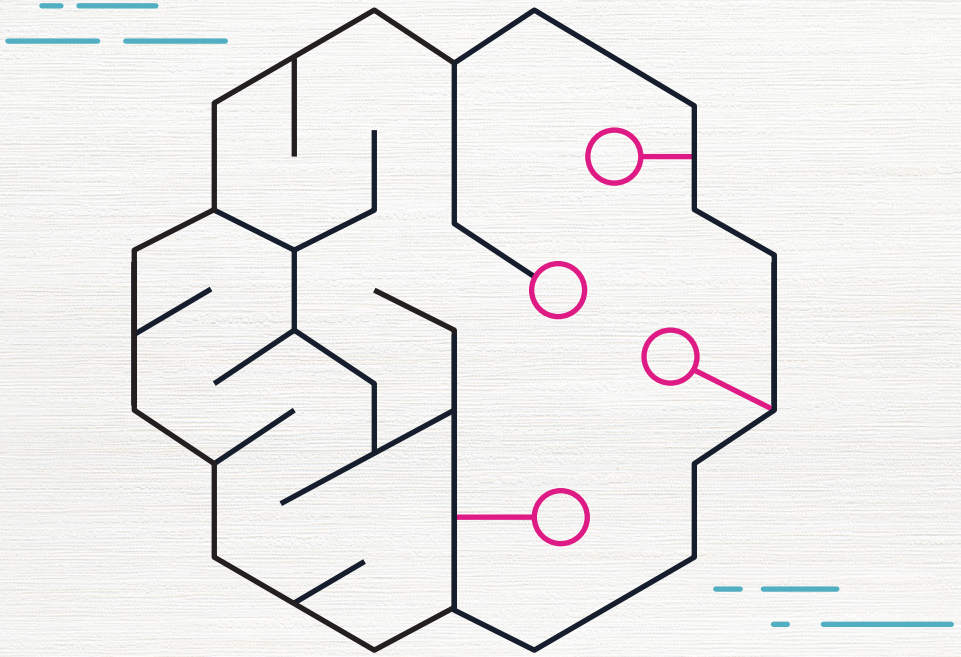
aws startups

기계 학습을 위한 데이터 프로세스 관리

기계 학습 사례를 위한 데이터 관리에 중점을 둔 참조 가이드



목차

서문 3

기계 학습 데이터의 이해. 4

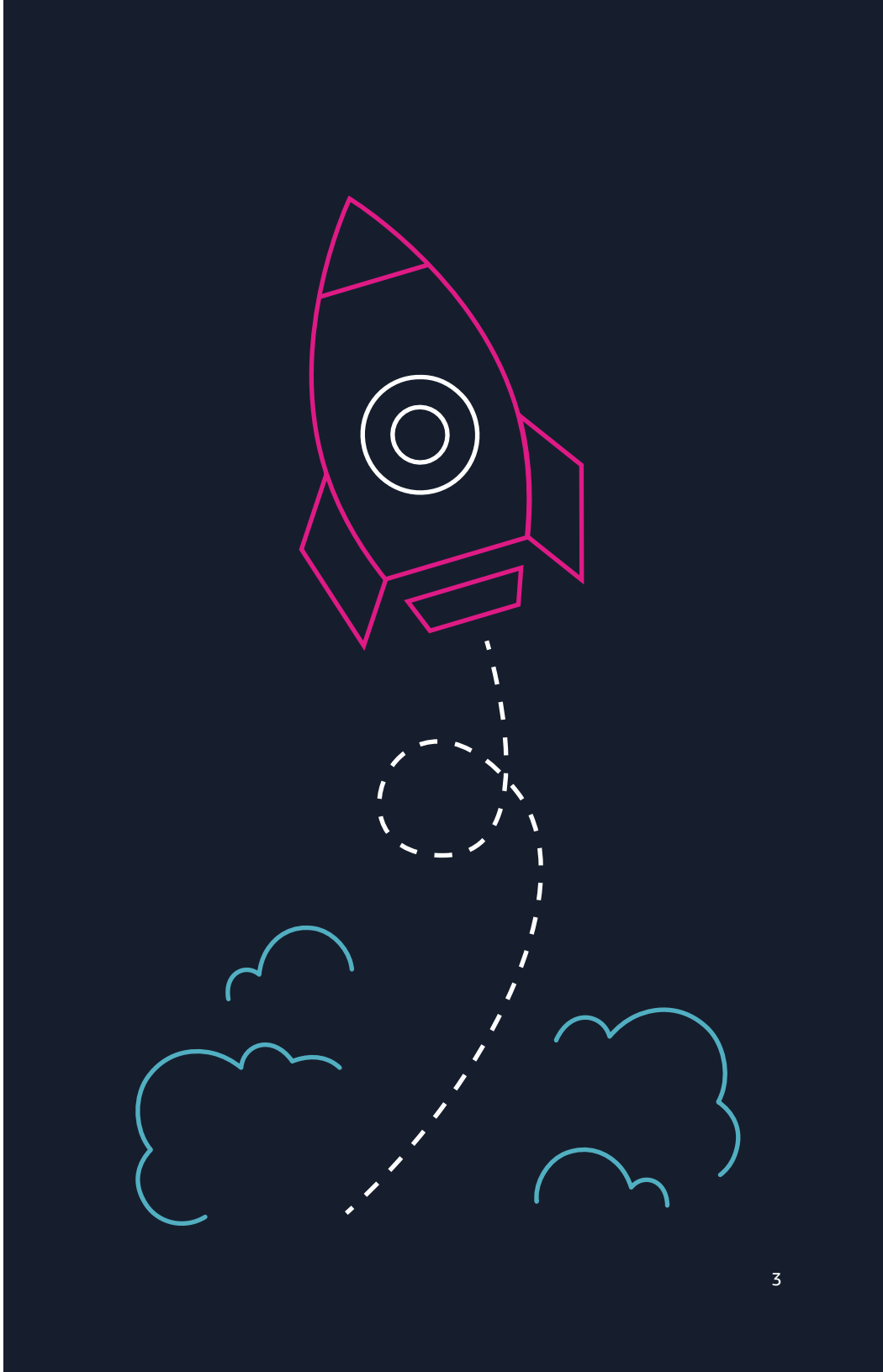
AWS로 데이터 관리 9

데이터 실용화 20

결론 24

서문

이 eBook은 기계 학습 사례의 핵심적인 부분인 데이터 처리에 대한 인사이트와 실용적 가이드를 제공합니다. 먼저 데이터 수집, 데이터 품질, 데이터 크기 조정 등 기계 학습 데이터에 대한 심도 깊은 이해를 제공하는 것으로 시작합니다. 그런 다음, AWS가 제공하는 관련 데이터 관리 및 처리 서비스를 다루는데, 기계 학습 데이터 관리에 이러한 서비스를 활용하는 성공적인 고객의 사례 연구도 함께 언급합니다. 이 eBook이 기계 학습 데이터 프로세스의 기반을 마련하고, 후속 단계의 기계 학습 사례에서 성공을 지속할 수 있는 길을 열어줄 것으로 기대합니다.



기계 학습 데이터의 이해

최신 기계 학습 기술은 데이터에 크게 의존하여 모델을 구축합니다. 기계 학습에 필요한 데이터의 다양한 측면을 이해하는 것은 기계 학습 도입의 성공 여부에 매우 중요합니다. 이 장에서는 기계 학습 프로젝트를 시작하는 여러분의 이해를 높이기 위해 기계 학습 데이터와 관련하여 이러한 몇 가지 중요한 측면을 살펴보겠습니다.

기계 학습의 데이터 원본

기계 학습 데이터 집합은 일반적으로 외부 데이터 원본 또는 내부 데이터 원본에서 가져올 수 있습니다. 외부 데이터 원본은 데이터 공급자가 제공하는 프라이빗 데이터 집합과 무료 개방형 데이터 집합으로 세분화할 수 있습니다.

외부 퍼블릭 기계 학습 데이터 원본

과거 운영 데이터, 연구 프로젝트 또는 기계 학습 과제 등을 통해 국제 기구 또는 기관, 국가 또는 지방 정부, 연구 기관 등의 조직에서 생성하거나 수집한 데이터 집합입니다. 흔히 사용되는 퍼블릭 기계 학습 데이터 집합의 몇 가지 예는 다음과 같습니다.

- AWS Data Exchange
- Registry of Open Data on AWS
- 데이터 | 세계은행
- 데이터와 맵 – 유럽환경청(EEA)
- 퍼블릭 데이터 집합: Amazon Web Services
- Google Public Data Explorer
- Kaggle 시합
- UCI 기계 학습 리포지토리

이 데이터 집합은 일반적으로 대중이 공개적으로 액세스, 사용, 기여할 수 있습니다. 기계 학습에 적합한 이 퍼블릭 데이터 원본 목록은 여기와 여기에서 찾을 수 있습니다.

외부 프라이빗 기계 학습 데이터 원본

특정 산업 또는 연구 영역에 더 독점적이고 잠재적으로 부가 가치가 높은 데이터 집합을 제공하는 공급업체입니다. 이러한 공급업체는 일반적으로 탐색 플랫폼(예: Explorium), 마켓플레이스(예: Datarade) 또는 기계 학습 SaaS 서비스(예: Calligo)를 통해 데이터 집합을 제공합니다. AWS가 제공하는 AWS Data Exchange 서비스를 통해 클라우드에서 서드 파티 데이터를 쉽게 찾고 구독하고 사용할 수도 있습니다.

내부 기계 학습 데이터 원본

고객 데이터, 비즈니스 운영 데이터 또는 지적 재산을 통해 획득한 조직 내의 데이터 집합입니다. 이 데이터 집합은 데이터베이스, 데이터 웨어하우스, 문서 스토어 같은 다양한 내부 시스템이나 CRM이나 ERP 시스템 같은 외부 공급업체의 SaaS 플랫폼에 저장할 수 있습니다. 이는 일반적으로 기업이 활용하려고 하는 주요 기계 학습 데이터 원본이지만 대개 여러 사일로에 분산되어 있으며 관리하기 어렵습니다. AWS 서비스를 활용하여 이 데이터 집합을 통합하고 관리하는 방법은 뒤에서 자세히 설명합니다.

기계 학습 데이터 집합의 데이터 품질

컴퓨팅 초기 시절부터 *"잘못된 데이터를 입력하면 잘못된 정보가 나온다(GIGO)"*는 말이 있었습니다. 이 말은 그 어느 때보다 인공 지능 시대에 의미가 있습니다. 인공 지능의 중요한 분과인 기계 학습은 방대한 양의 훈련 데이터를 사용하여 기계 학습 모델을 구축합니다. 본래 기계 학습 모델은 데이터 품질에 매우 민감합니다. Andrew Ng 박사의 데이터 중심 AI 참조 자료에 따르면, AI 개발자는 80%의 *시간을 데이터 준비에 사용하며, 데이터 최적화에 투자한 시간은 알고리즘 효율성보다 모델 성능을 높일 가능성이 높습니다*. 따라서 기계 학습 프로젝트의 첫 번째 단계로 기계 학습 데이터의 *일관성, 정확성, 호환성, 완전성, 적시성 뿐만 아니라 중복되거나 손상된 레코드를* 확인해야 합니다.

데이터 품질 측정 및 평가

데이터 과학의 세계에는 데이터 품질을 측정하고 평가하는 데 사용되는 많은 방법론이 있습니다. 예를 들어, 데이터 사이언티스트들은 **벤치마크**, **합의**, **Cronbach의 알파 테스트**, **검토**[1] 등 흔히 사용되는 데이터 품질 측정 프로세스를 따르고 데이터 품질을 평가하기 위해 정밀도, 편향, 정확도 같은 통계적 개념을 사용합니다[2]. 좀 더 실용적인 의미에서 데이터 품질에 영향을 미칠 수 있는 요소(예: 데이터 측정 및 수집 오류, 잡음, 특이값, 누락된 데이터)[2] 및 이를 측정할 좋은 지표 정의(예: 오류 대비 데이터 비율, 빈 값의 수)[3]를 이해하면 데이터 품질을 개선하는 데 도움이 됩니다.

80%

AI 개발자는 80%의 시간을 데이터 준비에 사용하며, 데이터 최적화에 투자한 시간은 알고리즘 효율성보다 모델 성능을 높일 가능성이 높습니다.

기계 학습 데이터 및 모델 품질 개선

기계 학습 모델을 개선하는 가장 좋은 방법은 훈련, 검증, 테스트를 위한 양질의 데이터 집합에서 시작됩니다. 따라서 데이터 집합의 사전 처리(예: 정리, 변환, 대치)는 중요한 단계입니다. Amazon SageMaker는 SageMaker Studio의 SageMaker Data Wrangler, SageMaker Processing 등 데이터를 시각화하고 준비하는 데 도움이 되는 여러 기능을 제공하여 기계 학습 데이터를 개선하도록 지원합니다. SageMaker Clarify 역시 기계 학습(ML) 모델의 편향을 감지하여 모델 품질을 개선하는 데 도움이 됩니다. 이러한 기능은 뒷부분에서 더 자세히 설명하겠습니다. 데이터 정리와 변환을 넘어서는 다른 고려 사항도 있습니다[4]. 고려해야 할 한 가지 중요한 점은 ML 데이터와 모델을 계속 모니터링하고(예: 모델 드리프트 방지), 최신 데이터 집합으로 모델을 다시 훈련하는 것입니다. 기계 학습[5]을 위한 자동화된 데이터 품질 관리에 대한 연구가 진전되면서, 기계 학습 모델을 지속적으로 개선하기 위한 Amazon SageMaker 데이터 품질 모니터링 및 모델 품질 모니터링 기능이 개발되었습니다.

기계 학습 모델 훈련을 위한 데이터 샘플링 크기

기계 학습은 데이터에 크게 의존합니다. *"얼마나 많은 데이터가 필요한가요?"*라는 질문도 많이 합니다. 다양한 문제(예: 분류, 예측, 이상 탐지, 클러스터링)를 해결하기 위한 다양한 학습 패러다임(예: 지도 학습 및 비지도 학습), 애플리케이션 도메인(예: 텍스트 분석, 자연어 처리 및 이미지 처리 등)만으로도 기계 학습은 이미 복잡합니다. 모델 훈련에 사용되는 기계 학습 알고리즘과 프레임워크 등 고려해야 할 다른 요소도 있습니다. 이러한 요소 외에도 모델 정확도, 훈련 시간, 비용에 대한 구체적인 요구 사항에 따라 필요한 데이터 크기도 결정됩니다. 따라서 **이 질문에 정해진 답은 없습니다. 경우에 따라 다릅니다. 하지만 아래와 같이 몇 가지 일반적인 경험 법칙이 있으며, 필요에 맞게 실험해보면서 조정할 수 있습니다**.

샘플링 데이터 크기 또는 기계 학습 훈련 데이터 집합을 주제로 많은 연구가 진행되었습니다[6,7,8]. 일반적 지침이 되는 단순한 경험 법칙을 예로 들면, 데이터 집합의 크기는 차원의 약 10배 이상이어야 하고 사용된 모델과 독립적이어야 합니다[9,10,11]. 다른 일반적인 기계 학습 도메인 또는 문제 유형의 경우, 샘플 데이터 크기에 대한 일반적인 가이드가 오른쪽 표에 나열되어 있습니다.

초기 기계 학습 프로젝트 계획을 위한 일반적인 가이드로 사용하십시오. 해결해야 할 구체적인 문제, 사용된 프레임워크와 알고리즘, 위에서 언급한 기타 고려해야 할 요구 사항을 기반으로 데이터 집합 크기를 조정해야 합니다.

이제 기계 학습 데이터의 가치와 다양한 측면을 이해했으므로 다음 장에서는 AWS 기계 학습 관련 서비스가 기계 학습 프로젝트에 착수하기 위해 기계 학습 데이터를 수집, 처리하고 관리하는 여러분에게 어떤 도움이 되는지 말씀드리겠습니다.

	학습 패러다임 또는 도메인	문제 유형	데이터 크기 경험 법칙	참고	참조
1	지도 학습	선형 회귀, 이진 또는 멀티클래스 분류	10:1 비율의 데이터 샘플: 차원	예: XGBoost	9, 10, 11
2	지도 학습	시계열 예측	50개~100개 관찰	예: ARIMA 모델 사용	12, 13
3	컴퓨터 비전	이미지 분류	클래스당 1,000개 이미지	예: ImageNet 데이터 집합 기반의 CNN 또는 DL	14, 15, 16
4	텍스트 분석	문서 분류	범주당 문서 100개 미만*	예: 문서 분류	17
5	자연어 처리(NLP)	감정 분석	200~300단어**	예: Yelp Academic Dataset를 사용한 감정 분석	18

*,** 참고: 이 수치는 대부분 실험 데이터를 기준으로 합니다.

AWS로 데이터 관리

기계 학습에서 데이터의 가치와 중요성에 대해 이해했으니 이제 기계 학습에 데이터를 효과적으로 사용하는 방법을 알아보겠습니다. 이 섹션에서는 기계 학습 프로세스의 일반적인 단계와 함께 각 단계를 쉽고 효율적으로 달성하는 데 도움이 되는 AWS 서비스를 살펴보겠습니다.

데이터 수집

앞서 언급했듯이 오늘날 기계 학습은 통찰력 있고 정확한 모델을 구축하기 위해 데이터에 의존합니다. 첫 단계는 기계 학습 모델에 필요한 데이터를 파악하고, 모델을 훈련하기 위해 그 데이터를 수집하는 데 사용할 수 있는 다양한 수단을 평가하는 것입니다. AWS는 정적 리소스에서 대량으로 또는 새롭게 동적으로 생성되는 리소스(예: 웹 사이트, 모바일 앱, 인터넷에 연결된 디바이스)로부터 데이터를 수집하는 여러 가지 방법을 제공합니다.

Amazon Kinesis 제품군: *스트리밍 데이터*는 여러 데이터 원본이 지속적으로 생성하는 데이터로서, 대개 작은 크기(킬로바이트 단위)의 데이터 레코드를 동시에 전송합니다. 스트리밍 데이터에는 애플리케이션 로그 파일, 전자 상거래 구매, 소셜 네트워크에서 얻은 정보 등 다양한 데이터가 포함됩니다. Kinesis는 다음 스트리밍 데이터를 수집, 처리하고 저장하는 데 도움이 됩니다.

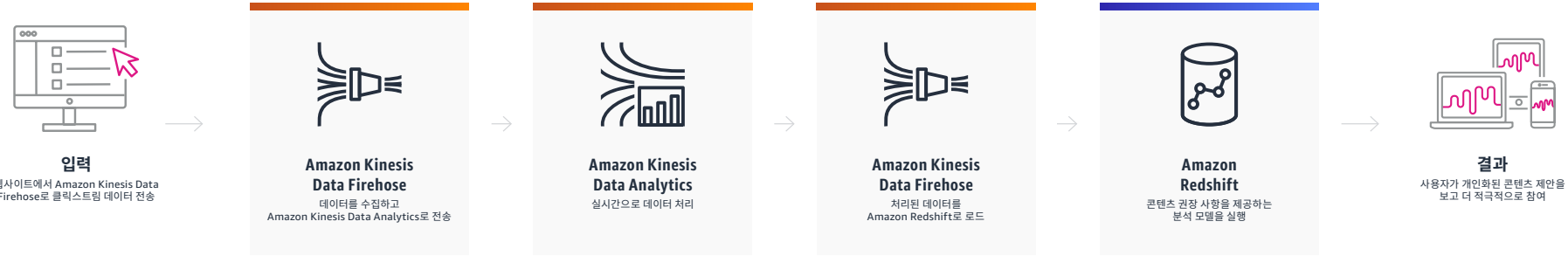
- Amazon Kinesis Data Streams 는 AWS의 확장 가능한 실시간 데이터 스트리밍 서비스입니다. KDS는 수천 개의 소스에서 초당 GB 규모의 데이터를 지속적으로 캡처하고, 수집한 데이터를 밀리초 이내에 제공하여 실시간 분석 사용 사례를 지원합니다.

- Amazon Kinesis Data Firehose를 사용하면 스트리밍 데이터를 데이터 레이크, 데이터 스토어 및 분석 서비스에 쉽게 로드할 수 있습니다. Kinesis Data Firehose는 완전관리형이며 데이터 처리량에 따라 자동으로 확장됩니다. 또한 로드하기 전에 데이터 스트림을 배치화하고 압축, 변환, 암호화할 수 있습니다.

- Amazon Kinesis Data Analytics를 사용하면 간단한 3단계(스트리밍 데이터 원본 설정, 쿼리 또는 스트리밍 애플리케이션 작성, 처리된 데이터의 대상 설정)에 따라 쿼리 및 정교한 스트리밍 애플리케이션을 쉽고 빠르게 구축할 수 있습니다.

예제: 스트리밍 소셜 미디어 데이터 분석

예: 클릭스트림 분석



클라우드 데이터 마이그레이션: 데이터를 클라우드로 옮기려면 다양한 사용 사례별로 데이터를 이동할 위치, 이동할 데이터 유형, 사용 가능한 네트워크 리소스 등을 고려해야 합니다. AWS는 하이브리드 클라우드 스토리지, 온라인 데이터 전송, 오프라인 데이터 전송의 필요성 등 데이터 마이그레이션 프로젝트에 적합한 솔루션을 지원하기 위해 데이터 전송 서비스 포트폴리오를 제공합니다. 자세한 내용은 클라우드 데이터 마이그레이션 페이지를 참조하십시오.

서드 파티 도구: AWS는 모든 고객과 각각의 사용 사례를 지원하고자 합니다. 일부 데이터는 AWS 클라우드 네이티브가 아닐 수 있습니다. 따라서 AWS는 서드 파티 데이터 플랫폼과의 통합도 허용합니다. 예를 들어 인기 있는 데이터 웨어하우스 플랫폼인 Snowflake는 대규모로 확장 가능한 객체 스토리지 솔루션인 Amazon Simple Storage Service(Amazon S3)와의 통합을 지원합니다. S3의 액세스 관리 정책을 통해 Snowflake에 S3 액세스 권한을 부여하면 되며, 이 내용은 뒤에서 다루겠습니다.

사례 연구: 기계 학습 모델을 훈련하기 위해 데이터를 수집할 수집 파이프라인 개발에 성공한 몇몇 스타트업을 살펴보겠습니다. 첫 번째는 독특한 패션 마켓플레이스를 주도하는 대안적 쇼핑 경험을 제공하는 Depop입니다[19]. Depop은 Amazon Kinesis Data Firehose, Amazon Managed Streaming for Kafka를 사용하여 2,500만 개 품목과 트랜잭션으로 구성된 방대한 인벤토리를 스트리밍합니다. 데이터 수집에 관리형 서비스를 활용함으로써 Depop은 인프라 관리보다 고객 서비스 개발에 집중할 수 있었습니다. 다음은 분석이 필요한 처방 및 청구 데이터를 수집하고, 전문 의료 종사자 팀의 개입이 필요한 사례를 위탁하는 axialHealthcare입니다[20]. axialHealthcare의 클라우드 기반 고객 센터는 Amazon Kinesis Data Streams를 사용하여 상담원 상태 이벤트를 모니터링하고 Amazon S3를 사용하여 통화 녹음을 객체로 저장합니다.

depop | axial Healthcare

데이터로 분석 수행

기계 학습 모델을 개발하기 위해서는 데이터를 이해해야 합니다. 탐색 데이터 분석[21]은 패턴 발견, 이상 발견, 가설 테스트 등을 위해 데이터에 대한 초기 조사를 수행하는 프로세스를 뜻합니다. 또한, 당사 모델의 맥락에서 특성을 분석함으로써 어떤 특성이 다른 특성들보다 중요한지에 대한 직관을 얻을 수 있습니다. 일부 특성은 모델 성능을 개선하는 반면, 다른 특성은 모델 성능을 개선하지 않거나 심지어 저해합니다. AWS는 그동안 수집하여 클라우드에 저장한 데이터를 바로 탐색할 수 있는 서비스를 제공합니다.

데이터를 이해할 때는 패턴 파악이 중요합니다. 이러한 패턴은 테이블에 있는 데이터만 보았을 때는 명확하게 드러나지 않는 경우가 많습니다. 적절한 시각화 도구를 사용하면 데이터에 대해 더욱 빠르게 심층적으로 이해할 수 있습니다. 도표 또는 그래프를 작성하기 전에 무엇을 표시할지 결정해야 합니다. 예를 들어, 차트는 핵심 성과 지표(KPI), 관계, 비교, 분포 또는 구성과 같은 정보를 전달할 수 있습니다.

Amazon Athena는 표준 SQL을 사용하여 Amazon S3에 저장된 데이터를 간편하게 분석할 수 있는 대화형 쿼리 서비스입니다. Athena는 서버리스 서비스이므로 관리할 인프라가 없으며 실행한 쿼리에 대해서만 비용을 지불하면 됩니다. Athena를 사용하면 별도의 변환 없이 원시 형식의 데이터를 직접 쿼리할 수 있습니다.

Amazon QuickSight는 대화형 대시보드를 쉽게 만들고 게시할 수 있는 기계 학습 기반의 확장 가능한 비즈니스 인텔리전스 서비스입니다. QuickSight가 함께 제공하는 ML Insights는 AWS의 검증된 기계 학습 및 자연어 기능을 활용하여 데이터에 숨은 인사이트와 추세를 발견하고, 핵심 동력을 파악하고, 비즈니스 지표를 예측하는 데 도움이 됩니다. Amazon SageMaker에 구축된 기계 학습 모델에 QuickSight를 연결하여 예측 대시보드를 생성할 수도 있습니다.

Amazon SageMaker Studio는 모든 기계 학습 개발 단계를 수행할 수 있는 단일의 웹 기반 시각적 인터페이스를 제공합니다. Studio는 빠르게 가동할 수 있는 원클릭 Jupyter 노트북을 제공합니다. 기본 컴퓨팅 리소스는 매우 탄력적이므로, 적절한 컴퓨팅 파워를 필요에 따라 쉽게 늘리거나 줄일 수 있습니다. 인기 있는 Python 오픈 소스 데이터 분석 및 조작 도구인 pandas 같은 도구로 노트북 환경에서 바로 데이터 집합 분석을 시작할 수 있습니다. 데이터 시각화를 위해 Matplotlib 및 Seaborn 같은 오픈 소스 라이브러리를 사용할 수 있습니다.

데이터 관리 - 데이터 레이크

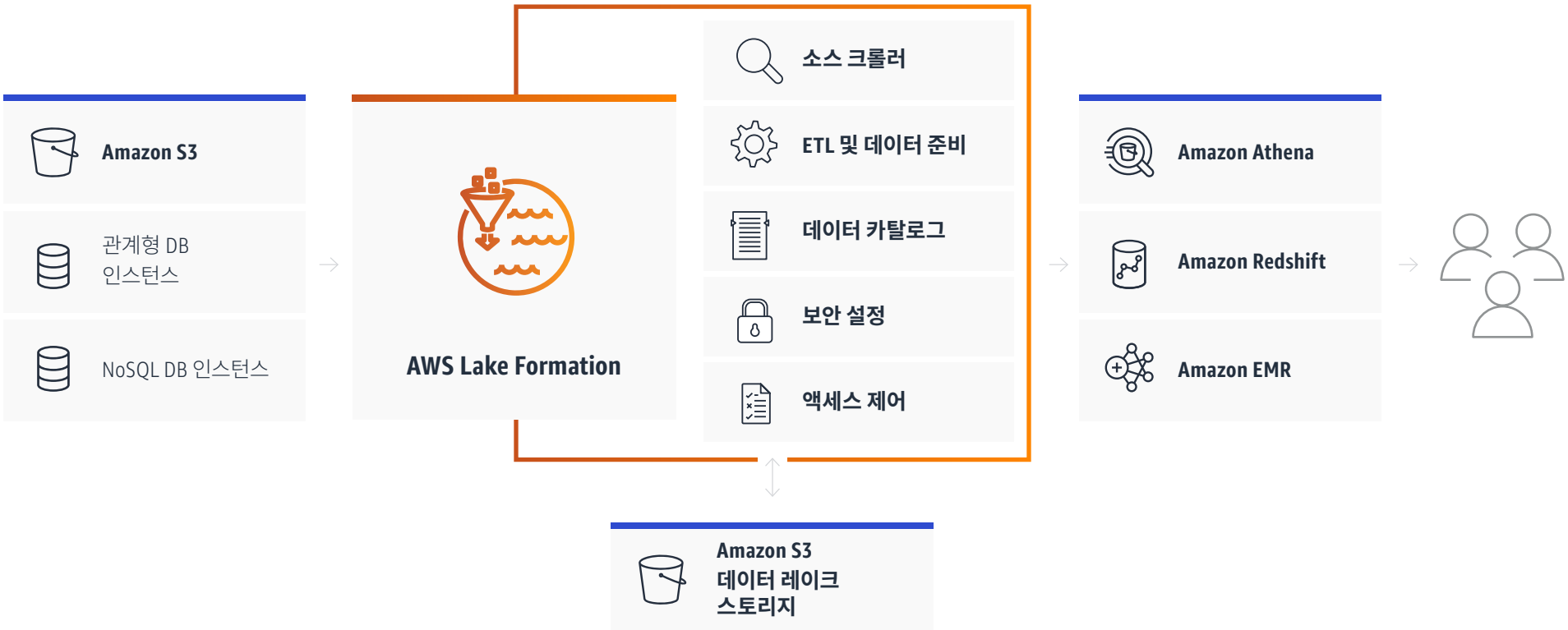
예전에는 누적된 운영 데이터가 다양한 데이터 사일로에 상주했기 때문에 분석을 수행하기가 매우 어려웠습니다. 데이터 사일로에는 많은 문제가 있습니다. 예를 들어 주어진 워크로드에 필요한 데이터가 여러 사일로에 분할되어 액세스가 불가능한 경우가 있습니다. 데이터가 상주하는 사일로가 해당 워크로드에 대한 비용 요건에 부합하지 않을 수 있습니다. 여러 사일로에 서로 다른 관리, 보안, 권한 부여 방식이 필요한 경우 운영 비용과 위험이 증가할 수 있습니다. 데이터 레이크를 사용하면 고도의 확장성, 가용성, 보안, 유연성을 보장하는 중앙의 한 데이터 스토어에 기업 전체의 데이터(정형 및 비정형)를 저장하고 카탈로그화하여 기계 학습, 분석 등의 사용 사례에 맞는 대용량 데이터 집합을 처리할 수 있습니다. AWS의 레이크 하우스 아키텍처를 통해 조직은 데이터 레이크 뿐만 아니라 주변에 목적별 분석 스토어 및 서비스를 구축함으로써 데이터 레이크와 주변의 목적별 분석 스토어 및 서비스 사이에 데이터 이동(내부에서 외부로, 외부에서 내부로, 측면으로)을 지원하여 데이터로부터 인사이트를 얻을 수 있습니다[22,23].

레이크 하우스 아키텍처의 핵심은 데이터 레이크이며 이는 Amazon S3에서 시작됩니다. Amazon S3는 정형 및 비정형 데이터에 맞게 최고의 규모와 성능을 자랑하는 객체 스토리지 서비스로, 데이터 레이크 구축에 최적화된 스토리지 서비스입니다[24]. 99.999999999%의 내구성으로 데이터가 보호되는 안전한 환경에서 규모에 관계없이 확장 가능하고 경제적인 데이터 레이크를 구축할 수 있습니다. Amazon S3는 다양한 데이터 액세스 수준을 그에 합당한 속도로 지원하는 비용 효율적이고 광범위한 S3 스토리지 클래스를 제공합니다. S3 스토리지 클래스 분석을 사용하면 액세스 패턴을 기반으로 더 낮은 비용의 스토리지 클래스로 옮겨야 하는 데이터를 검색하고, 전송을 수행할 S3 수명 주기 정책을 구성할 수 있습니다. 그리고 S3 Intelligent-Tiering에 액세스 패턴이 변경 중이거나 알려지지 않은 데이터도 저장할 수 있습니다. 그러면 액세스 패턴의 변화에 따라 객체를 계층화하므로 자연히 비용을 절감하는 효과가 있습니다. 고객들은 다른 어떤 클라우드 제공업체보다 오랫동안 Amazon S3에 데이터 레이크를 구축해 왔고, 현재 그 어느 곳보다 AWS에서 실행되고 있는 데이터 레이크가 더 많으며, 서버리스 대화형 쿼리 서비스인 Amazon Athena를 사용하여 모든 데이터를 분석하고 있습니다.

고객들은 데이터 레이크 외에도 Amazon EMR, Amazon Elasticsearch Service, Amazon Redshift 같은 AWS의 목적별 데이터베이스와 분석 서비스 조합을 사용하여 작업에 적합한 도구를 확보함으로써 최저 비용으로 뛰어난 성능과 확장성을 보장하고 있습니다. 고객은 서버리스 데이터 통합 서비스인 AWS Glue를 사용하여 이러한 시스템 간에 데이터를 이동합니다. AWS Lake Formation을 사용하는 고객은 데이터 레이크 또는 목적별 분석 스토어를 막론하고 모든 데이터의 보안과 거버넌스를 관리할 수 있습니다. 특히 AWS Lake Formation은 고객이 모든 데이터에 대한 단일 보안 및 거버넌스 제어, 세분화된 액세스 제어를 사용한 데이터 민감도 제어, 전체 데이터 레이크에 대한 중앙 집중식 감사 제어 기능을 갖춘 데이터 레이크를 구축할 수 있도록 지원합니다.

레이크 하우스 아키텍처에 제시된 목적별 데이터 스토어 중 하나는 기계 학습 사용 사례에 해당합니다. 이 데이터 여정은 데이터 레이크에 수집된 원시 데이터에서 시작되어 데이터 변환/특성 추출 프로세스를 거쳐 모델 훈련과 추론에 필요한 특성 생성으로 이어집니다. 이런 특성이 생성된 후에는 데이터 사이언티스트들이 각자의 모델 훈련/추론에 사용할 수 있도록 저장, 검색, 공유되어야 합니다. 여기에서 여러 모델이 여러 팀의 공통 특성을 공유할 수 있습니다(다음 그림 참조). 무엇보다 모델을 개선하려면 데이터 사이언티스트들이 계속 새로운 특성을 추가해야 합니다. 기존 특성을 다시 생성하거나 컴퓨팅하는 작업은 시간이 많이 걸리고 모델 예측에 대기 시간이 추가되기 때문에 누구도 원치 않습니다.

작동 방식



Amazon SageMaker 특성 저장소는 데이터 사이언티스트가 모델 훈련 및 추론을 위해 다른 팀의 데이터 사이언티스트들과 공통 데이터 특성을 공유하는 데 유용합니다. SageMaker 특성 저장소는 훈련을 위한 오프라인 특성 저장소와 온라인 추론을 위한 온라인 특성 저장소를 모두 지원합니다[25]. SageMaker 특성 저장소는 모델 훈련을 위해 대규모 배치로 특성을 제공할 뿐만 아니라, 실시간 추론 사용 사례에 맞게 밀리초 단위의 대기 시간으로 읽기가 가능한 특성을 제공합니다[26]. 특성 저장소는 모델 특성의 중앙 리포지토리 역할을 하므로 여러 특성이 일관성 있게 표시됩니다. 즉, 훈련과 추론에 정확히 동일한 특성을 사용할 수 있으므로 특성은 훈련과 추론 간에 동기화된 상태를 유지합니다. SageMaker Studio에서 특성을 시각적으로 검색하고 찾을 수 있습니다. 팀의 모든 구성원은 리포지토리에 특성을 공유하여 재사용을 촉진하고 재작업을 방지할 수 있습니다. 특성 저장소는 여러 팀에 걸쳐 통합된 특성 정의 세트도 제공하므로 여러 팀이 협업하기도 더 쉽습니다.

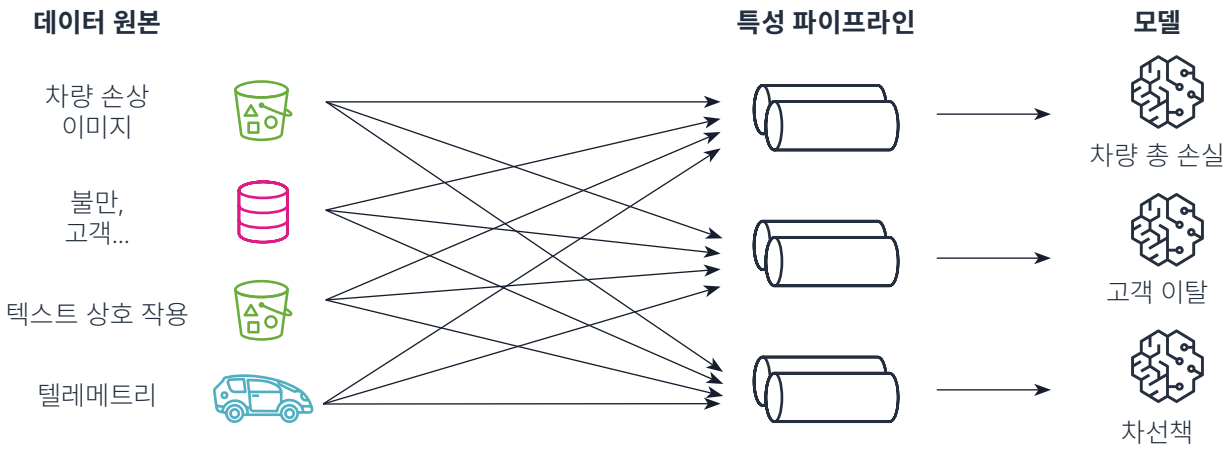
아래 다이어그램은 각각 전용 독립형 특성 추출과 공유 특성 저장소를 사용하는 모델 훈련을 개념적으로 비교하여 보여줍니다.

사례 연구: 온라인 식료품 플랫폼 HappyFresh는 S3[27] 기반의 데이터 레이크를 개발했습니다. 이 스타트업은 클릭스트림 데이터를 Amazon S3에 저장하고, AWS Glue를 사용해 데이터를 추출하고 고객 쇼핑 패턴을 분석하기 위해 처리합니다. S3로 데이터를 관리하는 HappyFresh가 개인화 서비스에 집중하자, 고객들은 제품 검색에 귀중한 시간을 낭비하지 않게 되었습니다.

happyfresh

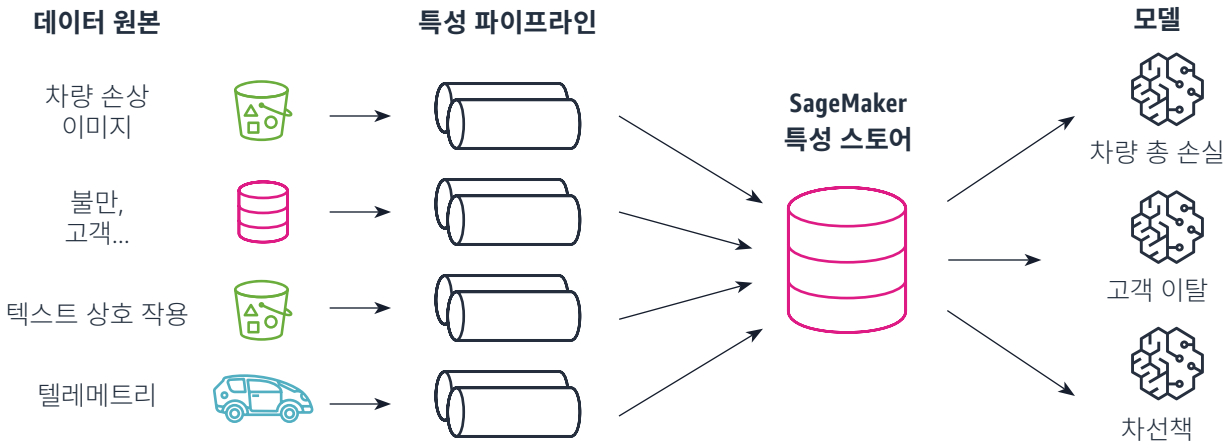
데이터 여정

각각의 새 모델에 맞는 독립형 특성 추출



- 특성 중복
- 출시 시간 지연
- 부정확한 예측

일단 특성을 구축한 후 특성 스토어를 활용해 여러 팀과 모델에 재사용



- 검색을 통해 찾을 수 있는 특성 그룹
- 재현 가능한 특성 변환
- 정확한 훈련 데이터 집합 추출
- 짧은 대기 시간으로 간섭 탐지
- 훈련 및 간섭에 대한 일관된 특성

데이터 처리

데이터를 수집한 후에는 모델이 효과적으로 사용할 수 있는 형태가 되도록 하려면 데이터를 어떻게 처리해야 할지를 고민해야 합니다. 처리 단계의 예에는 ML 알고리즘이 기대하는 입력 형식으로 데이터 변환하기, 열의 크기 조정 및 정규화, 텍스트 정리 및 토큰화 등이 포함됩니다.

기계 학습 모델은 이를 훈련하는 데 사용하는 데이터의 품질에 따라 성능이 달라집니다. 데이터를 수집한 후에는 해당 데이터의 통합, 준비, 처리 등이 중요합니다. 학습과 일반화에 최적화된 훈련 데이터가 성공적이고 정확한 모델을 만들 수 있는 열쇠입니다. 데이터 준비는 작고 통계적으로 유효한 샘플에서 시작해서 여러 데이터 준비 전략을 적용해 반복적으로 개선하는 한편, 데이터 무결성을 계속해서 관리해야 합니다[28]. AWS는 데이터에 주석을 달고 대규모로 데이터를 추출, 전송 및 로드(ETL)하는 데 사용할 수 있는 여러 가지 서비스를 제공합니다.

AWS Glue: 데이터를 입수했으면 일반적으로 '데이터 통합'이라고 하는 과정을 통해 분석과 기계 학습에 맞게 데이터를 준비하고 결합해야 합니다. AWS Glue는 시각적 인터페이스와 코드 기반 인터페이스를 모두 제공하므로 데이터 통합이 더 쉽습니다. 일정에 따라 또는 트리거를 통해 주기적으로 실행되도록 Glue를 구성할 수 있습니다. 예를 들어, 새 데이터가 S3 버킷에 도달하면 Glue가 실행되도록 할 수 있습니다.

AWS Batch: 처리 성능. AWS는 고객이 확장 가능 GPU 워크로드를 실행할 수 있도록 G4, P4 같은 GPU 인스턴스 패밀리를 제공합니다. AWS Batch를 사용하면 AWS에서 여러 배치 컴퓨팅 작업을 대규모로, 효율적으로 실행할 수 있습니다. 제출하는 배치 작업을 기반으로 최적화된 개수와 유형의 컴퓨팅 리소스를 동적으로 프로비저닝합니다. 게다가 AWS Batch에서는 제출된 작업의 일정이 예약되고 적절한 인스턴스에 배치되므로 작업의 수명 주기도 관리됩니다. 고객이 제공하는 AMI가 추가되면 AWS Batch 사용자는 GPU가 필요한 작업에 이러한 탄력성과 편의성을 누릴 수 있습니다.[29]

Amazon EMR: 많은 조직이 데이터 처리에 Spark를 사용합니다.[30] 이 상황에서 Spark 클러스터는 일반적으로 Hadoop 생태계 클러스터에 대한 관리형 서비스인 Amazon EMR에서 실행되므로, 자체 설정, 튜닝, 유지 관리의 필요성이 줄었습니다. EMR을 사용하면 원하는 컴퓨팅, 메모리 및 스토리지 파라미터로 사용자 지정 작업을 실행할 수 있습니다. 자동화된 클러스터 설정 및 자동 크기 조정을 제공하고, 비용 절감을 위해 스팟 인스턴스를 지원합니다. EMR을 사용하면 데이터 집합 가져오기, 내보내기, 결합 등 방대한 양의 데이터 작업을 빠르고 비용 효율적으로 수행할 수 있습니다.

사례 연구: Guru는 지식 관리 소프트웨어를 제공하는 스타트업입니다[31]. Guru는 Amazon OpenSearch Service를 사용하여 Elasticsearch 클러스터의 스토리지와 크기 조정을 관리합니다. 이 회사는 Amazon EMR을 사용하여 검색 엔진의 검색 결과 정확도를 개선하기 위한 실험 프레임워크도 개발할 수 있었습니다. 다음으로, AiCure는 환자 행동을 모니터링하고 임상 시험에 원격으로 환자가 참여할 수 있도록 하는 AI 및 고급 데이터 분석 회사입니다[32]. AiCure는 AWS Step Functions 및 AWS Batch를 활용해 AI 모델과 대규모 추론을 지속적으로 개선하여 데이터의 실효성을 높입니다.

Amazon SageMaker

데이터 스키마는 데이터 구성 및 작업을 시작할 때 사용하기 좋은 방법입니다. 그러나 스키마는 진화하고 코드는 구식이 되고 쿼리는 느려진다는 점을 명심하십시오. Data Centric AI는 기계 학습 엔지니어링 및 데이터 과학 팀을 비롯한 다운스트림 소비자에게 양질의 데이터를 제공하기 위해 데이터를 다듬는 데 중점을 두며, 이는 반복적인 접근 방식입니다. 데이터 품질로 인해 데이터 처리 파이프라인이 중단될 수 있습니다. 이 문제를 일찍 발견하지 못하면 오해의 소지가 있는 보고서, 편향된 AI/ML 모델, 기타 의도치 않은 데이터 제품이 만들어질 수 있습니다.

SageMaker Ground Truth: 정확하게 레이블이 지정된 데이터는 지도형 모델의 성공에 필수적입니다. 잘못된 레이블이 있는 경우 ML 모델은 '나쁜' 예를 학습하고, 이는 부정확한 예측으로 이어집니다. SageMaker Ground Truth는 데이터에 효율적이고 정확하게 레이블을 지정하는 데 도움이 됩니다. 자동과 수동 데이터 레이블 지정을 조합하여 사용합니다. 데이터 레이블 지정 작업에 대한 사용자 인터페이스(UI)를 정의하는 사용자 지정 워크플로를 구축할 수도 있습니다. 시작하는 데 도움을 드리기 위해 Amazon SageMaker는 이미지, 텍스트 및 오디오 데이터 레이블 지정 작업을 위한 사용자 지정 템플릿 블로그를 제공합니다.

SageMaker Data Wrangler는 기계 학습, 데이터 분석, 특성 추출, 특성 중요성 분석, 편향 탐지 용도로 특별히 설계되었습니다. Data Wrangler는 특성 추출 및 편향 완화를 위해 300개 이상의 데이터 변환을 기본 제공합니다.

SageMaker Processing Jobs는 scikit-learn 또는 Apache Spark처럼 친숙한 오픈 소스 도구를 사용하여 완전관리형 종량제 AWS 인프라에서 원하는 Python 스크립트 또는 사용자 지정 Docker 이미지를 실행할 수 있습니다. 이 서비스는 클러스터의 수많은 SageMaker 인스턴스에서 사용자 지정 스크립트 또는 Docker 이미지를 병렬화할 수 있습니다. SageMaker Processing을 사용하면 스크립트를 제공하고 인스턴스 유형과 클러스터 크기만 지정하면 됩니다.

SageMaker Clarify: SageMaker Clarify는 SageMaker Data Wrangler로 편향을 탐지할 뿐만 아니라, 모델 훈련에 최적의 열(즉 '특성')을 선택하고, 훈련 후 모델에서 편향을 찾아내고, 모델 예측을 설명하고, 모델 예측 입력 및 출력의 통계적 오차를 탐지하는 데 도움이 됩니다.

준비 →

SageMaker Ground Truth
기계 학습을 위한 훈련 데이터에 레이블 지정

SageMaker Data Wrangler 신규
기계 학습을 위한 데이터 집계 및 준비

SageMaker Processing
기본 제공 Python, BYO R/Spark

SageMaker 특성 저장소 신규
특성 저장, 업데이트, 검색 및 공유

SageMaker Clarify 신규
편향 탐지 및 모델 예측 이해

구축 →

SageMaker Studio Notebooks
탄력적 컴퓨팅 및 공유 기능이 있는 Jupyter 노트북

기본 제공 및 기존 보유 알고리즘
수십 개의 최적화된 알고리즘 또는 기존에 보유한 알고리즘 사용

로컬 모드
로컬 시스템에서 테스트 및 프로토타입 생성

SageMaker Autopilot
완벽한 가시성이 보장되는 기계 학습 모델 자동 생성

SageMaker Jumpstate 신규
일반 사용 사례에 적합한 사전 구축된 솔루션

훈련 및 튜닝 →

원클릭 훈련
분산된 인프라 관리

SageMaker Experiments
모든 단계를 캡처, 구성 및 비교

Automatic Model Tuning
하이퍼 파라미터 최적화

분산 훈련 라이브러리 신규
대규모 데이터 집합 및 모델에 대한 훈련

SageMaker Debugger 신규
훈련 실행 디버깅 및 프로파일링

Managed Spot Training
훈련 비용 90% 절감

배포 및 관리 →

원클릭 배포
완전관리형, 매우 짧은 대기 시간, 높은 처리량

Kubernetes 및 Kubeflow 통합
Kubernetes 기반 기계 학습 간소화

다중 모델 엔드포인트
인스턴스 하나당 여러 모델을 호스팅하여 비용 절감

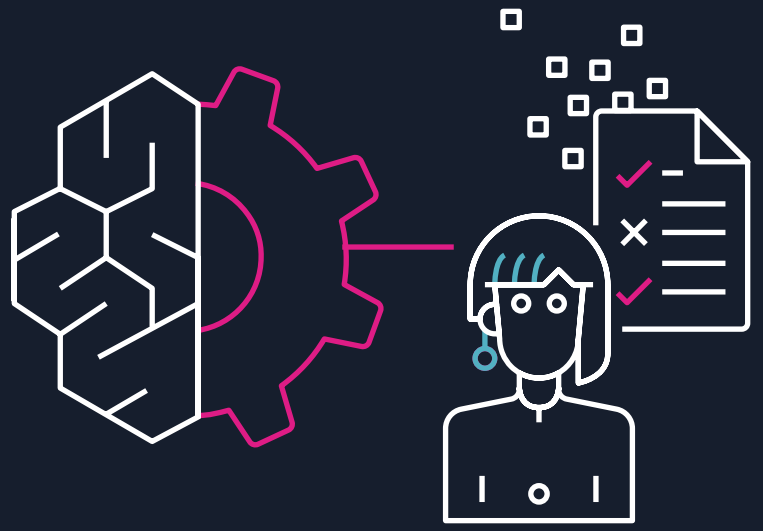
SageMaker Model Monitor
배포한 모델의 정확성 유지

SageMaker Edge Manager 신규
엣지 디바이스에서 모델 관리 및 모니터링

SageMaker Pipelines 신규
워크플로 오케스트레이션 및 자동화

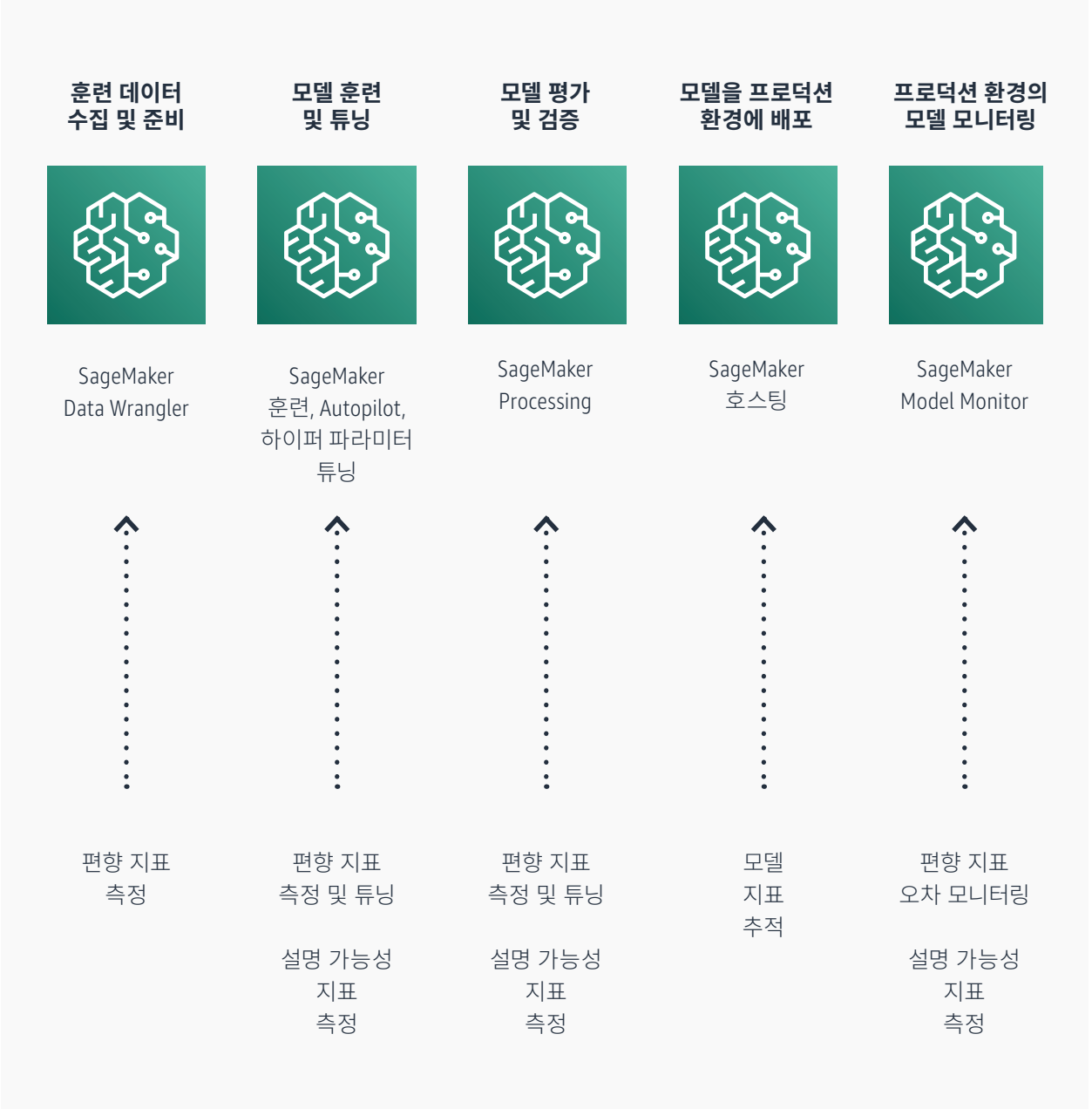
데이터 실용화

기계 학습을 수행할 때 데이터 관리를 통해 가치를 극대화하는 방법을 익혔으니 이제 모델을 훈련하고 튜닝할 때 유용한 AWS 도구를 살펴보겠습니다.



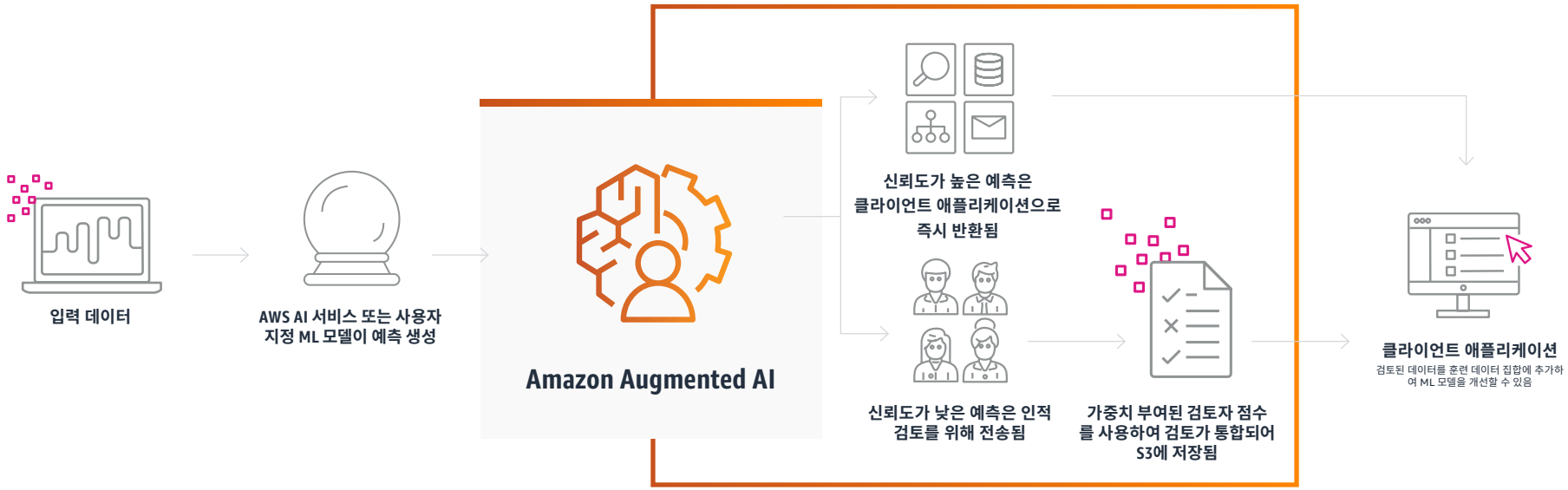
Amazon Augmented AI(Amazon A2I)를 사용하면 ML 예측 결과를 사람이 검토하는 데 필요한 워크플로를 쉽게 구축할 수 있습니다. Amazon A2I에서는 모든 개발자가 검토 기능을 사용할 수 있으므로 인적 검토 시스템을 구축하거나 많은 검토 인력을 관리하는 것과 관련된 획일적이고 번거로운 작업을 수행할 필요가 없습니다.

많은 기계 학습 애플리케이션에서는 신뢰도가 낮은 예측을 사람이 검토하여 결과가 올바른지 확인해야 합니다. 예를 들어 스캔된 주택 담보 대출 신청서 양식에서 정보를 추출할 때 스캔 품질이 좋지 않거나 손글씨를 알아보기 힘든 경우에는 사람이 직접 검토해야 할 수 있습니다. 그러나 인적 검토 시스템을 구축하려면 복잡한 프로세스 또는 '워크플로'를 구현하고 검토 작업 및 결과를 관리할 사용자 지정 소프트웨어를 작성해야 하며, 많은 경우 대규모 검토자 그룹을 관리해야 하기 때문에 많은 시간과 비용이 소모됩니다.



Amazon A2I를 사용하면 기계 학습 애플리케이션에 대한 인적 검토를 손쉽게 구축하고 관리할 수 있습니다. Amazon A2I는 일반적인 기계 학습 사용 사례(예: 콘텐츠 조정 및 문서의 텍스트 추출)를 위한 기본적인 인적 검토 워크플로를 제공하며 Amazon Rekognition과 Amazon Textract의 예측을 사용하여 검토 작업을 간소화합니다. Amazon SageMaker 또는 다른 모든 도구에 구축된 ML 모델에 사용할 자체 워크플로를 생성할 수도 있습니다. Amazon A2I를 사용하면 모델에서 신뢰도가 높은 예측을 생성할 수 없는 경우 검토 인력을 투입하거나 모델의 예측에 대한 지속적인 감사를 수행할 수 있습니다.

- Amazon A2I를 데이터 집합 레이블에 통합할 수도 있습니다. 이 블로그는 Amazon A2I를 사용하여 신뢰도가 낮은 데이터를 검토하고 보강하는 방법을 자세히 설명합니다.



Amazon SageMaker Studio: 앞서 언급했듯이 SageMaker Studio는 모든 기계 학습 개발 단계를 수행할 수 있는 웹 기반의 단일 시각적 인터페이스를 제공하여 데이터 과학 팀의 생산성을 개선합니다. 노트북, 실험 관리, 자동 모델 생성, 디버깅을 비롯한 모든 기계 학습 개발 활동을 Studio에서 수행할 수 있습니다. 이 가이드에서는 데이터 관리의 다양한 단계에 제공되는 다양한 AWS 서비스를 다뤘습니다. 아래 표에서 볼 수 있듯이 Studio는 기본적으로 SageMaker의 데이터 관리 및 기계 학습 기능들을 통합합니다. Studio를 사용하면 팀이 워크로드의 여러 부분을 하나의 보기에 통합하여 더 쉽게 관리하고 더 빠르게 협업할 수 있습니다.

- 파일 워크로드/성능이 팀에 중요한 경우 Amazon FSx for Lustre가 Amazon SageMaker의 입력 데이터 원본이 될 수 있습니다. FSx for Lustre를 입력 데이터 원본으로 사용하는 경우, Amazon SageMaker ML 훈련 작업은 초기 S3 다운로드 단계를 제거하여 가속화됩니다. S3에서 모든 기계 학습 훈련 데이터 집합을 다운로드할 필요 없이 S3 버킷과 연결된 FSx for Lustre 파일 시스템이 생성되는 즉시 SageMaker 작업이 시작될 수 있습니다. 데이터는 작업 처리를 위해 Amazon S3에서 필요에 따라 지연 로드됩니다. 또 하나의 이점은 동일한 데이터 집합에 대한 반복 작업에 공통된 객체를 반복적으로 다운로드할 필요가 없어 (S3 요청 비용 절감) TCO를 줄일 수 있다는 것입니다.

프로젝트	지속적 통합 및 지속적 전달(CI/CD)을 사용하여 모델 구축 및 배포 파이프라인을 자동화
Data Wrangler	기계 학습을 위한 데이터를 집계하고 준비
특성 스토어	기계 학습(ML) 특성을 저장, 업데이트, 검색, 공유하는 용도로 구축된 완전관리형 리포지토리
파이프라인	대규모 엔드 투 엔드 ML 워크플로를 생성, 자동화, 관리
실험 및 시도	기계 학습 실험의 구성, 추적, 비교 및 평가
모델 레지스트리	모델의 계보 및 메타데이터 추적
엔드포인트	예측 엔드포인트 추적

결론

오늘날의 기계 학습 환경에서 성공적인 모델을 구축하기 위해서는 데이터가 중요합니다. 이 가이드에서는 기계 학습 모델에 사용할 수 있는 데이터 원본 뿐만 아니라, 데이터 품질과 볼륨을 측정하는 방법을 다뤘습니다. 다음으로, 데이터를 관리하고 기계 학습을 위해 준비하는 데 필요한 여러 단계에 걸쳐 AWS 서비스와 기능이 어떤 도움이 될 수 있는지 논의했습니다.

데이터 관리 프로세스에 익숙해지고 기계 학습을 위한 데이터 사용을 모색하기 시작하는 단계라면, AWS의 기계 학습 페이지를 방문하여 기계 학습 서비스를 탐색하고 고객의 성공 사례를 확인하고 일반적인 사용 사례를 찾아보십시오. AWS의 사명은 분명합니다. 모든 개발자가 기계 학습을 사용할 있도록 하는 것입니다. 데이터를 이해하고 활용하는 데 기여하게 되어 기쁩니다.

AWS 계정 관리자 및 솔루션스 아키텍트에게 문의하시면 데이터 및 기계 학습 여정을 가속화하도록 도와드리겠습니다.

클라우드에서 기계 학습 애플리케이션을 구축, 배포하고 실행하는 데 필요한 무료 제품 및 서비스입니다.

시작하기 >>

참조

[1] 훈련 데이터 품질을 정의하고 측정하는 방법.
[2] 데이터 품질 평가.
[3] 데이터 품질을 측정하는 방법 - 데이터 품질을 평가하는 7가지 지표.
[4] 빅 데이터와 기계 학습을 위한 데이터 품질 고려 사항: 데이터 정리와 변환을 넘어.
[5] 기계 학습을 위한 자동화된 데이터 품질 관리를 목표로.
[6, 7] 분류 성능에 필요한 표본 크기 예측.
[8] 얼마나 큰 훈련 세트가 필요한가?
[9] 분류 모델의 표본 크기 계획.
[10] 데이터 집합 크기 vs 데이터 차원, 경험 법칙이 있는가?
[11] 얼마나 많은 훈련 데이터가 필요한가?
[12] 시계열 모델의 최소 관찰 횟수는 얼마여야 하는가?
[13] 경제 및 환경 문제에 적용한 개입 분석.
[14] 딥 러닝 모델을 훈련하는 데 필요한 최소 샘플 크기는 얼마인가(CNN)?
[15] 훈련 데이터가 충분한지 어떻게 알 수 있는가?
[16] 신경망 훈련에는 몇 개의 이미지가 필요한가?
[17] 텍스트 분석 입문: 문서 분류.
[18] 감정 분석에 실제로 얼마나 많은 텍스트가 필요한가?
[19] Depop 사례 연구.
[20] AxialHealthcare 사례 연구.
[21] 탐색 데이터 분석이란?
[22] 레이크 하우스 접근 방식이란?
[23] AWS 레이크 하우스에서 얻은 인사이트.
[24] AWS 기반의 데이터 레이크 스토리지.
[25] AWS 기반의 안전한 엔터프라이즈 기계 학습 플랫폼 구축.
[26] Amazon SageMaker 특성 저장소를 사용한 특성 생성, 저장 및 공유.
[27] HappyRefresh 사례 연구.
[28] Well-Architected 기계 학습 렌즈 - 데이터 준비.
[29] AWS Batch 기반의 딥 러닝.
[30] AI/ML의 데이터 처리 옵션.
[31] GURU사례 연구.
[32] AiCure 사례 연구.

aws
startups