

# 기업의 모든 데이터를 분석하는, AI 시대의 새로운 데이터 분석엔진

No ETL, No Migration. Data Lake를 넘어 Data Mesh로의 여정

SK Telecom

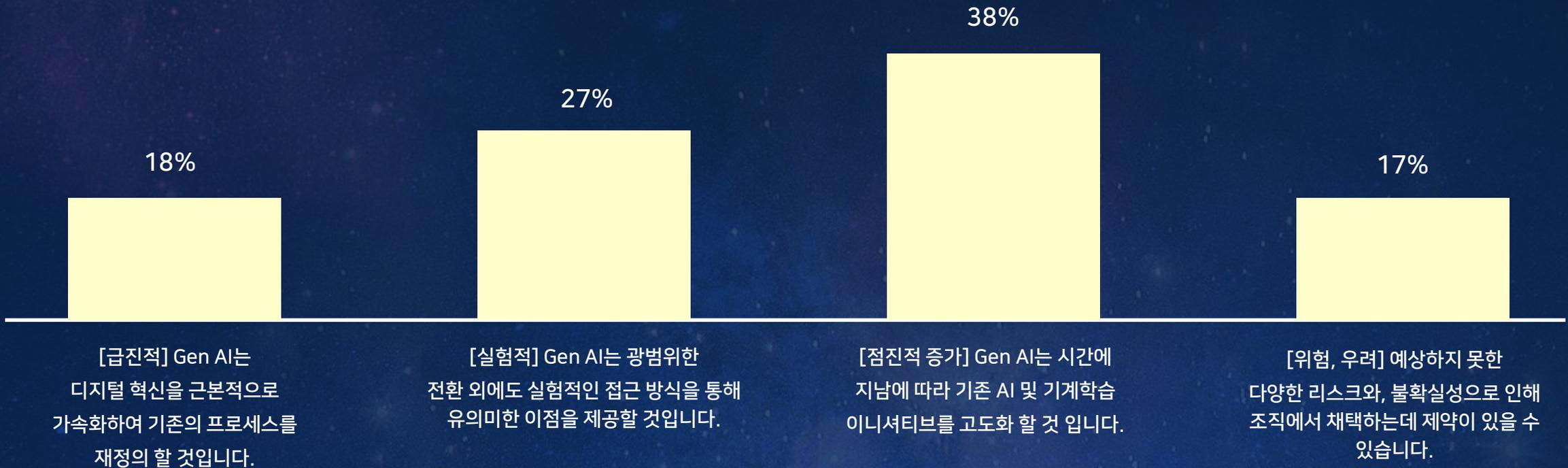
2024.

# Content

- 01 Gen AI and Data**
- 02 Starburst 소개**
- 03 활용 전략 및 구축 케이스**
- 04 About SK Telecom**

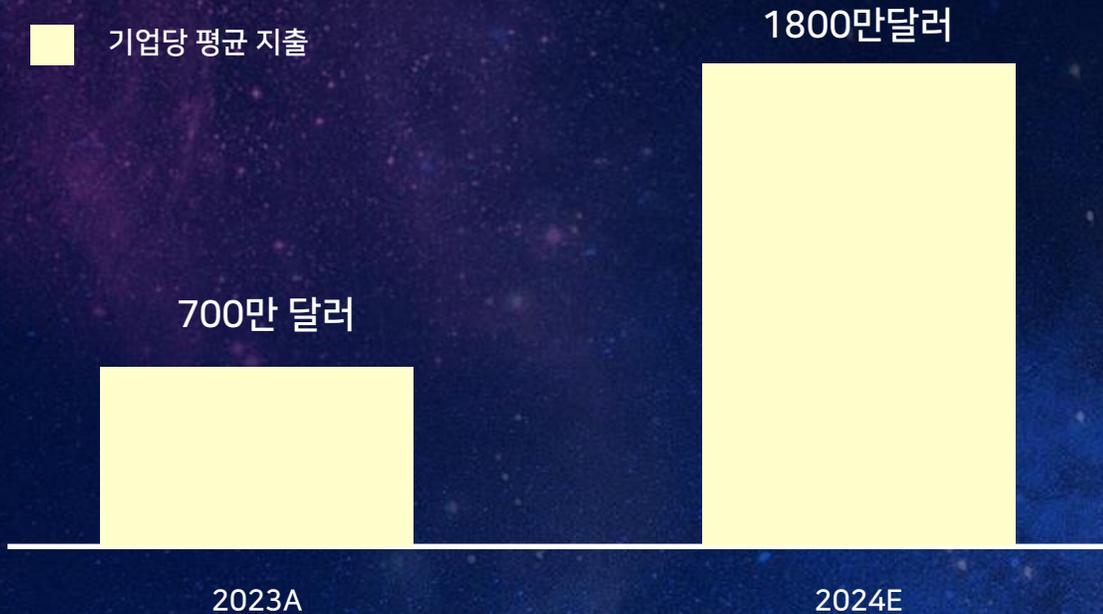
# Gen AI ?

Q : 다음 중 Gen AI 가 귀하의 기업에 미칠 영향도에 대해, 귀하의 견해를 가장 잘 설명하는 것은 무엇입니까?



# 성공적인 기업들은 AI에 대해 과감하게 투자하고 있습니다. 그러나...

LLM에 대한 글로벌 기업 평균 지출(실제 및 예상치)



AI 투자는 2025년까지 전 세계적으로 2,000억 달러에 이를 것으로 예상

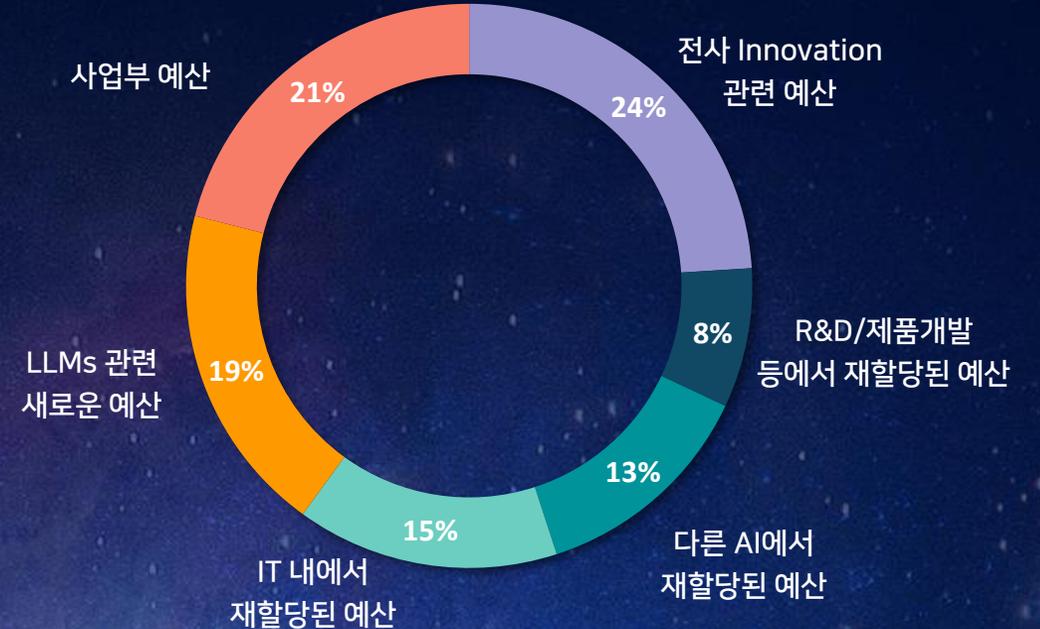
2X

2017년 이후 AI + 비즈니스 결합 니즈 증가로 인해 평균 애플리케이션 수행 작업이 2018년 1.9 에서 3.8로 증가

45%

설문조사 결과 내년 AI에 대규모 투자를 계획하고 있는 기업

예산 할당 : 생성 AI를 위한 돈은 어디서 나오나요?



그러나 설문조사에 참여한 의사 결정권자의 50%는, AI 적용 단계에서 기업 내 저조한 데이터 품질에 어려움을 겪고 있으며, 신뢰하지 못하고 있습니다. (Gartner)

# 현재 기업들의 AI는, 실패할 확률이 높습니다.

48%

설문조사 참여한 리더 중 거의 절반이 전사 AI의 문제점으로 데이터의 품질, 신뢰성에 대한 부족을 꼽았습니다.

69%

또한, AI를 실행하는데 필요한 모든 데이터를 액세스하는데 어려움을 겪고 있습니다.

많은 기대, 높은 성장 가능성에도 불구하고 오늘날 AI에 대한 투자는 데이터 문제로 인해 제대로 된 가치를 창출하지 못하고 있습니다.

## 데이터 관련 주요 이슈 및 과제

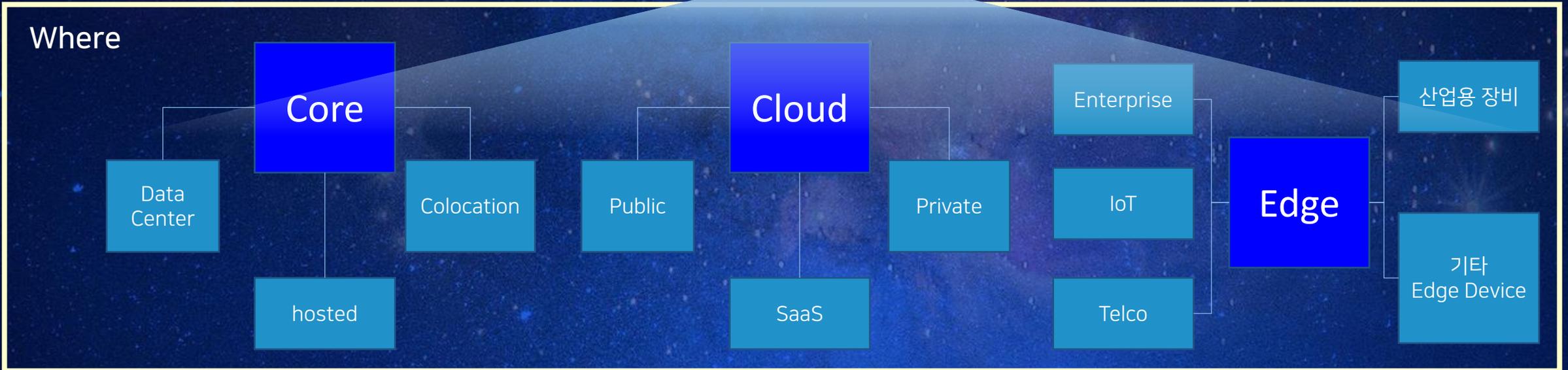
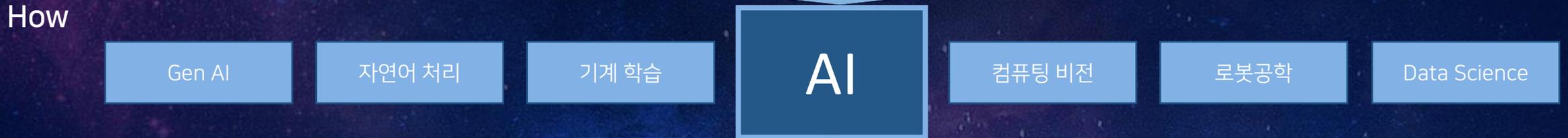


68%

기업은 데이터를 모든 데이터를 활용하고, 사용 가능한 형태로 정리하는데 어려움을 겪고 있습니다.

# 데이터는 어디에나 있으며, 각기 다른 형태와 방식으로 존재

What	Gov.	manufacturing	telco	Retail/ commerce	Agriculture	finance	energy
	교통 안전 공공 ...	전력 공정 품질 ...	통신망 보안 유동인구 ...	재고 개인화 가격, 상품 ...	스마트농업 일기예보 축산 ...	은행 카드 보험 ...	신재생 가스, 광물 ...



# 데이터를 AI로 가져오는 방법 ?

## AI를 위한 데이터 민주화 달성, Data Mesh Architecture 구현 방안

기술

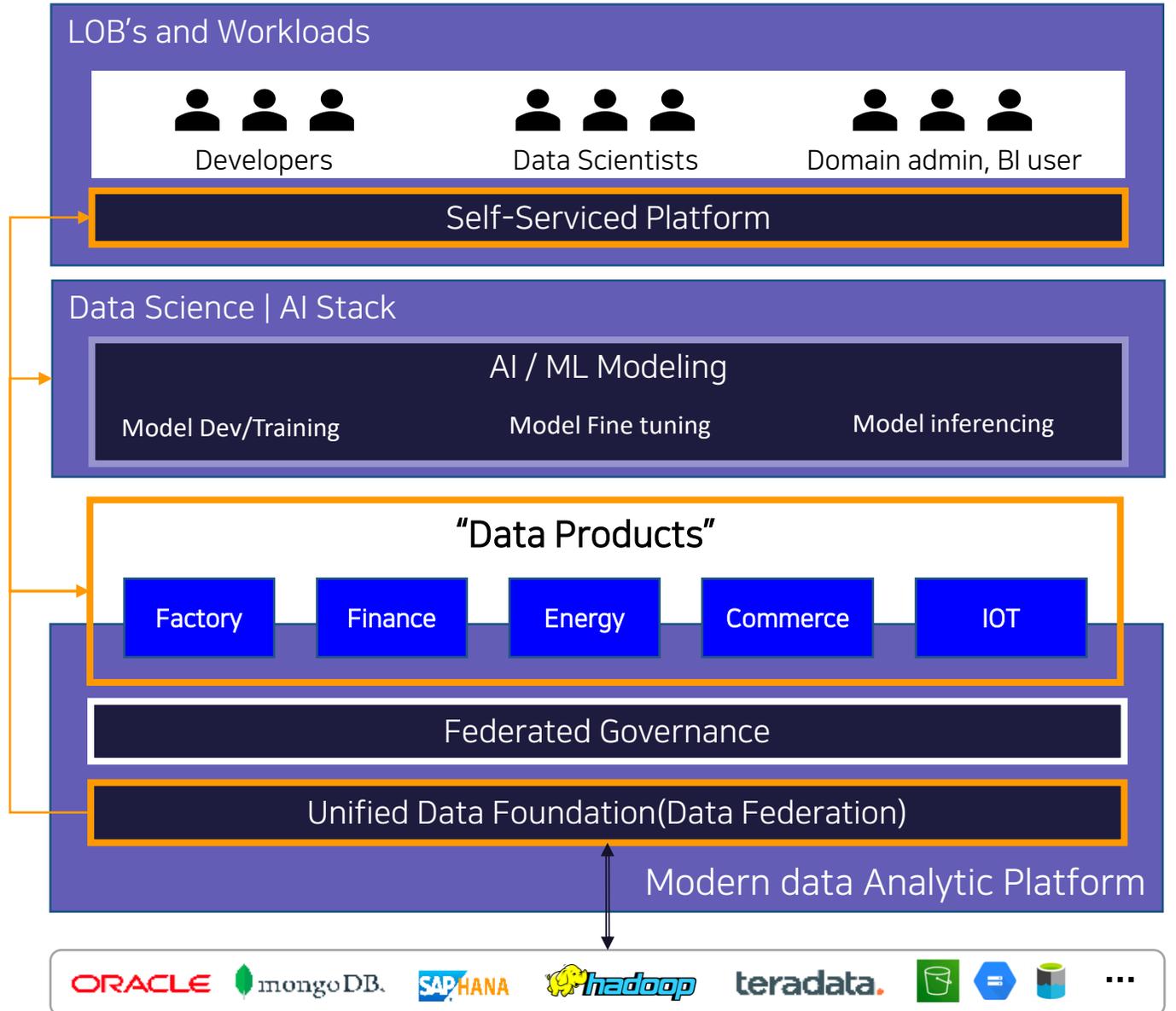
(Federation) 모든 데이터를 연결하여 분석/활용

(Data Product) 선별된 데이터 제품화 및 전달

프로세스

(Self-Service) 현업들이 직접 도메인의 소유권을 가지고 셀프로 데이터 작업 가능한 환경 제공

(Federated Governance) 분산된 조직, 분산된 권한을 통제할 수 있는 연합 거버넌스 체계 제공



모든 이기종 데이터를, ETL 없이  
단 하나의 SQL문으로 분석(Federated Query)

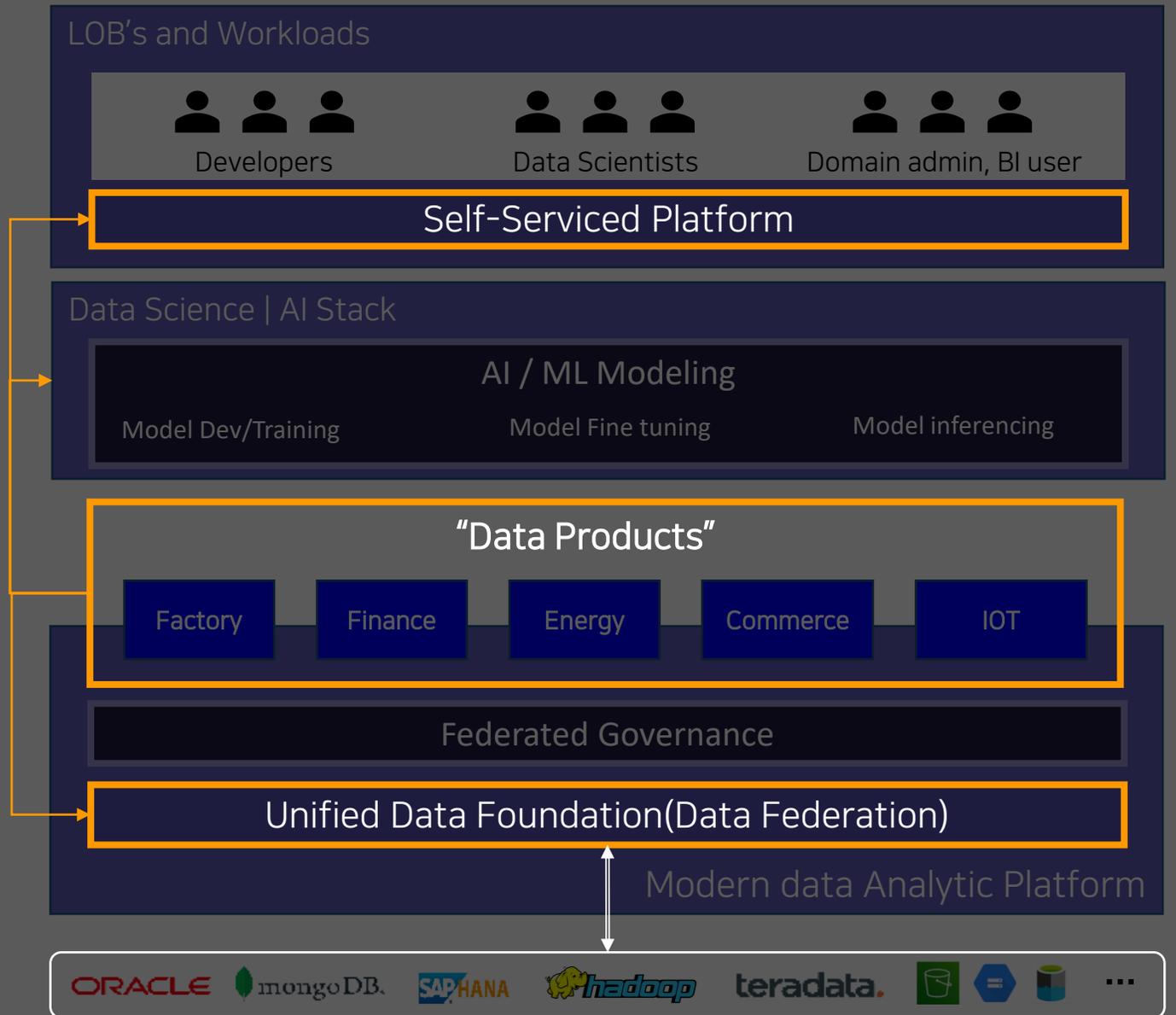


쿼리된 분석결과를 Data Product로 퍼블리싱

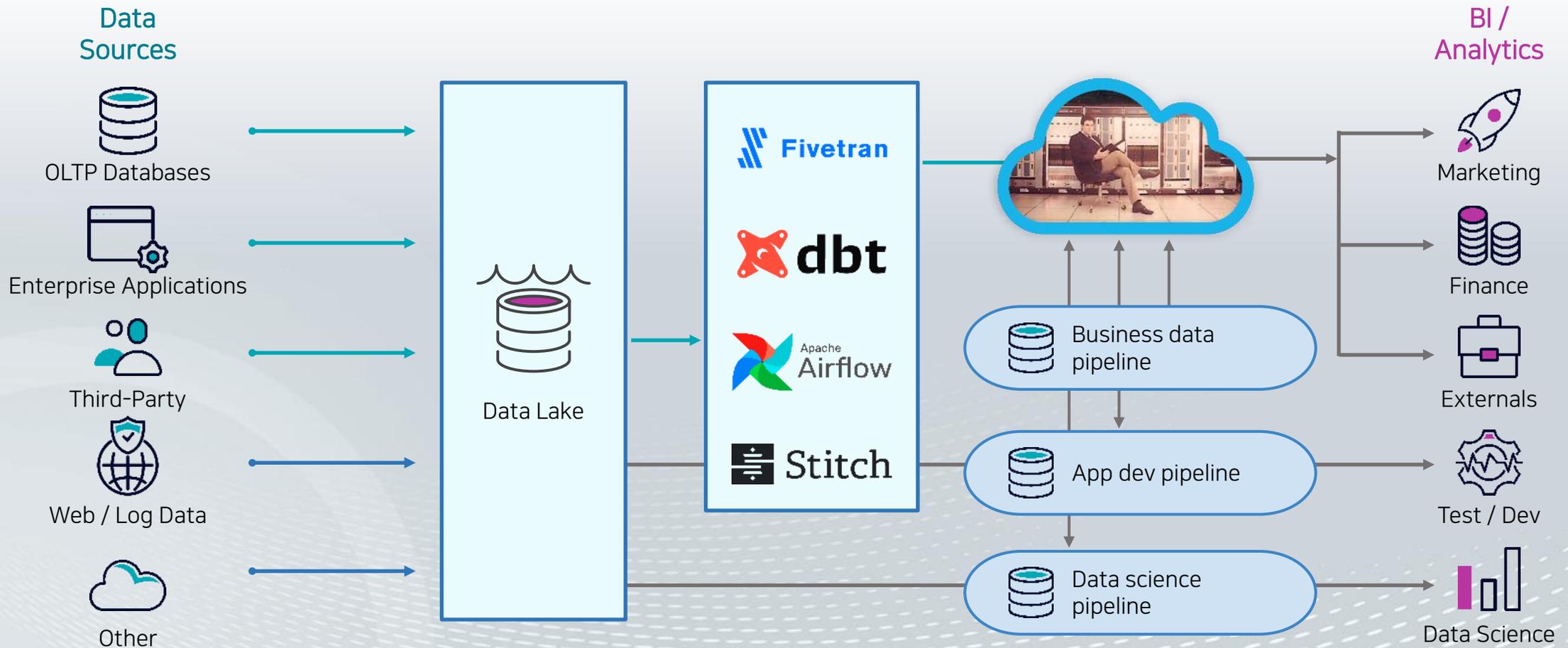


Data Product를 AI 모델에 적용, 품질 극대화  
(feat. Self-Serviced)

잘 짜여진 SQL문 만으로도,  
Modern Data Stack –  
Real AI Service의 완성



# 평균적으로 On-premise 환경과 Cloud 환경 간에 5.4개의 데이터 복사본을 운영 중



# 데이터 이동 및 복제 없이, 모든 데이터에 대한 액세스(Federated Query)와, 고품질의 Data Product를 제공하는 Analytic Engine





# Starburst 소개

# Why partner with Starburst



OUR CUSTOMERS LOVE US

**79 NPS**  
(S사 72 Nps)



Trino 창시자 구축/지원

**90%**  
of all Trino code  
commits



COST EFFECTIVE

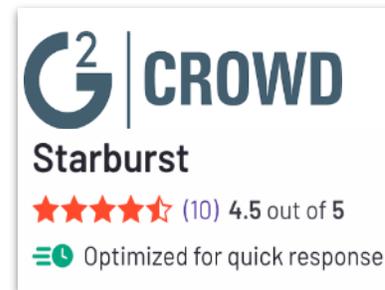
Up to **50%**  
lower TCO



BACKED BY EXPERTS

**1 million**  
combined hours  
Supporting Trino/Presto

# Awards & Recognition



# Starburst Growth

**\$414M**

Total Capital Raised

**\$3.35B**

Valuation

# Starburst는 \***“Analytics Query Accelerators”** 대표 리더 벤더

“Starburst”는 제일 많은 데이터

소스 연결과 성능이 보장된

Data Mesh 환경 구현

Most Commonly Supported Data Sources by Vendor

	Ahana	Alluxio	AtScale	ChaosSearch	Databricks	Data Virtuality	Denodo	Dremio	GridGain Systems	Incora	Jethro	Kylogence	Kyvos Insights	Starburst	Varada
CSV															
Apache Parquet															
Delta Lake															
JSON															
Apache ORC															
Proprietary/Other															
RDBMS (ODBC/JDBC)															
Apache Avro															
Apache Hudi															
Apache Iceberg															
Apache Kudu															
AWS Redshift															
Elasticsearch															
Google BigQuery															
MongoDB															
Snowflake															
Apache Hive															

Source: Gartner  
741458\_C

\* Analytic Query Accelerators (분석 쿼리 가속 시장) :

- Gartner에서 '22년에 신규로 지정한, Data lake, Data warehouse, Lakehouse 아키텍처 등 분산된 데이터 저장소에서 통합된 가치를 추출하기 위한 쿼리 가속 엔진 시장

# Starburst 적용 및 기대 효과



**90%**  
**데이터 인사이트  
확보 시간 단축**

기존 Data warehouse 및 중앙집중형  
아키텍처에서 데이터 처리 시간  
5주~8주 소요



해당 처리 프로세스를 1일 미만,  
몇 시간 단위로 단축 가능



**53%**  
**TCO 절감**

**인프라 비용 절감 -**

데이터 복사/통합X → 스토리지 절약  
데이터 이동X → 컴퓨팅 리소스 비용  
최소화

**FTE 시간 단축 -** 데이터 검색 시간을  
주 단위에서 시간단위로 줄여 비용 절감



**2%**  
**신규 수익 창출  
및 증대**

보다 더 큰 범위 + 최신 데이터  
처리/분석 가능

데이터 준비 단순화, 손쉬운 액세스로  
테스트 중인 가설에 대한 빠른 피드백  
제공 가능 → 더 많은 기회 발견 가능

# Starburst를 통한 데이터 파이프라인

## Step1. Connect

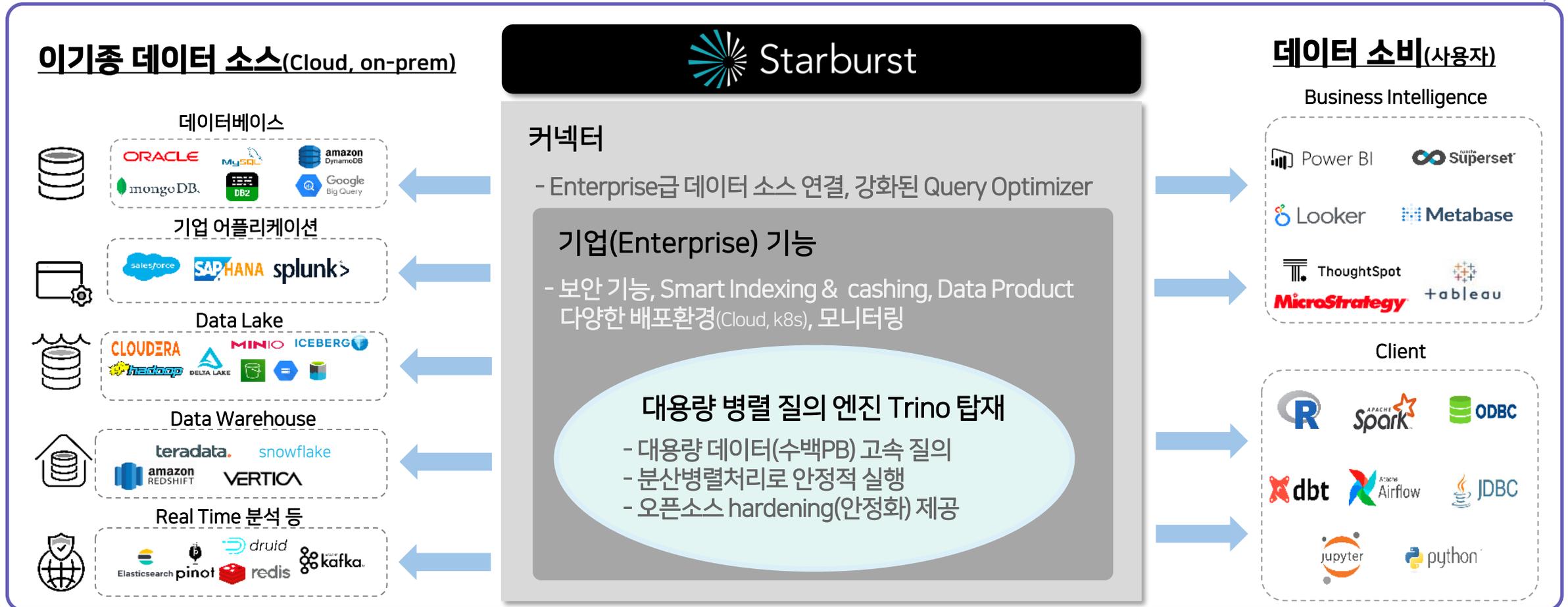
온프레미스 및 클라우드의 모든 데이터 소스에 대한 엔터프라이즈급 연결 제공

## Step2. Query

대규모 데이터 분석을 위한 분산병렬처리(MPP) 오픈소스 Trino 기반 임시 및 배치 워크로드 실행

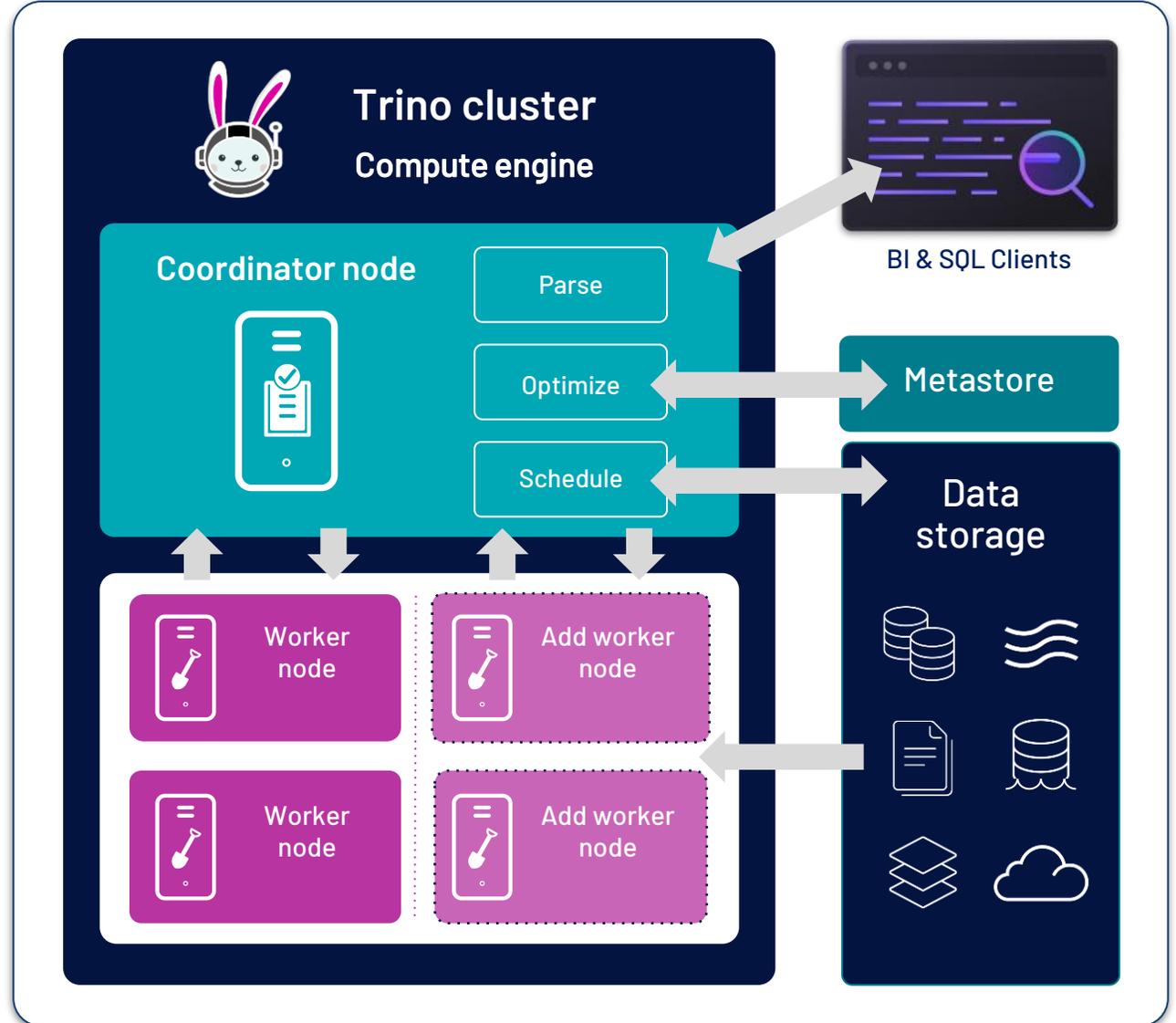
## Step3. Analyze and share

BI 및 Science Tool을 통한 데이터 분석/시각화, 데이터제품을 통한 손쉬운 데이터 공유



# 코어 엔진 : 대규모 병렬처리를 위한 빅데이터 분산 SQL 엔진 Trino

- '13년, Facebook의 300PB data를 분석 및 쿼리하기 위해 개발된 오픈소스 엔진(前 PrestoSQL)
- Trino(=PrestoSQL) 창시/개발자들이 Starburst를 설립
- 강력한 Cost-based resource optimizer 제공
- 표준 ANSI SQL 지원
- 쿼리를 여러 단계/조각으로 분리 후 플랜에 따라 병렬로 처리하여 안정적이고 빠르게 쿼리 수행
- Comping(수평 확장 가능)와 Storage의 진정한 분리
- Apple, Lyft, LinkedIn, Yahoo, Facebook, AWS Athena, **SK hynix(SKT 개발/구축/운영)** 등 사용 중  
(Trino 오픈소스 커뮤니티 9,000+ 멤버, 30,000+ Commits 등)



# Connector : 50종 이상의 최신 및 전통 데이터 소스를 실시간으로 연결

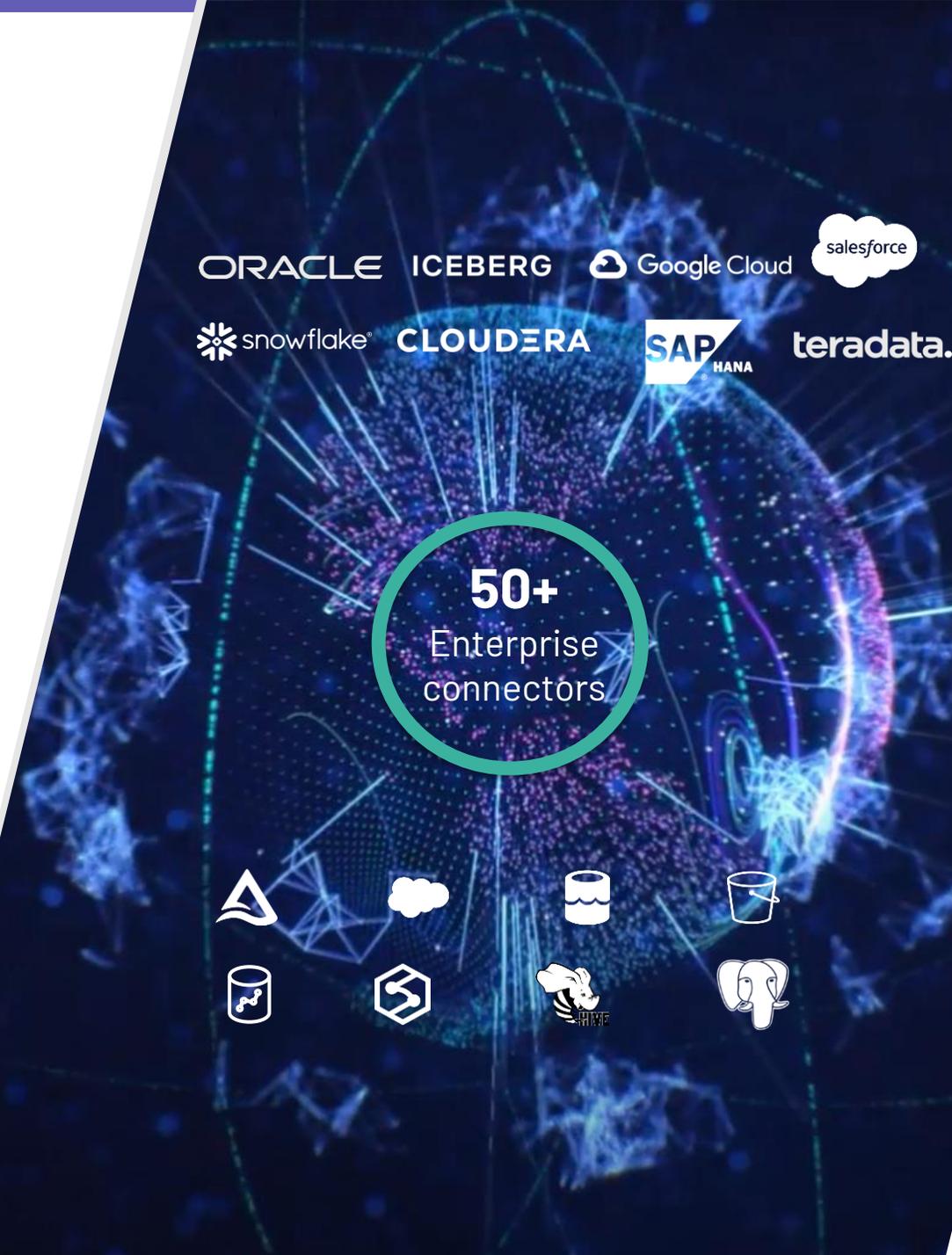
- 데이터소스에 직접 연결하여 실시간으로 조회
  - RDB, No-SQL, DW, Lake, Streaming, Ent App. 등 다양한 형식의 이기종 데이터소스 대상 연합 분석(federation) 제공
  - 성능 향상을 위한 다양한 옵티마이저 수행
  - 보안 및 인증 정책 연동
  - BI Tool, SQL Client, Science Tool 과의 강결합으로 데이터 소비 가속화
- 
- 주요 특화 기능

 **캐시 뷰(Cached Views)**  
Table redirection & materialized views & Smart Caching 등 차별화 캐시로 성능 강화

 **질의 푸시 다운 강화(Pushdown)**  
데이터에 가장 가까운 곳에서 질의로 성능 향상

 **테이블/관리 통계**  
테이블 등의 통계정보 추출로 연산 리소스/속도 최적화

 **다양한 보안 및 인증**  
Kerberos 등 다양한 자격증명, 인증 지원





# Starburst Stargate

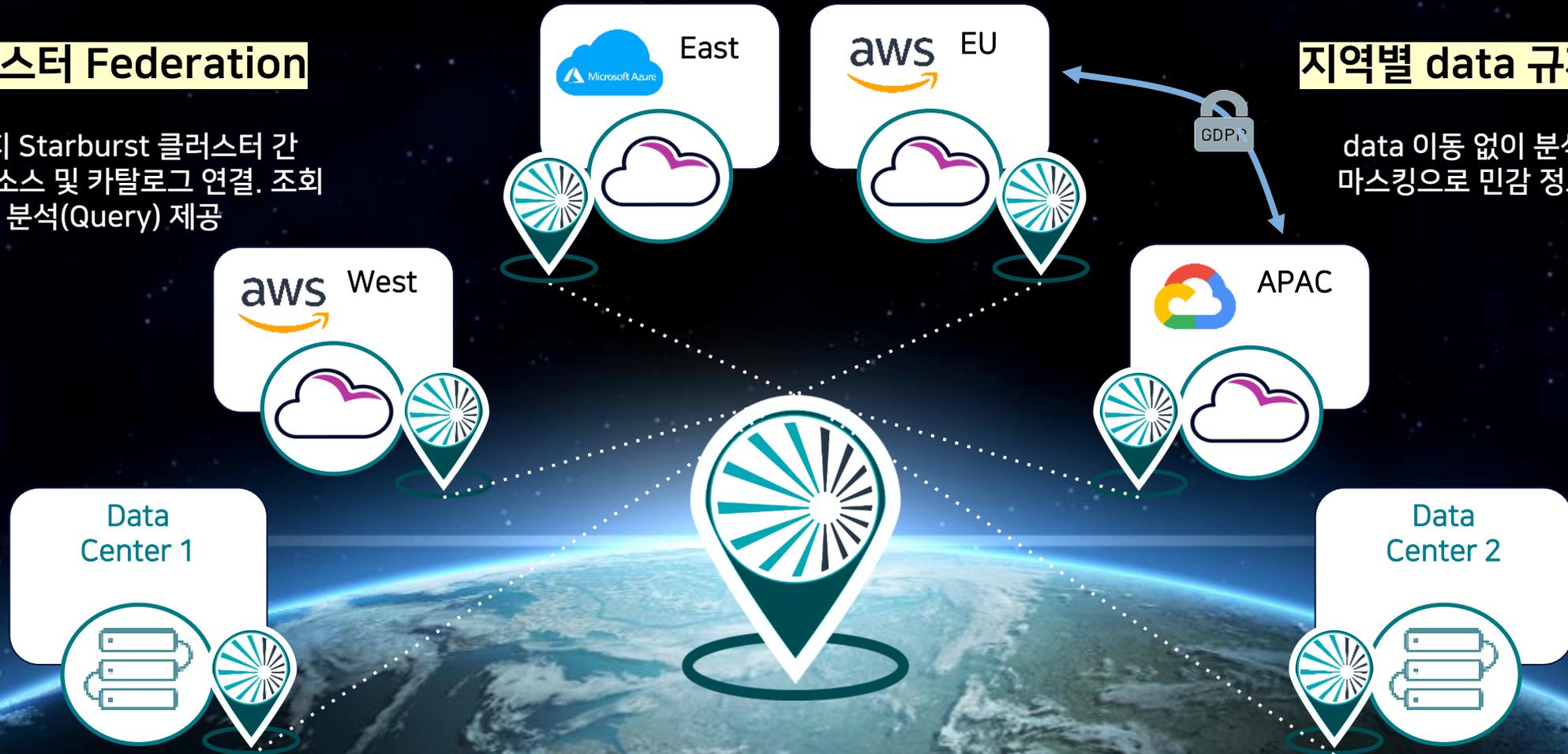
모든 원격지, 글로벌 데이터 간 물리적 이동 없이도,  
손쉽게 통합 분석 및 인사이트 창출 가능

## 클러스터 Federation

원격지 Starburst 클러스터 간  
데이터 소스 및 카탈로그 연결. 조회  
분석(Query) 제공

## 지역별 data 규제 준수

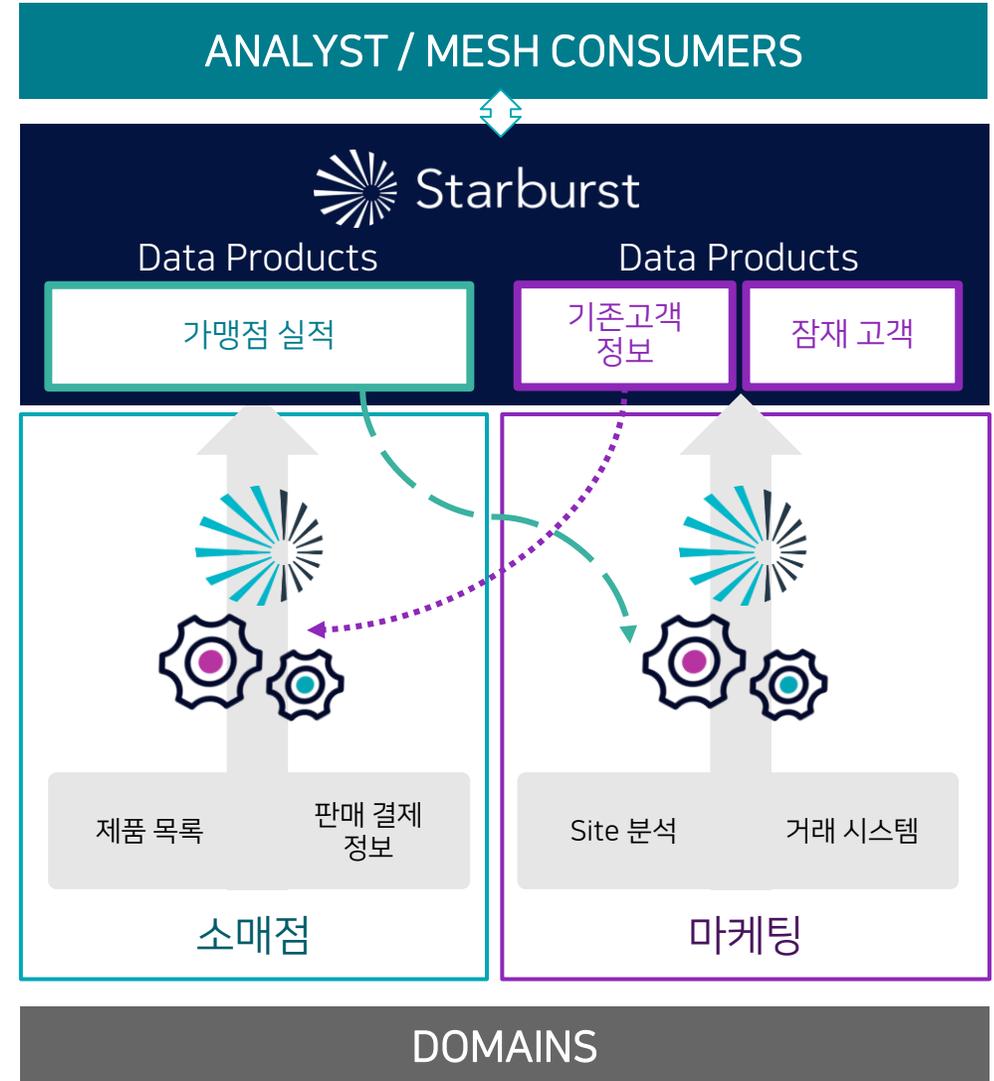
data 이동 없이 분석 가능  
마스킹으로 민감 정보 보호



# Enterprise Feature: Data Product

## - 도메인 중심 셀프 데이터 플랫폼 환경 구성

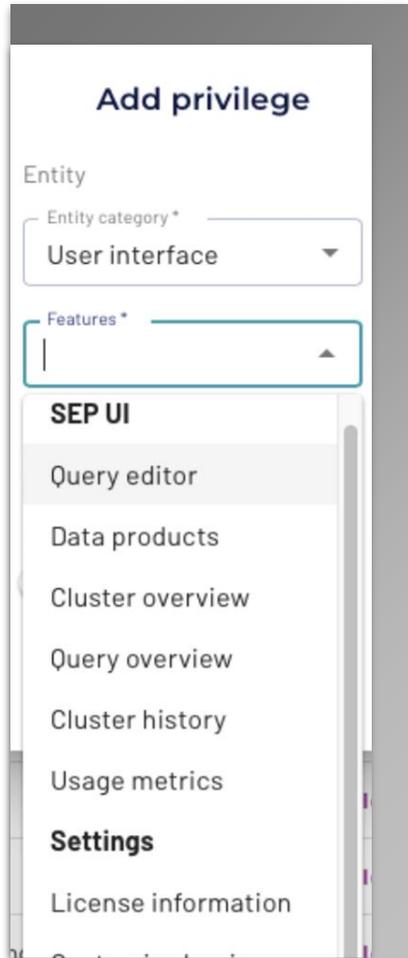
- ① 간소화된 가시성 : 도메인별 통찰력이 반영된 선별된 데이터셋을 손쉽게 게시, 관리하여 전사적인 데이터 소비 현황을 빠르게 파악 가능
- ② 일관된 거버넌스 제공 : 데이터 제품에 대한 명확한 권한/접근 관리를 제공하며 복잡한 데이터 중앙 집중화가 불필요
- ③ 최강의 접근성 : 빈번한 쿼리, 데이터셋을 제품화하여 파이프라인 업데이트 시간을 제거하고 타 도메인 게시 제품을 바로 사용 및 평가



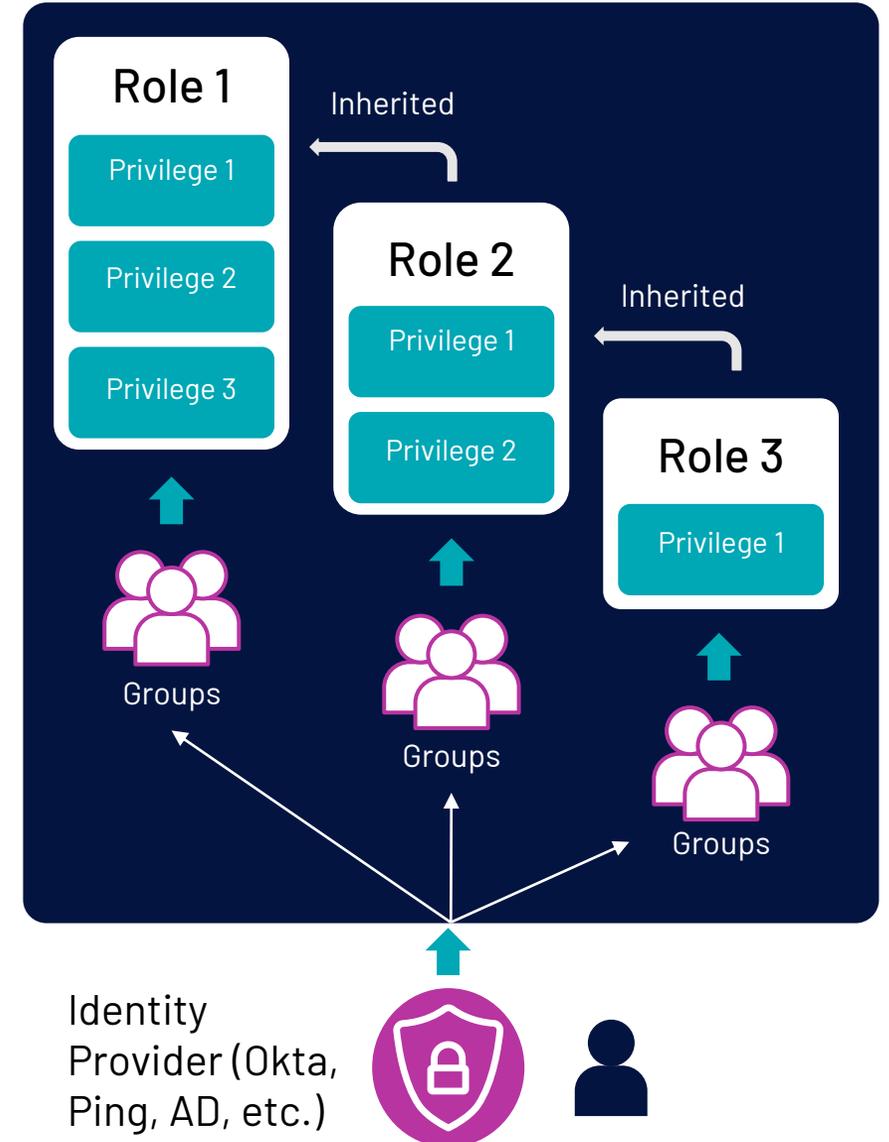


# Enterprise Feature: Security

## - Built in easy Access Control, Masking



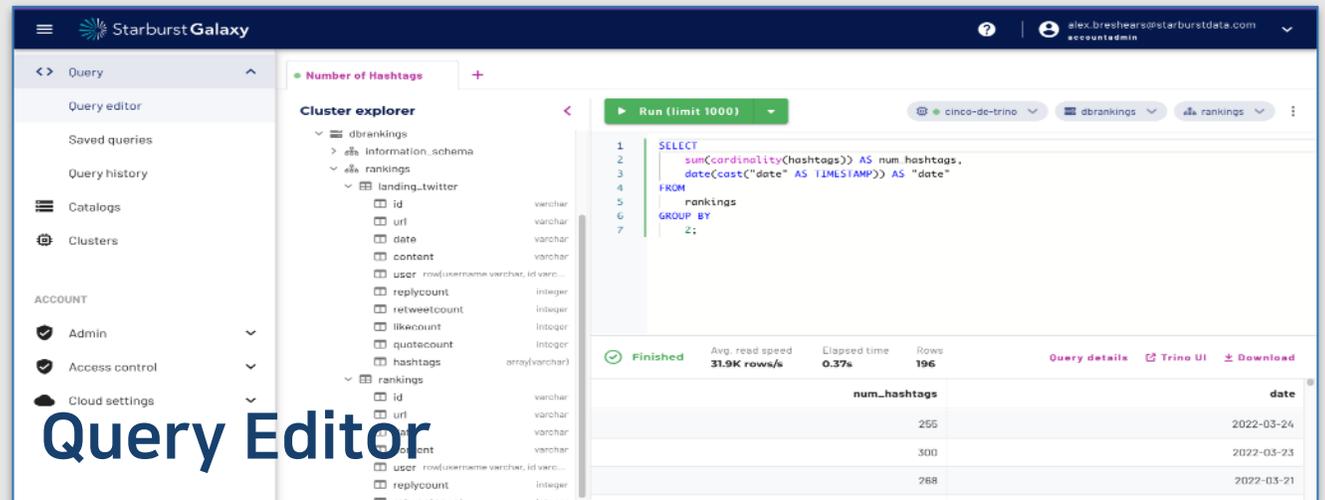
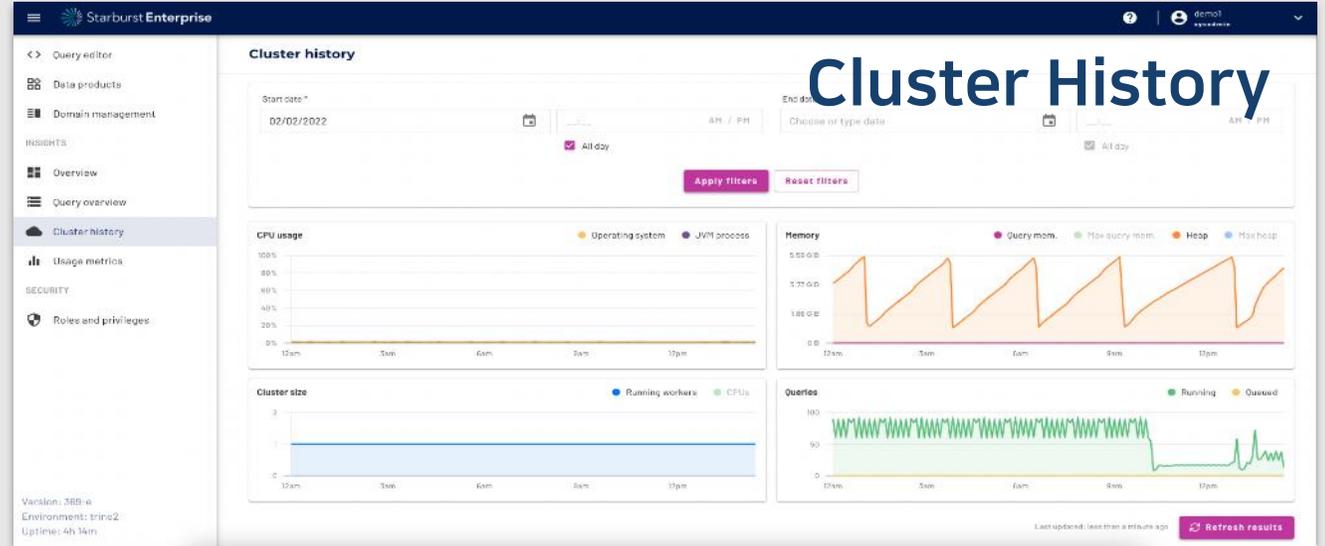
- TLS 암호화 프로토콜 기반 연결
- 카탈로그-스키마-테이블-특정 행/열 수준에 대한 손쉬운 액세스 제한/권한 관리
- Data Product, UI 화면에 대한 액세스 관리
- 모든 쿼리 로깅, 보안 변경 사항에 대한 감사 로그
- Row-level filtering 및 컬럼 마스킹 제공
- 다양한 인증, 자격증명 지원 : Kerberos, Password Credential Passthrough, SSO 등



# Enterprise Feature : 사용 편의성

## 운영 관리 용이성 보장, 직관적인 사용자 환경 제공

- 고품질, 대용량 데이터를 쉽게 검색, 활용이 가능한 직관적인 사용자 환경(UI)
- 클러스터 모니터링, 이벤트 로깅, Query and Usage History 제공 등
- 모두가 알고 있는 SQL문을 사용하고, 지원하는 완벽한 환경 : SQL 함수 및 문법 자동 완성 등
- 연결된 카탈로그, 스키마, 테이블 등에 대한 탐색 기능 등



# Federated Data to BI : 분석된 데이터에 대한 표면화, 시각화

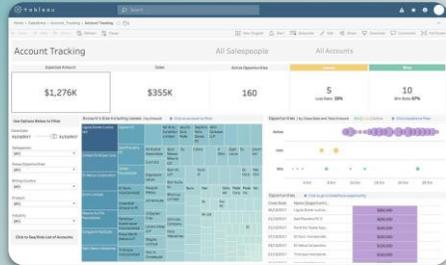


향상된 대시보드 성능 및 애드혹 분석 제공

고성능 라이브 데이터 액세스로 데이터 추출 작업 최소화

데이터를 이동, 복사, 혼합할 필요 없이 모든 데이터에 간편하고 안전하게 액세스

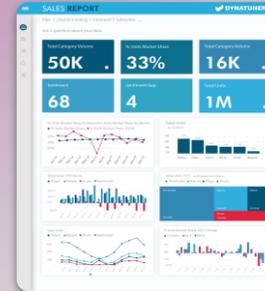
시각화 지연 요소 및 번수 최소화



TB에서 PB로의 확장성 보장

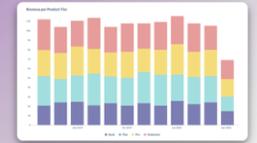
DirectQuery 모드를 통해 위치에 관계없이 소스에서 직접 데이터 쿼리

엄격한 세분화된 액세스 제어 및 강력한 보안 통합 제공



클로저와 상호 운용 가능

사용자가 스타버스트에서 쿼리된 데이터를 사용하는데 SQL 불필요



프리셋과 통합

50개 이상의 엔터프라이즈 데이터 원본에서 SQL 쿼리 작성, 새 테이블 및 시각화 만들기



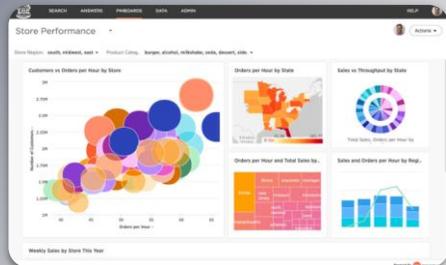
단일 소스 클라우드 데이터 웨어하우스 모델을 넘어 데이터에 대한 기본 연결 확장

더 깊은 인사이트 확보

비용 최적화 개선

하이브리드 클라우드 및 50개 이상의 데이터 소스 간 페더레이션 가능

클라우드 마이그레이션 지원



추출 또는 하위 집합 작업 제거

실시간에 가까운 완벽한 데이터로 운영 워크플로우 강화

데이터를 인메모리 DB로 가져옴과 동시에 스타버스트 직접 활용 가능



Community connectivity



# Deploy anywhere -

On-prem(베어메탈, 쿠버네티스), Cloud(public, private) 및 Hybrid 모든 환경에서 이용 가능



전문가 지원과 서비스가 제공되는 설치형 솔루션 (구축형)

On-prem, Cloud, Hybrid 모두 지원



Cross - Datacenter & Cloud

Data Center 1

Data Center 2

Data Center 3

Data Center 4

West



EU



East



APAC





# 활용 전략 및 구축 케이스

## [활용 예시 1.]

### 데이터가 어디에 있든, 바로 연결 및 분석이 가능한 Data Federation 환경 구현

- 기업 내 분리된 데이터 환경(On-premise, Public Cloud 등)에 대한 단일 액세스, 분석 환경 제공
- 데이터 아키텍처 내 분산되어있는 모든 이기종 data에 대한 단일 조회 환경 제공
- 불필요한 data 중복/이동/ETL 등을 줄이고 일관된 보안 정책을 손쉽게 구현하여 위험을 줄이고 비용을 절감
- Data 접근 경로를 일원화하여 데이터 민주화 기반 마련

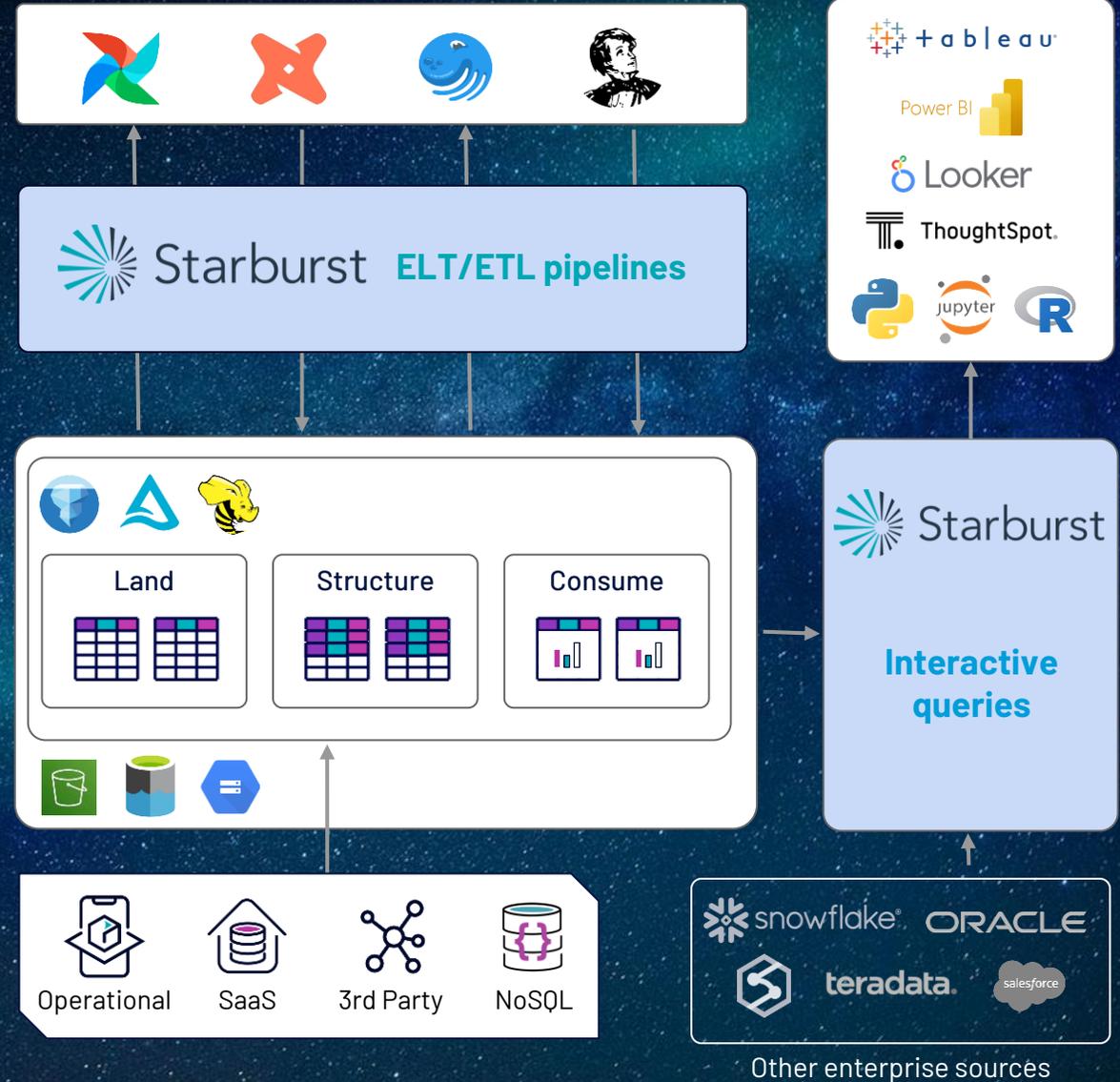


## [활용 예시 2.]

### ETL / batch processing 대체

- ANSI SQL만을 사용하여 모든 이기종 데이터 소스에 대한 파이프라인 구성 및 운영

- 대화형 및 배치 워크로드 모두를 지원하는 고성능 쿼리 엔진
- 익숙한 표준 SQL을 활용하여 누구라도 손쉽게 데이터 추출-변환-적재(ETL) 파이프라인 대체
- 커넥터를 활용하여 불필요한 data 복제/이동을 최소화
- 다양한 작업관리툴(dbt, airflow)과 Starburst fault tolerance 기능을 활용하여 성공 보장



# [활용 예시 3.]

## Data Lake house Analytics

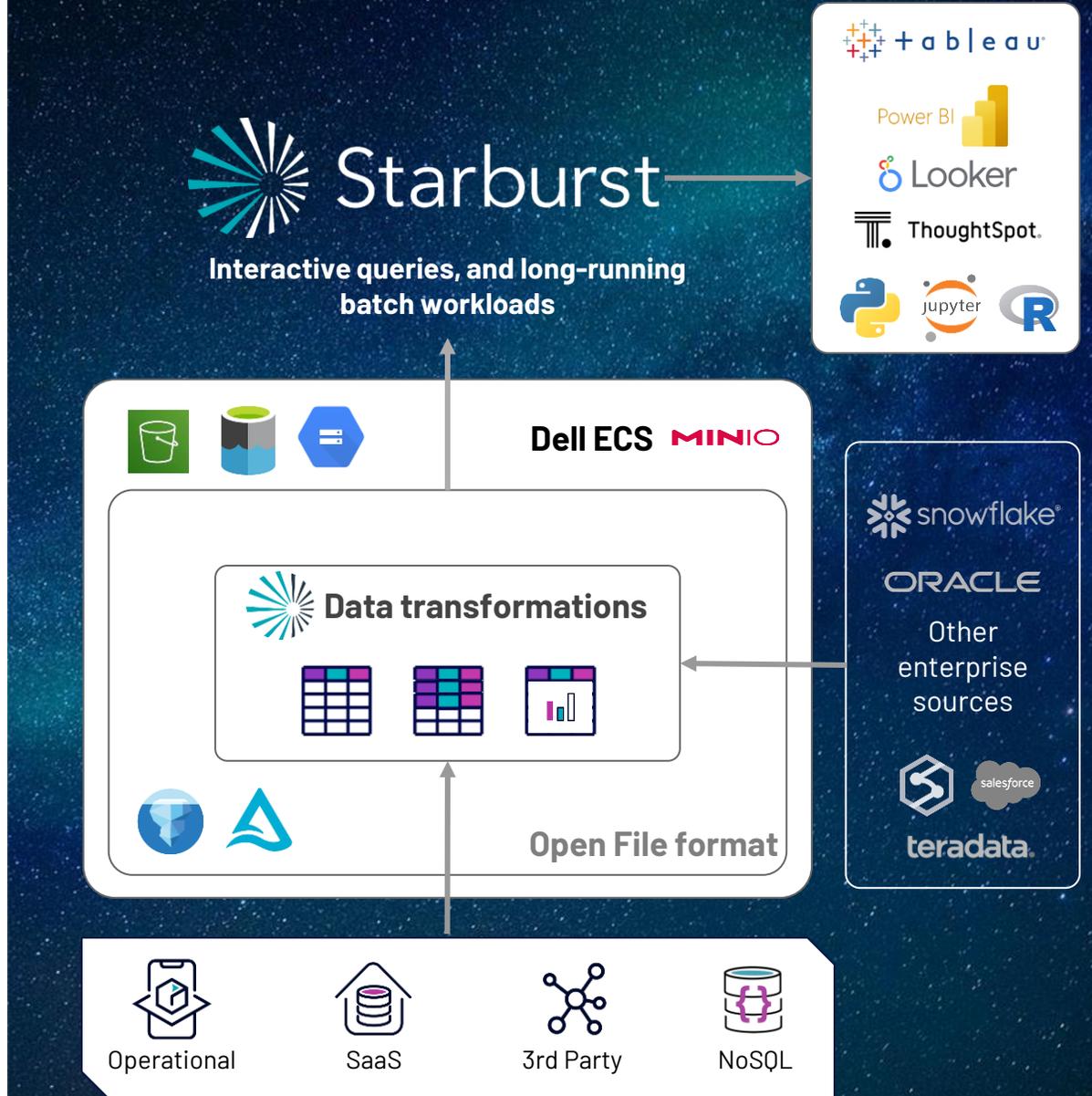


쉬운 분석 및 사용 -  
규모의 한계

Data Warehouse 와  
Data Lake 장점의 결합

손쉬운 수집/적재 -  
어려운 쿼리 및 분석

- 비용, 성능의 최적화를 위한 스토리지 컴퓨팅의 분리 구조
- 개방형 파일 포맷 지원을 통해 레이크하우스 아키텍처 강화
- 50개 이상 최신 및 레거시 소스 커넥터로 다양한 데이터의 수집 및 분석



# [활용 예시 3.] Dell Data Lakehouse for AI

AI에 최적화된 Dell HW 및 소프트웨어(Starburst etc) 기반의 완전 통합형 데이터 플랫폼

BI, AI/ML 등 광범위한  
도구 생태계에 연결

Federation을 지원하는,  
안전한 초고속 쿼리 엔진

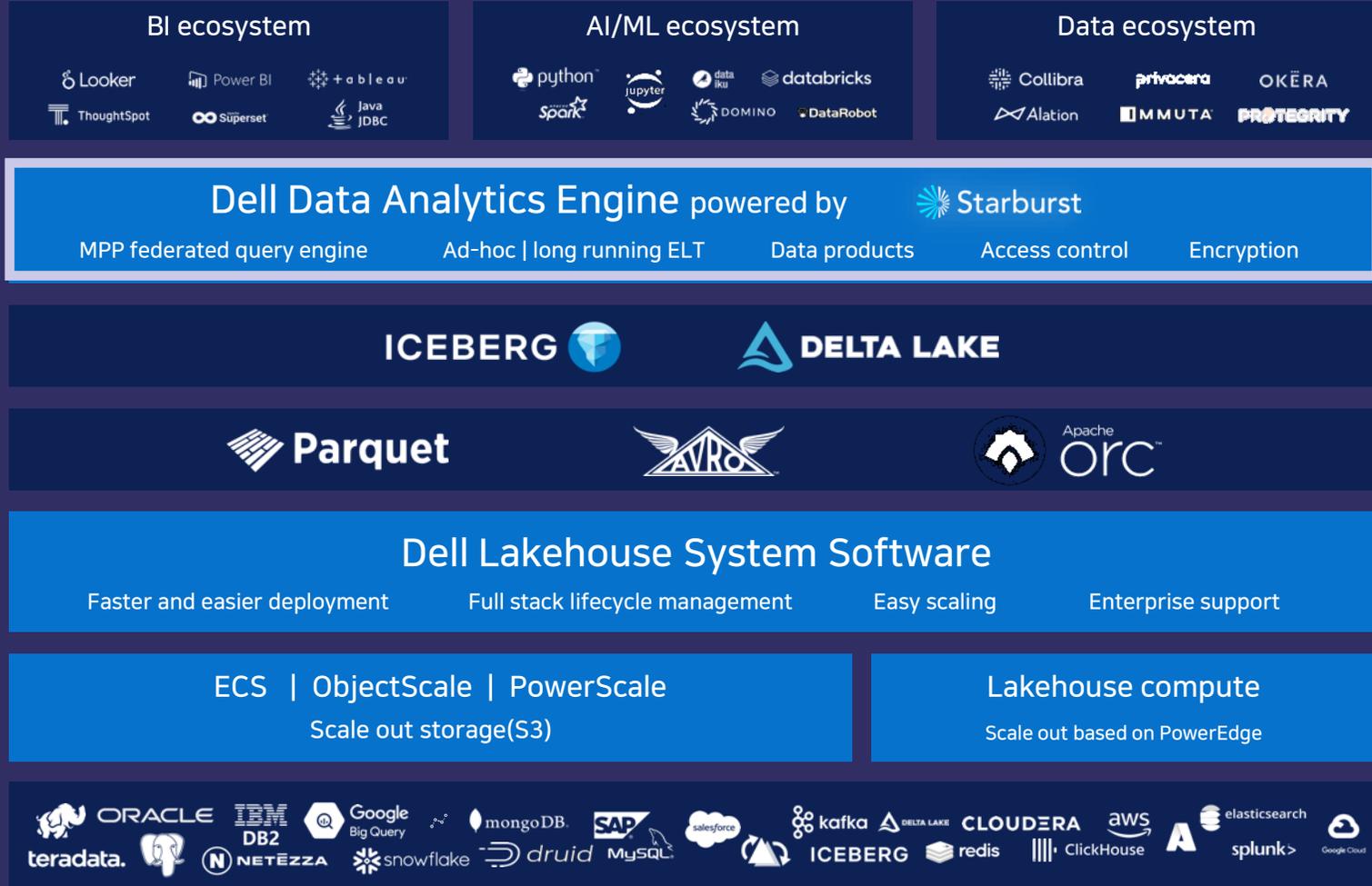
오픈 테이블 포맷

오픈 파일 포맷

레이크하우스 관리 및  
오케스트레이션 도구(k8s)

컴퓨터와 스토리지의 분리  
(Dell ECS & PowerEdge)

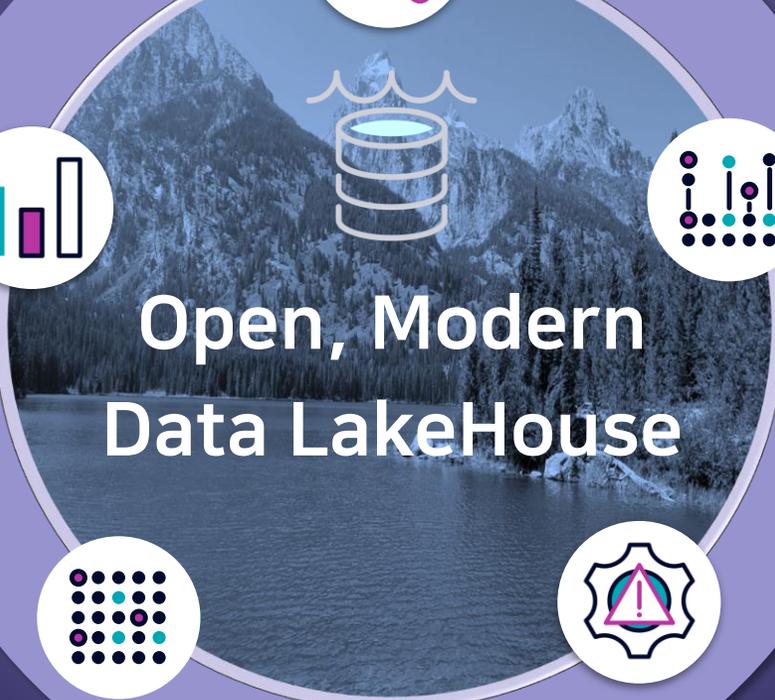
모든 데이터 소스 연결 및 통합



Federate all Data,  
스마트한 통합분석

High Performance  
S3 Storage

Open, Scale out  
Architecture



Open, Modern  
Data LakeHouse

Integrated Turnkey  
solution

Security and  
Governance

# [도입 사례] 미디어 분야 : 2억 달러 이상 수익 창출 및 아키텍처 확장 비용 절감



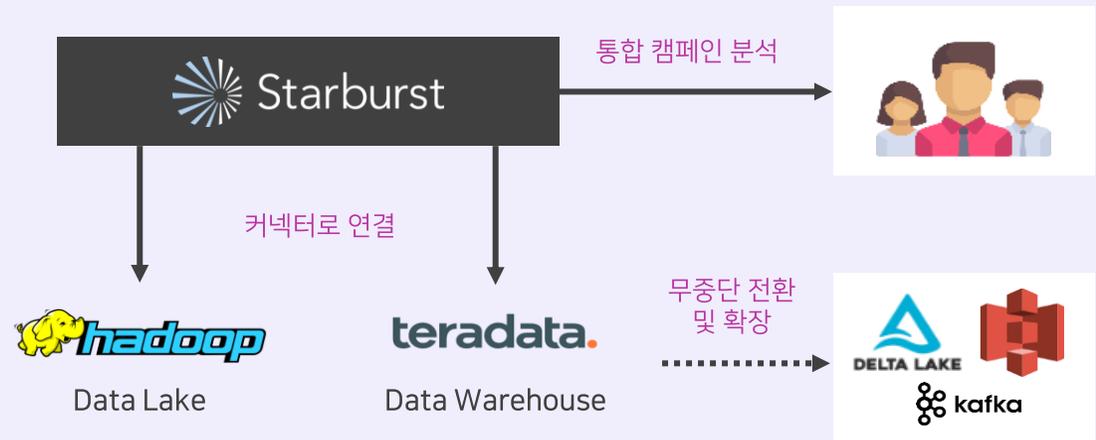
세계에서 두번째로 큰 방송 및 케이블 기업이자 미국 최대의 가정용 인터넷 서비스 제공 업체로 NBC Universal 모회사로 성장하였으며, 미국 내 40개 주와 컬럼비아 특별주 내의 모든 상업 고객에게 서비스 제공

## 배경 및 이슈

- ① **이원화된 데이터 아키텍처 보유**
  - 사용량 정보 : DataLake(Hadoop)
  - 과금 정보 : DataWareHouse(Tera data)
- ② **데이터 통합 분석 어려움**  
→ 통합 프로모션/캠페인 실행 불가
- ③ **Data를 한곳으로 통합 시 18개월의 장시간 소요**
  - 인적/시간/비용에 대한 과도한 리소스 발생
- ④ **기존 DataWarehouse(On-premise)의 확장 및 성능적 한계 봉착**

## Solution / Benefit

일일 약 250~300TB의 안정적 데이터 프로세싱 가능



**5 Week**

Data 통합 없이 바로 통합 캠페인 실행

**\$200M**  
(2,600억)  
신규 수익 창출

**TCO 61% 절감**  
데이터웨어 하우스, ETL, 인건비 절감

**93% 시간 단축**  
2시간 → 9분  
처리시간 단축

# [도입 사례] 플랫폼 분야 : 저장 비용 5x~10x 절감



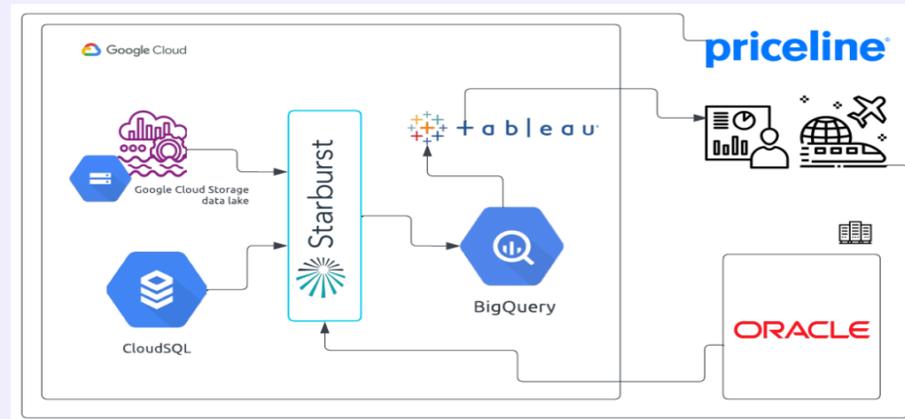
세계 최대 여행사 Booking.com의 자회사이자, 여행 관련 구매 할인율 정보를 맞춤형으로 제공하는 온라인 여행 플랫폼 기업으로, 사람들이 연간 10억 달러 이상을 절약할 수 있도록 다양한 할인 정보 제공

## 배경 및 이슈

- ① 글로벌 인지도 및 DAU 대비 임직원은 약 1,000명으로 인적 리소스 부족
- ② 매월 수백만 여행자에게 개인화된 추천 서비스를 제공
- ③ 다양한 환경에 저장되는 분산된 데이터
  - Oracle RDBMS (On-premise Warehouse)
  - Google Cloud Storage(GCS) Data lake
  - Google BigQuery
  - Google Cloud Platform의 Cloud SQL Instance 등

늘어나는 방대한 스트리밍 데이터에 대한 비용 이슈, 분산된 데이터 액세스에 대한 시간 단축 필요

## Solution / Benefit



### 저장 비용 5~10배 절감

- ✓ BigQuery의 비싼 컴퓨팅 비용 + 데이터 저장 비용 중복 발생
- ▶ Starburst 대체 및 통합으로 스토리지 비용 5~10배 절감

### 시간 및 리소스 절감

- ✓ 분산된 데이터에 대한 복잡한 ETL로, 1~2일 전 데이터 분석
- ▶ Starburst를 통한 ETL 최소화, 1~2일 → 분 단위로 단축으로 시간 및 리소스 절감

# [도입 사례] 금융 분야 : 고객 가입률 10% 상승 및 ETL 시간/비용 제거



미주 온라인보험플랫폼 기업으로, 수십만개의 보험 및 금융 서비스 상품 시장에서 개인별로 최적화된 이상적인 상품과 가치를 맞춤형으로 제공하는 플랫폼 기업

## 배경 및 이슈

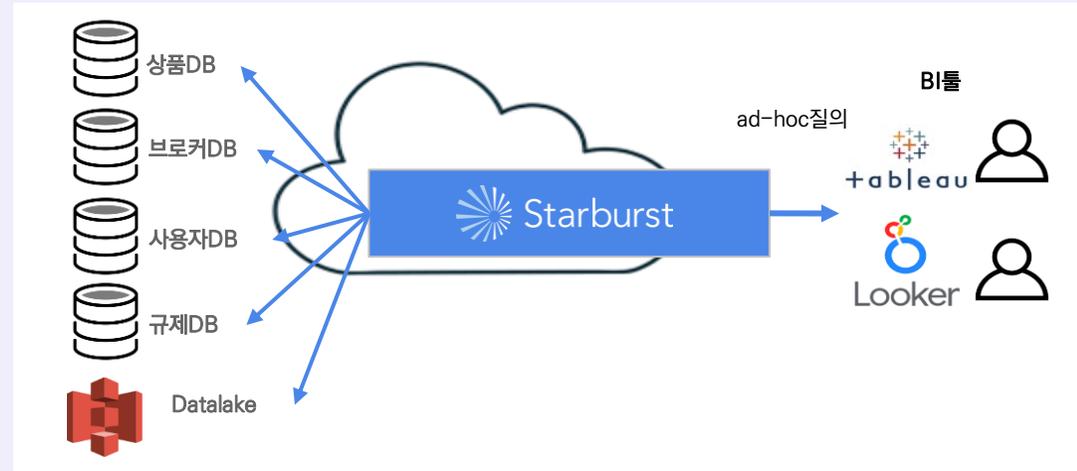
- MSA 아키텍처, 사일로된 Data 환경으로 새로운 가치 창출을 위한 Data 결합 및 분석이 어려움



→ 다수의 Postgres DB, AWS S3 별도 운영

- 통합 분석 시 반복적인 Data 이동, 복사 등의 작업 발생으로 단순한 TASK에도 Week 단위 소요시간 발생
- 일반 User 대상 인사이트 확보를 위한 BI 툴 사용 : Power BI, Tableau 등

## Solution / Benefit



1개월 → 1일로 단축

- ✓ ETL 최소화 및 추가적인 ETL TASK 불필요로 비용 절감
- ✓ Data 통합 분석에 대한 시간 단축 (3~4개월 -> 3~4일)

고객 가입률 10% 증대

- ✓ "손쉬운 Data 액세스/분석에 따른 실시간 대응으로 실제 10% 이상 고객 가입률이 증가했습니다."

- Shen Wang, Principal Data Engineer

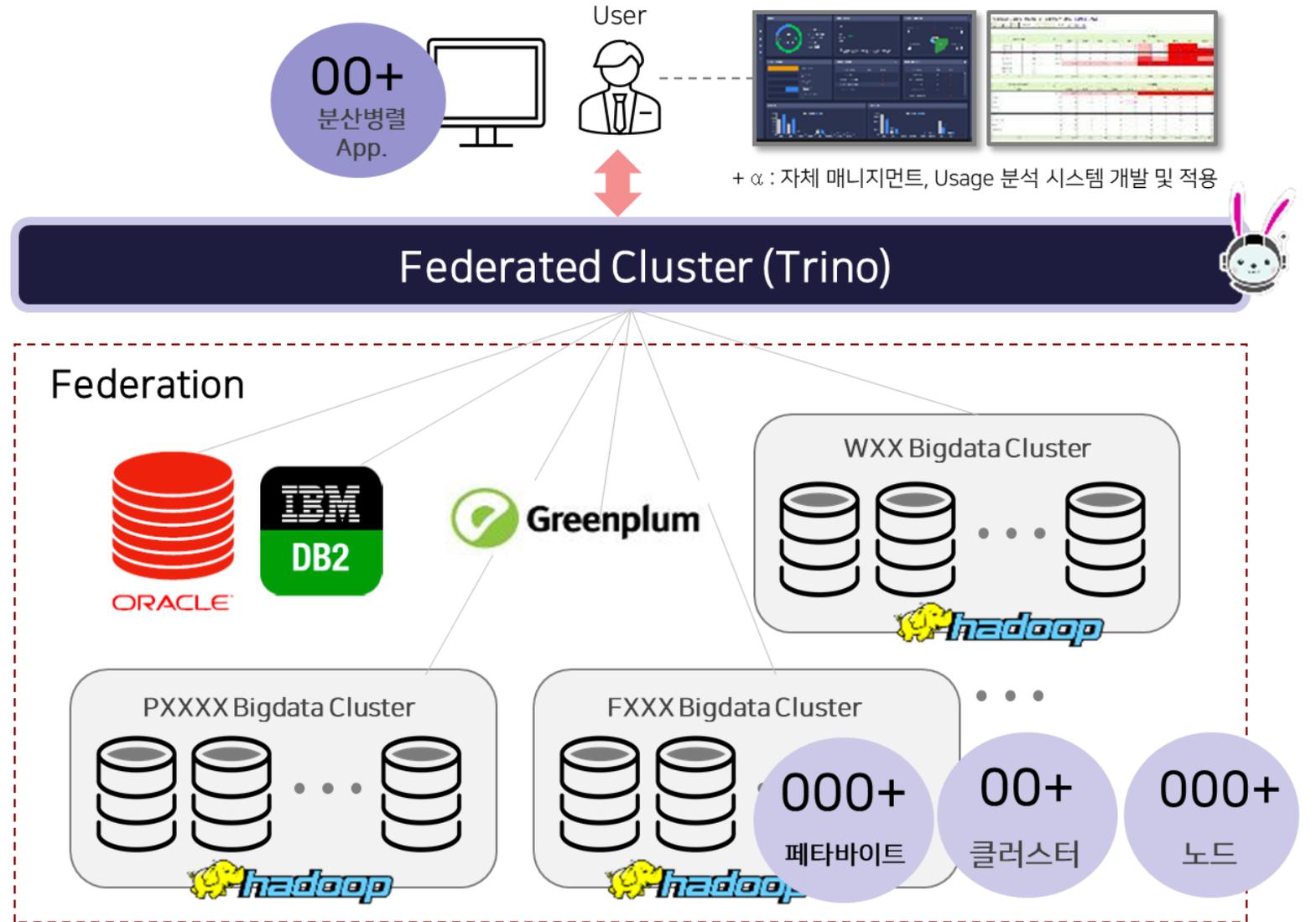
# [도입 사례] 제조 분야 : Trino 기반 자체 Federation 구현으로 분석 강화

## ✓ 글로벌 H사 Case :

이기종 데이터 소스에 대한  
대용량 데이터 핸들링 가능  
- 000M 파일/디렉토리, 00K 테이블

데이터에 대한 다양한 분석 가능,  
데이터 사용률 대폭 증가  
= 00K Trino 쿼리/일,  
TB단위 data read 쿼리 가능

다양한 공정 App. 내 필요 데이터에  
대한 손쉬운 Data preparation 제공



# Trino User / Staburst Customers

## Trino(Presto) Open Source Adoption and Contributors



## Starburst's Customer

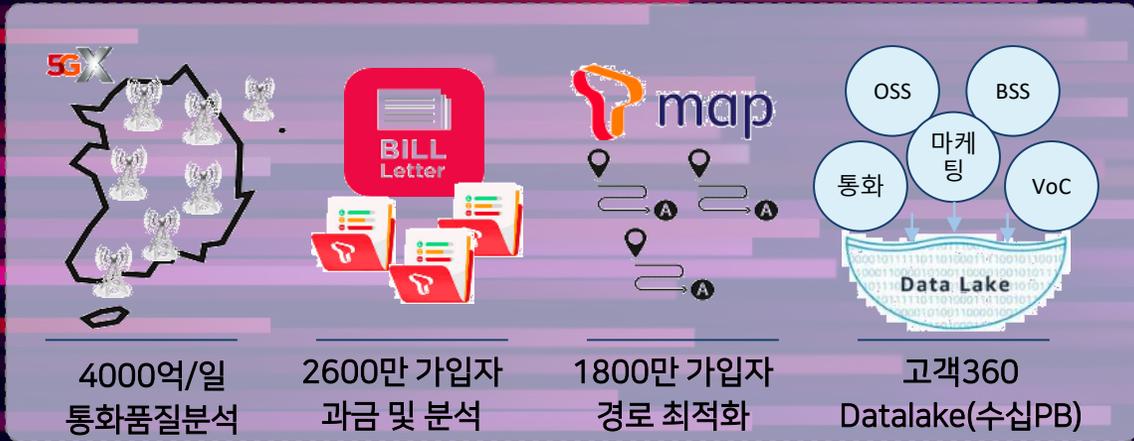


A series of white and light red lines on the left side of the slide, resembling a circuit board or data flow diagram. The lines are horizontal and vertical, with small circles at the ends, some of which are filled with a light red color. The lines extend from the left edge towards the center of the slide.

# About SK Telecom

# SK Telecom, 국내 최고의 Data 파트너

## 국내 최대 및 최난 Data 문제를 해결



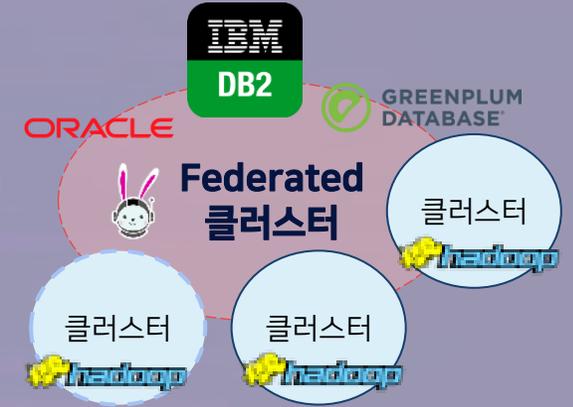
## 국내 최대/최신 데이터엔지니어링 역량 및 전문가 보유



## Trino, Hadoop 등 클러스터 구축 및 분산병렬처리 app. 개발

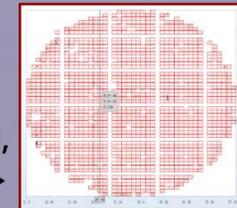
### Bigdata 클러스터 및 App 구축 요약 (총량 기준)

- Based on hadoop 3.1
- 00K vCPU, 00TB mem
- 00K Trino 쿼리/일
- TB 단위 data read 쿼리
- 총 운영 데이터 규모 약 000 PB 급



ex) 수천대 장비의 수백만 -인자 실시간 이상 감지 및 분류 ▼

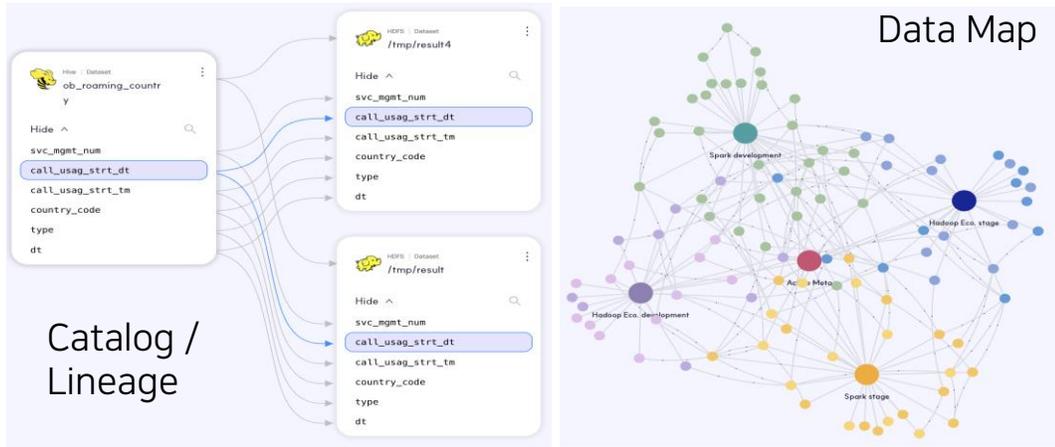
ex) 2천만 dot chart 구현, Zoom in&out Interface 제공 ▶



# Coming Soon ...

“ 모든 메타 정보에 실시간 추출 및 자동 추적으로,  
데이터 신뢰도, 품질 대폭 향상 ”

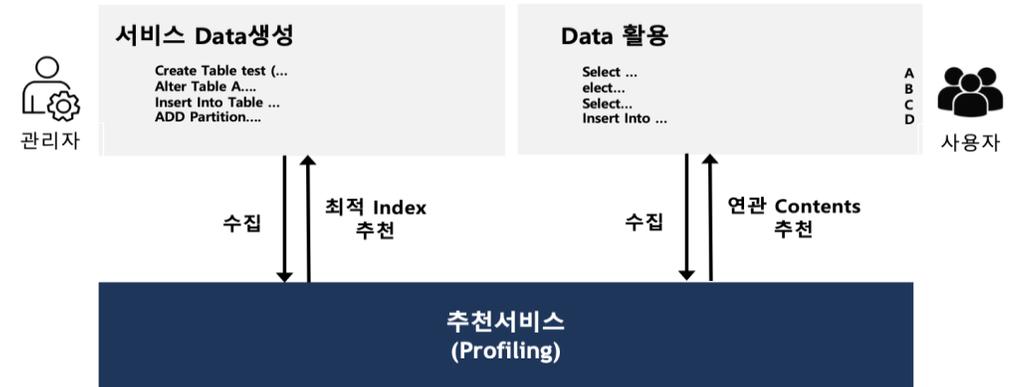
## Active Meta



- [타 기업 Case] 실시간이 아닌 배치 기준 데이터 관리, 모든 데이터 변화에 대한 추적 관리 불가
- [베타 버전 개발 완료]  
컬럼 단위 데이터 변경 사항에 대한 100% 실시간 추적 관리, 이기종 데이터 소스 작업에 대한 통합 모니터링

“ 어떤 데이터를 활용할지, 어떻게 효율적으로  
사용할지에 대한 자동 추천 서비스 ”

## Data 추천 서비스(Profiling)



- [기존] 사용자 Usage(Query문) Trend 분석 및 프로파일링에 따라 Data 검색 패턴 시기 추천/가이드
- [고도화 개발 진행 중]  
ML 기반의 사용자 데이터 사용정보 해석 및 분석을 통해 최적의 Index 구문 및 연관된 콘텐츠 추천

# SK텔레콤은 Starburst APAC 1호 파트너입니다.

## SK Telecom X Starburst 시너지 효과

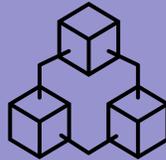
코어 엔진(Trino)에 대한  
상시적인 기술 지원  
대응 가능



대용량 데이터 병렬 처리  
기반 다양한 빅데이터 app.  
서비스 구축



기존 고객 데이터  
환경 분석 및 아키텍처  
컨설팅 제공



고객 맞춤형 Trino/  
Starburst 클러스터  
커스텀 구축 가능



국내 최고의 빅데이터  
엔지니어링 역량



글로벌에서 검증된  
Data Analytics  
Engine 개발





# Thank you.

SK Telecom(주)

starburst@sk.com

서울특별시 중구 을지로 65 T-타워

<http://sktenterprise.com>