IBM watsonx.data

기업 혁신을 위한 차세대 데이터 저장소

—
Data&AI | 김용민 CSM
kimym@kr.ibm.com
한국 IBM

Agenda

- Market trends
- Market growth
- Product overview
- Feature highlights
- Key components
- Demo

Introducing...

watsonx

What IBM offers

The platform for AI and data

watsonx

신뢰할 수 있는 데이터로 AI의 효과를 확장하고 가속화합니다.

watsonx.ai

Train, validate, tune and deploy AI models

watsonx.ai는 AI를 구축하려는 기업이 기존 머신 러닝과 함께 foundation model로 구동되는 새로운 생성형 AI 기능을 이용하여 AI 모델을 훈련, 검증, 튜닝 및 배포할 수 있는 차세대 엔터프라이즈 스튜디오입니다.

이를 통해 적은 양의 데이터로 짧은 시간 내에 AI 애플리케이션을 구축할 수 있습니다.

watsonx.data

Scale AI workloads, for all your data, anywhere

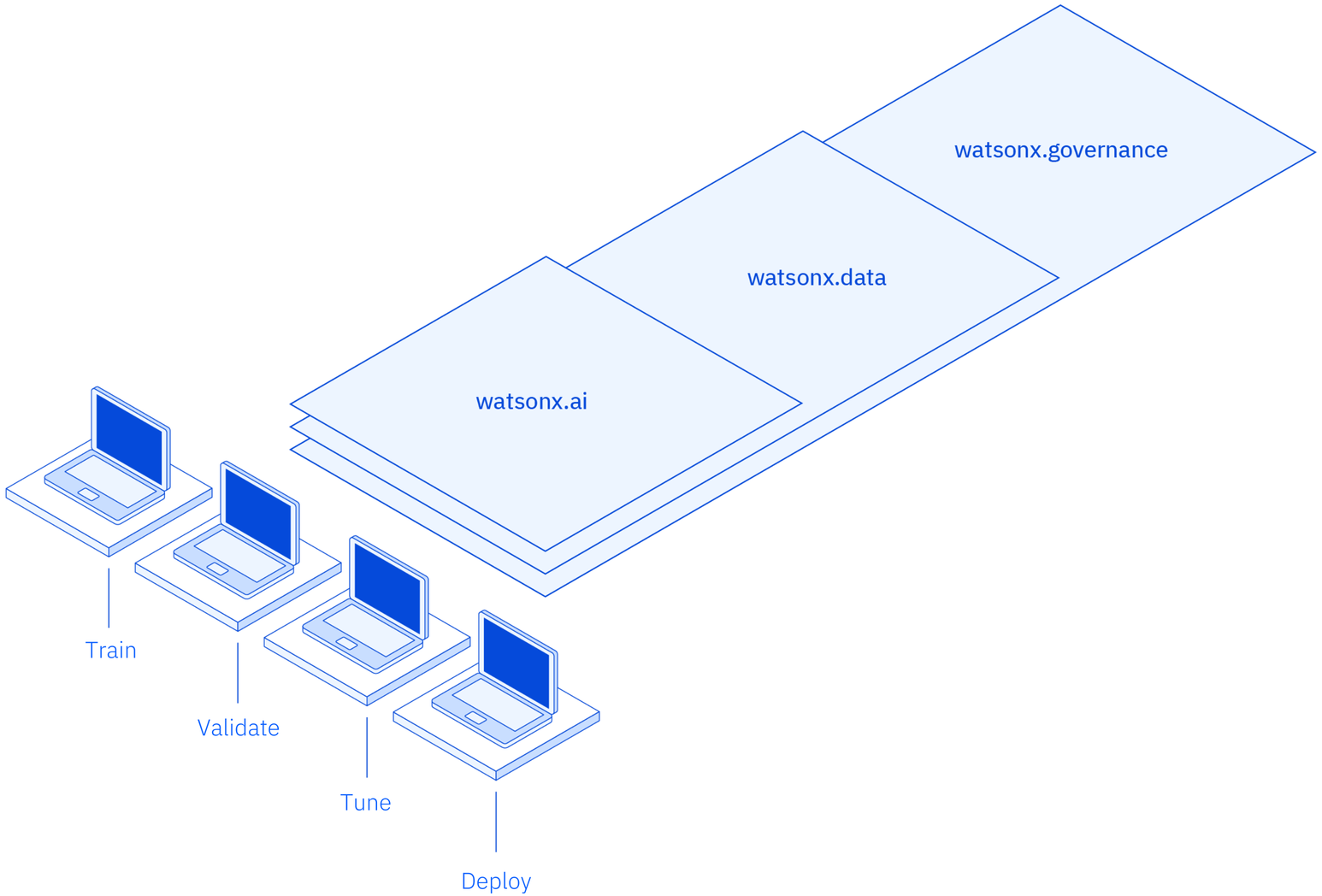
watsonx.data는 개방형 레이크하우스 아키텍처를 기반으로 구축된 목적 지향형 데이터 저장소로, Query, Governance 및 open data format을 통해 보다 쉽게 데이터에 액세스하고 공유할 수 있습니다.

watsonx.governance

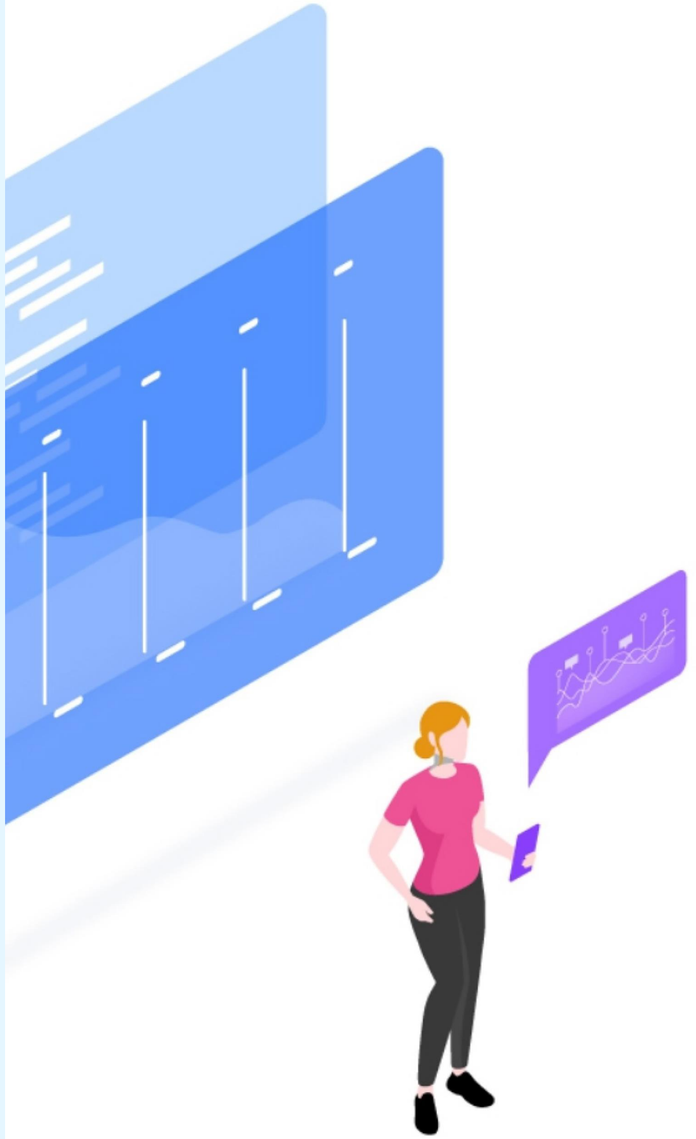
Enable responsible, transparent and explainable AI workflows

watsonx.governance는 데이터와 AI 거버넌스를 모두 포괄하는 엔드 투 엔드 툴킷으로 책임감 있고 투명하며 설명 가능한 AI 워크플로를 지원합니다.

Scale and accelerate the impact of AI with trusted data



Market trends

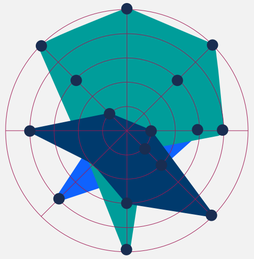


오늘날 기업은 여러 데이터 문제에 직면해 있습니다

더 많은 데이터

폭발적인 데이터 증가

저장된 데이터의 총 볼륨은 향후 5년 동안 250% 이상 증가할 것으로 예상됩니다



다양한 위치

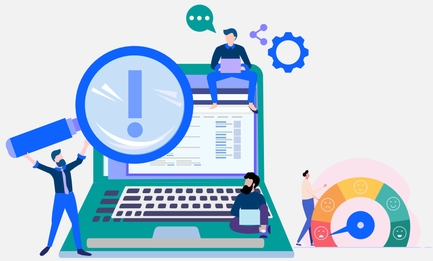
여러 위치, 클라우드, 애플리케이션에 위치한 데이터 사일로 문제

기업의 82%가 데이터 사일로로 인해 접근에 방해를 받고 있습니다.

복잡한 형태

문서, 이미지, 비디오

데이터 정리, 통합 및 준비에 소요되는 시간의 80%를 차지합니다.

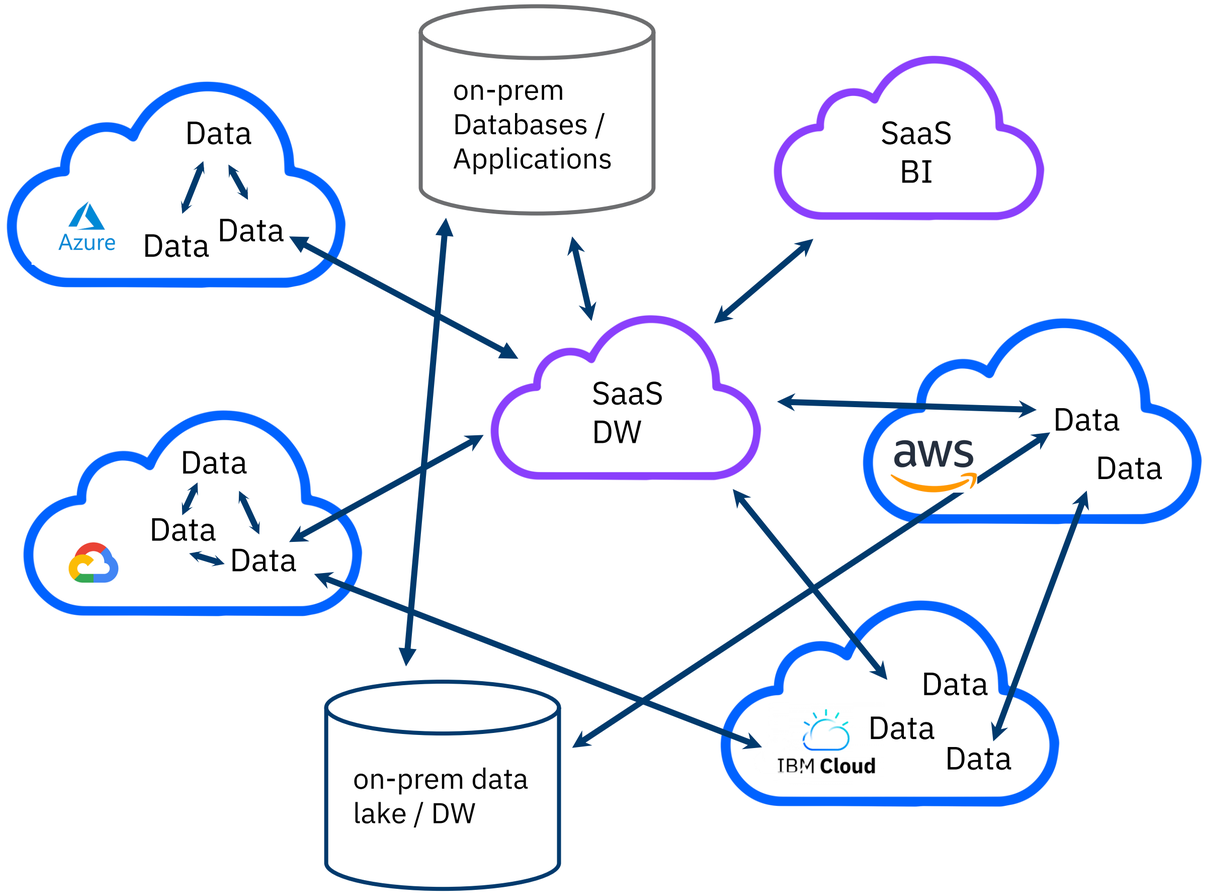


데이터 품질

오래되고 일관성이 없음

기업의 82%는 데이터 품질이 데이터 통합 프로젝트의 장벽이라고 말합니다.

데이터 관리 시장 동인



Hybrid-cloud data ecosystem

복원력(resiliency)을 개선하고 직간접 데이터 관리 비용을 줄이기 위한 인프라 현대화

- 하이브리드 클라우드 배포에서 데이터베이스, 앱 및 ETL에 대한 호환성
- 다운타임 없는 데이터베이스 마이그레이션 툴링 및 자동화
- 최저 TCO(No DBA)를 위한 완전 관리형 서비스 및 소비 기반 가격 책정

새로운 인사이트를 도출하여 비즈니스 성과를 개선하는 Digital Transformation

- 데이터의 저장 위치에 상관없이 손쉬운 정형 및 비정형 데이터 결합
- 다양한 페르소나 및 특정 워크로드에 대해 선택된 분석 도구 활용
- Zero-copy(ETL 없음)로 신뢰할 수 있는 단일 소스 및 가치 창출 시간 단축

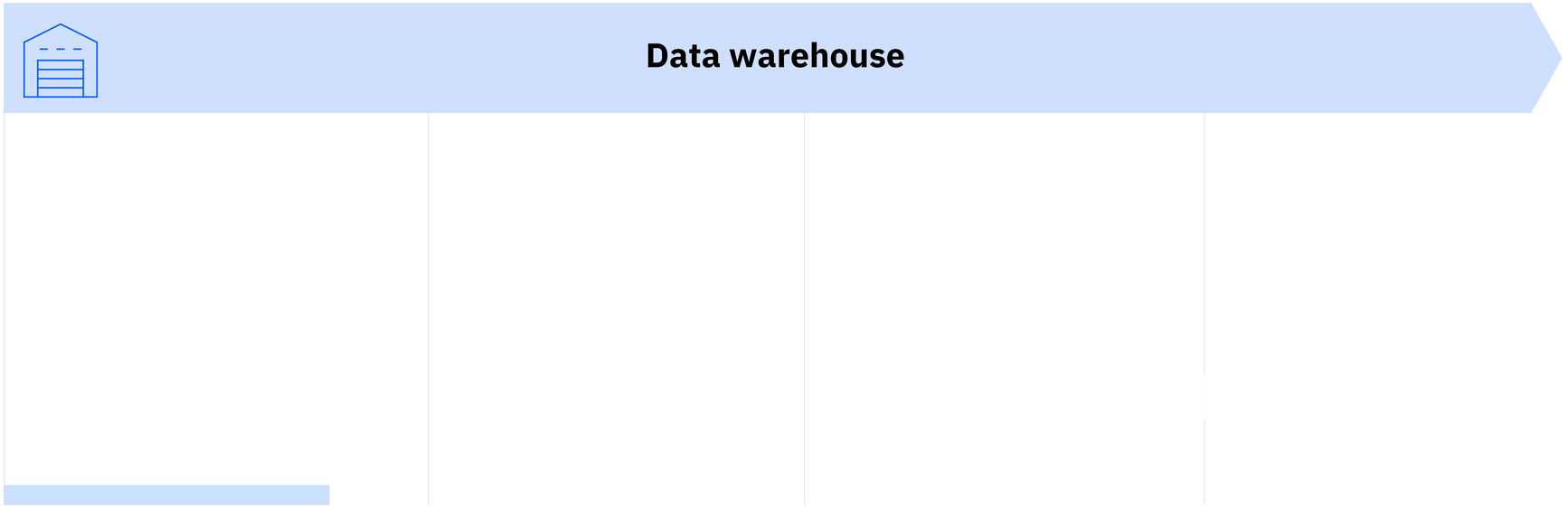
민첩성을 개선하고 손해 배상 및 이미지에 대한 손상을 방지하기 위한 규정 준수 및 보안

- 데이터 검색/카탈로그, 품질, 계보, 시맨틱 및 규정 준수/프라이버시를 위한 중앙 집중식 또는 통합 메타데이터 관리
- 모든 리포지토리 및 쿼리 엔진에서 관리되고 안전한 데이터 액세스 제어 지점을 통해 엔터프라이즈 전체에서 데이터 액세스를 손쉽게 함

Market growth

데이터 웨어하우스는 대부분의 조직에서 분석작업 중심으로 유지되고 있음

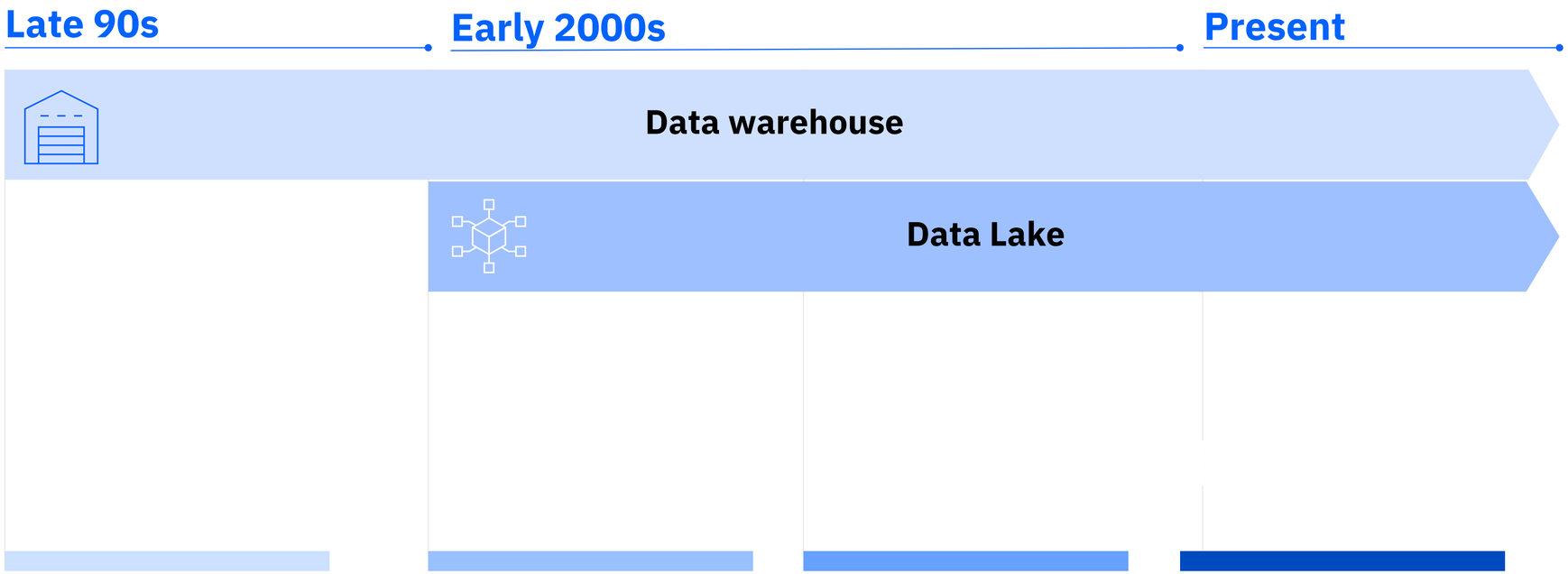
Late 90s | **Early 2000s** | **Present**



높은 초기 비용
구조화된 데이터만 가능
ETL 필요
공급업체 종속
제한된 확장성

- 데이터 웨어하우스는 데이터를 분석하는 지배적인 방법으로 부상
- 정규화되고 신뢰할 수 있는 데이터를 통해 쉽게 분석할 수 있었지만 비용이 많이 드는 선택
- 데이터 웨어하우스 기술은 Appliance 제품 폼 팩터에서 인메모리 기술에 이르기까지 지속적으로 개선

데이터 레이크의 출현

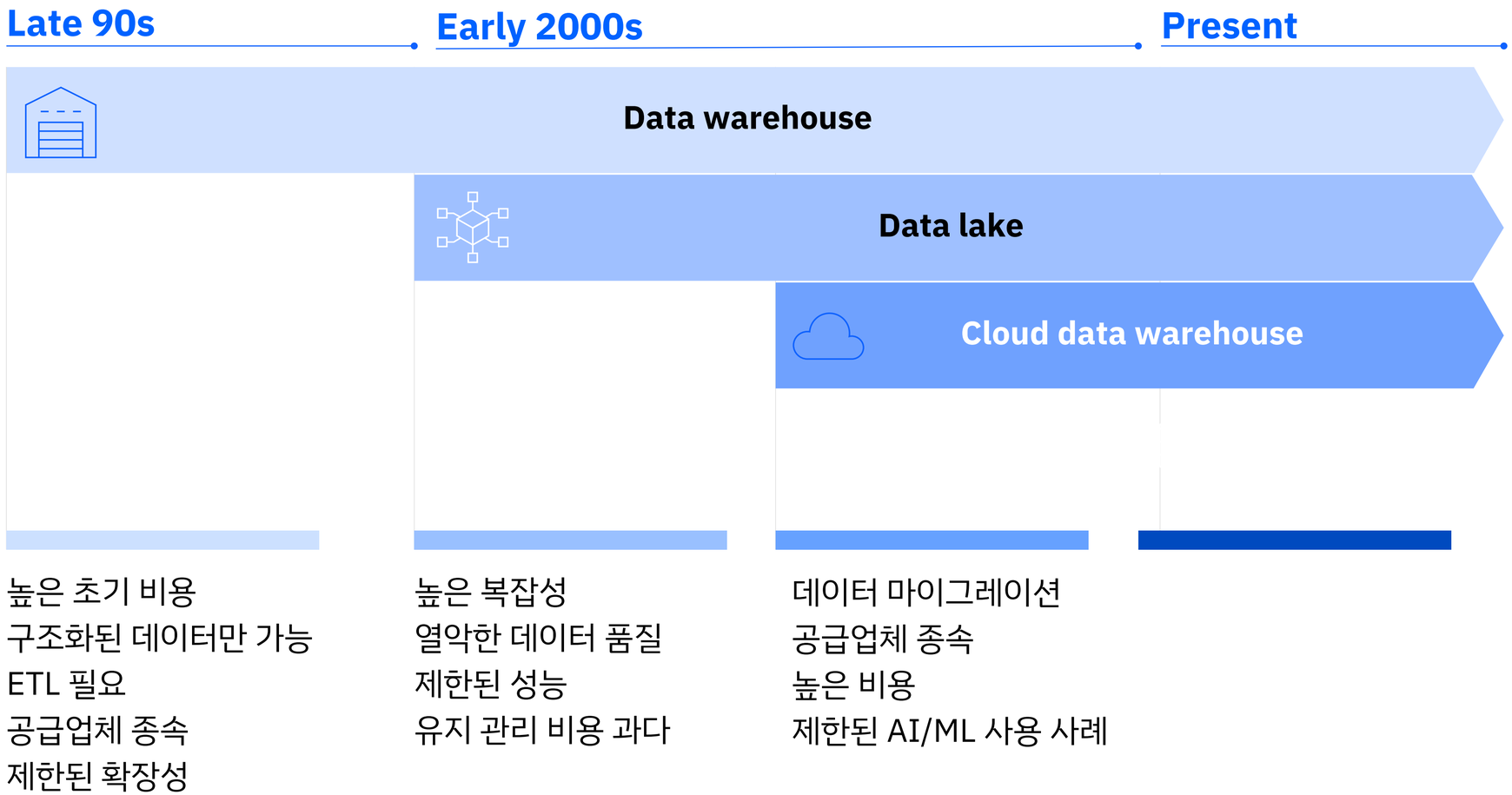


높은 초기 비용
구조화된 데이터만 가능
ETL 필요
공급업체 종속
제한된 확장성

높은 복잡성
열악한 데이터 품질
제한된 성능
유지 관리 비용 과다

- 데이터의 양, 속도 및 다양성이 증가함에 따라 데이터 레이크는 데이터 웨어하우스를 대체할 새로운 기술로 부상
- RAW 및 비정형 형식으로 저장된 데이터 = 대량의 데이터에 대한 비용 절감
- 뛰어난 유연성과 확장성
- 사용하기 어렵고 유지 관리가 복잡하며 데이터 과학자가 필요
- 결국 대부분의 데이터 레이크는 실패했고 2-tier 아키텍처가 필요

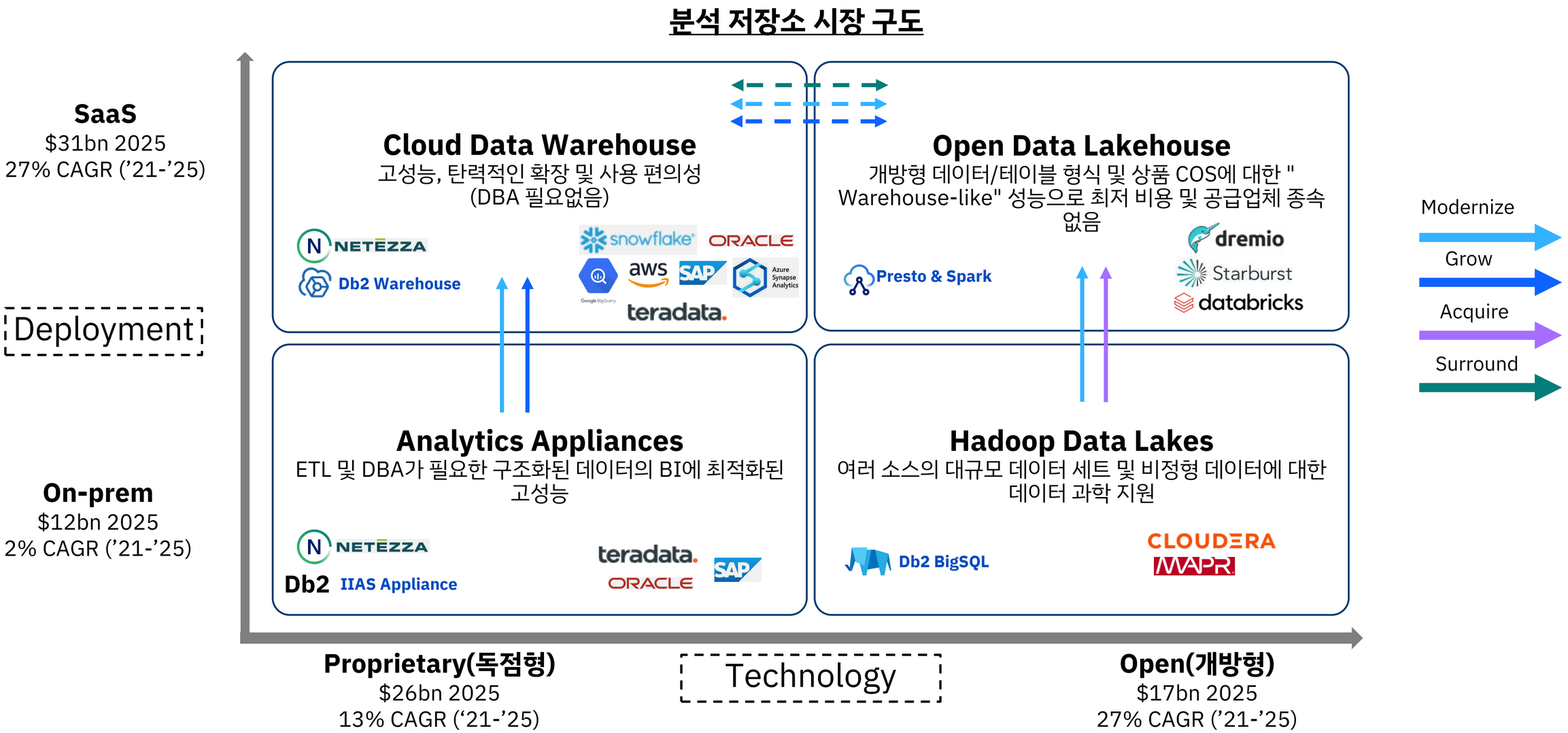
클라우드 데이터 웨어하우스는 데이터 웨어하우스의 문제를 해결하기 위해 등장



- 특히 클라우드 데이터 웨어하우스는 컴퓨팅과 스토리지를 분리 도입
- 기존 웨어하우스의 확장성 문제 해결 – 데이터 재배포 없음
- 문제를 해결하기 위해 동일한 데이터에 더 많은 컴퓨팅 리소스를 추가하는 기능
- 그러나 관리하기가 더 쉽지만 on-prem 웨어하우스보다 고비용

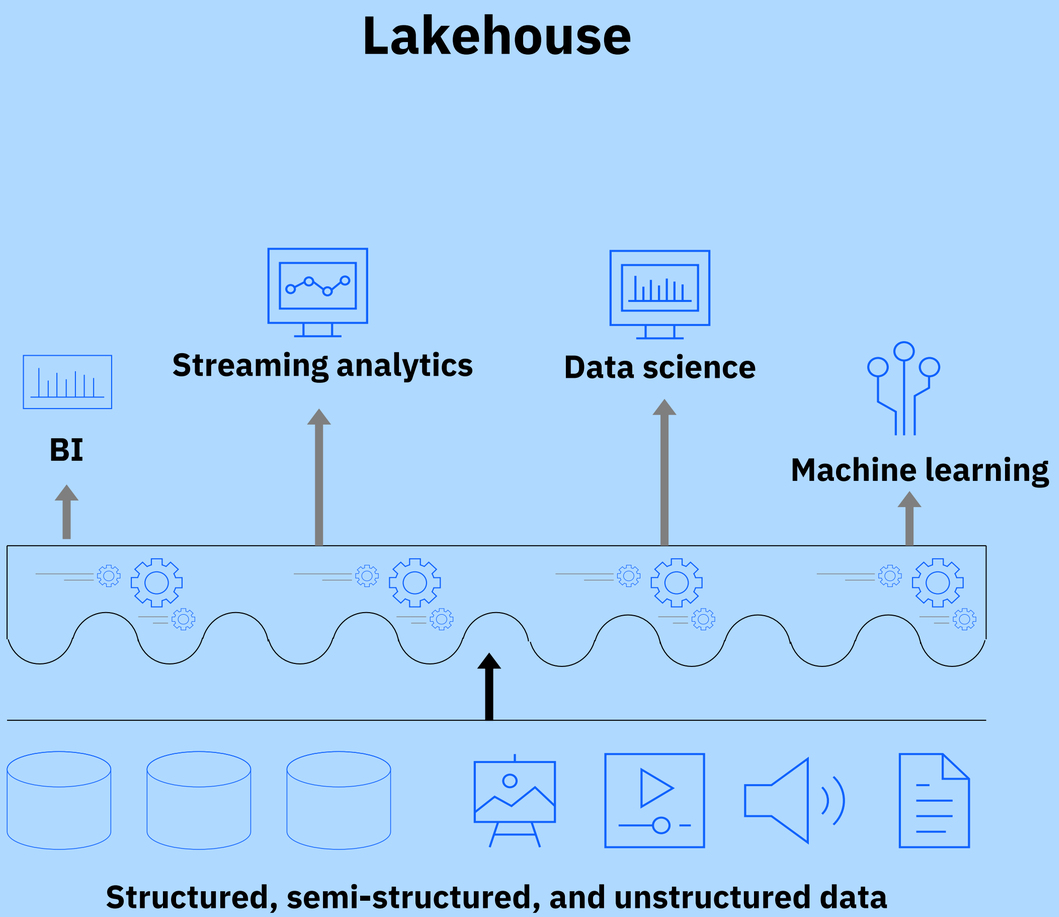
시장의 움직임

주요 혼란으로 인해 분석 저장소 시장이 on-prem에서 SaaS로, 독점에서 개방형 기술로 성장하고 있습니다

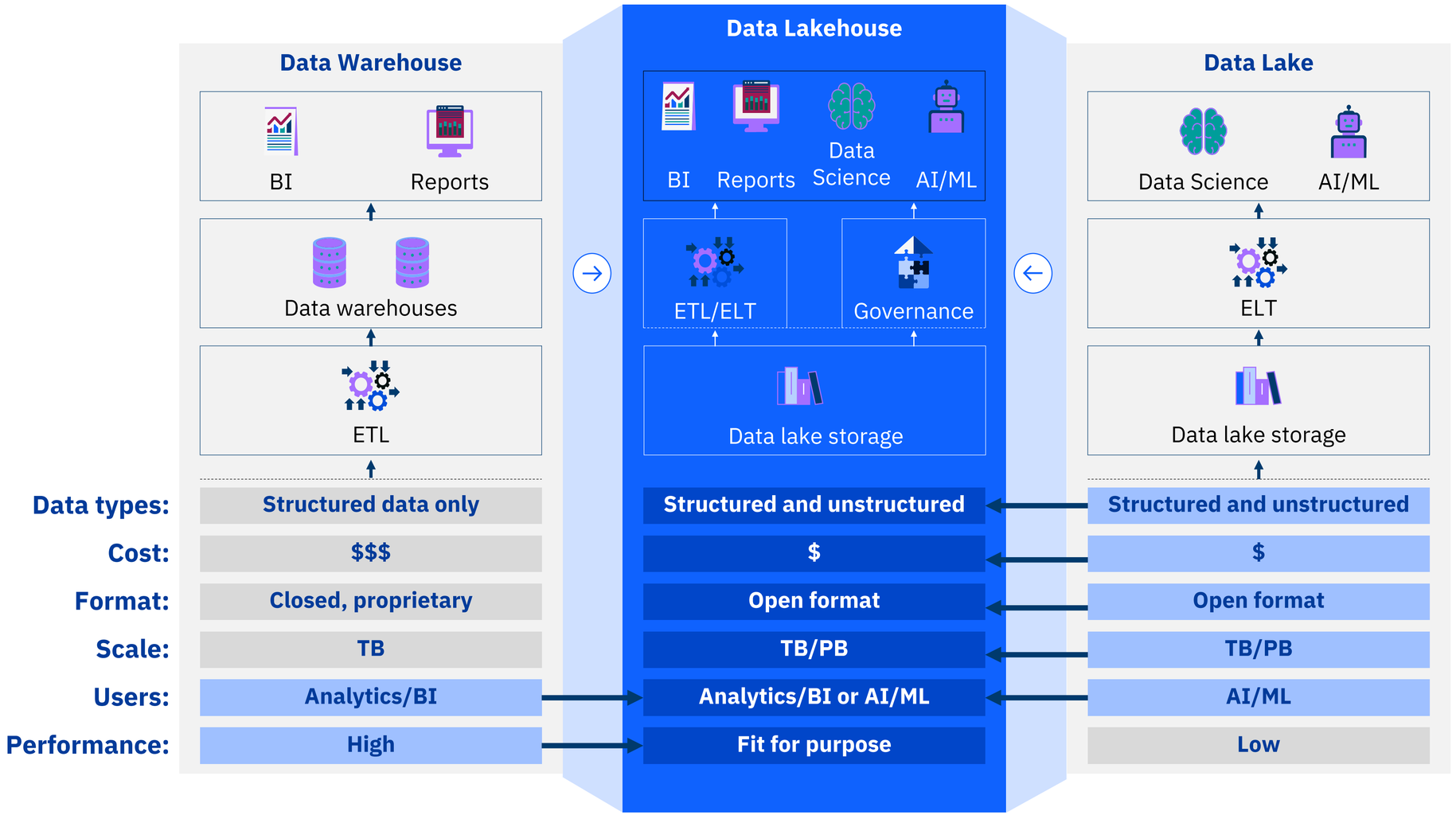


Lakehouse overview and first-generation lakehouses

레이크하우스는 데이터 레이크와 웨어하우스를 최적의 **단일 통합 플랫폼**으로 결합하여, 매우 복잡한 데이터 변환과 **다양한 사용 사례**를 지원합니다



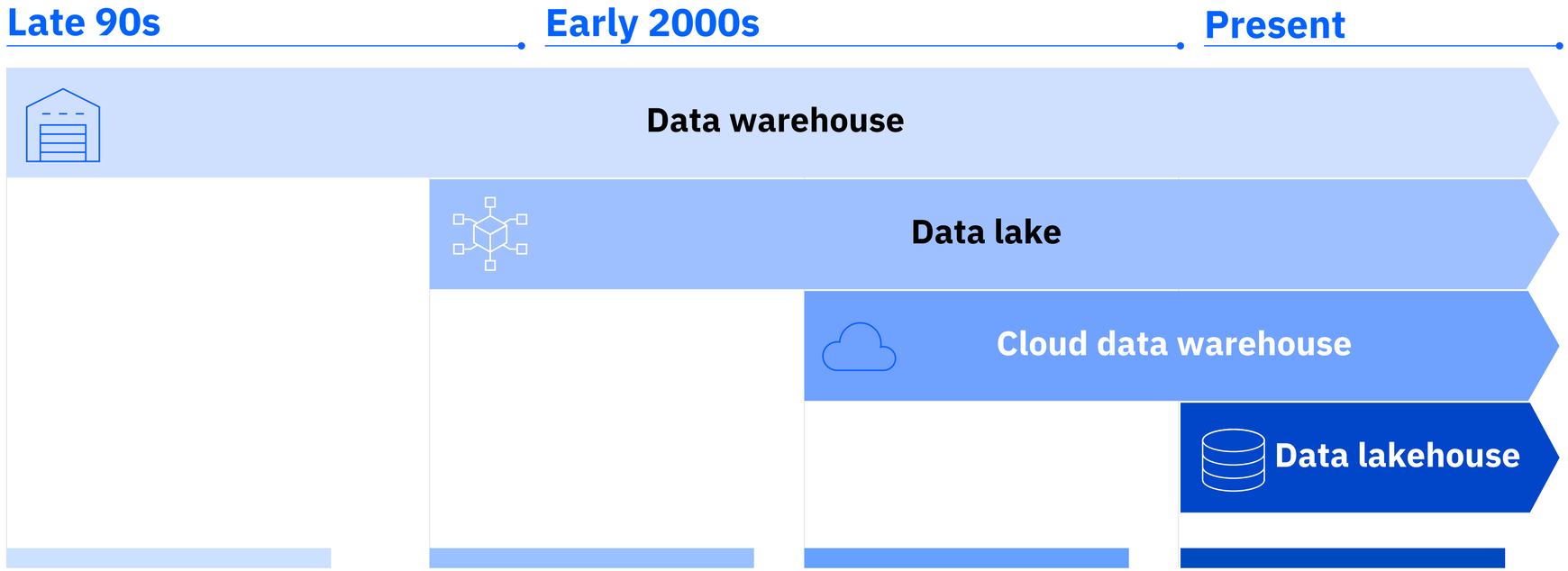
레이크하우스는 데이터 웨어하우스와 데이터 레이크의 장점을 결합한 새로운 차원의 데이터 저장소입니다



1세대 레이크하우스에는 여전히 비용 및 복잡성 문제의 해결을 제한하는 주요 제약 조건이 존재

1. 제한된 워크로드를 지원하도록 설정된 단일 쿼리 엔진(일반적으로 BI 또는 ML만 해당)
2. 일반적으로 멀티/하이브리드 클라우드 배포를 지원하지 않고 단일 클라우드를 통해서만 배포
3. 전체 에코시스템에 배포할 수 있는 최소한의 거버넌스 및 메타데이터 기능만 제공

데이터 레이크하우스는 분석의 새로운 시대를 열기 위한 신기술입니다



높은 초기 비용
구조화된 데이터만 가능
ETL 필요
공급업체 종속
제한된 확장성

높은 복잡성
열악한 데이터 품질
제한된 성능
유지 관리 비용 과다

데이터 마이그레이션
공급업체 종속
높은 비용
제한된 AI/ML 사용 사례

기술 curve의 초기 단계
BI와 AI를 한 곳에서
DW보다 저렴한 비용
개방성과 유연성

- 오늘날 대부분의 기업은 데이터 레이크와 여러 웨어하우스를 가지는 2-tier 아키텍처를 필요로 합니다
- 데이터는 레이크에서 웨어하우스로 이동 및 복제되며, 웨어하우스는 여전히 주요 데이터에 대한 액세스 계층입니다
- 데이터 레이크하우스는 웨어하우스와 데이터 레이크의 장점을 결합합니다.
- 데이터 웨어하우스 **엔진(성능)** + 데이터 레이크 **스토리지(비용)**

IBM watsonx.data feature highlights

IBM watsonx.data는 현재 1세대 레이크하우스에서 진화한 차세대 기술입니다.

	First generation lakehouse	IBM watsonx.data
	dremio databricks Starburst	
Deployment:	**Cloud or software** deployment **only**	**Hybrid and multi-cloud** deployment
Query engines:	**Single engine** with limited use case	**Multiple** fit-for-purpose **engines**
Governance:	**Separate** isolated functionality	Integrated and **enterprise** ready

IBM watsonx.data**는 Multi 쿼리 엔진**을 갖춘 유일한 레이크하우스로, 고객이 **워크로드를 최적의 엔진과 매칭하여 비용과 성능을 최적화**

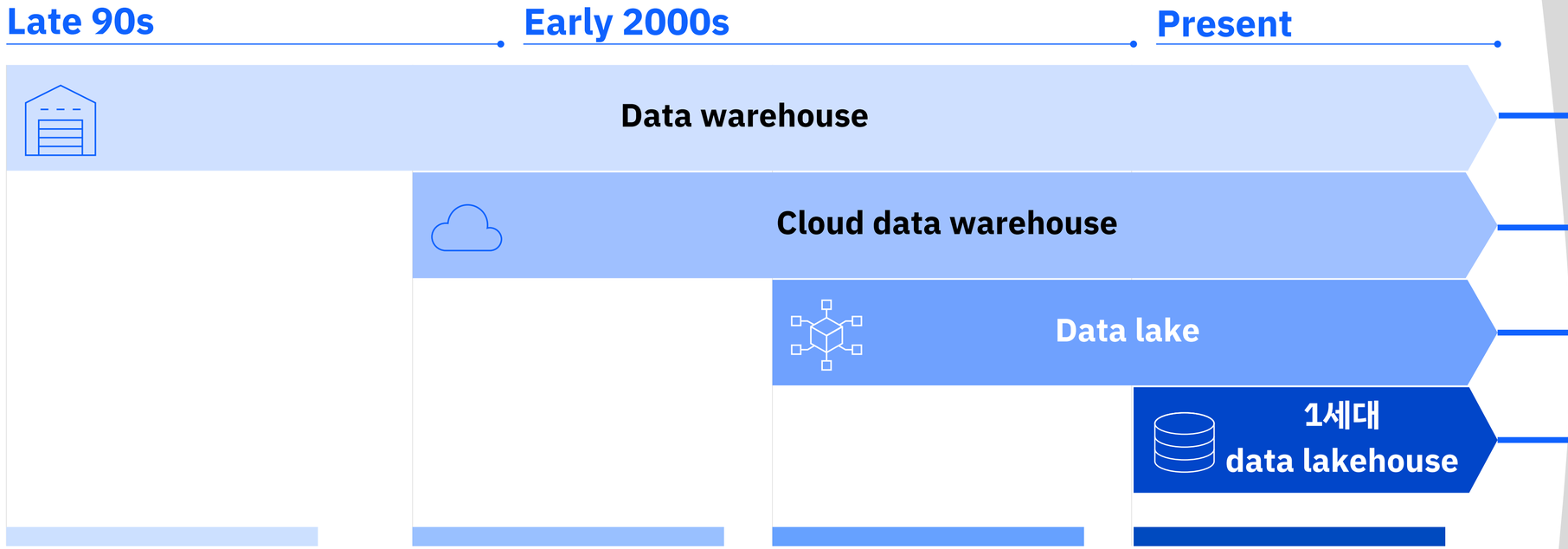
단일 창에서 모든 워크로드를 실행하여 비용과 성능을 향상시키면서 **편리함**까지 향상

하이브리드 클라우드 및 멀티클라우드 환경에 대한 완벽한 지원으로 어디에나 배포

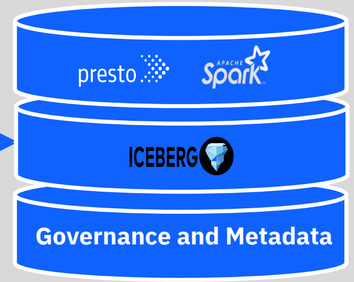
여러 엔진에서 메타데이터를 공유하여, 카탈로그를 다시 작성할 필요가 없이 가치 창출 **시간을 단축**하는 동시에 거버넌스를 보장하고, **비용**이 많이 드는 구현 노력을 **절약**

IBM watsonx.data는 기업들이 직면하고 있는 비용과 복잡성을 극복할 수 있도록 설계되었습니다

하이브리드 클라우드에서 관리되는 데이터 및 AI 워크로드에 최적화된 유일한 개방형 데이터 저장소



IBM watsonx.data



높은 초기 비용
구조화된 데이터만
ETL 필요
공급업체 종속
제한된 확장성

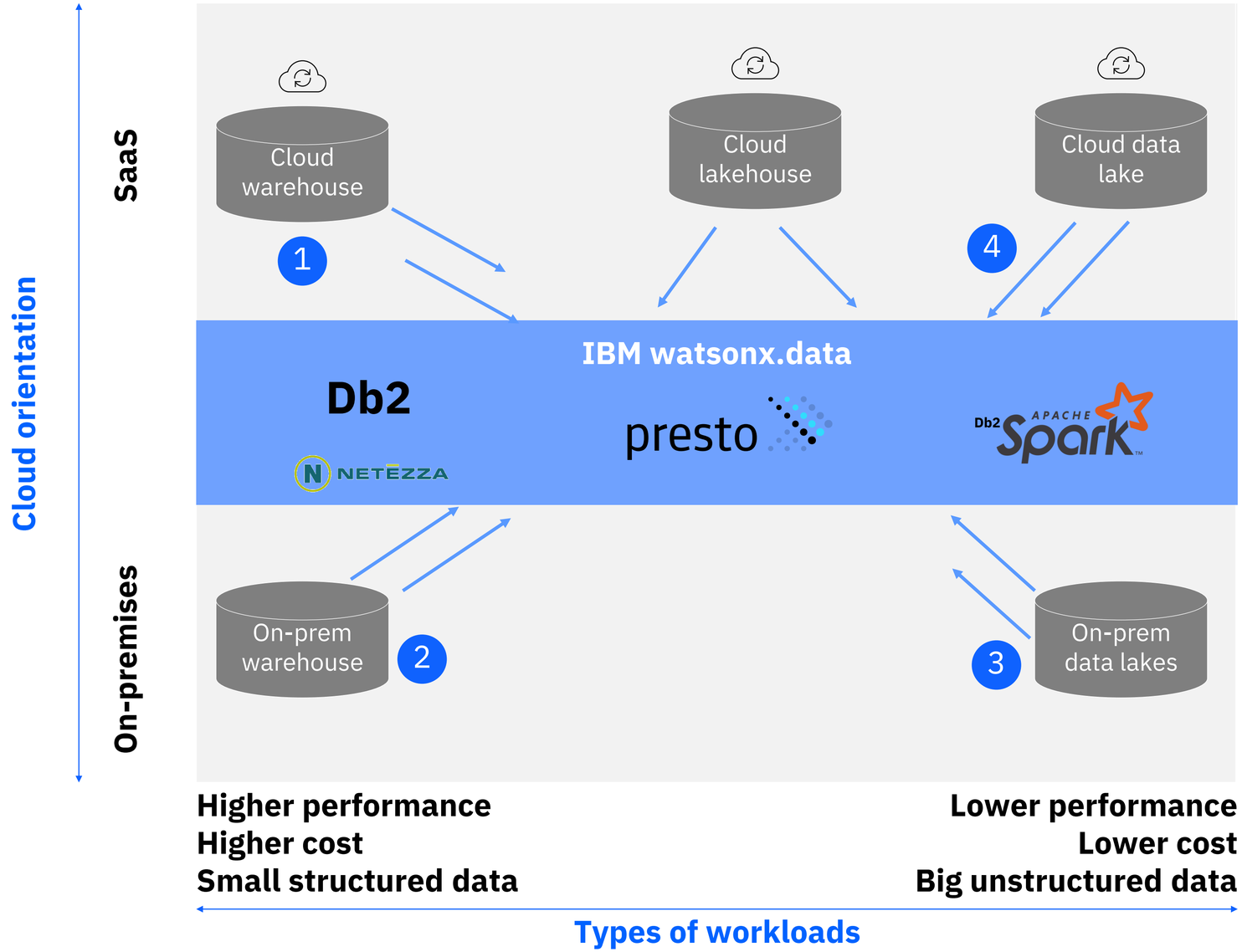
높은 복잡성
열악한 데이터 품질
제한된 성능
유지 관리 비용 과다

데이터 마이그레이션
공급업체 종속
높은 비용
제한된 AI/ML 사용 사례

제한된 사용 사례(예: BI 전용 또는 AI/ML 전용)

비용을 50% 절감하고,
데이터에 100% 액세스하고,
10분 안에 시작 가능

IBM watsonx.data를 사용하면 클라이언트 데이터에 100% 액세스할 수 있으며 전체 에코시스템에서 워크로드를 최적화



1 **비용이 많이 드는 클라우드 웨어하우스 최적화**

Snowflake(및 유사한) 워크로드를 최적화하여 목적에 맞는 쿼리 엔진과 컴퓨팅 리소스(예: 캐시 대 컴퓨팅 최적화)를 활용하여 비용을 절감

2 **On-prem 웨어하우스 최적화 및 액세스**

저비용 오브젝트 스토리지 및 목적에 맞는 엔진을 사용하여 온-프레미스 워크로드 최적화

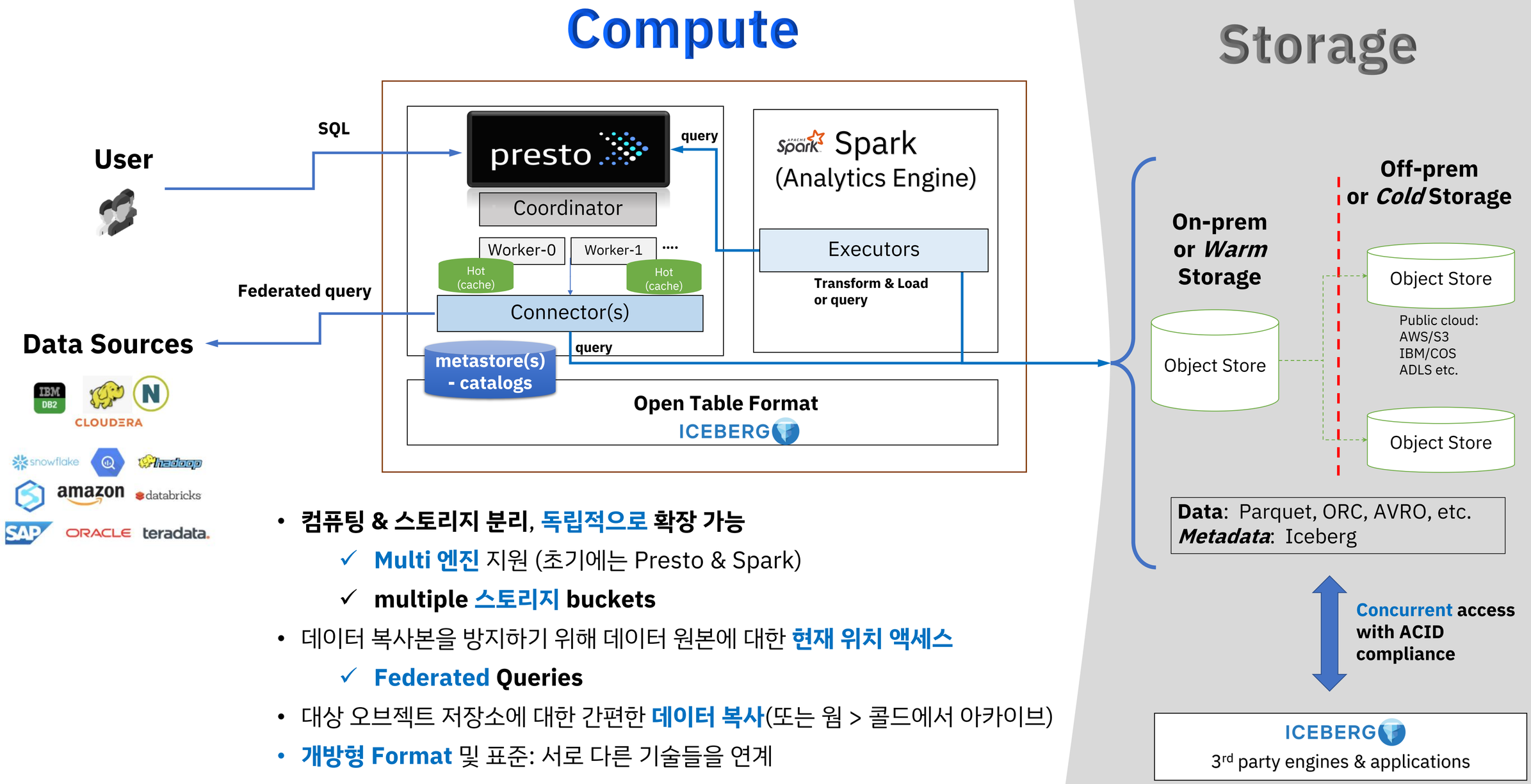
3 4 **데이터 레이크 현대화**

기존 기록 보고를 실행하고 Hadoop의 비용과 복잡성 없이 새로운 AI 워크로드를 지원

1 2 3 4 **하이브리드 클라우드와 멀티 클라우드에 배포**

10분안에 퍼블릭 클라우드와 고객의 기존 온프레미스 투자에 원활하게 배포

현대적인 레이크하우스 아키텍처



- **컴퓨팅 & 스토리지 분리, 독립적으로 확장 가능**
  - ✓ **Multi 엔진** 지원 (초기에는 Presto & Spark)
  - ✓ **multiple 스토리지 buckets**
- 데이터 복사본을 방지하기 위해 데이터 원본에 대한 **현재 위치 액세스**
  - ✓ **Federated Queries**
- 대상 오브젝트 저장소에 대한 간편한 **데이터 복사**(또는 웜 > 콜드에서 아카이브)
- **개방형 Format** 및 표준: 서로 다른 기술들을 연계

Apache Presto

Presto는 Facebook에서 개발한 빠르고 안정적이며 효율적인 오픈 소스 SQL 쿼리 엔진

Presto를 사용하면 대용량 데이터에 대해 1초 미만의 성능으로 대화형/Adhoc 쿼리를 실행

1. 대용량 데이터 처리

a.대용량 데이터 집합을 신속하게 처리

b.대규모 클러스터에서 병렬 처리 및 실시간 또는 배치 분석

2. 분산 쿼리 엔진

a.분산 환경에서 효율적인 데이터 처리

b.병렬 처리와 분산 컴퓨팅을 통한 빠른 응답 시간 제공

3. 다양한 데이터 소스 지원

a.관계형 데이터베이스, NoSQL 데이터베이스, 클라우드 스토리지, Hadoop HDFS 등 다양한 데이터 소스에 대한 쿼리 실행

4. ANSI SQL 호환성

a. ANSI SQL 표준 준수

b. SQL 쿼리 언어를 사용하여 데이터 질의 수행

5. 공통 데이터 엑세스 레이어

a.다양한 데이터 소스에 대한 일관된 데이터 엑세스 레이어 제공

b.데이터 통합 및 분석을 위한 일관성 확보

6. 커뮤니티 지원

a.오픈 소스 프로젝트로 활발한 커뮤니티 개발 및 지원

b.다양한 사용 사례에 대한 업데이트와 개발 보장

Apache Iceberg

1. Apache Iceberg

a.Netflix에서 개발됨

b.2018년 인큐베이팅 프로젝트 시작

c.2020년 Apache 정식 오픈소스 프로젝트로 인큐베이팅 졸업

2. 목표

: 페타바이트 규모의 테이블을 위한 개방형 테이블 형식 설계

Hive Table의 문제점을 해결. (ACID 지원, 병목현상 해결, 스키마 확장성, 파일관리 개선)

3. 기능 및 설계 원칙

a.파일 포맷 관리: 다양한 파일 포맷을 관리, 구성 및 추적

b.Snapshot 방식: 파일 관리와 버전 관리를 스냅샷(snapshot)을 통해 수행

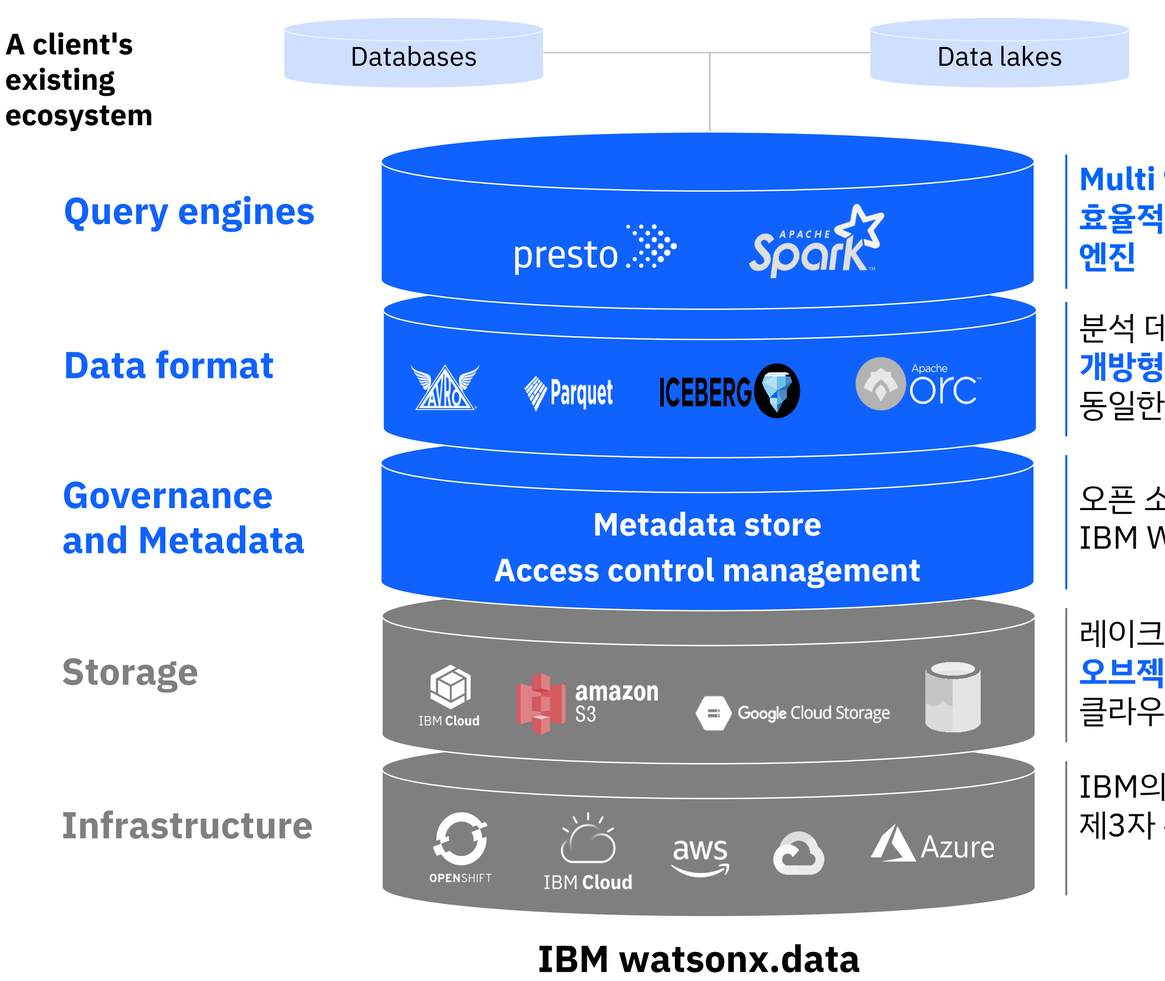
4. 활용 분야

a.대규모 데이터 테이블 관리

b.데이터 레이크 및 데이터 웨어하우스에서 활용

Key components

IBM watsonx.data의 주요 구성요소 Overview
: Multi 엔진, 개방형 테이블 형식 및 기본 제공 엔터프라이즈 거버넌스



Multi 엔진. Presto와 Spark는 **빠르고 안정적이며 효율적인** 대규모 처리를 제공하는 **오픈 소스 분산 엔진**

분석 데이터 세트에 대한 **공급업체에 종속되지 않는 개방형 format**을 통해 서로 다른 엔진이 동시에 동일한 데이터에 액세스하고 공유 가능

오픈 소스 솔루션에서 개발된 **Built-in 거버넌스.** IBM WKC를 포함한 기존 솔루션과도 호환 가능

레이크하우스를 위한 **비용 효율적이고 단순한 오브젝트 스토리지**. 하이브리드 클라우드 및 멀티 클라우드 환경에서 사용 가능

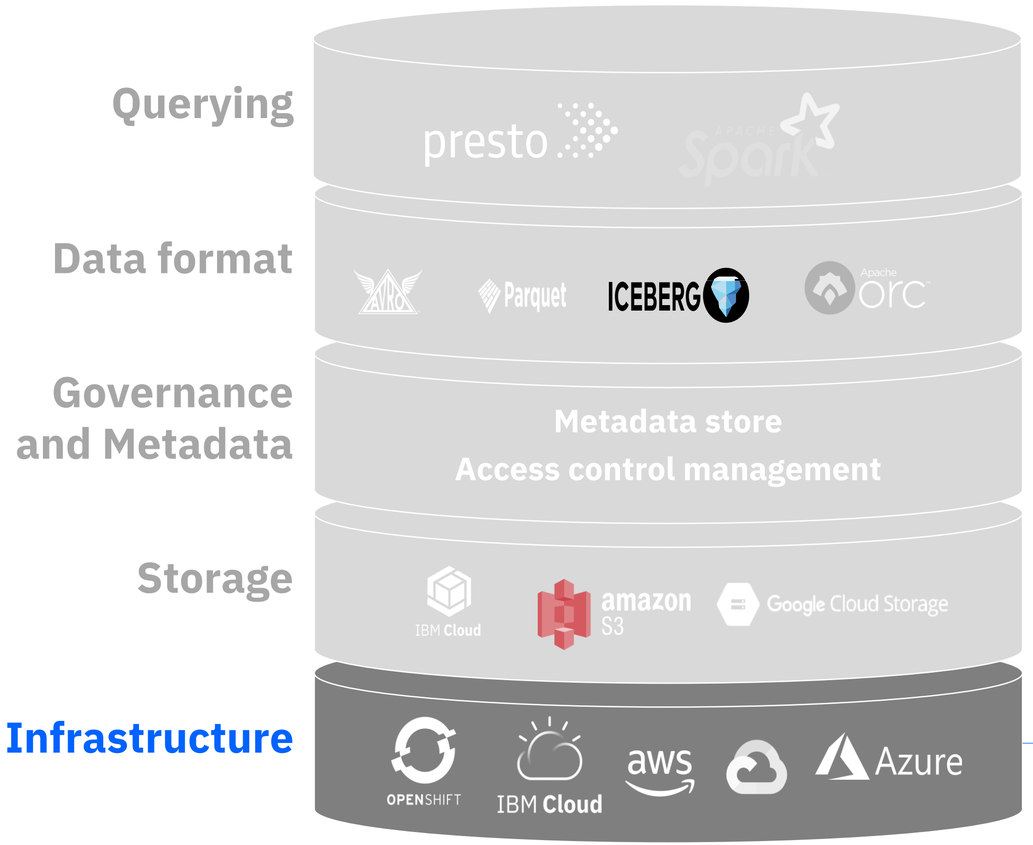
IBM의 쿠버네티스 플랫폼, OpenShift(또는 기타 제3자 쿠버네티스)와 **멀티 클라우드 통합**

Core lakehouse functionality

Ecosystem infrastructure

- **Multi 엔진 기능을 사용하여 워크로드 비용 및 성능 최적화**
- **lake, warehouse 및 다른 lakehouse의 데이터에 100% 액세스 가능**
- **거버넌스를 보장 및 단일 공유 메타데이터 및 액세스 관리를 통해 인사이트 확보 시간 단축**
- **비용이 많이 드는 관계형 warehouse 스토리지에 비해 스토리지 비용 절감**
- **필요한 것이 모두 갖추어진 어떤 인프라에도 배포 가능**

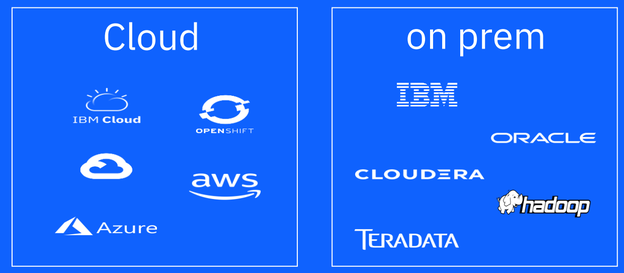
모든 클라우드 또는 On-prem 환경에서도 10분 이내에 원활하게 배포



1 10분 이내에 원활하고 완벽하게 관리되는 배포

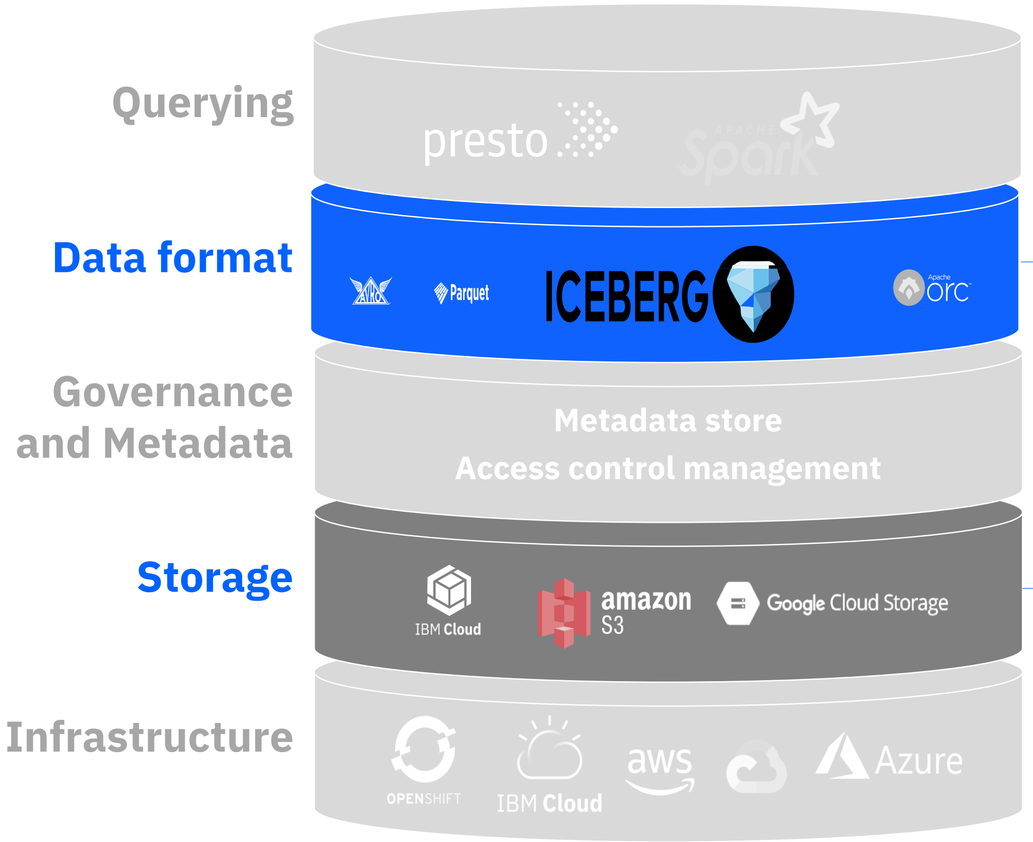
- 몇 분 안에 데이터 로드 준비
- Full 서비스 – 스토리지, 엔진 및 메타데이터 저장소가 모두 통합되고 사전 구성

2 모든 클라우드 또는 On-prem 환경에 배포



- On-prem 데이터와 클라우드 데이터를 결합하여 보다 쉽게 관리, 통제 및 분석할 수 있는 통합 데이터 View를 제공
- 민감한 데이터를 On-prem에 유지하면서 클라우드의 확장성과 비용 효율성을 활용

저비용 오브젝트 스토리지에 대량의 데이터를 저장하고, 고성능 분석을 위해 구축된 오픈 테이블 format을 통해 공유



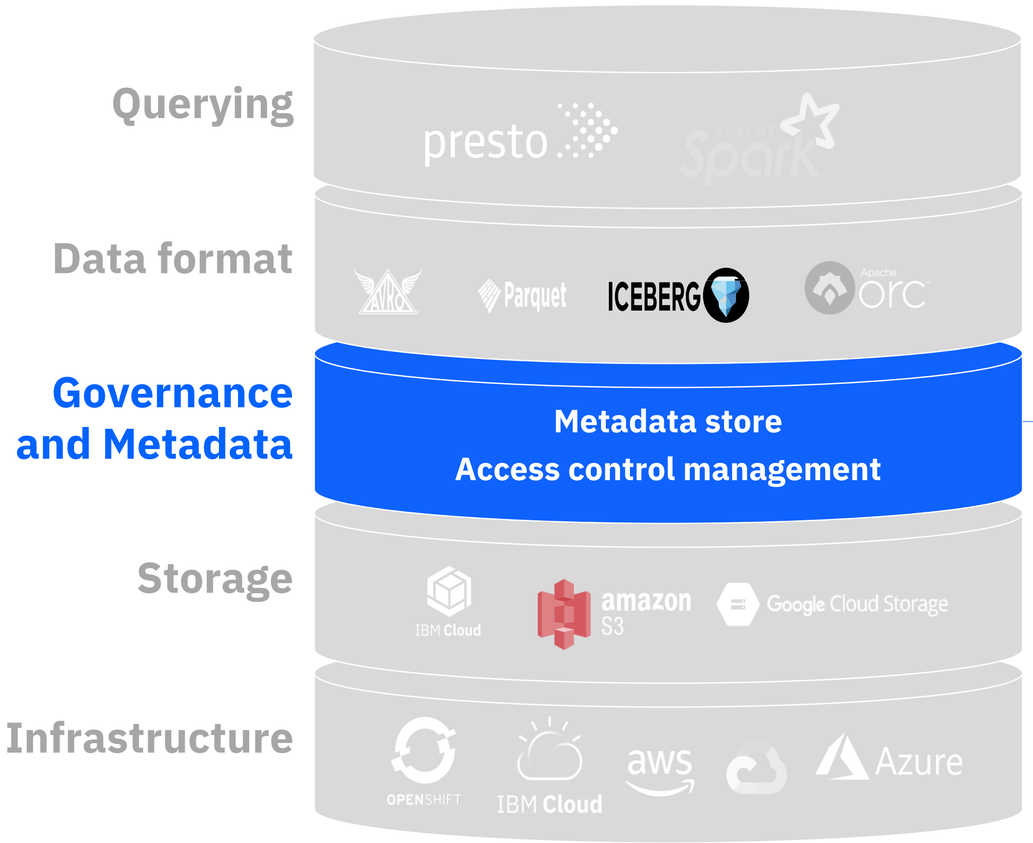
Apache Iceberg는 대규모 데이터 처리 및 분석에 탁월한 고성능 테이블 형식

- SQL의 안정성과 단순성을 빅 데이터에 제공
- 여러 엔진이 동시에 동일한 테이블에 액세스
- ACID 트랜잭션을 지원하여 데이터 일관성 보장
- 시간 이동 및 롤백 기능을 사용하면 재현 가능한 쿼리가 시간 경과에 따른 변경 사항을 검사하고 이전 상태로 빠르게 롤백 가능

간단한 오브젝트 스토리지를 활용하여 기존 블록 스토리지보다 훨씬 저렴한 비용으로 방대한 양의 데이터를 저장

- 용량 제약 없이 데이터를 저장할 수 있는 거의 무제한의 용량
- 고가용성 및 내구성 – 데이터 손실을 방지하기 위해 중복 복사본을 사용하여 가용성 도메인 간에 자동으로 복제
- 비용 효율성 – 오브젝트 스토리지는 기존 블록 스토리지 비용의 일부이며 여러 제공업체에서 사용 가능

내장된 통합 거버넌스 또는 기존 거버넌스 솔루션에 연결하여 기업 규정 준수 및 보안을 보장



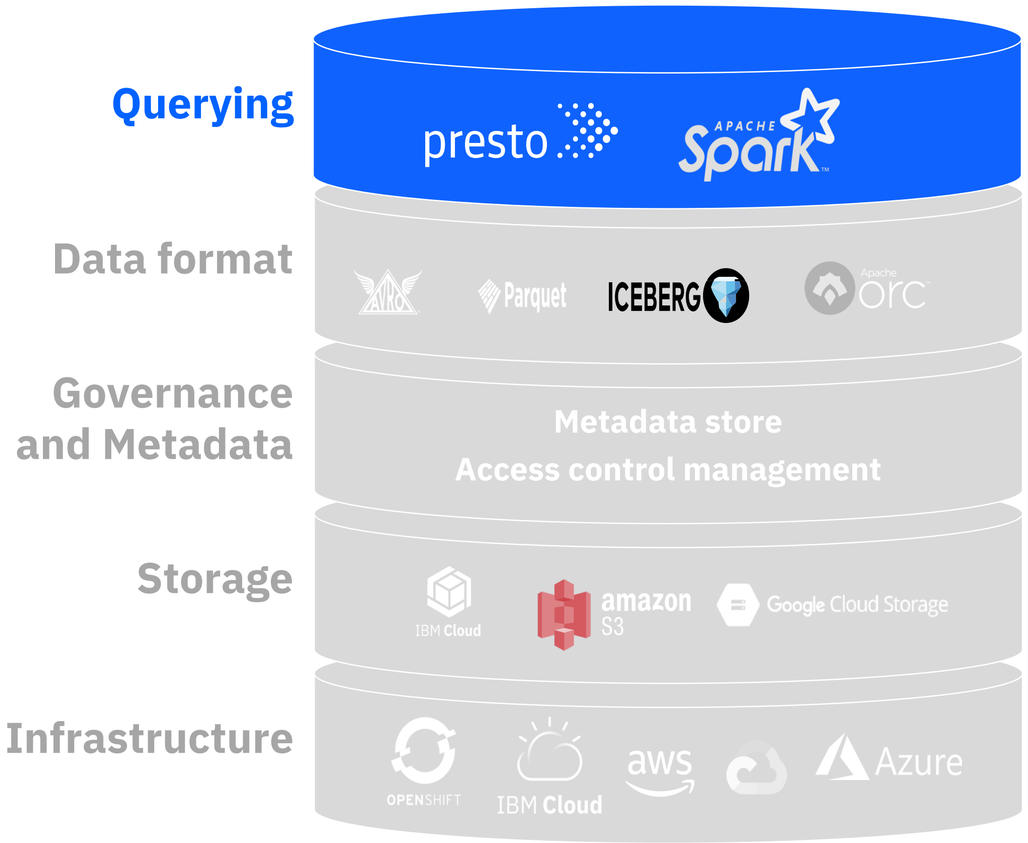
1 내장된 메타데이터 및 접근 제어 솔루션 활용

- 기본 제공 오픈 소스 기술을 활용하여 메타데이터, 계보, 액세스 및 카탈로그 작성을 포함한 거버넌스 및 보안을 관리

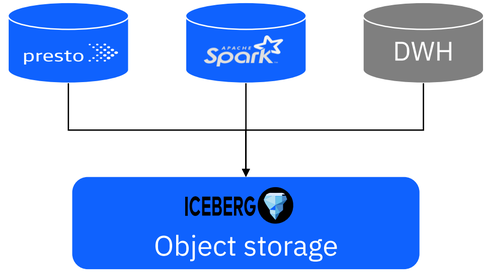
2 기존 거버넌스 솔루션에 연결

- 모든 레이크하우스 및 웨어하우스 엔진에서 단일 통합/공유 메타데이터 서비스를 활용하여 전체적인 액세스 관리 정책으로 액세스를 간소화
- IBM의 Watson Knowledge Catalog(WKC)와 직접 통합하고 메타데이터 서비스를 통해 레이크하우스에서 WKC 정책을 시행

고객의 요구에 따라 자동으로 Scale up/down 가능한 목적에 맞는 엔진을 사용하여 고비용 웨어하우스의 워크로드를 최적화



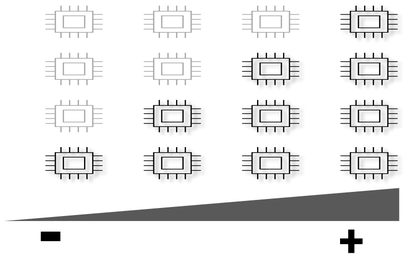
1 Iceberg 데이터 형식과 저비용 오브젝트 스토리지를 활용하여 여러 엔진에서 액세스할 수 있는 공유 메타데이터 계층을 생성



2 용도에 맞는 계산 및 쿼리 엔진을 사용하여 Lakehouse에서 워크로드 실행

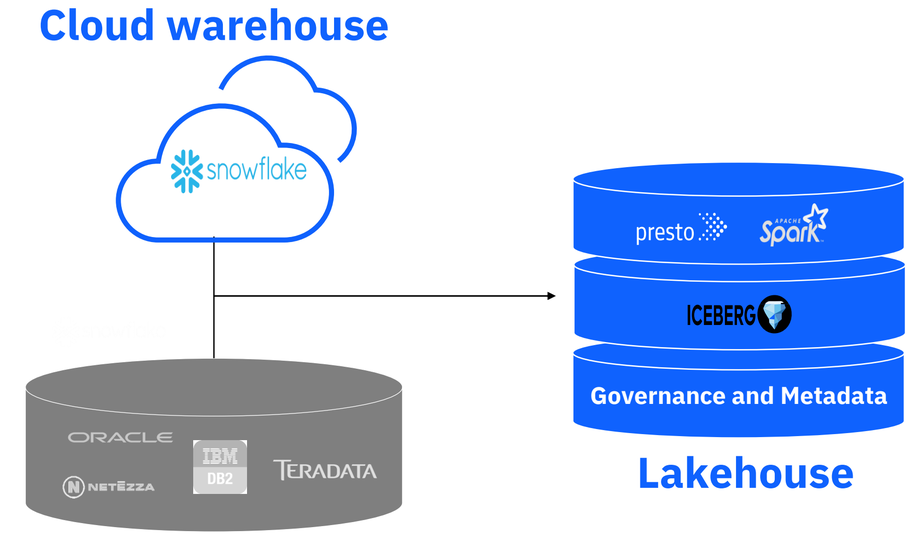
Use case	Query engine	Instance type
ELT/ETL	Spark	Compute
BI	presto	Cache
AI/ML	Spark	Compute

3 워크로드 요구 사항에 따라 자동으로 Scale up/down 되는 compute 리소스 활용



IBM watsonx.data로 데이터 웨어하우징 비용 최대 50% 절감

데이터 웨어하우스의 워크로드를 저비용 오브젝트 스토리지와 목적에 맞는 쿼리 엔진을 활용하여 최적화



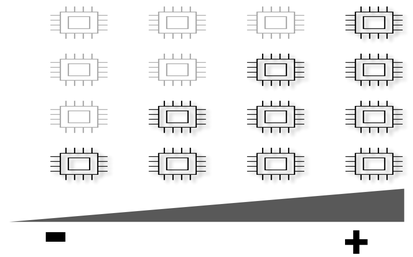
개방형 데이터 형식과 저비용 오브젝트 스토리지 활용

$$$$

용도에 맞는 쿼리 엔진을 사용하여 레이크하우스에서 워크로드 실행

Engine 1	Engine 2
Reporting Business intelligence	Data science Machine learning

워크로드 요구 사항에 따라 자동으로 scale up/down되는 리소스로 인프라 활용 개선



클라우드 데이터 웨어하우스에서 워크로드를 실행하는 것과 비교하여 최대 50% 절감

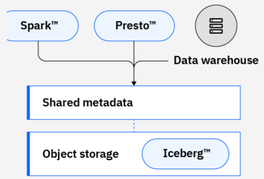
watsonx.data

어디서나 모든 데이터에 대해 AI 워크로드 확장

Data and AI 워크로드를 위한 개방형 레이크하우스 아키텍처 기반의 데이터 저장소

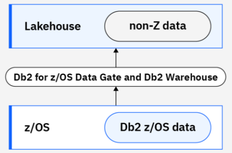
모든 데이터에 대한 single point 액세스

1. 데이터 중복을 최소화하기 위해 오픈 데이터 형식을 처리할 수 있는 도구를 활용하여 단일 데이터 스토리지 생성 및 공유



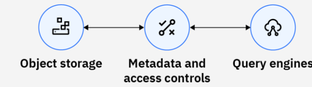
2. 원격지 소스를 캐시하는 기능을 통해 하이브리드 클라우드에서 원격으로 데이터에 연결하고 데이터에 액세스

3. 레이크하우스 분석을 위해 z/OS용 Db2 데이터를 동기화하고 통합



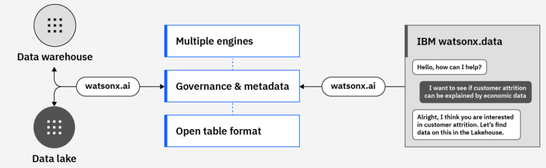
내장된 거버넌스, 보안 및 자동화로 몇 분 만에 사용

기존 분석 데이터에 연결 후 목적에 맞는 엔진을 몇 분 만에 배포



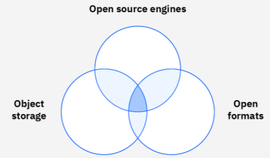
데이터 생태계 전반에 내장된 중앙 집중식 거버넌스를 사용하여 기업 규정 준수 및 보안 문제 해결

Foundation Model을 사용하여 watsonx.data 데이터 및 메타데이터를 검색, 보완, 정제 및 시각화

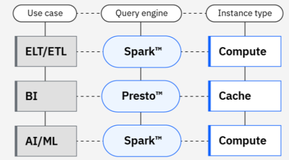


데이터 웨어하우스 비용을 최대 50%* 절감

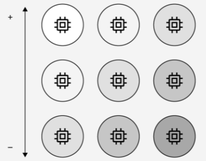
1. 여러 분석 엔진 간에 데이터 공유



2. 목적에 맞는 컴퓨팅 및 캐시 최적화 인스턴스 사용



3. 자동으로 Scale Up 및 Down

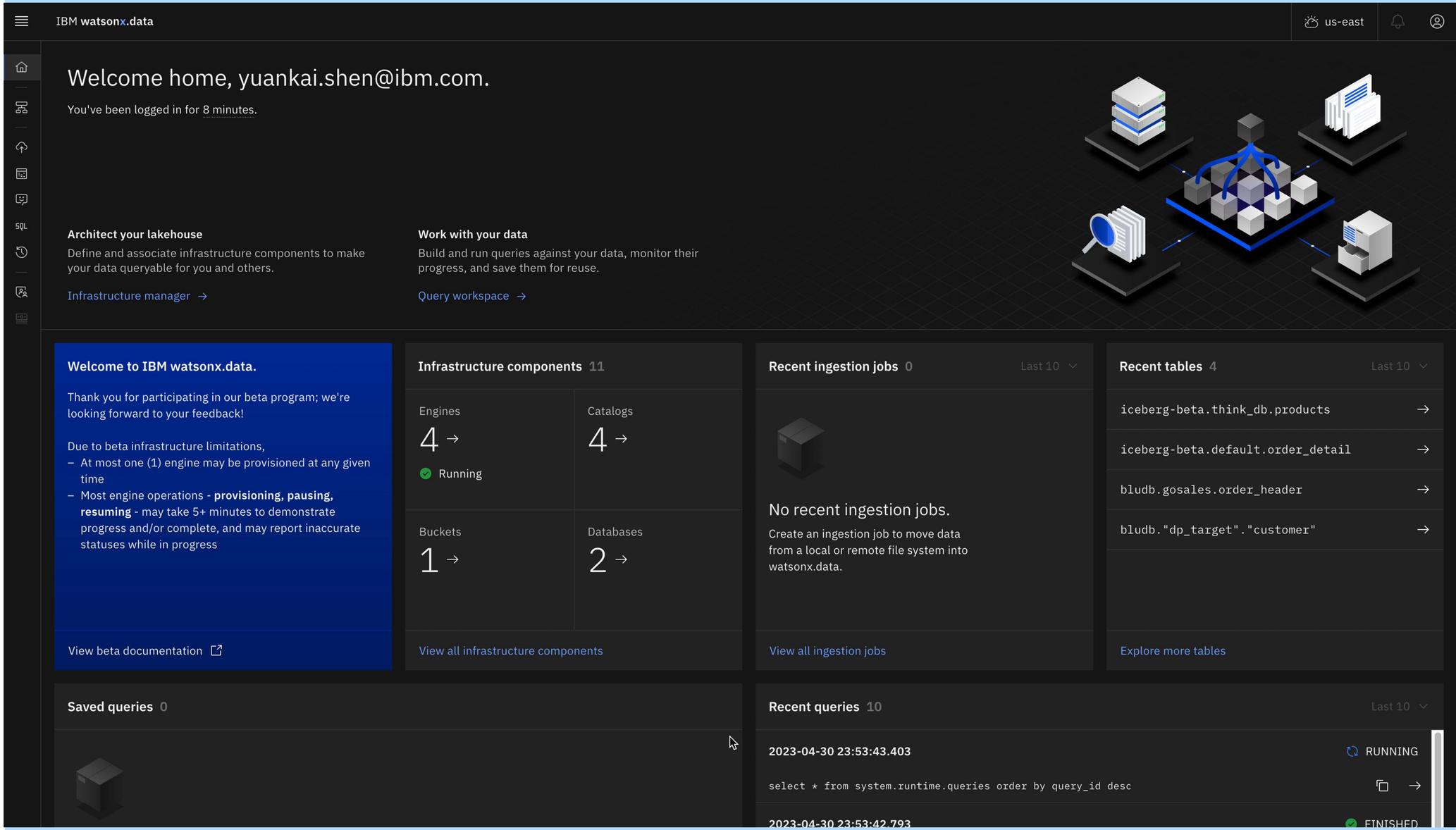


*When comparing published 2023 list prices normalized for VPC hours of IBM watsonx.data to several major cloud data warehouse vendors. Savings may vary depending on configurations, workloads and vendors.

Demo

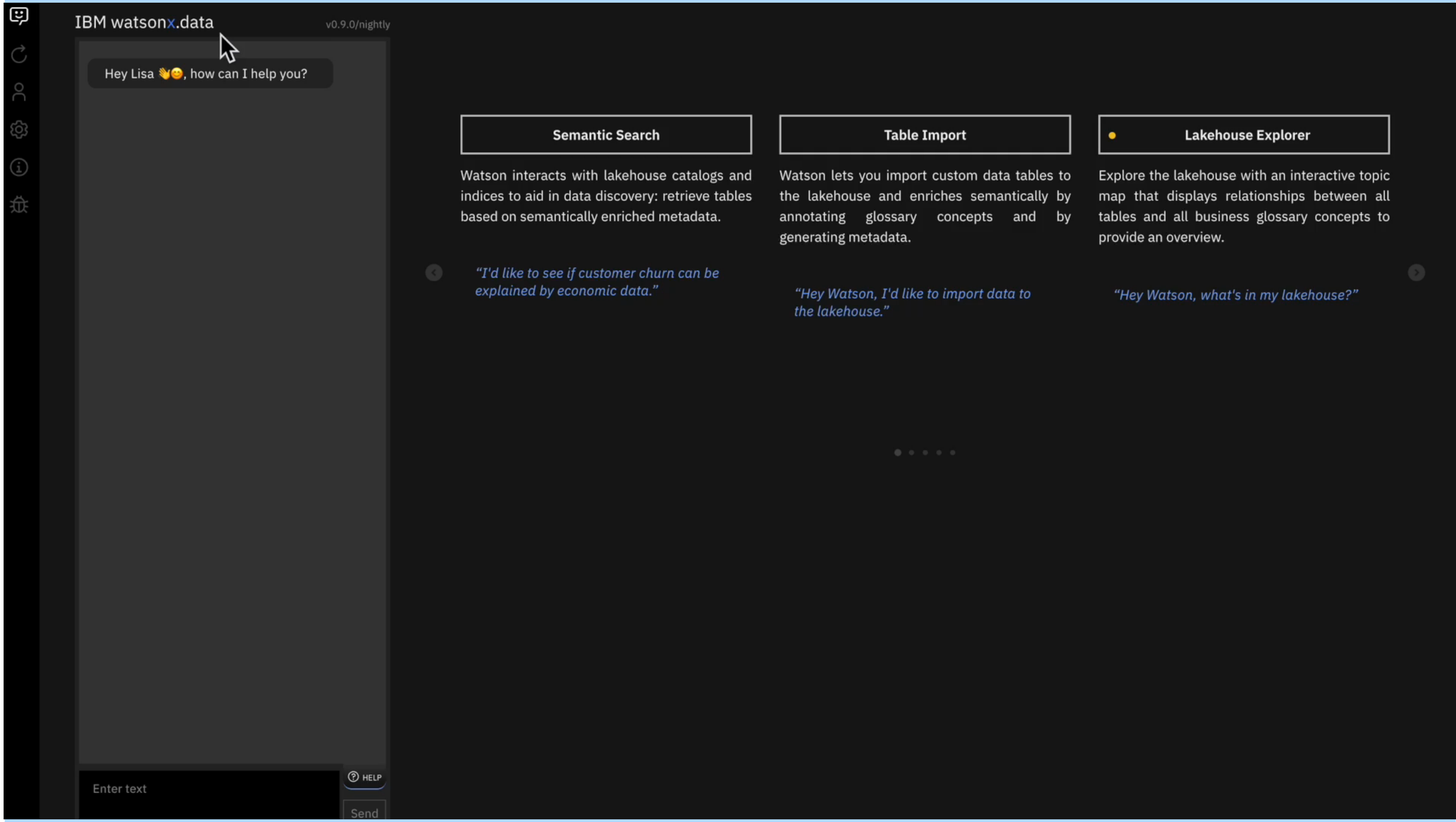
watsonx.data

UI Portal 및 주요기능



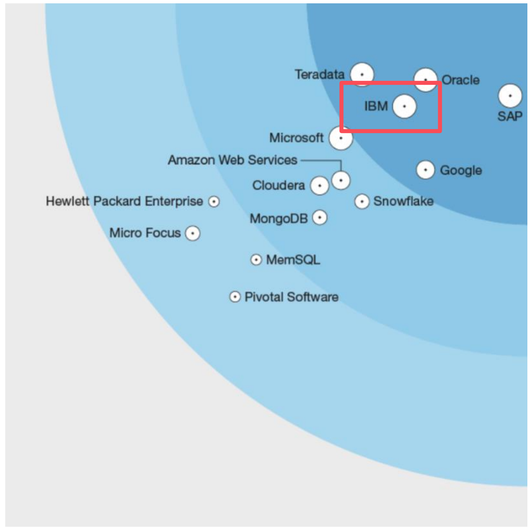
watsonx.data

Semantic Search



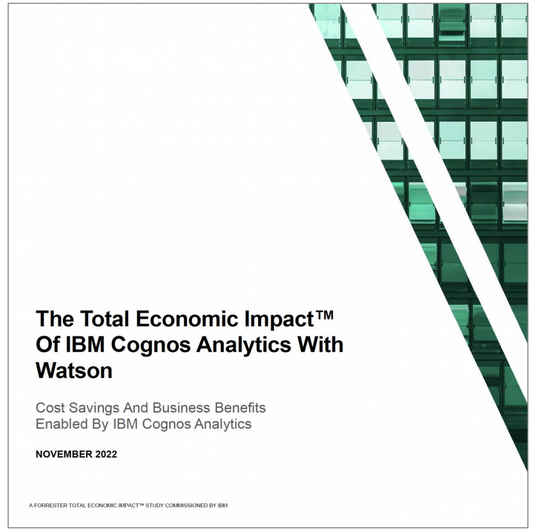
Closing

분석가들은 IBM이 데이터 관리 시장의 선두 주자라는 데 동의합니다



Forrester Wave:
분석을 위한 데이터 관리

Gartner Magic Quadrant : 클라우드 데이터베이스 관리 솔루션



2022년 Forrester TEI IBM Data Management 부문 수상

IBM Innovation Studio 프로그램

무료 평가판
사용하기

watsonx.data
브리핑 세션

- 데이터 레이크 및 웨어하우스 시장의 기술 현황
- 레이크하우스 개요 및 새로운 기술 소개
- IBM watsonx.data 기능 하이라이트
- 핵심 구성 요소
- watsonx.data를 위한 통합 및 IBM 에코시스템
- 데이터 레이크하우스 유스케이스

세션 종류 : 기술 브리핑
대상 : CxO, 현업, 전략기획, IT담당 등
소요 시간 : 60 분
진행방식 : 오프라인
(온라인 진행 가능)
장소 : IBM Innovation Studio
(고객사 방문 진행 가능)
이후과정 :

- watsonx.ai – IBM 생성형AI 기술 및 적용 방안 소개 브리핑
- watsonx.data - Discovery Workshop 워크샵

watsonx.data
디스커버리 워크샵

- 요구사항 정의
- Use cases 도출과 구체적인 정의
- 데이터와 리소스를 포함한 테크 솔루션 확인
- 우선 순위 설정 및 로드맵 정의
- (옵션) watsonx.data 와 유스 케이스 소개가 필요시 아젠다에 포함됩니다.

세션 종류 : 기술 브리핑
대상 : CxO, 현업, 전략기획, IT담당 등
소요 시간 : 0.5 ~ 1일
진행방식 : 오프라인 워크샵
장소 : IBM Innovation Studio
이후과정 :

- watsonx.ai – IBM 생성형AI 기술 및 적용 방안 소개 브리핑