



데이터 과학을 위한 SNOWFLAKE

소개: 모두 데이터에 관한 것

ML(머신 러닝) 기술이 주류에 진입했습니다. AI(인공지능) 및 ML 사용에 대한 2019년 TDWI 설문 조사에 따르면, 응답자의 92%가 머신 러닝 기술을 사용한다고 보고했으며 85%는 ML용 도구를 사용하여 예측 모델을 구축하고 있다고 말했습니다.¹

머신 러닝 모델을 구축하고 교육하기 위해 데이터 과학자는 막대한 양의 데이터를 필요로 합니다. AI 시대에, 데이터로의 빠르고 정확한 액세스는 중요한 경쟁 차별화 요소가 되었습니다. 데이터 관리(분석을 위한 데이터 검색, 액세스 보안, 정리, 결합 및 준비)는 일반적으로 프로세스에서 가장 시간이 많이 소요되는 측면으로 인식됩니다.

효율적인 데이터 플랫폼이 가장 중요

Forbes에 따르면, 과학자들이 소비하는 총 시간 중 최대 80%는 분석 및 교육을 위해 데이터를 찾고, 검색하고, 통합하고, 정리하고 준비하는 데 소비됩니다.² Forbes의 동일한 연구에 따르면 데이터 과학자는 데이터 마이닝이나 모델링보다 데이터 마사지에 대부분의 시간을 할애할 뿐만 아니라 이러한 고도로 숙련된 전문가의 76%가 데이터 준비를 작업에서 가장 즐겁지 않은 부분으로 간주합니다.³

이 백서에서는 오늘날의 데이터 과학 및 ML 이니셔티브를 주도하는 데이터 요구 사항을 식별하도록 돕고 Snowflake와 그 파트너가 제공하는 업계 최고의 도구를 지원하는 플랫폼으로 이러한 요구 사항을 충족할 수 있는 방법을 설명합니다.

데이터 과학을 위한 가장 완전한 플랫폼

Snowflake의 플랫폼은 데이터 웨어하우스의 힘, 빅 데이터 플랫폼의 유연성, 클라우드의 탄력성, 라이브 데이터 공유를 기존 데이터 플랫폼 솔루션보다 훨씬 적은 비용으로 결합합니다. Snowflake는 내부 사용과 데이터 익스체인지 생성을 위해 모든 데이터를 한 위치에 저장 및 분석하는 데 필요한 성능, 동시성 및 단순성을 제공합니다. 이 플랫폼이 다음과 같은 세 가지 필수 요구 사항을 충족하기 때문에 수천 명의 고객이 이 플랫폼에서 표준화를 진행하고 있습니다.

- **모든 데이터에 대한 단일 통합 소스:** Snowflake는 데이터 과학자가 하나의 일관된 소스에서 정형 및 반정형 데이터에 액세스할 수 있도록 지원해 귀사의 데이터 자산을 더 쉽게 찾고, 통합하고, 정리하고, 사용하도록 합니다. 데이터 과학으로 도출된 결과물은 비즈니스 사용자가 액세스할 수 있도록 Snowflake에 다시 원활하게 통합될 수 있습니다.
- **효율적이고 빠른 데이터 준비:** Snowflake는 다른 사용자나 부서에 영향을 주지 않고 SQL을 사용하여 데이터를 수집, 변환 및 쿼리할 수 있는 효율적인 전용 가상 웨어하우스를 제공합니다. Snowflake 내 SQL은 많은 경우 Spark 와 같은 다른 도구보다 10배 더 효율적으로 데이터를 준비하므로 ML 작업 간의 지연 시간이 감소합니다.
- **광범위한 파트너 생태계:** Snowflake에는 모든 기존 데이터 과학 기술과 새로운 데이터 과학 기술에 대한 커넥터가 있습니다. 이를 통해 고객은 필요에 가장 맞는 데이터 과학 도구를 선택할 수 있으며 모든 도구는 통합되고 일관된 데이터 플랫폼에 액세스할 수 있습니다. Snowflake는 데이터 과학 도구의 범용 액세스를 위해 Amazon S3 및 기타 Blob 저장소로 데이터를 원활하게 내보냅니다.

Snowflake 워크로드



데이터
웨어하우스



데이터
레이크



데이터
엔지니어링



데이터
익스체인지



데이터
애플리케이션



데이터
사이언스

그림 1: Snowflake는 데이터 과학 외에도 많은 사용 사례와 워크로드를 지원하므로 귀사는 데이터, 분석 및 예측 분석을 위한 단일 플랫폼의 강력한 기능을 활용할 수 있습니다.

주요 데이터 과학 개념 및 페르소나

데이터 과학자는 데이터의 패턴, 관계, 상관관계, 결과 및 추론을 식별하는 데 머신 러닝 기술을 사용합니다. 이러한 데이터 기반 발견은 사기를 감지하고, 유지 관리 주기를 예측하고, 고객 이탈을 완화하고, 판매를 예측하고, 기타 많은 미래 지향적인 작업을 자동화할 수 있는 모델에 통합됩니다. 이 프로세스 내 주요 역할 및 페르소나에는 다음이 포함됩니다.

- **데이터 과학자**는 모델을 구축하고 데이터로 모델을 교육합니다. 그들은 Jupyter 및 Zeppelin과 같은 노트북과 R, Python, Java 및 Scala와 같은 언어를 사용합니다.
- **데이터 분석가/시민 데이터 과학자**는 이러한 모델을 사용하여 머신 러닝에 대한 실제 이해를 기반으로 비즈니스 의사 결정을 위한 예측 및 처방 분석을 수행합니다.
- **데이터 엔지니어**는 데이터를 준비하고 지속적으로 ML 모델에 데이터를 제공하는 자동화된 데이터 파이프라인을 설정합니다.

데이터 과학에서 머신 러닝의 역할

머신 러닝은 주로 데이터 준비, 데이터 검색, 분석 및 데이터 모델링을 아우르는 훨씬 더 광범위한 데이터 과학 분야의 데이터 모델링 측면을 다룹니다. 오늘날의 ML 및 데이터 과학 도구는 데이터 구문 분석, 예측 및 처방 모델 생성, 모델을 생산에 배치하고 시간이 지남에 따라 이러한 모델을 유지 관리하는 등, 여러 측면을 처리할 수 있습니다. 예측 및 처방 분석 앱은 방문자에게 제품과 서비스를 추천하기 위해 웹 브라우징 패턴을 모니터링하는 등, 인간의 개입 없이 종종 스스로 결정을 내릴 수 있습니다.

데이터 과학자는 분석 도구를 사용하여 가설을 세운 다음 프로그래밍 언어와 ML 라이브러리를 사용하여 예측을 생성합니다. ML 유형에는 선형 회귀, 로지스틱 회귀, 분류, 의사 결정 트리, 딥 러닝 등이 있습니다. 일부 인기 있는 ML 라이브러리에는 XGBoost, TensorFlow, scikit-learn 및 PyTorch가 있습니다.

데이터 과학자가 신뢰할 수 있는 예측을 할 수 있는 모델을 만들고 이를 교육하는 일을 담당한다면, 데이터 엔지니어는 ML 모델에 유추에 필요한 데이터를 제공하는 데이터 파이프라인을 말합니다. 이러한 ML/AI 프로세스의 결과는 비즈니스 사용자가 데이터 기반 의사 결정을 내리는 데 사용할 수 있습니다.

머신 러닝 프로세스

머신 러닝 이니셔티브의 성공은 적시에 정확한 데이터를 올바른 모델로 가져오는 데 달려 있습니다. 대부분의 머신 러닝 주기는 발견 및 개발에서 생산에 이르는 여러 단계로 구성되기 때문에 이것은 항상 쉽지는 않습니다. 데이터는 ML 주기의 각 단계에서 여러 번 추가되고 준비되며 종종 서로 다른 데이터 요구 사항이 적용됩니다. ML의 성공은 적절한 조건의 정확한 데이터를 올바른 분석 플랫폼으로 가져와 비즈니스 결과를 생성하는 데 달려 있습니다.

그림 2와 같이 데이터 과학자는 데이터를 찾고, 수집하고, 이해하고, 준비하는 것으로 시작합니다(1-3단계). 데이터를 더 잘 이해하고 가설을 세우기 위해 비즈니스 인텔리전스 도구를 사용할 수도 있습니다. 데이터 과학자는 많은 데이터 세트를 이 반복적인 프로세스를 통해 실험합니다. 데이터 세트의 범위를 넓히거나 확장할 때마다 데이터 과학자는 데이터 엔지니어가 데이터를 로드하고 준비할 때까지 기다려야 합니다. 이로 인해 지연이 발생하고 반복 간에 상당한 대기 시간이 발생합니다. 또한 데이터를 정규화된 형식으로 “형성”해야 하며 많은 알고리즘에는 세분화된(nuanced) 형식이 필요합니다.

다음으로 이전 단계(4)에서 준비된 모델을 통해 교육 데이터를 실행하고 결과를 평가하여 각 모델의 유효성을 결정한 다음 기능 엔지니어링(3) 및 하이퍼 매개변수 조정(4) 주기를 통해 모델을 추가로 조정합니다.

그런 다음 훈련된 결과 모델을 생산(5)에 배포하여 비즈니스 사용자에게 예측 및 처방 도구를 제공합니다. 생산에 배포된 모델은 모델 드리프트를 식별하고 모델이 구식인지 여부를 확인하기 위해 지속적인 평가(6)를 받습니다. 모델은 새로운 교육 데이터로 주기적으로 재교육되어야 합니다. 이러한 모델 업데이트는 ML 주기의 또 다른 반복을 나타내며, 시간이 많이 소요되고 오류가 발생하기 쉬운 더 많은 데이터의 처리를 요합니다. 사용 사례에 따라 몇 시간, 며칠 또는 몇 주마다 모델을 재교육해야 할 수도 있습니다.

ML 워크플로우

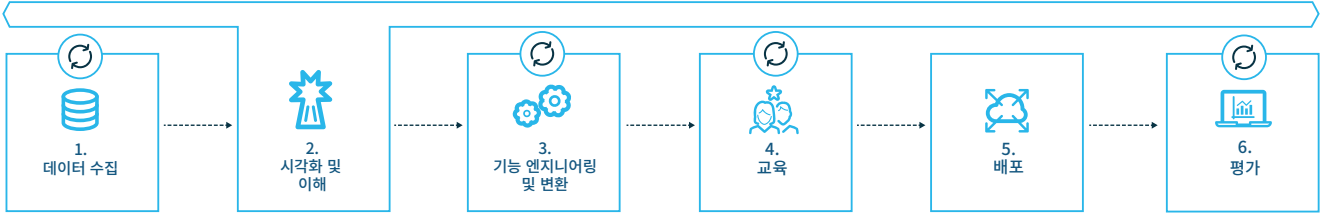


그림 2: 데이터가 수집 및 준비에서 교육, 예측 및 제품화에 이르는 ML 프로세스를 주도합니다.

데이터 클라우드의 역할

머신 러닝은 데이터 집약적인 활동이며 각 예측 모델의 성공은 다양한 방식으로 수집, 유지, 변환 및 제시해야 하는 다량의 각종 데이터에 달려 있습니다. 여기에는 여러 차원과 세부 정보로 특징지어지고 다양한 컨텍스트에서 발생하는 다량의 데이터가 포함됩니다. 예를 들어 귀사에서 고객 이탈을 예측하는 머신 러닝 모델을 구축한 경우 판매, 서비스, 구매 및 앱 상호 작용과 관련된 고객 행동에 대한 과거 및 실시간 데이터를 갖추고 있을 것으로 예상할 수 있습니다.

Snowflake의 데이터 클라우드는 데이터 웨어하우스, 데이터 마트 및 데이터 레이크를 여러 유형의 분석 및 데이터 과학 애플리케이션을 지원하는 단일 진실 공급원으로 통합할 수 있도록 지원합니다. 팀 구성원이 데이터를 복사하고 다른 곳으로 옮길 필요 없이 협업할 수 있도록 함으로써 다양한 팀이 관리형 데이터를 내부 및 외부에서 쉽게 공유할 수 있습니다. JSON, AVRO, XML, ORC 및 Parquet에 대한 기본 지원을 통해 데이터 과학 워크플로우에서 원시, 정형 및 반정형 데이터를 쉽게 검색하고 즉시 액세스할 수 있습니다.

하나의 도구 세트를 사용하여 정형 및 반정형 데이터를 모두 관리할 수 있으므로 데이터 검색 및 준비 주기가 단축됩니다. 또한 ML 알고리즘에서 도출된 결과 데이터는 원본 데이터와 함께 비즈니스 사용자가 액세스할 수 있도록 다시 리포지토리에 놓입니다. 즉, 비즈니스 사용자, 분석가 및 데이터 과학자를 위해 모든 데이터가 항상 최신 상태로 일관되게 유지 관리됩니다.

데이터 클라우드의 이점

TDWI는 머신 러닝, 인공 지능 및 예측 애플리케이션 개발의 전체 데이터 수명 주기를 충족할 수 있는 클라우드용으로 구축된 최신 데이터 플랫폼을 구입할 것을 권장합니다. 그런 플랫폼에서 무엇을 찾아야 할까요? 데이터 준비를 위해서는 대화식 응답 시간으로 대용량 데이터 세트로 작업할 수 있어야 합니다. 교육을 위해서는 이러한 데이터를 반복적으로 살펴보고자 할 것입니다. 생산의 경우 안정적이고 반복 가능하며 확장 가능한 데이터 파이프라인이 필요합니다.

데이터 과학과 관련된 시도에 Snowflake를 사용해야 하는 몇 가지 주요 이유는 다음과 같습니다.

- **단순성:** 여러 컴퓨팅 플랫폼을 관리하고 지속적으로 통합을 유지할 필요가 없습니다.
- **보안:** 사용자 자격 증명을 신중하게 관리하고 모든 전송을 암호화하는 Snowflake 환경에 단 하나의 데이터 사본이 안전하게 저장됩니다.
- **성능:** 쿼리 결과는 캐시되며 ML 프로세스뿐만 아니라 분석용으로 반복적으로 사용할 수 있습니다.
- **워크로드 격리:** 각 사용자와 워크로드는 전용 컴퓨팅 리소스를 받을 수 있습니다.
- **탄력성:** 대용량 데이터 처리 작업을 수용할 수 있도록 용량을 확장하는 데 몇 초밖에 걸리지 않으며 완료되면 쉽게 릴리즈할 수 있고 초당 가격 책정을 통한 지불로 비용 부담을 최소화할 수 있습니다.
- **정형 및 반정형 데이터 지원:** 통합 리포지토리 내에서 모든 유형의 데이터를 쉽게 로드, 통합 및 분석합니다.
- **동시성:** 공유 데이터 전반에 걸쳐 동시 워크로드를 대규모로 실행합니다.

머신 러닝 및 분석을 위한 데이터 통합

머신 러닝 애플리케이션을 위해 데이터를 프로비저닝하는 방법에는 여러 가지가 있으며 유연성이 필수적입니다. 예를 들어, 몇몇 조직에서는 데이터 웨어하우스와 데이터 레이크를 함께 사용합니다. 이를 통해 방대한 양의 원시 데이터를 기본 형식으로 저장할 수 있으며 필요할 때 광범위한 분석을 위해 용도를 변경할 수 있습니다. 대부분의 데이터 과학 도구는 데이터 소스로 데이터 레이크를 사용하지만, 오늘날의 분석 전략에서는 빅 데이터 플랫폼, 클라우드, 데이터 레이크 및 데이터 웨어하우스가 혼합된 멀티 플랫폼 데이터 아키텍처를 점점 더 많이 사용합니다. 여러 선도적인 조직에서 데이터 레이크를 통째로 건너뛰고, 대신 데이터를 클라우드 데이터 플랫폼으로 완전히 통합하고 있습니다. 이 접근 방식은 별도의 데이터 레이크를 관리하는 복잡성을 제거하고 데이터 레이크와 데이터 웨어하우스 간의 데이터 변환 파이프라인 역시 요하지 않습니다. 다용도 클라우드 데이터 플랫폼을 기반으로 하는 통합 리포지토리를 사용하면 각 데이터 세트와 워크로드에 적합한 스토리지, 처리 및 경제성을 선택하여 ML 및 분석을 위한 옵션을 최적화할 수 있습니다.

일단 데이터를 수집하고 준비하고 나면, 분석 및 예측 분석 도구를 통해 패턴과 통찰력을 발견할 수 있어야 합니다. Snowflake를 사용하면 일반 분석을 예측 분석과 결합할 수 있으므로 동일한 관리형 데이터에 대해 비즈니스 인텔리전스 도구와 데이터 과학 도구가 일관된 뷰를 갖게 됩니다. 모든 데이터 과학 도구가 동일한 데이터 정의를 참조하므로 쿼리, 예측, 대시보드 및 보고서의 콘텐츠를 일관되게 재현할 수 있습니다. 쉬운 액세스를 위해 원시 데이터와 ML 결과는 모두 데이터 플랫폼에 있습니다. 이 통합된 접근 방식을 통해 데이터 과학자는 머신 러닝 활동의 결과를 범용 분석을 위한 데이터 플랫폼으로 다시 출력할 수 있을 뿐만 아니라 의사 결정 프로세스에 이러한 결과를 포함할 수 있습니다.

공통된 의미, 데이터 정의 및 데이터 모델을 사용하면 계속해서 모든 사람이 동일하게 이해하게 됩니다. 예를 들어, 영업 관리자가 영업 팀의 과거 성과를 보여주는 BI 보고서를 볼 수도 있습니다. ML 모델은 타겟 계정의 경향을 기반으로 다가오는 분기의 예상 판매 결과를 예측하고 동일한 보고서를 통해 예약된(booked) 수익과 예측된 수익을 강조 표시할 수도 있습니다.

자동화된 데이터 엔지니어링, 데이터 통합 및 데이터 세이핑

ML로 성공을 거둔다는 것은 비즈니스 사용자가 사용하는 앱과 서비스를 채우는 것은 물론, 정확하고 시기적절한 데이터를 비즈니스 사용자에게 공급하는 효율적이고 안정적인 데이터 파이프라인을 만드는 것을 의미합니다.

데이터 수집하기

Snowflake는 데이터를 비동기 방식으로 로드하여 즉시 사용할 수 있도록 하는 Snowpipe라는 서버리스 수집 서비스를 포함하고 있습니다. 수동 데이터 평면화 작업은 완전히 자동화됩니다. 플랫폼은 데이터를 각 타겟 테이블에 필요한 유형과 세이프로 변환합니다.

표준 커넥터 및 어댑터를 사용하면 Kafka 및 기타 메시징 시스템에서 이벤트 스트림을 쉽게 수집할 수 있고 Snowflake 스트림 및 작업을 사용하면 SQL 작업용 데이터 로드를 쉽게 예약할 수 있습니다.

자동화된 데이터 수집 서비스로 ML 모델을 “상품화”함으로써 파이프라인은 복잡한 데이터 통합 작업을 단순화합니다. 데이터 과학자는 테스트 사이에 며칠 또는 몇 시간을 기다리지 않고 온디맨드 방식으로 데이터를 찾고 준비할 수 있습니다. 자동화된 데이터 파이프라인 서비스가 생산에 도입되면 데이터 레이크에서 ETL을 요구하지 않고도 원시 데이터를 즉시 사용할 수 있습니다. 데이터가 들어오면 모델을 통해 자동으로 실행되어 예측을 합니다. 또한 모두 클라우드를 기반으로 하기 때문에 데이터 과학자가 다른 사용자에게 영향을 주지 않고 전용 가상 웨어하우스 컴퓨팅 리소스를 사용할 수 있습니다.

범용 SQL 기능

Snowflake 고객은 강력하고 효율적인 ETL 및 ELT 워크로드를 지원하는 범용 SQL 기능을 통해 중앙 진실 공급원을 활용할 수 있습니다. SQL을 이용한 데이터 변환은 Spark를 이용한 동일한 작업보다 빠르고 쉬우며 비용도 더 저렴합니다. 데이터가 SQL 쿼리의 일부로 변환될 수 있으므로 변환은 분석의 일부가 됩니다. Snowflake의 아키텍처와 압축 덕분에 대량의 스트리밍 데이터를 빠르게 수집하고 얼마 안 되는 비용으로 무기한 저장할 수 있습니다. 데이터 엔지니어는 Alteryx, Alooka, Matillion, Fivetran, Alation, Informatica 등을 비롯한 여러 유형의 통합 도구를 활용할 수 있습니다.

전용 컴퓨팅 리소스

Snowflake를 사용하면 ML 데이터 수집, 데이터 관리 및 데이터 준비 워크로드가 비 ML 데이터 엔지니어링 및 분석 워크로드와 결합하지 않는 전용 리소스를 받습니다. 리소스에 대한 결합이 사라지므로, 라이브 데이터를 스트림에서 수집 및 변환하고 즉시 분석에 사용할 수 있습니다. 각 워크로드에 대해 데이터 웨어하우스의 크기를 사용자 지정하고, 필요에 따라 확장하고, 완료되면 클라우드 서비스를 해제할 수 있습니다. 선형 배율 조정 덕분에 예측 가능한 기간 내에 쿼리를 실행하는 데 필요한 정확한 양의 리소스를 요청할 수 있습니다. 즉각적인 탄력성과 초당 청구를 통해 각 사용자와 작업 그룹은 각자가 사용한 정확한 양의 컴퓨팅 리소스에 대해서만 비용을 지불합니다. 궁극적으로 이 아키텍처를 사용하면 일관된 데이터를 제공하면서 각 팀의 성능과 효율성을 극대화할 수 있습니다.

강력한 데이터 보안

많은 레거시 데이터 과학 프로젝트가 분산 프레임워크에서 데이터를 저장하고 처리하기 위한 오픈 소스 프레임워크인 Apache Hadoop에 의존합니다. 그러나 Hadoop 아키텍처는 가장 기본적인 액세스 제어만 사용하며 HIPAA, PCI DSS 및 GDPR을 포함해 데이터의 보안 및 개인 정보 보호를 관리하는 중요한 산업 표준을 준수하도록 설계되지 않았습니다. 다른 데이터 과학 프로젝트는 Amazon S3와 같은 범용 객체 저장소를 활용하는데 이는 강력한 데이터 보안 부분이 부족합니다.

이와는 대조적으로 Snowflake의 데이터 클라우드는 암호화, 액세스 제어, 네트워크 모니터링 및 물리적 보안 조치와 포괄적인 모니터링, 경고 및 사이버 보안 관행을 포함하는 다계층 보안 기반에 구축되었습니다. Snowflake는 ISO/IEC 27001 및 SOC 1/SOC 2 Type 2와 같은 산업 표준 기술 인증 외에도 PCI DSS, HIPAA/HITRUST(Health Information Trust Alliance) 및 FedRAMP 인증과 같은 중요한 정부 및 산업 규정을 준수합니다. Snowflake 고객은 모든 유형의 데이터 과학 활동을 위한 데이터에 안전하게 액세스할 수 있습니다.

효율적인 데이터 공유

Snowflake 플랫폼은 Snowflake 데이터 마켓플레이스 및 데이터 익스체인지지를 통해 파트너, 제공업체, 공급업체 및 고객 간의 데이터 교환을 단순화합니다. 이를 통해 모델의 효율성을 높이고 추가적인 기능 엔지니어링 기회를 제공할 수 있는 고유한 데이터 세트에 액세스할 수 있습니다. Snowflake의 안전한 데이터 공유는 FTP를 통한 데이터 전송이나 애플리케이션 연결을 위한 API 구성을 필요로 하지 않습니다. ETL 통합을 단순화하고 데이터 공급자와 데이터 소비자 간에 “라이브” 데이터를 자동으로 동기화합니다. 원본 데이터가 복사되는 대신 공유되기 때문에 소비자는 추가 클라우드 스토리지가 필요하지 않습니다. Snowflake 데이터 마켓플레이스 및 데이터 익스체인지지를 통해 데이터 과학자가 원시 데이터와 처리된 데이터를 공유하여 모델에 대해 쉽게 협업할 수 있습니다.

Snowflake 데이터 마켓플레이스 및 데이터 익스체인지지

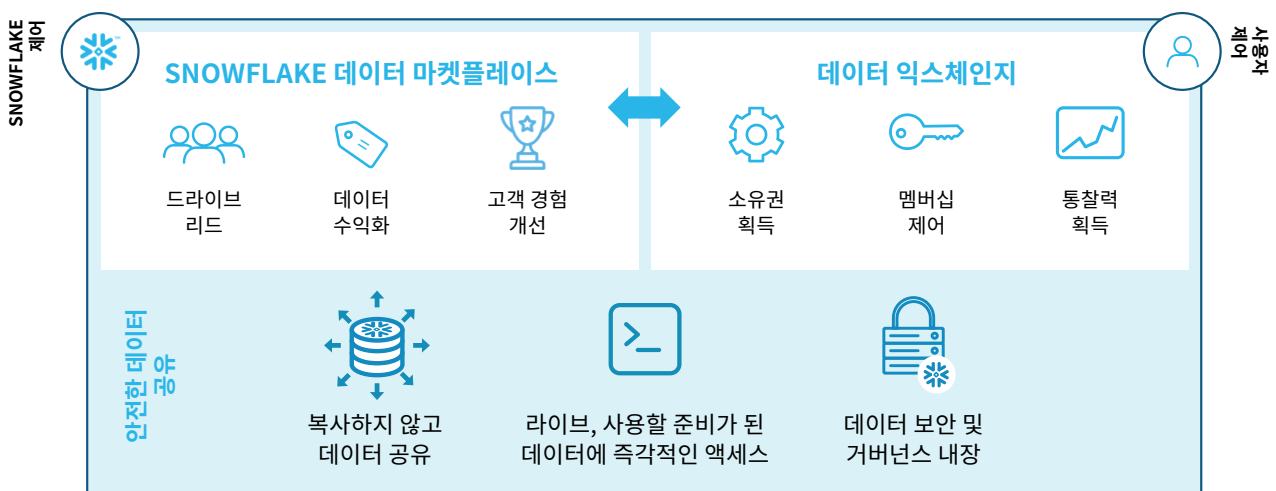


그림 3: Snowflake 안전한 데이터 공유를 사용하면 Snowflake 데이터 마켓플레이스를 통해 외부에 데이터를 공유하고 고객, 공급업체 및 기타 비즈니스 파트너와 자체적인 데이터 익스체인지지를 생성할 수 있습니다.

광범위한 파트너 생태계

머신 러닝 분야는 매년 새로운 도구가 추가되면서 빠르게 진화하고 있습니다. Snowflake의 광범위한 파트너 생태계를 통해 고객은 모든 기존 및 새로 등장한 데이터 과학 도구, Python, R, Java 및 Scala와 같은 언어, PyTorch, XGBoost, TensorFlow, scikit-learn 과 같은 오픈 소스 라이브러리, upyter 및 Zeppelin과 같은 노트북, Data Robot, Dataiku, H2O.ai, Amazon Sagemaker 등과 같은 플랫폼에 대한 직접 연결을 활용할 수 있습니다. Snowflake는 데이터에 대해 일관된 단일 리포지토리를 제공함으로써 도구, 언어 또는 라이브러리를 전환할 때마다 기본 데이터를 재정비할 필요가 없습니다. 이러한 활동 결과물이 또한 Snowflake에 쉽게 반영되며 비전문가인 사용자가 이에 액세스해 비즈니스 가치를 창출할 수 있습니다.

노트북 기반 ML 도구

Jupyter 및 Zeppelin과 같은 기존 ML 노트북은 Amazon Sagemaker, Dataiku, Zepl 등을 비롯한 오늘날의 선도적인 데이터 과학 도구를 지원합니다. 이 접근 방식을 통해 데이터 과학자는 자신이 선택한 프레임워크 및 알고리즘을 궁극적으로 제어하고, 심층적인 기능 엔지니어링을 수행하고, 초매개변수 (hyperparameter)를 조정하고, ML 모델을 반복적으로 생성, 평가 및 상품화할 수 있습니다. 알고리즘을 반복적으로 실험하고 그 성능에 점수를 매기고 새로운 모델을 선택하고 수정함으로써 직관을 정확한 예측으로 바꿀 수 있습니다. Amazon SageMaker 사용자가 Snowflake Python 커넥터를 사용하여 Pandas DataFrames를 직접 채울 수 있습니다. 이 고속 연결은 ANSI SQL의 모든 기능을 활용하는 최적화된 데이터 준비 및 기능 엔지니어링 주기뿐만 아니라 교육의 가속화를 초래합니다.

AutoML 도구

또는 RapidMiner, BigSquid, H2o.ai 및 DataRobot 과 같은 AutoML 도구가 알고리즘을 자동으로 선택하고, 모델 교육을 수행하고, 최상의 모델을 선택할 수 있습니다. 이러한 도구는 데이터 분석가가 고급 프로그래밍 기술이나 심오한 수학/통계 지식 없이도 ML 기능을 수행할 수 있도록 하여 고급 분석에 대한 액세스를 민주화하는

훌륭한 방법입니다. 몇 가지 도구로 이 두 가지 접근 방식을 연결하여 데이터 과학자가 AutoML 프로세스를 사용자화할 수 있도록 합니다. AutoML 분야의 선두주자인 DataRobot에는 사용자가 DataRobot 계정을 Snowflake에 빠르게 연결하고 이를 데이터 저장소로 사용할 수 있는 Snowflake 통합이 내장되어 있습니다.

분석 및 클라우드 파트너

어떤 ML 접근 방식을 선택하든 Snowflake를 사용하면 Tableau, Looker, ThoughtSpot 및 Sigma와 같은 다른 생태계 파트너와의 연결을 활용하여 대시보드, 보고서 및 비즈니스 분석 도구를 통해 결과를 사용할 수 있습니다. 또한 Snowflake는 Amazon, Microsoft 및 Google의 인기 제품을 포함하여 모든 클라우드의 모든 지역에서 데이터를 저장하고 복제할 수 있도록 지원합니다. Snowflake는 Amazon S3, Azure Blob 및 Google Cloud Storage에서 유지 관리하는 외부 테이블로 데이터를 원활하게 내보내 어떤 도구로든 범용 액세스가 가능합니다. 예를 들어 Snowflake 를 사용하여 AWS에서 데이터 레이크를 보완한 다음 Amazon SageMaker와 연결하여 ML 모델을 대규모로 개발, 테스트 및 배포할 수 있습니다. 이 플랫폼은 데이터 저장 및 처리부터 트랜잭션 관리, 보안, 거버넌스 및 메타데이터 관리에 이르기까지 모든 것을 자동화합니다.

몇 분 안에 시작하기

Snowflake 및 ML을 시작하는 가장 빠른 방법을 찾고 있다면 선별된 기술 파트너와 사전 구성된 통합을 통해 배포를 간소화하는 Snowflake Partner Connect 프로그램을 고려해 보십시오. 몇 분 만에 파트너 애플리케이션을 자동으로 프로비저닝 및 구성하고 Snowflake에 데이터를 로드하여 바로 사용할 수 있습니다.

실제 사례

ConsumerTrack은 수백 개의 웹 사이트에서 CNN, MSN과 같은 포털에 이르기까지 웹 사이트 성능 데이터를 집계하여 판매하는 디지털 광고주이자 게시자입니다. 이 회사의 기존 데이터 과학 팀은 MySQL과 다양한 오케스트레이션 도구를 사용하는 ML 환경으로 어려움을 겪었고, 이로 인해 데이터 병목 지점과 대기 시간 문제가 발생했습니다.

ConsumerTrack은 기존의 데이터 레이크를 Snowflake로 보강하고 ML 워크플로우 자동화를 위한 완전 관리형 서비스로 Amazon SageMaker를 선택했습니다. 데이터에 레이블을 지정하여 준비하고, 알고리즘을 선택하고, 모델을 교육하고, 배포를 위해 모델을 조정 및 최적화하고, 예측한 다음 조치를 취합니다.

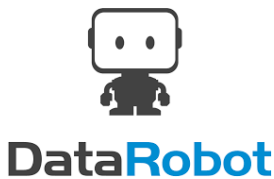
이제 데이터는 AWS Lambda 및 AWS Glue를 사용하는 자동화된 파이프라인을 통해 데이터 레이크로 흐릅니다. 데이터는 선별된 후 Snowflake에 로드되며 데이터 스트림은 사용자 지정 알림으로 구성됩니다. Amazon SageMaker를 Snowflake에 연결하여 ML 모델의 개발, 테스트 및 구축을 간소화합니다.

ConsumerTrack은 병목 지점을 제거했으며, 통찰력을 얻는 시간을 몇 시간에서 몇 분으로 단축했습니다. Snowflake가 데이터 검색 및 준비에 소요되는 시간을 상당히 줄입니다. Snowflake의 광범위한 생태계를 통해 ConsumerTrack은 Python용 기본 커넥터를 포함하여 다양한 유형의 데이터 과학 플랫폼 및 도구와 연결할 수 있습니다. 필요하다면 데이터 과학 팀이 범용 액세스를 위해 모든 Blob 스토리지에 데이터를 내보낼 수 있습니다.

다음 단계

머신 러닝에 대해 더 자세히 알아보려면, [Snowflake 데이터 과학 페이지](#) 및 [Snowflake 플랫폼 페이지](#)를 확인하십시오.

추천 데이터 과학 파트너



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/?lang=ko



© 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.

인용

- ¹ “모범 사례 보고서: AI 및 머신 러닝을 사용한 디지털 혁신 추진”(tdwi.org/bpreports).
- ² Forbes “빅 데이터 정리: 설문 조사에 따르면, 가장 시간이 오래 걸리고 가장 즐겁지 않은 데이터 과학 작업”(bit.ly/38EbXmN).
- ³ Forbes “빅 데이터 정리: 설문 조사에 따르면, 가장 시간이 오래 걸리고 가장 즐겁지 않은 데이터 과학 작업”(bit.ly/38EbXmN).