제공:

databricks

초보자를 위한 데이터 인텔리전스 플랫폼

for dummies®
A Wiley Brand

인텔리전스를 통한
데이터 및 AI 민주화

AI를 통한 엔터프라이즈
데이터 이해

ETL, DW, BI 및 AI를
통한 혁신 가속화

Ari Kaplan
Stephanie Diamond

Databricks 특별판

Databricks 소개

Databricks는 데이터 및 AI 전문 기업입니다. Comcast, Condé Nast, Grammarly, Fortune 500대 기업의 60% 이상을 포함한 전 세계 수천 개의 조직이 데이터, 분석, AI를 통합 및 민주화하기 위해 Databricks 데이터 인텔리전스 플랫폼을 사용하고 있습니다. Databricks는 샌프란시스코에 본사가 있고 전 세계에 지사를 두고 있으며 Lakehouse, Apache SparkTM, Delta Lake, MLflow의 최초 개발자들이 설립했습니다. 자세한 내용을 알아보려면 SNS에서 Databricks를 팔로우하세요.


x.com/databricks


linkedin.com/company/databricks


facebook.com/databricksinc

초보자를 위한 데이터 인텔리전스 플랫폼

Databricks 특별판

저자: Ari Kaplan, Stephanie Diamond

for dummies®
A Wiley Brand

초보자를 위한 데이터 인텔리전스 플랫폼, Databricks 특별판

발행인

John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN: 978-1-394-32296-1 (pbk); ISBN: 978-1-394-32298-5 (ebk);
ISBN: 978-1-394-32299-2 (ePub). 인쇄 버전의 일부 빈 페이지는 ePDF 버전에 포함되지 않을 수 있습니다.

발행인 감사의 글

다음은 이 책을 발간할 수 있도록 도움을 주신 분들입니다.

프로젝트 관리자 겸 편집자:
Carrie Burchfield-Leighton

선임 관리 편집자: Rev Mengle

원고 검토 편집자: Traci Martin

선임 고객 계정 관리자:
Matt Cox

목차

서론

조직의 성공은 데이터를 효과적으로 사용하여 현명한 의사결정을 하고 비즈니스 성장을 도모하는 데 따라 좌우됩니다. 여기에는 분석 또는 인공지능(AI)에 사용할 수 있는 전략적 자산으로 원시 데이터를 변환하는 작업이 포함됩니다. AI를 적용한 데이터 인텔리전스는 조직이 더 현명한 의사결정을 하고 기업의 성공을 이룰 수 있는 강력함을 부여합니다.

Databricks 데이터 인텔리전스 플랫폼은 데이터와 AI를 위한 통합 플랫폼을 제공하여 조직이 데이터를 민주화하고 AI 애플리케이션을 구축할 수 있도록 합니다. 각 팀에서는 협업을 통해 데이터 사일로를 허물어 데이터 기반의 의사결정 문화를 만들 수 있습니다. 기존 AI와 생성형 AI, 데이터 웨어하우징, 비즈니스 인텔리전스, 거버넌스를 활용해 자체 데이터 자산을 최대한 이용합니다.

이 책에 대한 정보

초보자를 위한 데이터 인텔리전스 플랫폼, Databricks 특별판에서는 기업이 사후 대응 전략에서 사전 예방 전략으로 전환하고 데이터를 경쟁력 있는 자산으로 활용할 수 있는 중요한 방법을 알아봅니다.

- 데이터 인텔리전스의 가치와 AI의 능력
- Databricks 데이터 인텔리전스 플랫폼
- 기존 AI와 생성형 AI로 애플리케이션 구축
- 데이터 인텔리전스 플랫폼이 필요한 이유

이런 분들을 위한 문서

이 책을 쓰면서 여러분에 대해 몇 가지 가정을 세워봤습니다.

- 복잡한 문제를 해결하기 위해 AI를 활용하고 AI와 데이터를 통합하는 솔루션을 원하시는 분
- 효율성과 혁신, 경쟁 우위를 확보할 수 있는 통합형이면서 개방적이고 확장 가능한 플랫폼을 찾고 있는 의사결정자이신 분

- 데이터 거버넌스, 보안 및 규정 준수를 보장할 책임이 있으신 분 Databricks가 어떻게 지원하는지 알려드립니다.
- 데이터 볼륨 및 처리 요구 증가에 따라 효과적으로 확장 가능한 솔루션을 원하시는 분
- 운영, 전략, 미션 크리티컬 영역에 상관없이 새로운 방식으로 문제를 해결하고자 하시는 분
- Databricks 플랫폼이 기존 시스템, 데이터 인프라, 분석 도구와 어떻게 통합되는지 궁금하신 분

이 중 하나라도 해당되는 분을 위한 문서입니다.

이 책에서 사용된 아이콘

이 책 전체에 걸쳐 중요한 정보를 강조하기 위해 다양한 아이콘이 사용됩니다. 각각의 의미는 다음과 같습니다.


팁

이 아이콘은 작업을 더 쉽고 빠르게 수행할 수 있는 정보를 표시합니다.


기억하세요

이 아이콘은 메모리 뱅크를 검색할 때 기억해야 할 내용을 다룹니다.


경고

이 아이콘은 독자 또는 독자의 회사가 주의해야 할 내용을 다룹니다.

추가 자료

이 책을 통해 데이터 인텔리전스 플랫폼을 자세히 알아볼 수 있으나, 이 책의 내용보다 많은 리소스가 필요한 경우 다음을 통해 더 많은 인사이트를 얻을 수 있습니다.

- Databricks에서 데모, 제품 둘러보기, 튜토리얼을 확인하세요. Databricks.com/resources/demos에 방문하세요.
- community.databricks.com에서 10만 명 이상이 모인 강력한 Databricks 커뮤니티에 참여하세요.

주요 내용

- » 데이터 인텔리전스의 가치 알아보기
- » 데이터 인텔리전스 플랫폼의 주요 기능에 대해 알아보기
- » 여러 산업에서의 사용 사례 살펴보기

1장 데이터 인텔리전스에 대한 이해

데이터 인텔리전스를 효과적으로 활용하면 모든 사용자가 데이터에 더 수월하게 액세스할 수 있게 되어 조직 내 의사결정과 데이터 상호작용에 혁신을 불러올 수 있습니다. 생성형 인공지능(AI)을 결합하여 더 큰 인사이트를 얻고 전략적 의사결정을 지원할 수 있도록 데이터 분석의 수준을 한 단계 끌어올립니다. 비기술적인 사용자도 자연어로 조직에 대한 질문을 할 수 있습니다.

데이터 인텔리전스는 AI를 적용하여 조직 데이터의 고유성을 이해함으로써 더 큰 인사이트를 얻고 실행 가능한 인사이트를 산출하는 것을 의미합니다. 이 프로세스에는 생성형 AI를 사용해 방대한 양의 데이터를 선별하고 이해함으로써 의사결정에 정보를 제공하고 서비스, 투자 및 전반적인 비즈니스 전략을 개선할 수 있는 지능적인 인사이트를 조직이 도출할 수 있도록 합니다.

1장에서는 조직이 데이터를 실행 가능한 지식으로 전환할 수 있는 데이터 인텔리전스의 정의, 이점 및 영향에 대해 알아봅니다.

데이터 인텔리전스에 대해 알아보기

경쟁 우위를 확보하고자 하는 기업은 기업이 보유한 데이터에 대한 이해가 필요합니다. 생성형 AI로 강화된 데이터 인텔리전스는 데이터 수집 및 분석부터 실제 문제 해결을 위한 데이터 인사이트 적용까지 다양한 활동을 아우릅니다. 기업은 데이터 인텔리전스를 활용하여 패턴을 발견하고, 트렌드를 예측하며, 증거 기반의 의사결정을 할 수 있습니다. 이 섹션에서는 데이터 인텔리전스를 통해 기업의 성공에 필요한 도구를 지원하는 방법을 알아봅니다.

지능형

데이터 인텔리전스는 생성형 AI와 레이크하우스 아키텍처의 통합 이점을 결합하여 데이터의 고유한 시맨틱을 이해하는 데이터 지능을 활성화시킵니다. 이로써 Databricks 데이터 인텔리전스 플랫폼은 자동으로 성능을 최적화하고 비즈니스에 고유한 방식으로 인프라를 관리할 수 있습니다.

단순함

자연어는 사용자 경험을 단순화합니다. 데이터 인텔리전스는 조직의 언어를 이해하기 때문에 동료에게 질문하듯이 새로운 데이터를 쉽게 검색하고 발견할 수 있습니다. 게다가 자연어를 지원함으로써 코드를 작성하고 오류를 수정하며 해답을 찾는 등 새로운 데이터와 애플리케이션을 개발하는 속도가 빨라집니다.

프라이빗

데이터 및 AI 애플리케이션에는 강력한 거버넌스와 보안이 필요하며, 특히 생성형 AI가 등장하면서 그 필요성이 더욱 커지고 있습니다. Databricks는 거버넌스와 보안에 대한 통합된 접근 방식을 기반으로 구축된 엔드투엔드 MLOps 및 AI 개발 솔루션을 제공합니다. 데이터 프라이버시 및 IP 제어를 손상시키지 않으면서 OpenAI와 같은 API 사용부터 맞춤형 모델까지 모든 AI 이니셔티브를 추진할 수 있습니다.

기억하세요

*생성형 AI*는 자체적으로 새로운 콘텐츠를 해석하거나 생성할 수 있는 모든 유형의 AI입니다. 생성형 AI 콘텐츠에는 텍스트, 이미지, 동영상, 음악, 번역물, 요약, 코드가 있습니다. 또한 개방형 질문에 답하고 채팅에 참여하는 등 특정 작업을 완료할 수 있습니다. 일반 대중에게 ChatGPT, DALL-E와 같은 솔루션을 통해 생성형 AI가 가진 의미를 소개하면서 기술에 대한 인지도가 크게 상승했습니다.

팁

데이터 인텔리전스가 없는 플랫폼은 유용하지 않습니다. 데이터 웨어하우스(DW)와 같은 플랫폼은 고도로 숙련된 엔지니어가 인프라를 수동으로 유지관리하며 최적화해야 하기 때문에 지능적이지 않습니다. 이 예시에서 데이터 인텔리전스는 플랫폼의 사용 및 트렌드에 대해 학습하고 학습한 내용을 적용하여 보다 효율적이면서 더 나은 플랫폼을 만들 수 있습니다.

인사이트 수집, 분석 및 적용

경쟁 우위를 확보하고자 하는 기업은 데이터를 효과적으로 수집 및 해석해야 합니다. 데이터 인텔리전스는 더 나은 데이터 수집 및 결합을 통해 강력한 데이터 세트를 만들고, 분석 및 AI로 데이터를 분석하여 실제 의사결정에 도움을 줍니다.

데이터에 대한 기업의 이해 돕기

데이터 인텔리전스는 기업이 데이터를 이해하는 데 중요한 역할을 합니다. 고급 도구를 사용하면 기업이 고객과 시장을 더 잘 이해할 수 있습니다. 이는 더 현명한 의사결정을 하는 데 도움이 됩니다. 다음은 이를 촉진하는 몇 가지 예시입니다.

- **데이터 민주화 향상:** 이제 의사결정자는 자연어를 사용해 기술 프로그래머들의 도움을 받지 않고 데이터에 대해 자주적으로 질문할 수 있습니다. 이렇게 하면 훨씬 더 많은 사용자들이 기업의 데이터 자산에서 가치를 얻을 수 있습니다.
- **작업 간소화:** 데이터 인텔리전스는 자동으로 성능을 최적화하고 비즈니스에 고유한 방식으로 인프라를 관리할 수 있습니다.
- **데이터 거버넌스 및 규정준수 보장:** 데이터에 대한 이해란 데이터의 출처를 파악하고, 데이터를 어떻게 사용할지 판단하며, 법적 및 윤리적 기준을 준수하는지 확인하는 것입니다. 데이터 인텔리전스는 효과적인 데이터 거버넌스를 위한 도구를 제공함으로써 기업이 데이터 품질 및 보안을 관리하여 규정을 준수하도록 지원합니다.

레이크하우스 아키텍처 기반의 구축

데이터 인텔리전스로 구축된 통합, 개방, 확장형 레이크하우스 아키텍처는 데이터 관련 기능을 일관된 단일 환경으로 통합하는 포괄적인 시스템으로서의 역할을 합니다.


팁

조직은 통합 플랫폼을 활용하여 더욱 효율적인 데이터 관리 및 분석 접근 방식의 이점을 누릴 수 있습니다. 데이터 사일로를 없애고 모든 데이터 자산을 위한 중앙 집중형의 단일 리포지토리를 제공합니다. 이로써 조직 전반적으로 일관성, 정확성, 거버넌스를 보장합니다.

데이터 인텔리전스의 이점 극대화

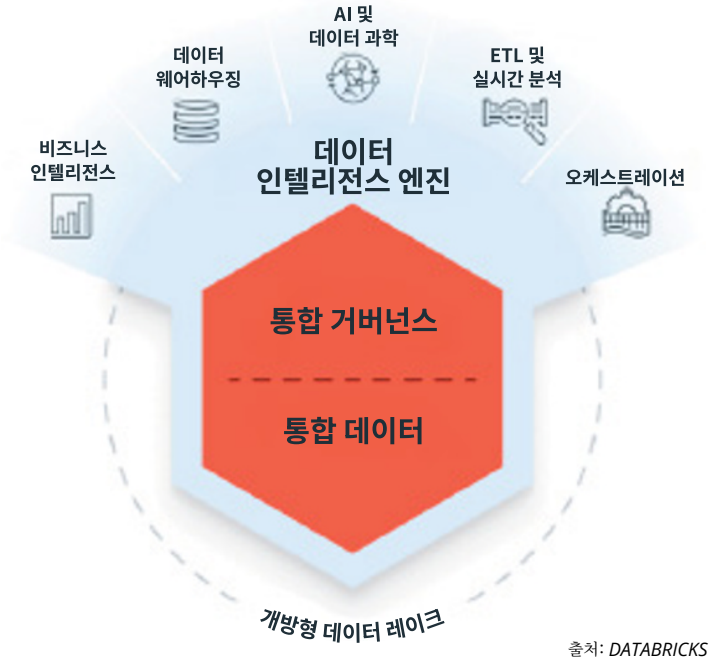
데이터 인텔리전스는 조직에 대한 핵심 전략으로 발전하여 데이터가 가진 능력을 활용할 수 있도록 합니다. 데이터 인텔리전스는 숙련된 직원이 필요하지 않은 개발을 간소화합니다. 이 섹션에서는 이러한 이니셔티브가 조직에 제공할 수 있는 몇 가지 이점에 대해 간략하게 알아봅니다.

데이터를 쉽게 검색하고 이해할 수 있도록 하기

데이터 인텔리전스는 조직의 언어를 이해하기 때문에 동료에게 질문하듯이 새로운 데이터를 쉽게 검색하고 발견할 수 있습니다. 그림 1-1에서 볼 수 있듯이, 데이터 인텔리전스는 단순한 키워드 일치뿐만 아니라 검색의 맥락을 이해하여 정보를 쉽게 발견하고 검색할 수 있도록 합니다. 또한 자연어 처리(NLP) 도구를 채택하면 사용자가 일반어를 사용해 데이터를 질의할 수 있습니다.


기억하세요

*자연어*는 데이터 인텔리전스에서 중요한 기능을 하며 시스템이 언어를 이해하고 해석할 수 있도록 합니다. 이를 통해 시맨틱을 사용하면 대량의 텍스트에서 중요한 정보를 더 쉽게 추출하여 의사결정 및 고객 인사이트를 향상하는 목적으로 사용할 수 있습니다.



출처: *DATABRICKS*

그림 1-1: 데이터 인텔리전스 엔진은 엔드투엔드 데이터 플랫폼의 모든 단계에 통합 거버넌스 및 데이터를 추가합니다.

사일로화된 데이터를 단일 플랫폼으로 통합

사일로화된 데이터를 단일 플랫폼으로 통합하면 다수의 조직에서 직면하는 문제, 즉 여러 시스템, 부서, 위치에서 발생하는 데이터 단편화 문제를 해결할 수 있습니다. 데이터가 사일로화되면 다른 관련 데이터와 격리되어 고객 행동이나 시장 추세와 같은 개념에 대한 이해를 거의 할 수 없게 됩니다.

경고

단편화된 데이터는 관리자가 시야를 넓게 볼 수 없기 때문에 비효율적일 수 있으며 더 중요한 문제는 기회를 놓칠 수 있다는 점입니다. 데이터를 통합된 단일 플랫폼으로 결합하면, 기업이 사일로를 허물고 데이터의 흐름 및 분석이 가능해집니다.

기억하세요

통합 데이터 플랫폼은 모든 데이터(정형 또는 비정형)를 저장 및 분석할 수 있는 중앙 집중형 리포지토리를 제공합니다. 이로써 데이터의 정확성을 보장하고 고급 분석 및 AI 애플리케이션을 더 효과적으로 사용할 수 있습니다. 결과적으로, 조직은 데이터 자산을 최대한 활용할 수 있습니다.

비전문 사용자를 위한 데이터 인사이트 획득

데이터에 더 간편하게 액세스할 수 있는 점은 비전문 사용자도 데이터에 더 쉽게 액세스할 수 있음을 의미합니다. 이를 통해 IT 부서에 의존하지 않고 인사이트를 얻을 수 있습니다. 모든 직원이 데이터 분석에 대한 기본 사항과 이를 간편히 수행할 수 있게 하는 도구를 숙지하도록 합니다.

기업 운영 간소화 및 비용 절감

데이터 인텔리전스는 기업의 기술 운영을 간소화함으로써 비용을 절감하도록 합니다. 예측 분석을 통해 트렌드를 예상하고 기업이 전략을 조정할 수 있습니다.

팁

AI는 시간 소모적인 수동 프로세스를 자동화하는 등 기술 효율성과 비용 절감을 위한 새로운 기회를 찾아낼 수 있습니다. 예를 들어, 부적절하게 활용되는 리소스를 파악하여 재할당하거나 데이터 저장 방식을 재구성하여 더 크고 빠르게 확장할 수 있도록 합니다. 핵심은 이제 AI가 원시 데이터 소프트웨어 엔지니어링에서 분석까지 모든 기술 단계의 운영을 개선할 수 있다는 점입니다.

협업 증진

데이터 인텔리전스 도구는 모든 작업에 공통된 환경을 제공함으로써 여러 팀의 협업을 가능하게 합니다. 팀은 동일한 데이터 세트에서 동시에 작업하고, 함께 코드를 개발하며, 대시보드 및 보고 인사이트를 공유하고, 공동의 의사결정을 할 수 있습니다. 이와 같은 협업 환경은 팀원들이 공동의 목표를 향해 나아가도록 장려합니다.

비즈니스 전반에 미치는 영향

데이터 인텔리전스는 비즈니스의 모든 측면에 대한 기능과 효율성을 개선합니다. 이는 인간의 요구 및 윤리적 기준에 더 부응할 수 있도록 하는 진화의 원동력이 됩니다. 이를 통해 전체 데이터 및 AI 생태계를 향상하여 복잡한 문제를 해결할 수 있도록 합니다. 이 섹션에서는 데이터 인텔리전스가 지형을 형성하는 주요 방식에 대해 설명합니다.

데이터 품질 및 무결성 개선

데이터의 무결성 및 품질은 AI 시스템과 분석 프로세스의 유효성에 있어 기초적인 요소입니다. 데이터 인텔리전스는 다양한 데이터 소스에 대해 데이터 검증, 정리 및 일관된 관리를 위한 메커니즘을 지원하여 이러한 측면을 개선합니다.

혁신 및 새로운 비즈니스 모델 추진

데이터 인텔리전스는 혁신을 지원하며 새로운 비즈니스 모델 형성에 매우 중요합니다. 기업은 데이터를 분석하여 새로운 트렌드와 충족되지 않은 시장의 니즈를 파악할 수 있습니다. 이를 통해 새롭고 혁신적인 제품과 서비스에 대한 기회를 발견할 수 있습니다.


팁

기업은 데이터 기반의 접근 방식을 통해 구독 서비스 또는 온디맨드 플랫폼과 같은 비즈니스 모델을 실험할 수 있으며, 이러한 과정을 통해 경쟁 우위를 확보할 수 있습니다. 데이터에서 얻은 인사이트는 새로운 수익원 및 혁신적인 전략으로 이어질 수 있습니다.

AI 및 ML 가속화

데이터 인텔리전스는 데이터를 AI 및 ML이 사용할 수 있는 형식으로 준비 및 변환하여 이러한 기술을 위한 기반을 지원합니다. 정확하고 신뢰성 있는 AI 모델을 훈련하려면 관리가 잘 된 고품질의 데이터가 필수적입니다.

데이터 인텔리전스 플랫폼의 주요 기능 평가

데이터 인텔리전스 플랫폼은 기업이 데이터를 가치 있는 비즈니스 자산으로 만드는 데 필요한 도구를 제공합니다. 데이터를 통합 플랫폼에 모아서 분석한 다음 효과적인 전략을 수립합니다. 이러한 플랫폼의 주요 기능을 파악하면 데이터 요구사항과 목표에 맞는 플랫폼을 선택하는 데 도움이 됩니다.


팁

데이터 인텔리전스 플랫폼 평가 시 조직은 확장성, 성능, 사용 편의성, 통합 기능과 같은 요소를 고려하여 선택한 플랫폼이 특정 비즈니스 요건 및 기술 인프라에 부합하는지 확인해야 합니다.

NLP 사용


팁

NLP 기술은 번역 소프트웨어, 챗봇, 검색 엔진과 같은 도구의 핵심입니다. 데이터 인텔리전스 플랫폼은 NLP를 활용하여 고객 리뷰, SNS 게시물, 지원 티켓과 같은 기업의 비정형 데이터에 숨어 있는 잠재력을 최대한 활용할 수 있습니다.

성장을 위한 데이터 보안 및 확장성 보장

보안 및 확장성은 모든 조직이 성장하는 데 필수적인 요소입니다. 데이터 인텔리전스는 강력한 보안 기능을 제공하여 개인 정보를 보호하고 데이터 규정을 준수하도록 합니다. 또한, 증가하는 데이터와 늘어나는 조직의 니즈를 수용할 수 있도록 확장할 수 있어야 합니다.

다양한 사용자가 사용할 수 있는 플랫폼 만들기

데이터 인텔리전스 플랫폼은 다양한 기술 전문성 수준을 가진 사람들이 액세스할 수 있어야 합니다. 이렇게 하면 조직 내 데이터 과학자부터 비즈니스 분석가에 이르는 다양한 사용자들이 플랫폼의 기능을 활용할 수 있습니다.


팁

코드가 없는 도구와 직관적인 인터페이스를 통해 사용자 환경을 단순화하면 데이터 사용 범위를 넓히고 더 많은 이해관계자가 정보에 근거한 데이터 기반의 의사결정을 할 수 있습니다.

데이터 프로세스 자동화

데이터 인텔리전스 플랫폼의 자동화는 기업이 방대한 양의 데이터를 처리하는 방식을 혁신합니다. 기업은 데이터 프로세스에 자동화를 통합하여 효율성, 정확성, 속도를 크게 향상할 수 있습니다. 워크플로우를 간소화하고, 수동적 개입을 줄이며, 전반적인 데이터 관리 환경을 개선합니다.


기억하세요

자동화의 가장 중요한 이점 중 하나는 수동 데이터 처리의 필요성을 줄여준다는 것입니다. 수동 작업은 시간을 소모하며 오류가 발생하기 쉽습니다. 자동화는 사람이 입력해야 하는 필요성을 최소화하여 데이터 처리 실수가 발생할 위험을 줄여줍니다.

팁

기업은 데이터 수집, 정리, 처리와 같은 작업을 자동화하여 운영 효율성을 개선하고 시간을 절약하며 전략적 이니셔티브에 리소스를 적용할 수 있습니다.

다양한 산업에서의 데이터 인텔리전스 사용 사례 살펴보기

데이터 인텔리전스는 금융부터 의료, 에너지까지 다양한 산업에서 사용됩니다. 데이터 기반의 인사이트를 사용하면 비즈니스 운영 방식이 달라집니다. 이 섹션에서는 여러 분야의 다양한 사용 사례를 접해 보고, 기업들이 데이터 인텔리전스를 통해 고객 이해, 프로세스 개선, 사기 적발 측면에서 어떻게 도움을 얻는지 알아봅니다. 다음은 몇 가지 예시입니다

- **금융:** 이 분야에서는 재무 위험 처리, 경제 동향 예측, 규정 준수에 데이터 인텔리전스를 사용합니다. 은행 및 기타 금융 기관은 신용도 평가, 사기 적발, 고객 분류에 데이터 분석을 사용합니다.
- **리테일 및 CPG:** 고객 선호도 파악, 원활한 재고 관리, 공급망 최적화, 개인 구매자를 위한 맞춤형 마케팅에 데이터 인텔리전스를 활용합니다.
- **공공 부문:** 공공 부문에서 데이터 인텔리전스는 서비스 개선 및 정책 선택에 중요한 역할을 합니다. 정부 기관은 데이터를 사용하여 변화하는 경제를 모니터링하고 더 유익한 서비스를 제공합니다.
- **보험:** 보험 회사는 위험 평가, 보험 상품의 가격 책정, 거짓 청구 발견에 데이터 인텔리전스를 사용합니다. 대량의 데이터를 분석하여 위험을 더 명확하게 파악하고 청구서 제출을 더 효율적으로 처리할 수 있습니다.
- **의료:** 의료 기관은 환자 치료 개선, 비용 관리, 연구 수행에 데이터 인텔리전스를 사용합니다. 데이터 분석은 의료 결정 및 효과적인 치료 방법을 찾는 데 도움이 됩니다.
- **에너지:** 에너지 회사는 에너지 사용 추적 및 예측, 전력망 효율성 개선에 데이터 분석을 사용합니다.

기억하세요

데이터 인텔리전스 애플리케이션은 산업별로 다를 수 있으나, 공통된 목표는 동일합니다. 데이터에서 가치 있는 인사이트를 추출하고 이를 활용하여 비즈니스 성장을 촉진하고 고객 경험을 향상하는 것입니다.

2장 주요 내용

- » 레이크하우스가 없는 경우의 문제점 살펴보기
- » 레이크하우스, 데이터 웨어하우스 및 데이터 레이크 알아보기
- » 생성형 AI와 기존 AI 살펴보기
- » AI의 잠재력 이해하기
- » 레이크하우스 아키텍처와 생성형 AI 사용하기
- » 데이터 인텔리전스 플랫폼 배포하기

2장
레이크하우스와 생성형 AI 및 기존 AI 알아보기

데이터 인텔리전스 플랫폼은 생성형 인공지능(AI)을 갖춘 레이크하우스 아키텍처를 기반으로 구축되며 조직에서 데이터와 AI를 민주화할 수 있는 강력한 방식을 제안합니다. 레이크하우스 아키텍처는 방대한 양의 정형 또는 비정형 데이터를 통합된 단일 환경에 저장 및 처리하여 데이터 웨어하우징, 비즈니스 인텔리전스, 기존 AI 및 생성형 AI를 새로운 차원으로 끌어올립니다.

2장에서는 레이크하우스, 데이터 웨어하우스(DW), 데이터 레이크 간의 차이점과 레이크하우스에 생성형 AI 및 기존 AI를 추가하여 조직에 대한 가치를 향상하는 방법을 알아봅니다.

레이크하우스가 없는 경우의 문제점 알아보기

대다수의 기업은 비즈니스 목표를 이루기 위해 데이터와 AI를 효과적으로 결합하는 데 문제를 겪습니다. 이러한 문제에는 데이터 인텔리전스 생태계에 각기 중요한 역할을 하는 다양한 구성 요소가 포함되어 있습니다. 데이터 관리와 AI를 결합하는 데 다음과 같은 문제가 있습니다.

- 데이터와 AI가 사일로화됩니다. 데이터 사일로는 운영 비용을 높입니다.
- 데이터 프라이버시 및 제어에 문제가 생깁니다. 일관성 없는 정책으로 데이터에 대한 신뢰도가 저하됩니다.
- 고도로 숙련된 엔지니어에게 의존합니다. 서로 다른 도구는 팀 간 생산성을 떨어뜨립니다.

제 기능을 하려면 여러 서비스를 함께 연결해야 합니다. 각 구성 요소에는 저마다의 문제점이 있습니다. 그림 2-1을 통해 데이터 레이크부터 시계 방향으로 각 구성 요소와 그에 대한 문제점을 알 수 있습니다.

- **데이터 레이크:** 문제는 방대한 양의 비정형 데이터 세트를 저장 및 관리하는 데 있습니다.
- **머신 러닝(ML):** 복잡한 알고리즘의 정확성을 실험, 개발, 적용 및 모니터링해야 하는 문제가 있습니다.
- **스트리밍:** 연속적인 데이터 스트림을 실시간으로 처리해야 하는 기술이 필요합니다.
- **생성형 AI:** AI 기술을 통해 새롭고 현실성 있는 콘텐츠를 만드는 데 복잡한 문제가 있습니다.
- **데이터 웨어하우스:** 문제는 분석을 위해 구조화된 데이터를 중앙 집중화하는 것인데 이렇게 하려면 복잡해지고 비용이 증가할 수 있습니다.
- **비즈니스 인텔리전스(BI):** 데이터를 효과적으로 시각화하고 비즈니스 인사이트를 추출하는 데 어려움이 있습니다.
- **오케스트레이션 및 추출, 변환, 로드(ETL):** 여기에는 데이터 준비와 이동을 조정하는 작업이 수반됩니다.

- **거버넌스:** 규정을 탐색하고 강력한 데이터 관리 및 보안 제어를 구현하는 것이 문제입니다.
- **데이터 과학:** 과학적 방법을 사용하여 데이터를 탐색 및 분석하는 경우의 작업은 복잡합니다.

그림 2-1: 데이터 인텔리전스 생태계의 문제점.

이러한 구성 요소와 Databricks 데이터 인텔리전스 플랫폼이 이 문제점을 해결하는 방법에 대한 자세한 내용은 3장에서 확인할 수 있습니다.

데이터 인텔리전스 플랫폼에는 모든 데이터 유형을 저장 및 관리하기 위한 개방형 데이터 레이크, 신뢰성과 공유를 위한 통합 데이터 스토리지, 통합 보안, 거버넌스, 카탈로그 환경, 그리고 데이터의 시맨틱을 이해하는 AI 기반 엔진 등 조직 전체에 도움이 되는 다양한 요소가 있습니다. 데이터 인텔리전스 플랫폼은 데이터 과학, AI, ETL, 실시간 분석, 오케스트레이션, 데이터 웨어하우징의 경험을 개선합니다.

레이크하우스와 데이터 웨어하우스 및 데이터 레이크의 비교

레이크하우스는 데이터 저장 및 분석에 대한 접근 방식이 DW 및 데이터 레이크와 확연히 다릅니다. DW와 데이터 레이크에서 레이크하우스로의 전환은 방대한 양의 정형 및 비정형 데이터를 관리하기 위해 더욱 확장형의 개방적이면서 비용 효율적인 솔루션이 필요함에 따라 일어납니다. 이 섹션에서는 차이점을 알아봅니다.

개방형 아키텍처

Databricks를 사용하면 독점 형식과 폐쇄적인 생태계를 벗어나 언제나 데이터를 제어할 수 있습니다. Databricks 레이크하우스는 널리 채택된 오픈 소스 프로젝트인 Apache Spark, Delta Lake, MLflow를 기반으로 합니다. 이외에도 Delta Sharing은 비용이 많이 드는 복제와 복잡한 ETL 없이 레이크하우스에서 모든 컴퓨팅 플랫폼으로 라이브 데이터를 안전하게 공유할 수 있는 개방형 솔루션을 제공합니다.

통합 아키텍처

레이크하우스 아키텍처는 모든 통합, 스토리지, 처리, 거버넌스, 공유, 분석, AI를 통합합니다. 정형 및 비정형 데이터 작업을 위한 접근 방식, 데이터 계보 및 출처에 대한 엔드투엔드 보기, 모든 Python, R, Scala, SQL을 위한 하나의 노트북, 배치 및 스트리밍을 위한 하나의 소스, 3대 클라우드 제공업체를 위한 하나의 플랫폼이 있습니다.

확장형

레이크하우스는 기존 DW와 데이터 레이크보다 더 크게 확장할 수 있으며 낮아진 비용과 높아진 성능으로 레코드를 최대 수 조개까지 확장할 수 있습니다. 성능을 위한 자동 최적화를 제공하며, 스토리지는 세계 최고 수준의 성능을 갖추어 모든 데이터 플랫폼 중 가장 낮은 총소유비용(TCO)을 보장합니다.

데이터 거버넌스 및 보안 개선

레이크하우스는 조직의 모든 데이터 및 AI 액세스에 대해 하나의 보안 및 거버넌스 모델을 사용하여 데이터를 관리 및 보호하는 방식을 개선합니다. 하나의 통합 거버넌스 플랫폼을 사용하면 여러 개의 분산된 거버넌스 솔루션과 다양한 데이터 종류가 있어 일관된 정책과 보호 기능을 적용하기 어려웠던 기존 DW 및 데이터 레이크보다 더 쉽게 규정을 준수하고 데이터를 더욱 안전하게 보관할 수 있습니다.

생성형 AI와 기존 AI의 구분

생성형 AI와 기존 AI는 AI의 두 가지 분야로 나뉩니다. 기존 AI는 각 매장의 향후 매출을 예측하거나 수백만 명의 고객을 여러 세그먼트로 분류하는 등 수치로 예측하거나 항목을 분류해야 하는 분야에서 유용합니다. 반면에 생성형 AI는 텍스트 요약을 만들거나 PowerPoint, PDF, Word 문서와 같은 비정형 데이터에서 일반적인 질문에 답변하기에 유용합니다.

생성형 AI는 트레이닝 데이터에서 학습한 패턴을 기반으로 새로운 콘텐츠를 생성합니다. 생성형 AI 시스템은 데이터를 분석하는 것뿐만 아니라 텍스트, 이미지, 오디오, 동영상이나 기타 미디어를 해석하고 검색할 수 있습니다. 또한 생성형 AI는 내부적으로 새로운 텍스트를 생성하고 소프트웨어 코드를 작성 및 편집하는 용도로 사용할 수 있습니다.

기억하세요

생성형 AI의 핵심 기능은 스타일과 구조가 트레이닝 데이터와 유사하지만 직접 복사하지 않은 새로운 콘텐츠를 만들 수 있는 것입니다. 대규모 언어 모델(LLM)과 같은 생성 모델은 데이터의 기초 패턴을 식별하고 학습함으로써 개념을 결합하여 독창적인 창작물을 만들 수 있습니다. 다음은 비즈니스를 위한 생성형 AI의 몇 가지 예시입니다.

- **텍스트 생성:** 사람처럼 문장과 설명을 작성할 수 있으며 기업의 자체 데이터로 학습됩니다
- **텍스트 요약:** 대량의 문서를 가져와 사람이 쉽게 해석할 수 있도록 시놉시스 또는 등급을 부여할 수 있습니다
- **소프트웨어 코드 작성:** 간단한 프롬프트를 가져와 SQL, Python, Scala, R로 코드를 작성할 수 있습니다
- **데이터 자산 문서화:** 더 나은 시맨틱 검색을 위해 테이블과 열에 대한 내용을 설명할 수 있습니다

데이터 인텔리전스 향상에 대한 AI 중요성 인식

데이터 인텔리전스를 AI와 결합하면 기업이 데이터를 분석, 이해, 활용하는 방식이 큰 발전을 이룰 수 있습니다. 이러한 영향은 조직의 전략적 역량을 강화합니다. 이를 통해 기업은 시장 변화와 소비자 니즈에 더 민첩하게 적응할 수 있습니다.

기억하세요

데이터 인텔리전스 향상에 대한 AI의 중요성은 기술적인 능력뿐만 아니라 다양한 분야에서 혁신을 주도할 수 있는 능력에 있습니다.

레이크하우스 아키텍처와 생성형 AI의 활용

레이크하우스 아키텍처와 생성형 AI를 통합하면 두 기술의 역량이 향상됩니다. 이러한 통합으로 데이터 처리 및 분석을 위한 더 강력한 환경이 조성됩니다.

레이크하우스 아키텍처 활용

레이크하우스 아키텍처는 데이터 레이크와 데이터 웨어하우스가 가진 최고의 기능을 결합하여 데이터 저장 및 관리 기반을 제공합니다. 이로써 조직은 정형 및 비정형 데이터를 단일 리포지토리에 저장하면서 분석 및 ML 작업을 수행할 수 있습니다.

기억하세요

생성형 AI를 통합하면 데이터 기반 인사이트를 분석 및 생성하는 새로운 방식이 도입됩니다. 이 기술을 사용하여 조직은 데이터 품질을 개선하고 더 정확한 예측 모델을 개발할 수 있으며, 이를 통해 경쟁 우위를 확보할 수 있습니다.

조직은 레이크하우스 아키텍처와 생성형 AI를 통합함으로써 데이터를 더욱 효과적으로 관리하고 데이터 기반 의사결정을 위한 새로운 가능성을 찾을 수 있습니다.

개방형 데이터 스토리지 활용

개방형 데이터 스토리지를 활용하면 신뢰성과 데이터 공유에 도움이 됩니다. 이는 레이크하우스 아키텍처를 활용하고 생성형 AI 기능을 통합하는 데 필수적입니다. 이는 효율적인 데이터 스토리지 계층을 제공하여 조직이 생성형 AI 애플리케이션을 쉽고 효율적으로 개발 및 배포할 수 있도록 지원하는 기술입니다. 조직은 이를 활용하여 생성형 AI의 능력을 이용해 가장 효과적으로 혁신을 추진할 수 있습니다.

레이크하우스에 생성형 AI 기능 통합

레이크하우스 아키텍처에 생성형 AI 기능을 통합하면 데이터 분석이 향상됩니다. 조직은 레이크하우스에 결합된 강점을 활용하여 데이터 인텔리전스 활동을 한 단계 높일 수 있습니다. 이 통합의 몇 가지 주요 이점은 다음과 같습니다.

- 데이터 작업 자동화: 생성형 AI는 레이크하우스 내 데이터 운영을 간소화할 수 있습니다. 기존 AI는 데이터 정리를 자동화할 수 있으나, 생성형 AI는 모델 테스트 및 훈련을 위한 인공 데이터를 생성하여 강력한 분석과 AI 애플리케이션을 보장함으로써 추가적인 지원을 할 수 있습니다.
- 검색 기능 강화: AI는 레이크하우스 내 지능형 검색 기능을 강화합니다. 사용자는 자연어 쿼리를 활용하여 데이터 자산 간의 관계를 효율적으로 발견하고 이해할 수 있습니다. 이렇게 하면 단순한 키워드 검색에 그치지 않고 데이터 검색을 간소화하며 분석에 적합한 데이터 세트에 쉽게 액세스할 수 있습니다.
- 맞춤형 AI 애플리케이션 개발: 조직은 레이크하우스 프레임워크에 AI를 통합하여 특정 니즈에 맞는 애플리케이션을 만들 수 있습니다. 예를 들어, 기업의 자체 데이터에 대한 LLM을 만들고, 예측 모델을 개발하며, 추천 엔진을 사용자 지정하거나 복잡한 보고 작업을 자동화하는 등의 작업을 수행할 수 있습니다.

기억하세요

레이크하우스 아키텍처에 생성형 AI 기능을 통합하면 조직은 데이터에서 더 많은 가치를 창출할 수 있습니다.

데이터팀과의 협업 활성화

레이크하우스 아키텍처와 생성형 AI 기능의 결합은 데이터팀의 협업 가능성을 크게 높입니다. 이를 통해 더욱 활발한 아이디어 교환이 가능해집니다. 이는 혁신 문화를 조성합니다. 데이터팀은 함께 협력하여 AI 모델을 더욱 효율적으로 구축, 훈련 및 배포할 수 있습니다. 이로써 레이크하우스의 데이터 관리 강점과 생성형 AI의 창의적인 잠재력을 활용하여 비즈니스 성장을 촉진할 수 있습니다.

AI를 통한 데이터 분석 및 인사이트 강화

AI를 통해 데이터 분석을 강화하려면 다양한 AI 기반 도구와 기술을 활용하여 여러 단계로 구성된 데이터 분석 프로세스를 자동화 및 간소화합니다. AI는 다음과 같은 방식으로 데이터 분석의 여러 단계에 통합될 수 있습니다.

- **데이터 준비:** AI는 데이터 정리, 체계화, 전처리를 포함한 데이터 준비 단계를 자동화할 수 있습니다. AI 도구는 데이터 품질 문제를 감지 및 수정하고, 비정형 데이터에서 정보를 추출하며, 다른 형식의 데이터를 결합할 수 있습니다.
- **데이터 탐색:** AI 알고리즘은 자연어를 사용해 데이터를 탐색할 수 있습니다. 이렇게 하면 사람에게는 보이지 않는 인사이트를 발견할 수 있습니다.
- **데이터 해석:** AI는 데이터에서 요약, 인사이트 또는 스토리를 생성하여 데이터 해석을 개선할 수 있습니다. 데이터를 기반으로 인과관계를 파악하고 향후 결과 또는 행동을 예측할 수 있습니다.
- **데이터 품질:** AI는 데이터 및 모델 품질이 왜곡된 경우에 이를 감지하여 자동으로 플래그를 지정하고 수정할 수 있도록 지원합니다.

복잡한 데이터 작업 및 프로세스 자동화

복잡한 데이터 관련 작업 및 프로세스의 자동화는 기술을 사용하여 데이터를 더 효율적으로 처리 및 분석하는 것을 의미합니다. 데이터 레이크에서 대량의 비정형 데이터를 체계화하고, 생성형 AI와 ML 모델을 구축 및 적용하며, 지속적인 데이터 흐름을 실시간으로 처리합니다.

오케스트레이션 작업의 경우에는 AI가 요건에 맞는 적합한 인스턴스와 시작 시간을 자동으로 선택할 수 있습니다. 자동 확장 및 오류 수정과 같은 작업을 사용자 대신 처리합니다.

테이블에 대한 파일 크기 최적화 등 데이터 엔지니어링의 여러 측면은 AI의 이점을 활용하고 자동화할 수 있습니다. 일반적으로 엔지니어는 데이터를 읽거나 쓰는 목적으로 최적화된 파일 크기를 파악하기 위해 무수한 시간과 전문 지식을 투자하는데, 이는 유의미한 성능 개선으로 이어질 수 있습니다. 이 복잡한 작업을 자동화하면 판도를 바꿀 게임 체인저가 될 수 있습니다.

ETL 프로세스의 지능형 자동 확장 기능은 데이터 볼륨과 처리 요구사항에 따라 지정된 한도까지 리소스를 자동을 조정하여 클러스터 활용도를 최적화하고 스트리밍 워크로드의 엔드투엔드 지연을 최소화합니다. 데이터 도착이 처리 속도를 능가하는 경우에는 효율적으로 확장하고, 저부하인 경우에는 축소하여 종료 전에 작업을 완료해 인프라 비용을 절감합니다.

데이터 인텔리전스 플랫폼 배포

데이터 인텔리전스 플랫폼은 데이터 과학자, 데이터 엔지니어, 아키텍트, 비즈니스 분석가 등 다양한 사용자를 아우르는 통합 플랫폼으로 조직이 혁신할 수 있도록 지원합니다. 여러 단계의 데이터를 통합된 단일 환경으로 결합합니다. 다음은 이 통합을 활성화하는 기능입니다.

- **데이터 통합:** 다양한 소스의 데이터를 한 곳으로 가져올 수 있습니다. 데이터베이스, 데이터 웨어하우스, 데이터 레이크, 스트리밍 데이터 소스의 데이터 통합을 지원하기 때문에 하나의 플랫폼에서 모든 데이터로 더 수월하게 작업할 수 있습니다.
- **처리 및 분석:** 데이터를 플랫폼에 저장한 다음에 데이터를 처리 및 분석할 수 있습니다. 이 플랫폼은 Python, R, Scala, SQL 등 사용자가 선호하는 모든 주요 언어를 지원합니다. 기본으로 제공되는 기능과 라이브러리를 사용하여 데이터를 더 쉽게 정리, 변환, 분석할 수 있습니다.
- **워크스페이스에서의 협업:** 이 플랫폼은 여러 팀원이 동일한 데이터와 프로젝트에서 함께 작업할 수 있는 공유 워크스페이스를 제공합니다. 데이터 엔지니어, 데이터 과학자, 분석가는 모두 동일한 플랫폼에서 시각화 및 대시보드를 만들 수 있습니다. 공유 워크스페이스는 버전 관리를 유지관리할 수 있기 때문에 모든 사용자가 최신 데이터와 분석을 사용할 수 있습니다.

- **통합된 장소 사용:** 한 곳에서 전체 데이터 분석 워크플로우를 관리하여 데이터에 대한 액세스를 제어하고, 리소스를 관리하며, 작업을 모니터링할 수 있습니다. 이를 통해 진행 중인 작업에 대한 리소스 할당과 모니터링을 개선할 수 있습니다.
- **원활한 배포:** 데이터 분석 솔루션을 구축한 다음에는 프로덕션에 쉽게 배포할 수 있습니다. 원활한 배포를 통해 조직은 데이터 인텔리전스 플랫폼이 없을 때 일반적으로 발생하는 문제점 없이 데이터 프로젝트를 개발에서 프로덕션으로 이동할 수 있습니다.

이 플랫폼에 대한 자세한 내용은 3장을 참조하세요.

주요 내용

- Databricks 데이터 인텔리전스 플랫폼 검토하기
- Databricks 플랫폼 아키텍처 살펴보기
- 개발자 활동 지원하기

3장 Databricks 데이터 인텔리전스 플랫폼 시작하기

기업은 데이터 아키텍처를 간소화하면서 유의미한 인사이트를 얻을 수 있는 역량을 향상할 수 있는 방법을 끊임없이 찾고 있습니다. 3장에서는 Databricks 데이터 인텔리전스 플랫폼을 시작하기 위한 기본적인 측면을 살펴봅니다.

Databricks 데이터 인텔리전스 플랫폼 도입

Databricks 데이터 인텔리전스 플랫폼으로 조직 전체가 데이터와 인공지능(AI)을 사용할 수 있습니다. 레이크하우스를 기반으로 구축되어 모든 데이터, AI, 거버넌스 요구사항을 위한 개방적이면서 통합된 기반을 지원하며 데이터의 고유성을 이해하는 데이터 인텔리전스 엔진으로 구동됩니다.

추출, 변환, 로드(ETL)에서 데이터 웨어하우징, 생성형 AI에 이르기까지 Databricks는 데이터 및 AI 목표를 간소화하고 가속화할 수 있도록 합니다.

DatabricksIQ를 통한 데이터 인텔리전스 제공

Databricks는 생성형 AI의 강력함에 레이크하우스 아키텍처의 포괄적인 기능을 결합하여 데이터 인텔리전스 엔진인 DatabricksIQ를 만들어냈습니다. DatabricksIQ로 비즈니스와 데이터의 고유한 뉘앙스를 학습하여 다양한 사용 사례에서 자연어 액세스를 강화합니다. 조직의 전 직원이 자연어로 데이터를 검색하고, 이해하며, 질의할 수 있습니다. DatabricksIQ는 데이터, 사용 패턴, 트렌드에 대한 정보를 사용해 비즈니스 전문용어와 고유한 데이터 환경을 이해하고 번역, 요약, 질의응답, 텍스트 생성 등 광범위한 언어 관련 작업을 수행할 수 있는 생성형 AI의 한 형태인 대규모 언어 모델(LLM)을 단순하게 사용하는 것보다 훨씬 더 나은 답변을 제공합니다.

물론 LLM은 데이터에 언어 인터페이스를 지원하기로 했으며 대다수의 데이터 기업에서 AI 어시스턴트를 추가하고 있기는 하지만, 실제로 이러한 솔루션 중 상당수는 엔터프라이즈 데이터에 미치지 못합니다. 모든 기업은 비즈니스 질문에 답변하는 데 필요한 고유한 데이터 세트, 전문용어와 내부 지식을 보유하고 있으며, 단순히 인터넷에서 학습된 LLM을 통해 질문에 답변하는 것은 잘못된 결과를 초래하기도 합니다. 고객 정의 또는 회계연도처럼 간단한 것조차 기업마다 다릅니다.

DatabricksIQ는 기업의 비즈니스와 데이터 개념을 자동으로 학습하여 이 문제를 직접 해결하는 데이터 인텔리전스 엔진입니다. Unity 카탈로그(UC), 대시보드, 노트북, 데이터 파이프라인, 문서 등 Databricks 플랫폼의 시그널을 사용하여 Databricks 플랫폼의 고유한 엔드투엔드 특성을 활용해 데이터가 실제로 사용되는 방식을 확인할 수 있습니다. DatabricksIQ는 기업을 위한 매우 정확한 전문 모델을 구축합니다.

자연어를 통한 사용자 경험 간소화

자연어 처리(NLP)를 사용하면 Databricks에 대한 사용자 경험이 크게 간소화됩니다. Databricks 데이터 인텔리전스 플랫폼은 조직에서 사용되는 특정 용어를 이해할 수 있도록 설계되었습니다. 따라서 동료에게 질문하는 것만큼이나 간단하게 데이터를 검색하고 찾을 수 있습니다.

팁

NLP를 사용하면 시스템이 언어를 이해하고 해석할 수 있습니다. 이 기능은 새로운 데이터 애플리케이션을 개발하는 데까지 확장됩니다. 코드 작성, 오류 수정, 답변 제공을 지원합니다. 이렇게 하면 개발 프로세스의 속도를 높일 수 있습니다.

프라이버시 및 거버넌스 확보

데이터와 AI 애플리케이션에서 강력한 거버넌스와 보안의 필요성이 그 어느 때보다 중요해졌습니다. Databricks는 거버넌스와 보안에 대한 통합 접근 방식으로 지원되는 머신러닝 운영(MLOps) 및 AI 개발을 위한 포괄적인 솔루션을 제공합니다. 이 솔루션을 통해 다양한 AI 이니셔티브를 추진할 수 있으며 지적 재산에 대한 프라이버시 보호 및 제어를 유지관리할 수 있습니다.

기억하세요

MLOps는 ML 엔지니어링의 한 기능입니다. ML 모델을 프로덕션에 적용하고 데이터 변화에 따라 유지관리 및 모니터링하는 프로세스를 간소화하는 데 중점을 둡니다.

데이터 인텔리전스 플랫폼 사용

Databricks는 데이터 레이크하우스와 생성형 AI의 능력을 활용하는 데이터 인텔리전스 플랫폼을 만들었습니다. 데이터 레이크하우스 플랫폼 내에서 AI의 잠재력을 탐구하는 데 상당한 진전을 이뤘습니다. Databricks 데이터 인텔리전스 플랫폼은 데이터와 AI를 모두 포괄하는 통합 거버넌스 계층이 두드러집니다. 또한 ETL, SQL, ML, 비즈니스 인텔리전스(BI)를 아우르는 단일 쿼리 엔진도 있습니다. 그리고 Mosaic AI를 통합하면 DatabricksIQ를 지원하는 AI 모델 개발에 도움이 됩니다. 이러한 통합은 조직의 전 직원이 데이터에 액세스할 수 있도록 하는 것이 매우 중요합니다.

기억하세요

Databricks는 레이크하우스 개념을 정립했습니다. 모든 데이터와 거버넌스 요구사항을 충족하는 개방형의 통합 아키텍처를 제공합니다. 또한 조직은 하나의 시스템에서 정형 및 비정형 데이터를 저장하고 관리할 수 있습니다.

팁

Databricks는 Photon(합리적인 비용으로 굉장히 신속한 쿼리 성능을 선사하는 차세대 엔진)과 같은 성능 향상 기법을 개발하여 플랫폼의 확장성과 효율성을 높였습니다. 이렇게 하면 Databricks는 대규모 데이터 워크로드도 처리할 수 있습니다.

Databricks 데이터 인텔리전스 플랫폼의 아키텍처를 시각적으로 표현한 그림 3-1을 확인할 수 있습니다. 이 섹션에서는 그림 아래쪽부터 각 구성 요소를 살펴보고 모든 구성 요소가 어떻게 들어맞는지 알아봅니다.

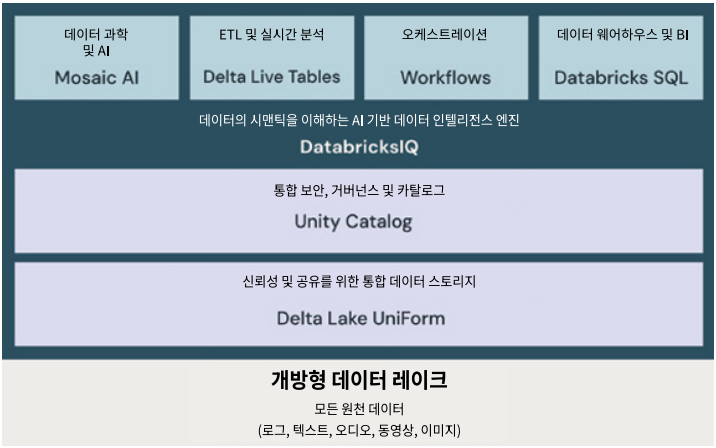


그림 3-1: **Databricks** 데이터 인텔리전스 플랫폼

개방형 데이터 레이크

Databricks를 사용하면 독점 형식과 폐쇄적인 생태계를 벗어나 언제나 데이터를 제어할 수 있습니다. 데이터 레이크는 이미지, 동영상, 오디오, 반정형 데이터, 텍스트 등 여러 새로운 데이터 애플리케이션에 필요한 데이터 유형을 저장, 정제, 분석, 액세스하는 데 사용할 수 있습니다.

Delta Lake UniForm

Delta Lake Universal Format(UniForm)을 사용하면 즐겨 쓰는 Iceberg 또는 Hudi 클라이언트로 UC 엔드포인트를 통해 Delta 테이블을 읽을 수 있습니다. DatabricksIQ는 AI 모델을 사용하여 일반적인 데이터 스토리지 문제를 해결하기 때문에 시간이 지나 테이블이 바뀌더라도 수동으로 관리할 필요 없이 더 빨라진 성능을 경험할 수 있습니다.

기억하세요

스토리지에는 Delta Lake, Apache Iceberg, Apache Hudi와 같이 크게 세 가지 형식이 있습니다. 이전에는 기업들이 여러 장소와 형식으로 다수의 복사본을 복제하여 보관했습니다. 이 방식은 비용과 시간이 많이 드는 만큼 비용과 수고가 두 배로 가중됩니다.

Databricks는 데이터를 반복적으로 복사하지 않고도 이 세 가지 형식 중 하나에 데이터를 저장하고 처리(비즈니스 인텔리전스, AI 등)할 수 있도록 Delta Lake UniForm을 도입했습니다.

Unity Catalog

Databricks UC는 Databricks 데이터 인텔리전스 플랫폼 내에서 데이터와 AI를 위한 통합 거버넌스 계층을 제공합니다. UC를 사용하면 조직은 모든 클라우드 또는 플랫폼에서 정형 및 비정형 데이터, ML 모델, 노트북, 대시보드, 파일을 원활하게 관리할 수 있습니다. 데이터 과학자, 분석가, 엔지니어는 UC를 사용함으로써 신뢰할 수 있는 데이터와 AI 자산을 안전하게 검색, 액세스, 협업하고 AI를 활용하여 생산성을 높이며 레이크하우스 아키텍처의 잠재력을 최대한 활용할 수 있습니다. 거버넌스에 대한 통합 접근 방식은 데이터와 AI 이니셔티브를 가속화하면서 규정 준수를 간소화합니다.

DatabricksIQ

DatabricksIQ는 UC를 기반으로 구축되고 관리됩니다. DatabricksIQ는 UC의 모든 데이터 자산에 대한 설명과 태그를 자동으로 삽입하여 UC의 거버넌스를 개선합니다. 그리고 이러한 자산을 활용하여 전체 플랫폼에서 전문용어, 약어, 메트릭, 시맨틱을 인식할 수 있도록 합니다. 이 프로세스로 시맨틱 검색, AI 어시스턴트 품질, 그리고 거버넌스 수행 능력이 향상됩니다.

또한 DatabricksIQ는 Databricks의 제품 내 검색 기능을 크게 향상합니다. 새로운 검색 엔진은 단순히 데이터를 찾는 데 그치지 않고 데이터를 해석하고 정렬하며 실행 가능한 컨텍스트 형식으로 제시하여 모든 사용자가 데이터로 더 빠르게 시작할 수 있도록 합니다.

팁

자산이 UC에 등록되면 DatabricksIQ는 사용자가 자연어(NL)와 회사별 용어를 사용해 검색할 수 있도록 하여 데이터의 검색 가능성을 크게 향상함으로써 사용자가 조직 내에서 데이터 자산을 더 쉽게 사용할 수 있도록 합니다.

Mosaic AI

Databricks는 Mosaic을 인수하여 데이터 인텔리전스 플랫폼에 통합함으로써 LLM 관련 기능을 대폭 강화했습니다. 이를 통해 사용자는 특정 니즈에 맞는 맞춤형 생성형 AI 애플리케이션을 세밀하게 조정하거나 만들 수 있습니다. 이러한 통합으로 사용자는 처음부터 새롭게 시작하거나 기존 모델을 개선할 수 있으며, 독점 데이터의 프라이버시 보호 및 제어를 확보할 수 있습니다.

기억하세요

플랫폼의 생성형 AI 활용은 데이터 이해도를 높여, 지능형 검색 기능을 강화하고, SQL 코드 생성 및 수정을 지원하며, 데이터 테이블과 열에 대한 자세한 설명을 자동으로 생성하는 시맨틱 이해가 가능하도록 합니다. Mosaic과 Databricks의 생성형 AI 기능의 특징을 결합하면 데이터 보안과 사용자 자율성에 중점을 둔 강력한 AI 애플리케이션 개발 환경을 구축할 수 있습니다.

Mosaic AI에 대한 자세한 내용과 이 플랫폼을 통해 자체적인 생성형 AI 애플리케이션을 구축하는 방법은 4장에서 확인할 수 있습니다.

Delta Live Tables

Delta Live Tables(DLT)는 데이터팀이 비용 효율적으로 스트리밍과 배치 ETL을 간소화할 수 있도록 지원하는 Databricks 데이터 인텔리전스 플랫폼을 위한 선언적 ETL 프레임워크입니다. 데이터에서 수행할 변환을 정의하기만 하면 DLT 파이프라인이 작업 오케스트레이션, 클러스터 관리, 모니터링, 데이터 품질 및 오류 처리를 자동으로 관리합니다.

DLT를 통해 사용자가 ETL이 해야 할 작업을 설명하면 데이터 인텔리전스 엔진이 데이터 및 변환을 이해하고 처리하기 위해 워크로드를 자동으로 확장합니다. DatabricksIQ는 모든 것을 처리하며, 최적의 총소유비용을 위해 필요한 사항만 업데이트합니다. 또한 새로운 데이터가 추가되면 엔진이 기초 테이블을 업데이트하는 최선의 방법을 찾아내 스트리밍/실시간 ETL을 합리적으로 할 수 있습니다. 기본적인 데이터 품질 및 모니터링은 다운스트림 비즈니스 앱을 활성화하는 데 필수적입니다.

기억하세요

ETL은 엔지니어가 다양한 소스에서 데이터를 추출하기 위해 사용하는 프로세스 데이터입니다. 그런 다음 데이터를 사용 가능하고 신뢰할 수 있는 리소스로 변환합니다. 마지막으로 최종 사용자가 액세스하여 다운스트림에서 비즈니스 문제를 해결하는 데 사용할 수 있는 시스템으로 해당 데이터를 로드합니다.

Databricks Workflows

Databricks Workflows는 Databricks 데이터 인텔리전스 플랫폼에서 데이터 처리, ML, 분석 파이프라인을 오케스트레이션합니다. 데이터팀은 광범위한 작업 유형과 심층적인 가시성 기능, 높은 신뢰성을 통해 서버리스 컴퓨팅에서 모든 파이프라인의 자동화 및 오케스트레이션을 향상할 수 있는 도구를 제공받습니다.

데이터 인텔리전스를 핵심으로 하는 Databricks Workflows는 잠재적인 해결책을 제안하여 디버깅과 알림을 간소화하고 모든 상호작용을 분석할 수 있어 데이터 처리를 담당하는 업무와 팀을 쉽게 파악할 수 있습니다. 이렇게 하면 데이터 처리 및 통합 가시성이 간소화됩니다. 작업을 완료하지 못한 경우 워크플로우가 작업을 지능적으로 복구하고 필요한 부분만 다시 실행하므로 총소유비용이 훨씬 절감되고 지능이 향상됩니다.

Databricks SQL

Databricks SQL은 서버리스 데이터 웨어하우스의 선두주자 중 하나입니다. 몇 가지 기능은 다음과 같습니다

- UC를 통한 거버넌스의 추가적인 이점과 함께 ETL 워크로드 및 BI 실행
- 최적의 가격 및 성능으로 확장 가능한 오픈 소스 기반의 아키텍처 사용
- 쿼리 실행 속도의 최적화로 더 쉬워진 데이터 분석
- 쿼리 및 보고서 작성 도구 등 고급 기술 사용으로 데이터 액세스 속도 증가

Databricks SQL은 차세대 벡터화 쿼리 엔진 Photon을 활용하며 수천 가지의 최적화를 통해 모든 도구, 쿼리 유형, 실제 애플리케이션에 최상의 성능을 제공합니다. 여기에 신경망을 통해 지능적으로 데이터를 미리 가져와 인덱싱과 같은 성능 튜닝을 하지 않아도 되는 AI 기반 예측 I/O가 있습니다.

SQL은 다양성, 효율성 및 광범위한 사용으로 인해 데이터 분석에 매우 중요합니다. 단순성으로 대규모 데이터 집합을 신속하게 검색, 조작, 관리할 수 있습니다. 데이터 분석을 위해 SQL에 AI 기능을 통합하면 효율성이 향상되어 비즈니스에서 신속하게 인사이트를 추출할 수 있습니다.

AI 함수는 내장된 Databricks SQL 함수로, SQL에서 LLM에 직접 액세스할 수 있습니다. AI 함수는 LLM 호출의 기술적인 복잡성을 추상화하여 분석가와 데이터 과학자가 기초 인프라에 대한 걱정 없이 이 모델을 사용할 수 있도록 합니다.

DATABRICKS AI/BI 살펴보기

Databricks AI/BI는 데이터 인텔리전스를 기반으로 구축된 동종 최초의 분석 제품으로, 조직의 전 직원이 BI를 사용할 수 있도록 지원합니다. Databricks를 위한 데이터 인텔리전스 엔진인 DatabricksIQ를 기반으로 하는 AI/BI는 고유한 데이터와 비즈니스 개념을 이해합니다. 이를 위해 데이터 플랫폼의 시그널을 선별된 지침과 함께 자동으로 캡처하고 사전에 설명을 찾아 통합하여 사용자가 복잡한 실제 데이터에서 관련성 있고 정확한 AI 생성 인사이트를 얻을 수 있도록 합니다.

대시보드는 분석가가 자연어를 사용하여 비즈니스팀을 위한 대화형 데이터 시각화를 신속하게 구축할 수 있도록 하며 Genie는 비즈니스 사용자가 자체 분석을 직접 실행할 수 있는 대화형 환경을 제공합니다. 경험이 많은 동료에게 질문하듯이 같은 방식으로 질문할 수 있어 기술 전문가에게 의존하지 않고도 데이터에서 직접 신뢰할 수 있는 답변을 구할 수 있습니다.

Databricks AI/BI는 데이터 인텔리전스 플랫폼에 기본으로 탑재되어 데이터 확장에 따른 인터랙티브 성능 저하 없이 즉각적인 인사이트를 제공하면서 Unity 카탈로그를 통해 통합 거버넌스와 세분화된 보안을 확보합니다.

기억하세요

Databricks는 오픈 소스 기술을 기반으로 구축되어 독점 시스템을 사용하여 특정 공급업체에 종속될 수 있는 일부 경쟁사와 차별화를 두었습니다. 개방적인 접근 방식은 오픈 소스 커뮤니티의 기여를 통해 혁신을 촉진합니다.

프로그래머 지원을 위한 DatabricksIQ 사용

Databricks Assistant는 Databricks 노트북, SQL 편집기, 파일 편집기에서 기본으로 제공되는 상황인지형 AI 어시스턴트입니다. Databricks Assistant로 대화형 인터페이스를 통해 데이터를 쿼리할 수 있어 Databricks 내에서 생산성을 높일 수 있습니다. 영어로 작업에 대한 설명을 하면 Assistant가 SQL 쿼리를 생성한 다음 복잡한 코드를 설명하고 자동으로 오류를 수정하도록 할 수 있습니다. Assistant는 UC 메타데이터를 활용해 기업 전반의 테이블, 열, 설명과 인기 있는 데이터 자산을 이해하여 사용자에게 맞춤화된 응답을 제공합니다.

SQL, Python, R, Scala 코드 생성

생성형 AI를 통해 코드 생성을 지원하여 데이터 쿼리 프로세스를 간소화합니다. 데이터의 시맨틱과 사용자 쿼리의 의도를 이해하여 코드를 자동으로 생성할 수 있습니다. 이렇게 하면 데이터 작업 수행에 필요한 시간과 노력이 줄어듭니다.

예를 들어, 자전거가 가장 많이 팔린 10개 도시를 찾는 SQL 프로그램을 작성하거나 직원 연봉을 격주급으로 나누는 Python 프로그램을 작성합니다.

다른 언어로 코드 변환

생성형 AI의 기능 중 하나인 한 프로그래밍 언어에서 다른 프로그래밍 언어로 변환하는 기능으로, 여러 프로그래밍 언어를 사용하는 환경에서 유용합니다. 이 기능으로 시스템과 애플리케이션 간의 원활한 통합 및 상호운용성이 가능합니다.

기존 코드 문서화 또는 설명

특히 복잡한 프로젝트에서 기존 코드를 이해하는 건 버거울 수 있습니다. 생성형 AI는 코드를 문서화하거나 설명할 수 있도록 하며 특정 코드 세그먼트가 수행하는 작업에 대해 분명하고 간결한 설명을 할 수 있습니다. 이는 새 팀원 온보딩뿐만 아니라 코드베이스를 유지관리 및 업데이트하는 데 도움이 됩니다.

문제와 오류 디버깅 및 수정

생성형 AI는 잠재적인 코드의 문제와 오류를 식별하여 디버깅 및 수정을 할 수 있도록 제안합니다. 이러한 오류 감지 및 해결에 대한 접근 방식은 개발 시간을 단축하고 소프트웨어 품질을 높일 수 있습니다. 프롬프트에 '/fix'를 입력하기만 하면 코드가 수정되며 이에 대한 자세한 내용을 간단한 설명 및 링크를 통해 확인할 수 있습니다.

상황에 맞는 응답 받기

생성형 AI에서는 응답이 개발자를 위한 혁신적인 도구가 될 수 있습니다. 생성형 AI는 각자의 코딩 습관, 프로젝트 특성, 데이터 의미에 맞추어 모든 조언 및 지원이 비즈니스와 최신 데이터에 관련성이 있고 직접적으로 적용될 수 있도록 합니다. 이로써 개발 프로세스가 더 쉬워지고 직관성과 관련성이 높아집니다.

4장 주요 내용

- » 기존 AI로 애플리케이션 구축하기
- » 기존 AI 개발 문제점 관리하기
- » 모델 관리 시작하기
- » 생성형 AI로 애플리케이션 구축하기
- » 모두 통합하기

4장
데이터 인텔리전스 플랫폼 기반 AI 애플리케이션 구축

기존 인공지능과 생성형 인공지능(AI)이 거의 모든 비즈니스 기술의 측면에 통합되면서 기업은 고객에게 더 나은 서비스를 제공하고 경쟁 우위를 확보하기 위해 자체적인 맞춤형 AI 애플리케이션을 개발해야 합니다.

4장에서는 Databricks 데이터 인텔리전스 플랫폼이 기존 AI 및 생성형 AI 애플리케이션 개발과 머신러닝 운영(MLOps) 및 대규모 언어 모델 운영(LLMOps)을 통한 관리를 어떻게 지원하는지 살펴봅니다. 원활한 피처 엔지니어링, 모델 생성, 모델 실험 추적, ML 자동화, AI 애플리케이션 배포를 지원하는 플랫폼의 도구와 기능에 대해 자세히 설명합니다. Databricks 데이터 인텔리전스 플랫폼은 예측 모델 구축부터 최신 생성 AI와 LLM에 이르기까지 AI 및 ML 솔루션을 구축, 배포 및 모니터링할 수 있는 통합 도구를 제공합니다.

기존 AI 애플리케이션 개발

기억하세요

기존 AI는 명시적 프로그래밍 알고리즘을 기반으로 구축된 모델을 사용합니다. 이러한 AI 형태는 논리적 규칙과 의사결정 프로세스에 대한 사람의 지침에 의존합니다. 기존 AI 모델은 예측 및 분류와 같은 구체적이면서 구조화된 작업에 최적화되어 있습니다. 데이터 인텔리전스 플랫폼의 강점 중 하나는 모델 개발, 실험 추적, 모델 관리, 모델 배포, 기초 데이터의 변화에 따른 모든 AI 모델의 상태 모니터링까지 AI 모델의 전체 수명 주기를 지원하는 MLOps와 LLMOps입니다.

Delta Live Tables 및 Databricks Workflows 사용

Delta Live Tables 및 Databricks Workflows는 기존 AI 애플리케이션의 개발 및 배포를 향상합니다.

- **Delta Live Tables:** 이 기능으로 대규모의 안정적인 데이터 파이프라인을 구축하고 유지관리하는 작업을 간소화합니다. 추출, 변환, 로드(ETL) 프로세스의 여러 측면을 자동화하여 데이터 무결성을 확보하고 수동 관리의 필요성을 줄입니다. Delta Live Tables을 통해 작업을 정의하여 올바른 순서대로 조율하고, 데이터 품질 규칙을 설정하여 데이터가 들어올 때 생기는 문제를 처리하며, 강력한 오류 처리 기능을 제공하여 계획대로 진행되지 않으면 근본적인 원인을 신속하게 찾아내고, 이벤트 로그를 통해 전체 파이프라인을 모니터링할 수 있습니다. 쉽고, 확장 가능하며, 분산 환경에서 높은 데이터 품질을 보장합니다.
- **Databricks Workflows:** 이 기능으로 기초 인프라를 미리 구성하고 관리하지 않아도 작업을 실행할 수 있습니다. 워크로드에 맞게 리소스를 자동으로 최적화하고 확장하며 환경을 곧바로 시작하기 때문에 데이터 처리 및 분석 파이프를 더 수월하게 구현할 수 있습니다. 또한 비용을 절감하면서 높은 성능을 유지할 수 있습니다.

거버넌스, 보안 및 규정 준수 보장

AI 애플리케이션은 비즈니스 운영에서 갈수록 중요해지고 있으며 강력한 거버넌스, 보안 및 규정 준수 조치에 대한 필요성도 커지고 있습니다. Databricks는 Unity 카탈로그(UC)를 통해 데이터 프라이버시 보호 및 규정 준수를 보장하는 포괄적인 거버넌스와 보안 기능을 제공합니다.

기억하세요

이러한 기능은 민감하거나 독점적인 정보를 처리하는 기존 AI 및 생성형 AI 애플리케이션에 필수적입니다. 원시 데이터부터 AI 모델, 노트북, 애플리케이션까지 모든 것을 책임감 있게 사용되도록 하고 무단 액세스로부터 보호할 수 있습니다.

기존 AI 개발의 문제점 해결

기존 AI 애플리케이션 개발에 여러 문제점이 있습니다.

- **데이터 품질 및 가용성 저하:** 모든 AI 모델의 기반은 데이터입니다. 데이터 품질이 저하되고 데이터가 충분하지 않으면 AI 모델의 성능, 정확성, 신뢰도를 저해할 수 있습니다.
- **모델 복잡성:** 기존 AI 모델은 복잡해져서 이해, 신뢰, 관리, 확장이 어려워질 수 있습니다.
- **더 광범위한 생태계와의 통합:** AI 애플리케이션을 여러 내외부 비즈니스 시스템 및 워크플로우와 통합하면 더 광범위한 사용자 지정 및 구성이 가능해질 수 있습니다.

Databricks는 AI 개발 수명 주기를 간소화하도록 설계된 기능을 사용하여 이러한 문제를 해결합니다.

- **데이터 관리:** Databricks는 데이터 통합, 처리 및 품질 관리를 위한 포괄적인 도구를 제공하여 AI 모델이 고품질 데이터에 액세스할 수 있도록 보장합니다.
- AI **워크플로우 간소화:** Databricks의 통합 접근 방식은 복잡한 AI 모델 및 워크플로우에 대한 관리를 간소화합니다. 생성형 AI는 데이터 준비 및 초기 데이터 분석 등의 작업을 자동화하여 이러한 작업에 더 수월하게 액세스할 수 있도록 합니다.
- **완벽한 통합:** Databricks는 AI 애플리케이션을 기존 시스템과 쉽게 통합할 수 있는 다양한 애플리케이션 프로그래밍 인터페이스(API)와 커넥터를 제공하여 원활한 배포를 보장합니다.

모델 관리 및 MLOps/LLMOps 고려 사항

AI 모델은 갑자기 생겨나지 않았습니다. 모델을 구축, 배포 및 관리하기 위한 엔드투엔드 프로세스가 갖추어져 있으며, Databricks 데이터 인텔리전스 플랫폼이 모든 단계에서 도움이 됩니다. 이 섹션에서는 MLOps/LLMOps 세계를 자세히 살펴봅니다.

피처 엔지니어링 개선

모델은 Databricks 데이터 인텔리전스 플랫폼에서 직접 또는 제3자 환경에 연결하여 레이크하우스 환경의 데이터를 기반으로 구축됩니다. Databricks 데이터 인텔리전스 플랫폼은 데이터의 품질을 보장하여 양질의 모델이 구축될 수 있도록 합니다. 데이터를 변환하여 레이크하우스에 로드한 다음에 모델을 만들 수 있습니다.

이러한 모델은 기존 피처의 처리에 기반하여 새 피처(변수에 대한 다른 워드)을 만들어 개선 및 향상할 수 있습니다. 예를 들어, 금융 분야에서는 가격과 수익의 비율에 따라 새로운 수익당 가격 피처를 만들 수 있습니다. 결과로 나타난 비율은 원래의 가격 및 수익 피처보다 예측 가능성이 훨씬 더 높습니다.

모델 개발

데이터가 확보되면 모델을 개발할 수 있습니다. MLflow는 선도적인 오픈 소스 모델 개발 솔루션으로, Databricks와 함께 사용할 수 있습니다. 특정 데이터에 사용할 모델 유형을 선택할 수 있습니다. 예를 들어, 선형 회귀와 같은 일부 모델은 최적의 제품 가격을 예측하기에 적합한 반면에, 로지스틱 회귀와 같은 다른 모델은 정해진 이자율로 대출을 받아야 하는지 예측하기에 적합합니다.


팁

*하이퍼파라미터*는 각 모델을 실행하는 방법에 대한 지침입니다. 이는 돌리는 다이얼과 당기는 레버처럼 특정한 방식으로 작동하도록 모델에 지시하는 것입니다. 예를 들어, 고객을 8개의 세그먼트로 분류해야 할지, 아니면 20개의 세그먼트로 분류해야 할지 같은 것입니다.

모델 유형과 해당 하이퍼파라미터를 선택하면 MLflow가 모델을 실행하고 주어진 데이터에 대한 모델 예측의 정확도와 같은 정보를 제공합니다. 정확도는 80%일까요, 99%일까요?

실험 추적 문서화

모델을 개발한 것으로 끝나지 않습니다. 다양한 기능과 하이퍼파라미터를 사용하여 많은 실험을 하고 가장 정확한 접근 방식은 무엇인지 확인해야 합니다. MLflow는 모델이 생성된 시기의 정확도에 대한 다양한 메트릭을 제공할 수 있습니다.

기억하세요

모델이 마음에 들면 MLflow 추적 서버를 통해 UC에 등록합니다. 서버는 모델이 생성된 날짜, 버전 번호, 사용된 하이퍼파라미터, 그리고 정확도 메트릭과 같은 정보를 저장합니다.

여러 회차의 실험을 거듭하여 다른 모델에 비해 우수한 챌린저 모델이 무엇인지 확인한 다음에야 챔피언 모델을 대체할 수 있습니다. 충분한 권한이 있으면 이 비교 실험을 통해 가장 우수한 모델이 승리하도록 할 수 있습니다.

AutoML을 통한 간소화

자동화는 확장에 필수적이며, 자동화 머신러닝(AutoML)은 다양한 모델 유형과 하이퍼파라미터 조합을 시도하면서 수많은 모델 시나리오를 실행할 수 있습니다. 수백, 수천 가지의 조합을 자동으로 시도합니다. 이 과정을 거치면 AutoML이 주어진 데이터와 목표에 가장 정확한 모델을 결정할 수 있는 리더보드가 있습니다. Databricks의 AutoML은 데이터 과학자가 시간이 많이 소요되는 반복적인 작업에서 벗어나 이보다 복잡한 작업을 할 수 있도록 합니다. AutoML을 사용하면 비기술적 비즈니스 분석가도 전문 박사급 프로그래머만이 할 수 있는 영역이었던 모델을 만들 수 있습니다.

모델 설명가능성 및 투명성 명확화

일반적인 사람들은 모델의 결정의 근거에 대한 이해 없이 맹목적으로 모델의 권장사항을 구현해서는 안 됩니다. 비즈니스는 성과 창출에 가장 중요한 기능을 파악할 수 있는 가시성을 필요로 합니다. 예를 들어, 특정 잠재 고객에게 특정한 이자율로 은행 대출을 내주어야 하는 권장사항에 대하여 찬성 및 반대하는 이유는 무엇인가요?

이러한 설명가능성이 없으면 의사결정은 블랙박스와 같아서 모델에 대한 사람의 신뢰는 당연히 제한적일 수밖에 없습니다. 이로 인해 모델이 실제로 실용화되지 못하는 경우가 흔히 있습니다.


기억하세요

Databricks는 계보 추적을 통해 데이터부터 모델의 최종 결과에 이르기까지 완전한 투명성을 제공합니다. 이렇게 하면 데이터 여정의 모든 단계를 투명하게 파악할 수 있습니다.

모델 배포

모델을 확보하면 데이터 과학자 또는 ML 엔지니어가 Databricks를 사용하여 수월하게 프로덕션에 배포할 수 있습니다. 이는 엔드포인트를 제공하는 모델을 생성하는 것으로, 더 이상 많은 경험이 요구되지는 않습니다.

ML 워크플로우는 실험 및 개발에서 스테이징, 그리고 궁극적으로 실질적인 환경에서 사용할 수 있도록 모델을 프로덕션으로 이동할 수 있도록 합니다.

모델 거버넌스 준수

시간이 지나면서 비즈니스는 프로덕션의 첫 번째 모델에서 두 번째 모델, 50번째 모델로 성장하며, 어느새 수백, 수천 개의 모델이 프로덕션에 있게 됩니다. 핵심은 모델의 수명 주기를 관리할 수 있는 것이며, Databricks는 이 작업을 간단하게 합니다.

또한 UC를 통한 모델 수명 주기 관리도 중요합니다. 적합한 사용자만 모델을 프로덕션에 배치할 수 있도록 하며, 허용된 사용자만 모델이 구축되는 데이터에 액세스할 수 있도록 해야 합니다. Unity 카탈로그의 감사 로그와 시스템 테이블에는 누가, 언제, 어떤 데이터로 무슨 모델을 실행했는지 모든 세부 정보가 표시됩니다.

모델 및 데이터 드리프트 모니터링

금리가 변동하고 쇼핑 패턴이 바뀌며 고객의 지출에 변화가 생기는 등 시간에 따라 데이터가 바뀔 가능성이 높아집니다. 이를 *데이터 드리프트*라고 하며, 데이터가 어느 정도 바뀌면 모델의 정확도가 떨어지거나 모든 가치를 잃게 될 수도 있습니다.

Databricks SQL 대시보드를 사용하면 각 모델의 전체 상태, 모델의 실패 여부(예: 데이터 소스 작동 중지), 모델의 리프레시 필요 여부를 알 수 있습니다. 레이크하우스 모니터링을 통해 사용자 지정 메트릭을 설정하는 옵션을 사용하여 개발하는 도중에 메트릭을 설정할 수 있습니다.


팁

Databricks는 모델이 오래되면 데이터 과학자에게 모델을 재보정해야 하는 시기를 알려주는 SQL 알림을 보내거나 사람이 개입하지 않고 자동으로 모델이 리프레시되도록 설정할 수 있습니다.

생성형 AI 애플리케이션 개발

생성형 AI는 모든 산업의 비즈니스에서 새로운 애플리케이션을 개발하는 방식을 혁신하고 있습니다. 이 기술을 통해 기업은 혁신의 속도를 높이고 제품을 맞춤화하며 복잡한 문제를 해결할 수 있습니다. 기업은 생성형 AI를 도입하여 개발 시간을 단축하고 솔루션의 확장성을 넓힐 수 있습니다. 이는 더 나은 서비스와 제품을 제공할 수 있도록 합니다.

맞춤형 생성형 AI 애플리케이션 제작

Databricks의 Mosaic AI는 Databricks 생태계의 일부로, 처음부터 생성형 AI 애플리케이션을 구축할 수 있도록 합니다. 데이터 경계 외부로 기밀 정보를 유출하지 않고 원시 데이터로 시작해 비즈니스의 컨텍스트와 독점 데이터를 위해 특별히 설계된 AI 모델을 개발할 수 있습니다.

다음은 Databricks 데이터 인텔리전스 플랫폼에서 Mosaic AI의 주요 기능입니다.

» **LLM 훈련 맞춤화:** Mosaic AI를 통해 조직의 독점 데이터를 사용하여 LLM을 사용자 지정할 수 있습니다. 이렇게 하면 모델에 대한 지식이 특정 도메인과 밀접하게 연계되어 관련성이 높고 정확한 결과물을 제공할 수 있습니다.


기억하세요

LLM**은 빅 데이터를 사용하여 사람처럼 텍스트를 이해하고 생성합니다. 언어 작업을 수행하기 위해 방대한 양의 정보를 학습하며, 학습한 데이터에서 인식한 패턴을 기반으로 텍스트를 생성합니다.**

- **훈련 비용 절감:** 이 플랫폼은 맞춤형 LLM 훈련 비용을 크게 절감하는 최적화된 훈련 솔루션을 제안합니다. 이렇게 하여 대다수의 기업이 모델 품질에 영향을 미치지 않으면서 맞춤형 AI 솔루션에 투자할 수 있게 되었습니다.
- **포괄적인 모델 지원:** 모델이 훈련되면 Mosaic AI는 이러한 AI 모델을 배포, 관리 및 쿼리할 수 있는 통합 서비스를 제공합니다. 맞춤형 ML과 기초 모델이 여기에 포함되어 비즈니스 애플리케이션 및 워크플로우에 원활하게 통합됩니다.
- **데이터 보안 및 거버넌스 강화:** Mosaic AI는 모든 데이터와 지적 재산이 조직의 통제 범위 내에 있도록 하여 데이터 프라이버시 보호 및 규정 준수 위험을 줄입니다. 이는 의료기관이나 금융기관처럼 민감한 정보를 취급하는 기업에서 특히 중요하게 여겨집니다. 또한 조직은 데이터에서 프로덕션 모델까지 강력한 제어, 엔드투엔드 계보 및 감사를 확보할 수 있습니다.
- **완전한 제어:** 모델 및 데이터에 대한 소유권을 유지합니다. 조직은 Databricks로 고유한 엔터프라이즈 데이터를 사용해 생성형 AI 솔루션을 구축할 수 있습니다.
- **다양한 GenAI 아키텍처 패턴 지원:** Databricks는 프롬프트 엔지니어링, 검색 증강 생성(RAG), 미세 조정, 사전 훈련 맞춤형 LLM 등 다양한 생성형 AI 아키텍처를 지원합니다. 이러한 유연성으로 기업에서 특정 사용 사례에 가장 적합한 접근 방식을 선택하며, 변화하는 요건에 따라 개발할 수 있습니다.

RAG 애플리케이션 설계

RAG 애플리케이션은 LLM과 맞춤형 엔터프라이즈 데이터를 결합하여 AI 생성 응답의 정확성 및 관련성을 향상합니다. 쿼리와 관련된 데이터를 검색하여 LLM에 컨텍스트로 제공합니다.

팁

RAG는 최신 정보 상태를 유지관리하거나 도메인별 지식에 액세스해야 하는 챗봇과 Q&A 시스템을 성공적으로 지원하고 있습니다. RAG는 도메인별 애플리케이션에 알맞은 언어 모델을 조정하기 위해 전체 모델을 미세 조정하는 등 다른 방법과 비교하여 비용 효과적이면서 효율적인 솔루션을 제안합니다. RAG를 사용하면 조직에서 기본 LLM 모델을 수정하지 않고 외부 데이터를 활용할 수 있기 때문에 특히 데이터를 수시로 업데이트해야 하는 경우에 유용합니다. 또한 RAG는 모델의 답변이 오래된 훈련 데이터일 가능성이 있는 경우에는 이를 대신하여 최신 정보를 기반으로 이루어지도록 합니다.

기존 모델 미세 조정 (파인 튜닝)

이미 오픈 소스 LLM 모델을 보유하고 있는 경우에는 Databricks Mosaic AI를 사용하면 데이터를 통해 모델을 파인 튜닝 할 수 있습니다. 사전에 훈련된 생성형 AI 모델을 특정 데이터 세트 또는 도메인에 맞도록 조정합니다. 즉, 데이터 세트를 더 원활하게 반영되도록 조정하여 모델의 성능을 개선할 수 있습니다. 파인 튜닝을 하면 데이터에 대한 제어권을 유지하면서 데이터가 보안 환경 외부로 벗어나지 않으므로 프라이버시를 보호할 수 있습니다.

처음부터 모델 구축하기

Databricks Mosaic AI를 사용하면 처음부터 맞춤형 LLM 모델을 구축하여 데이터의 고유한 특성에 맞는 AI 솔루션을 만들 수 있습니다. 이 프로세스에는 자체적인 데이터 세트를 사용해 새 모델 전체를 훈련시키고, 결과물인 AI 애플리케이션이 비즈니스 프로세스와 통합되어 인사이트를 제공할 수 있는지 확인하는 작업이 포함됩니다.

LLM 훈련은 일반적으로 복잡하고 난이도가 높으며 광범위한 전문 지식을 갖추고 있어야 합니다. 그러나 Mosaic AI Foundation Model Training을 통해 데이터 소스를 지정하는 것만으로도 누구나 쉽고 효율적으로 나만의 맞춤형 LLM을 훈련할 수 있습니다. Foundation Model Training은 수백 개의 GPU로 확장, 모니터링, 자동 복구 등의 나머지 작업을 처리합니다. 수십억 개 매개변수 LLM 훈련을 단 며칠 만에 완료할 수 있습니다.

가능한 경우를 예로 들자면, Databricks는 기초 모델 훈련과 Mosaic AI의 강력함을 사용해 최첨단 LLM인 DBRX를 훈련시켰습니다. DBRX는 전문가 혼합(MOE) 아키텍처로 구축되었으며 발표 당시에 품질 및 가격 대비 성능 측면에서 선도적인 오픈 소스 LLM이었습니다. DBRX의 모든 기술과 최적화, 그리고 Foundation Model Training을 통해 모든 조직에서 합리적인 가격으로 자사 데이터에 완전히 맞춤화된 자체 LLM을 구축할 수 있게 되었습니다. 그 결과, 고유한 차별성을 가지는 조직의 IP에 대해 훈련된 맞춤형 모델을 얻습니다.

모든 것을 하나로 통합

Databricks 데이터 인텔리전스 플랫폼을 통해 엔터프라이즈 AI 애플리케이션 개발을 대폭 간소화할 수 있습니다. DatabricksIQ는 AI 플랫폼인 Mosaic AI와 직접적으로 통합되어 기업이 데이터를 이해하는 AI 애플리케이션을 수월하게 만들 수 있도록 합니다. 기업 데이터를 AI 시스템에 직접 통합하기 위해 Mosaic AI는 다음과 같이 다양한 기능을 제공합니다.

- **맞춤형 데이터에 고급 대화형 에이전트를 구축하는 엔드투엔드** RAG
- **대상 도메인에 대한 심층적인 이해를 기반으로** AI **애플리케이션을 더욱 향상하기 위해 처음부터 조직 데이터에 대해 맞춤형 모델을 훈련하거나** DBRX, MPT, Llama 3 **등 기존 모델을 지속적으로 사전 훈련**
- **엔터프라이즈 데이터에 대한 효율적이고 안전한 서버리스 추론과** UC**의 거버넌스 및 품질 모니터링 기능에 연결**
- **생성된 모든 모델과 데이터를 레이크하우스에서 자동으로 실행, 추적 및 모니터링할 수 있는 인기 있는** MLflow **오픈 소스 프로젝트 기반의 엔드투엔드** MLOps

주요 내용

- 통합 데이터 플랫폼의 이점
- 데이터 인사이트 발견
- 자체 데이터 및 지적 재산 소유권 확보
- 비용 절감

5장 데이터 인텔리전스 플랫폼이 필요한 이유

기업은 매일 다양한 소스에서 방대한 양의 데이터를 수집합니다. 그러나 방대한 양의 데이터에 액세스하는 것만으로는 충분하지 않기 때문에 데이터 자산의 잠재력을 최대한 활용하기 위한 강력한 도구가 있어야 합니다. 데이터 인텔리전스 플랫폼이 필요한 이유는 다음과 같습니다.

- **통합 데이터 플랫폼 제공:** 이 중심적인 장소는 모든 데이터 유형에 대한 하나의 위치로서의 역할을 합니다. 일관된 데이터 관리가 가능하며 데이터 사일로를 제거합니다.
- **데이터 및 AI 보안 강화:** 데이터 인텔리전스 플랫폼을 사용하면 데이터 및 인공지능(AI) 보안을 강화할 수 있습니다. 이 플랫폼은 강력한 보안 기능을 갖추었으며 조직이 민감한 데이터를 보호하고 규정 준수 요건을 충족할 수 있도록 합니다.

  기업에서 AI 및 분석을 사용하는 경우에는 정보 유출이 발생하지 않도록 해야 합니다.

기억하세요

- **데이터 및 지적 재산(IP)의 완전한 소유:** 플랫폼을 통합하면 자체 데이터를 기반으로 애플리케이션, 솔루션 또는 분석을 만들거나 강화할 수 있습니다. 이렇게 하면 가장 강력한 경쟁력과 경제적 이점을 확보할 수 있습니다.

- **검색 개선:** 데이터 인텔리전스 플랫폼으로 데이터 자산을 더 쉽게 검색하고 데이터의 의미에 대하여 더 나은 컨텍스트를 찾을 수 있습니다. 이 시나리오에서는 모든 데이터 인사이트가 더욱 포괄적으로 제공되며 조직 내 다양한 사용자가 액세스할 수 있습니다.

- 더욱 지능적인 데이터 기반 인사이트 발견 조직은 데이터 인텔리전스 플랫폼에서 데이터에 숨겨진 인사이트와 트렌드를 발견할 수 있습니다. 데이터 정리, 검증, 강화 기능을 제공함으로써 데이터 품질을 보장하므로 AI가 더 나은 컨텍스트 인식 정보를 제공한다는 점이 중요합니다.

- 자동화된 통합 워크플로우로 데이터 작업 가속화 데이터 엔지니어링, 데이터 과학, 머신러닝 등 작업의 속도를 높여주는 단일 플랫폼 안에서 작업하므로 원활한 협업과 효율적인 워크플로우를 지원합니다. 데이터 인텔리전스 플랫폼은 데이터 파이프라인과 인프라 관리를 자동화하여 수작업을 줄이고 오류를 최소화하며 확장성을 개선합니다.

- 조직의 전 직원이 데이터에 더 쉽게 액세스 가능 데이터 인텔리전스는 소프트웨어 코드 작성 방법을 모르는 비전문가도 자연어로 자체 데이터를 쿼리할 수 있습니다. 이로써 비즈니스 분석가, 경영진, 비즈니스 부문 관리자 등 누구나 그 어느 때보다 더 쉽고 빠르게 훨씬 지능적인 인사이트를 얻을 수 있게 되었습니다.

- 더 나은 협업 지원 이 플랫폼은 팀 및 사용자 간의 원활한 파트너십을 통해 인사이트, 코드 및 결과물을 공유할 수 있도록 합니다. 이러한 팀워크는 데이터 기반의 의사결정을 가속화하기 때문에 비즈니스에 큰 이점입니다.

- **대규모 확장:** 이 플랫폼으로 대규모 데이터 프로세스를 처리하여 조직이 방대한 양의 데이터를 더욱 효율적으로 처리할 수 있도록 합니다. 하나의 플랫폼에서 정형, 반정형, 비정형 데이터를 모두 사용할 수 있습니다.

- **ROI 개선:** 비용 절감은 매우 중요합니다. 데이터 관리 및 분석 도구를 단일 플랫폼에 통합하여 조직에서 비용을 절감하고 데이터 인프라를 간소화할 수 있습니다.

데이터+AI에 대한 기업의 잠재력 극대화

데이터 인텔리전스를 통해 데이터와 AI에 대한 조직의 잠재력을 극대화하세요. Databricks 데이터 인텔리전스 플랫폼은 개방형, 통합형, 관리형 레이크하우스 아키텍처를 기반으로 하며 데이터 인텔리전스 엔진으로 구동됩니다. AI를 사용하면 플랫폼에서 자체 엔터프라이즈 데이터를 추론하여 고유한 데이터 자산의 가치를 최대한 활용할 수 있도록 지원합니다. ETL, 데이터 웨어하우징, BI, 기존 AI 또는 생성형 AI 등 데이터 인텔리전스는 데이터 기반의 성공을 향한 여정을 간소화하고 가속화합니다.

내용

- 데이터 인텔리전스의 가치
- AI의 능력과 잠재력
- 데이터 인텔리전스 플랫폼의 기능
- AI 애플리케이션 구축
- 데이터 인텔리전스 플랫폼이 필요한 이유

databricks

Ari Kaplan is Databricks' Head of Technical Evangelism. He is Caltech's "Alumni of the Decade" and created the Chicago Cubs' & Baltimore Orioles' analytics departments. **Stephanie Diamond,** former AOL marketing director, is founder of Digital Media Works. She's authored dozens of marketing and custom e-books.

ISBN: 978-1-394-32296-1
재판매 금지

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.