E북:

Databricks의 데이터 관리 개론

Databricks에서 어떻게 데이터 관리 생명 주기를 간소화하는지 알아봅니다.

databricks

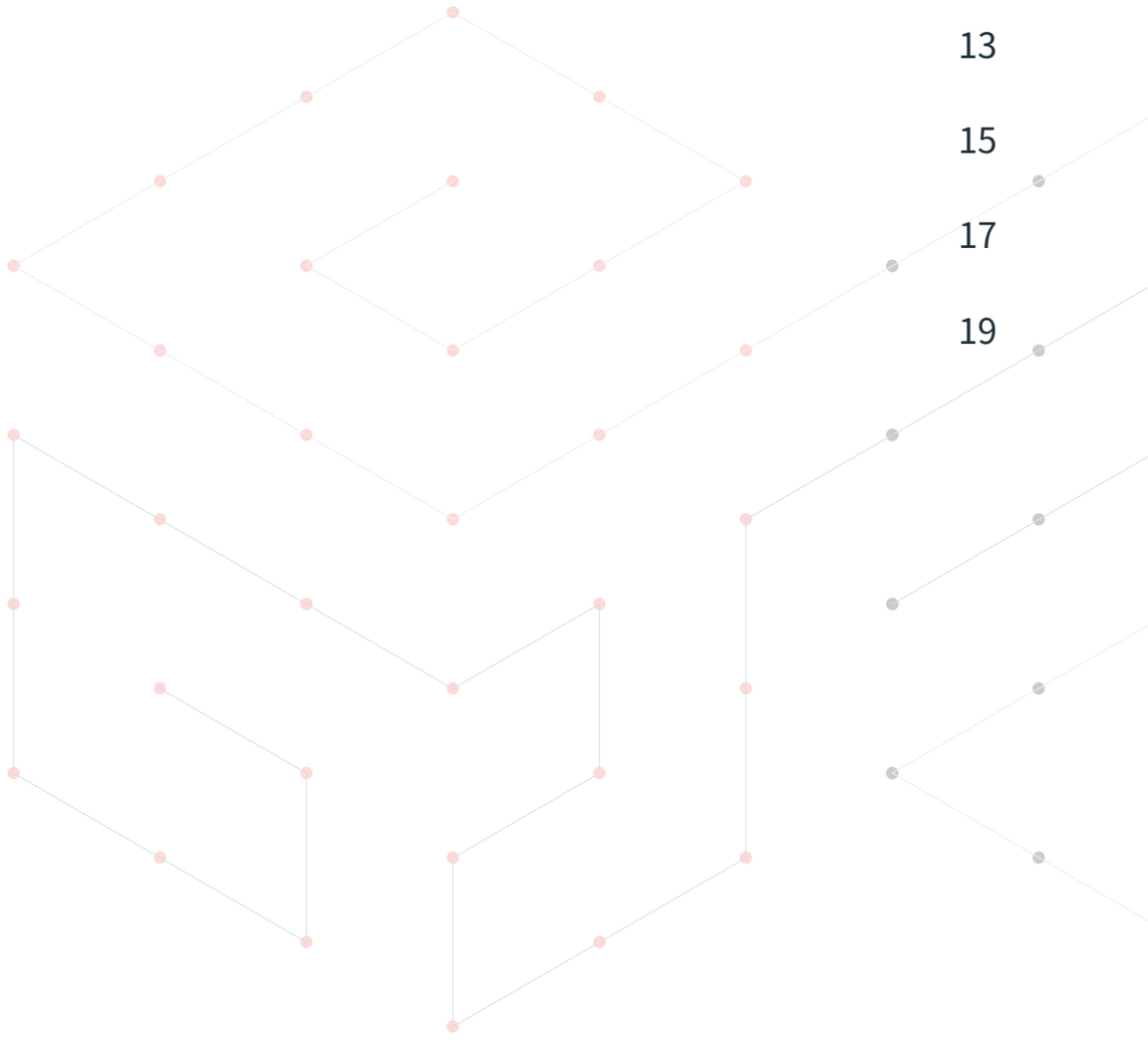
서론

원격 근무와 새로운 채널들의 증가로 변화하는 작업 환경 속에서, 데이터 관리의 중요성이 더 커지고 있습니다.

Gartner에서는 “중앙 집중형에서 분산형 작업으로 전환하려면 조직이 지금보다 빠르게, 더 많은 장소에서 데이터를 생성하고 관리할 수 있어야 한다”라고 합니다.

조직마다 지칭하는 용어는 다를 수 있어도, 데이터 관리는 수년 동안 업계 전반에 걸쳐 일반적인 관행으로 자리 잡았습니다. Databricks는 데이터 수집, 데이터 처리, 데이터 관리, 데이터 공유, 분석을 포함하여, 데이터를 전략적 가치가 있는 리소스로 관리하는 데 관련된 모든 분야를 아우르고, 이 모든 것을 비용 효율적이고 효과적이며 신뢰할 수 있는 방식으로 수행하는 것을 데이터 관리라고 생각합니다.

목차



데이터 관리의 과제

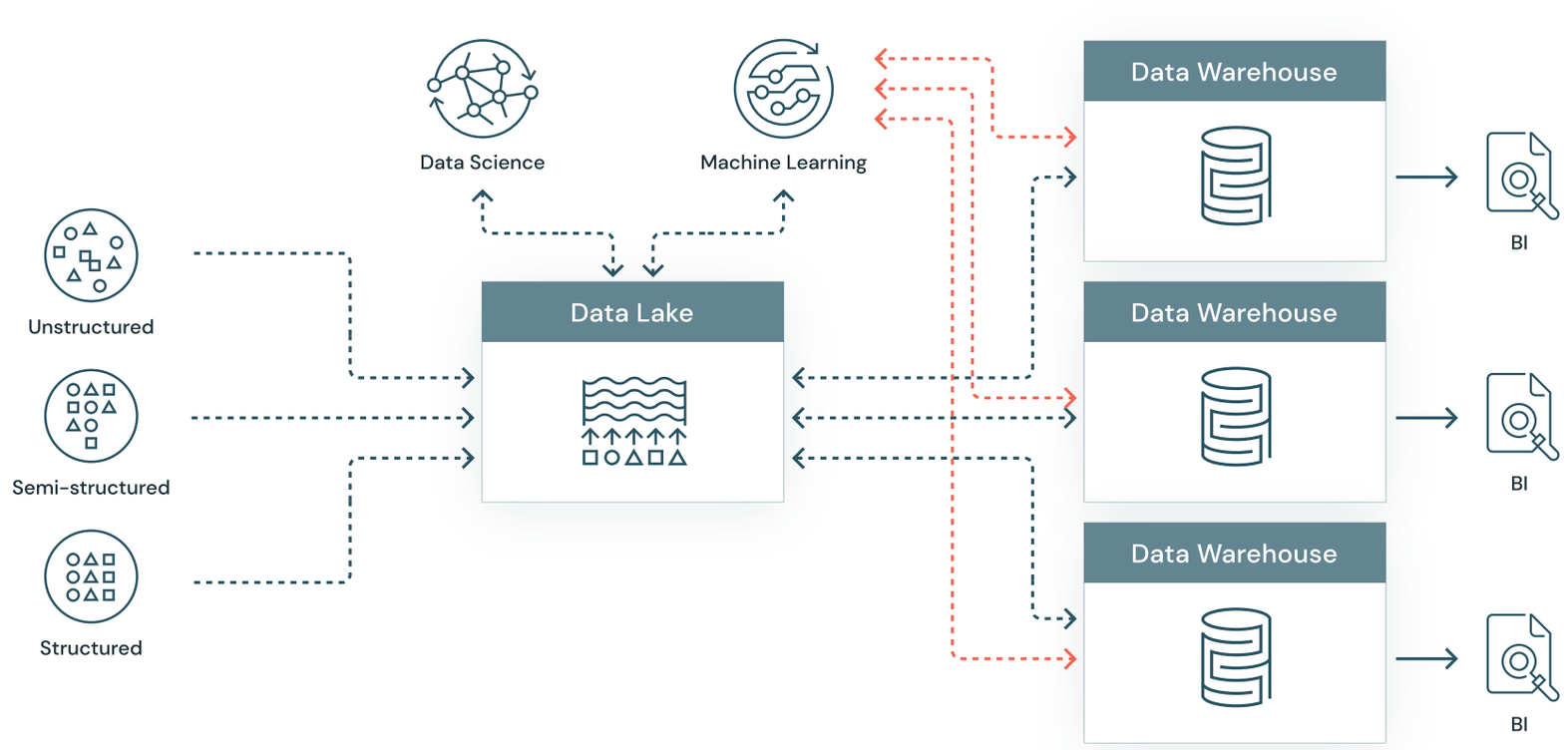
그리고 기업들은 데이터를 기반으로 제품을 결정하거나, 협업을 강화하거나, 새로운 채널로 신속하게 이동하는 등을 통해 데이터의 가치를 깨달았지만, 대부분은 데이터를 올바르게 관리하고 활용하는 데 어려움을 겪고 있습니다.

Forrester에 따르면, 기업 데이터의 최대 73%가 분석 및 의사 결정에 사용되지 않으며, 이는 비즈니스의 성공을 저해하는 지표입니다.

오늘날 기업 데이터는 대부분 데이터 레이크로 유입되며, 여기서 다운스트림 데이터 과학 및 머신 러닝 이니셔티브를 지원하기 위해 데이터 준비 및 검증을 수행합니다. 그와 동시에, 비즈니스 인텔리전스(BI)를 위해 엄청난 양의 데이터가 변환되어 다양한 다운스트림 데이터 웨어하우스로 전송됩니다. 기존 데이터 레이크는 너무 느리고 BI 워크로드에 대해 안정적이지 못하기 때문입니다.

워크로드에 따라 데이터 웨어하우스에서 데이터 레이크로 데이터를 다시 이동해야 하는 경우도 있습니다. 그리고 머신 러닝 워크로드도 데이터 웨어하우스를 쓰는 경우가 증가하고 있습니다. 이러한 종류의 데이터 관리가 어려운 이유는 데이터 레이크와 데이터 웨어하우스 간에 본질적인 차이가 있기 때문입니다.

데이터 레이크는 개방적 형식과 에코시스템을 통해 머신 러닝을 지원하는 데 큰 역할을 하지만 비즈니스 인텔리전스에 대한 지원이 부족하고 복잡한 데이터 품질 문제로 인한 어려움이 있습니다. 반면에 데이터 웨어하우스는 BI 애플리케이션에 적합하지만 머신 러닝 워크로드에 대한 지원이 제한적이며 SQL 인터페이스 전용 시스템입니다.

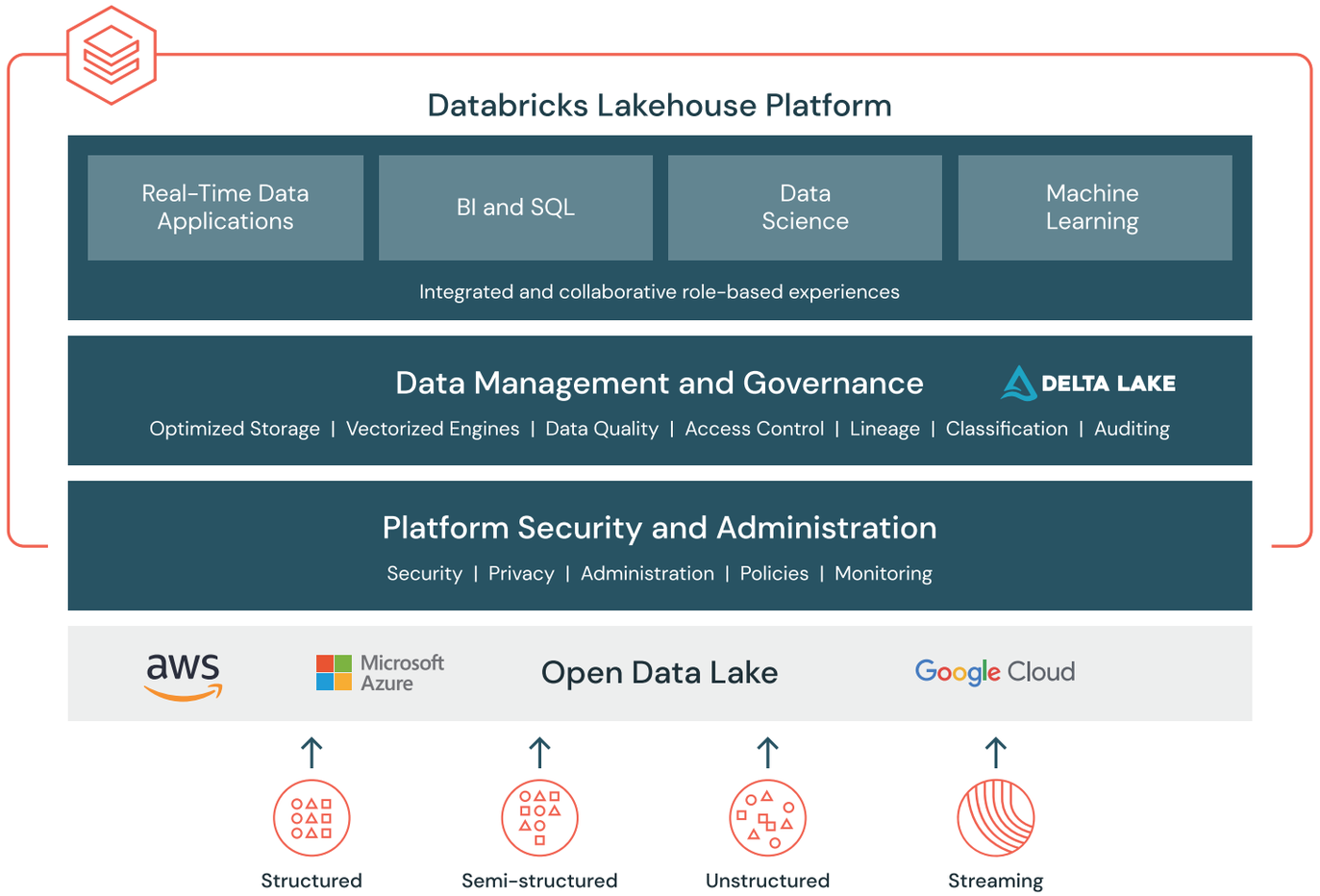


Databricks의 데이터 관리

이 시스템들을 통합하면 데이터에 대한 사고방식을 바꿀 수 있습니다. 그리고 Databricks Lakehouse Platform은 다양한 워크로드, 팀, 데이터를 통합하고, 데이터 관리 수명 주기의 모든 단계에 대해 종단 간 데이터 관리 솔루션을 제공합니다. 그리고 Delta Lake가 데이터 레이크에 안정성, 성능, 보안을 제공하고 레이크하우스의 기반을 형성하므로 데이터 엔지니어는 이러한 아키텍처 문제를 피할 수 있습니다. 그러면 Databricks의 데이터 관리 단계를 살펴보겠습니다.

더 알아보기
Databricks Lakehouse Platform

Delta Lake 더 알아보기

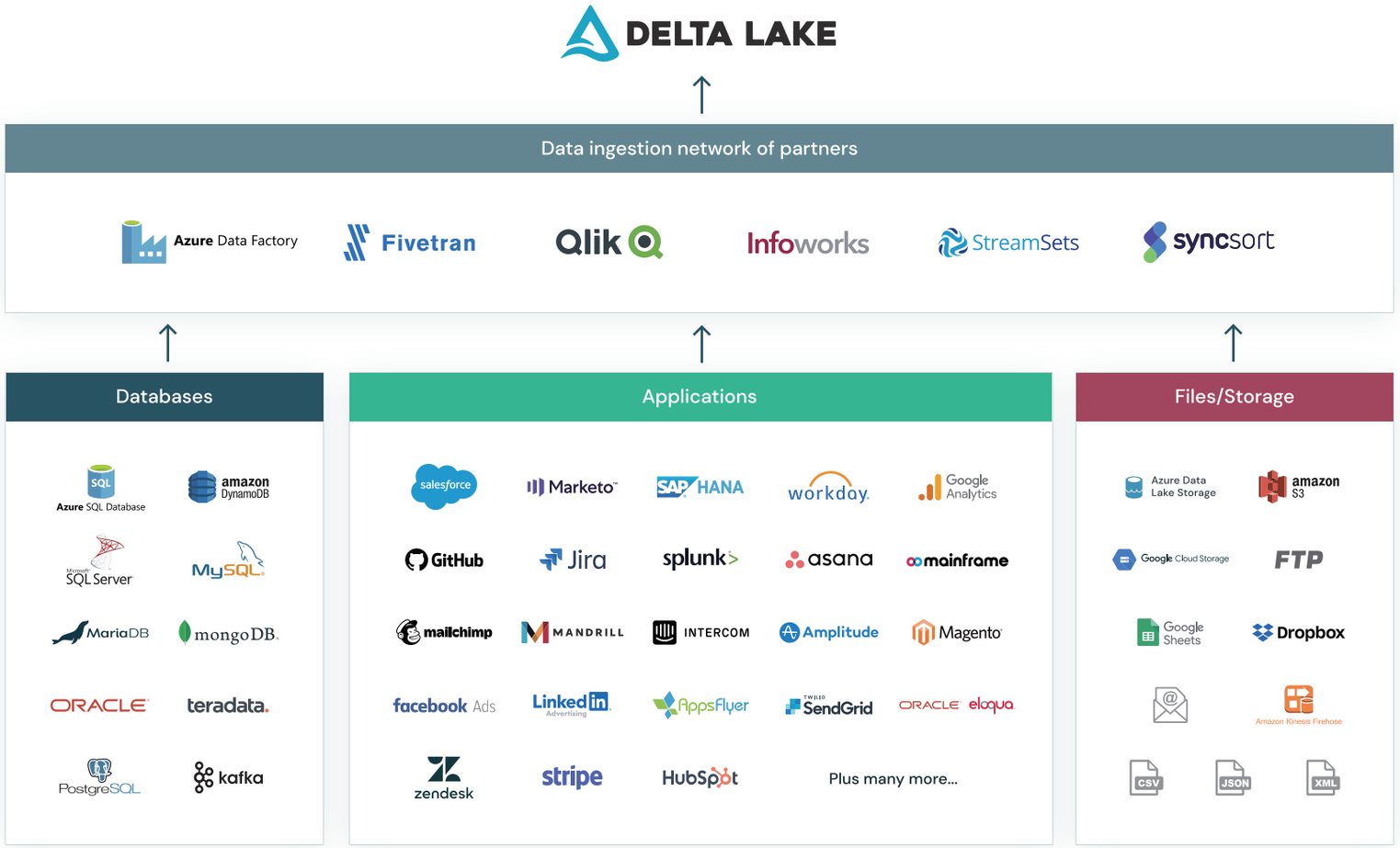


데이터 수집

오늘날 IT 기업들은 다양한 온프레미스 애플리케이션 시스템, 데이터베이스, 데이터 웨어하우스 및 SaaS 애플리케이션에서 발생한 데이터 사일로로 넘쳐나고 있습니다. 이러한 분열로 인해 분석이나 머신 러닝에 대한 새로운 사용 사례를 지원하기가 어렵습니다. 이러한 새로운 사용 사례가 등장하고 데이터의 양과 복잡성이 커지자 이를 지원하기 위해 많은 IT 팀에서는 이제 오픈 형식 스토리지 계층인 Delta Lake를 기반으로 구축된 레이크하우스 아키텍처를 사용하여 모든 데이터를 중앙 집중화하는 방법을 모색하고 있습니다.

그러나 레이크하우스 아키텍처를 지원하는 데 있어 데이터 엔지니어들이 직면한 가장 큰 문제는 다양한 시스템에서 레이크하우스로 데이터를 효율적으로 이동하는 작업입니다. Databricks는 레이크하우스로 쉽게 데이터를 수집하는 방법을 두 가지 제공합니다. 데이터 수집 파트너 네트워크를 통하거나 Auto Loader를 사용하여 Delta Lake로 쉽게 데이터를 수집하는 것입니다.

데이터 수집 파트너 네트워크를 통해 다양한 사일로 시스템에서 레이크로 데이터를 이동할 수 있습니다. 파트너들은 Databricks와의 기본 통합을 구축하여 데이터를 수집하고 Delta Lake에 저장함으로써, 데이터 팀이 쉽게 데이터에 액세스하여 작업하도록 지원합니다.

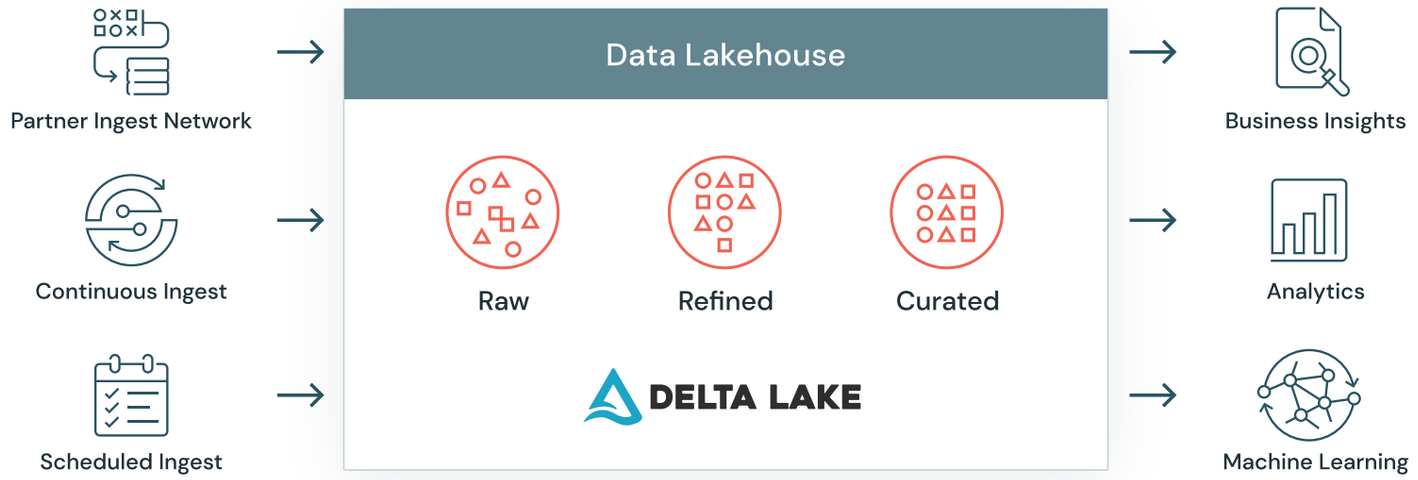


한편, 많은 IT 기업이 AWS S3, Microsoft Azure Data Lake Storage, Google Cloud Storage와 같은 클라우드 스토리지를 사용하며 다양한 시스템에서 데이터를 수집하는 방법을 구현해 왔습니다. Databricks Auto Loader는 단 한 번 만에, 대기 시간이 짧으면서도 최소한의 DevOps 작업을 통해 저렴한 비용으로 파일 소스를 최적화하고 스키마를 추론하며, 클라우드 저장소에 도착하는 새 데이터를 점진적으로 처리합니다.

데이터 엔지니어는 Auto Loader를 사용하여 소스 디렉터리 경로를 제공하고 수집 작업을 시작합니다. “cloudFiles”라는 새로운 Structured Streaming 소스는 입력 디렉터리에서 파일 이벤트를 구독하고, 새 파일이 도착하면 처리하는 파일 알림 서비스를 자동으로 설정하고, 해당 디렉터리의 기존 파일도 처리할 수 있는 옵션을 제공합니다.

Databricks의 데이터 수집

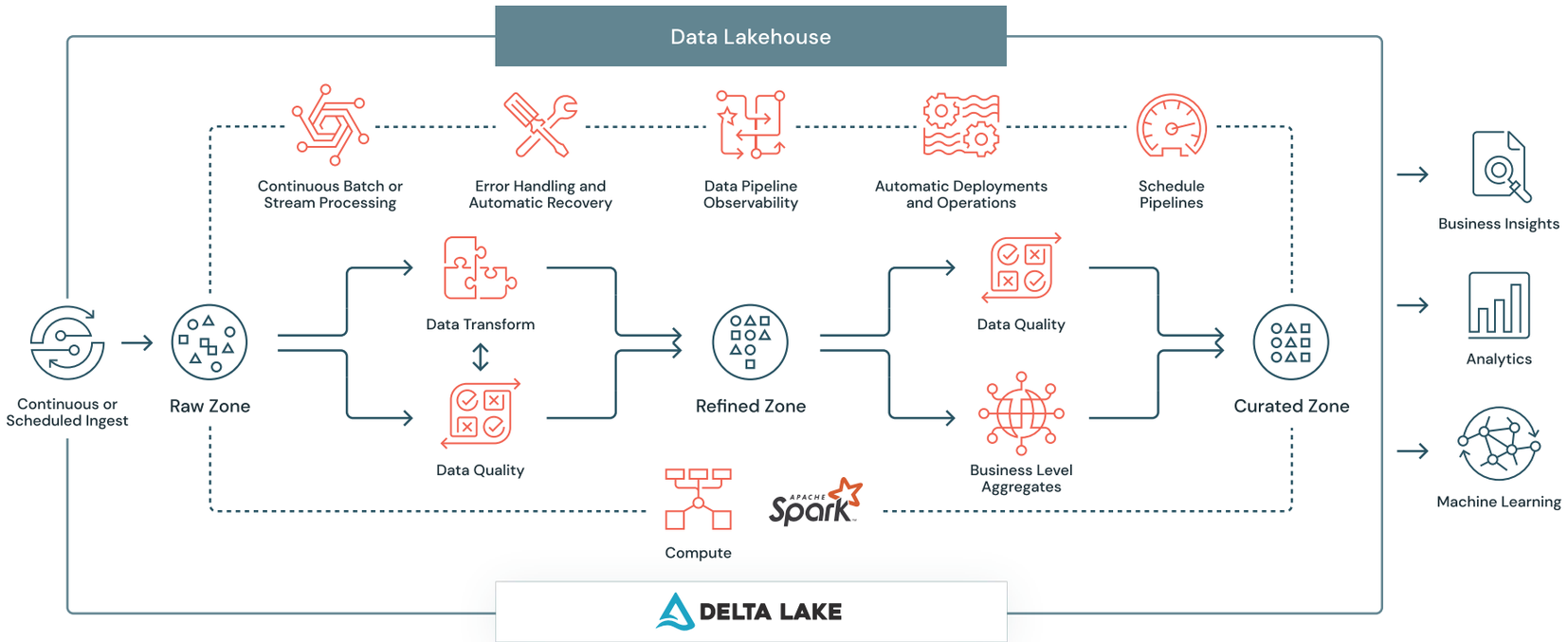
더 알아보기



머신 러닝과 분석을 통합하려면 모든 데이터를 레이크하우스로 가져오는 것이 중요합니다. 데이터 엔지니어링팀은 Databricks Auto Loader와 광범위한 파트너 통합 기술을 통해, 어떤 데이터 타입도 데이터 레이크로 효율적으로 이동할 수 있습니다.

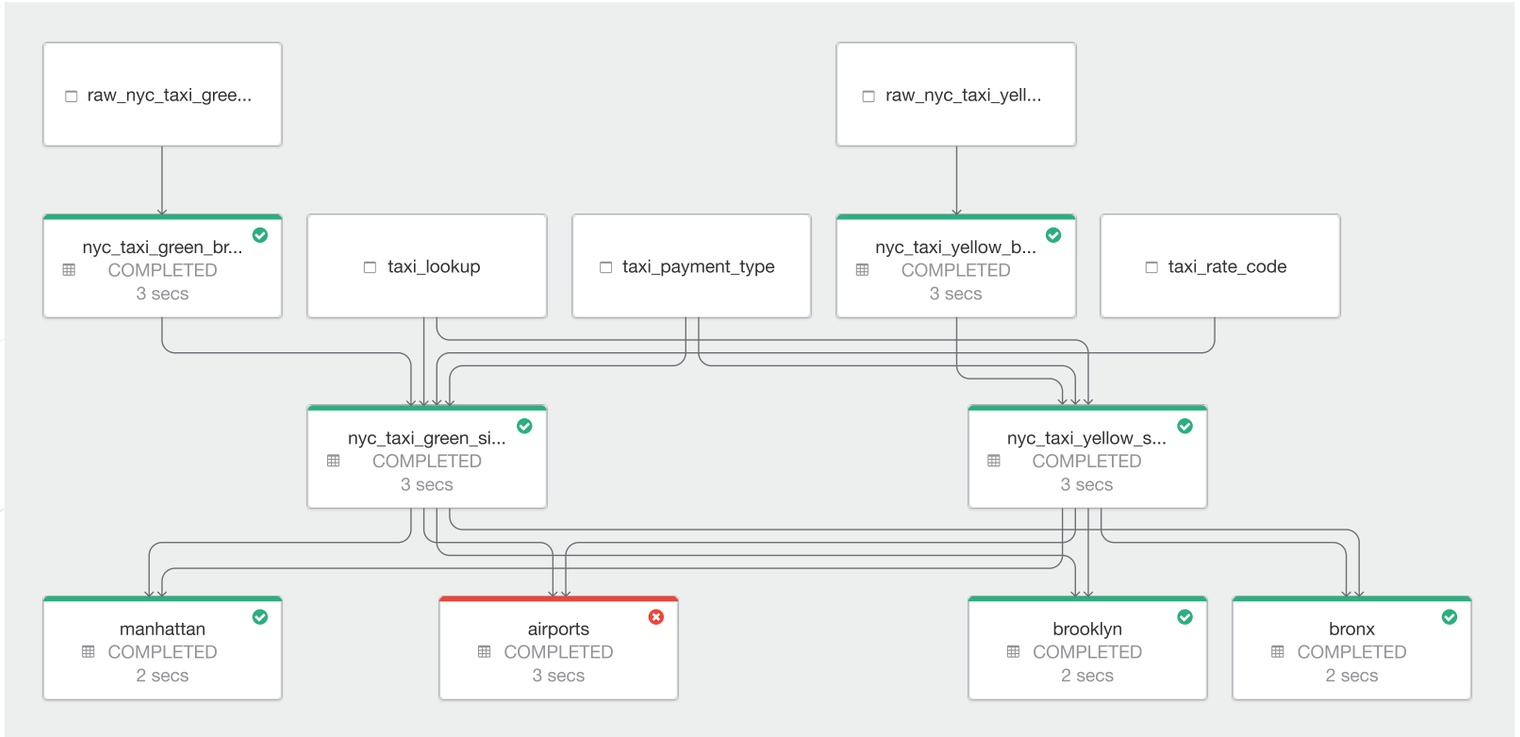
데이터 변환, 품질 및 처리

데이터를 레이크하우스로 이동하면 데이터 관리 문제 중 하나가 해결됩니다. 하지만 데이터 애널리스트나 데이터 과학자가 데이터를 사용하려면 데이터도 깨끗하고 신뢰할 수 있는 소스로 변환해야 합니다. 오래되었거나 신뢰할 수 없는 데이터는 실수나 오류, 인사이트에 대한 불신으로 이어질 수 있기 때문에 이는 중요한 단계입니다.



데이터 엔지니어들은 복잡하고 다양한 데이터를 정리하여 분석, 보고하거나 머신 러닝에 적합한 형식으로 변환하는 힘들고 까다로운 작업을 수행합니다. 이를 위해서 데이터 엔지니어는 데이터 인프라 플랫폼의 내/외부를 알아야 하며, 다양한 언어로 복잡한 쿼리(변환)를 구축하고 프로덕션을 위해 쿼리를 연결해야 합니다. 데이터 관리 단계에서 이러한 복잡성이 존재하기 때문에 많은 조직에서 다운스트림 분석, 데이터 과학 및 머신 러닝에 대한 역량을 제대로 발휘하지 못합니다.

이러한 복잡성을 제거하기 위해, Databricks Delta Live Tables(DLT)는 데이터 엔지니어링팀에게 SQL 또는 Python으로 선언적 데이터 파이프라인을 구축할 수 있는 대규모 확장 가능한 ETL 프레임워크를 제공합니다. 데이터 엔지니어는 DLT를 통해 인라인 데이터 품질 매개변수를 적용하고, 여러 클라우드에 걸쳐 안전한 완전 관리형 레이크하우스 플랫폼에서의 데이터 파이프라인 운영에 대한 자세한 정보를 확보함으로써 거버넌스와 규정 준수 상태를 관리할 수 있습니다.



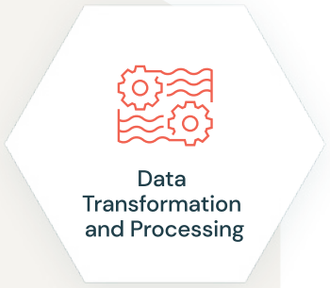
DLT는 ETL을 생성, 표준화 및 유지 관리하는 간단한 방법을 제공합니다. DLT 데이터 파이프라인은 데이터나 코드, 환경 변화에 맞추어 자동으로 조정되므로, 데이터 엔지니어는 변환 중인 데이터를 개발, 검증 및 테스트하는 데 집중할 수 있습니다. 데이터 엔지니어는 신뢰할 수 있는 데이터를 제공하기 위해 데이터 파이프라인 내에서 예상되는 데이터 품질에 대한 규칙을 정의합니다. DLT를 통해 데이터 품질을 지속해서 분석하고 모니터링하여 부정확하고 일관성 없는 데이터의 확산을 줄일 수 있습니다.

"Delta Live Tables는은 우리 팀이 데이터를 대규모로 관리하는 데 드는 시간과 노력을 절약하는 데 도움이 되었습니다. Databricks는 기존 레이크하우스 아키텍처를 보강하는 이 기능으로 우리 같은 회사에서 중요하게 여기는 ETL 및 데이터 웨어하우스 시장에 파괴적인 혁신을 일으키고 있습니다."

— Shell 데이터 과학 총괄 매니저, Dan Jeavons

성공적인 데이터 엔지니어링 구현의 핵심은 엔지니어가 ETL 개발 및 테스트에 집중하게 하고 인프라 구축에 소요되는 시간은 줄이는 것입니다. Delta Live Tables는 파이프라인을 실행하는 동안 기본 데이터 파이프라인의 정의를 추출합니다. 즉, 파이프라인 실행 시 DLT는 파이프라인을 최적화하고, 기본 데이터 파이프라인 쿼리에 대한 실행 그래프를 자동으로 구축하고, 동적 리소스로 인프라를 관리하고, 성능, 대기 시간, 품질 등에 대한 전체 파이프라인 상태에 대하여 파이프라인 전체적인 가시성을 확보할 수 있는 시각적 그래프를 제공합니다.

데이터 엔지니어는 이 모든 DLT 구성 요소를 통해 머신 러닝 및 분석을 위한 고품질 데이터를 변환, 정리하여 제공하는 데에만 집중할 수 있습니다.

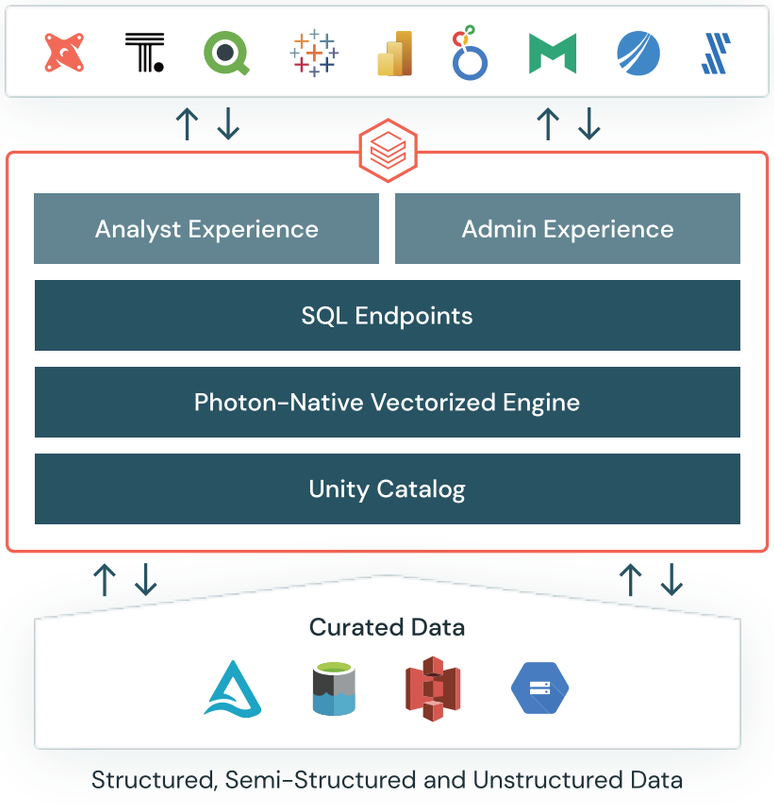


Delta Live Tables를 통한
Databricks의 데이터 변환

더 알아보기

데이터 분석

이제 데이터 애널리스트는 데이터를 사용해서 비즈니스 의사 결정을 내리는 데 필요한 인사이트를 얻을 수 있습니다. 일반적으로 데이터 레이크 내에서 잘 구성된 데이터에 접근하려면 애널리스트가 Apache Spark™를 활용하거나 개발자 인터페이스를 사용하여 데이터에 접근해야 합니다. **Databricks SQL**은 데이터 애널리스트가 SQL 네이티브 경험을 통한 심층 분석을 수행하여 멀티클라우드 레이크하우스 아키텍처에서 BI 및 SQL 워크로드를 실행하도록 지원함으로써 레이크하우스 접근과 쿼리를 단순화합니다. Databricks SQL은 데이터 애널리스트와 데이터 과학자가 Databricks 내에서 직접 데이터 레이크 데이터를 쿼리할 수 있는 SQL 네이티브 인터페이스로 기존 BI 도구를 보완합니다.



전용 SQL 작업 공간은 레이크하우스에서 임시 쿼리를 실행하고, 자세한 시각화를 생성하여 다양한 관점에서 쿼리를 탐색하며, 이러한 시각화를 조직 전체의 관계자들과 공유할 수 있는 드래그 앤 드롭 방식 대시보드로 구성할 수 있어 데이터 애널리스트에게는 친숙합니다. 애널리스트는 작업 공간에서 스키마를 탐색하고, 재사용을 위해 쿼리를 코드 조각으로 저장하고, 쿼리의 자동 새로 고침을 예약할 수 있습니다.

고객은 Databricks SQL 엔드포인트를 통해 선호하는 BI 도구를 레이크하우스에 연결하여 기존 투자를 극대화할 수 있습니다. 재설계되고 최적화된 커넥터는 데이터 레이크에서 빠른 성능, 짧은 대기 시간 및 높은 사용자 동시성을 보장합니다. 따라서 애널리스트는 ETL 및 데이터 사일로를 최소화하면서 통합 정보 출처에서 데이터 작업을 처리하기에 가장 적합한 도구를 사용할 수 있게 됩니다.

"조직에는 그 어느 때보다도 속도와 민첩성을 적응시킬 수 있는 데이터 전략이 필요합니다. 조직들이 데이터를 클라우드로 빠르게 이동하면서, 데이터 레이크의 분석 작업에 대한 관심이 커지고 있습니다. Databricks SQL는 고객에게 필요한 성능, 안정성 및 규모를 갖추고,방대한 데이터에서 얻은 인사이트를 활용할 수 있는 완전히 새로운 경험을 제공합니다. 우리는 Databricks와 협력하여 그 기회를 실현하게 된 것을 자랑스럽게 생각합니다"

—Tableau 최고 제품 책임자, Francois Ajenstat

마지막으로, 관리자는 거버넌스 및 관리를 위해 테이블에 SQL 데이터 액세스 제어를 적용함으로써 레이크하우스 전체적인 분석 작업에 대해 데이터 사용 및 액세스 방식을 세밀하게 제어하고, 자세한 가시성을 확보할 수 있습니다. 관리자는 성능, 각 쿼리가 실행된 위치, 쿼리가 실행된 시간 및 워크로드를 실행한 사용자 등을 이해하기 위해 실행한 모든 쿼리를 비롯해 Databricks SQL 사용 내역에 대한 가시성을 얻게 됩니다. 이 모든 정보는 캡처되고, 관리자가 쉽게 분류하고 문제를 해결하고 성능을 이해할 수 있도록 제공됩니다.

Databricks SQL을 통한
Databricks의 데이터 분석

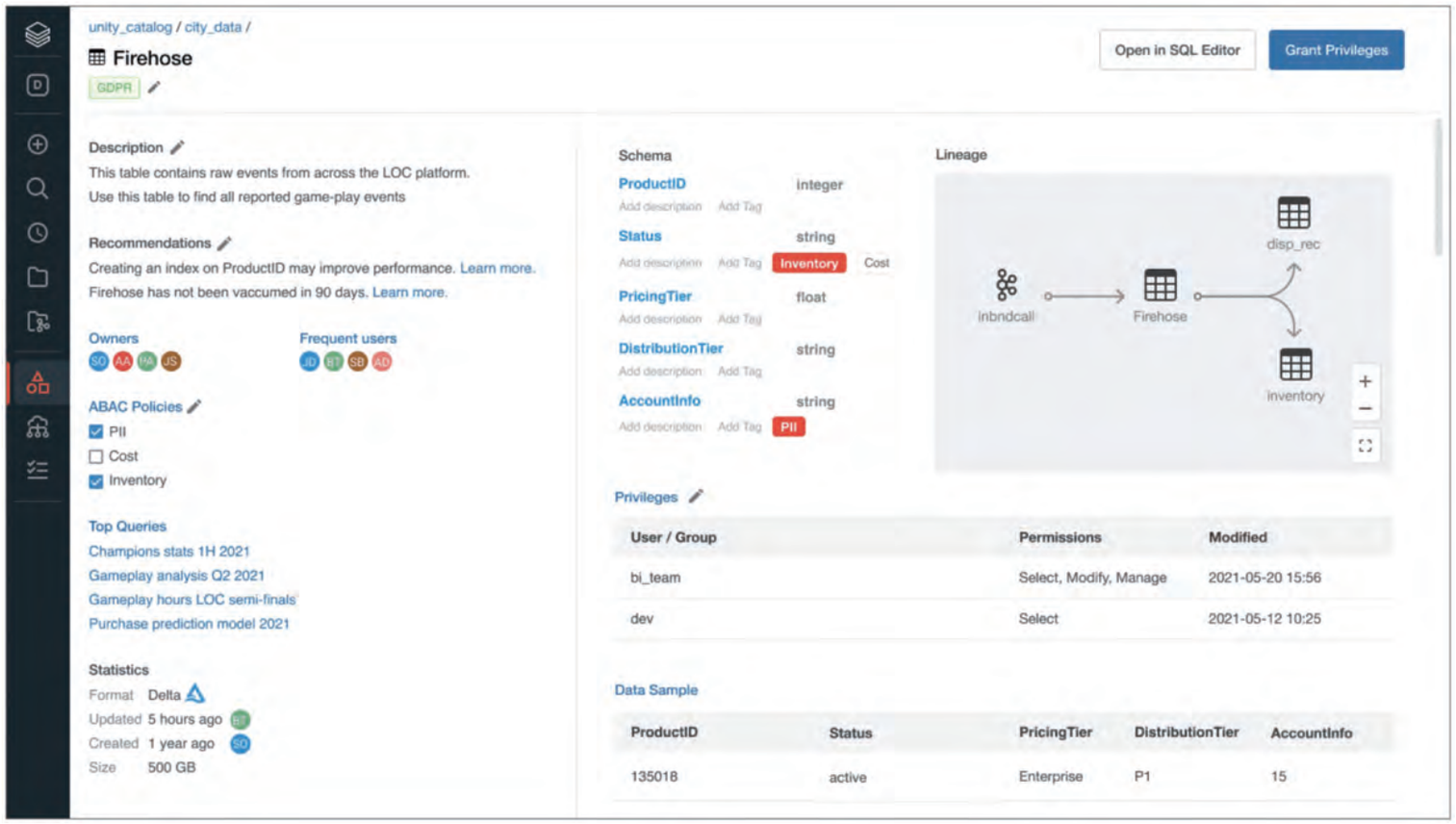
더 알아보기

데이터 거버넌스

분석 및 머신 러닝을 해결하기 위한 수단으로 데이터 레이크를 구축하고, 데이터 거버넌스를 나중에 고려하는 조직이 많습니다. 그러나 레이크하우스 아키텍처가 빠른 속도로 도입되면서 조직 전반에 데이터가 민주화되고 액세스되고 있습니다. 관리자는 IAM 역할이나 RBAC와 같은 클라우드 공급 업체별 보안 제어 및 파일 지향 액세스 제어에 의존하여 데이터를 관리하는 방식으로 데이터 레이크를 관리했습니다. 그러나 이 기술 보안 메커니즘은 데이터 거버넌스 및 데이터 팀의 요구 사항을 해결할 수 없습니다. 데이터 거버넌스는 조직 내에서 데이터 자산에 대한 권한을 가지고 제어할 수 있는 사람과 이러한 자산이 사용되는 방식을 정의합니다.

데이터를 보다 효과적으로 관리하기 위해, Databricks **Unity Catalog**는 표준 ANSI SQL나 간단한 UI를 사용하여 레이크하우스에 세부적인 거버넌스와 보안을 제공함으로써, 데이터 관리자가 레이크하우스를 안전하게 개방하여 내부 전체에서 광범위하게 사용할 수 있도록 합니다. 데이터 관리자는 SQL 기반 인터페이스를 통해 속성 기반 액세스 제어를 적용하여 동일한 속성을 가진 유사한 데이터 개체에 태그를 지정하고, 정책을 적용할 수 있습니다. 또한, 데이터 관리자는 동일한 인터페이스 내에서 ML 모델, 대시보드, 외부 데이터 소스와 같은 다른 데이터 자산에 대해 강력한 거버넌스를 적용할 수 있습니다.

조직에서 데이터 플랫폼을 온프레미스에서 클라우드로 현대화함에 따라, 많은 조직이 데이터 관리를 위해 단일 클라우드 환경 너머로 이동하고 있습니다. 대신, 여러 지역에 걸쳐 AWS, Azure 및 GCP 등 클라우드 제공업체 주요 3사와 협력하는 멀티클라우드 전략을 선택하고 있습니다. 조직에서 데이터를 민주화하려면 멀티클라우드 플랫폼, 스토리지, 기타 카탈로그에서 이 모든 데이터를 관리하는 문제를 해결해야 합니다. Unity Catalog는 데이터 추적을 중앙에서 관리, 추적 및 감사할 수 있는 안전한 단일 제어 지점 역할을 합니다.



Data Governance

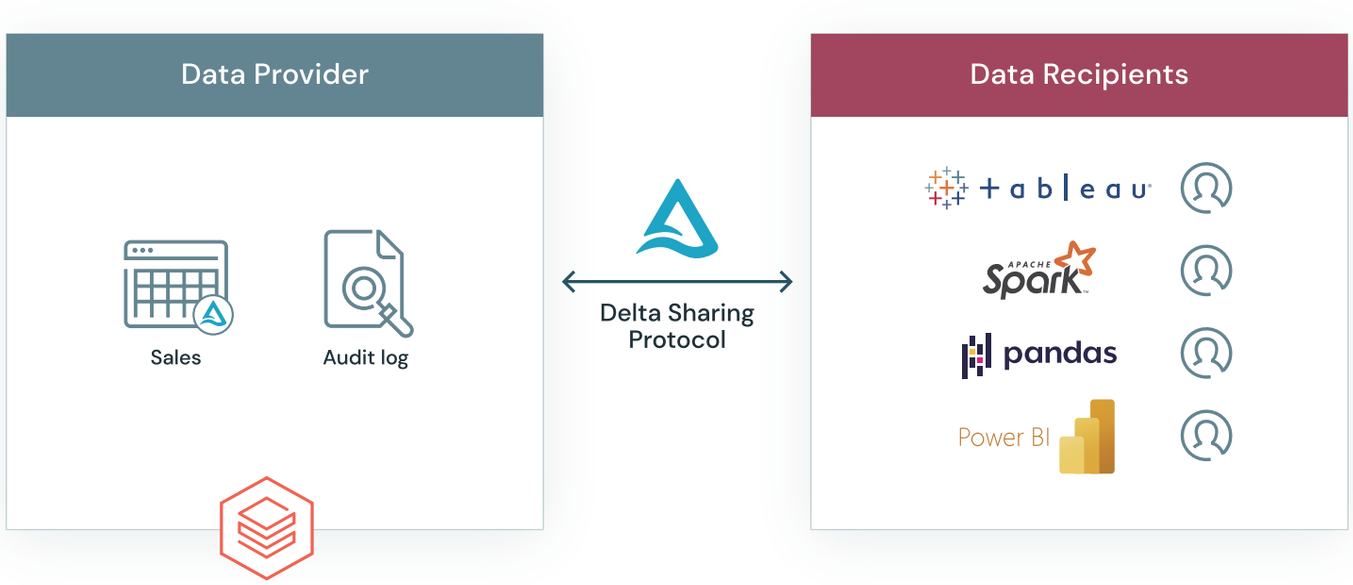
Unity Catalog를 통한
Databricks의 데이터 거버넌스

더 알아보기

마지막으로, Unity Catalog는 중앙 한 지점에서 데이터 자산을 쉽게 검색하여 기술하고, 감사 및 관리를 지원합니다. 데이터 관리자는 모든 권한을 시각적으로 설정하거나 검토할 수 있으며, 카탈로그에서는 각 데이터 자산이 어떻게 생성 및 액세스 되었는지를 보여주는 감사 및 계보 정보를 캡처합니다. 데이터 관리자는 데이터 계보, 역할 기반 보안 정책, 테이블 또는 열 수준 태그, 중앙 감사 기능을 사용하여 레이크하우스에서 직접 데이터 액세스를 관리하고 보호하고, 규정 준수 및 개인 정보 보호 요구 사항을 해결할 수 있습니다. UI는 협업이 용이하도록 설계되어, 데이터 사용자가 각 자산을 문서화하고 누가 사용하는지 확인할 수 있습니다.

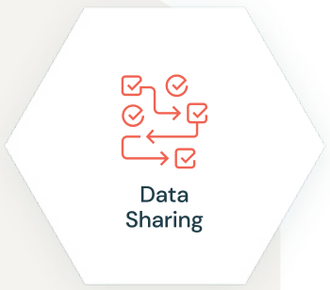
데이터 공유

조직에서 레이크하우스 아키텍처를 구축하고 있기 때문에, 정제되고 신뢰할 수 있는 데이터의 공급과 수요는 분석과 머신 러닝 외에도 발생합니다. 오늘날의 데이터 중심 경제를 살아가는 대부분 IT 리더가 이미 알고 있겠지만, 더 의미 있는 인사이트를 얻으려면 조직 전반에 걸쳐 고객, 파트너 및 공급업체와 데이터를 공유하는 것이 성공을 좌우합니다. 그러나 많은 조직에서 기준이 없고, 시스템이나 도구로 구성된 대규모 에코시스템에서 대량의 데이터 세트로 작업할 때 협업하기가 어렵고, 데이터 공유 시 위험을 완화하지 못해 데이터 공유에 실패합니다. 안전한 실시간 데이터 공유를 위한 개방형 프로토콜인 Delta Sharing은 이러한 문제를 해결하기 위해 조직 간 데이터 공유를 단순화합니다.



Databricks 레이크하우스 플랫폼과 통합된 Delta Sharing을 통해, 공급자는 다른 서버나 클라우드 개체 저장소에 복사하지 않고도 기존 데이터나 워크플로를 쉽게 사용하여 Delta Lake 또는 Apache Parquet 형식의 라이브 데이터를 안전하게 공유할 수 있습니다. 데이터 소비자는 Delta Sharing의 개방형 프로토콜을 통해, 오픈 소스 클라이언트(예: pandas) 또는 상용 BI, 분석 또는 거버넌스 클라이언트를 사용하여 공유 데이터에 직접 쉽게 액세스할 수 있습니다. 데이터 소비자가 공급자와 동일한 플랫폼에 있을 필요가 없습니다. 프로토콜은 개인 정보 보호 및 규정 준수 요구 사항에 맞추어 설계됩니다. Delta Sharing은 단일 적용 지점에서 공유 데이터에 대한 액세스 권한을 부여하고 추적. 감사에 필요한 보안 및 개인 정보 제어 기능을 관리자에게 제공합니다.

Delta Sharing은 안전한 데이터 공유를 위한 업계 최초의 개방형 프로토콜로써, 사용하는 컴퓨팅 플랫폼과 관계없이 간단하게 다른 조직과 데이터를 공유할 수 있습니다. Delta Sharing은 Apache Parquet 및 Delta Lake 형식을 기반으로 하는 기존의 대규모 데이터 세트를 원활하게 공유할 수 있으며, Delta Lake를 지원하는 기존 엔진을 쉽게 구현할 수 있도록 Delta Lake 오픈 소스 프로젝트에서 지원됩니다.



Delta Sharing을 통한
Databricks의 데이터 공유

Learn more

databricks

결론

새로운 작업 방식으로 전환하고, 새로운 기술을 채택하고, 운영을 확장해나가는 동안, 현대화의 병목 현상을 제거하려면 효과적인 데이터 관리에 대한 투자가 매우 중요합니다. Databricks Lakehouse Platform을 사용하면, 수집에서 분석에 이르기까지 데이터를 관리할 수 있고, 데이터, 분석 및 AI를 진정으로 통합할 수 있습니다.

Databricks의 데이터 관리에 대해 더 알아보세요.
바로 보기

데모 허브에 방문하세요. 데모 영상 보기

Databricks 소개

Databricks는 데이터 및 AI 회사입니다. Comcast, Condé Nast, H&M 및 Fortune 선정 500대 기업의 40% 이상을 포함하여 전 세계적으로 5,000개 이상의 기업이 Databricks 레이크하우스 플랫폼을 사용하여 데이터, 분석 및 AI를 통합합니다. Databricks는 샌프란시스코에 본사가 있으며 전 세계에 지사를 두고 있습니다. Apache Spark™, Delta Lake 및 MLflow의 원 제작자가 설립한 Databricks는 데이터팀이 세계의 어려운 문제들을 해결할 수 있도록 지원하겠다는 사명을 가지고 있습니다. Databricks에 대해 자세히 알고싶다면 Twitter, LinkedIn,Facebook을 팔로우하세요.

databricks