손쉬운 학습

Databricks 특별판

데이터 웨어하우스에서 데이터 레이크하우스로 마이그레이션

for dummies®
A Wiley Brand

데이터 레이크하우스 이해

데이터 및 AI 전략 현대화

마이그레이션 계획 수립

제공

databricks

Stephanie Diamond

Databricks 소개

Databricks는 데이터 및 AI 전문 회사입니다. Comcast, Condé Nast, Nationwide, 및 H&M을 비롯한 전 세계 수천 개의 조직이 데이터 엔지니어링, 머신 러닝 및 분석을 위해 Databricks의 개방형 통합 플랫폼에 의존하고 있습니다. Databricks는 벤처 지원을 받는 회사로, 샌프란시스코에 본사가 있고 전 세계에 지사를 두고 있습니다. Apache Spark, Delta Lake 및 MLflow의 최초 제작자들이 설립한 Databricks는 데이터 팀이 세계에서 가장 어려운 문제를 해결할 수 있도록 지원하는 임무를 수행하고 있습니다. 자세히 알아보려면 소셜 미디어에서 Databricks를 팔로우하세요.

 twitter.com/databricks

 www.linkedin.com/company/databricks

 www.facebook.com/databricksinc

데이터 웨어하우스에서 데이터 레이크하우스로 마이그레이션

Databricks 특별판

저자: Stephanie Diamond

for dummies®
A Wiley Brand

데이터 웨어하우스에서 데이터 레이크하우스로 마이그레이션 For Dummies®, Databricks 특별판

발행인
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN: 978-1-394-16168-3 (pbk); ISBN: 978-1-394-16169-0 (ebk). 인쇄 버전의 일부 빈 페이지는 ePDF 버전에 포함되지 않을 수 있습니다.

발행인 감사의 글

다음은 이 책을 발간할 수 있도록 도움을 주신 분들입니다.

프로젝트 책임자:
Carrie Burchfield-Leighton

선임 관리 편집자: Rev Mengle

원고 검토 편집자: Ashley Coffey

선임 고객 계정 관리자: Matt Cox

콘텐츠 개선 전문가: Tamilmani Varadharaj

목차

서론

데이터 레이크하우스는 직관적인 사용자 인터페이스(UI)를 통해 데이터 처리를 위한 강력한 엔진과 개발자, 분석가, 데이터 과학자 및 비즈니스 사용자를 위한 간단하고 직관적인 도구를 제공하는 데이터 관리용 클라우드 네이티브 플랫폼입니다. 이를 통해 몇 시간 또는 며칠이 아닌 몇 분 만에 분석 애플리케이션을 구축, 배포, 확장하고 관리할 수 있습니다. 데이터 레이크하우스는 최고의 데이터 웨어하우스와 데이터 레이크를 단일 플랫폼에 통합한 개방형 데이터 아키텍처입니다.

데이터 레이크하우스를 사용하면 먼저 다른 시스템으로 이동하지 않고도 모든 데이터를 한 곳에서 분석할 수 있습니다. 이 혁신적인 플랫폼 활용의 핵심은 현재 시스템을 데이터 레이크하우스로 마이그레이션하는 것입니다. 이 책에서는 기업이 미래에 대비하기 위해 엔터프라이즈 데이터 웨어하우스(EDW)에서 데이터 레이크하우스로 마이그레이션하는 이유와 그 방법을 살펴봅니다.

이 책에 대한 정보

이 책은 리더가 데이터 관리의 새로운 과제를 해결하기 위해 알아야 할 사항을 설명합니다. 다음을 포함하여 여러 주제를 다룹니다.

- 정형, 비정형, 반정형 데이터 등 관리해야 하는 고유한 데이터 유형
- 데이터 및 인공 지능(AI) 성숙도 곡선에서 회사가 어느 위치에 있는지 판별하는 방법
- EDW에서 데이터 레이크하우스로의 진화
- 레이크하우스로 마이그레이션하는 회사에서 발생하는 이점
- 레이크하우스로 마이그레이션할 때 고려해야 할 사항
- Databricks 레이크하우스로 마이그레이션해야 하는 10가지 이유

이 책에서 사용된 아이콘

이 책 전체에 걸쳐 중요한 정보를 강조하기 위해 다양한 아이콘이 사용됩니다. 각각의 의미는 다음과 같습니다.

팁

이 아이콘은 프로세스를 더 빠르고 쉽게 관리하는 데 도움이 되는 추가 정보를 제공합니다.

기억하세요

이 아이콘은 메모리 뱅크를 검색할 때 기억해야 할 내용을 다룹니다.

경고

이 아이콘은 귀하 또는 귀사에 해로울 수 있는 정보를 알려줍니다.

기술 자료

가끔 기초 단계를 넘어선 연구나 사실에 대한 몇 가지 간단한 정보를 제공합니다. 따라서 기술적 세부사항을 알고 싶다면 이 아이콘을 주목하세요.

추가 자료

이 책은 EDW에서 데이터 레이크하우스로 마이그레이션하는 방법에 대해 자세히 알아보는 데 도움이 될 수 있지만, 이 책에서 제공하는 것 이상의 정보가 필요한 분은 다음 링크를 참조하십시오.

- `databricks.com/discoverlakehouse`: 기존 데이터 웨어하우스가 오늘날의 향상된 요구사항을 지원할 수 없는 이유를 알아보세요.
- `databricks.com/product/data-lakehouse`: 레이크하우스 플랫폼이 간단한 개방형 멀티클라우드 플랫폼에서 모든 데이터, 분석 및 AI 사용 사례를 어떻게 지원할 수 있는지 알아보세요.
- `databricks.com/product/Databricks-sql`: Databricks SQL을 사용하여 레이크하우스 아키텍처에서 SQL 워크로드를 실행하는 방법을 알아보세요.
- `databricks.com/p/ebook/building-the-data-lakehouse`: 성공적인 데이터 레이크하우스를 위한 5단계를 다운로드하세요.
- `databricks.com/p/ebook/data-lakehouse-is-your-next-data-warehouse`: 레이크하우스의 내부 작동 원리를 다운로드하여 Databricks SQL의 내부 구조와 Databricks 레이크하우스가 어떻게 차세대 데이터 웨어하우스가 될 수 있는지 알아보세요.

이 장에서 다루는 내용

» 데이터 관리의 난제 살펴보기

» 새로운 데이터 유형 검토

» 클라우드로 데이터 이동

1장
한 시대의 종말

비즈니스 데이터는 계속해서 기업이 보유하고 있는 가장 귀중한 자산 중 하나가 되고 있습니다. 데이터 가용성이 계속해서 폭발적으로 증가함에 따라, 엔터프라이즈 데이터의 극대화, 최적화 및 세분화는 번창하는 비즈니스의 핵심으로 간주됩니다. 그러나 기업이 증가하는 데이터량을 따라잡기는 어렵습니다. 따라서 위험을 최소화하고 재무 목표를 달성할 수 있도록 잘 설계된 데이터 관리 아키텍처가 필요합니다.

이 장에서는 데이터 레이크하우스의 채택을 주도하는 전개 상황을 살펴보고 데이터 웨어하우스에서 오늘날의 레이크하우스로 진화한 과정을 검토합니다.

당면한 오늘날의 도전 과제

모든 형태의 데이터를 효과적으로 관리하는 것은 조직의 미래 경쟁력을 확보하고자 하는 리더에게 매우 중요합니다. 머신 러닝(ML), 인공 지능(AI) 및 데이터 과학을 수용하려면 데이터를 통합해야 합니다. 어떤 기술을 구축할지 결정할 때 "이것이 우리가 원하는 곳으로 데려다 줄 수 있는가?"라고 자문해야 합니다.

저비용 클라우드 스토리지, 오픈 소스 소프트웨어, ML 및 AI의 출현으로 조직이 데이터를 활용하는 방식이 크게 바뀌었습니다. 또한 COVID-19 팬데믹으로 인해 기업은 원격지에 분산된 인력에 적응해야 했습니다. 그 결과 클라우드 도입이 급증했습니다. 과거의 엔터프라이즈 데이터 웨어하우스(EDW)는 다음과 같은 기능을 포함한 현대의 데이터 관리 문제를 수용하기에 적합하지 않은 폐쇄형 독점 시스템이었습니다.

- ML, 데이터 과학 및 AI를 수행하고 예측에 필요한 기타 새로운 데이터 소스 지원
- 오디오 및 비디오 데이터 세트 저장
- 실시간 작업을 위한 스트리밍 지원
- 유연한 방식으로 확장
- 형식에 관계없이 원시 데이터 관리

문제의 범위를 이해하려면, 기술이 발전함에 따라 다양한 유형의 데이터를 사용할 수 있게 되었고 기업에서 데이터의 중요한 가치를 인식했다는 사실을 알아야 합니다. 기업에서는 정형 데이터뿐만 아니라 증가하는 반정형 및 비정형 데이터를 저장하고 분석할 수 있는 통합된 장소가 필요하다는 것을 깨닫게 되었습니다.

- **반정형:** 이러한 데이터에는 로그, 클릭스트림, CVS, JSON 및 XML이 포함됩니다.
- **비정형:** 이러한 데이터는 문서, 이메일, 서신 형식의 조직 내부 대화에서 나오며, 온도계, 드론 및 공장 기계 등에서 나오는 사물 인터넷(IoT) 데이터와 같은 기타 비정형 데이터와 이미지, 비디오 및 아날로그 기반 데이터가 포함됩니다.

데이터 관리의 초창기 회고

기술이 어떻게 발전했는지 이해하려면 초창기 데이터 관리를 살펴보십시오. 중앙 저장소에 데이터를 보관할 필요가 정말로 없었습니다. 사용 가능한 데이터가 계층적이며 데이터베이스 테이블에 저장되었기 때문에 SQL(구조화된 질의 언어)을 사용하는 관계형 데이터베이스가 구축되었습니다. 오랫동안 이 방법은 필요한 재무 보고서 및 기타 비즈니스 보고서를 작성하는 데 적합했습니다.

데이터 웨어하우징의 기원 이해

데이터량이 증가함에 따라 보고 및 분석을 위해 운영체제의 데이터와 외부 데이터 소스를 결합하는 중앙 장소로 데이터 웨어하우스가 구축되었습니다. 따라서 IT 부서는 구조를 잘 알고 효율적으로 사용하는 방법을 알고 있었습니다. 데이터 웨어하우스 사용의 이점 중 일부에는 다음이 포함됩니다.

- **다양한 소스의 데이터 통합:** 최적화하고 질의할 수 있는 데이터 소스를 한 군데로 모았습니다. 이는 모든 데이터에 대한 단일 지점 역할을 했습니다.
- **과거 인텔리전스 획득:** 트랜잭션 데이터베이스에서 분석 처리를 분리하여 과거 인텔리전스를 추출할 수 있었습니다.
- **데이터 품질, 일관성 및 정확성 유지:** 다양한 제품 유형, 언어 및 통화에 대한 명명 규칙 코드가 일관되었습니다.
- **보고 및 비즈니스 인텔리전스(BI) 분석 지원:** 조직에서 이 구조를 사용하여 필요한 보고서와 BI를 얻을 수 있었습니다.
- **고속 검색 지원:** 프로덕션 데이터를 고속 데이터 입력 설계에서 고속 검색을 지원하는 설계로 변환하는 신중하게 설계된 데이터 모델을 중심으로 구축되었습니다.

데이터 웨어하우스의 부적절성 검토

기업이 데이터 웨어하우스를 구축하기 시작하면서 새로운 형태의 데이터가 등장하자, 데이터 웨어하우스의 부적절성이 곧바로 분명해졌습니다. 고려해야 할 데이터 웨어하우스의 몇 가지 문제:

- 데이터가 여러 소스(여러 데이터베이스 및 여러 주제 영역 기반 EDW 및 데이터 마트)에 흩어져 있었습니다.
- 각 소스의 데이터에는 고유한 스키마가 있고 각 비즈니스 애플리케이션은 고유한 스키마를 사용했습니다. 이 경우 표준화된 데이터 모델로 로드하기 위해 광범위하고 복잡한 ETL(추출, 변환, 로드)이 필요했으며, 다른 비즈니스 팀에서 다른 형식으로 다시 복사해야 했습니다.
- 데이터 웨어하우징을 로드하기 위한 ETL에는 광범위한 모델링과 수개월의 노력이 필요합니다. 데이터를 분석할 준비가 되었을 때 비즈니스 요구사항이 이미 충족되었거나 변경되었으며, 해당 데이터가 쓸모없게 되는 경우가 자주 있었습니다.
- 확장 비용이 기하급수적으로 더 비쌌습니다.
- 데이터 과학, ML, 실시간 분석, 반정형 또는 비정형 데이터 세트에 대한 지원이 없었습니다.

데이터 레이크 추가

EDW는 특정 보고서 및 대시보드에 대해 기업 수준에서 데이터를 집계하는 정형화된 데이터 모델이었습니다. 일반적으로 EDW에는 비즈니스 팀이 셀프 서비스 분석과 탐색·고급 ML/AI 요구사항을 위해 필요로 했던 세부적인 원시 데이터가 없었습니다. 또한 스토리지와 컴퓨팅을 확장하여 기업 전체의 모든 데이터(정형 및 비정형)를 수용할 수 있는 용량도 없었습니다. 그 결과 2011년경 Hadoop의 등장과 함께 데이터 레이크가 출현했습니다.

기억하세요

데이터 레이크는 비정형 원시 데이터의 저장소로, 일반적으로 다양한 목적으로 생성된 파일들이 저장되는 곳입니다. Apache Hadoop은 비용 절감을 위해 구축되었으며 ETL을 사용했습니다. 이제 Apache Spark는 클라우드에서 데이터 레이크를 실행합니다. 데이터 레이크는 저렴했고 개방형 표준 형식을 기반으로 모든 종류의 데이터를 저장할 수 있었습니다. 즉, 데이터 레이크와 외부 소스 간에 병목 현상이 없었습니다.

데이터 레이크의 즉각적인 이점에는 다음이 포함되었습니다.

- 모든 데이터가 최신 상태로 유지되었으며 새로운 데이터 소스를 쉽게 추가할 수 있었습니다.
- 데이터 세트의 여러 사본을 유지할 필요가 없었습니다.
- 대규모 데이터 정리 및 변환이 가능했습니다.
- 전체 데이터 세트에 대해 임시 쿼리를 실행할 수 있었습니다.
- 데이터 레이크에서 데이터를 쉽게 추출하여 다른 위치로 보낼 수 있었습니다.
- 오픈 소스 ML 라이브러리를 지원했습니다.

데이터 레이크에는 불가피하게 몇 가지 문제도 존재했습니다. 다음과 같은 문제들이 있었습니다.

- ACID(원자성, 일관성, 격리 및 내구성) 트랜잭션에 대한 지원 없음
- 데이터 품질 또는 거버넌스의 시행 없음
- 작업 실패 및 데이터 누락
- 열악한 BI 지원
- 열악한 성능

경고

데이터 레이크의 또 다른 문제는 데이터 레이크가 모든 가용 데이터의 폐기장이 되는 경우가 많았다는 점이었습니다. 부적절한 데이터 거버넌스와 데이터 품질에 대한 관심 부족 등 중대한 결점으로 인해 데이터 레이크는 때때로 데이터 늪으로 불렸습니다. 3장에서 새로운 데이터 레이크하우스를 지원하기 위해 데이터 레이크가 어떻게 진화했는지 자세히 알아보십시오.

클라우드를 통한 기존 데이터 웨어하우스 고려

현대의 엔터프라이즈 데이터 스토리지가 도입되고 몇 년 후, 기업이 극복해야 하는 몇 가지 심각한 문제가 있음이 분명해졌습니다. 특히, 조직은 값비싼 하드웨어와 소프트웨어로 인해 복잡한 대규모 데이터 사일로, 불충분한 스토리지, 효율성 부족으로 어려움을 겪었습니다. 이렇게 되어서 클라우드 데이터 웨어하우스로 진입합니다.

클라우드 데이터 웨어하우스를 채택하게 된 동기는 무엇일까요? 기업이 성공하는 데 필요한 몇 가지 핵심적인 기본 기능을 제공했기 때문입니다.

- **탄력성:** 즉각적인 프로비저닝을 제공하므로 필요에 따라 확장/축소할 수 있었습니다.
- **손쉬운 관리:** 인프라 비용 절감 외에도 관리 리소스를 확보하여 가장 중요한 일에 집중할 수 있었습니다.
- **혁신의 속도:** 즉각적인 혁신 환경으로 비즈니스를 성장시킬 수 있었습니다.

이 장에서 다루는 내용

- 데이터 및 기술 경영진의 주요 목표 살펴보기
- 데이터 문화 육성
- 성숙도 모델 평가

2장 데이터 및 AI 전략의 우선순위 지정

오늘날의 경쟁 환경에서는 조직의 데이터를 지원하는 데 적합한 아키텍처를 갖추는 것만으로는 충분하지 않습니다. 조직의 모든 필수 구성요소를 지원하는 포괄적인 전략도 필요합니다. 이 전략에는 사람, 비즈니스 목표 및 기술 활용이 포함되어야 합니다. 이는 장기적인 비즈니스 성공의 열쇠입니다. 궁극적으로 기술은 전략의 원동력이 되어야 하며 다른 방식이 되어서는 안 됩니다.

이 장에서는 데이터 및 기술 경영진이 달성하고자 하는 상위 3가지 전략적 목표와 데이터 문화 구축의 이점을 살펴봅니다. 또한 여러분의 조직이 어디에 해당하는지 판별하기 위한 성숙도 모델과 진행하기 위해 취할 수 있는 조치들을 살펴봅니다.

최우선 전략 목표 검토

데이터 및 인공 지능(AI) 전략을 구체화하기를 원하는 데이터 및 기술 경영진은 다음 세 가지 목표에 우선순위를 둘 것입니다.

- 데이터에서 새로운 비즈니스 가치 창출
- 리스크 줄이기
- 비용 통제

이 섹션에서는 각 목표를 살펴봅니다.

데이터에서 새로운 비즈니스 가치 창출

사용할 수 있는 고유한 형식의 데이터(예: 웹 및 모바일 장치 또는 소셜 미디어 게시물의 고객 상호작용을 포함하는 반정형 데이터)가 훨씬 더 많아짐에 따라, 기업은 레거시 플랫폼을 확장할 수 없고 더 나은 데이터 분석에 대한 증가하는 요구를 충족할 수 없다는 것을 인식하고 있습니다.

데이터 및 기술 경영진은 해당 데이터를 사용하여 더 나은 통찰력을 얻고 비즈니스 경쟁력을 높이기를 원합니다. 특히, 그들은 사용자 경험을 개선하고 데이터 페르소나 전반에 걸쳐 협업을 늘리는 저비용 접근방식을 추구합니다. 이 목표로 인해 복잡하고 값비싼 온프레미스 엔터프라이즈 데이터 웨어하우스(EDW) 아키텍처에서 멀어지게 됩니다.

리스크 줄이기

조직의 리더를 위한 또 다른 전략적 목표는 취약한 데이터 관리, 실패한 IT 프로젝트, 첨단 분석 플랫폼의 부족으로 인한 혁신의 실패, 항상 존재하는 사이버 공격의 위협과 같은 여러 잠재적 리스크를 줄이는 것입니다. 이러한 위협으로 인해 데이터를 저장, 처리, 관리 및 보호하기 위한 일관된 방식이 필요합니다. 그러나 다음과 같은 요인으로 인해 이 목표는 더 복잡해집니다.

- 일반 데이터 보호 규정(GDPR) 및 캘리포니아주 소비자 개인정보 보호법(CCPA) 등 진화하는 개인정보 보호 규제 환경을 준수해야 할 필요성
- Google 및 Apple과 같은 기업의 새로운 개인정보 보호 지침에 대응하기 위해 데이터 관리를 조정해야 할 필요성
- 일반적인 행동, 인구통계 및 가용 데이터를 대체할 수 있는 새로운 데이터 소스를 활용하는 방법을 파악해야 할 필요성

비용 통제

리더는 항상 비용 통제의 필요성과 씨름해야 합니다. 관리하는 데이터의 양이 늘어남에 따라 데이터 웨어하우스의 비용이 매우 빠르게 증가할 수 있습니다. 게다가 데이터 센터 장비, 데이터베이스 관리 운영 및 유지보수, 많은 고정된 벤더 계약에서 발생하는 비용도 있습니다.

이 문제(현재 데이터 및 AI 이니셔티브 수용)에 대한 해결책은 변화하는 비즈니스 요구에 적응할 수 있는 탄력적이고 유연한 클라우드 아키텍처를 구현하고 데이터 크기가 증가함에 따라 클라우드에서 우수한 가격 대비 성능을 제공하는 것입니다.

기억하세요

더 간단한 아키텍처를 활용하면 민첩성이 향상되어 데이터 및 기술 경영진이 지연이나 IT 부서의 개입 없이 통합할 수 있고 실행 가능한 인사이트를 얻을 수 있습니다.

데이터 문화 조성에 집중

조직에서 데이터의 가치를 인식함에 따라 기업은 데이터 문화를 구축하고 유지해야 할 필요성이 생겼습니다. 강력한 데이터 문화는 조직이 미래에 대비할 수 있도록 보장합니다.

다음 두 가지 측면에서 강력한 데이터 문화를 가진 회사를 판별할 수 있습니다.

- 전체 조직이 사용 가능한 관련 데이터를 사용하여 매일 정보에 입각한 비즈니스 결정을 내림.
- 데이터가 경험, 직관 또는 근속 기간보다 더 큰 비중을 차지함. 데이터와 인사이트는 근거를 제공합니다.

팁

여러분의 조직이 여기에 해당되나요? 그렇지 않다면 "데이터 및 AI 성숙도 곡선 조사" 섹션에서 진행 방향을 안내해주는 데이터 성숙도 모델을 확인하십시오.

더 나은 데이터 문화를 구축하는 데 집중하려면 다음 아이디어를 고려하십시오.

- **비즈니스 목표 및 결과를 달성하는 방법에 대해 명확히 하십시오.** 오늘날 비즈니스 리더는 데이터와 AI가 목표 달성에 도움이 된다는 것을 알고 있습니다. 조직에서 이들을 어떻게 지원할 것인지 자세히 명시하십시오.
- **사람에게 투자하십시오.** 오늘날의 시장에서 직원들은 장기적인 성공의 열쇠입니다. 최고의 인재를 채용하기 위한 경쟁이 치열하므로 다음 두 가지에 집중해야 합니다.
  - 최고의 인재가 찾아올 수 있도록 최대한 원활한 데이터 환경 구축
  - 최고의 직원이 가진 기술을 활용하기 위한 재교육
- **기술을 현대화하십시오.** 기업이 복잡한 독점 솔루션에서 멀어짐에 따라 데이터 환경은 개방형 표준, 저비용 스토리지 및 주문형 컴퓨팅을 지원해야 합니다.

데이터 민주화

MIT Tech Review는 Databricks와 협력하여 351명의 최고 데이터 책임자, 최고 분석 책임자, 최고 정보 책임자 및 기타 고위 기술 경영자를 대상으로 고성능 데이터 및 AI 조직을 구축하는 데 어떻게 성공했는지(또는 실패했는지) 판별하기 위해 글로벌 설문조사(2021)를 실시했습니다.

주요 발견 사항 중 하나는 데이터를 민주화해야 한다는 것이었습니다. 이를 달성하기 위해 그들이 권장한 사항은 다음과 같습니다.

- **데이터 과학자를 사업부에 직접 포함:** 이 프로세스를 통해 데이터 사용자와 직접 상호작용할 수 있습니다.
- **사용자에게 분석에 대한 접근 권한 부여:** 이 조치를 통해 사용자가 스스로 인사이트를 얻을 수 있습니다.
- **간단한 인터페이스로 고위 리더가 시각적 도구에 접근할 수 있도록 함:** 이 방법을 통해 고위 리더가 필요에 따라 데이터 인사이트를 얻을 수 있습니다.

데이터에서 필요한 정보 얻기

많은 사람들이 데이터가 신종 석유라고 말합니다. 그러나 석유는 한 번만 정제할 수 있습니다. 데이터의 가치는 무수히 많은 방법으로 재분석하여 필요한 답변을 생성할 수 있다는 것입니다. 이러한 이유로 엔터프라이즈 데이터 플랫폼 아키텍처는 데이터를 분석하여 중요한 질문에 답변할 수 있어야 합니다.

기억하세요

데이터 플랫폼의 고객이 알고 싶어하는 사항

- **무슨 일이 일어났고 그 이유는 무엇인가?** 이 질문에 답하려면 보고서와 대시보드, 분석가, 자체 분석을 수행할 수 있는 역량을 갖춘 주제별 전문가가 필요합니다.
- **무슨 일이 일어날 것인가?** 데이터 과학, 머신 러닝(ML) 및 인공 지능(AI)에 대한 역량이 필요합니다.
- **적시에 어떤 조치를 취할 수 있는가?** 고객은 대응 방법을 알려주는 처방적 분석을 원합니다(예를 들어 사기 분석처럼 때로는 몇 초 만에 자동화된 방식으로 제공). 비즈니스 관리자, 데이터 과학자 및 데이터 엔지니어는 협업 플랫폼에서 함께 협력하는 방식으로 이러한 답변을 도출할 수 있습니다.

그림 2-1은 이러한 질문들이 함께 작동하는 방식을 보여줍니다.

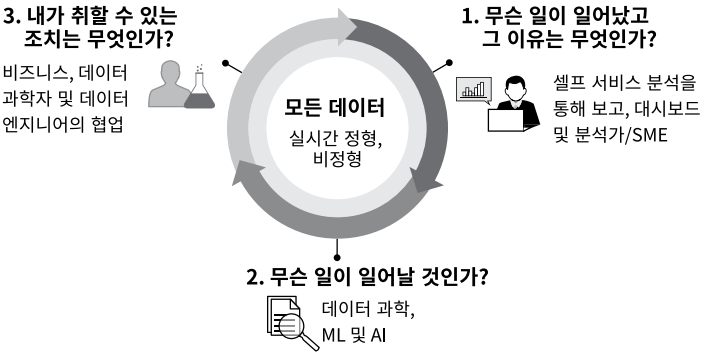


그림 2-1: 데이터 플랫폼 고객이 묻는 질문.

데이터 및 AI 성숙도 곡선 조사

미래에 대비하고 정의된 성숙도 곡선을 따라 움직이기 위해 기업은 데이터 자산과 AI 사고방식을 채택해야 합니다. 목표는 서술형에서 규범형으로 전환하는 것입니다. 이를 위해 Databricks는 조직이 데이터 및 AI 성숙도를 향한 여정의 현재 상태를 이해하는 데 사용할 수 있는 성숙도 모델을 만들었습니다.

모델은 다음과 같습니다.

- **탐색:** 이 단계에서 조직은 빅 데이터 및 AI를 탐색하고 몇 가지 시작 프로젝트 및 실험의 가능성과 잠재력을 이해하기 시작합니다.
- **실험:** 이 단계의 조직은 보다 광범위한 데이터 및 AI 전략을 탐색하기 위한 기본 기능과 기반을 구축하지만 비전, 장기 목표 또는 리더십 동의가 부족한 상태입니다.
- **공식화:** 이 단계에서는 데이터와 AI의 핵심 원칙이 기업 전략에 통합됨에 따라 데이터와 AI가 특정 프로젝트 및 이니셔티브와 연계된 비즈니스 사용자를 위한 가치의 원동력으로 싹틉니다.
- **최적화:** 데이터와 AI가 조직 전반에 걸쳐 가치의 핵심 원동력입니다. 이것들이 조직 전체에서 비즈니스 요구사항과 동의를 충족하는 확장 가능한 아키텍처를 통해 체계화되고 기업 전략의 핵심으로 자리잡습니다.
- **혁신:** 이 단계에서는 데이터와 AI가 기업 전략의 핵심이며 경쟁 우위의 귀중한 차별화 요소이자 원동력입니다.

여러분의 조직이 이 단계들 중 하나에 속한다고 인식하시나요? 그렇다면 더 높은 성숙도를 달성하기 위해 앞으로 나아가는 방법을 알고 계십니까? databricks.com/p/business-value-assessment-databricks에서 맞춤형 비즈니스 가치 평가를 예약할 수 있습니다.

이 장에서 다루는 내용

- » 레이크하우스의 진화 이해하기
- » 조직의 미래 경쟁력 구축
- » 레이크하우스 아키텍처 살펴보기

3장 레이크하우스 시대의 도래

엔터프라이즈 데이터 웨어하우스(EDW) 및 데이터 레이크를 사용하면서 얻은 교훈은 레이크하우스의 최신 클라우드 기반 데이터 아키텍처로 향하는 길을 열었습니다. 레이크하우스는 최고의 속성과 기능을 결합하여 과거보다 훨씬 더 강력하고 유연한 데이터 플랫폼을 제공합니다.

이 장에서는 최신 클라우드 기반 레이크하우스의 진화 과정과 조직의 미래 경쟁력에 대한 필요성을 살펴봅니다.

조직의 미래 경쟁력 구축

기업 리더의 중요한 요구사항 중 하나는 조직이 미래에 대비할 수 있도록 성공적으로 준비하는 것입니다. 구식 프로세스에 의존하여 데이터를 관리하면 조직을 정체시키고 경쟁에서 뒤처질 수 있습니다.


기술 자료

Statista의 조사에 따르면 2020년에서 2024년까지 전 세계적으로 조직에서 생성, 저장, 복사 및 소비되는 데이터의 총량이 149제타바이트로 152.5% 증가할 것으로 예상됩니다.

귀하의 조직은 데이터 레이크, EDW, 비즈니스 인텔리전스(BI), 데이터 과학, 머신러닝(ML) 및 스트리밍 플랫폼의 복잡한 기술 스택과 이들 간의 데이터 이동 및 다양한 보안 패러다임 관리의 복잡성을 운영하고 관리할 준비가 되어 있습니까? 아니면 복잡한 플랫폼 및 보안 관리 대신 데이터로 비즈니스 과제를 해결하는 데 집중하고 대비하기 위해 관리가 간편한 단일 레이크하우스 플랫폼으로 단순화하는 것을 고려하시겠습니까? 새로운 도전에 대비할 수 있도록 레이크하우스로 마이그레이션하는 것을 고려하십시오.

EDW에서 데이터 레이크, 레이크하우스로 진화

그림 3-1과 같이 레이크하우스의 최신 데이터 아키텍처는 1980년대의 EDW와 2000년대 중반의 Hadoop 스타일 데이터 레이크의 진화에서 찾아볼 수 있습니다.

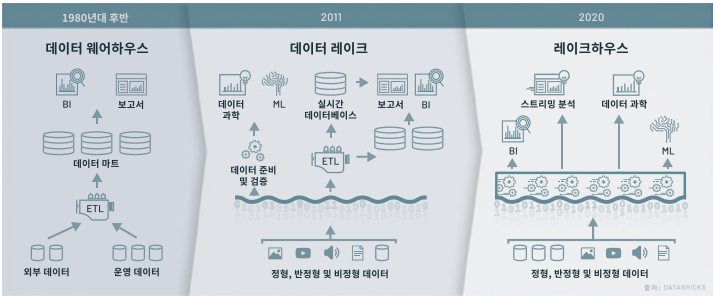


그림 3-1: 데이터 관리의 진화.

초기 데이터 웨어하우스는 분석에 최적화되었지만 비정형 데이터에 대해서는 최적화되지 않았습니다. 마찬가지로 데이터 레이크는 전통적으로 비정형 데이터를 저장하는 데 사용되었지만 분석에 최적화되지 않았습니다. 결과: 고객은 민첩성과 거버넌스 중 하나를 선택해야 했습니다. 레이크하우스 아키텍처의 가치는 데이터 웨어하우스의 속도와 거버넌스, 데이터 레이크의 유연성 덕분에 이제 데이터 팀이 모든 데이터를 단일 플랫폼에 저장할 수 있다는 것입니다.

레이크하우스란 무엇인지 살펴보기

레이크하우스는 데이터 웨어하우스와 데이터 레이크에서 최고의 요소들을 가져와서 두 환경의 장점을 모두 제공하는 단일 플랫폼으로 결합합니다. 다음은 레이크하우스 아키텍처 운영 시의 이점입니다.

- 모든 데이터에 대해 단일 소스에서 모든 데이터 사용 사례를 관리합니다.
- 보다 신속하게 대응하고 새로운 인사이트를 더 빨리 찾습니다.
- 모든 사람이 동일한 버전의 데이터를 보도록 합니다.

- 사일로와 관리해야 하는 시스템 및 도구의 수를 줄여 기존 아키텍처와 보안을 단순화합니다.
- 데이터 마트 및 EDW를 다른 비정형 데이터와 통합 및 연결하여 강화하고 혁신적인 데이터 제품을 생성할 수 있습니다.
- 데이터 레이크하우스 내의 데이터에 대해 추출, 변환, 로드(ETL) 작업을 수행합니다.

기억하세요

레이크하우스 아키텍처의 특징

- **단순함:** 데이터, 분석 및 AI를 단일 플랫폼에 통합할 수 있습니다.
- **개방형:** 개방형 표준 및 형식으로 데이터 생태계를 통합하여 벤더에 종속되는 것을 방지합니다.
- **멀티 클라우드:** 모든 클라우드에서 일관된 관리, 보안 및 거버넌스 경험을 제공하고 팀이 모든 데이터를 활용하여 새로운 인사이트를 발견하는 데 집중할 수 있도록 합니다.

예를 들어, 그림 3-2는 Databricks 레이크하우스 플랫폼을 보여줍니다.

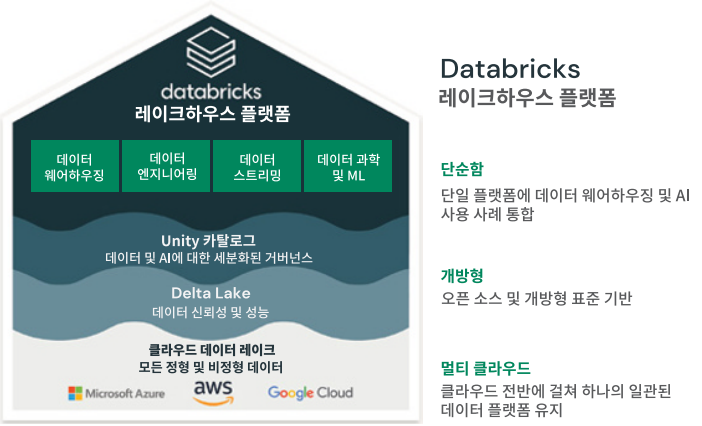


그림 3-2: **Databricks** 레이크하우스 플랫폼.

레이크하우스가 최신 클라우드 데이터 웨어하우스 이상인 이유 이해

레이크하우스는 다양한 소스의 모든 데이터가 함께 저장되는 시스템입니다. 이러한 방식으로 단일 쿼리로 서로 다른 사업부나 출처의 여러 데이터 요소에 한 번에 액세스할 수 있습니다.

팁

데이터 레이크하우스는 기존 데이터 웨어하우스에 비해 몇 가지 중요한 이점이 있습니다.

- **더 유연합니다.** 비즈니스에 적합한 형식으로 데이터를 저장할 수 있습니다.
- **항상 최신 상태입니다.** ELT(추출, 로드, 변환) 패턴 덕분에 분석가가 추가 ETL 없이 최신 데이터에 액세스할 수 있습니다. 실시간 스트리밍 파이프라인은 레이크하우스에서 기본적으로 지원됩니다.
- **단일 데이터 사본만 필요합니다.** 여러 데이터 사본을 서로 다른 스키마와 형식으로 유지할 필요가 없습니다.
- **새 데이터 소스를 쉽게 추가할 수 있습니다.** 광범위한 ETL 없이 데이터를 쉽게 수집할 수 있습니다.
- **진정한 엔터프라이즈급입니다.** 많은 주제 영역별 데이터 웨어하우스 및 데이터 마트를 유지관리하는 것과 달리, 데이터 레이크하우스는 모든 EDW, 데이터 마트를 원시, 비정형 및 반정형 데이터 바로 옆에 보관합니다.

이 장에서 다루는 내용

- 데이터 리더의 4가지 진실 살펴보기
- 기술 향상
- 더 나은 건강성 제공

4장
레이크하우스 마이그레이션의 이점

조직의 이질적인 시스템의 모든 원시 데이터에서 지식을 추출하면 엄청난 경쟁 우위를 확보할 수 있습니다. 과거로 돌아갈 수 있다면 현재의 과제를 지원하기 위해 데이터를 관리하는 방법에 대해 다른 결정을 내릴 수 있을 것입니다. 이 장에서는 레이크하우스 아키텍처를 배포하고 Bread 및 Amgen과 같은 회사가 이러한 전환을 통해 어떤 이익을 얻었는지 검토함으로써 조직을 혁신할 수 있는 방법을 살펴봅니다.

조직 혁신

조직이 미래에 대비하기 위해 변화를 모색할 때 데이터 리더는 다음 사항들을 고려합니다.

- 기존 솔루션에서 마이그레이션하는 비용
- 데이터 팀이 직면하게 될 학습 곡선
- 솔루션을 중심으로 구축된 광범위한 생태계

그들은 다음 네 가지 사실이 결정의 지침이 되어야 한다고 믿습니다.

- **머신 러닝(ML)과 인공 지능(AI)은 미래다.** 리더는 데이터에서 더 나은 통찰력을 얻기 위해 성숙도 곡선을 높여 ML과 AI를 최대한 활용하기를 원합니다. 자세한 내용은 2장을 참조하세요.
- **오픈 소스와 개방형 형식 사용.** 리더는 특정 벤더에 얽매이는 것을 불안해하며 가능하면 오픈 소스를 사용하기를 원합니다.
- **클라우드에 대한 계획.** 대부분의 회사는 비용 절감, 유연성 및 강점을 구축할 수 있는 역량을 확보하기 위해 클라우드로 이동합니다.
- **간단한 데이터 아키텍처 사용.** 리더는 복잡성을 제거하고 사일로를 최소화하며 동일한 데이터의 여러 사본을 피하기를 원합니다. 그들은 모든 데이터에 대해 하나의 거버넌스, 보안 및 계보 모델을 원합니다.

위의 사실들과 관련하여 레이크하우스 아키텍처가 이러한 각각의 요구사항을 어떻게 충족하는지 살펴봅니다.

- **ML과 AI는 미래다.** 레이크하우스는 처음부터 ML과 AI를 사용합니다.
- **오픈 소스 및 개방형 형식 사용.** 레이크하우스는 개방형 형식과 표준을 사용하여 데이터 이식성을 높이고 벤더 종속을 피합니다.
- **클라우드에 대한 계획.** 레이크하우스는 저렴한 클라우드 개체 저장소를 활용하여 모든 엔터프라이즈 데이터를 저장합니다.
- **간단한 데이터 아키텍처 사용.** 레이크하우스는 데이터 엔지니어링, 데이터 웨어하우징, 실시간 스트리밍, 데이터 과학 및 ML을 포함한 모든 사용 사례 플랫폼을 지원합니다.

고객 성공 사례

다른 회사에서 레이크하우스를 어떻게 사용했는지 궁금한 분들을 위해 이 섹션에서는 레이크하우스, 특히 Databricks 레이크하우스 플랫폼으로 마이그레이션함으로써 상당한 성공을 거둔 두 회사의 사례를 보여줍니다.

Bread

Bread는 Alliance Data Systems의 한 부서입니다. 판매자 및 파트너와 통합하여 고객을 위한 맞춤형 결제 옵션을 제공하는 기술 기반 결제 회사입니다. Bread의 데이터 웨어하우스는 기가바이트에서 테라바이트로 증가하는 데이터를 처리할 수 없었습니다. 데이터를 쿼리하는 데 몇 시간이 걸렸습니다. 이 회사는 또한 일괄 처리에서 스트리밍 데이터로 전환하고 실시간 인사이트와 결과를 제공하는 데 어려움을 겪었습니다.

Bread는 AWS의 Databricks 레이크하우스 플랫폼으로 마이그레이션하여 POS 시스템에서 트랜잭션 데이터를 효율적으로 수집하고 ELT(추출, 로드, 변환) 시스템을 Delta Lake로 전환했습니다. 회사는 이를 수행하면서 분석 엔지니어와 데이터 과학자가 신용 위험, 손실 추정 및 사기 사용 사례에 대한 데이터 세트를 민주화하도록 지원했습니다. Bread는 이제 이 솔루션을 사용하여 데이터 수집 및 성능을 저하시키는 대용량 데이터에 대해 걱정하지 않고 데이터 분석 및 ML을 확장할 수 있게 되었습니다. 그 결과 이 회사는 다운스트림 비즈니스 보고, 분석 및 ML 사용 사례를 위해 테라바이트의 데이터를 분석할 수 있게 되었으며, 이들은 모두 비즈니스 의사 결정과 고객 경험을 개선하도록 설계되었습니다.

실제로 Bread의 직원 데이터 엔지니어인 크리스티나 테일러 씨는 회사의 오래된 클라우드 컴퓨팅 기반의 데이터 웨어하우징 솔루션에서 새로운 Databricks의 레이크하우스로 전환함으로써 Bread가 비즈니스 활동을 추진하는 방식이 가장 완전한 최신 데이터 보기를 기반으로 하는 방식으로 변화했다고 부언했습니다. 이는 이전의 데이터 웨어하우스에서는 불가능했던 것입니다.

다른 결과로는 과거 데이터 웨어하우스와 비교하여 컴퓨팅 비용이 90% 절감되었고, 단 1.5배의 비용으로 데이터 볼륨이 140배 증가했으며, 비즈니스 보고를 위해 실행 가능한 인사이트의 성능이 23% 향상되었습니다. 전체 고객 사례와 클라우드 데이터 웨어하우스에서 Databricks 레이크하우스에 이르는 여정에 대해 알아보려면 databricks.com/customers/bread-finance를 방문하세요.

Amgen

Amgen은 세계 최대의 독립 생명공학 회사입니다. 지난 40년 동안 방대한 양의 데이터는 신약 제조 프로세스를 개척하고 생명을 구하는 의약품을 개발하는 데 도움이 되었습니다. 하지만 데이터의 크기가 늘어남에 따라 회사에서 비즈니스의 다양한 측면들을 짜맞추고 확장할 수 없었습니다. Amgen은 데이터에 있는 다양한 관점을 활용하기 위해 부서 간 협업을 확장해야 했습니다. Amgen은 디지털 혁신 여정을 지원하기 위해 Databricks 레이크하우스 플랫폼을 사용하기로 결정했습니다.

Databricks 레이크하우스 플랫폼의 구현은 회사가 환자에게 서비스를 제공하고 의약품 개발 수명주기를 개선하는 데 도움이 되었습니다. Amgen의 데이터 수집 속도가 크게 증가하여 처리 시간이 75% 향상되었으며, 정적 Hadoop 클러스터에 비해 컴퓨팅 비용이 약 25% 감소하는 동시에 비즈니스에 인사이트를 2배 더 빠르게 제공할 수 있었습니다. 2017년 Databricks와 협력 관계를 맺은 이후 데이터 엔지니어링에서 분석가에 이르기까지 2,000명 이상의 데이터 사용자가 Databricks를 통해 400TB(테라바이트)의 데이터에 액세스하여 40개 이상의 데이터 레이크 프로젝트와 240개 이상의 데이터 과학 프로젝트를 지원했습니다.

성능 기록 수립

Databricks는 데이터 레이크에서 직접 완벽한 데이터 웨어하우징 기능을 지원하고 단일 데이터 아키텍처에서 두 환경의 장점을 모두 제공하는 데이터 레이크하우스를 신속하게 개발했습니다. 2020년 11월 Databricks SQL이라는 이름으로 전체 데이터 웨어하우징 기능 제품군을 발표했습니다. 그 이후로 대두된 질문은 레이크하우스를 기반으로 하는 개방형 아키텍처가 기존 클라우드 데이터 웨어하우스의 성능과 속도 및 비용을 제공할 수 있는지 여부였습니다.

인기있는 데이터 웨어하우스에서 TPC-DS를 자주 실행하는 바르셀로나 슈퍼컴퓨팅 센터는 Databricks SQL이 데이터 웨어하우징의 표준 성능 벤치마크인 100TB TPC-DS에서 2021년 11월 세계 신기록을 수립했다는 것을 알게 되었습니다. Databricks SQL은 이전 기록을 2.2배 앞질렀습니다. 또한 이 연구에서 Databricks가 다음과 같은 측면에서 경쟁사보다 우수한 것으로 나타났습니다.

- **데이터 웨어하우스로서 탁월한 가격 대비 성능:** 다른 데이터 웨어하우스에 비해 최대 12배 더 나은 가격대 성능비를 통해 상당한 비용 절감 효과를 얻을 수 있습니다.
- **더 빠른 인사이트:** 주요 클라우드 데이터 웨어하우스에 비해 대용량 데이터를 2.7배 더 빠르게 처리합니다.

여기에서 전체 스토리를 읽을 수 있습니다: databricks.com/blog/2021/11/02/databricks-sets-official-data-warehousing-performance-record.html.

이 장에서 다루는 내용

- 마이그레이션에 대한 애자일 접근방식
- 리프트 앤 시프트가 작동하지 않는 이유 알아보기
- 마이그레이션의 5가지 기둥 살펴보기

5장 레이크하우스로 마이그레이션해야 하는 이유 검토

새로운 아키텍처로 마이그레이션하는 것은 복잡한 프로세스가 될 수 있습니다. 데이터 웨어하우스 현대화 전략을 고려할 때는 진행하기 전에 몇 가지 필수 마이그레이션 요소를 계획하십시오. 아키텍처 선택에 내재된 차이점을 이해하면 현대화 이니셔티브를 가장 잘 진행하는 방법에 대해 정보에 입각한 결정을 내리는 데 도움이 됩니다. 이 장에서는 마이그레이션에 대한 반복적인 애자일 접근방식의 가치를 살펴보고 마이그레이션 여정을 계획하고 실행하는 방법을 제안합니다.

애자일 접근방식 사용

데이터를 레이크하우스로 마이그레이션하는 방법은 중요한 결정입니다. 어떤 길을 가더라도 균형이 필요합니다. Databricks는 다음과 같은 단계적 애자일 방식을 권장합니다.

- **플랫폼 관점에서 레이크하우스로 현대화 하기와 기존 EDW를 또 다른 클라우드 EDW로 리프트 앤 시프트 하기를 병행하십시오.** 플랫폼의 리프트 앤 시프트는 동일한 문제와 단점을 클라우드로 옮기는 결과를 가져옵니다. 코드와 애플리케이션 재설계 관점에서 리프트 앤 시프트 및 현대화에 대한 균형 잡힌 접근방식을 취하십시오.

- **리프트 앤 시프트 및 현대화에 대한 균형 잡힌 접근방식을 구현하십시오.** 한 번의 반복으로 코드를 리프트 앤 시프트하고 현대화하십시오. 자동화된 코드 변환기와 함께 리프트 앤 시프트를 사용하고 최적의 Databricks 패턴으로 즉시 현대화하십시오.


기술 자료

  리프트 앤 시프트는 대대적인 변경 없이 한 환경에서 다른 환경으로 애플리케이션 디자인과 코드를 이동하는 것을 의미합니다. 그러나 나중에 현대화하고 재설계할 때까지 기다리지 마십시오. 모든 모범 사례를 즉시 재설계하고 적용하십시오. 재설계가 필요한 항목과 리프트 앤 시프트 방식의 이점을 얻을 수 있는 코드를 결정하십시오.

- **효과가 있는 것과 그렇지 않은 것을 파악하고 반복하십시오.** 진행하면서 사용 사례와 워크로드를 추가하십시오.
- **단거리 경주에서 성공을 보여주고 적응하십시오.** 이런 방식으로 이해관계자에게 즉시 성공을 보여주고 학습 및 피드백을 통해 다음 마이그레이션 반복을 개선할 수 있습니다.


경고

단순한 리프트 앤 시프트가 답이 되는 경우는 드뭅니다. 리프트 앤 시프트만으로는 다음 세 가지 기본적 이점을 얻을 수 없습니다.

- 레이크하우스를 완전히 활용하지 못합니다.
- 모든 것을 있는 그대로 옮기기만 하면 발견할 수 있는 모든 혁신을 잃게 됩니다.
- 비용 절감, 민첩성 및 규모를 최적화하기 위해 기술 전략을 개선할 기회가 없습니다.

리프트 앤 시프트와 토털 리엔지니어링 모두 장단점이 있습니다.

- **리프트 앤 시프트:** 장점: 수백만 개의 데이터 웨어하우스 라이선스 갱신이 예정되어 있는 경우 이 방식이 더 빠르고 확실합니다. 단점: 디자인과 코드를 리엔지니어링하고 리팩토링할 기회를 얻지 못할 수도 있습니다.
- **토털 리엔지니어링:** 장점: 최고의 품질을 제공합니다. 단점: 완료하는 데 몇 년이 걸릴 수 있으며 초기 비용이 많이 듭니다.

마이그레이션 여정 계획

마이그레이션 여정을 고려할 때는 각 단계를 신중하게 계획하십시오. 여정은 일련의 단계들로 묘사될 수 있습니다(그림 5-1 참조). 각 단계의 작동 방식은 다음과 같습니다.

1. **발견 단계: 내부 질문을 합니다.**

   이 단계의 핵심은 두 가지 질문에 답하는 것입니다. 나는 지금 어디에 있고 어디로 가야 하는가? 모든 데이터 팀, 최고 정보 책임자 및 기타 관련 이해관계자로부터 설문지를 수집해야 합니다. 팀이 가정을 테스트하고 검증할 때는 새로운 학습과 자기 발견을 많이 할 준비를 하십시오.

2. **평가 단계: 마이그레이션 평가를 수행합니다.**

   표에 있는 솔루션을 수정하고 평가합니다. 모든 마이그레이션 항목의 목록을 가져와서 사용 사례의 우선순위를 지정합니다. 마이그레이션 평가를 완료하면 일정을 보다 명확하게 파악하고 원래 계획된 일정에 맞출 수 있게 됩니다.

3. **전략 단계: 기술 계획을 수행합니다.**

   대상 아키텍처를 검토하고 장기적으로 비즈니스를 지원하는지 확인합니다. 이 단계에서는 수집 전략 및 기술, 추출, 변환, 로드(ETL) 패턴 및 도구, 레이크하우스의 데이터 구성 원칙, 의미 및 보고 계층 아키텍처와 도구 선택에 대한 중요한 결정을 내립니다.

4. **프로덕션 파일럿 단계: 평가 및 구현을 완료합니다.**

   새로운 플랫폼이 무엇을 제공해야 하는지 이해합니다. 접근방식을 검증하는 데 도움이 되는 표적 데모 또는 계획을 수행합니다.

5. **실행 단계: 마이그레이션을 실행합니다.**

   현장에 적용하고 해당 마이그레이션이 처음부터 올바로 작동하는지 확인합니다.

기억하세요

마이그레이션을 더 빨리 실행할수록 더 빨리 분석 능력을 확장하고 비용을 절감하고 전체 팀 생산성을 높일 수 있습니다.

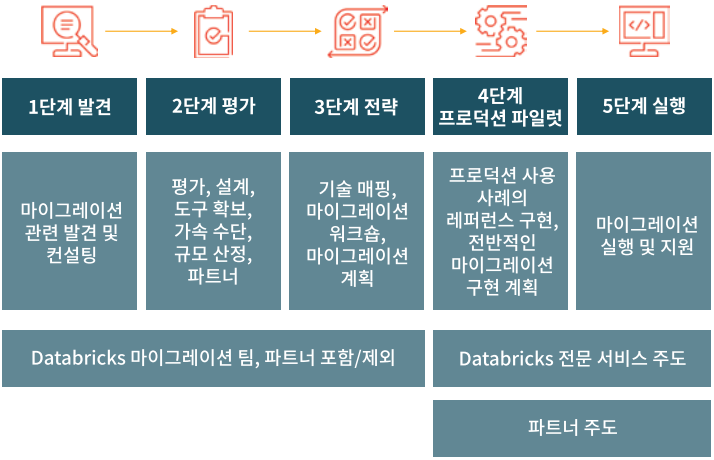


그림 5-1: 마이그레이션 방법론의 단계.

마이그레이션의 5가지 기둥

레이크하우스의 기본 아키텍처는 단일 프레임워크의 5가지 기둥을 따릅니다.

- **아키텍처/인프라:** 배포 아키텍처를 설정하고 보안 및 거버넌스 프레임워크를 구현합니다.
- **데이터 마이그레이션:** 데이터 구조 및 레이아웃을 매핑하고, 일회성 로드를 완료하고, 증분 로드 접근방식을 마무리합니다.
- **ETL 및 파이프라인:** 이 단계에서는 데이터 변환 및 파이프라인 코드, 오케스트레이션 및 작업을 마이그레이션합니다. 자동화 도구를 사용하고 결과를 온프레미스 데이터 및 예상 결과와 비교하여 마이그레이션 속도를 높입니다.
- **분석:** 비즈니스 분석 및 비즈니스 결과를 위한 보고서와 분석을 재지정합니다. 보고 의미 계층 및 온라인 분석 처리(OLAP) 큐브도 ODBC(Open Database Connectivity) 및 JDBC(Java Database Connectivity)를 통해 레이크하우스로 재지정해야 합니다.
- **데이터 과학/머신 러닝(ML):** ML 도구 및 온보드 데이터 과학 팀에 대한 연결을 설정합니다.

이 장에서 다루는 내용

- 미래에 대비한 투자
- 어디서나 최신 데이터 추출
- 모든 데이터에 대해 단일 소스 사용

6장
Databricks 레이크하우스로 마이그레이션해야 하는 10가지 이유

고객이 레이크하우스로 마이그레이션하기로 결정할 때 Databricks 레이크하우스를 플랫폼으로 선택해야 하는 10가지 이유가 있습니다. 그 이유는 다음과 같습니다.

- **고객은 개방적이고 유연한 플랫폼을 원합니다.** 플랫폼에 종속되는 것은 몇 년마다 마이그레이션하는 근본 원인입니다. 최신 데이터 스택과 원활하게 작동하는 개방형 멀티 클라우드, 고도로 혁신적인 레이크하우스 플랫폼을 사용하여 벤더에 종속되는 것을 피하고 미래에 대비한 투자를 보장할 수 있습니다. 미래 보장에 대한 자세한 내용은 3장을 참조하세요.
- **고객은 더 빠른 인사이트를 실현하기를 원합니다.** 모두를 위한 거의 실시간의 자체 서비스 분석을 통해 비즈니스를 지원하고 가장 완벽한 최신 데이터에서 새로운 인사이트를 발견할 수 있습니다.
- **고객은 운영 비용을 낮추고 모든 데이터에 대해 단일 거버넌스 계층을 구축하기를 원합니다.** 통합 아키텍처 및 거버넌스 모델을 구축합니다.
- **고객은 최고의 가격대 성능비를 원합니다.** 레이크하우스 아키텍처를 운영하면 다른 클라우드 데이터 웨어하우스보다 최대 12배 더 나은 가격대 성능비를 제공합니다.

- **고객은 어디서나 데이터를 쉽게 수집하고 최신 데이터에 액세스하기를 원합니다.** Databricks SQL은 장소에 관계없이 데이터를 처리합니다. Databricks에는 원활한 파일 수집을 위한 자동 로더 기능과 PartnerConnect에 내장된 많은 수집 파트너 도구 통합 기능이 있습니다. 이 통합 기능은 클라우드 스토리지 및 엔터프라이즈 데이터에서 Salesforce 또는 Marketo와 같은 엔터프라이즈 애플리케이션으로 데이터를 수집하는 턴키 기능을 제공합니다. 클릭 한 번이면 됩니다.

- **선택한 도구를 통한 최신 분석이 필요합니다.** Databricks SQL은 dbt, Tableau, Power BI, Looker와 같은 가장 널리 사용되는 비즈니스 인텔리전스(BI) 및 SQL 도구와 원활하게 작동합니다. 따라서 분석가는 자신이 선호하는 도구를 사용하여 가장 완전한 최신 데이터에서 새로운 비즈니스 인사이트를 발견할 수 있습니다.

- **최고의 SQL 개발 환경을 제공합니다.** 분석가가 Databricks SQL 쿼리 편집기를 통해 익숙한 구문(ANSI SQL)으로 쿼리를 작성하고 레이크하우스에 있는 데이터를 쉽게 탐색할 수 있습니다. 분석가가 다양하고 풍부한 시각화를 통해 쿼리 결과를 쉽게 이해하고 대시보드를 신속하게 구축하고 이해관계자와 공유할 수 있습니다.

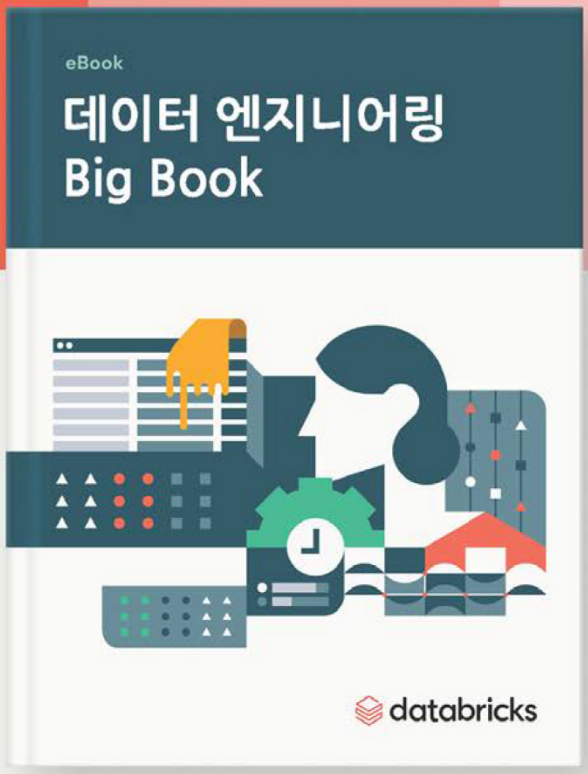
- **인프라 관리가 필요 없습니다.** 서버리스 SQL 컴퓨팅으로 비용을 절감하고 클라우드 인프라를 관리, 구성 또는 확장할 필요가 없습니다. 이를 통해 데이터 팀이 가장 잘하는 일에 집중할 수 있습니다.

- **레이크하우스에서 세분화된 거버넌스를 실행할 수 있습니다.** 세분화된 거버넌스를 통해 레이크하우스의 데이터 액세스를 확실히 관리하고 보호할 수 있습니다. 또한 모든 데이터 자산에 대한 데이터 계보, 역할 기반 보안 정책, 테이블 또는 열 수준 태그를 통해 규제 요구사항을 충족할 수 있습니다.

- **모든 데이터에 대해 단일 소스를 가집니다.** 엔터프라이즈 데이터 웨어하우스(EDW)와 달리 Databricks 레이크하우스 플랫폼은 기존 데이터 레이크의 모든 데이터 유형에 대해 단일 공용 스토리지 및 데이터 관리 프레임워크를 제공합니다.

팁

Databricks 전체 성능 벤치마크는 다음 링크를 참조하세요: databricks.com/product/databricks-sql.



데이터 엔지니어링 방법 가이드 다운로드

코드 샘플, 노트북 및 사용 사례가 있는 기술 블로그 컬렉션

이 eBook에서 다루는 내용

- 원시 데이터를 실행 가능한 데이터로 바꾸는 법 - 데이터 세트, 코드 샘플 및 리더와 전문가가 전하는 모범 사례 포함
- Databricks 레이크하우스 플랫폼의 데이터 수명 주기 - 데이터 수집부터 데이터 처리, 분석 및 머신 러닝까지
- J.B.Hunt, ABN, AMRO 및 Atlassian과 같은 선도 기업의 전체적인 실제 사용 사례

databricks

eBook 다운로드 →

레이크하우스로 데이터 전략을 발전시키십시오.

데이터는 현대 비즈니스에 가장 귀중한 자원입니다. 데이터 가용성이 계속해서 급증함에 따라, 기업들은 여러 데이터 소스에 산재한 방대한 양의 데이터를 관리하는 데 필요한 업무를 처리하기 위해 고군분투하고 있습니다. 데이터 레이크하우스를 사용하면 모든 데이터를 한 곳에서 관리하고 분석할 수 있으며, 며칠이 아닌 몇 분 만에 분석 애플리케이션을 구축, 배포, 확장할 수 있습니다.

내용

- 고유한 데이터 유형을 관리하는 방법
- 회사의 데이터 전략 평가
- 레이크하우스의 이점
- 마이그레이션 계획

스테파니 다이아몬드 씨는 전 AOL 마케팅 이사이자 기업이 숨겨진 수익을 찾도록 돕는 온라인 마케팅 회사인 Digital Media Works의 설립자입니다. 그녀는 *Facebook Marketing For Dummies*를 포함하여 25권 이상의 마케팅 서적과 맞춤형 전자책을 저술했습니다.

Dummies.com으로 이동하여 동영상, 단계별 사진, 사용법 정보를 보거나 구매하십시오!

ISBN: 978-1-394-16168-3
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.