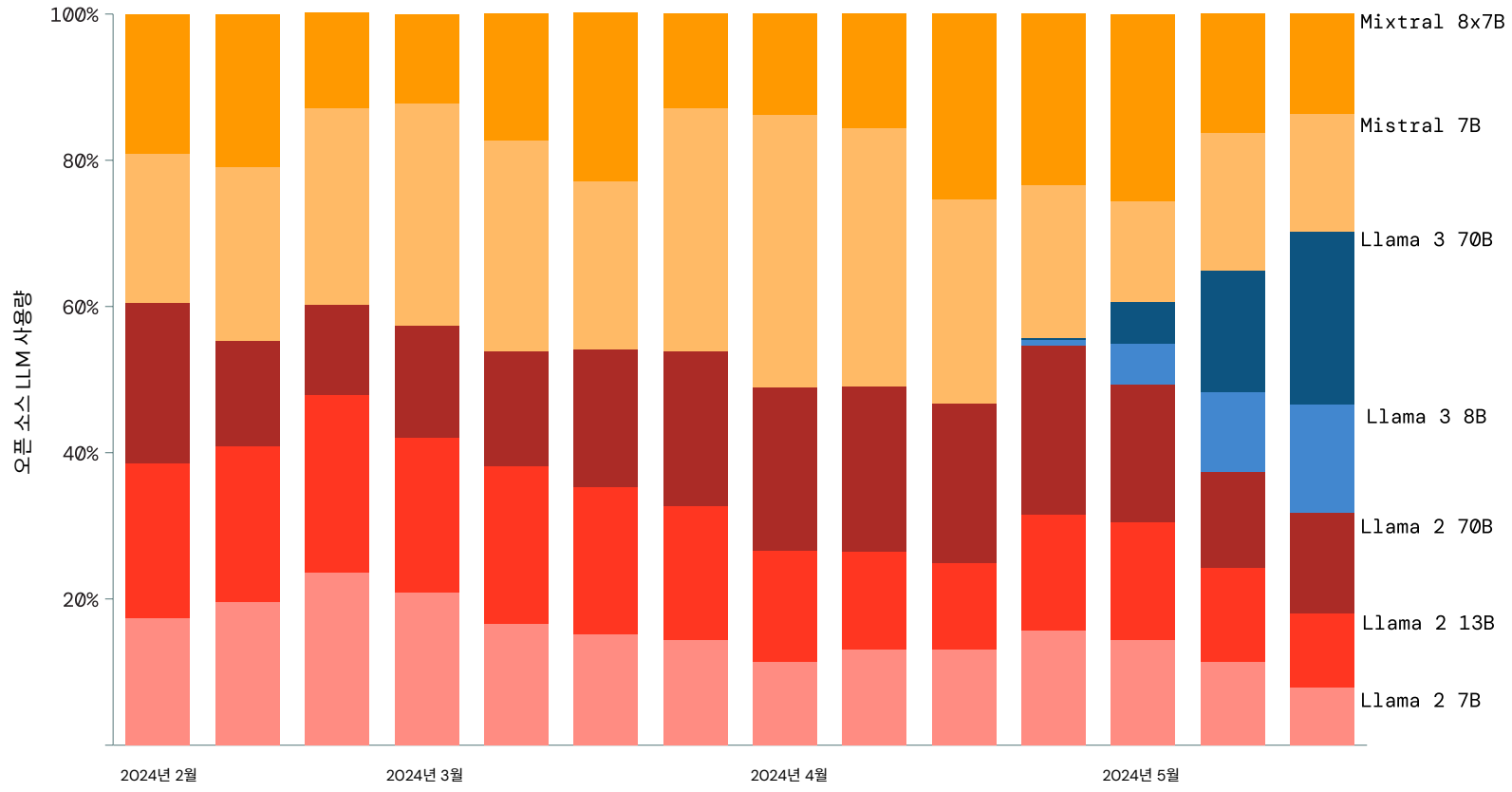


그림 7: Databricks 파운데이션 모델 API에 Mistral 및 Meta Llama 오픈 소스 모델 도입

참고: Meta Llama 3 출시를 반영하기 위해 차트의 날짜가 2024년 5월 19일로 연장되었습니다.

오픈 소스 LLM의 가장 큰 장점 중 하나는 특히 기업 환경에서 특정 사용 사례에 맞게 맞춤 구성할 수 있다는 점입니다. 실제로 *"가장 인기 있는 오픈 소스 모델이 무엇인가요?"라는 질문을 가장 많이 받습니다.* 실제로 고객들은 여러 모델과 모델 제품군을 사용해 봅니다. 이에 가장 큰 두 가지 오픈 소스 모델인 Meta Llama와 Mistral의 사용 현황을 분석했습니다. 데이터에 따르면 개방형 LLM 공간이 유동적이기 때문에 새로운 최신 모델이 빠르게 채택되고 있습니다.

각 모델마다 비용, 레이턴시, 성능 측면의 장단점이 있습니다. 가장 작은 두 Meta Llama 2 모델(70억 및 130억 매개변수)의 사용량을 합친 것이 가장 큰 Meta Llama 2 모델(700억 매개변수)의 사용량보다 훨씬 높습니다. Across Meta Llama 2, Llama 3 및 Mistral 전체 사용자 중 77%가 130억 이하의 매개변수 모델을 선택한 것으로 나타났습니다. 이는 기업들이 비용과 레이턴시에 대해 상당한 관심을 가지고 있음을 시사합니다.

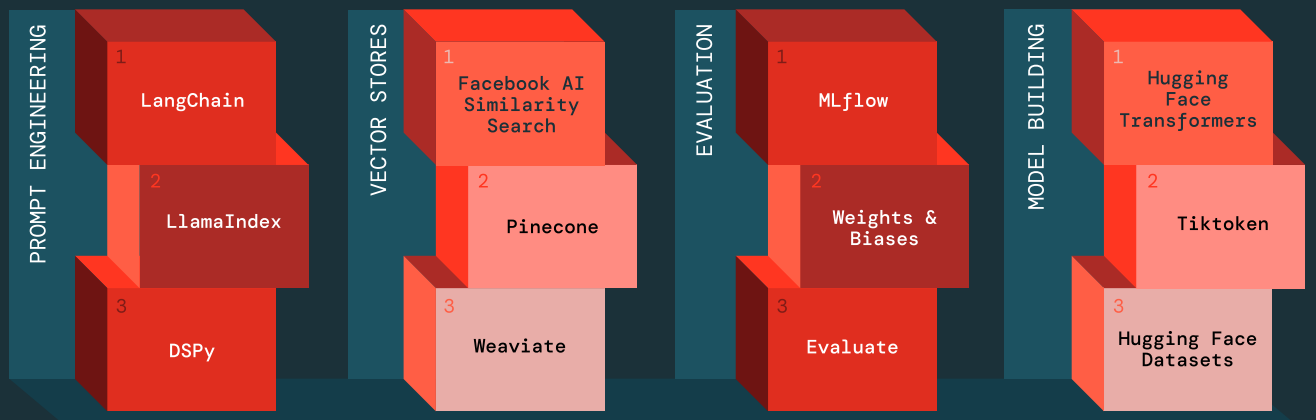
76% LLM을 사용하는 기업 중 오픈 소스를 선택하고, 종종 독점 모델과 함께 사용하는 비율

70% GenAI를 활용하는 기업 중 도구, 검색 및 벡터 데이터베이스를 사용하여 모델을 맞춤 구성하는 비율

새로운 모델을 빠르게 시도하는 기업들

Meta Llama 3는 2024년 4월 18일에 출시되었습니다. 출시 첫 주 만에 이미 다른 모델 및 다른 제공업체를 제치고 더 많은 조직에서 활용하기 시작했습니다. Llama 3는 출시 4주 만에 전에 전체 오픈 소스 LLM 사용량의 39%를 차지했습니다.

상위 GenAI Python 패키지



생성형 AI

GenAI를 빠르게 도입하는, 규제가 엄격한 산업

규제가 엄격한 산업은 위험을 회피하고 새로운 기술 도입을 주저하는 것으로 유명합니다. 엄격한 규정 준수 요건, 교체 비용이 많이 드는 고착화된 레거시 시스템, 구현 전 규제 당국의 승인 요구 등 그 이유는 여러 가지가 있습니다.

모든 산업이 새로운 AI 혁신을 수용하고 있지만, 그 중에서도 규제가 엄격한 금융 서비스와 의료 및 생명과학 분야는 규제가 덜한 다른 산업과 보조를 맞추고 있고 때로는 이를 능가하는 경우도 있습니다.

2023년 12월, Databricks는 파운데이션 모델 API를 출시하여 Meta Llama 및 MPT 모델과 같은 인기 오픈 소스 LLM에 대한 즉각적인 액세스를 제공했습니다. 최근 출시된 Llama 3에서 볼 수 있듯이 모델이 계속해서 빠르게 개선됨에 따라 오픈 소스에 대한 관심이 크게 증가할 것으로 예상됩니다.

산업별 요구에 맞는 오픈 LLM 모델의 활용

제조 및 자동차, 의료 및 생명과학이 고객당 가장 높은 평균 사용량으로 파운데이션 모델 API의 채택을 주도하고 있습니다. 제조 분야에서는 공급망 최적화, 품질 관리 및 효율성이 가장 유망한 사용 사례로 꼽힙니다.

MIT Tech Review Insights의 최근 보고서에 따르면 설문 조사에 참여한 의료 및 생명과학 분야의 CIO들은 GenAI가 조직에 가치를 가져다 줄 것으로 믿고 있습니다. 오픈 소스 LLM을 사용하면 의료 및 생명과학과 같이 규제가 엄격한 산업에서 데이터를 최대한 제어하면서 GenAI를 통합할 수 있습니다.

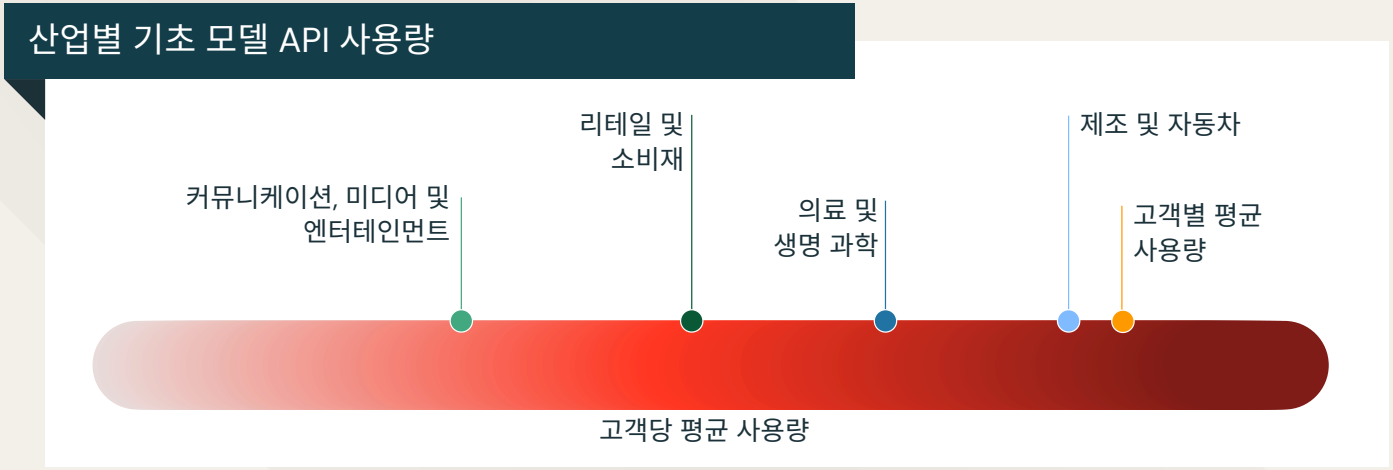


그림 8:제조 및 자동차, 의료 및 생명과학이 고객당 가장 높은 평균 사용량으로 파운데이션 모델 API의 채택을 주도하고 있습니다.

참고: 날짜 범위: 2024년 1월 – 2024년 3월

CPU과 GPU 비교: 6개월 만에 88% 성장한 LLM에 대한 금융 서비스의 노력

CPU는 다양한 작업을 빠르게 처리하도록 설계된 범용 프로세서이지만 동시에 처리할 수 있는 작업 수에 제한이 있습니다. CPU는 기존의 머신 러닝에 사용됩니다. GPU는 수천 또는 수백만 개의 개별 작업을 한 번에 동시에 처리할 수 있는 특수 프로세서입니다. GPU는 LLM을 훈련하고 서비스하는 데 필요합니다.

Databricks는 모델 서빙 고객들의 CPU 및 GPU 사용량과 성장세를 살펴봄으로써 고객들이 AI에 어떻게 접근하고 있는지 파악했습니다. 데이터에 따르면 GPU는 주로 LLM과 연관되어 있습니다.

GPU 사용량이 가장 많은 금융 서비스

규제가 가장 엄격한 산업 중 하나인 금융 서비스 업계도 GenAI를 외면하지 않고 있습니다. 지난 6개월 동안 기업당 평균 GPU 사용량이 가장 높을 뿐만 아니라 GPU 성장률도 88%로 가장 높았습니다. LLM은 사기 탐지, 자산 관리, 투자자 및 애널리스트 애플리케이션 등 비즈니스에 중요한 사용 사례를 지원합니다.

모델 서빙, 기존 ML 대 LLM

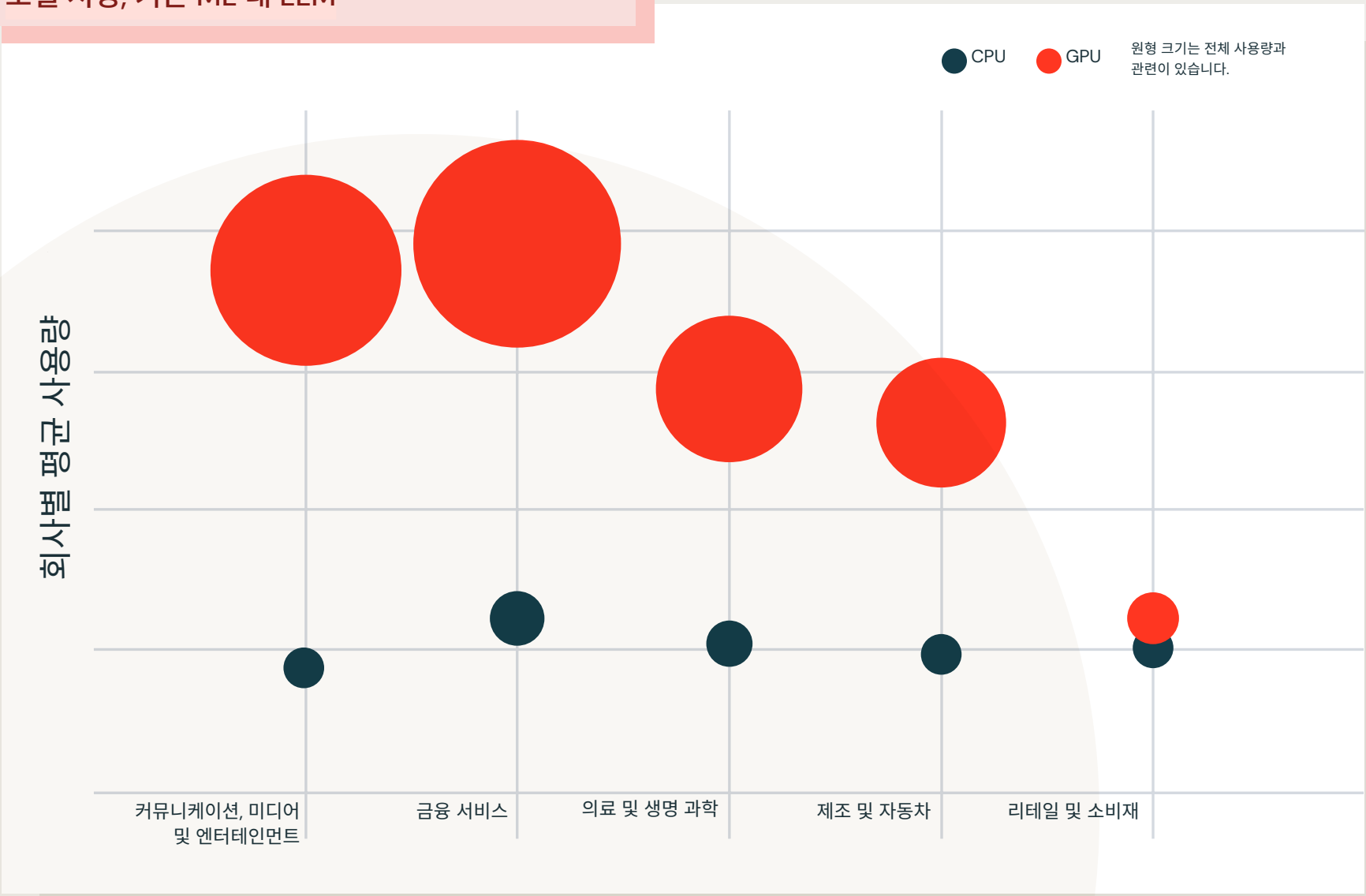


그림 9: 금융 서비스는 CPU와 GPU의 평균 사용량이 모두 가장 높습니다.

참고: 데이터 범위: 2024년 1월 – 2024년 3월

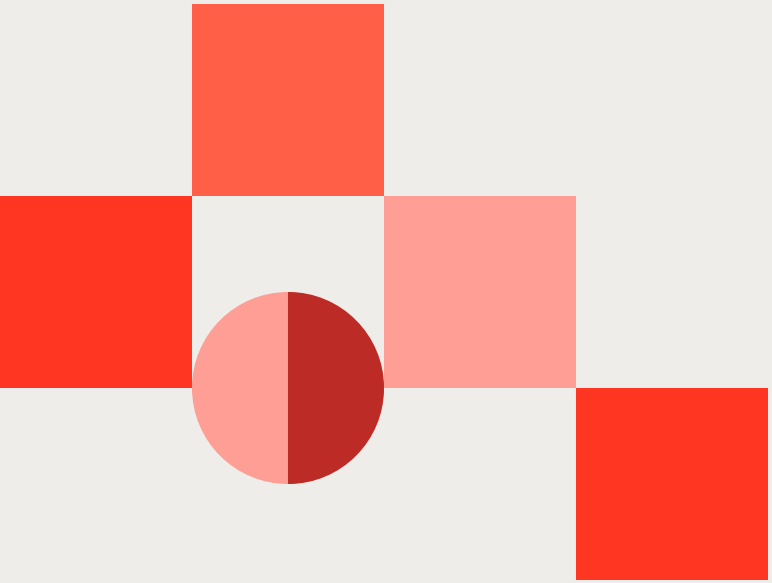
규제가 엄격한 산업에서 통합 거버넌스 도입 선도

AI 보안과 거버넌스는 조직의 AI 이니셔티브에 대한 신뢰를 구축하는 데 매우 중요합니다. 이는 데이터 실무자가 정확한 지침과 표준을 준수하면서 제품을 개발하고 유지 관리할 수 있도록 도와줍니다. Databricks Unity Catalog와 같은 통합 거버넌스 솔루션은 모든 데이터와 AI 자산을 포괄하며, 조직이 비공개 데이터로 GenAI 모델을 더 쉽게 훈련하고 배포할 수 있도록 지원합니다.

Gartner에 따르면 2024년에 비즈니스 및 기술 결정에 영향을 미칠 주요 추세는 AI 신뢰, 위험 및 보안 관리입니다. 이제 그 어느 때보다 많은 리더들이 데이터와 AI를 활용하여 조직을 혁신하고자 합니다. 이는 고객들이 통합 거버넌스를 도입하는 데에도 반영되어 있습니다.

데이터 및 AI 거버넌스의 최전선에 있는 금융 서비스

규제 및 보안 규정 준수는 금융 서비스 조직의 문화에 뿌리내리고 있습니다. MIT Technology Review Insights CIO 비전 2025 보고서의 설문조사 데이터에 따르면 금융 기관은 데이터 관리 및 인프라 부문에서 가장 높은 투자 성장률을 보일 것으로 예상되는데, "금융 업계 응답자에 따르면 현재와 2025년 사이에 74% 성장해 표본 전체의 52%에 비해 가장 높은 투자 성장을 보일 것"으로 추정됩니다.



산업별 UNITY CATALOG 도입

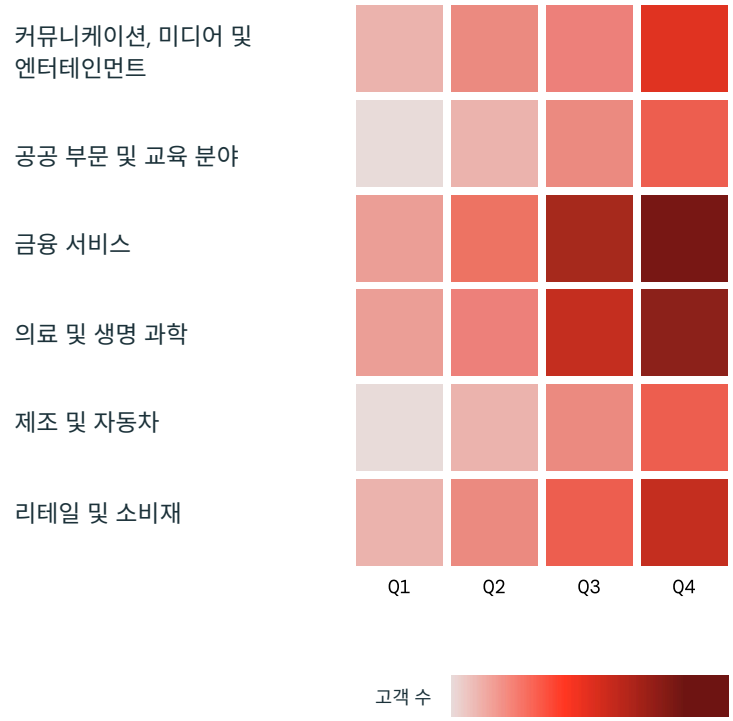


그림 10:금융 서비스는 통합 데이터 및 AI 거버넌스를 위해 Unity Catalog를 선도적으로 도입하고 있습니다.

참고: 데이터 범위: 2023년 2월 1일 ~ 2024년 1월 31일

산업별 서버리스 모델 서빙 도입

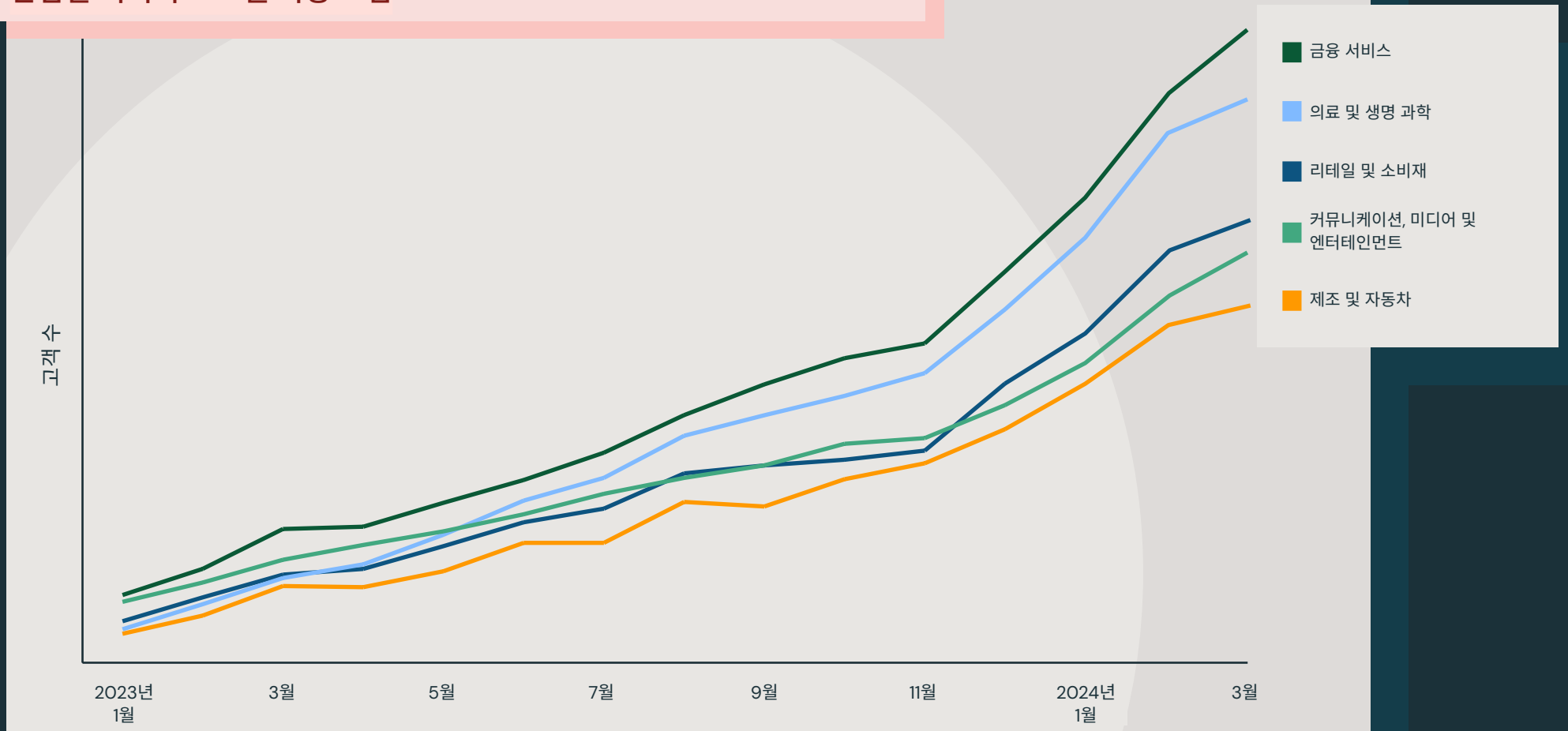


그림 11: 금융 서비스가 서버리스 제품 도입을 주도하고 있으며, 의료 및 생명과학이 그 뒤를 잇고 있습니다.

참고: 여기에는 서버리스 엔드포인트에서의 모델 서빙, Databricks SQL, 레이크하우스 모니터링 및 서버리스 작업이 포함됩니다. 2023년 11월 서버리스는 추가 지역 클라우드 플랫폼에서 사용할 수 있게 되었습니다.

실시간 ML 애플리케이션 구축을 위해 서버리스로 전환하는 기업

실시간 ML 시스템은 들어오는 데이터를 기반으로 즉각적인 예측이나 조치를 취할 수 있는 기능을 제공함으로써 비즈니스 운영 방식을 혁신적으로 변화시키고 있습니다. 하지만 이러한 시스템을 구축하고 유지 관리하려면 빠르고 확장 가능한 서비스 인프라가 필요한데, 이를 위해서는 전문 지식이 필요합니다.

서버리스 모델 서빙은 수요 변화에 따라 자동으로 확장 또는 축소되므로 기업은 사용한 만큼만 비용을 지불하고 비용을 절감할 수 있습니다. 기업은 개인화된 추천부터 사기 탐지에 이르기까지 다양한 실시간 ML 애플리케이션을 구축할 수 있습니다. 모델 서빙은 사용자 상호작용을 위한 머신 러닝 애플리케이션 지원에도 도움이 됩니다.

서버리스 데이터 웨어하우징 및 모니터링의 도입이 꾸준히 증가하고 있으며, 이 역시 수요에 따라 확장됩니다.

서버리스 제품을 가장 많이 채택한 금융 서비스는 6개월 동안 사용량이 131% 증가했습니다. 이 업계는 시장 예측을 위해 노력하고 있으며, 실시간 예측을 통해 더욱 강력한 시장 분석을 제공합니다.

의료 및 생명과학은 6개월 동안 서버리스 제품 사용량이 132% 증가했습니다. 이 산업은 지난 1년 동안 4위에서 2위로 올라섰습니다. 의료 및 생명과학은 특히 피크 시간대나 대용량 데이터 세트(예: 게놈 데이터 또는 의료 영상)를 처리할 때 데이터 처리 요구사항에 큰 변동이 발생합니다.

결론

데이터 사이언스와 AI는 기업의 효율성을 향상시키고 있으며, GenAI는 새로운 가능성의 지평을 열어가고 있습니다. Data Intelligence Platform은 조직 전체가 데이터와 AI를 사용할 수 있는 응집력 있고 관리되는 공간입니다. 데이터에 따르면 모든 산업 분야의 기업들이 이러한 도구를 도입하고 있으며, 예상치 못한 산업이 얼리 어답터가 될 수도 있습니다.

조직은 ML 모델을 프로덕션에 활용하여 측정 가능한 이점을 실현했습니다. 데이터에서 인사이트를 얻기 위해 NLP를 채택하고 사용하는 기업이 점점 더 많아지고 있습니다. 기업들은 벡터 데이터베이스와 RAG 애플리케이션을 사용해 자체 엔터프라이즈 데이터를 LLM에 통합하고 있습니다. 오픈 소스 도구는 가장 인기 있는 제품 중에서 계속해서 높은 순위를 차지하고 있기 때문에 앞으로도 대세를 이어 갈 것입니다. 기업은 통합 데이터와 AI 거버넌스를 통해 전략을 수립하고 있습니다.

핵심 내용: 모든 업계의 승자들은 데이터와 AI를 가장 효과적으로 활용하는 기업들일 것입니다.

Databricks 정보

Databricks는 데이터 및 AI 기업입니다. BLOCK Comcast, Condé Nast, Rivian, Shell 및 Fortune 500대 기업의 60% 이상을 포함하여 전 세계적으로 10,000개 이상의 조직이 Databricks 데이터 인텔리전스 플랫폼을 사용하여 데이터를 제어하고 이를 AI와 함께 활용하고 있습니다. Databricks는 샌프란시스코에 본사를 두고 세계 각지에 지사를 운영하고 있으며 Lakehouse, Apache Spark™, Delta Lake와 MLflow를 개발한 주역들이 모여 설립했습니다.

더 자세히 알아보려면 LinkedIn, X, Facebook에서 Databricks를 팔로우하세요.

databricks

© Databricks 2024. All rights reserved. Apache, Apache Spark, Spark 및 the Spark 로고는 Apache Software Foundation의 상표입니다. | 이용약관