



데이터 엔지니어링을 위한 모범 사례

데이터의 크기 조정 및 처리를 더 빨리 수행하는 방법



TABLE OF CONTENTS

- 2 소개
- 3 #1: 동시 워크로드를 처리하도록 파이프라인 활성화
- 3 #2: 기존 기술을 활용하여 작업 완료
- 3 #3: 일괄 수집 대신 데이터 스트리밍 사용
- 4 #4: 파이프라인 개발 프로세스 간소화
- 4 #5: 파이프라인 개발 운영
- 4 #6: 연결 기능이 내장된 도구에 투자
- 4 #7: 확장성 통합
- 5 #8: 파이프라인에서 데이터 공유 활성화
- 5 #9: 데이터 랭글링에 적합한 도구 선택
- 5 #10: 엔지니어링 전략에 데이터 카탈로그 구축
- 5 #11: 데이터 소유자에 기대어 보안 정책 설정
- 6 Snowflake 소개

소개

데이터 엔지니어에게 이보다 더 좋은 시기는 없었습니다.

그러나 직업의 매개변수는 빠르게 변하고 있습니다. 데이터베이스와 데이터 웨어하우스가 클라우드로 이동하고 있고 새로운 도구와 데이터 파이프라인이 ETL 코드를 수동으로 작성하고 데이터를 정리하는 것과 같은 기존의 데이터 엔지니어링 작업을 대체하고 있습니다. 결과적으로 기업은 엔지니어에게 데이터 전략 및 파이프라인 최적화에 대한 지침을 제공하도록 요청하고 있습니다. 또한 정보가 기하급수적으로 증가하고 데이터 소스와 유형이 더욱 복잡해짐에 따라 비즈니스에서 수익성 및 성장을 증대하기 위해 해당 데이터를 활용할 수 있도록 하는 최신 전략과 도구를 엔지니어가 알고 있어야 합니다.

“데이터 엔지니어는 비즈니스 목표를 위해 데이터의 가치를 활용할 수 있는 귀중한 자원이 되었으며, 이는 궁극적으로 전체 조직에 필수적인 복잡한 환경에서의 전략적 역할을 수행합니다.” 빅데이터 뉴스 포털 Datanami에서는 이렇게 말합니다. “데이터 요구를 이해하고 탐색하면 데이터 엔지니어가 조직을 번창하는 데이터 우선 기업으로 발전시킬 수 있도록 힘을 실어줄 수 있습니다.”¹

조직의 데이터 전략 및 도구에 대해 올바른 결정을 내리고자 하는 데이터 엔지니어라면, 여기 수익성과 손실의 차이를 의미할 수 있는 데이터 엔지니어링을 위한 11가지 모범 사례가 있습니다.



데이터 엔지니어링을 위한 모범 사례

1. 동시 워크로드를 처리하도록 파이프라인 활성화

기업이 수익을 내기 위해서는 여러 데이터 분석 프로세스를 동시에 실행해야 하며, 수요를 따라갈 수 있는 시스템이 필요합니다. 데이터는 웹, 모바일 장치 및 IoT(사물 인터넷) 장치를 통해 하루 24시간 연중무휴로 기업에 유입됩니다. 데이터 파이프라인이 해당 데이터를 로드하고 처리해야 하며, 과학자가 데이터를 분석하고 다운스트림 애플리케이션이 추가적인 활용을 위해 이를 처리합니다. 클라우드에 있는 최신 데이터 파이프라인은 동시 워크로드를 처리할 수 있는 탄력적인 멀티 클러스터 공유 데이터 아키텍처를 특징으로 합니다. 리소스 경합 없이 동일한 데이터를 동시에 공유하면서 처리, 데이터 로드, 변환 및 분석을 위한 여러 개의 독립적이고 격리된 클러스터를 할당할 수 있습니다.

2. 기존 기술을 활용하여 작업 완료

많은 파이프라인이 데이터 엔지니어가 Apache Spark, Apache Kafka 또는 Python을 사용해야 하는 복잡한 알고리즘을 사용합니다. 그러나 문제를 해결하기 위해 새로운 플랫폼을 배울 필요는 없습니다. 대신 현재 기술을 사용할 방법을 찾으십시오. 예를 들어 최신 ETL을 사용하면 Kafka를 사용하는 대신 직접 SQL 문을 사용하여 스트림 처리 작업을 수행할 수 있습니다.

새로운 것을 배우는 데 리소스를 투자하기 전에 현재 기술을 최대한 활용하십시오.

3. 일괄 수집 대신 데이터 스트리밍 사용

데이터가 귀사로 하루 종일 유입되므로 주기적인 일괄 수집을 수행하는 경우 최근 이벤트를 놓칠 수도 있습니다. 이는 사기 또는 데이터 침해 감지 실패와 같은 치명적인 결과를 초래할 수 있습니다. 오래된 데이터는 수익성에도 영향을 줄 수 있습니다. 예를 들어, 온라인 쇼핑 이벤트를 운영하는 회사는 가장 많이 본 제품, 가장 많이 구매한 제품, 인기가 가장 적은 제품에 대한 즉각적인 통찰력을 통해 웹사이트 레이아웃을 변경하는 등 더 많은 판매를

유도하는 조치를 신속하게 취할 수 있기를 바랍니다. 연속 스트리밍 수집을 설정하여 파이프라인 지연 시간을 줄이고 비즈니스에서 하루 전이 아닌 몇 분 전의 데이터를 사용할 수 있도록 합니다. 사용 가능한 스트리밍 기능과 이것이 다양한 아키텍처에서 작동하는 방식을 이해하고 일괄 처리 및 스트리밍 데이터를 모두 처리할 수 있는 파이프라인을 구현합니다.



4. 파이프라인 개발 프로세스 간소화

생산 데이터의 유효성을 확인하려면, 코드와 알고리즘을 생산 환경에 사용할 준비가 될 때까지 반복적으로 테스트할 수 있는 테스트 환경에서 파이프라인을 구축하십시오. 데이터 파이프라인을 실행하기 위한 기반으로 클라우드 데이터 플랫폼을 사용하는 경우, 테스트 환경을 생성하는 것이 새로운 데이터베이스와 인프라를 관리하는 번거로움 없이 기존 환경의 복제본을 생성하는 것만큼 간단할 수 있습니다. 이렇게 하면 개발부터 테스트, 생산으로 옮겨가는 시간을 크게 단축하여 온프레미스에서 동일한 파이프라인을 구축하는 것보다 훨씬 빠르게 진행할 수 있습니다.

5. 파이프라인 개발 운영

파이프라인을 생성한 후 더 많은 데이터 소스를 수용하도록 파이프라인을 수정하거나 확장해야 할 수 있습니다. 파이프라인을 쉽게 수정하거나 확장할 수 있도록 설계하십시오. 이 개념은 “DataOps” 또는 데이터용 DevOps로 알려져 있으며 자동화와 경우에 따라 AI(인공 지능)를 사용하여 파이프라인에 지속적인 통합, 전달 및 배포를 구축하는 것으로 구성됩니다. 파이프라인에 DataOps를 통합하면 데이터의 신뢰성과 가용성이 높아집니다.

6. 연결 기능이 내장된 도구에 투자

최신 클라우드 기반 데이터 파이프라인은 서로 통신해야 하는 많은 도구와 플랫폼을 수용합니다. 원본 시스템, 데이터 웨어하우스, 데이터 레이크 및 분석 애플리케이션 간의 연결을 구축하려면 시간, 노동 및 비용이 소요됩니다. 그렇게 하는 대신 서로 간의 연결이 내장되어 있는 도구에 투자하십시오. 도구가 연결되어 있지 않은

경우 다른 도구에서 이를 찾을 수 있도록 Amazon Simple Storage Service(S3)에서 사용하는 형식과 같은 일반 형식으로 데이터를 저장하는 추가 단계를 수행하십시오.

7. 확장성 통합

조직은 데이터에서 의미를 도출하기 위해 다양한 도구를 사용합니다. 예를 들어, 조직은 이미지를 스캔하고 이미지에서 텍스트를 추출하기 위해 사용자 지정 API를 작성할 수 있습니다. 사용자 지정 알고리즘의 또 다른 예는 고객 서비스 채팅의 감정 분석을 수행하는 것입니다. 이 코드를 활용할 수 있는 최신 파이프라인을 구축해야 합니다. API 및 파이프라인 도구를 사용하면 외부 코드를 원활하게 사용하는 데이터 흐름을 생성할 수 있습니다.



8. 파이프라인에서 데이터 공유 활성화

종종 조직 내부와 외부의 여러 그룹이 분석을 수행하기 위해 동일한 핵심 데이터를 필요로 합니다. 예를 들어, 한 개의 소매업체가 판매 데이터를 세 개의 서로 다른 공급업체와 공유해야 할 수 있습니다. 동일한 데이터로 각각의 파이프라인을 구축하려면 시간과 비용이 소요됩니다. 그 대안으로 클라우드에서 최신 도구를 사용하면 데이터에 액세스할 수 있는 사용자를 관리할 수 있도록 하는 공유 파이프라인을 생성할 수 있습니다. 공유 파이프라인은 적절한 사람들에게 딱 맞는 정보를 신속하게 제공합니다.

9. 데이터 랭글링에 적합한 도구 선택

데이터 랭글링 도구는 데이터 세트 내의 필드, 행 또는 데이터값과 같은 고유한 엔티티를 변환하여 데이터의 불일치를 수정함으로써 활용하기 쉽도록 만들 수 있습니다. 예를 들어, 상점 이름 “Giantmart”는 “Giant-Mart”, “Giantmart Megacenter” 및 “Giant-mart Inc.”와 같은 다양한 소스에서 파이프라인에 도착할 수 있습니다. 이로 인해 데이터가 로드 및 분석될 때 문제가 발생할 수 있습니다. 더 잘 정리된 데이터가 비즈니스 의사 결정을 위한 더 좋고 더 정확한 통찰력을 제공합니다.

10. 엔지니어링 전략에 데이터 카탈로그 구축

분석가가 데이터의 출처, 액세스한 사람 또는 소유한 비즈니스 프로세스와 같은 파이프라인의 데이터에 관해 질문할 수도 있습니다. 데이터 과학자가 정확성을

보장하기 위해 데이터를 원시 형식으로 봐야 할 수도 있습니다. 최종 사용자가 신뢰할 수 있는 데이터 세트와 진행 중인 데이터 세트를 알고 싶어할 수도 있습니다. 필요한 경우 데이터를 추적할 수 있도록 데이터 계보를 추적하는 데이터 카탈로그를 구축합니다. 이렇게 하면 데이터에 대한 최종 사용자의 신뢰가 높아지고 데이터의 정확성도 향상됩니다.

11. 데이터 소유자에 기대어 보안 정책 설정

데이터 엔지니어가 보안 정책을 설정하는 방법(누가 볼 수 있고 그에 대해 어떤 종류의 액세스 권한이 있는지)을 이해하지 못할 수도 있습니다. 예를 들어, 데이터를 특정 사용자에게 전송하기 전에 어떤 데이터 필드를 단독 처리해야 한다는 사실을 깨닫지 못하여 보안 또는 규제 문제가 발생할 수 있습니다. 이 시나리오를 방지하려면 데이터 소유자 또는 생산자가 보안 정책을 설정해야 합니다. 다른 사람들도 권장 사항을 제시할 수 있지만 궁극적으로, 데이터를 배포하기 전에 이것을 어떻게 보호해야 하는지 가장 잘 알고 있는 사람은 소유자입니다.

데이터 엔지니어링의 세계는 빠르게 변화하고 있습니다. IoT, AI, 클라우드와 같은 기술은 데이터 파이프라인을 변화시키고 기존의 데이터 관리 방식을 뒤엎고 있습니다. 크든 작든 데이터 파이프라인에 대한 결정은 비즈니스에 상당한 영향을 미칠 수 있습니다. 잘못된 선택은 불필요한 작업에 소요되는 비용과 시간의 증가를 의미합니다. 올바른 결정은 비즈니스가 데이터의 힘을 강화하여 향후 몇 년간 수익성과 성장을 달성할 수 있도록 합니다.





SNOWFLAKE 소개

Snowflake 클라우드 데이터 플랫폼은 조직이 데이터의 진정한 가치를 활용하는 데 방해가 되는 장애물을 해결해 드립니다. 수천 명의 고객이 Snowflake를 배포하여 모든 비즈니스 사용자가 모든 데이터에서 모든 통찰력을 도출함으로써 한때 가능했던 수준 이상으로 비즈니스를 발전시킵니다. Snowflake는 모든 클라우드용으로 구축된 유일한 데이터 웨어하우스, 전체 데이터 네트워크에 대한 즉각적이고 안전하며 통제된 액세스, 최신 데이터 애플리케이션 개발을 위한 단일 플랫폼을 포함하여 다른 많은 유형의 데이터 워크로드를 사용 가능하게 하는 핵심 아키텍처를 제공하는 단일 통합 플랫폼을 조직이 갖추도록 합니다.

Snowflake: 제한 없는 데이터. [snowflake.com/?lang=ko](https://www.snowflake.com/?lang=ko)에서 더 자세히 알아보십시오.



© 2022 Snowflake, Inc. All rights reserved. [snowflake.com/?lang=ko](https://www.snowflake.com/?lang=ko) #MobilizeYourData

인용

¹ datanami.com/2019/07/18/data-engineers-the-c-suites-savior