**AZ-레이크하우스 참조 아키텍처

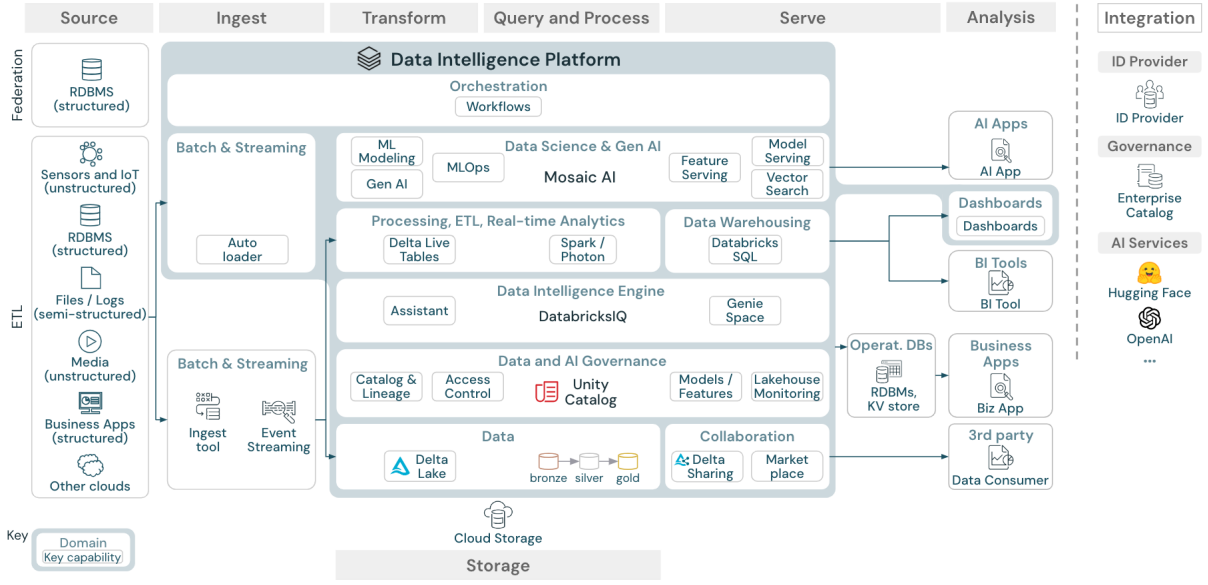
레이크하우스 참조 아키텍처 다운로드

이 문서의 내용

이 문서에서는 데이터 원본, 수집, 변환, 쿼리 및 처리, 서비스, 분석/출력 및 스토리지 측면에서 Lakehouse에 대한 아키텍처 지침을 설명합니다.

각 참조 아키텍처에는 11 x 17(A3) 형식의 다운로드 가능한 PDF가 있습니다.

제네릭 참조 아키텍처



다운로드: Databricks에 대한 일반 레이크하우스 참조 아키텍처(PDF)

참조 아키텍처 구성

참조 아키텍처는 스윔 레인 *원본, 수집, 변환, 쿼리 및 프로세스, 서비스, 분석 및 스토리지를 따라 구성됩니다*.

- **Source**

  아키텍처는 반구조적 데이터와 구조화되지 않은 데이터(센서 및 IoT, 미디어, 파일/로그) 및 구조적 데이터(RDBMS, 비즈니스 애플리케이션)를 구분합니다. RDBMS(SQL 원본)는 레이크하우스 페더레이션을 통해 ETL 없이 레이크하우스 및 Unity 카탈로그에 통합할 수도 있습니다. 또한 다른 클라우드 공급자에서 데이터를 로드할 수도 있습니다.

- **수집**

  일괄 처리 또는 스트리밍을 통해 레이크하우스로 데이터를 수집할 수 있습니다.

  - 클라우드 스토리지에 배달된 파일은 Databricks 자동 로더를 사용하여 직접 로드할 수 있습니다.
  - 엔터프라이즈 애플리케이션에서 Delta Lake로 데이터를 일괄 수집하기 위해 Databricks Lakehouse는 이러한 레코드 시스템에 대한 특정 어댑터를 사용하는 파트너 수집 도구를 사용합니다.
  - 스트리밍 이벤트는 Databricks 구조적 스트리밍을 사용하여 Kafka와 같은 이벤트 스트리밍 시스템에서 직접 수집할 수 있습니다. 스트리밍 원본은 센서, IoT 또는 변경 데이터 캡처 프로세스일 수 있습니다.

- **스토리지**

  데이터는 일반적으로 ETL 파이프라인이 medallion 아키텍처를 사용하여 데이터를 델타 파일/테이블로 큐레이팅된 방식으로 저장하는 클라우드 스토리지 시스템에 저장됩니다.

- **변환** 및 **쿼리 및 처리**

  Databricks Lakehouse는 모든 변환 및 쿼리에 Apache Spark 및 Photon 엔진을 사용합니다.

  단순성으로 인해 선언적 프레임워크 DLT(Delta Live Tables)는 안정적이고 유지 관리 가능하며 테스트 가능한 데이터 처리 파이프라인을 빌드하는 데 적합합니다.

  Apache Spark 및 Photon에서 제공하는 Databricks Data Intelligence 플랫폼은 SQL 웨어하우스를 통한 SQL 쿼리와 작업 영역 클러스터를 통한 SQL, Python 및 Scala 워크로드의 두 가지 유형의 워크로드를 모두 지원합니다.

  데이터 과학(ML 모델링 및 Gen AI)의 경우 Databricks AI 및 Machine Learning 플랫폼은 AutoML 및 코딩 ML 작업을 위한 특수한 ML 런타임을 제공합니다. 모든 데이터 과학 및 MLOps 워크플로는 MLflow에서 가장 잘 지원됩니다.

- **제공하다**

DWH 및 BI 사용 사례의 경우 Databricks Lakehouse는 Databricks SQL, SQL 웨어하우스에서 제공하는 데이터 웨어하우스 및 서버리스 SQL 웨어하우스를 제공합니다.

기계 학습 의 경우 모델 제공 은 Databricks 컨트롤 플레인에서 호스트되는 기능을 제공하는 확장성 있는 실시간 엔터프라이즈급 모델입니다.

운영 데이터베이스: 운영 데이터베이스와 같은 외부 시스템을 사용하여 최종 데이터 제품을 저장하고 사용자 애플리케이션에 제공할 수 있습니다.

협업: 비즈니스 파트너는 델타 공유를 통해 필요한 데이터에 안전하게 액세스할 수 있습니다. 델타 공유 를 기반으로 하는 Databricks Marketplace 는 데이터 제품을 교환하기 위한 공개 포럼입니다.

- **분석**

  최종 비즈니스 애플리케이션은 이 스윔 레인에 있습니다. 예를 들어 실시간 유추를 위해 Mosaic AI Model Service에 연결된 AI 애플리케이션 또는 레이크하우스에서 운영 데이터베이스로 푸시된 데이터에 액세스하는 애플리케이션과 같은 사용자 지정 클라이언트가 있습니다.

  BI 사용 사례의 경우 분석가는 일반적으로 BI 도구를 사용하여 데이터 웨어하우스에 액세스합니다. SQL 개발자는 쿼리 및 대시보드에 Databricks SQL 편집기 (다이어그램에 표시되지 않음)를 추가로 사용할 수 있습니다.

  데이터 인텔리전스 플랫폼은 데이터 시각화를 빌드하고 인사이트를 공유하는 대시보드도 제공합니다.

워크로드에 대한 기능

또한 Databricks Lakehouse에는 모든 워크로드를 지원하는 관리 기능이 제공됩니다.

- **데이터 및 AI 거버넌스**

  Databricks Data Intelligence Platform의 중앙 데이터 및 AI 거버넌스 시스템은 Unity 카탈로그입니다. Unity 카탈로그는 모든 작업 영역에 적용되는 데이터 액세스 정책을 관리할 수 있는 단일 위치를 제공하며 테이블, 볼륨, 기능(기능 저장소) 및 모델(모델 레지스트리)과 같이 레이크하우스에서 만들거나 사용하는 모든 자산을 지원합니다. Unity 카탈로그를 사용하여 Databricks에서 실행되는 쿼리에서 런타임 데이터 계보 를 캡처할 수도 있습니다.

  Databricks Lakehouse 모니터링 을 사용하면 계정의 모든 테이블에서 데이터의 품질을 모니터링할 수 있습니다. 또한 기계 학습 모델 및 모델 제공 엔드포인트의 성능을 추적할 수 있습니다.

  관찰을 위해 시스템 테이블 은 계정 운영 데이터의 Databricks 호스팅 분석 저장소입니다. 시스템 테이블은 계정 전체에서 기록 관찰에 사용할 수 있습니다.

- **데이터 인텔리전스 엔진**

  Databricks Data Intelligence 플랫폼을 사용하면 전체 조직에서 데이터 및 AI를 사용할 수 있습니다. DatabricksIQ를 통해 구동되며, 생성 AI를 레이크하우스의 통합 이점과 결합하여 데이터의 고유한 의미 체계를 이해합니다.

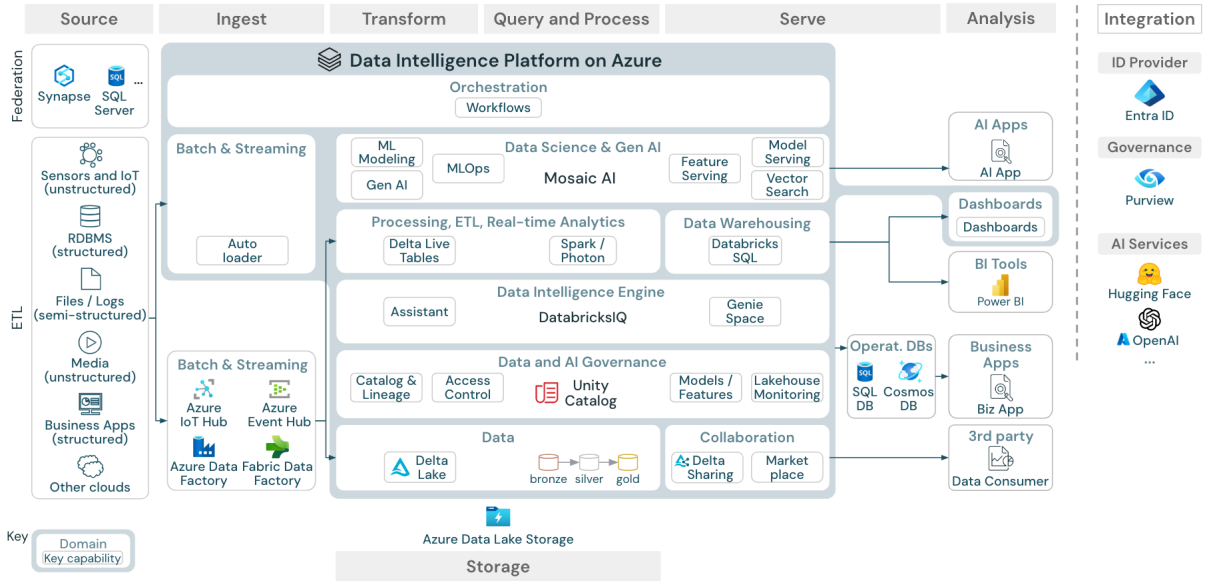
  Databricks Assistant는 Databricks Notebook, SQL 편집기 및 파일 편집기에서 개발자를 위한 컨텍스트 인식 AI 도우미로 사용할 수 있습니다.

- **오케스트레이션**

  Databricks 작업은 Databricks Data Intelligence 플랫폼에서 데이터 처리, 기계 학습 및 분석 파이프라인을 오케스트레이션합니다. Delta Live Tables 를 사용하면 선언적 구문을 사용하여 안정적이고 유지 관리 가능한 ETL 파이프라인을 빌드할 수 있습니다.

Azure의 데이터 인텔리전스 플랫폼 참조 아키텍처

Azure Databricks 참조 아키텍처는 원본, 수집, 서비스, 분석/출력 및 스토리지 요소에 대한 Azure 관련 서비스를 추가하여 일반 참조 아키텍처에서 파생됩니다.



다운로드: Azure의 Databricks Lakehouse에 대한 참조 아키텍처

Azure 참조 아키텍처는 수집, 스토리지, 서비스 및 분석/출력을 위한 다음과 같은 Azure 관련 서비스를 보여 줍니다.

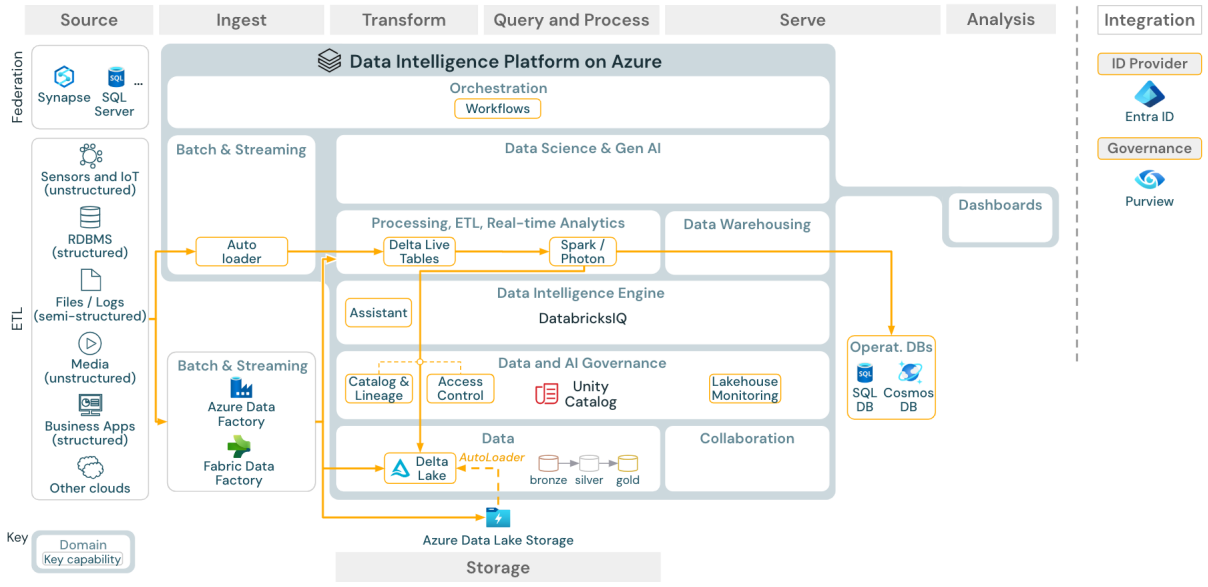
- Lakehouse Federation의 원본 시스템으로 Azure Synapse 및 SQL Server
- 스트리밍 수집을 위한 Azure IoT Hub 및 Azure Event Hubs
- 일괄 처리 수집을 위한 Azure Data Factory

- 개체 스토리지로 Azure Data Lake Storage Gen 2(ADLS)
- Azure SQL DB 및 Azure Cosmos DB를 운영 데이터베이스로
- UC가 스키마 및 계보 정보를 내보낼 엔터프라이즈 카탈로그인 Azure Purview
- POWER BI를 BI 도구로 사용

참고

- 참조 아키텍처의 이 보기는 Azure 서비스 및 Databricks Lakehouse에만 중점을 둡니다. Databricks의 레이크하우스는 대규모 파트너 도구 **에코시스템과** 통합되는 개방형 플랫폼입니다.
- 표시된 클라우드 공급자 서비스는 완전하지 않습니다. 개념을 설명하기 위해 선택됩니다.

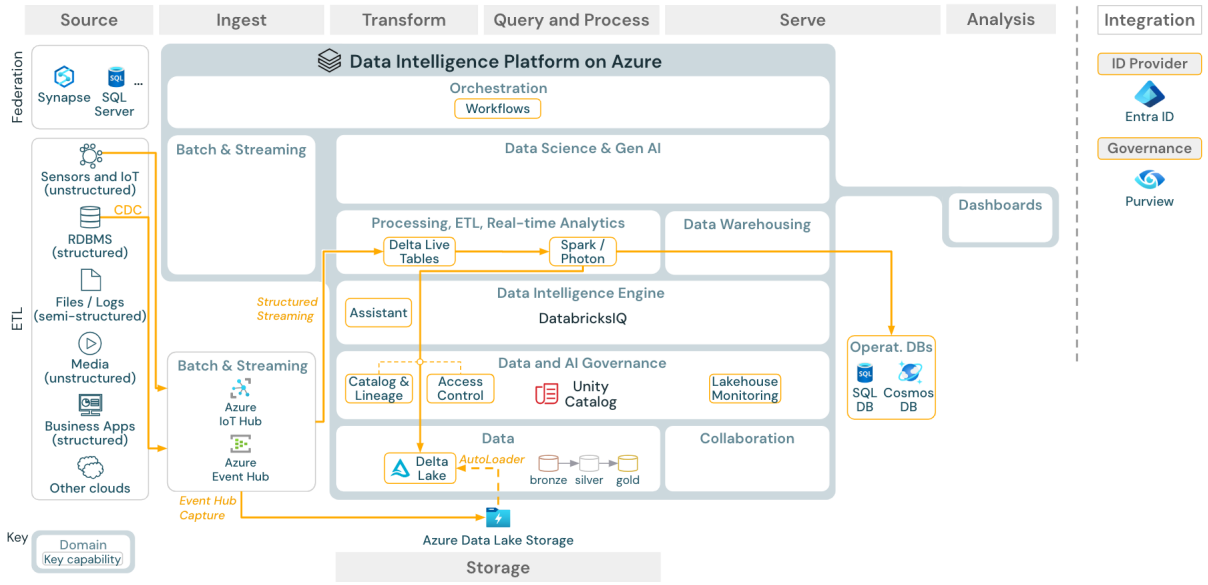
사용 사례: Batch ETL



다운로드: Azure Databricks에 대한 Batch ETL 참조 아키텍처

수집 도구는 원본별 어댑터를 사용하여 원본에서 데이터를 읽은 다음 자동 로더가 읽을 수 있는 클라우드 스토리지에 저장하거나 Databricks를 직접 호출합니다(예: Databricks Lakehouse에 통합된 파트너 수집 도구 사용). 데이터를 로드하기 위해 DLT를 통해 Databricks ETL 및 처리 엔진이 쿼리를 실행합니다. 단일 또는 멀티태스크 워크플로는 Databricks 작업에서 오케스트레이션하고 Unity 카탈로그(액세스 제어, 감사, 계보 등)로 제어할 수 있습니다. 대기 시간이 짧은 운영 시스템에서 특정 골든 테이블에 액세스해야 하는 경우 ETL 파이프라인의 끝에 있는 RDBMS 또는 키-값 저장소와 같은 운영 데이터베이스로 내보낼 수 있습니다.

사용 사례: 스트리밍 및 변경 데이터 캡처(CDC)



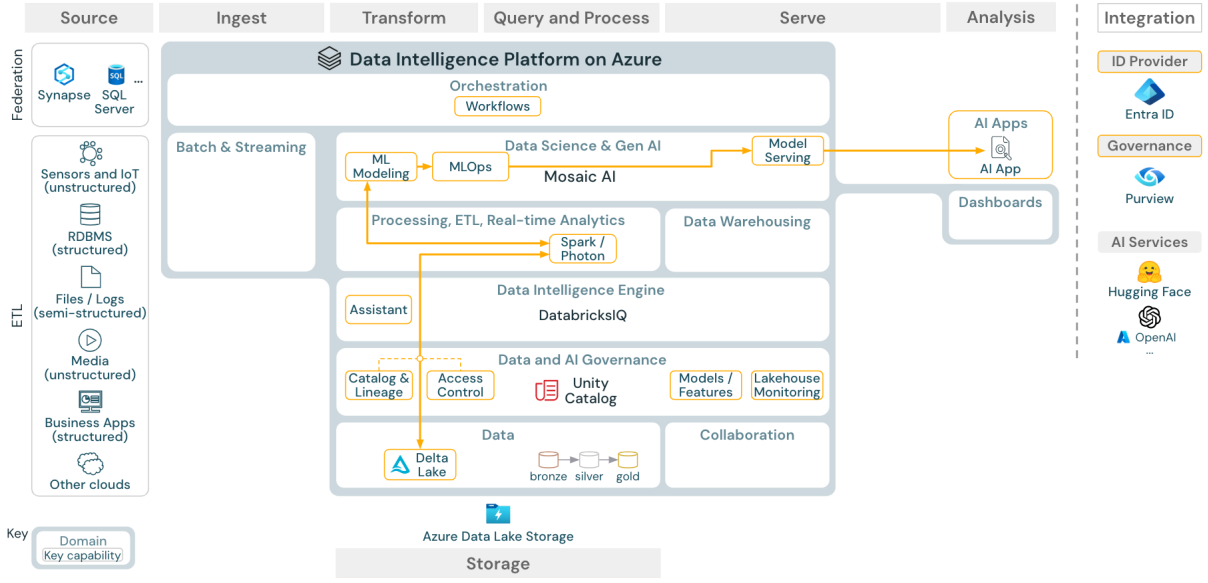
다운로드: Azure Databricks에 대한 Spark 구조적 스트리밍 아키텍처

Databricks ETL 엔진은 Spark 구조적 스트리밍을 사용하여 Apache Kafka 또는 Azure Event Hub와 같은 이벤트 큐에서 읽습니다. 다운스트림 단계는 위의 Batch 사용 사례 방식을 따릅니다.

CDC(실시간 변경 데이터 캡처)는 일반적으로 이벤트 큐를 사용하여 추출된 이벤트를 저장합니다. 여기에서 사용 사례는 스트리밍 사용 사례를 따릅니다.

추출된 레코드가 클라우드 스토리지에 먼저 저장되는 일괄 처리로 CDC가 수행되는 경우 Databricks 자동 로더는 이를 읽을 수 있으며 사용 사례는 Batch ETL을 따릅니다.

사용 사례: 기계 학습 및 AI



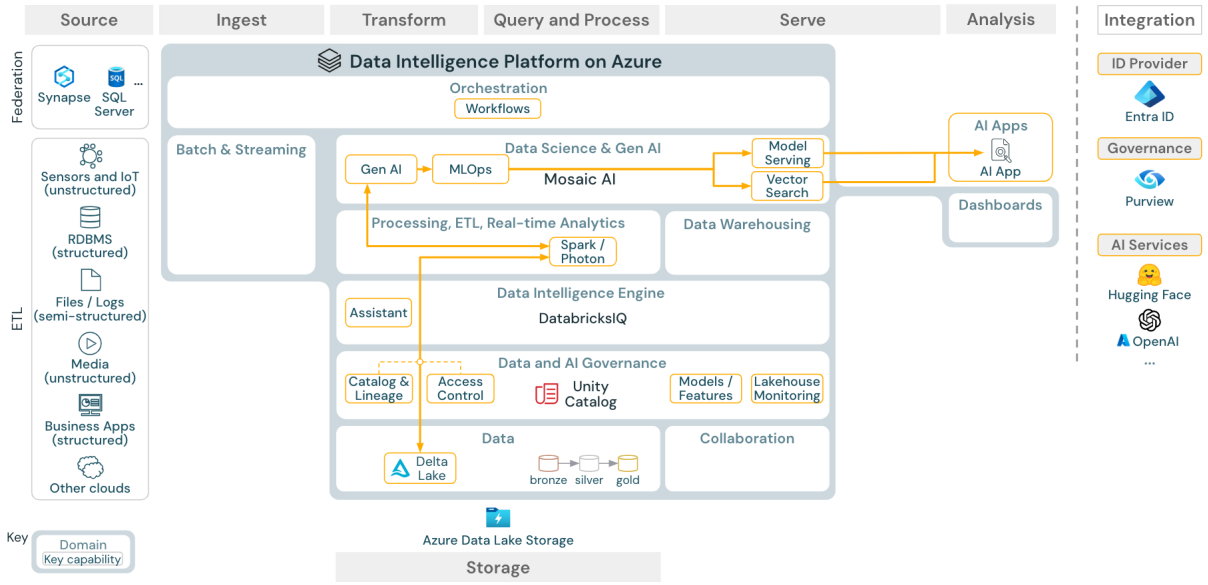
다운로드: Azure Databricks에 대한 기계 학습 및 AI 참조 아키텍처

기계 학습을 위해 Databricks Data Intelligence 플랫폼은 최신 기계 및 딥 러닝 라이브러리와 함께 제공되는 모자이크 AI를 제공합니다. 기능 저장소 및 모델 레지스트리(둘 다 Unity 카탈로그에 통합됨), AutoML을 사용하는 하위 코드 기능, 데이터 과학 수명 주기에 MLflow 통합과 같은 기능을 제공합니다.

모든 데이터 과학 관련 자산(테이블, 기능 및 모델)은 Unity 카탈로그에 의해 관리되며 데이터 과학자는 Databricks 작업을 사용하여 작업을 오케스트레이션할 수 있습니다.

확장 가능하고 엔터프라이즈급 방식으로 모델을 배포하려면 MLOps 기능을 사용하여 모델 제공에 모델을 게시합니다.

사용 사례: Gen AI(검색 증강 세대)

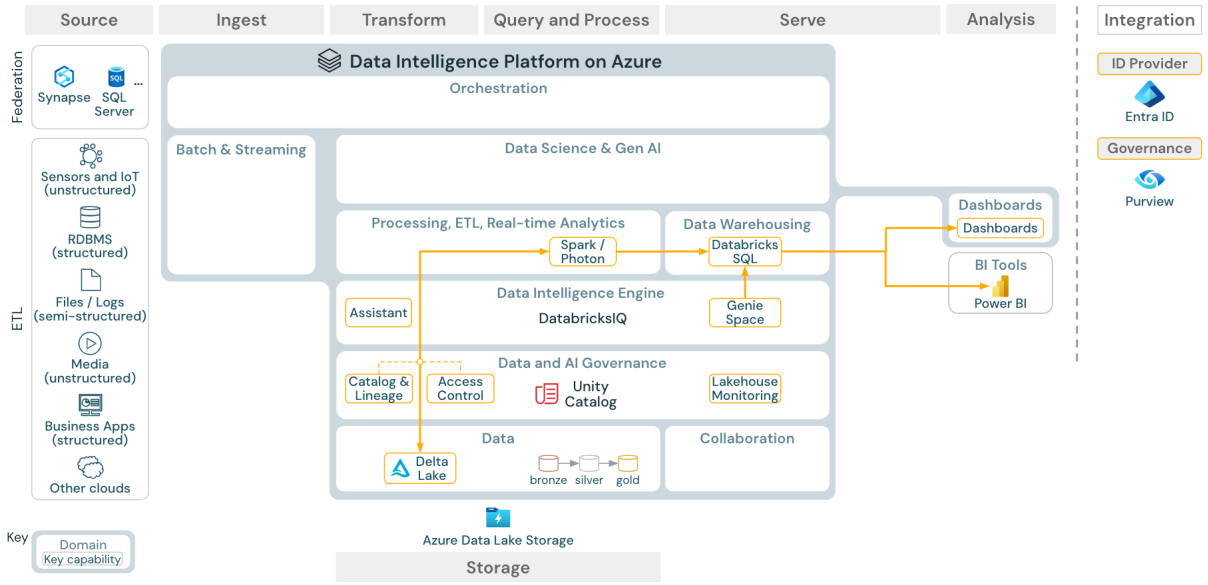


다운로드: Azure Databricks에 대한 Gen AI RAG 참조 아키텍처

생성된 AI 사용 사례의 경우 Mosaic AI에는 프롬프트 엔지니어링에서 기존 모델의 미세 조정 및 처음부터 사전 학습에 이르는 최신 라이브러리 및 특정 Gen AI 기능이 함께 제공됩니다. 위의 아키텍처는 RAG(검색 보강 세대) AI 애플리케이션을 만들기 위해 벡터 검색을 통합하는 방법의 예를 보여 줍니다.

확장 가능하고 엔터프라이즈급 방식으로 모델을 배포하려면 MLOps 기능을 사용하여 모델 제공에 모델을 게시합니다.

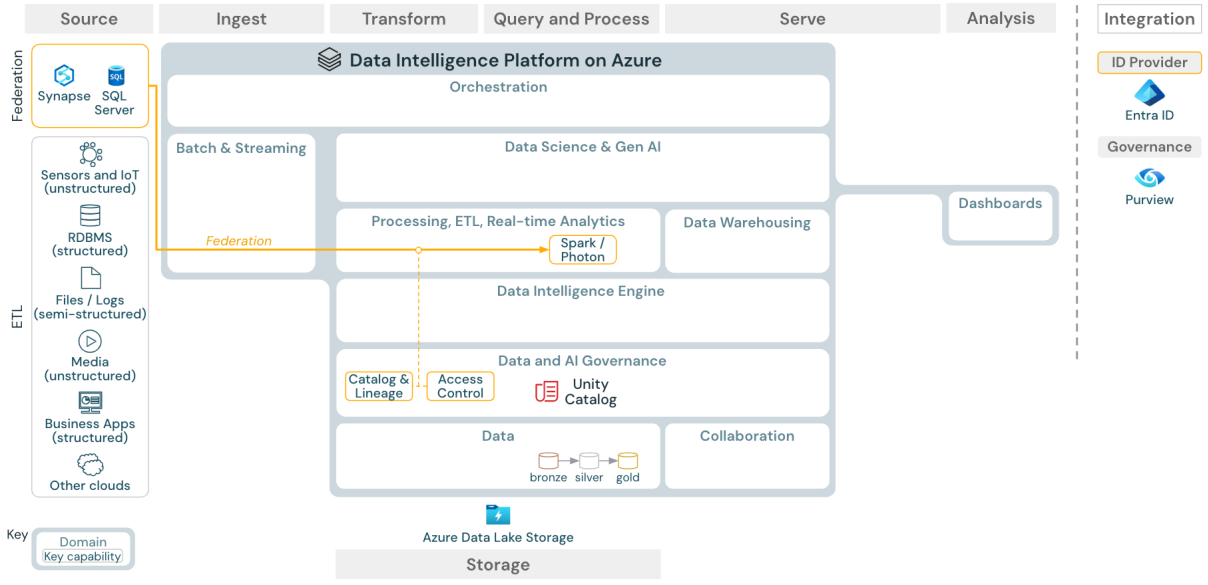
사용 사례: BI 및 SQL 분석



다운로드: Azure Databricks에 대한 BI 및 SQL 분석 참조 아키텍처

BI 사용 사례의 경우 비즈니스 분석가는 대시보드, Databricks SQL 편집기 또는 Tableau 또는 Power BI와 같은 특정 BI 도구를 사용할 수 있습니다. 모든 경우에 엔진은 Databricks SQL(서버리스 또는 비 서버리스)이며, Unity 카탈로그에서 데이터 검색, 탐색 및 액세스 제어를 제공합니다.

사용 사례: Lakehouse 페더레이션

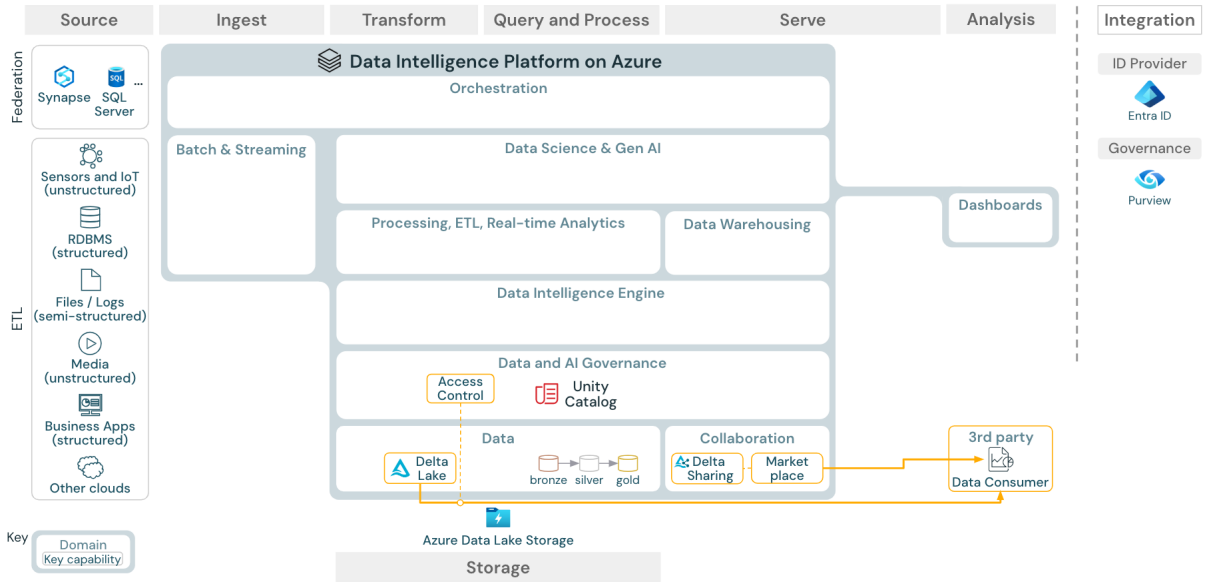


다운로드: Azure Databricks에 대한 Lakehouse 페더레이션 참조 아키텍처

Lakehouse 페더레이션을 사용하면 외부 데이터 SQL 데이터베이스(예: MySQL, Postgres, SQL Server 또는 Azure Synapse)를 Databricks와 통합할 수 있습니다.

모든 워크로드(AI, DWH 및 BI)는 데이터를 먼저 개체 스토리지에 ETL할 필요 없이 이 이점을 활용할 수 있습니다. 외부 원본 카탈로그는 Unity 카탈로그에 매핑되고 세분화된 액세스 제어를 Databricks 플랫폼을 통해 액세스에 적용할 수 있습니다.

사용 사례: 엔터프라이즈 데이터 공유



다운로드: Azure Databricks에 대한 엔터프라이즈 데이터 공유 참조 아키텍처

엔터프라이즈급 데이터 공유는 델타 공유에서 제공합니다. Unity 카탈로그로 보호되는 개체 저장소의 데이터에 직접 액세스할 수 있으며 Databricks Marketplace는 데이터 제품을 교환하기 위한 공개 포럼입니다.