



현장의 ETL에서 클라우드 기반 ELT로 전환

데이터 파이프라인의 가치와 효율성을 극대화하기 위한 모범 사례



챔피언
가이드

TABLE OF CONTENTS

- 2 종합 요약
- 3 기본 용어 및 개념 이해
- 4 ELT의 부상
- 6 다용도 데이터 관리 전략 수립
- 7 언제 ELT를 고려해야 할까요?
- 7 데이터 파이프라인 선택
- 9 Snowflake로 데이터 처리
- 10 결론
- 11 Snowflake 소개

종합 요약

이전의 데이터 파이프라인은 온프레미스 비즈니스 애플리케이션에서 예측 가능하고 느리게 이동하며 쉽게 범주화된 데이터를 수용하기 위해 설계되었습니다. 이는 다양한 소스에서 데이터를 캡처하고, 이를 유용한 형식으로 변환하고, 데이터 웨어하우스와 같은 타겟 대상에 로드하는 ETL (추출, 변환 및 로드) 프로세스에 의존합니다. 이러한 레거시 파이프라인은 엔터프라이즈 애플리케이션에서 가져온 정형 데이터 소스에 대해서는 잘 작동하지만 최신 데이터 환경을 특징짓는 다양한 데이터 유형 및 수집 스타일에는 더 이상 적합하지 않습니다.

오늘날의 최신 파이프라인은 데이터를 먼저 추출하고 로드한 다음 데이터가 대상에 도달하면 변환하도록 설계되어 있습니다(ELT로 알려진 주기). 최신 ELT 시스템은 변환 워크로드를 클라우드로 전환하여 훨씬 뛰어난 확장성과 유연성을 지원합니다. 기존 온프레미스 환경에서 ETL 작업은 동일한 인프라에서 실행되는 다른 워크로드와 리소스를 놓고 경합합니다. ELT를 사용하면 데이터를 원시 형식으로 로드한 다음, 데이터를 사용할 방식이 명확해지면 여러 방식으로 변환할 수 있습니다.

ELT 파이프라인을 사용하면, 여러 유형의 원시 데이터를 클라우드 데이터 플랫폼과 같은 클라우드 기반 저장소로 로드할 수 있습니다. 이 플랫폼은 조직 전반에 걸쳐 데이터를 수집, 변환 및 공유할 수 있는 속도를 개선합니다. 이를 통해 리소스 집약적인 변환 워크로드를 클라우드로 실행할 수 있으므로, 확장 가능한 클라우드 리소스의 처리 능력과 용량을 극대화할 수 있습니다.

다음 페이지에서 보듯이, ELT는 다음과 같은 상황에서 좋은 선택입니다.

- **방대한 데이터 요구 사항이 있는 경우:** ELT는 클라우드에서 대량의 정형 및 비정형 데이터를 빠르게 처리할 수 있습니다.
- **분석 실험의 경우:** ELT는 분석가와 데이터 과학자가 데이터의 잠재력을 탐색하여 이를 특정 프로젝트에 필요한 대로 변환하는 과정에서 옵션을 극대화합니다.
- **낮은 대기 시간 데이터 파이프라인의 경우:** ELT는 데이터를 즉시 전송하므로 낮은 대기 시간 분석 및 준실시간 사용 사례에 유용할 수 있습니다.

각 상황 및 워크로드에 적합한 유형의 변환 방법을 사용하여 귀사 데이터 파이프라인의 가치를 극대화할 방법을 알아보려면 계속 읽으십시오.



기본 용어 및 개념 이해

ETL은 다양한 소스에서 데이터를 추출하고, 스테이징 서버에서 데이터를 변환하고, 데이터 웨어하우스, 데이터 레이크 또는 클라우드 데이터 플랫폼과 같은 타겟 대상에 데이터를 로드하는 것을 뜻하는 소프트웨어 통합 프로세스입니다. 기존

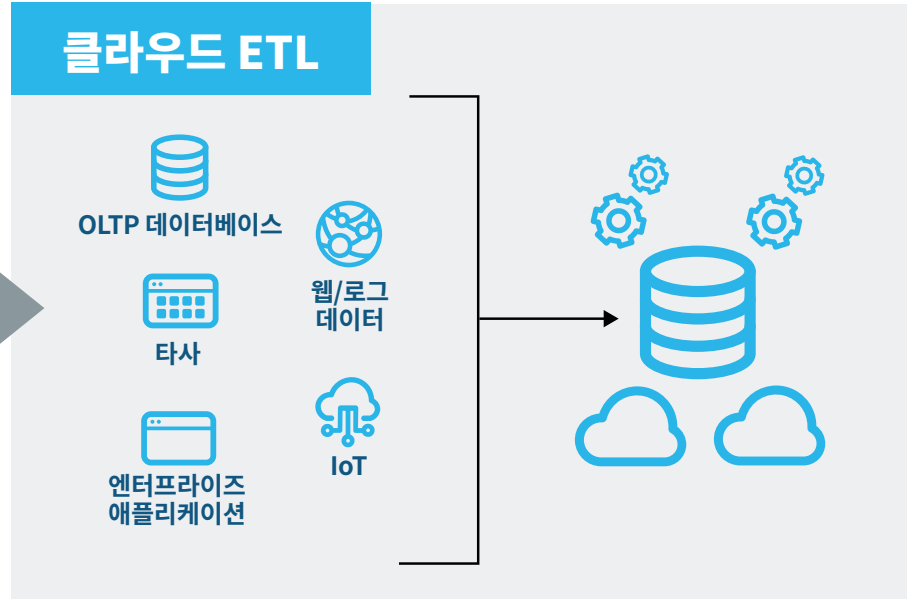
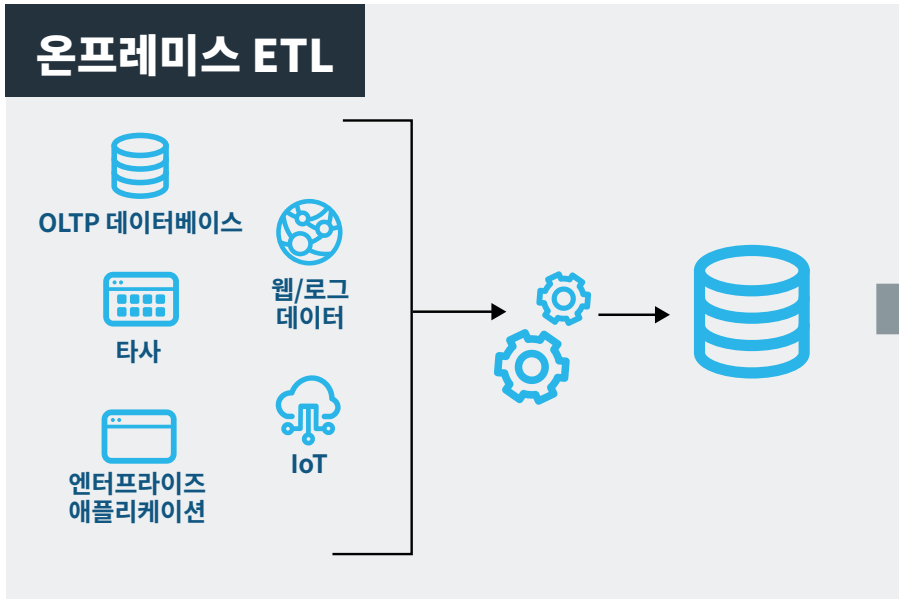
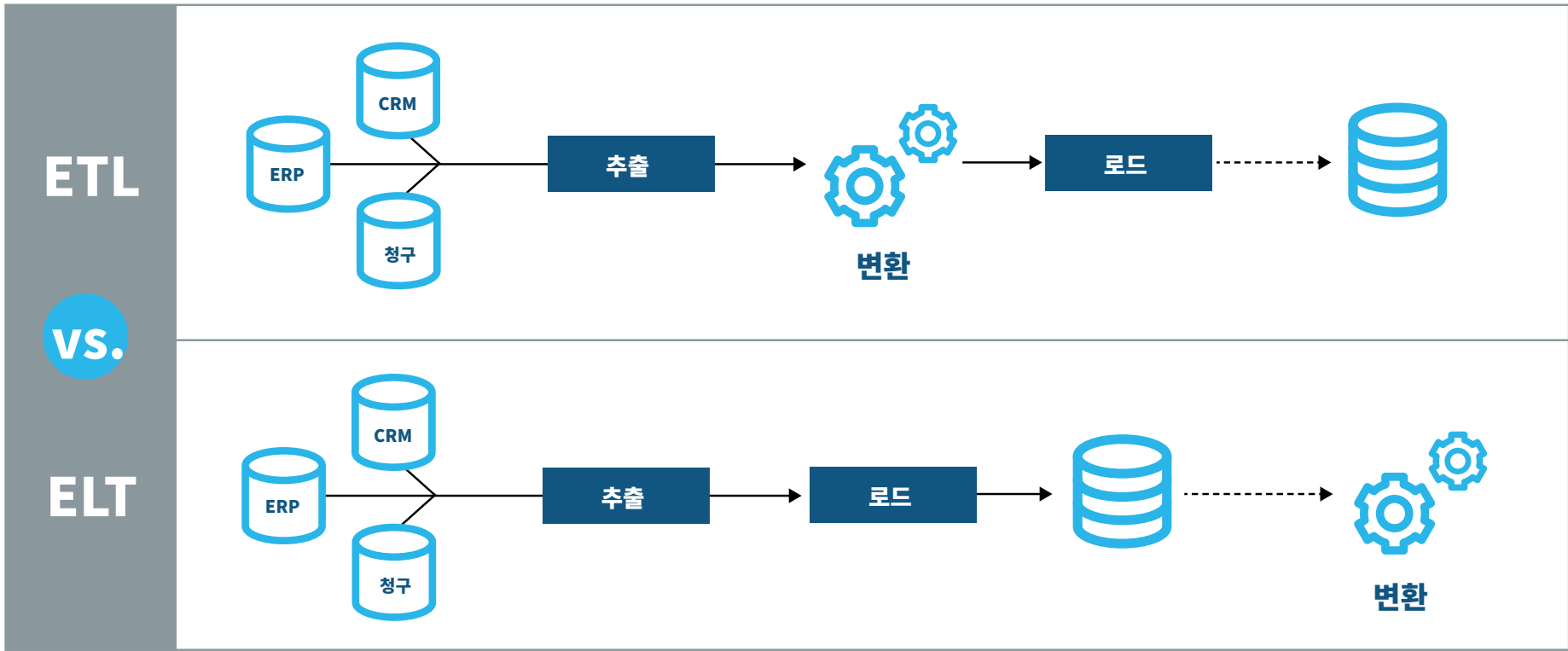
데이터 웨어하우스에서, 데이터는 관계형 데이터 모델에 맞게 매핑됩니다. 대상 데이터베이스에 로드되기 전에 정리, 강화 및 공통 형식으로 변환될 수도 있습니다.

그 데이터를 구조화하고 변환하면 SQL 기반 비즈니스 인텔리전스(BI) 도구를 사용하여 빠르고 효율적으로 분석할 수 있지만, 변환 과정에서 원시 데이터의 일부 아티팩트가 손실되기 때문에 데이터 사용 방법이 제한됩니다. 대부분의 ETL 워크플로우에서 데이터는 원본 데이터베이스에서 캡처되어 데이터 웨어하우스로 스테이징됩니다. 스테이징 서버는 필터링, 마스킹, 강화,

매핑, 중복 제거 및 여러 소스의 데이터 통합을 포함할 수 있는 변환 로직을 실행합니다.

데이터 엔지니어는 데이터 파이프라인을 생성하여 일괄 데이터 업로드의 이동을 조정하고 데이터를 지속적으로 스트리밍합니다. 이러한 파이프라인은 애플리케이션, 장치 및 이벤트 스트림에서 데이터를 추출합니다. ETL 파이프라인은 기본 ETL 워크플로우의 일부로서 데이터를 비즈니스 준비 형식으로 변환합니다. 비즈니스 요구 사항이 명확할 때는 괜찮습니다. 그러나 머신 러닝 및 데이터 과학 등 오늘날 널리 사용되는 일부 워크로드의 경우, 데이터 형식 요구 사항이 항상 사전에 알려져 있지 않습니다. 예를 들어, 데이터 과학자는 데이터를 원시(또는 덜 처리된) 상태로 유지한 다음 다양한 유형의 모델, 예측 엔진 및 분석 시나리오를 수용하기 위한 다양한 형식으로 변환하는 것을 선호할 수 있습니다.





다용도 데이터 관리 전략 수립

ETL은 데이터가 예측 가능하고 관리 가능하며 정기적인 간격으로 업데이트되는 경우 실행 가능한 옵션입니다. 이러한 일괄 수집 프로세스는 지속적으로 새로 고쳐질 필요가 없는 애플리케이션 데이터에 일반적으로 사용됩니다. 예를 들어 소매 POS 데이터는 일일 수익 보고서를 반영하기 위해 하루를 마칠 때 데이터 웨어하우스에서 업데이트될 필요가 있습니다. CRM 시스템의 고객 데이터는 현재 판매 및 서비스 거래를 반영하기 위해 시간당 한 번씩 콜 센터 대시보드에 업로드될 필요가 있습니다. 스마트 미터에서 수집된 전기 사용량 데이터는 사용 시간 청구 프로그램을 지원하기 위해 15분마다 새로 고쳐질 필요가 있습니다.

그러나 상황이 빠르게 복잡해질 수 있습니다. ETL 시스템 및 레거시 데이터 아키텍처를 사용하면, 개별 시스템의 데이터가 여러 서로 다른 위치에 격리됩니다. 예를 들어, 각 데이터 유형은 특정 요구 사항에 맞게 설계 및 모델링된 고유한 시스템에 빠지게 될 수 있습니다. 그 결과 여러 서로 다른 리포지토리가 생성되어 유지 관리의 악몽으로 빠르게 변할 수 있습니다. 온프레미스 또는 클라우드에서, 각 생산 애플리케이션은 마케팅 자동화 시스템의 마케팅 데이터, CRM 시스템의 판매 데이터, ERP 시스템의 재무 데이터, 창고 관리 시스템의 재고 데이터 등 자체 데이터 사일로를 만듭니다. 이러한 각 앱은 생산 시스템에서 데이터를 수집하고 분석을 위해 해당 데이터를 변환하기 위한 특수화된 ETL 도구와 고유한 소프트웨어 절차에 의존할 수 있습니다.

오늘날의 비즈니스 및 분석 요구 사항을 고려할 때, 모든 데이터를 다양한 작업 그룹, 응용 프로그램 및 도구에서 보편적으로 액세스할 수 있도록 설계된 단일 진실 공급원으로서 하나의 장소에 통합해야 합니다.



언제 ETL을 고려해야 할까요?

ETL 기술은 일괄 처리 모드에서 데이터가 한 시스템에서 다른 시스템으로 이동되는 상황에 널리 사용됩니다. 그러나 대부분의 기존 ETL 솔루션이 모든 유형의 데이터를 처리할 수는 없습니다. 엔터프라이즈 애플리케이션의 정형 데이터에는 잘 작동하지만, IoT 시스템의 머신 생성 데이터, 소셜 미디어 피드의 스트리밍 데이터, JSON 이벤트 데이터, 인터넷 및 모바일 앱의 웹로그 데이터에는 적합하지 않습니다.

어떤 접근 방식을 사용할지 결정하려면 다음 기본 지침을 기억하십시오.

- ETL 프로세스는 테이블 형식 구조를 유지해야 하는 관계형 데이터에 적합합니다.
- ETL은 특정 분석 사용 사례가 고안될 때까지 원시 또는 기본 형식으로 유지되어야 하는 반정형 데이터에 더 좋은 방법입니다.

다른 고려 사항으로는 처리할 데이터의 양과 다운스트림 분석을 위해 데이터가 얼마나 빨리 준비되어야 하는지가 있습니다. 변환 프로세스에는 여러 번의 계산 주기가 필요합니다. ETL을 사용하면, 자동 확장성이 각 작업을 지원하는 데 필요한 리소스를 즉시 프로비저닝합니다. ETL 프로세스를 사용하면 클라우드의 무한한 리소스를 사용하여 데이터를 빠르고 효율적으로 처리하고 변환할 수 있습니다. 또한 데이터를 독립 서버나 스토리지 메커니즘으로 이동하는 대신 데이터가 있는 곳에서 처리할 수 있으므로 데이터 이동을 최소화합니다.

처리 엔진을 실행할 위치, 사용 가능한 인프라 리소스 및 필요한 성능을 정하십시오. 제한된 서버 용량과 같은 확장성 또는 동시성 문제가 있습니까? 데이터가 생성되는 곳이 OLTP(온라인 거래 처리) 시스템, 웹사이트 상호 작용, SaaS 앱, 장비 센서든 또는 소셜 미디어 스트림이든 관계없이 데이터 엔지니어는 데이터 파이프라인을 개발하여 해당 데이터를 캡처하고, 데이터 저장소로 수집하고, 이를 비즈니스 커뮤니티에 액세스할 수 있도록 만들어야 합니다. 많은 경우, 클라우드에서 대량 데이터베이스의 처리 능력을 활용함으로써 데이터 파이프라인 작업이 향상됩니다.

데이터 파이프라인 선택

다음과 같은 경우 ETL을 사용합니다.

- 처리해야 하는 데이터의 총 크기가 상대적으로 작을 때
- 소스와 타겟의 데이터베이스가 서로 다른 데이터 유형을 필요로 할 때
- 주로 정형 데이터를 처리하고 있을 때

다음과 같은 경우 ETL을 사용하지 않습니다.

- 다량의 데이터를 처리해야 할 때
- 소스와 타겟의 데이터베이스가 동일한 유형일 때
- 데이터가 반정형 또는 비정형일 때



SNOWFLAKE 고객 사례 연구

조직: Paciolan은 해마다 1억 2천만 장 이상의 티켓을 판매하는 500개 이상의 라이브 공연 조직을 지원하는 티켓팅, 모금, 마케팅, 분석 및 기술 솔루션의 선두 주자입니다.

문제: 반정형 데이터를 관계형 데이터로 변환하기 위해 Paciolan은 데이터를 구문 분석하고 정규화하는 독점적인 ETL 코드를 작성했습니다. 온프레미스 데이터 웨어하우스에서 5만 개의 레코드가 1백만 개의 행으로 바뀔 수 있었습니다. 매일 100GB에 가까운 데이터로 구성된 ETL 프로세스를 완료하는 데 30~60분이

걸렸습니다. 제한된 리소스로 인해 분석가가 데이터를 효과적으로 요약하고 롤업할 수 없었습니다.

솔루션: Paciolan은 이제 반정형 JSON 데이터를 Snowflake 플랫폼에 VARIANT 데이터 유형으로 저장합니다. 현대적이고 민첩한 엔터프라이즈 데이터 웨어하우스를 지원하기 위한 특정 데이터 모델 디자인 패턴을 포함하는 아키텍처 접근 방식인 Data Vault를 통해, Snowflake를 데이터 레이크 및 데이터 웨어하우스 둘 다로 사용합니다.

결과: 레거시 데이터 웨어하우스에서 완료하는 데 한 시간 정도 걸리던 ETL 프로세스가 Snowflake 데이터 파이프라인을 사용하면 이제 단 몇 분이면 됩니다. 개발자는 간단한 Python 스크립트를 사용하여 명령문을 동적으로 삽입할 수 있습니다.

혜택

- 스토리지와 컴퓨팅의 분리가 성능 안정성과 비용 가시성 제공
- 즉각적인 탄력성으로 사실상 모든 사용자에게 대해 거의 무제한에 가까운 컴퓨팅 성능 구현
- 반정형 데이터를 다양한 데이터 유형으로 저장할 수 있도록 지원하여 더 풍부한 데이터 통찰 제공

“전후 수치를 비교한 결과 Snowflake를 사용하면 ETL 프로세스에 사용되는 코드가 90% 감소한 것으로 나타났습니다. 이는 우리에게 대규모 수익을 의미합니다.”

Ashkan Khoshcheshmi
수석 소프트웨어 엔지니어
Paciolan



SNOWFLAKE로 데이터 처리

Snowflake 플랫폼에는 기본 서비스의 일부로 유연하고 확장 가능한 데이터 파이프라인 기능이 포함되어 있습니다. 원시 데이터를 Snowflake로 바로 수집할 수 있으므로 데이터를 다른 형식으로 변환하기 위한 파이프라인을 만들 필요가 없습니다. Snowflake는 이러한 변환을 자동으로 수행하여, 스토리지 및 컴퓨팅 비용을 최소화합니다.

Snowflake는 또한 다음과 같이 데이터 사일로를 제거하여 데이터 관리를 단순화합니다. 여러 다운스트림 애플리케이션에 대해 여러 데이터 복사본을 유지 관리할 필요가 없습니다. 원시 데이터의 원래 형태를 유지하지만 고도로 최적화된 저장 기술을 투명하게 적용하여 분석 및 데이터 변환이 아주 잘 이뤄집니다.

가장 중요한 것은 Snowflake가 클라우드의 고유한 속성을 최대한 활용하도록 설계되었다는 점입니다. 이는 대규모 데이터 변환을 수용하기 위해 컴퓨팅 리소스와 스토리지 리소스를 분리하는 멀티 클러스터 및 공유 데이터 아키텍처를 기반으로 합니다. 각 리소스 유형은 개별 애플리케이션의 특정 요구 사항을 수용하기 위해 독립적으로 확장될 수 있습니다.

Snowflake의 플랫폼은 데이터 엔지니어링 파이프라인을 통해 데이터를 수집하는 동시에 동일한 데이터를 사용하도록 머신 러닝 모델을 교육하는 것과 같이, 성능 저하 없이 모든 유형의 워크로드를 처리하도록 설계된 강력한 처리 엔진을 기반으로 구축되었습니다. 확장 가능한 파이프라인 서비스는 이러한 다른 워크로드의 성능에 영향을 주지 않고 지속적으로

데이터를 수집할 수 있습니다. 데이터 엔지니어가 각 데이터 수집 프로세스에 할당할 컴퓨팅 성능을 결정할 수도 있고 또는 시스템이 자동으로 확장되도록 할 수도 있습니다.

Snowflake는 또한 데이터 엔지니어가 다양한 언어 및 수집 스트림 관리를 위한 통합 도구를 사용하여 데이터 파이프라인을 구축할 수 있도록 합니다. 일괄 통합 및 Apache Kafka와의 스트리밍 통합을 포함해 널리 사용되는 다양한 데이터 수집 스타일을 지원할 수 있습니다. 또한 Snowflake를 사용하면 데이터 처리의 공용어인 표준 SQL을 사용하여 다양한 유형의 데이터를 쉽고 효율적으로 수집할 수 있습니다.



결론

데이터의 양, 다양성 및 속도가 꾸준히 증가함에 따라 새로운 유형의 데이터 파이프라인이 필요하고 데이터를 캡처하여 작동하도록 하는 보다 발전된 클라우드 기반 데이터 처리 엔진이 필요합니다.

ETL 프로세스는 종종 고정 용량, 제한된 대역폭 및 유한한 CPU 주기 세트를 갖춘 온프레미스 서버에서 처리됩니다. 최신 데이터 통합 워크로드는 클라우드 데이터베이스 및 마음대로 확장할 수 있는 클라우드 데이터 플랫폼의 처리 능력을 활용하여 향상됩니다.

이러한 클라우드 리소스를 활용하기 위해 점점 더 많은 조직에서 데이터를 추출하여 클라우드 데이터베이스에 로드한 다음 데이터가 대상에 도달하면 변환하는, ELT 라고 알려진 주기의 데이터 파이프라인을 설계하고 있습니다.

이 접근 방식은 최신 데이터 처리 엔진의 성능을 활용하고 불필요한 데이터 이동을 줄이기 때문에 기존 ETL 프로세스보다 빠릅니다.

ELT 프로세스는 다음 두 가지 주요 이유를 위해 리소스 집약적인 변환 워크로드를 클라우드로 푸시합니다.

1. 클라우드의 거의 무제한 리소스를 사용하여 데이터를 빠르고 효율적으로 처리하고 변환
2. 비즈니스 요구 사항을 완전히 이해할 때까지 데이터를 원시 상태로 유지

가능하면 ETL 대신 ELT를 사용하여 리소스 집약적인 변환 작업을 클라우드 기반 대상 플랫폼으로 푸시합니다. 이 접근 방식은 귀사의 데이터 파이프라인을 단순화하고, 데이터 이동을 최소화하며, 데이터 사일로의 수를 줄이고, 데이터가 궁극적으로 사용되는 방법에 대한 선택지를 최대화합니다.

Snowflake의 데이터 파이프라인 솔루션에 대해 자세히 알아보려면 [Snowflake.com/workloads/data-engineering](https://www.snowflake.com/workloads/data-engineering)을 방문하세요.





SNOWFLAKE 소개

Snowflake가 제공하는 데이터 클라우드의 거의 무제한의 규모, 동시성, 성능을 통해 수천 개의 조직이 데이터를 모으는 글로벌 네트워크입니다. 데이터 클라우드 내에서 조직은 사일로된 데이터를 통합하고, 관리형 데이터를 쉽게 검색하고 안전하게 공유하며, 다양한 분석 워크로드를 실행합니다. 데이터나 사용자가 어디에 있든 Snowflake는 여러 공용 클라우드에서 단일하고 원활한 경험을 제공합니다. Snowflake의 플랫폼은 데이터 클라우드에 대한 액세스를 지원하고 제공하는 엔진입니다. 데이터 클라우드에서는 데이터 웨어하우징, 데이터 레이크, 데이터 엔지니어링, 데이터 사이언스, 데이터 애플리케이션 개발 및 데이터 공유를 위한 솔루션을 만듭니다. 이미 데이터 클라우드의 새로운 영역으로 비즈니스를 추진하고 있는 Snowflake 고객, 파트너 및 데이터 공급자에 합류하십시오. [Snowflake.com](https://www.snowflake.com).



©2022 Snowflake Inc. All rights reserved. 여기에 언급된 Snowflake, Snowflake 로고 및 기타 모든 Snowflake 제품, 기능 및 서비스 이름은 미국 및 기타 국가에서 Snowflake Inc.의 등록 상표 또는 상표입니다. 여기에 언급되거나 사용된 기타 모든 브랜드 이름 또는 로고는 식별 목적으로만 사용되며 해당 소유자의 상표일 수 있습니다. Snowflake는 그러한 소유자와 연관되거나 후원 또는 보증을 받지 않습니다.