**AZ-레이크하우스의 원칙

레이크하우스의 원칙 안내

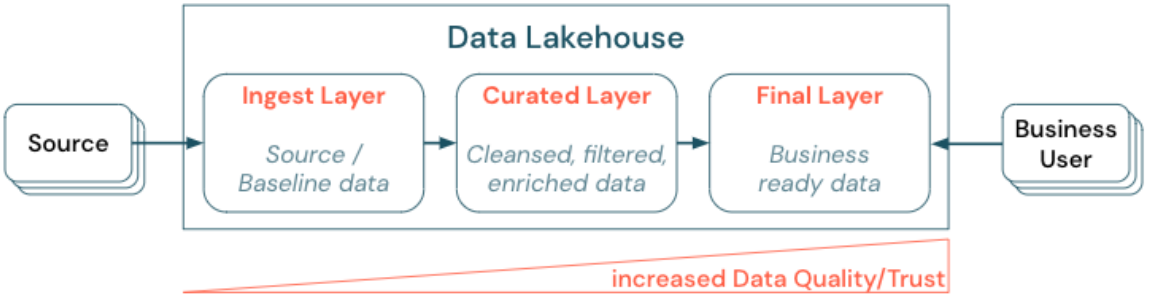
이 문서의 내용

1.
데이터를 큐레이팅하고 신뢰할 수 있는 데이터 제품 제공
2.
데이터 사일로 제거 및 데이터 이동 최소화
3.
셀프 서비스를 통해 가치 창출 민주화
4.
조직 전체의 데이터 거버넌스 전략 채택**2개 더 표시**

지침 원칙은 아키텍처를 정의하고 영향을 주는 수준 0 규칙입니다. 현재와 미래의 비즈니스 성공을 돕는 데이터 레이크하우스를 구축하려면 조직의 이해 관계자 간의 합의가 중요합니다.

데이터를 큐레이팅하고 신뢰할 수 있는 데이터 제품 제공

데이터를 큐레이팅하는 것은 BI 및 ML/AI용 고부가가치 데이터 레이크를 만드는 데 필수적입니다. 명확한 정의, 스키마 및 수명 주기를 사용하여 데이터를 제품처럼 처리합니다. 비즈니스 사용자가 데이터를 완전히 신뢰할 수 있도록 의미 체계 일관성을 보장하고 데이터 품질이 계층에서 계층으로 개선되도록 합니다.



데이터 팀이 품질 수준에 따라 데이터를 구조화하고 계층당 역할 및 책임을 정의할 수 있으므로 계층화된(또는 다중 홉) 아키텍처를 설정하여 데이터를 큐레이팅하는 것은 레이크하우스에 중요한 모범 사례입니다. 일반적인 계층화 방법은 다음과 같습니다.

- **수집 계층**: 원본 데이터가 레이크하우스에 첫 번째 계층으로 수집되고 이 계층에 유지되어야 합니다. 수집 계층에서 모든 다운스트림 데이터를 만들 때 필요한 경우 이 계층에서

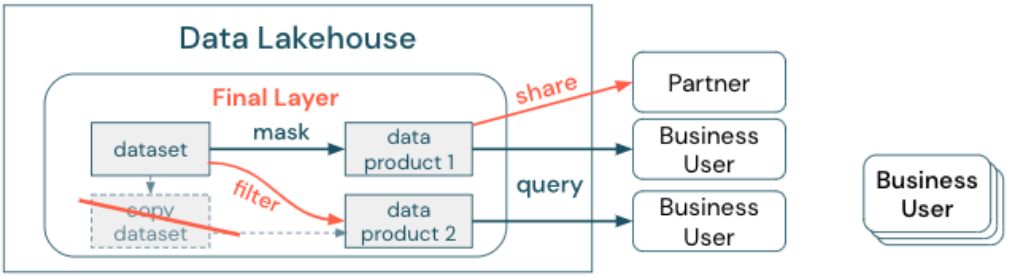
후속 계층을 다시 빌드할 수 있습니다.

- **큐레이팅된 계층**: 두 번째 계층의 목적은 정리되고, 구체화되고, 필터링되고, 집계된 데이터를 보유하는 것입니다. 이 계층의 목표는 모든 역할 및 함수에서 분석 및 보고서를 위한 건전하고 신뢰할 수 있는 기반을 제공하는 것입니다.
- **최종 계층: 세 번째 계층**은 비즈니스 또는 프로젝트 요구 사항을 중심으로 만들어집니다. 다른 사업부 또는 프로젝트에 대한 데이터 제품으로 다른 보기를 제공하고, 보안 요구 사항(예: 익명화된 데이터)을 중심으로 데이터를 준비하거나, 성능 최적화(미리 집계된 뷰 포함)를 제공합니다. 이 계층의 데이터 제품은 비즈니스의 진실로 간주됩니다.

모든 계층의 파이프라인은 데이터 품질 제약 조건이 충족되도록 해야 합니다. 즉, 데이터가 동시 읽기 및 쓰기 중에도 항상 정확하고 완전하며 액세스 가능하며 일관성이 있어야 합니다. 새 데이터의 유효성 검사는 큐레이팅된 계층에 데이터를 입력할 때 수행되며 다음 ETL 단계는 이 데이터의 품질을 개선하기 위해 작동합니다. 데이터가 계층을 통해 진행됨에 따라 데이터 품질이 개선되어야 하며, 따라서 데이터에 대한 신뢰가 비즈니스 관점에서 증가합니다.

데이터 사일로 제거 및 데이터 이동 최소화

이러한 서로 다른 복사본을 사용하는 비즈니스 프로세스가 있는 데이터 세트의 복사본을 만들지 마세요. 복사본은 동기화되지 않는 데이터 사일로가 되어 데이터 레이크의 품질이 낮아지고 마지막으로 오래된 정보나 잘못된 인사이트가 될 수 있습니다. 또한 외부 파트너와 데이터를 공유하려면 안전한 방식으로 데이터에 직접 액세스할 수 있는 엔터프라이즈 공유 메커니즘을 사용합니다.



데이터 복사와 데이터 사일로를 구분하기 위해 독립 실행형 또는 버려진 데이터 복사본은 자체적으로 해롭지 않습니다. 민첩성, 실험 및 혁신을 촉진하는 데 필요한 경우도 있습니다. 그러나 이러한 복사본이 종속된 다운스트림 비즈니스 데이터 제품으로 작동하면 데이터 사일로가 됩니다.

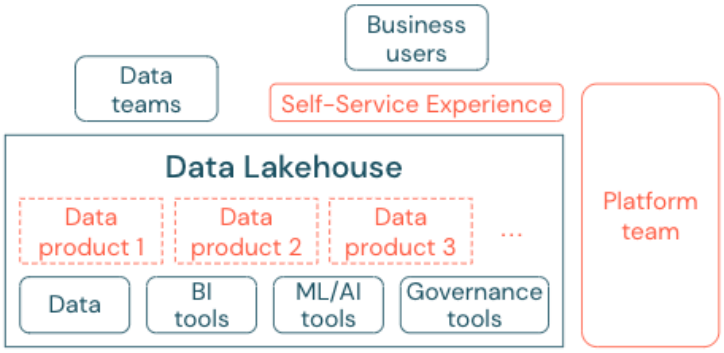
데이터 사일로를 방지하기 위해 데이터 팀은 일반적으로 모든 복사본을 원본과 동기화된 상태로 유지하는 메커니즘 또는 데이터 파이프라인을 빌드하려고 합니다. 이는 일관되게 발생할 가능성이 낮기 때문에 결국 데이터 품질이 저하됩니다. 이로 인해 비용이 높아지고 사용

자의 신뢰가 크게 손실 될 수 있습니다. 반면, 여러 비즈니스 사용 사례에는 파트너 또는 공급업체와 데이터 공유가 필요합니다.

중요한 측면은 최신 버전의 데이터 세트를 안전하고 안정적으로 공유하는 것입니다. 데이터 세트의 복사본은 빠르게 동기화되지 않을 수 있으므로 충분하지 않은 경우가 많습니다. 대신 엔터프라이즈 데이터 공유 도구를 통해 데이터를 공유해야 합니다.

셀프 서비스를 통해 가치 창출 민주화

사용자가 BI 및 ML/AI 작업에 대한 플랫폼 또는 데이터에 쉽게 액세스할 수 없는 경우 최상의 데이터 레이크는 충분한 가치를 제공할 수 없습니다. 모든 사업부의 데이터 및 플랫폼 액세스 장벽을 낮춥니다. 린 데이터 관리 프로세스를 고려하고 플랫폼 및 기본 데이터에 대한 셀프 서비스 액세스를 제공합니다.



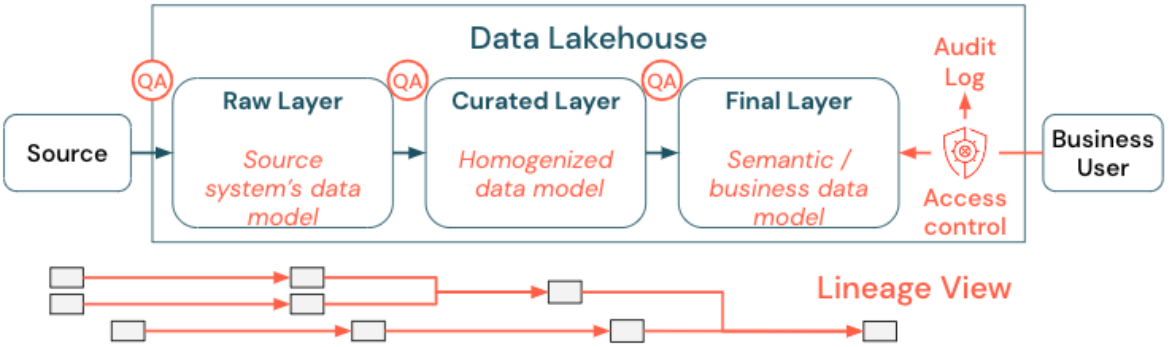
데이터 기반 문화로 성공적으로 이전한 기업은 번창할 것입니다. 즉, 모든 사업부는 분석 모델 또는 자체 또는 중앙에서 제공된 데이터를 분석하여 의사 결정을 도출합니다. 소비자의 경우 데이터를 쉽게 검색하고 안전하게 액세스할 수 있어야 합니다.

데이터 생산자에게 좋은 개념은 "제품으로서의 데이터"입니다. 데이터는 제품과 같은 한 사업부 또는 비즈니스 파트너가 제공하고 유지 관리하며 적절한 권한 제어를 가진 다른 당사자가 사용합니다. 중앙 팀과 잠재적으로 느린 요청 프로세스에 의존하는 대신 셀프 서비스 환경에서 이러한 데이터 제품을 만들고, 제공하고, 검색하고, 사용해야 합니다.

그러나 중요한 것은 데이터만이 아닙니다. 데이터의 민주화를 위해서는 모든 사람이 데이터를 생성하거나 사용하고 이해할 수 있는 올바른 도구가 필요합니다. 이를 위해 데이터 레이크하우스는 다른 도구 스택을 설정하는 노력을 복제하지 않고도 데이터 제품을 빌드하기 위한 인프라 및 도구를 제공하는 최신 데이터 및 AI 플랫폼이어야 합니다.

조직 전체의 데이터 거버넌스 전략 채택

데이터는 모든 조직의 중요한 자산이지만 모든 사용자에게 모든 데이터에 대한 액세스 권한을 부여할 수는 없습니다. 데이터 액세스는 적극적으로 관리되어야 합니다. 액세스 제어, 감사 및 계보 추적은 데이터의 정확하고 안전한 사용을 위한 핵심입니다.



데이터 거버넌스는 광범위한 주제입니다. 레이크하우스는 다음과 같은 차원을 다룹니다.

- **데이터 품질**

  정확하고 의미 있는 보고서, 분석 결과 및 모델의 가장 중요한 필수 조건은 고품질 데이터입니다. QA(품질 보증)는 모든 파이프라인 단계에 존재해야 합니다. 이를 구현하는 방법의 예로는 데이터 계약, SLA 모임, 스키마 안정적 유지, 제어된 방식으로 진화 등이 있습니다.

- **데이터 카탈로그**

  또 다른 중요한 측면은 데이터 검색입니다. 특히 셀프 서비스 모델에서 모든 비즈니스 영역의 사용자는 관련 데이터를 쉽게 검색할 수 있어야 합니다. 따라서 레이크하우스에는 모든 비즈니스 관련 데이터를 포함하는 데이터 카탈로그가 필요합니다. 데이터 카탈로그의 기본 목표는 다음과 같습니다.

  - 동일한 비즈니스 개념이 비즈니스 전체에서 균일하게 호출되고 선언되는지 확인합니다. 큐레이팅된 최종 계층에서 의미 체계 모델로 생각할 수 있습니다.
  - 사용자가 이러한 데이터가 현재 모양과 형태에 어떻게 도착했는지 설명할 수 있도록 데이터 계보를 정확하게 추적합니다.
  - 데이터의 적절한 사용을 위해 데이터 자체만큼 중요한 고품질 메타데이터를 유지 관리합니다.

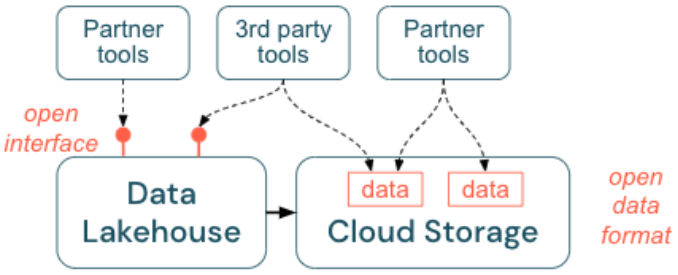
- **액세스 제어**

  레이크하우스의 데이터에서 가치를 창출하는 것이 모든 비즈니스 영역에서 이루어지기 때문에 레이크하우스는 일류 시민으로서 보안을 통해 건설되어야 합니다. 기업은 보다 개방적인 데이터 액세스 정책을 사용하거나 최소 권한의 원칙을 엄격하게 따를 수 있습니다. 데이터 액세스 제어와는 별개로 모든 계층에 데이터 액세스 제어가 있어야 합니다. 처음부터 고급 권한 체계(열 및 행 수준 액세스 제어, 역할 기반 또는 특성 기반 액세스 제어)를 구현하는 것이 중요합니다. 기업은 덜 엄격한 규칙으로 시작할 수 있습니다. 그러나 레이크하우스 플랫폼이 성장함에 따라 보다 정교한 보안 체제를 위한 모든 메커니

증과 프로세스가 이미 마련되어 있어야 합니다. 또한 레이크하우스의 데이터에 대한 모든 액세스는 get-go의 감사 로그에 의해 제어되어야 합니다.

열린 인터페이스 및 열린 형식 권장

개방형 인터페이스 및 데이터 형식은 Lakehouse와 다른 도구 간의 상호 운용성에 매우 중요합니다. 기존 시스템과의 통합을 간소화하고 도구를 플랫폼과 통합한 파트너의 에코시스템을 엽니다.



개방형 인터페이스는 상호 운용성을 활성화하고 단일 공급업체에 대한 종속성을 방지하는 데 중요합니다. 전통적으로 공급업체는 데이터를 저장, 처리 및 공유할 수 있는 방식으로 기업을 제한하는 독점 기술과 폐쇄된 인터페이스를 구축했습니다.

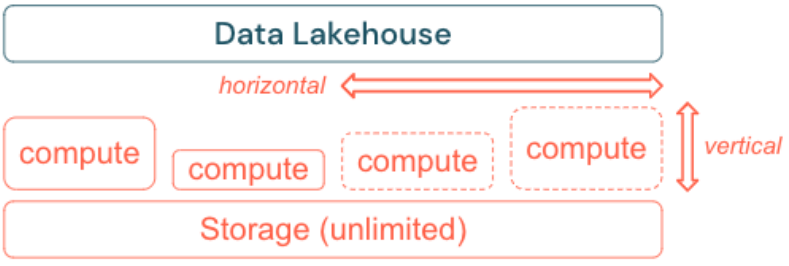
개방형 인터페이스를 기반으로 구축하면 미래를 위해 빌드할 수 있습니다.

- 더 많은 애플리케이션 및 더 많은 사용 사례에 사용할 수 있도록 데이터의 수명과 이식성을 향상시킵니다.
- 개방형 인터페이스를 신속하게 활용하여 도구를 Lakehouse 플랫폼에 통합할 수 있는 파트너의 에코시스템을 엽니다.

마지막으로 데이터에 대한 개방형 형식을 표준화하면 총 비용이 상당히 낮아질 것입니다. 높은 송신 및 계산 비용이 발생할 수 있는 독점 플랫폼을 통해 파이프할 필요 없이 클라우드 스토리지에서 직접 데이터에 액세스할 수 있습니다.

성능 및 비용에 맞게 크기 조정 및 최적화를 위한 빌드

데이터는 필연적으로 계속 증가하고 더 복잡해집니다. 향후 요구 사항에 맞게 조직을 구성하려면 레이크하우스를 확장할 수 있어야 합니다. 예를 들어 요청 시 쉽게 새 리소스를 추가할 수 있어야 합니다. 비용은 실제 소비로 제한되어야 합니다.



표준 ETL 프로세스, 비즈니스 보고서 및 대시보드에는 종종 메모리 및 계산 관점에서 예측 가능한 리소스가 필요합니다. 그러나 새 프로젝트, 계절 작업 또는 모델 학습(변동, 예측, 유지 관리)과 같은 최신 접근 방식은 리소스 요구의 최고를 생성합니다. 비즈니스에서 이러한 모든 워크로드를 수행할 수 있도록 하려면 메모리 및 계산을 위한 확장 가능한 플랫폼이 필요합니다. 새 리소스는 주문형으로 쉽게 추가해야 하며 실제 소비량만 비용을 발생시켜야 합니다. 피크가 끝나면 리소스를 다시 해제하고 그에 따라 비용을 절감할 수 있습니다. 이를 수평적 크기 조정(더 적은 노드) 및 수직 크기 조정(더 크거나 작은 노드)이라고도 합니다.

또한 크기를 조정하면 더 많은 리소스가 있는 노드 또는 더 많은 노드가 있는 클러스터를 선택하여 쿼리 성능을 향상시킬 수 있습니다. 그러나 대규모 머신과 클러스터를 영구적으로 제공하는 대신 전체 성능 대 비용 비율을 최적화하는 데 필요한 시간 동안만 주문형으로 프로비전할 수 있습니다. 최적화의 또 다른 측면은 스토리지와 컴퓨팅 리소스입니다. 이 데이터를 사용하는 데이터 볼륨과 워크로드 간에는 명확한 관계가 없으므로(예: 데이터의 일부만 사용하거나 작은 데이터에 대해 집중적인 계산을 수행) 스토리지 및 컴퓨팅 리소스를 분리하는 인프라 플랫폼에 정착하는 것이 좋습니다.