

# Databricks

## Databricks 모멘텀

제목: Databricks 모멘텀: 규모와 영향력 핵심 내용 (3~4개 주요 블렛):

- 성장: 상반기 예상 매출 40억 달러 초과 (전년 대비 +50% 성장).
- 에코시스템: 전 세계 8,000명 이상의 직원, 30개 사무소, 5,800개 이상의 파트너사.
- 기술 리더십: Delta Lake, Apache Spark, Unity Catalog, MLflow 등 선도적인 OSS 데이터 프로젝트 제작.
- 전략적 파트너십: Gemini/Google Cloud, Anthropic 및 SAP 등과의 새로운 협력 강화.

## Databricks의 사명

제목: Databricks의 사명: 데이터 보편화 핵심 메시지:

- 슬로건: 데이터와 AI의 통합을 통해 모든 데이터에 대한 접근성을 높이고, AI를 보편화합니다.
- 목표: 데이터와 AI를 별도로 관리하는 복잡성을 해소하고, 하나의 플랫폼에서 운영합니다.

## AI 도입의 도전 과제

제목: AI 도입의 장벽: "일반적인" 접근 방식의 문제 핵심 내용:

- 문제 1: 평가의 어려움: 모델 제공자에게 종속되어 도메인별 평가가 어려움.
- 문제 2: 통제력 및 거버넌스 부재: 데이터와 모델 간의 연관성 및 거버넌스 확보 난항.
- 해결책: 모델을 데이터로 가져옵니다. (데이터 중심 접근)

## 데이터 관리의 복잡성

제목: 데이터 관리의 악몽: 사일로화된 인프라 문제점:

- 데이터 웨어하우스, 데이터 레이크, ETL, 머신러닝, 생성형 AI 등 기능별로 분리된 수많은 시스템.
- 결과: 높은 비용과 독점적 종속으로 인한 복잡성. 해결책 제안: 오픈 포맷과 통합된 거버넌스로 모든 것을 통합합니다.

## 데이터 인텔리전스 플랫폼의 기반

제목: 모든 데이터와 AI를 위한 거버넌스: Unity Catalog 핵심 내용:

- 단일 거버넌스 계층: 모든 자산(테이블, AI 모델, 파일, 노트북, 대시보드)을 통합 관리.
- 주요 기능: 액세스 제어, 리니지(계보), 감사, 비용 통제, 안전한 데이터 공유.
- 기반 기술: 오픈 포맷 (DELTA LAKE, ICEBERG).

## Databricks 데이터 인텔리전스 플랫폼

제목: Databricks 데이터 인텔리전스 플랫폼 아키텍처 주요 구성 요소:

- 거버넌스: Unity Catalog (DELTA LAKE, ICEBERG 기반)
- AI 에이전트: Agent Bricks

- 트랜잭션 DB: Lakebase
- 데이터 웨어하우징: DB SQL
- 통합 데이터 엔지니어링: LakeFlow
- 분석: AI/BI (비즈니스 인텔리전스)

## AI 에이전트의 시대: Agent Bricks

제목: AI 에이전트, 실제 운영 환경에 적용하기 도입 배경: AI 에이전트는 어디에나 과장되어 있지만, 88%의 AI 프로젝트가 프로덕션에 도달하지 못합니다. **Agent Bricks** 역할: 정확한 도메인별 결과를 제공하는 에이전트 시스템 구축을 간소화합니다. 핵심 단계: 데이터 준비 -> 에이전트 구축 -> 에이전트 배포 -> 에이전트 평가 -> 에이전트 관리

## Lakebase: 트랜잭션 데이터베이스 엔진

제목: 다음 시대를 위해 설계된 트랜잭션 처리: Lakebase 문제점: 기존 데이터베이스는 대규모 종속, 비싼 비용, 온프레미스용으로 설계됨. **Lakebase** 특징:

- 데이터 저장: 저비용 레이크에 저장된 데이터 (컴퓨팅-스토리지 분리).
- 성능: 매우 짧은 레이턴시, 매우 높은 QPS.
- 기술: 오픈 소스 기반(Postgres 기반), AI를 위해 구축됨.
- 새로운 사용 사례: 기본 제공 ML 및 감사를 통한 실시간 가격 책정, 통합 사례 관리.

## LakeFlow: 통합 데이터 엔지니어링

제목: 데이터 처리를 위한 두 가지 영역 통합: LakeFlow 기존 문제: 데이터 엔지니어(복잡한 데이터 파이프라인)와 비즈니스 분석가(스프레드시트)의 사일로화된 워크플로. **LakeFlow**의 가치:

- **LakeFlow Designer**: 운영 환경 품질 ETL, 코딩 필요 없음.
- 협업: 분석가와 엔지니어가 함께 구축하고 재작성 필요 없음.
- **AI**: 메타데이터, 리니지, 컨텍스트에 기반한 AI로 생산성 향상.

## AI/BI: 비즈니스 인텔리전스의 확장

제목: 데이터 인텔리전스를 사용하여 BI를 모든 팀에 확장 현황: AI/BI 사용자 수가 지난 1년간 +500% 급증 (2024년 6월 → 2025년 6월). **Genie**의 역할: 대시보드에서 예상하지 못한 질문에 대해 데이터와 직접 대화할 수 있는 공개 Q&A 기능 제공. 혜택: 추가 비용 없이 BI를 모두에게 확장, 매우 빠른 결과 도출, Unity Catalog로 보안 유지.

## DB SQL: 뛰어난 데이터 웨어하우스

제목: 뛰어난 BI를 위한 뛰어난 데이터 웨어하우스: DB SQL 도입 현황: 전 세계적으로 대규모 도입, 사용량이 지난 1년간 2배 증가. 성능 향상: DB SQL 쿼리 속도 5배 더 빠름 (2022년 대비 2025년). 비용 효율성: 업계 최고의 TCO (모든 주요 CDW 대비 가격/성능 모든 영역에서 전체 선두).

## Lakebridge: 웨어하우스 마이그레이션 가속화

제목: 마이그레이션의 혐난한 여정, Lakebridge로 극복 문제점: 마이그레이션의 절반 이상이 예산과 일정 초과 (높은 복잡성, 노동 집약적). **Lakebridge** 구성:

- 분석: 심층 스캐너 (범위 영역 평가)
- 코드 변환기: 고급 LLM 변환기 (DB SQL 지원 코드 생성)

- 데이터 마이그레이션: Lakeflow Connect
- 데이터 검증: 조정기 (마이그레이션 완료) 지원: Teradata, Oracle, SQL Server 등 20개 이상의 레거시 웨어하우스.

## Lakebridge의 가치

제목: Lakebridge의 가치: 마이그레이션 비용과 시간 절감 주요 혜택:

- 코드 정확성 향상.
- 마이그레이션 가속화.
- 비용 절감.
- 플랫폼 현대화 가속.
- 강점: 무료 제공, 100회+ 마이그레이션으로 입증됨.

## 최종 요약 및 마무리

제목: 데이터 인텔리전스 플랫폼으로 미래를 준비하세요. 핵심 메시지 (**Databricks** 플랫폼의 5가지 핵심):

- Unity Catalog:** 통합 거버넌스
- Agent Bricks:** AI 에이전트
- Lakebase:** 트랜잭션 데이터베이스
- LakeFlow:** 데이터 엔지니어링 및 ETL
- AI/BI (DB SQL):** 데이터 웨어하우징 및 비즈니스 인텔리전스
- Lakebridge :** 마이그레이션
- 마무리: 감사합니다.

# Databricks Free Training

<https://www.databricks.com/resources/learn/training/databricks-fundamentals>

Data intelligence platforms have created a paradigm shift by unifying data, analytics and AI on a single, open platform while applying AI to your data to create custom AI applications and democratize productivity across your entire organization.

Data Intelligence Platform은 데이터와 분석, 그리고 AI를 하나의 개방형 플랫폼에 통합하는 동시에 데이터에 AI를 적용하여 맞춤형 AI 애플리케이션을 구축하고 조직 전체의 생산성을 높여 패러다임의 변화를 일으켰습니다.

## 오늘의 주제

1. 데이터 관리 플랫폼의 역사
2. Databricks 데이터 인텔리전스 플랫폼이란?
3. Databricks DI 플랫폼 아키텍처 및 보안 기본 사항
  - a. 플랫폼 아키텍처 개요
  - b. 데이터 거버넌스
  - c. 보안, 안정성과 성능
  - d. 데이터 인텔리전스 엔진(출시 예정!)
4. Databricks DI 플랫폼에서 지원되는 워크로드
  - a. 데이터 웨어하우징
  - b. 오키스트레이션
  - c. ETL 및 실시간 분석
  - d. 데이터 사이언스 & AI
5. 요약 및 다음 단계

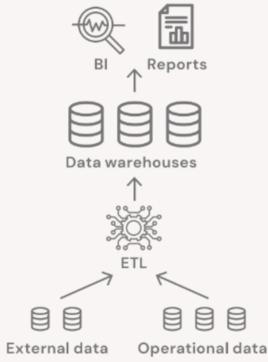
## 데이터 관리 플랫폼의 역사

### 데이터 관리 플랫폼의 기원과 목적

기업들은 비즈니스 의사결정과 혁신을 위해 데이터 기반 인사이트 활용 필요  
이를 위해 단순한 관계형 데이터베이스를 넘어 빠른 속도로 수집 생성되는 대량의 데이터를 관리하고  
분석하는 시스템이 필요

### 데이터웨어하우스

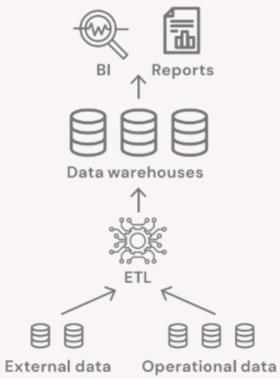
## Data warehouse



장점:

- 비즈니스 인텔리전스(BI)
- 분석
- 정형 및 정제된 데이터
- 사전 정의된 스키마

## Data warehouse



단점:

- 반정형 또는 비정형 데이터를 지원하지 않음
- 융통성 없는 스키마
- 볼륨과 속도 증가에 따른 어려움
- 긴 처리 시간

## 데이터 레이크

2000년대 빅데이터의 폭발적 증가

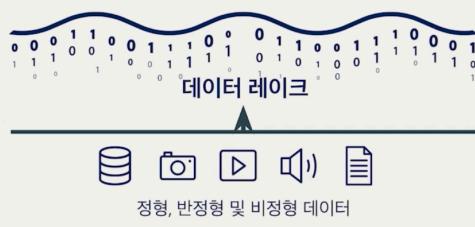
데이터 레이크는 정형, 반정형, 비정형 데이터로 데이터 생성 속도 지원

여러 데이터 유형을 데이터 레이크에 저장 : 웹로그나 센서 데이터 등 다양한 소스에서 생성된 데이터를 데이터 레이크에 빠르게 저장

## 데이터 레이크

장점:

- 유연한 데이터 스토리지
- 스트리밍 지원
- 클라우드에서 비용 효율적
- AI 및 기계 학습 지원



## 데이터 레이크

단점:

- 트랜잭션 지원 없음
- 데이터 신뢰성 저하
- 느린 분석 성능
- 데이터 거버넌스 문제
- 데이터 웨어하우스가 여전히 필요



빅데이터 관리 및 사용의 과제

데이터 관리 플랫폼이 어떻게 데이터 인텔리전스 플랫폼으로 발전했는지 설명

## 레이크하우스

데이터 웨어하우스와 데이터 레이크의 장점을 결합



데이터 레이크하우스 아키텍쳐

비즈니스팀은 서로 다른 플랫폼, 레이크하우스, 데이터 웨어하우스로 워크로드로 분할하면서 단절된 사일로에서 작업

따라서 속도와 효율성을 제공하는 단일 통합 기술 스택과 단순성에 대한 필요성에 충족 필요

데이터브릭스는 개방형 아키텍처로 개발된 데이터 레이크하우스의 데이터 관리 플랫폼 아키텍처를 지지  
데이터 레이크의 장점과 데이터 웨어하우스의 분석기능 및 제어기능을 결합

데이터 레이크하우스 플랫폼은

- 레이크하우스 위에 구축되어 모든 데이터 유형을 함께 저장. 신뢰할 수 있는 단일 데이터소스가 됨
- 통합된 보안 거버넌스 및 카탈로그 구성요소와 함께 작동

비즈니스에 필요한 모든 데이터를 위한 개방적인 통합 기반이 됨

## 데이터 레이크하우스의 주요 기능:

- 트랜잭션 지원
  - 스키마 적용과 관리
  - 데이터 거버넌스
  - BI 지원
  - 컴퓨팅에서 스토리지 분리
- 개방형 저장 포맷
  - 다양한 데이터 유형 지원
  - 다양한 워크로드 지원
  - 엔드투엔드 스트리밍

기업은 정형, 반정형, 비정형 데이터를 한곳에서 저장, 정제, 분석 및 액세스 할 수 있으며 데이터사이언스, 머신러닝 및 분석을 위한 광범위한 워크로드를 지원

레이크하우스는 실시간 보고를 위한 엔드투엔드 스트리밍 지원 → 실시간 데이터 애플리케이션을 위한 별도의 시스템 필요 없음

생성형 AI 의 부상

### 데이터 레이크하우스

모든 데이터를 위한 개방적이고 통합된 기반



**생성형 AI**

데이터와 AI를 쉽게 확장하고 사용

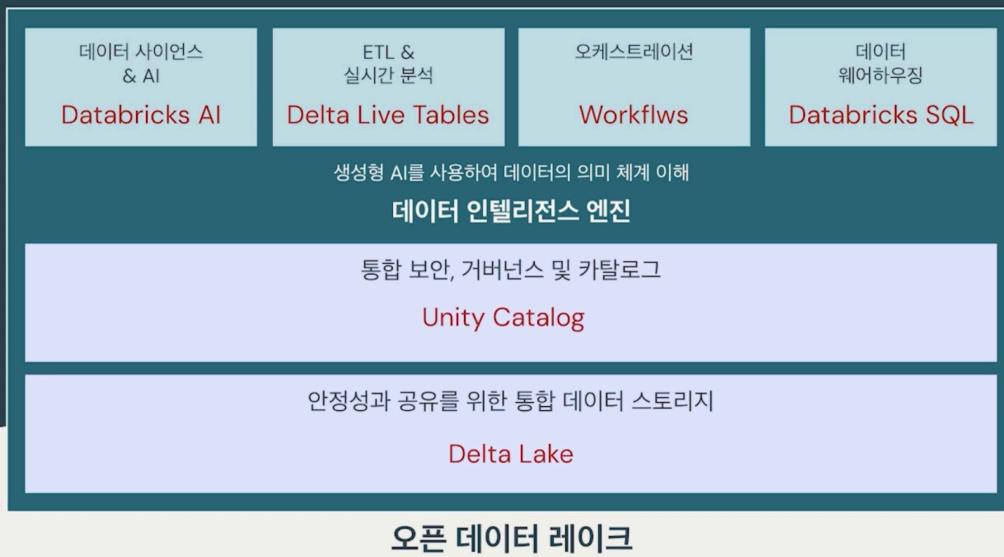


### 데이터 인텔리전스 플랫폼

조직 전체에서  
데이터 + AI의 민주화

데이터브릭스는 데이터 레이크하우스와 생성형 AI 의 두가지 기술을 결합 → 데이터와 AI 대중화 / 민주화

# Databricks 데이터 인텔리전스 플랫폼



# 1. 강의 주제 개요

이 세션은 데이터 관리 플랫폼의 발전 과정을 설명하고, Databricks가 제시하는 새로운 통합 모델인 **Data Intelligence Platform**을 소개하는 내용으로 구성되어 있습니다.

주요 흐름

1. 데이터 관리 플랫폼의 역사
  2. Databricks 데이터 인텔리전스 플랫폼 개요
  3. Databricks 플랫폼 아키텍처 및 보안 개념
  4. Databricks에서 지원되는 주요 워크로드
    - 데이터 웨어하우징
    - 오케스트레이션
    - 실시간 분석
    - 데이터 사이언스 및 AI
  5. 요약 및 다음 단계
- 

## 2. 데이터 관리 플랫폼의 역사

### 2.1 데이터 관리의 필요성

기업은 비즈니스 의사결정과 혁신을 위해 데이터 기반 인사이트를 확보해야 하며, 이를 위해 방대한 양의 데이터를 빠르고 효율적으로 수집·저장·분석할 수 있는 시스템이 필요해졌습니다.

---

### 2.2 데이터 웨어하우스 (**Data Warehouse**)

장점

- 비즈니스 인텔리전스(BI)와 분석에 적합
- 정제되고 구조화된 데이터
- 사전에 정의된 스키마 기반
- 높은 정확성과 일관성

단점

- 반정형·비정형 데이터 지원 부족

- 확장성과 유연성 한계
  - 처리 속도 저하
  - ETL 과정 복잡
- 

## 2.3 데이터 레이크 (**Data Lake**)

등장 배경

2000년대 빅데이터 확산으로 다양한 소스(웹로그, IoT 센서 등)에서 데이터가 폭발적으로 증가하며, 정형·반정형·비정형 데이터를 모두 저장할 수 있는 환경이 필요하게 됨.

장점

- 다양한 데이터 형식 저장 가능
- 스트리밍 및 실시간 데이터 지원
- 클라우드 기반의 비용 효율성
- AI 및 머신러닝 학습 데이터로 활용 가능

단점

- 트랜잭션 미지원
  - 데이터 신뢰성 낮음 (Data Swamp 문제)
  - 데이터 거버넌스 부재
  - 여전히 별도의 웨어하우스 필요
- 

## 3. 데이터 레이크하우스 (**Data Lakehouse**)

데이터 레이크하우스는 데이터 웨어하우스와 데이터 레이크의 장점을 결합한 통합형 구조입니다.

특징

- 정형·비정형 데이터를 함께 저장 및 분석
- BI, 실시간 분석, AI 워크로드를 단일 플랫폼에서 수행
- 통합 보안, 거버넌스, 카탈로그 제공
- 클라우드 네이티브 및 오픈 포맷 기반

주요 기능

- 트랜잭션(ACID) 지원
  - 스키마 자동 적용 및 관리
  - 데이터 거버넌스(통합 카탈로그)
  - BI 및 실시간 분석 지원
  - 다양한 데이터 형식 및 워크로드 통합
  - 엔드투엔드 스트리밍 처리
- 

## 4. 데이터 인텔리전스 플랫폼 (Databricks Data Intelligence Platform)

Databricks는 Lakehouse 개념을 확장하여 데이터 + 생성형 AI(Generative AI)를 결합한 **Data Intelligence Platform**으로 진화시켰습니다.

핵심 개념

Data Lakehouse + Generative AI = Data Intelligence Platform

특징

- 데이터, 분석, AI를 하나의 통합 플랫폼에서 활용 가능
- 모든 데이터 유형을 오픈 포맷으로 저장 및 관리
- Unity Catalog를 통한 중앙 집중형 거버넌스
- 조직 전체의 데이터와 AI 활용을 민주화

## 5. 발전 단계 요약

단계	플랫폼	주요 특징	한계
1단계	Data Warehouse	정형 데이터 기반 BI 중심	비정형 데이터 처리 불가
2단계	Data Lake	다양한 데이터 저장, AI 학습 활용	거버넌스 및 신뢰성 부족
3단계	Data Lakehouse	통합 거버넌스 및 확장성 확보	관리 복잡성 일부 존재
4단계	Databricks Data Intelligence Platform	데이터 + AI 완전 통합, 생성형 AI 활용	

## 6. 요약

Databricks는 데이터 웨어하우스의 안정성과 데이터 레이크의 유연성을 결합한 **Lakehouse Architecture** 위에

AI 기능을 통합하여, 조직 전체가 데이터와 인공지능을 하나의 환경에서 활용할 수 있도록 지원하는 데이터 인텔리전스 플랫폼입니다.

즉, Databricks는 단순한 데이터 관리 플랫폼이 아니라,

데이터를 인텔리전스로 전환하고 조직의 의사결정 및 혁신을 가속화하는 통합형 데이터·AI 플랫폼으로 정의할 수 있습니다.

## Databricks Data Intelligence Platform 이란?

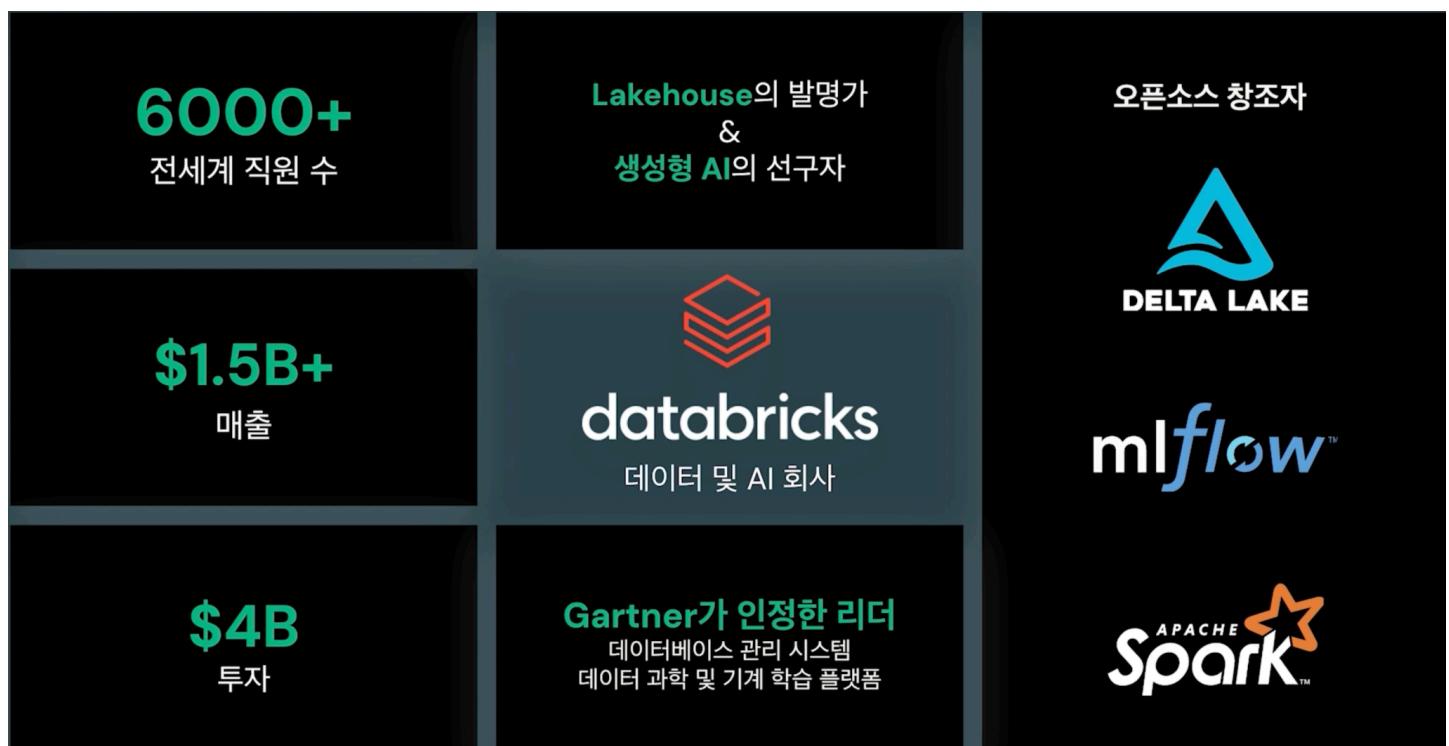
### 학습 목표

Databricks의 비즈니스 설명

Databricks Data Intelligence 플랫폼 설명

Databricks DI 플랫폼이 비즈니스 과제를 어떻게 해결하는지 사례 설명

Databricks DI 플랫폼이 데이터 실무자와 비기술 데이터 사용자에게의 이점 설명



데이터와 데이터분석 뿐만 아니라 AI 까지 포함되도록 접근 방식 확장

모든 산업에서의 승자는 데이터와 AI 기업이 될 것이라 믿음 → 그러나 데이터와 AI 활용이 현실에서는 쉽지 않음



여러가지 기술 스택을 활용하여 하나의 데이터 생태계를 만드는 것은 도전적인 과제임  
 데이터웨어하우스로 시작, 시각화를 위한 BI 플랫폼, 실시간 요구 사항을 처리하기 위한 스트리밍 플랫폼, 머신러닝, 데이터 사이언스, 다양한 데이터 유형 (텍스트, 이미지, 고객 데이터 등)을 위한 데이터 레이크, 거버넌스, 생성형 AI

저장소로 시작했지만, 거대한 데이터 에코시스템으로 확대

조직의 속도를 늦추는 세가지 문제



# 데이터 레이크하우스

## 데이터 레이크하우스

모든 데이터를 위한 개방적이고 통합된 기반

### 오픈 데이터 레이크

모든 원시 데이터  
(로그, 텍스트, 오디오, 비디오, 이미지)

데이터브릭스에서 기업이 데이터 및 AI 요구사항을 보다 쉽게 충족할 수 있도록 통합하는 개방형 접근 방식인 데이터레이크 패러다임 발표

### 통합 거버넌스 / 카탈로그

통합된 거버넌스 및 카탈로그 기능으로 데이터레이크에 저장된 데이터가 안전하고, 관리 가능하며, 카탈로그화 되는 것이 보장

## 데이터 레이크하우스

모든 데이터를 위한 개방적이고 통합된 기반

통합 보안, 거버넌스 및 카탈로그

안정성과 공유를 위한 통합 데이터 스토리지

### 오픈 데이터 레이크

모든 원시 데이터

데이터 사이언스 및 AI, ETL, 실시간 분석, 오케스트레이션, 데이터 웨어하우징 등 다양한 워크로드 지원

# 데이터 레이크하우스

모든 데이터를 위한 개방적이고 통합된 기반

데이터 사이언스  
& AI

ETL & 실시간  
분석

오케스트레이션

데이터  
웨어하우징

통합 보안, 거버넌스 및 카탈로그

안정성과 공유를 위한 통합 데이터 스토리지

## 오픈 데이터 레이크

모든 원시 데이터

방향 : 데이터를 다양한 데이터플랫폼에 사일로화 하지 않고, 오픈 소스로 제공되는 클라우드 데이터레이크에 저장하는 것 → 데이터를 제어하고 필요한 사람들이 사용 가능

# 데이터 레이크하우스

모든 데이터를 위한 개방적이고 통합된 기반

데이터 사이언스  
& AI  
**Databricks AI**

ETL &  
실시간 분석  
**Delta Live Tables**

오케스트레이션  
**Workflows**

데이터  
웨어하우징  
**Databricks SQL**

통합 보안, 거버넌스 및 카탈로그

**Unity Catalog**

안정성과 공유를 위한 통합 데이터 스토리지

**Delta Lake**

## 2020

Databricks는  
레이크하우스  
아키텍처를 개척

## 오늘날

글로벌 기업의 **74%**가  
레이크하우스를 채택

MIT Technology Review  
인사이트, 2023

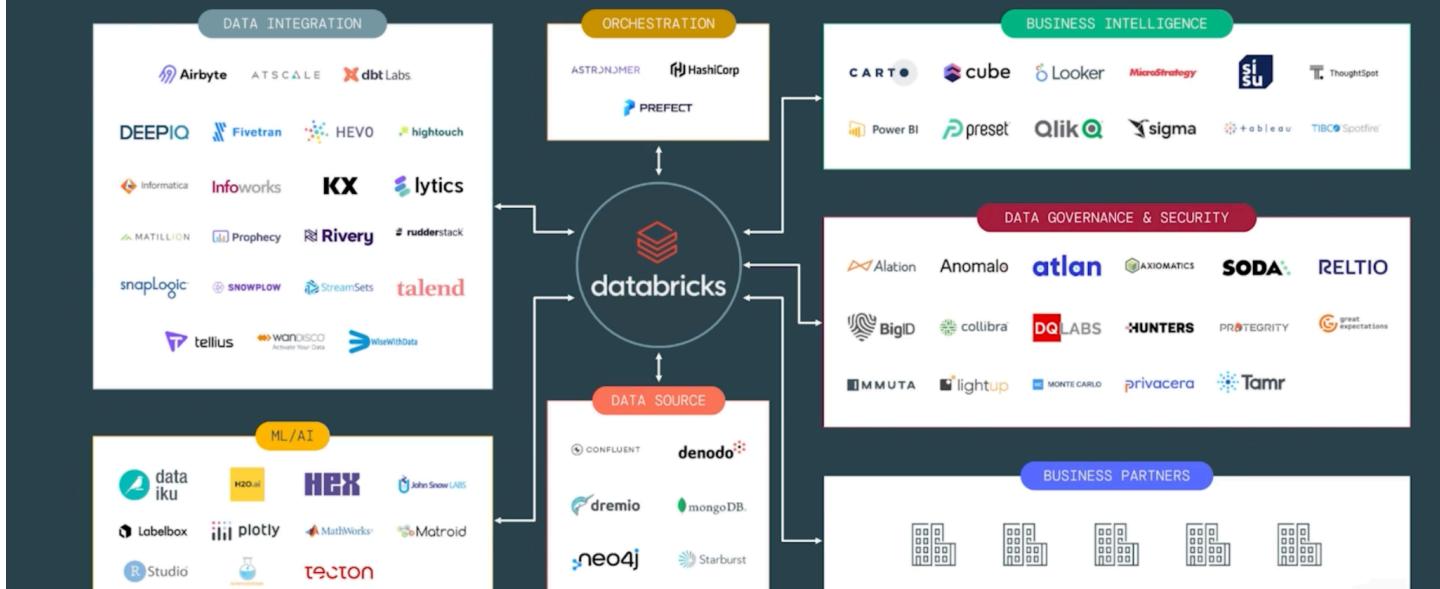
## 오픈 데이터 레이크

모든 원시 데이터

개방형 기반위에 구축되어 전체 데이터와 AI 에코 시스템(ML, AI, BI, 데이터거버넌스) 과의 통합이 쉬움

# 개방형 기반 위에 구축

전체 데이터 및 AI 에코시스템과 쉽게 통합

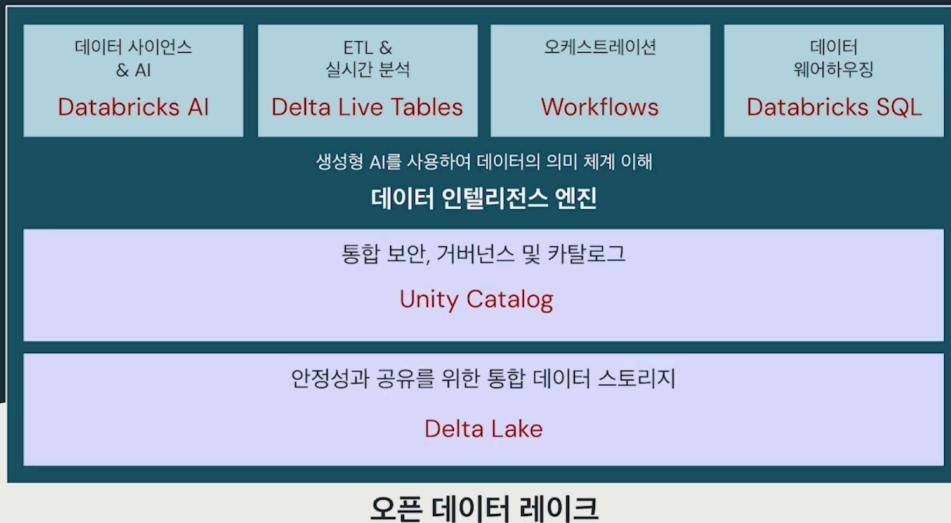


## 데이터 인텔리전스 플랫폼



레이크하우스와 생성형 AI 인텔리전스를 결함  
데이터 인텔리전스 플랫폼은 레이크하우스 토대위에 구축

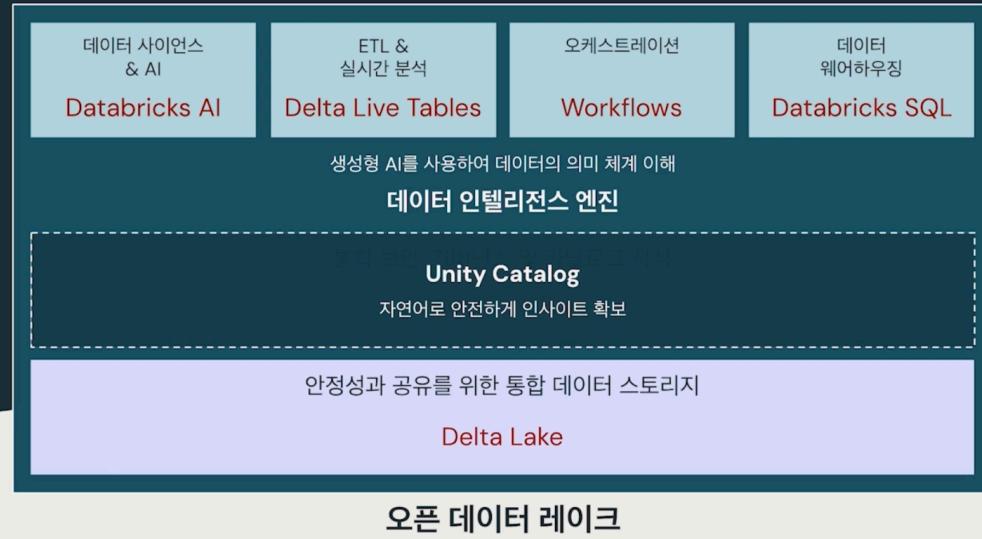
# Databricks 데이터 인텔리전스 플랫폼



모든 영역에서 생성형 AI 활용

생성형 AI가 모든 데이터를 이해하고, 모든 플랫폼에 AI 기능 주입

# Databricks 데이터 인텔리전스 플랫폼



# Databricks 데이터 인텔리전스 플랫폼



# Databricks 데이터 인텔리전스 플랫폼



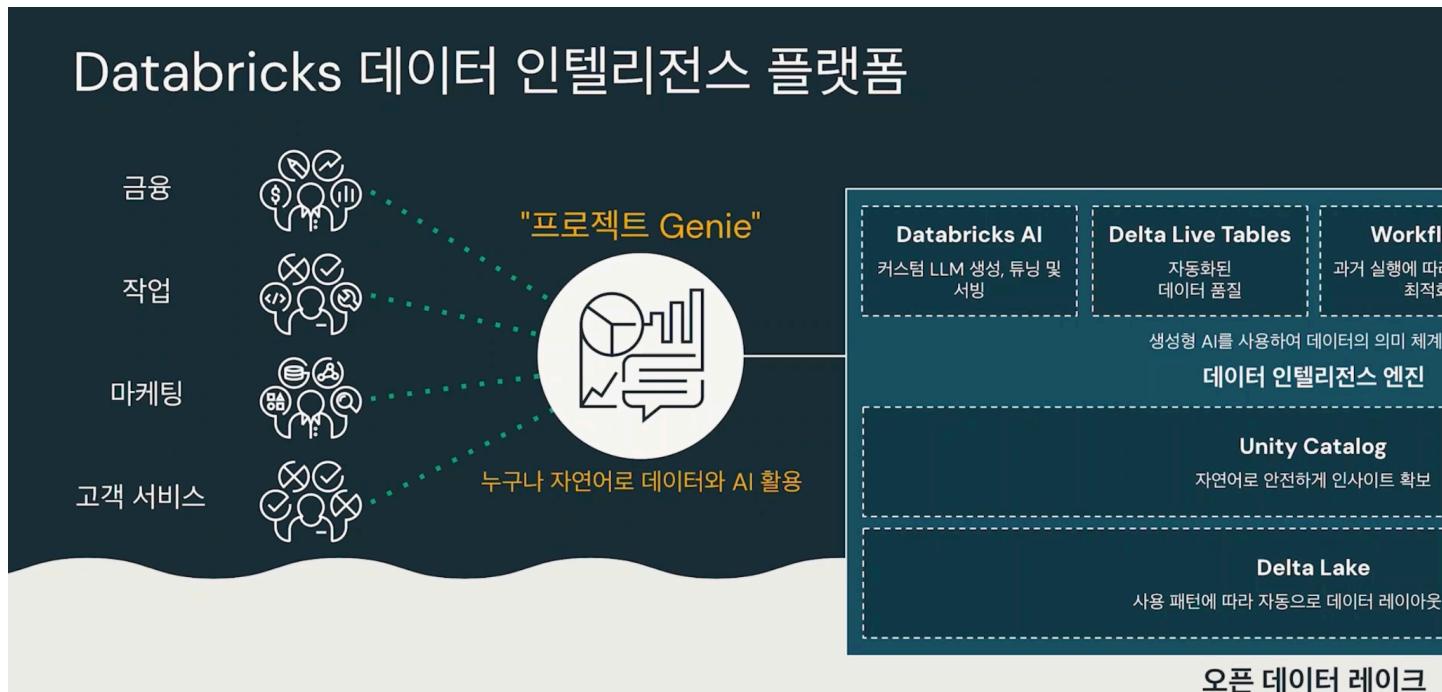
# Databricks AI

## Databricks 데이터 인텔리전스 플랫폼



특화된 AI 모델 구축  
모델 서빙  
RAG

데이터 실무자와 비기술 데이터 사용자에게의 이점



지니를 사용하면 자연어만으로 데이터와 AI 활용  
비기술 직군도 기술 지원 없이 자연어 기반으로 답변과 데이터 시각화 가능

## 데이터 인텔리전스 플랫폼은 진정한 데이터와 AI의 민주화를 가능케 합니다

### ➤ **Simple**

자연어로 모두에게 사용 편의성과 효율성 제공

### ➤ **Intelligent**

고유한 데이터를 이해하기 위해 엔드 투 엔드로 통합된 AI

### ➤ **Private**

내부 데이터를 기반으로 커스텀 모델을 쉽게 구축

# 데이터브릭스 보안 기본 사항

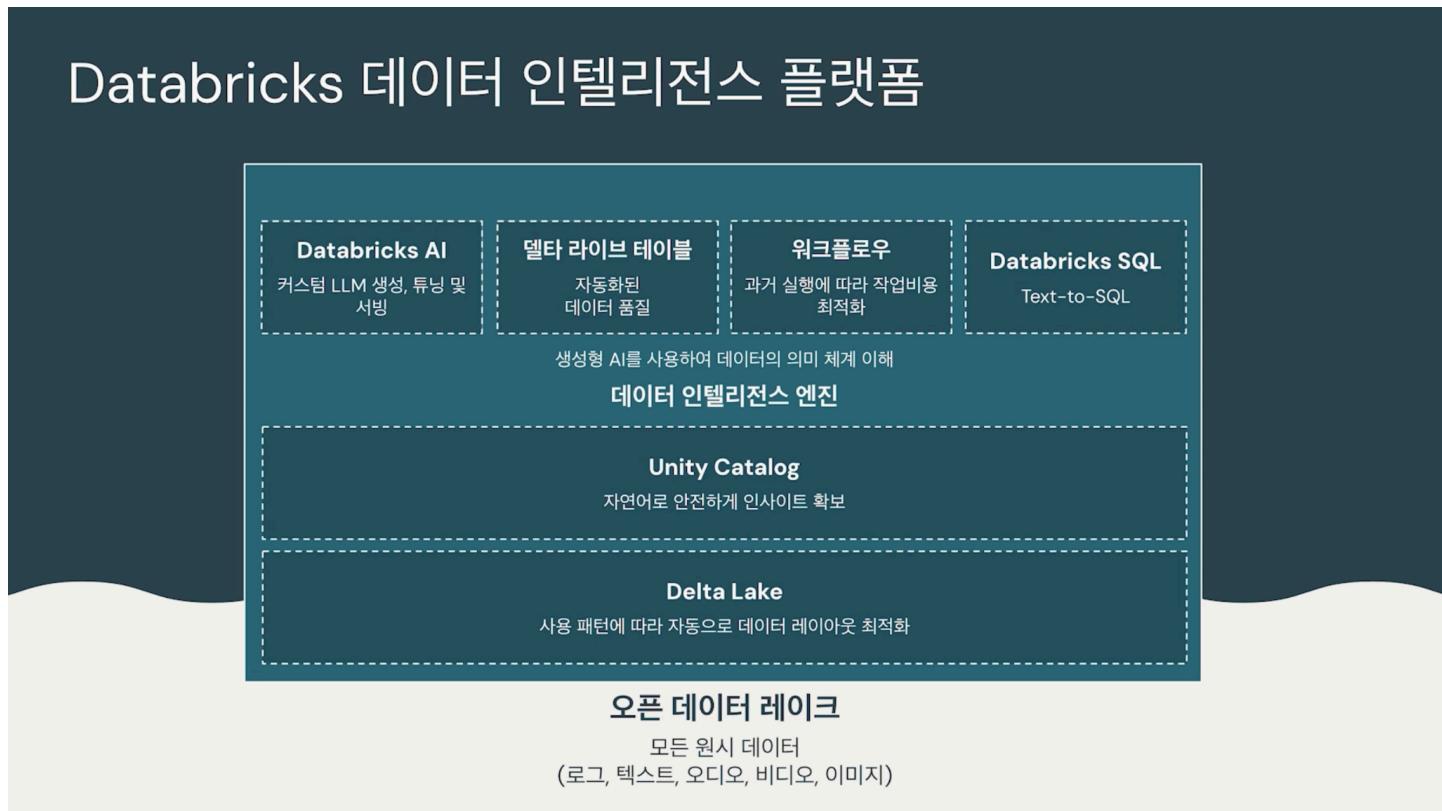
## 학습 목표

데이터브릭스 Data Intelligence 의 기본 아키텍쳐 설명

Delta Lake 특징과 이점 및 중요성 설명

Unity Catalog 의 특징과 이점 및 중요성 설명

## Databricks DI 기본 아키텍쳐

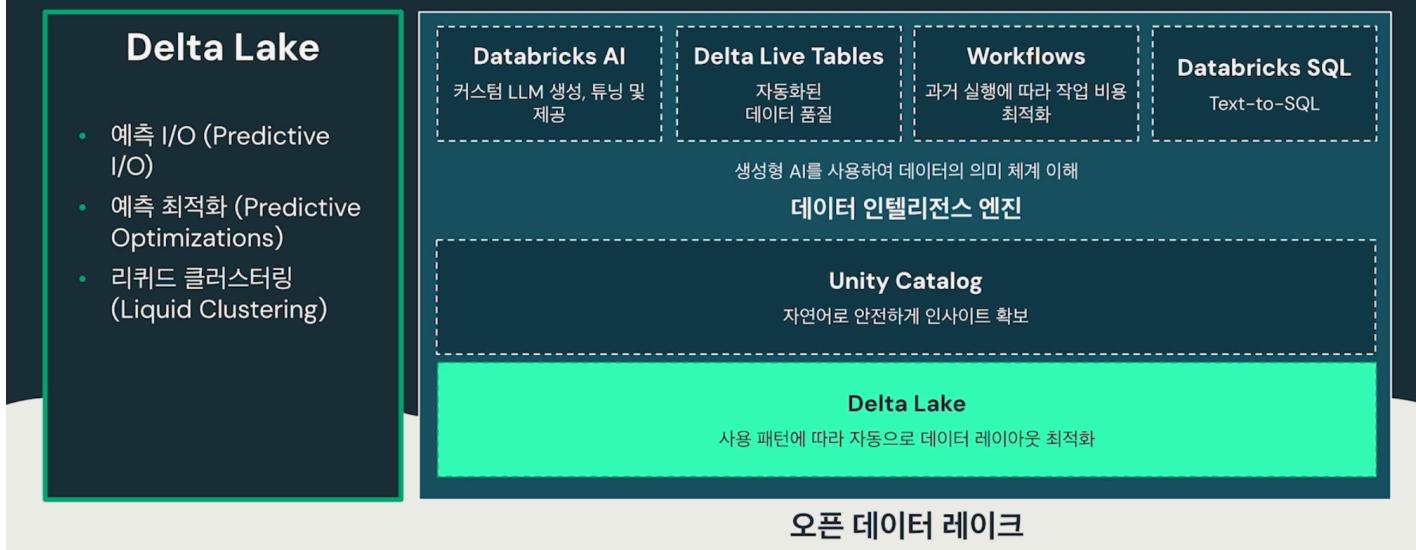


## Delta Lake

unified data storage layer

조직의 사용 패턴에 따라 자동적으로 최적화

# Databricks 데이터 인텔리전스 플랫폼



file based open source storage format

ACID 트랜잭션 보장

확장 가능한 데이터 및 메타데이터 처리

트랜잭션 로그를 활용한 감사 기록 및 time travel (시간 이동)

스키마 적용 및 스키마 진화

삭제, 업데이트 및 병합 지원

통합 스트리밍 및 배치 데이터 처리

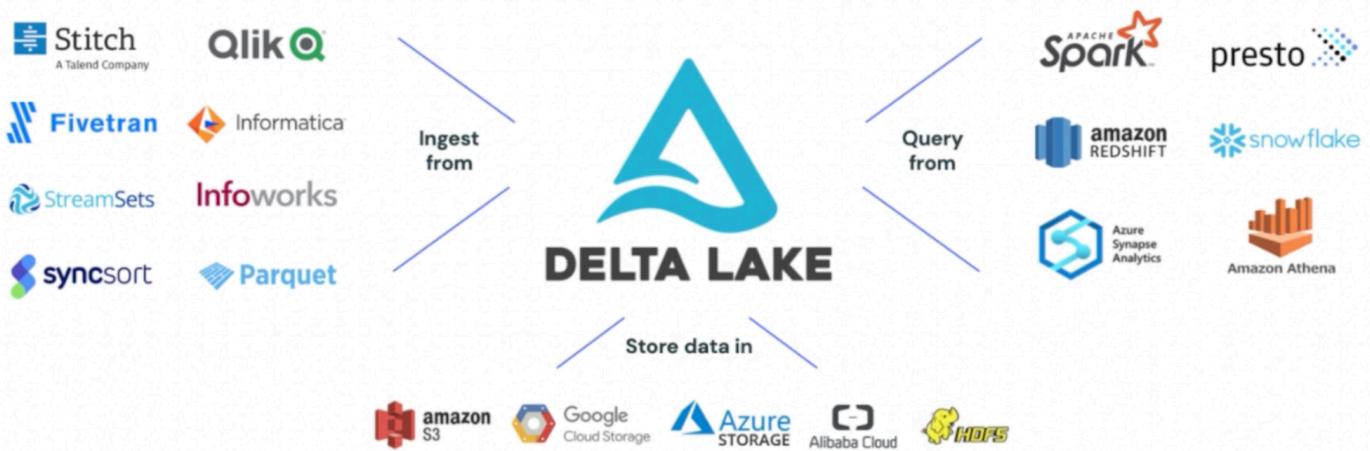
Apache Spark 와 호환

델타 테이블 사용. Apache Parquet 기반. common format

정형, 반정형, 비정형 데이터 지원

## Delta Lake integrates with all major analytics tools

Eliminates unnecessary data movement and duplication



## Unity Catalog

Unified Governance, security, cataloging layer  
data access control

# Databricks 데이터 인텔리전스 플랫폼

## Unity Catalog

- 문맥을 이해하는 검색
- 테이블 및 컬럼 자동 설명
- 자동화된 리니지
- 엔드-투-엔드 가시성 및 모니터링
- AI 모델 공유

데이터 사이언스 & AI

Databricks AI

ETL & 실시간 분석

Delta Live Tables

오케스트레이션

Workflows

데이터 웨어하우징

Databricks SQL

생성형 AI를 사용하여 데이터의 의미 체계 이해

데이터 인텔리전스 엔진

Unity Catalog

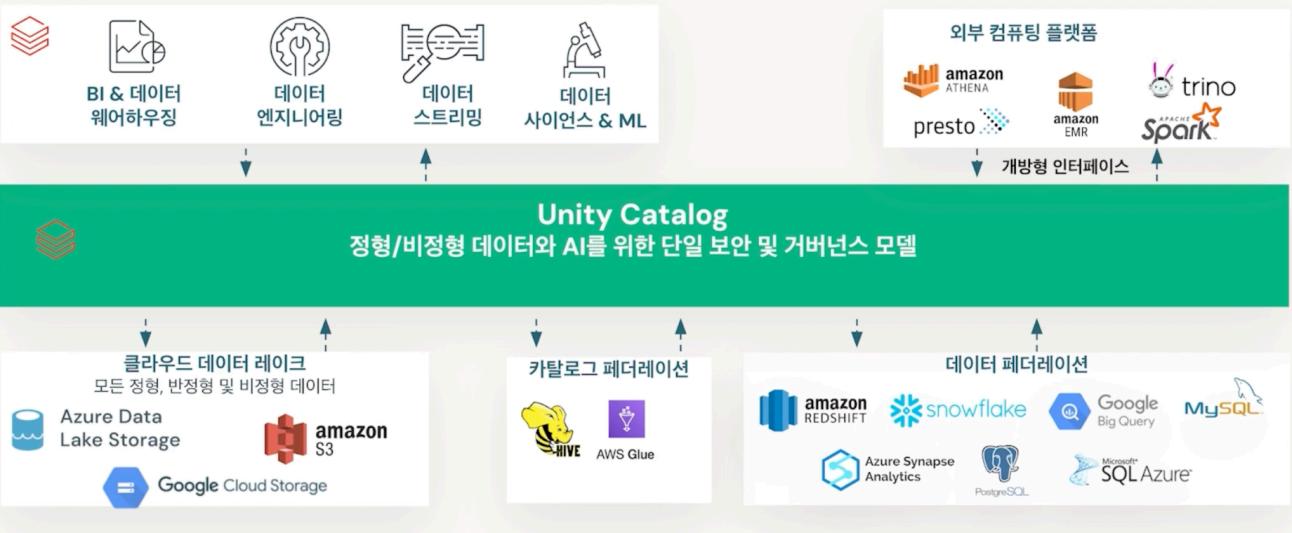
자연어로 안전하게 인사이트 확보

Delta Lake

사용 패턴에 따라 자동으로 데이터 레이아웃 최적화

오픈 데이터 레이크

## Unity Catalog는 데이터와 AI 거버넌스를 통합



# Databricks Unity Catalog

The screenshot shows the Databricks Unity Catalog interface. At the top, there are five navigation tabs: 사용자 경영 (User Management), 접근 컨트롤 (Access Control), 데이터 협통 (Data Integration), 자동 모니터링 (Automatic Monitoring), and 감사 (Audit). Below these tabs is a search bar with placeholder text: 데이터 검색 및 분류 (파일 | 테이블 | 노트북 | 모델 레지스트리 | 기능 저장소). The main area is currently empty.

## 데이터, 분석, AI를 위한 통합 거버넌스

- 데이터와 AI 자산에 대한 통합 뷰
- 데이터와 AI에 대한 단일 권한 모델
- AI 기반 모니터링과 리포팅
- 개방형 데이터 공유와 협업

## 데이터와 AI 자산에 대한 통합 뷰

- 정형, 비정형 데이터뿐만 아니라 노트북, ML 모델, ML 피처와 파일을 한 곳에서 검색, 분류 및 구성합니다.
- 데이터 페더레이션을 활용하여 수집 없이 외부 데이터 원본의 데이터를 등록하고 쿼리하여 분석 기능을 확장합니다.
- 효율적인 태그 기반 검색으로 데이터를 더 잘 이해하고 더 빠르게 인사이트를 얻을 수 있습니다

The screenshot shows the Databricks Unity Catalog interface with a search results page. On the left, there is a sidebar with a tree view of catalog structures. A specific section under 'models/default' is highlighted with a red box. On the right, a modal window titled 'Create a new connection' is open, listing various database types like SNOWFLAKE, DATABRICKS, MySQL, etc. The URL field contains 'jerry\_s\_t@prod.west-us.azuredatabricks.net:443?'. The modal has a 'Create' button at the bottom right.

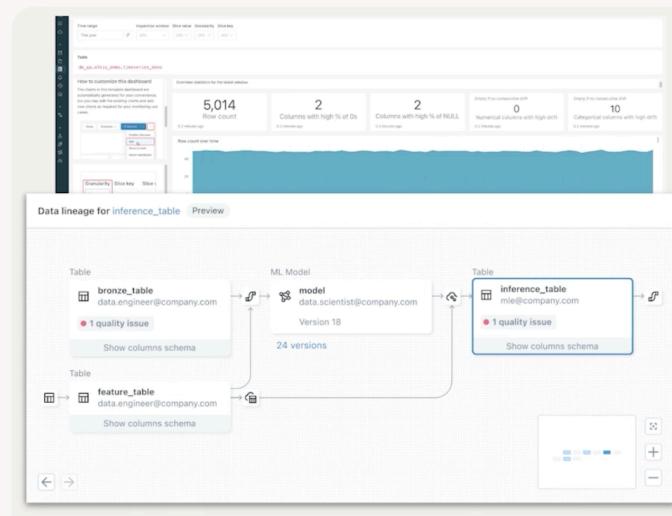
## 데이터와 AI에 대한 단일 권한 모델

- 모든 데이터와 AI 자산에 대한 액세스 정책을 관리하는 통합 인터페이스로 데이터 자산을 보호
- 행과 열에 대한 세분화된 액세스 제어 사용으로 보안 강화
- 개방형 인터페이스를 사용하여 다양한 컴퓨팅 플랫폼에서 안전하게 데이터에 접근

The screenshot shows the Databricks Data Explorer interface. A 'Grant on prod.finance' dialog box is open over the main Data Explorer window. The dialog shows the schema 'prod.finance' selected. It includes sections for 'Granted privileges will be inherited by applicable objects (e.g. tables, views) in this schema. Learn more.', 'Users also require USE CATALOG on the parent catalog to perform actions in this schema.', 'Users and groups' (set to 'ANALYST\_USA'), 'Privilege presets' (set to 'Data Reader (Can read from any object in the schema)'), and 'Privileges' (checkboxes for USE SCHEMA, SELECT, MODIFY, EXECUTE, CREATE TABLE, CREATE FUNCTION, CREATE MODEL, and ALL PRIVILEGES). There are 'Cancel' and 'Grant' buttons at the bottom right.

# AI 기반 모니터링과 리포팅

- 데이터와 ML 모델 파이프라인의 품질 문제 및 오류에 대해 사전 경고
- 효율적인 근본 원인 분석과 오류 디버깅을 위해 컬럼 수준까지 실시간 데이터 리니지 활용
- 자동 생성된 대시보드를 활용하여 데이터 및 ML 품질 보고서를 쉽게 공유
- 데이터 흐름과 소비에 대한 명확한 **엔드 투 엔드 뷰**를 확보하여 규정 준수 및 감사 준비 상태 보장



# Data Intelligence Architecture 및 보안 기본 사항

## 학습 목표

플랫폼 아키텍처에서 통합 아키텍처와 보안의 중요성

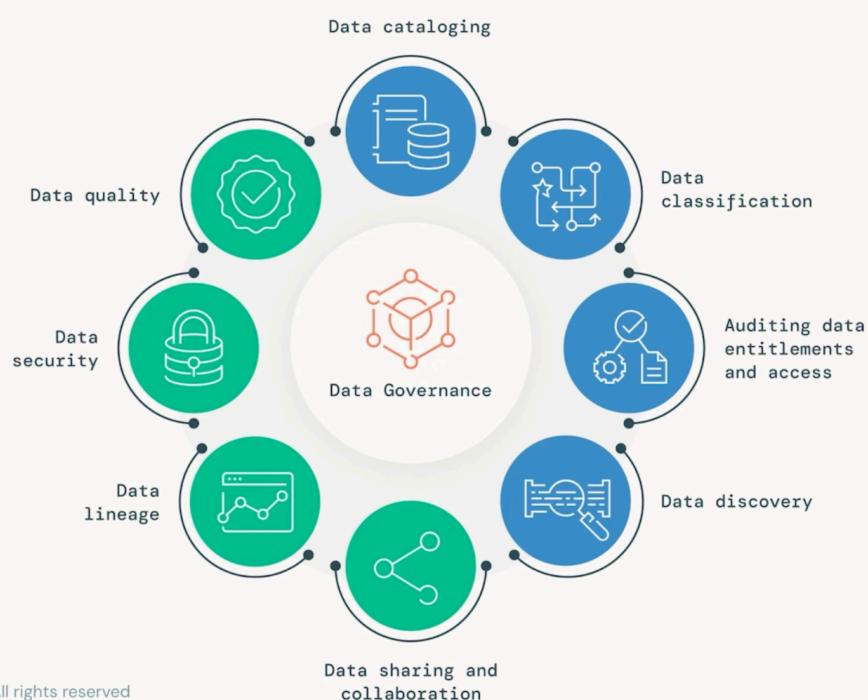
Unity Catalog가 데이터브릭스 DI 플랫폼의 필수 데이터 거버넌스 기능인 이유

Unity Catalog를 데이터 인텔리전스 엔진과 함께 사용할 때 데이터 거버넌스가 향상되는 이유

Delta Sharing 및 Databricks Marketplace의 중요성 설명

## Data Intelligence Platform과 보안

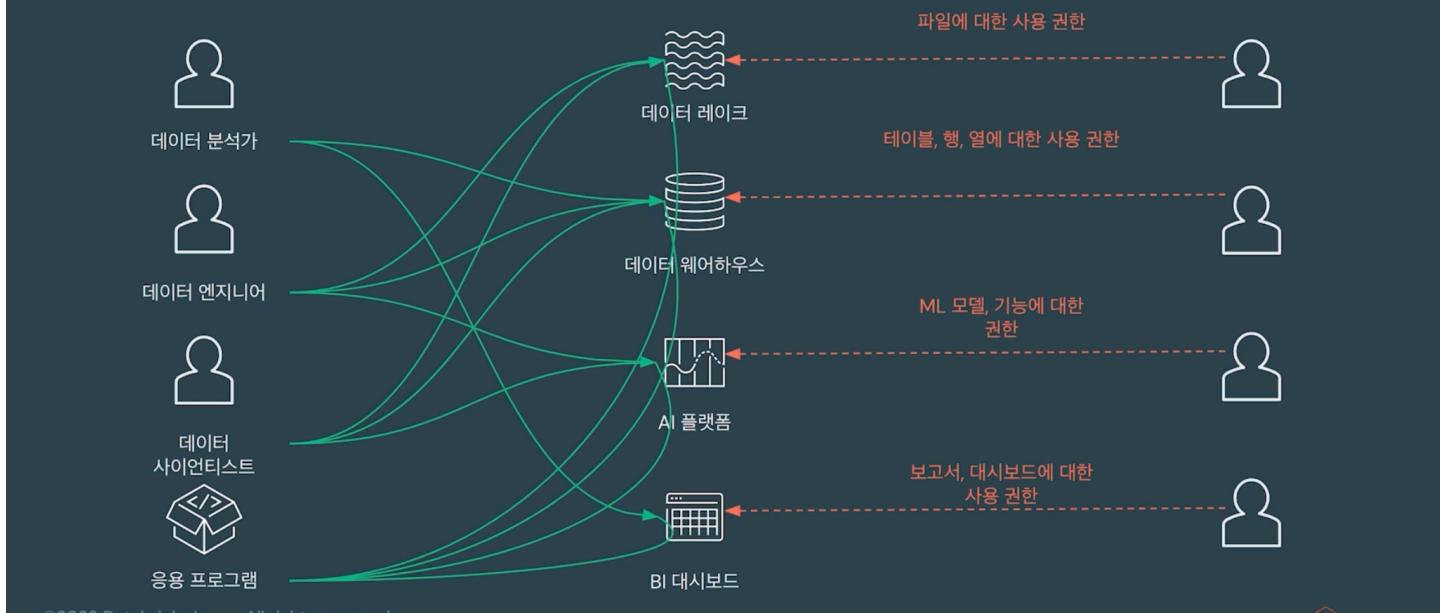
### 데이터 거버넌스의 핵심 요소



# 오늘날 데이터 및 AI 거버넌스는 복잡합니다

데이터 소비자

데이터 거버넌스/보안 팀

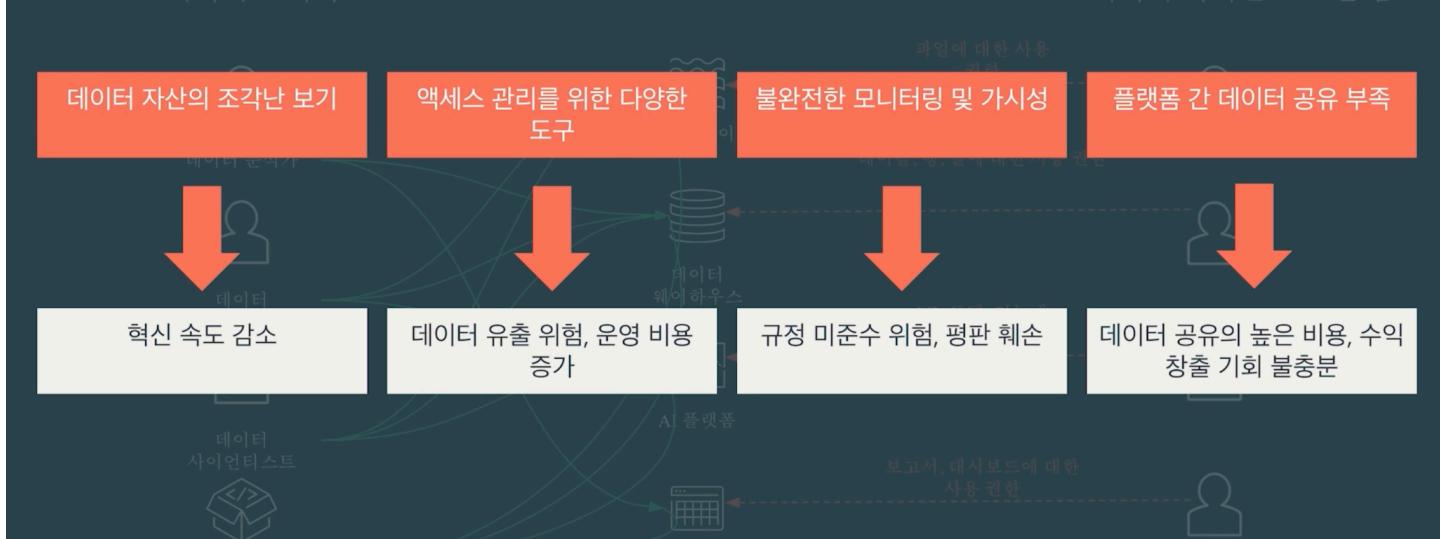


데이터 거버넌스를 지속적으로 운영하는 것이 현실에서는 어려운 도전과제입니다. 정책이 변경될 때..

# 오늘날 데이터 및 AI 거버넌스는 복잡합니다

데이터 소비자

데이터 거버넌스/보안 팀



# Databricks 데이터 거버넌스 제품들

## Unity Catalog

통합 거버넌스 및 보안

## Delta Sharing

조직 간 공유

## Databricks Marketplace

데이터 자산의 상용화

## Databricks Cleanroom

프라이빗, 보안 컴퓨팅

## Unity Catalog는 복잡성을 줄여줍니다.

데이터 자산의 파편화된 뷰

액세스 관리를 위한 다양한 도구

불완전한 모니터링 및 가시성

플랫폼 간 데이터 공유 부족



데이터 자산의 통합 뷰

데이터 및 AI에 대한 단일 권한 모델

AI 기반 모니터링 및 리포팅

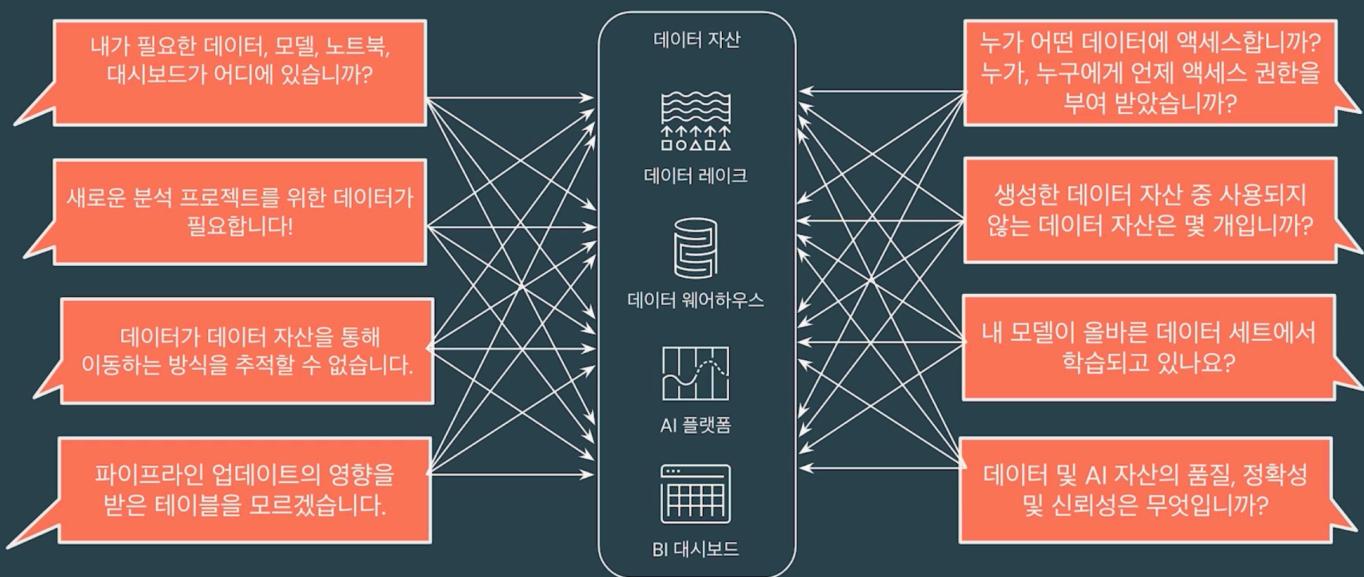
플랫폼에 기본 제공되는 Delta Sharing

전체 플랫폼이 Generative AI 를 활용

# Databricks 데이터 인텔리전스 플랫폼



## 거버넌스는 인텔리전스 없이는 복잡 합니다.

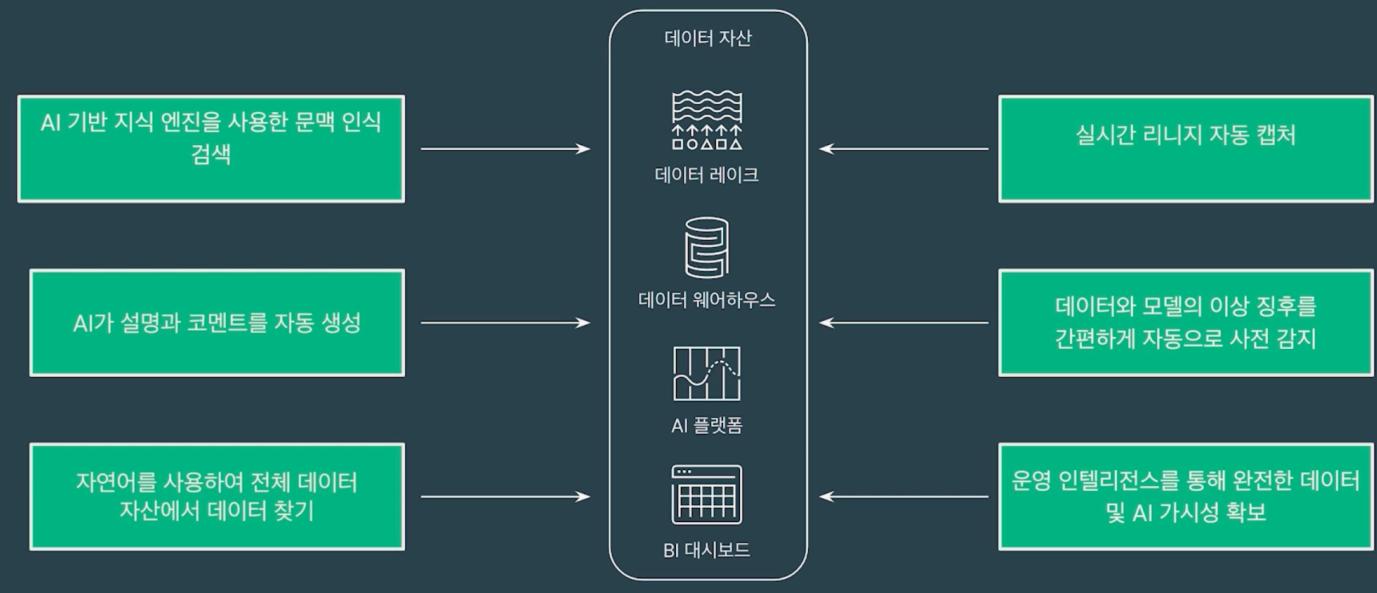


©2023 Databricks Inc. — All rights reserved

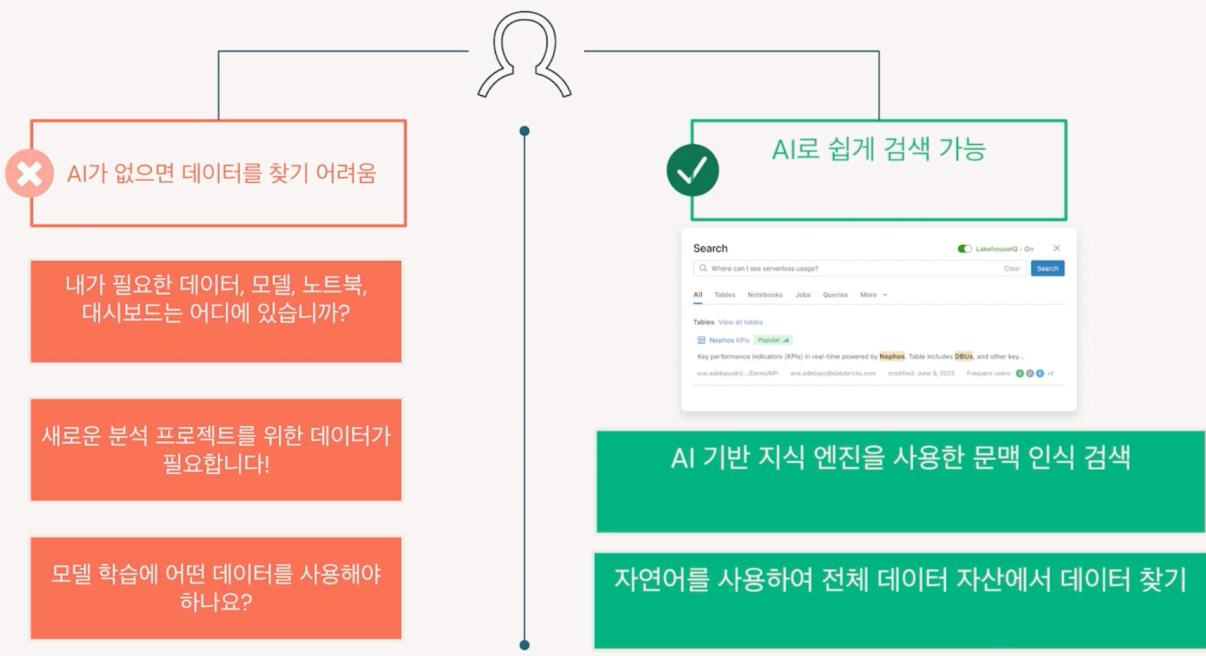


플랫폼 적으로 지원하더라도 적절하게 데이터 거버넌스를 운영하는 것은 어려운 일. 인텔리전스 필요성

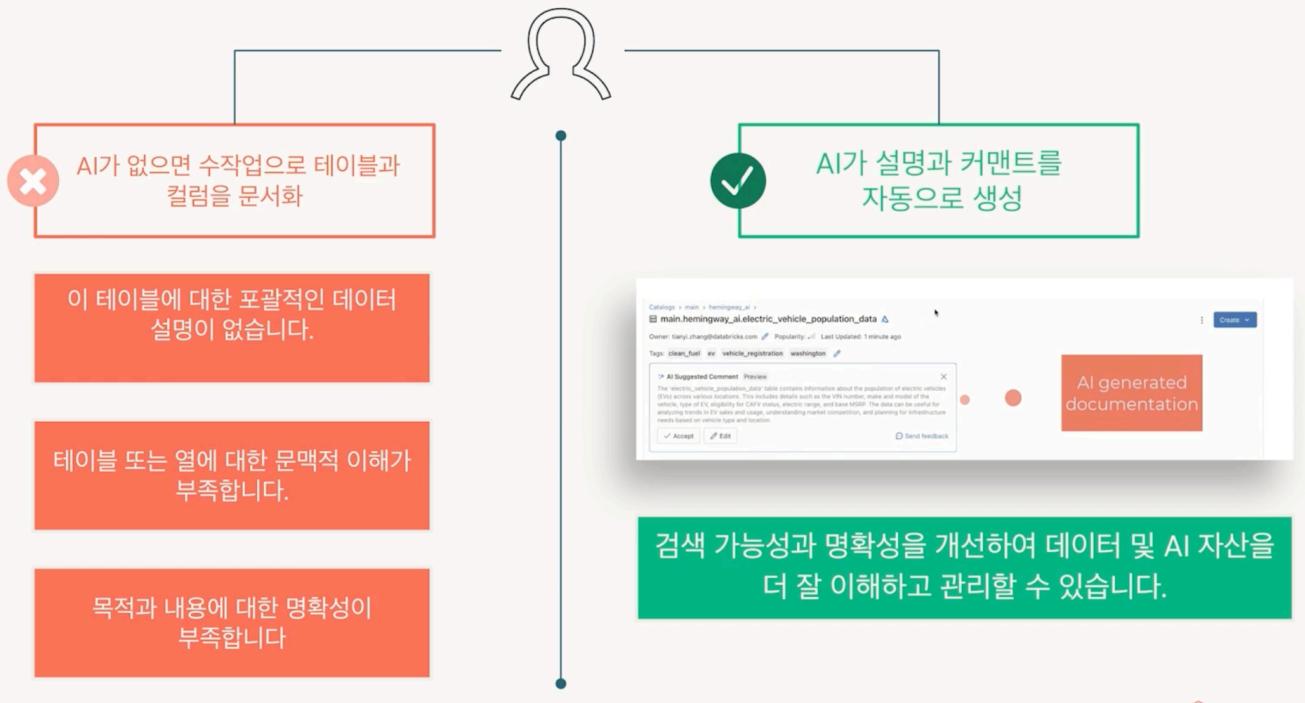
# 인텔리전스를 통한 거버넌스 간소화



## 데이터와 AI 자산을 자동으로 찾고 검색



# AI로 데이터 문서화 강화



## 모든 워크로드에 대한 자동화된 리니지



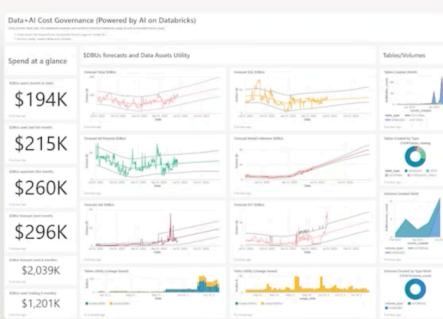
# AI 기반 모니터링과 가시성

누가 어떤 데이터에 액세스합니까?  
누가, 누구에게 언제 액세스 권한을  
부여받았습니까?

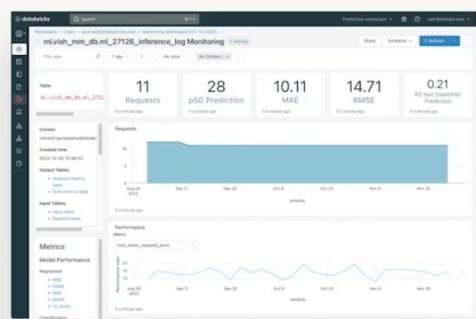
생성한 데이터 자산 중 사용되지  
않는 데이터 자산은 몇 개입니까?

내 모델이 올바른 데이터 세트에서  
학습되고 있나요?

데이터 및 AI 자산의 품질, 정확성  
및 신뢰성은 무엇입니까?



빌링, 감사, 리니지 등을 위한 운영 인텔리전스를 통해  
완전한 데이터 및 AI 가시성을 확보



데이터와 모델의 이상 징후에 대해 자동화되고 사전  
예방적이며 단순화된 감지

## Delta Sharing 과 Databricks Marketplace

### Delta Sharing의 장점

- 개방형 방식으로 플랫폼 간 공유
- 복사 없이 라이브 데이터를 공유
- 중앙 집중식 관리 및 거버넌스
- 데이터 제품을 위한 마켓플레이스
- 개인 정보 보호 데이터 클린룸



# Databricks 마켓플레이스

모든 데이터, 분석 및 AI를 위한 개방형 마켓플레이스

공급자

모든 플랫폼의  
사용자에게 도달

데이터 뿐만 아니라  
그 이상의 거래 가능

안전하게 데이터 공유

소비자

데이터 뿐만 아니라  
그 이상의 뷰

빠른 데이터 제품 평가

벤더 종속 방지



## Databricks Clean Rooms

협력자 1



기존 테이블



Delta Sharing  
프로토콜

DATABRICKS TRUSTED COMPUTE



신뢰할 수 있는 컴퓨팅에서  
상호 승인된 작업

협력자 N



기존 테이블



Delta Sharing  
프로토콜

자유로운 컴퓨팅

Python, SQL, R, Java 등으로 모든 컴퓨팅  
실행

데이터 복제 없음

Delta Sharing은 복제 없이 지역 간 또는  
클라우드 간 공유를 제공

필요에 따라 확장 가능

모든 크기의 데이터에 대해 여러 공동  
작업자로 쉽게 확장

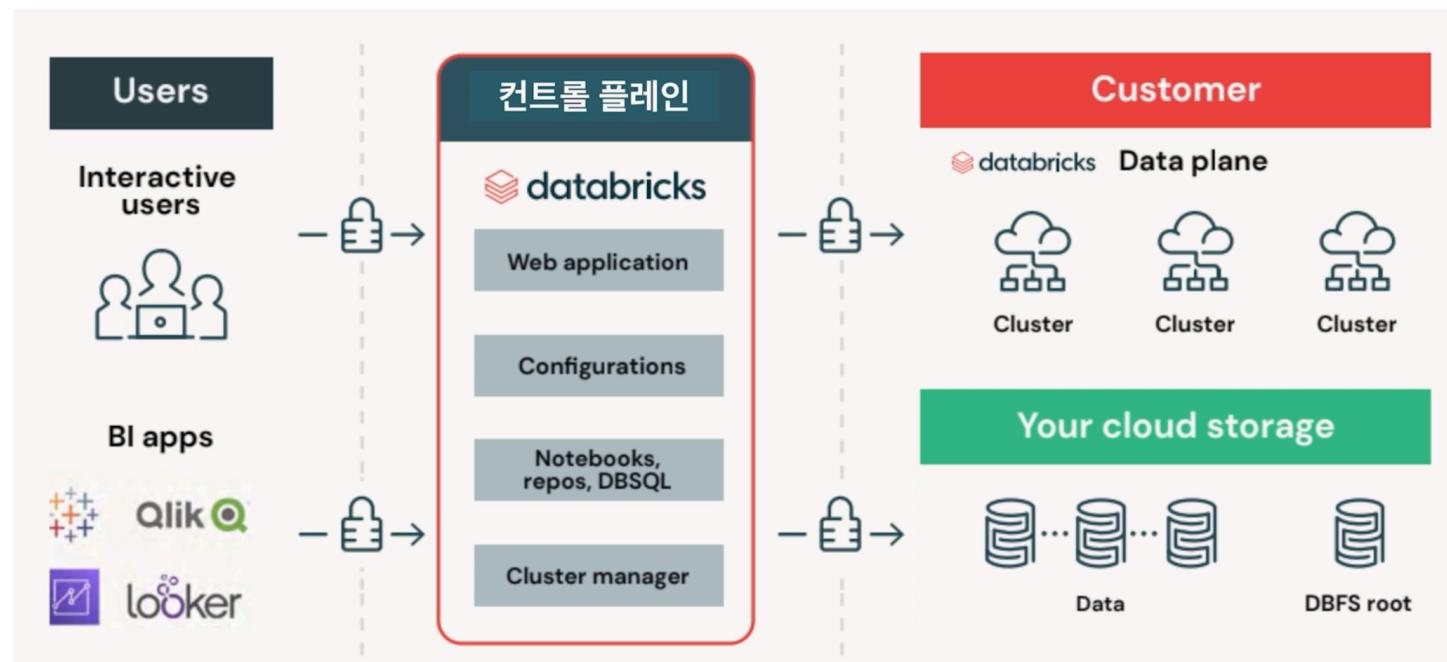
# Databricks 보안, 안정성 및 성능

## 학습목표

- 컨트롤 플레인과 데이터 플레인을 구분
- Databricks DI 플랫폼의 보안 기능 파악
- Databricks DI 플랫폼의 컴퓨팅 리소스 설명
- 서버리스 컴퓨팅 정의
- Databricks Serverless SQL 사용의 이점
- Photon을 활용한 Databricks DI 플랫폼의 성능 개선 방법

## 보안 아키텍처

컨트롤 플레인 및 데이터 플레인 : 아키텍처를 컨트롤 플레인과 데이터 플레인으로 분리



컨트롤 플레인은 데이터브릭스가 제공하는 관리형 백엔드 서비스로 구성  
데이터 플레인은 데이터가 처리되는 곳

## 사용자 ID 및 액세스

- 테이블 ACL 기능
- IAM 인스턴스 프로파일
- 안전하게 저장된 액세스 키
- 시크릿 API



## 데이터 보안

Databricks encryption capabilities are in place both at rest and in motion

### For data-at-rest encryption:

- Control plane is encrypted
- Data plane supports local encryption
- Customers can use encrypted storage buckets
- Customers at some tiers can configure customer-managed keys for managed services

### For data-in-motion encryption:

- Control plane <-> data plane is encrypted
- Offers optional intra-cluster encryption
- Customer code can be written to avoid unencrypted services (e.g., FTP)

데이터브릭스는 거버넌스 및 보안 구조를 통해 암호화, 격리, 감사를 제공함

각 팀이나 부서에서 서로 다른 워크스페이스를 사용할 경우 워크스페이스 등 다양한 수준에서 사용자를 격리 가능

클러스터 수준에서는 클러스터 ACL을 통해 노트북에서 접속하는 클러스터 제어

# 컴플라이언스

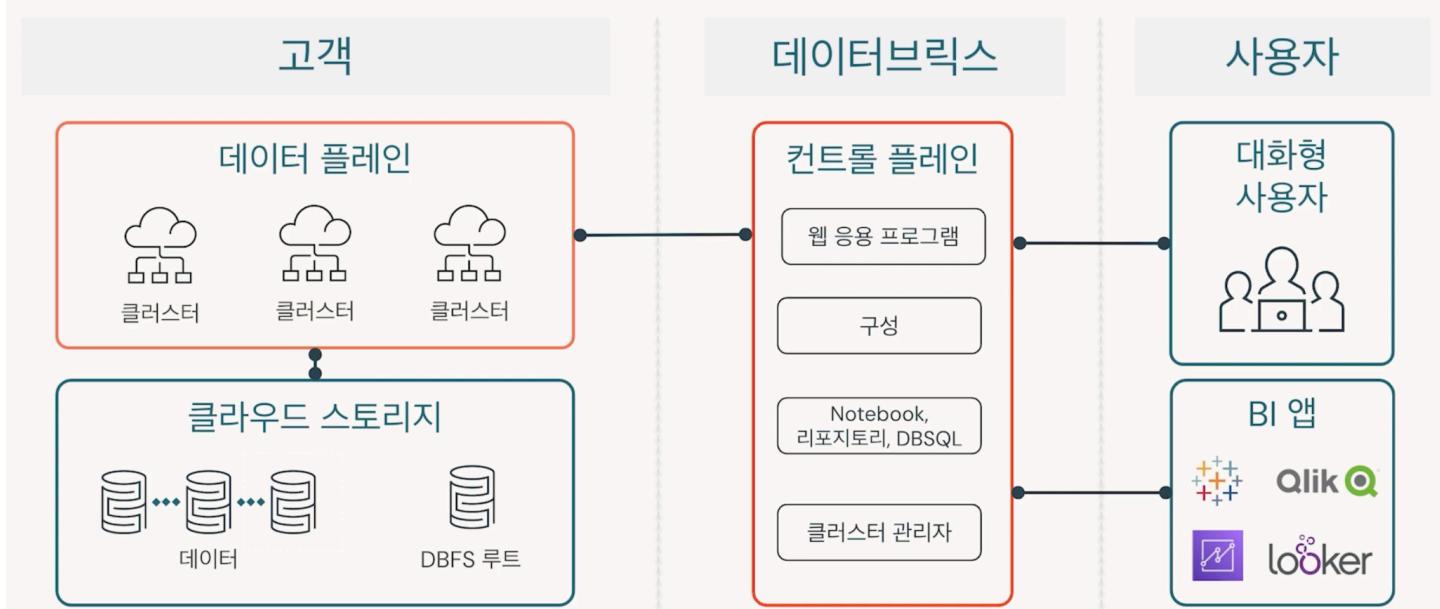
- SOC 2 Type II
- ISO 27001
- ISO 27017
- ISO 27018

- FedRAMP High
- HITRUST
- HIPAA
- PCI

GDPR 및 CCPA 대비 완료



## 클래식 데이터 플레인



## 컴퓨팅 리소스의 도전 과제

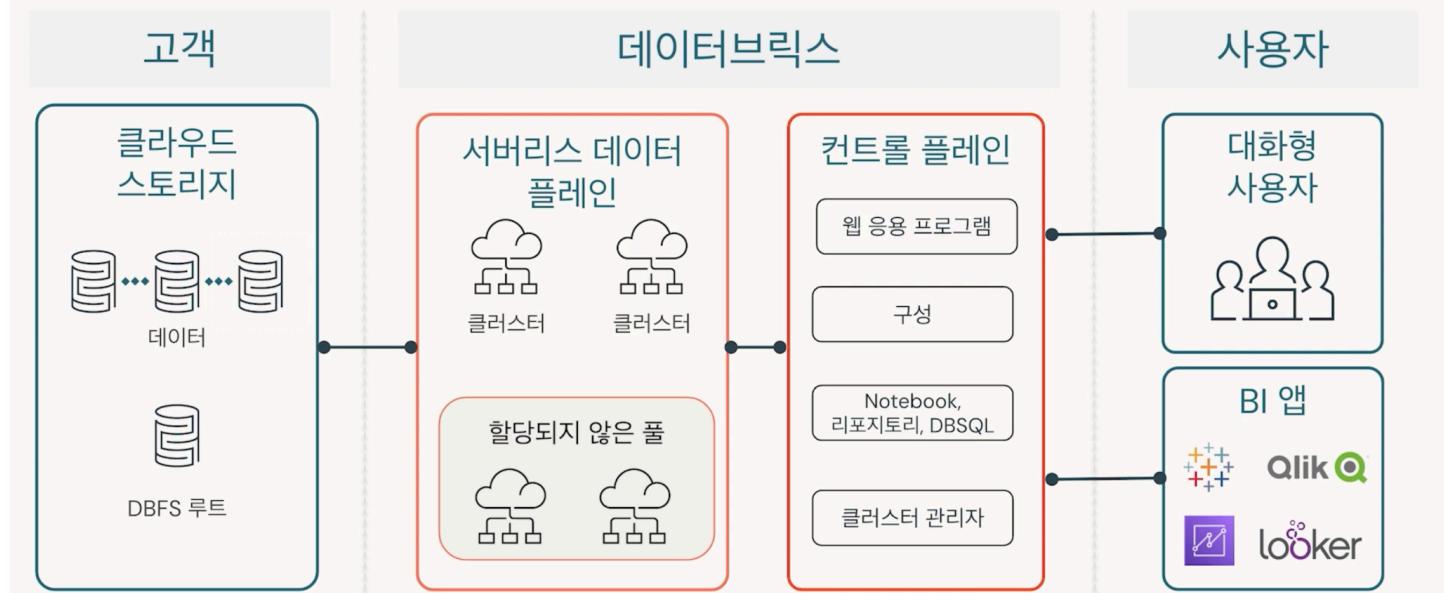
- 클러스터 생성이 복잡
- 환경 시작이 느림
- 클라우드 계정 제한 사항 및 리소스 옵션
- 장기 실행 클러스터
- 리소스의 과잉 프로비저닝
- 더 높은 리소스 비용
- 높은 관리 오버헤드
- 생산성 낮은 사용자

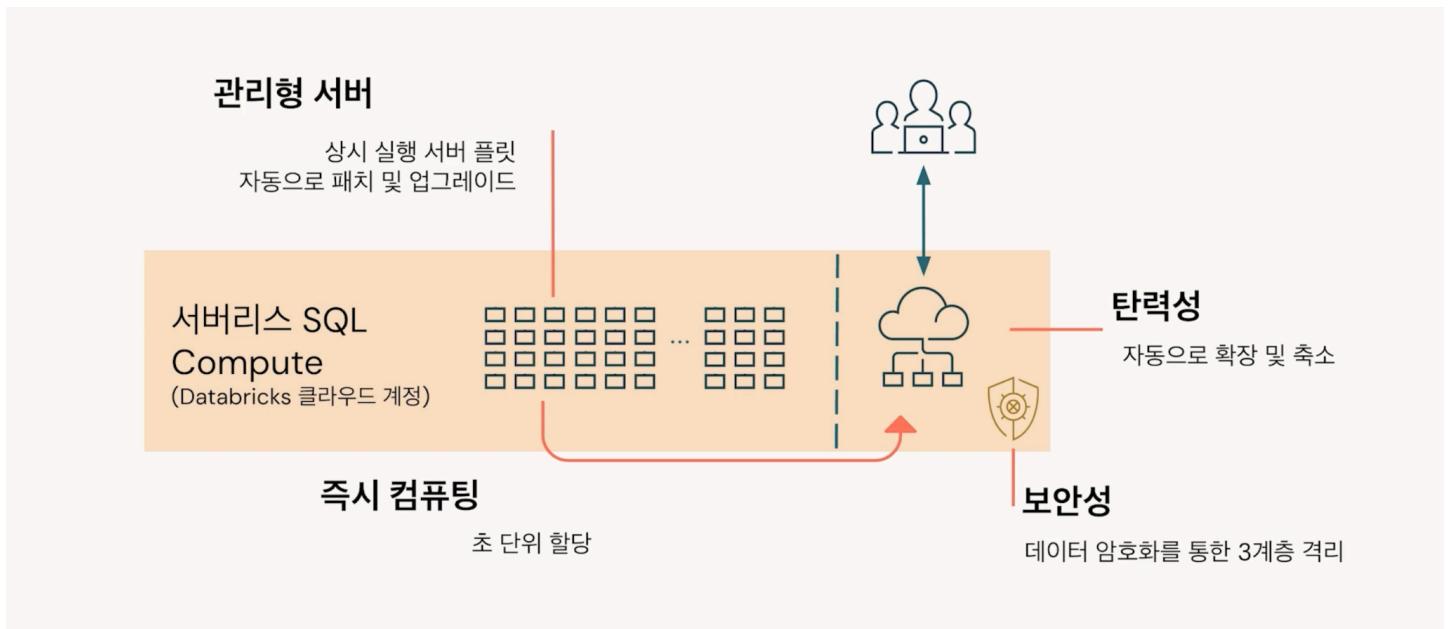
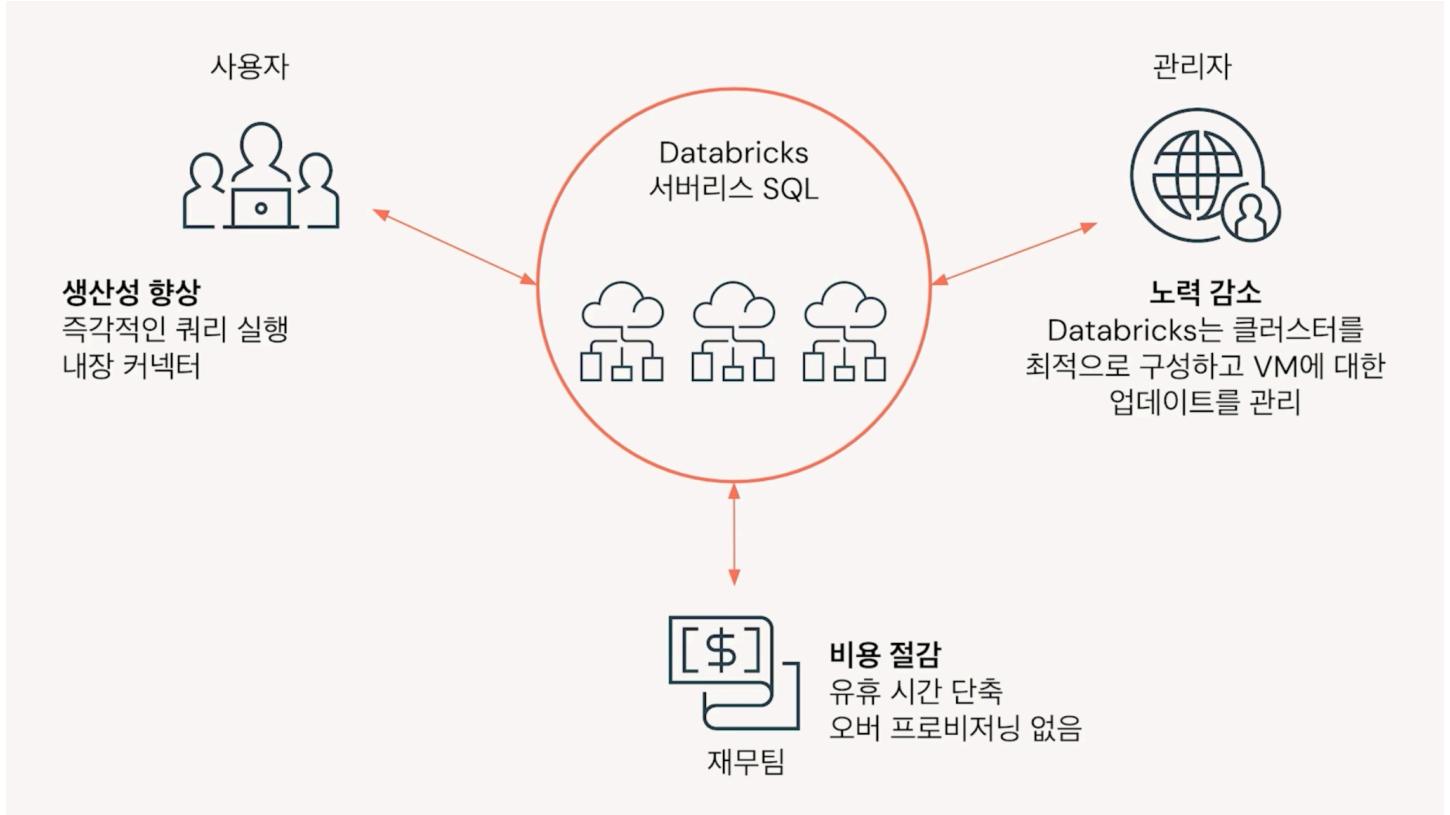
### 직면 과제

- 클러스터 생성의 어려움. 적합한 크기, 인스턴스 유형. 프로비저닝 어려움
- 리소스 관리
- 비용 관리

직면 과제를 위한 하나의 옵션 서비스 데이터 플레이인

## 서비스 데이터 플레이인

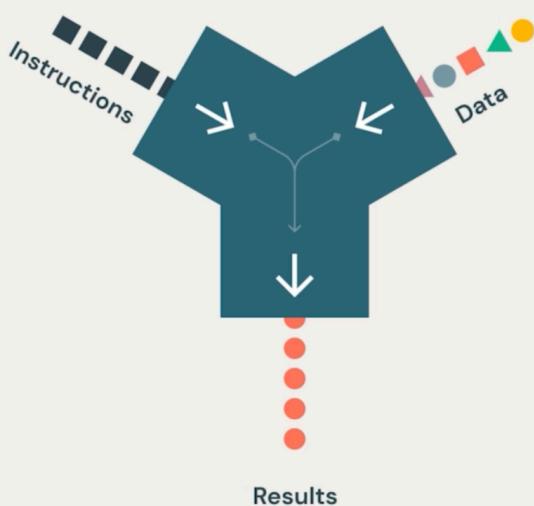




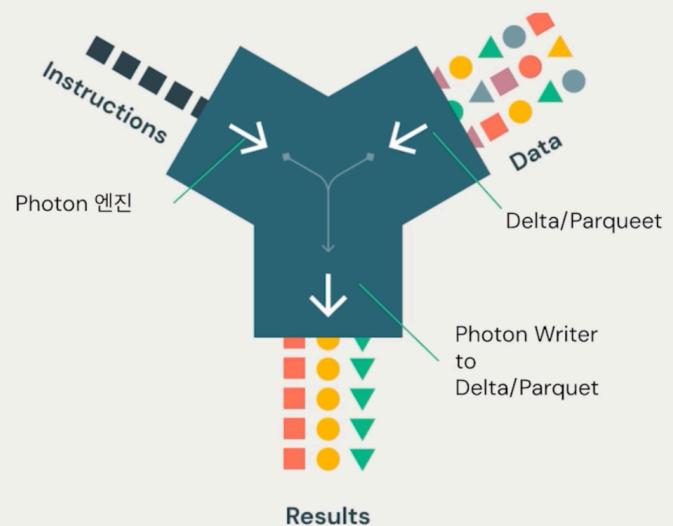
## Photon

Next Generation 쿼리 엔진  
인프라 비용 절감, 성능 효과 제공

## Spark 인스트럭션



## Photon 인스트럭션

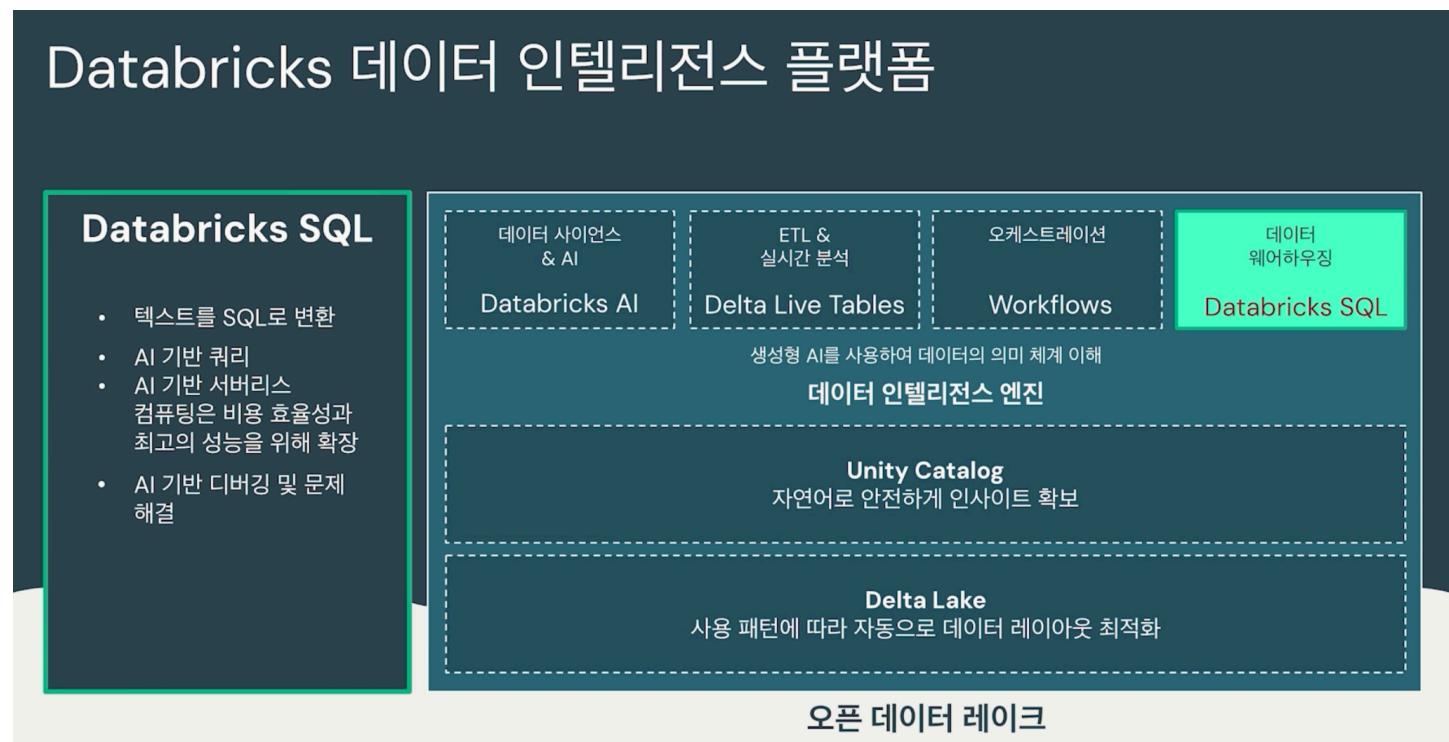


# Databricks DI 플랫폼에서 지원되는 워크로드

## 데이터웨어하우징

### 학습 목표

Databricks DI 플랫폼에서 Databricks SQL 을 사용하여 데이터웨어하우징을 지원하는 방법  
Databricks DI 플랫폼에서 데이터웨어하우징 워크로드를 실행할 때의 이점



SQL 분석, BI, 데이터 변환, 쿼리, 대시보드 생성, 실시간 비즈니스 인사이트 등 모두 효율적으로 처리 가능 → 데이터 전문가가 효과적으로 작업하고 비용 효율적이면서 적시에 비즈니스 인사이트 제공  
모든 활동들은 자연어로 처리 가능  
AI 활용 비용 절감, 최고의 성능 속도 지원  
AI 기반 디버깅 지원

# AI가 없는 데이터 웨어하우징 문제



잘못된 자원 할당



확장의 어려움



더 높은 운영 비용

비효율적인 리소스 할당으로 인프라  
과다 사용 또는 과소 사용

더 큰 데이터 볼륨과 복잡한 쿼리를  
위한 확장의 어려움

수작업 증가로 운영 비용 증가

## 높은 성능을 위한 AI 기반 데이터 웨어하우징

### 자동 튜닝

관리형 테이블의 쓰기를  
자동으로 최적화하고  
스토리지를 압축하여 대기  
시간과 비용을 줄임

### 지능형 워크로드 관리

기계 학습을 활용하여 쿼리를  
효율적으로 라우팅하고 클러스터를  
확장하여 비용/성능을 극대화

### 예측 I/O

고비용의 검색 최적화  
인덱스에 필적하는 성능을  
자동으로 제공

## BI를 위한 Databricks Assistant

**텍스트에서 SQL로 변환:** 자연어를 SQL 쿼리  
로 손쉽게 변환

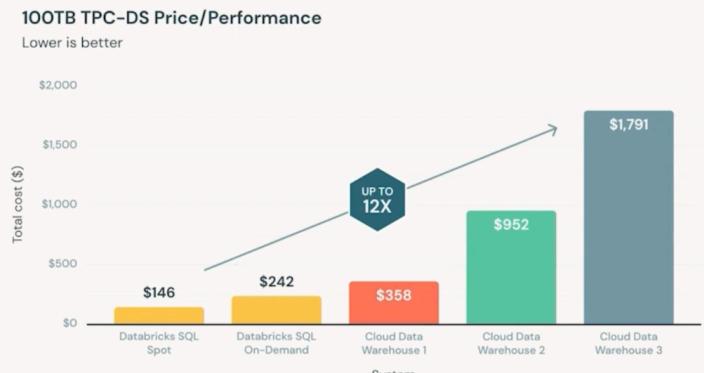
**코드/쿼리 자동 생성 및 자동 완성:** 지능형  
자동화로 코딩 간소화

**문제 진단 및 해결:** 문제를 식별하고  
해결방법을 제공

**Unity Catalog와 통합:**  
데이터 자산에 따라 문맥에 맞는 결과를 제공

# 데이터 웨어하우징을 위한 최고의 TCO 및 성능

- 최적의 가격/성능을 위한 인스턴스 유형 및 구성 자동 결정(최대 12배)
- 높은 **동시성** 기본 제공, 자동 부하 분산
- 지능형 **워크로드 관리**와 더 빠른 읽기
- 서비스를 통한 즉각적인 시작, 가용성 향상, 평균 40% 비용 절감



# 데이터 오케스트레이션 Workflow

## 학습 목표

데이터 오케스트레이션 작업에 지능형 자동화를 사용할 때의 이점 설명  
Databricks workflow 가 데이터 오케스트레이션을 지원하는 방법

## Databricks 데이터 인텔리전스 플랫폼

### Workflows

- 지능형 ETL 처리
- AI 기반 디버깅 및 문제 해결
- 엔드-투-엔드 가시성과 모니터링
- 광범위한 애코시스템 통합

데이터 사이언스 & AI

Databricks AI

ETL & 실시간 분석

Delta Live Tables

오케스트레이션

Workflows

데이터 웨어하우징

Databricks SQL

생성형 AI를 사용하여 데이터의 의미 체계 이해

데이터 인텔리전스 엔진

Unity Catalog

자연어로 안전하게 인사이트 확보

Delta Lake

사용 패턴에 따라 자동으로 데이터 레이아웃 최적화

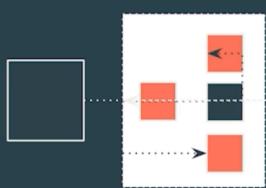
### 오픈 데이터 레이크

모든 원시 데이터  
(로그, 텍스트, 오디오, 비디오, 이미지)

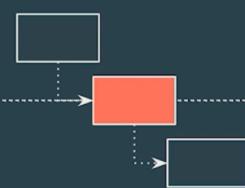
데이터 파이프라인을 좀 더 효과적으로 관리. 스마트 ETL 프로세스와 AI 를 활용하여 문제를 찾고 해결  
Databricks workflow 는 데이터 파이프라인과 툴을 완벽하게 파악하여 문제를 신속하게 해결할 수 있도록 지원

## 자동화가 없는 오케스트레이션은 비효율적

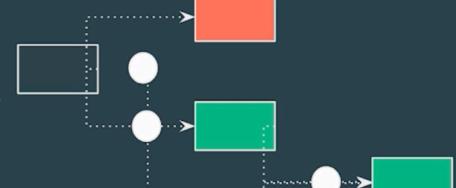
모듈형



순차형



조건부



↑  
파이프라인의 수동 시작

모듈 모듈  
비효율적인 자원 할당

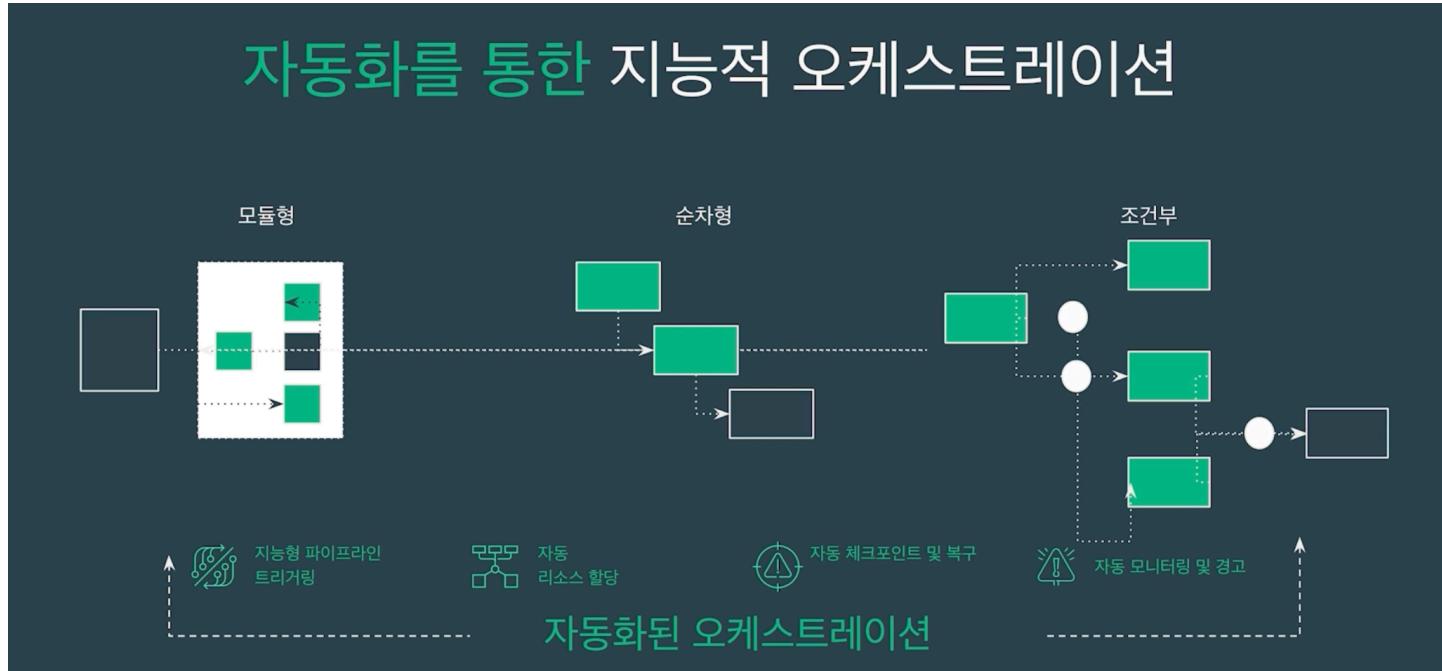
자동 복구 부족

자동 경고 또는 모니터링  
없음

오케스트레이션 과제

자동화가 없이 운영하는 것은 비효율적 → 파이프라인을 수동으로 재시작. 자동 복구 / 리소스 할당 어려움.  
모니터링 어려움

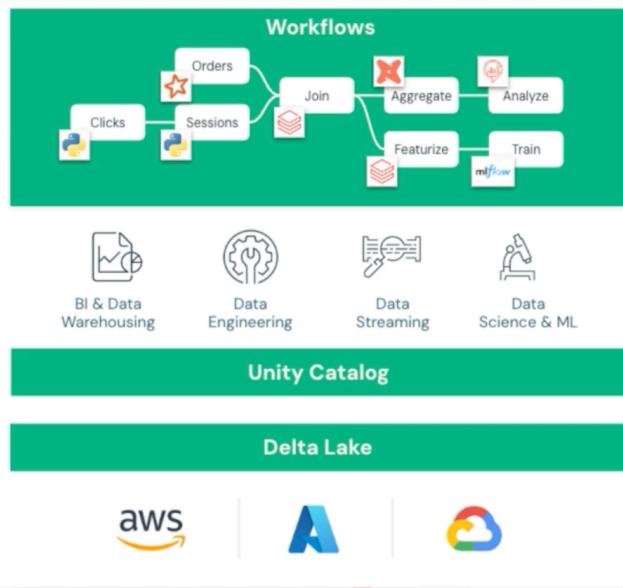
# 자동화를 통한 지능적 오케스트레이션



## 간소화되고 자동화된 오케스트레이션



# 자동화된 오케스트레이션의 이점



- **효율적인 트리거:** 자동화된 파이프라인 트리거는 원활한 프로세스 시작을 보장하여 수동 오류의 위험을 최소화
- **최적의 리소스 활용:** 자동 리소스 할당으로 컴퓨팅 리소스를 최적화하여 비용효율성 제고, 서비스 워크플로우를 통한 성능 향상
- **향상된 신뢰성:** 자동 체크포인트 및 복구는 안정적인 장애 대응으로 프로세스 연속성을 보장하고 장애 발생 시 마지막 성공 지점부터 원활하게 재개
- **사전 예방적 문제 해결 및 모니터링:** 모니터링 기능이 강화된 자동 경고는 시스템 관리자에게 문제를 신속하게 알려 빠르게 대응하고 다운타임을 줄이며 시스템 복원력을 보장

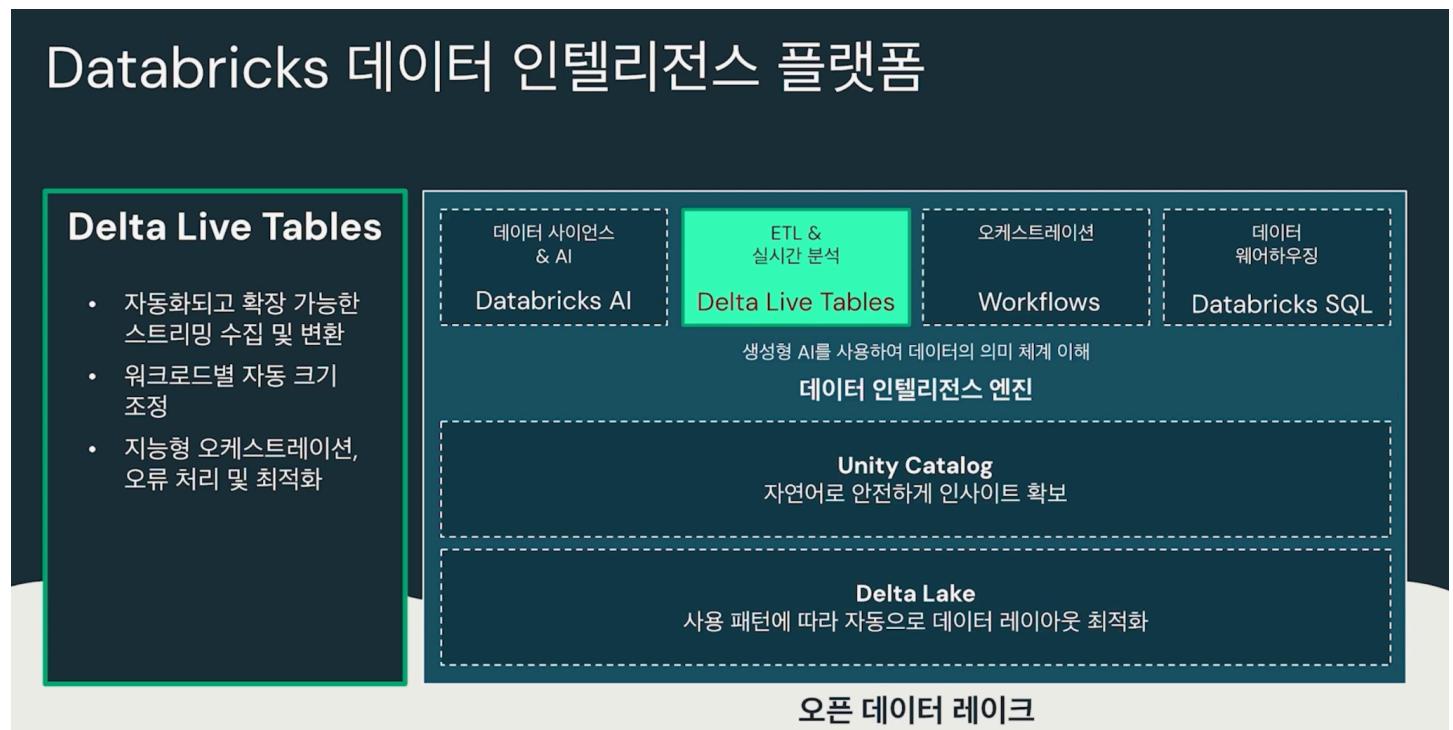
# Databricks 레이크하우스 플랫폼에서 지원되는 워크로드

## ETL 및 실시간 분석

### 학습 목표

DLT(Delta Live Table) 가 ETL 을 지원하는 방법

ETL 파이프라인을 최적화하는데 사용되는 DLT 및 AI 기능 통합의 이점



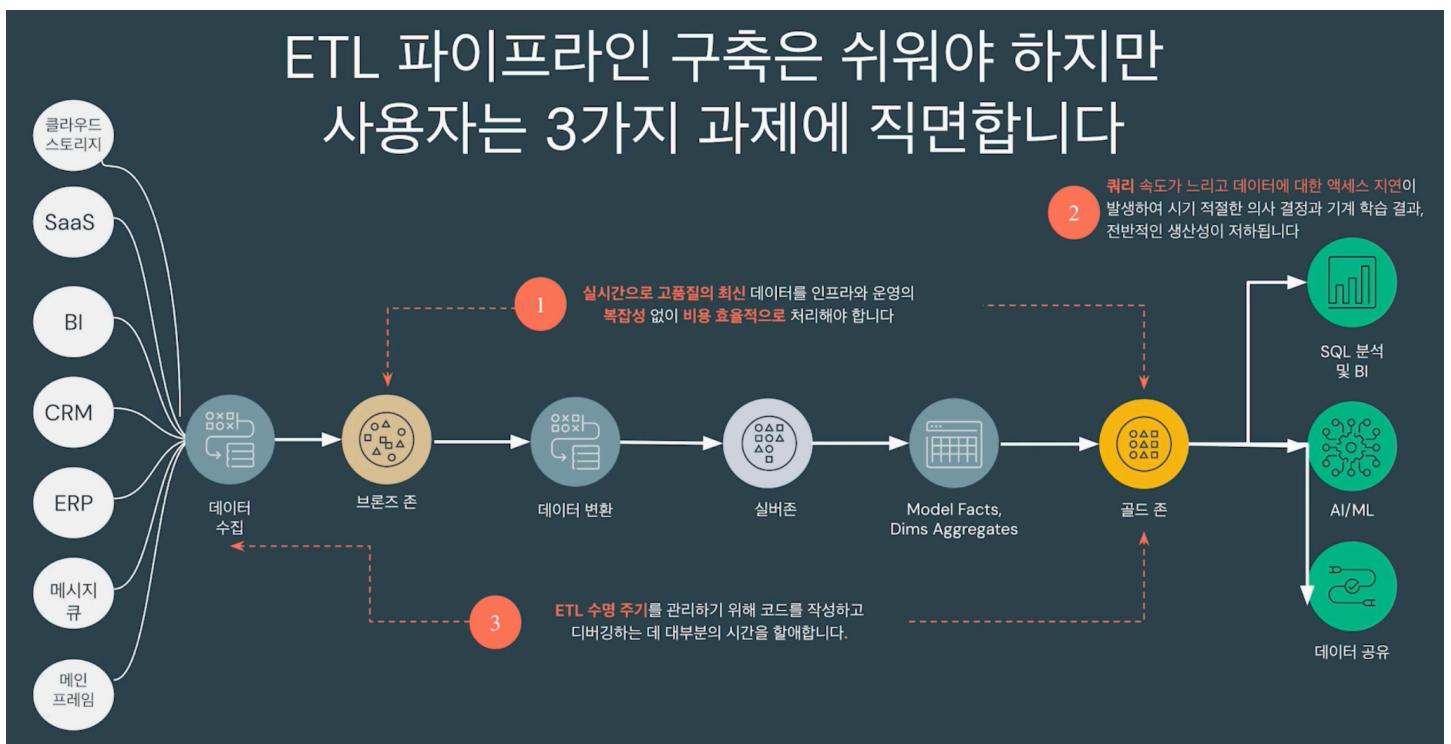
ETL 프로세스를 단순화하여 데이터 엔지니어가 작업이 아닌 데이터 변환에 집중

DLT는 워크로드별 자동 조정 기능과 파이프라인 모니터링 제공

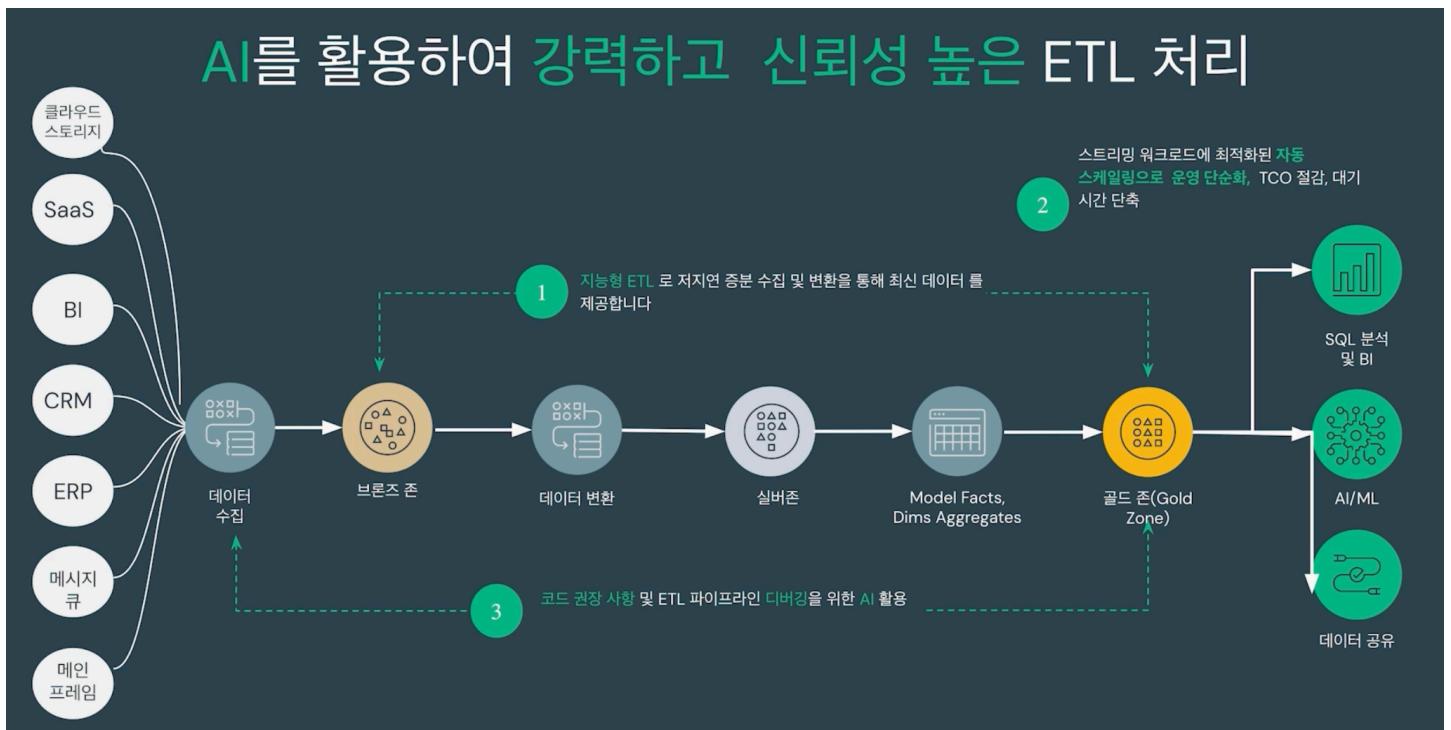
더 크고 복잡한 데이터 워크로드를 보다 효율적으로 처리 가능

ETL 파이프라인 관리의 어려움

# ETL 파이프라인 구축은 쉬워야 하지만 사용자는 3가지 과제에 직면합니다

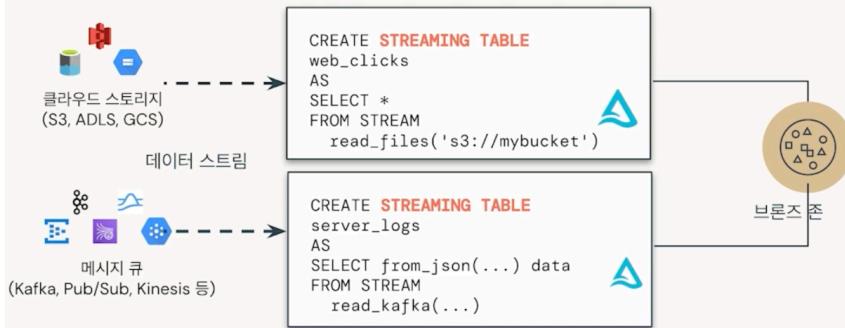


AI를 활용하여 강력하고 신뢰성 있는 ETL 처리 → 고품질의 데이터를 적시에 활용



AI를 활용하여 운영을 단순화, 자동화 및 최적화하여 TCO를 절감하고 스트리밍 워크로드의 지연을 최소화  
AI 추천으로 데이터 엔지니어의 운영 및 디버깅 추천 → 데이터에 집중하고, 운영하는 시간 절감

# 자동화되고 확장 가능한 데이터 수집



- 간단한 SQL 구문으로 모든 데이터 엔지니어와 분석가가 데이터 스트리밍에 액세스 가능
- 증분 처리와 대규모 배치를 통해 대량의 데이터를 효율적으로 처리하기 위한 **자동 수집 확장성**
- 스트리밍 데이터를 이용한 실시간 분석/BI, 기계 학습 및 운영 사용 사례 지원

## 증분 ETL을 위한 지능형 최적화

### MERGE updates to specific rows

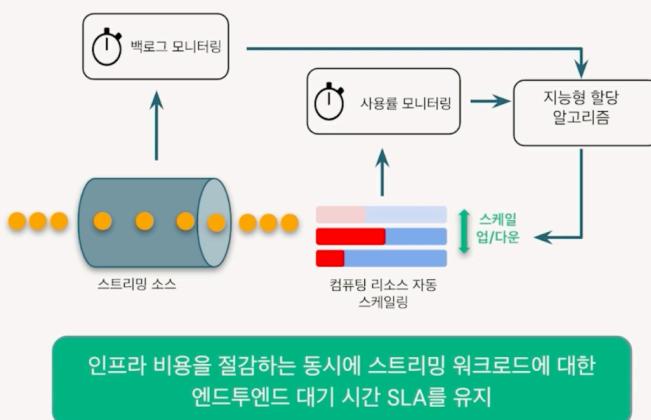
Use techniques from databases literature to compute changes to results

date	amount
2022-06-01	\$10
2022-06-01	\$46
2022-06-02	\$324
2022-06-02	\$24
2022-06-03	\$32
2022-06-03	\$18

date	sum
2022-06-01	\$56
2022-06-02	\$348
2022-06-03	\$50

- 엔드투엔드 ETL 파이프라인을 위한 **최적화된 자동 증분 수집 및 변환**
- 쿼리 계획 및 데이터를 기반으로 최상의 증분 전략을 **지능적으로** 평가합니다.
- 증분 새로 고침을 위한 가장 효율적인 기술 **자동 적용**

## AI로 강화된 자동 확장



- 컴퓨팅 및 스케줄링을 **동적으로 최적화**하여 TCO 및 대기 시간 단축
- 변화하는 스트리밍 워크로드에 **맞게 자동으로** 조정되어 인프라 지출을 최적화하는 동시에 다운스트림 SLA를 보장
- 급증 가능하고 예측할 수 없는 스트리밍 워크로드를 처리하도록 설계

# ETL을 위한 DatabricksIQ 실행

A screenshot of the Databricks Notebook interface. On the left, there's a sidebar titled 'Assisted' with a 'SQL' tab. The main area shows a Python notebook cell with code for creating a database and a table. A tooltip box is overlaid on the code, suggesting the use of 'spark.createDataFrame()' instead of 'createTable()'. The tooltip also provides a link to documentation.

```
1 # Import necessary libraries
2 from pyspark.sql import SparkSession
3
4 # Create Spark session
5 spark = SparkSession.builder.appName("CSV to DataFrame").getOrCreate()
6
7 # Read CSV file into DataFrame
8 df = spark.read.option("header", "true").load("path/to/csv")
9
10 # Show DataFrame
11 df.show()
12
13 # Stop Spark session
14 spark.stop()
```

AI 지원으로 ETL 코드 추천

A screenshot of the Databricks Notebook interface. A tooltip box from the 'Databricks Assistant' is shown, providing suggestions for fixing an error related to reading a CSV file with pandas. It suggests changing the file path and fixing an intentional error about setting a non-existent column as index.

```
# Read the CSV file into a DataFrame
file_path = "/databricks-datasets/RDatasets/data-edb/cv/ggplot2/diamonds.csv"
# Intentional Error: Setting a non-existent column as index, which could be a more subtle.
data = pd.read_csv(file_path, index_col="non_existent_column")
```

AI 기반으로 디버깅 및 문제 해결

코드, 지능형 디버깅, 수정을 추천하는 AI 기반 지식 엔진으로 ETL 파이프라인 개발

# 데이터 사이언스 & AI

## 학습 목표

기업이 머신러닝과 AI를 활용하려고 할 때 직면하는 과제 설명  
Databricks DI 플랫폼이 데이터 사이언스와 AI를 지원하는 방법



오늘날 실무환경에서 GenAI는 어렵고 비용이 많이 듭니다



데이터 또는 모델에 대해 제어  
할 수 없음



GenAI를 프로덕션에 도입하기  
어려움



규모 확장시 비용 증가

데이터 유출 우려

제어권과 소유권 부족

예측할 수 없는 성능

자동화 및 확장 필요

대규모 기초 모델은 고비용

LLM 구축 비용이 높음

# 오늘날 실무환경에서 GenAI는 어렵고 비용이 많이 듭니다



엔터프라이즈 데이터로  
안전하게 훈련된 GenAI  
모델 소유



GenAI 모델을 프로덕션으로  
빠르게 이동



LLM 학습 및 배포는 규모에  
따라 비용 효율적입니다.

+  
데이터 및 모델 소유  
향상된 개인 정보 보호

+  
기본 제공 모델 모니터링  
사용 사례 전반에 걸쳐 프로덕션으로  
확장할 수 있는 표준화된 운영

+  
자체 LLM을 구축하고 배포하는 데 비용  
효율적입니다.

## Databricks에서 더 나은 GenAI 솔루션 구축

좋은 모델 그 이상이 필요합니다

완벽한 제어



모델 및 데이터에 대한  
완전한 소유권

생산 품질



여러 사용 사례에서 더  
빠르고 안정적인 배포

비용 절감



비용효율적으로 대규모  
LLM 구축

## Databricks에서 Gen AI 애플리케이션 구축

### Databricks AI

#### 생성형 AI

- 커스텀 모델
- 모델 서빙
- RAG

#### 엔드투엔드 AI

- MLOps(MLflow)
- AutoML
- 모니터링
- 거버넌스

데이터 사이언스  
& AI  
**Databricks AI**

ETL &  
실시간 분석  
**Delta Live Tables**

오키스트레이션  
**Workflows**

데이터  
웨어하우징  
**Databricks SQL**

생성형 AI를 사용하여 데이터의 의미 체계 이해

데이터 인텔리전스 엔진

#### Unity Catalog

자연어로 안전하게 인사이트 확보

#### Delta Lake

사용 패턴에 따라 자동으로 데이터 레이아웃 최적화

### 오픈 데이터 레이크

커스텀 모델 : 데이터브릭스 AI는 포괄적인 요구사항 지원 → 맞춤형 모델 개발 가능

모델 서빙 : 간소화  
원전한 AI 워크플로우 제공  
AutoML : 성능 저하 없이 최적의 모델 검색

## 생성형 AI를 위한 Databricks AI



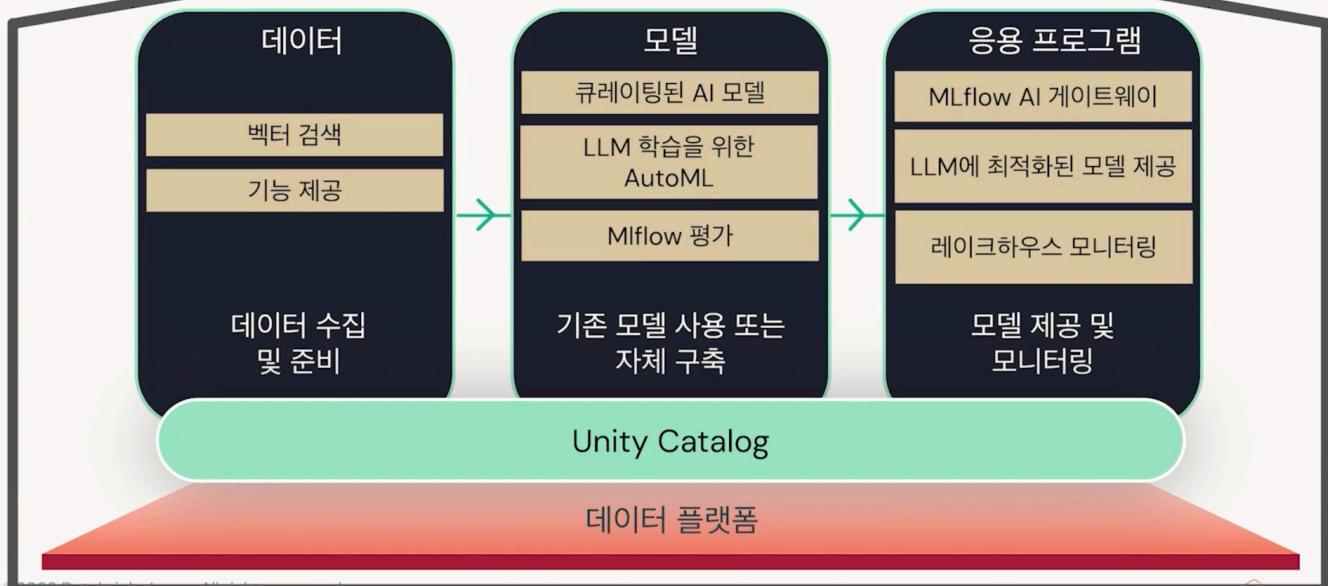
1. 데이터 및 AI 주도 성장의 토대 마련, MIT Technology Review, <https://www.databricks.com/resources/analyst-papers/laying-foundation-data-and-ai-led-growth>  
©2023 Databricks Inc. All rights reserved

2023년에 MosaicML 인수 → RAG : GenAI 모델을 미세조정 가능

## Databricks AI – 데이터 중심 AI 플랫폼

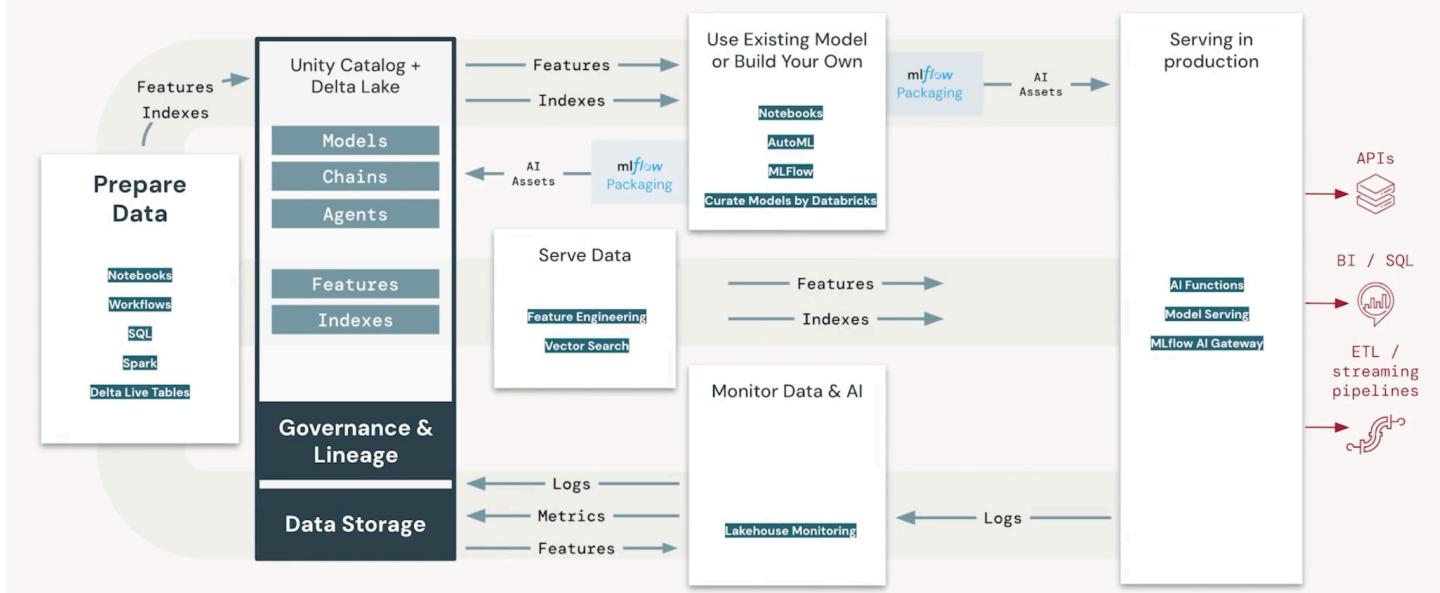


# Databricks AI – 생성형 AI에 최적화



Databricks AI 는 엔드투엔드 AI  
지원을 제공합니다.

## Databricks AI 의 기능



# Databricks AI는 모든 AI 모델에서 작동합니다.

클래식, 딥, 독점 또는 오픈 소스 생성형 AI + LLM

딥 러닝 모델



TensorFlow



기존 ML  
알고리즘



XGBoost

독점 LLM



ANTHROPIC



오픈소스  
생성형 AI +  
LLM



stability.ai

체인 &  
에이전트



haystack  
by deepset



사용 사례에 가장 적합한 모델 선택

