데이터 현대화와 분석의 진화

박경규

데이터 트렌드

데이터 분석과 AI 결합, 개방형 아키텍처, 컴퓨팅과 데이터 분리등으로 확장성과 유용성이 대폭 증가

머신러닝의 부상(Rise)

- 단시간에 방대한 양의 데이터를 처리하고 분석가능
- 데이터 패턴을 인식하도록 훈련된 알고리즘으로 예측 수행

광범위하게 예측 분석 사용

- 예측 분석은 빅데이터의 가장 큰 이점 중 하나
- 고객반응 이해와 미래추세 예측에 예측 분석을 광범위하게 사용

스트리밍 분석의 진화

- 교통정보, 예약, 영상서비스 등 상당수 서비스는 실시간 정보 기반
- 대량의 데이터를 실시간으로 처리하고 분석하는 기술의 발전

셀프서비스 분석

- 시민데이터과학자(Citizen Data Scientist)의 출현
- 실무자는 셀프서비스 분석도구로 데이터를 직접 분석하여 활용

클라우드 스토리지 사용 증가

- 중요한 빅 데이터 추세 중 하나는 클라우드 마이그레이션 증가 추세
- 클라우드 전환함으로써 비용절감, 효율성, 보안 문제를 해결

데이터 레이크

- 모든 규모의 정형 및 비정형 데이터를 관리 , 탐색 , 공유 , 분석 할 수 있는 리포지토리
- 컴퓨팅에서 스토리지 분리

데이터 패브릭

- 하이브리드 멀티 클라우드에서 빅데이터 모범 사례를 표준화
- 일관된 기능을 제공하는 데이터 서비스의 프레임워크/컬렉션

데이터 품질/보안 강화

- 데이터 유출 증가로 보안에 막대한 투자 발생
- 고객과 브랜드 가치 유지를 위해 고객 민감 데이터 보호 강화 필요

데이터 분석 트렌드

데이터 센트릭 비즈니스 운영과 고객 경험 개선을 위해 주기적으로 방대한 양의 데이터를 캡처, 저장하고 분석
데이터 기반 의사 결정 선호도가 증가하고 있으며 데이터 분석 시장이 지속 성장하고 있음

분석의 보편화와 구성 가능한 분석 선호

셀프 서비스 분석 모델은 데이터 실무자가 추구하는 분야로
여러 분석 솔루션의 구성 요소를 융합한 비즈니스 애플리케이션을
구축할 것임

구성 가능한 분석 모델 구축에 대한 명확한 전략이 없으면
노력과 데이터의 중복으로 인해 더 많은 비용 초과가 발생할 수 있음

온디맨드 분석 플랫폼인 클라우드에서 셀프분석 수요를 충족하는
기능을 제공하지만, 아직까지는 고비용 발생

더 많은 비즈니스에서 AI를 운용

AI 및 ML 기술을 데이터 분석 및 비즈니스 인텔리전스(BI) 도구와
통합하거나 결합하여 복잡한 데이터 유형을 처리하고 대규모 비정형
데이터의 숨겨진 가치를 발견

대규모 언어 모델 애플리케이션(ChatGPT 등)이 분석 공간에 어떤
영향을 미칠지 관심을 가져야 함.(자연어에서 SQL 쿼리를 생성 등)

메타 데이터 기반 데이터 패브릭의 증가

메타데이터로 데이터 구조를 강화함으로써 분석가는 데이터를 더 깊이
있고 의미 있게 이해할 수 있음

서로 다른 시스템을 통합 및 자동화하고 AI/ML 기술을 활용 하여 방대한
데이터풀을 분석함에 따라 데이터 패브릭 개념이 도입됨

데이터 패브릭은 다른 시스템(온프레미스, 다중 클라우드, 소셜 미디어,
IoT 장치, 모바일 애플리케이션 등)의 데이터를 통합 처리하고 분석 하는
데 도움이 됨

적응형 분석과 실시간 의사 결정

AI 및 ML 기술 활용으로 분석이 상황에 따라 지속적으로 이루어짐에
따라 적응력도 높아질 것임

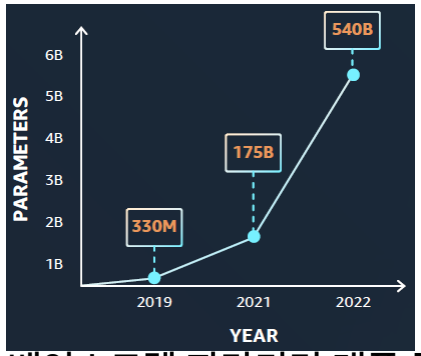
분석은 과거 데이터에만 초점을 맞추는 것이 아니라, 컨텍스트를 이해
하면서 데이터를 실시간으로 처리

적응형 분석의 핵심 이점은 매우 높은 정확도로 실시간 데이터를
기반으로 의사 결정을 내릴 수 있다는 것임

AI 트렌드

AI를 업무에 적용하고 대부분의 디바이스에 AI 기능이 탑재되는 등 AI 활용이 보편화됨
DALL-E2, ChatGPT, AlphaCode 등 생성형 AI 기술의 진격으로 사회와 산업 전반에 격변 발생 예상

AI 모델 정교화



베이스모델 파라미터 대폭 증가

생성형 AI의 진격

ChatGPT

Stable Diffusion

머신러닝의산업화

머신러닝 유즈케이스

고객 경험 증진

보다 빠르고
더 나은 의사결정

비즈니스 운영 개선

데이터 활용사례

데이터를 활용하지 않는 곳은 별로 없지만, 데이터를 성공적으로 활용하는 곳은 많지 않음

‘소비자 입체 분석’ 아마존

- 20년간 수집한 광범위한 데이터 보유
- 다양한 데이터 분석 기술로 개인화 추천
- 온라인과 오프라인 데이터를 모두 가지고 입체적으로 소비자를 분석

‘데이터 비즈니스의 신’ 스타벅스

- 개인화된 프로모션부터 신제품 개발에 데이터 분석 적극 활용
- 신규 매장 장소 선정을 데이터 기반으로 결정
- 기계 고장 및 유지보수 필요성을 미리 예측

‘취향저격 콘텐츠 추천’ 넷플릭스

- 정교한 빅데이터 기술을 이용하여 만든 추천 알고리즘
- 영화별로 수많은 메타데이터를 태그하여 다양한 마이크로 장르를 카테고리화
- 이용자들의 콘텐츠 시청 빅데이터를 분석하고 이를 바탕으로 고객 기호에 맞는 콘텐츠 직접 제작

‘유행 트렌드 즉시 반영’ ZARA

- 유행하는 패션 트렌드를 즉시에 반영하여 단기간 동안 다품종 소량 생산
- 데이터 분석으로 상품 수요를 예측하고 매장별 적정 재고를 산출
- 상품별 가격 결정과 운송 계획까지 실시간 데이터 분석에 의존

데이터 현대화

방대한 데이터를 확보해도 성과를 내지 못하는 이유는 데이터의 속성을 이해하지 못해
실제로 필요한 데이터를 보유하지 못했거나, 데이터 현대화의 새로운 환경이 필요하기 때문

데이터의 함정

- ✓ 데이터가 비즈니스 전략과 일치하지 않는 데이터
- ✓ 데이터 사일로로 인한 데이터 접근·활용성 부족
- ✓ 데이터 분석의 깊이, 철저함, 정확도가 결여된 약한 분석

데이터 현대화

- ✓ 현장에서 자산을 캡처하기 위한 실시간 데이터 프로세싱
- ✓ 데이터와 AI 민주화로 데이터 사용 극대화
- ✓ 고비용인 데이터 이동과 ETL을 최소화하고 분산 프로세싱

데이터 현대화 절차

마이그레이션
디지털 트랜스포메이션과 마이그레이션으로 레거시 기술에 대한 의존성 제거

데이터 & 앱 현대화
고급 분석과 AI 활용이 가능하도록 데이터와 앱 현대화

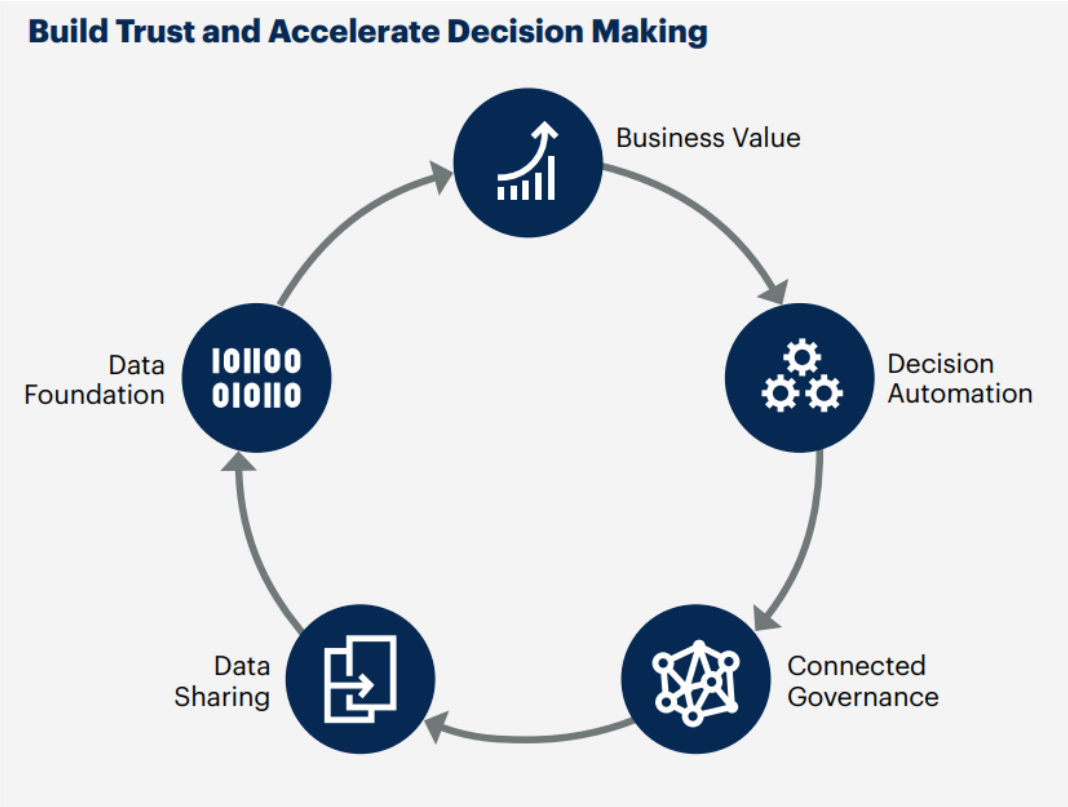
분석 현대화
360도 고객뷰를 구축하고 최신 분석 기술로 인사이트 발굴

AI로 혁신
AI 활용과 워크플로우 효율화로 자동화를 추진하고 높은 생산성 지원

민첩한 데이터 분석

민첩한 데이터와 분석은 현재 과제를 해결하고 미래의 기회를 다루는 핵심 원동력
데이터와 분석 이니셔티브를 측정 가능한 비즈니스 목표에 연결하고, 자동화를 통해 반복 가능한 프로세스를 생성해야 함

데이터 분석 전략



출처 : 가트너

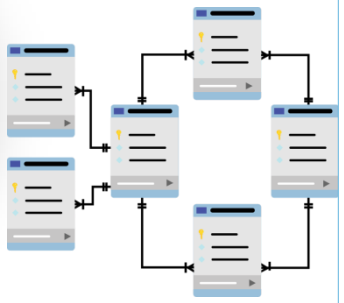
민첩한 데이터 분석

- 데이터 분석/활용 역량이 기업의 '비즈니스 핵심 역량'
- 데이터로부터 고객 경험을 차별화할 수 있는 상품과 서비스를 발굴
- 시장의 반응을 통해 데이터 분석의 실패와 성공으로부터 상품과 서비스를 수정하고 강화
- 고객으로부터 생성된 다양한 고객 경험 데이터는 고객 경험을 더욱 차별화하는 인사이트로 이어지는 선순환 구조를 구축

데이터 분석 기법

빅데이터 분석이 각광받는 이유는 방대한 양의 데이터 수집하고 처리할 수 있게 되었고,
이전에 처리하기 까다로웠던 비정형데이터 분석을 할 수 있게 되었고, 의미 있는 정보를 발굴할 수 있기 때문

정형데이터



정형데이터
비정형 데이터
반정형 데이터

분석기법

- 유형 분석
- 회귀 분석
- 시계열 분석
- 머신러닝 분석
- 유전 알고리즘
- 연관 규칙 학습
- 딥러닝 분석
- 감정 분석
- 군집 분석
- 그래프 분석

분석 활용

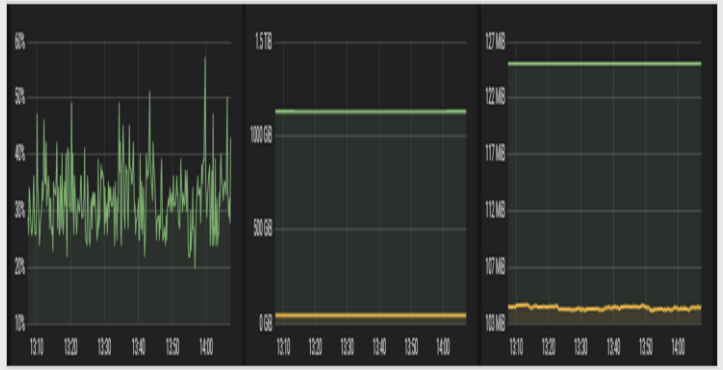
신선식품 산업에서의 **고객수요예측**
이커머스에서의 고객구매패턴에 따른 **상품 큐레이션**
데이터에 기반한 **공급망 관리**및 스케줄링
A/B테스팅을 통한 **사용자 경험 개선**

실시간으로 사용자 특성에 따라 광고를 배정하는 **애드테크(Adtech)**
사용자 유입 및 앱/웹 내 활동내역을 추적해주는 **마케팅 어트리뷰션 서비스**
사내 데이터 파이프라인 구축 및 데이터 활용을 도와주는 **데이터컨설팅**
데이터저장, 축적, 분석플랫폼을 서비스화 해서 제공하는 **클라우드 솔루션**

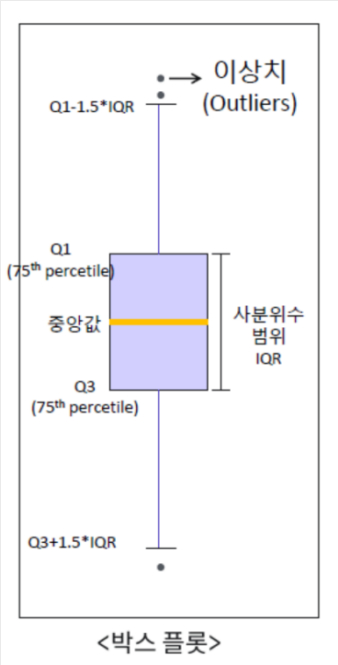
탐색적 데이터 분석

데이터의 분포와 값을 다양한 각도에서 관찰하고 다양한 기준에서 데이터를 살펴보는 과정을 통해
미처 발견하지 못한 다양한 패턴을 발견하고 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 추가

탐색적 데이터 분석



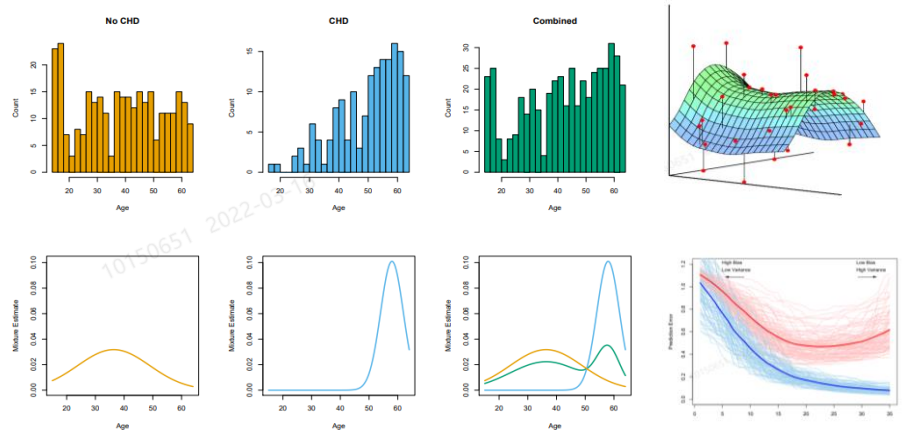
이상치 탐지



<박스 플롯>

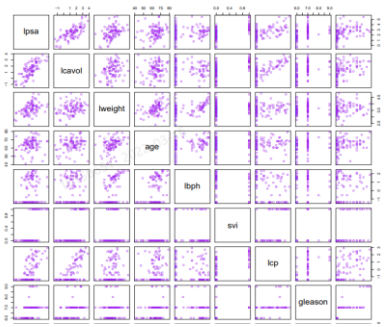
데이터 시각화

패턴 파악: 경향성

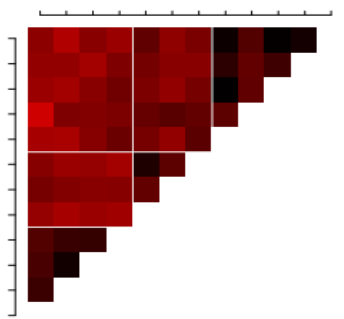


패턴 파악: 변수간 상관관계

1. Pair Plot



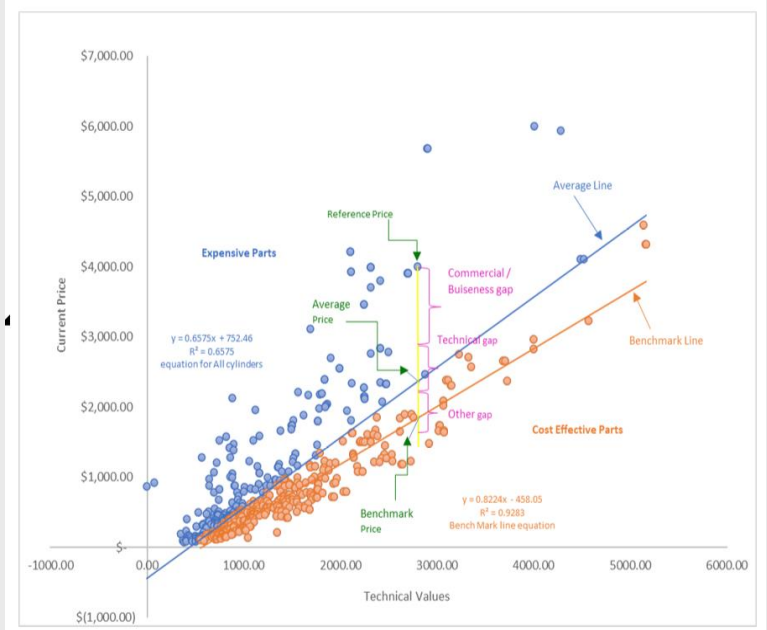
2. Correlation Plot



통계 분석

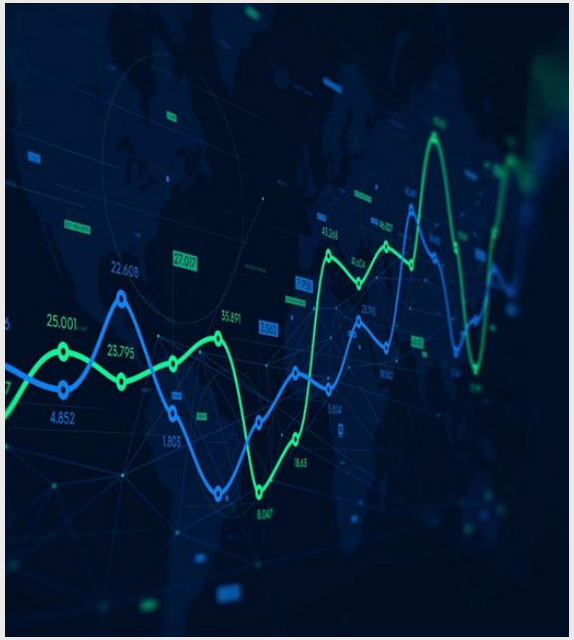
통계 기법으로 독립변수들과 종속변수간의 관계를 추론하고 분석 데이터를 기반으로 예측

회귀 분석



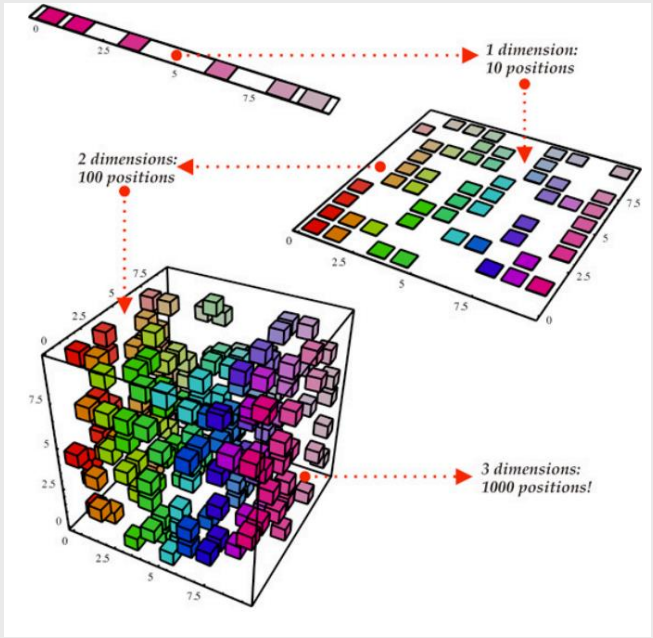
주어진 독립변수들과 종속변수간의
관계를 추론하는 통계 기법

시계열 분석



시간에 흐름에 따라 기록된 데이터를 바탕으로
미래의 변화에 대한 추세를 분석하는 방법

주성분 분석

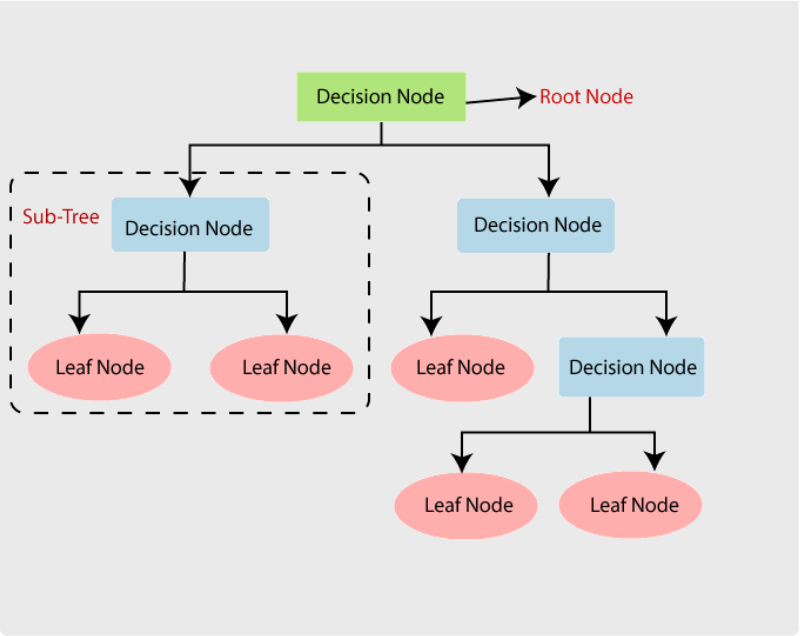


특성이 많은 데이터셋의
차원을 줄이는 차원 축소 기법

머신러닝 분석

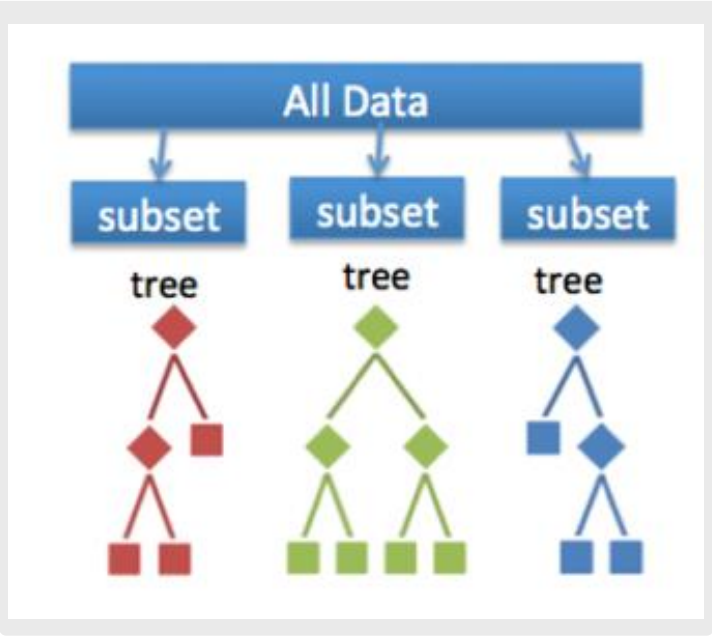
머신러닝은 컴퓨터 알고리즘이 데이터를 학습하여 입력데이터와 출력간의 관계를 찾음

결정트리 알고리즘



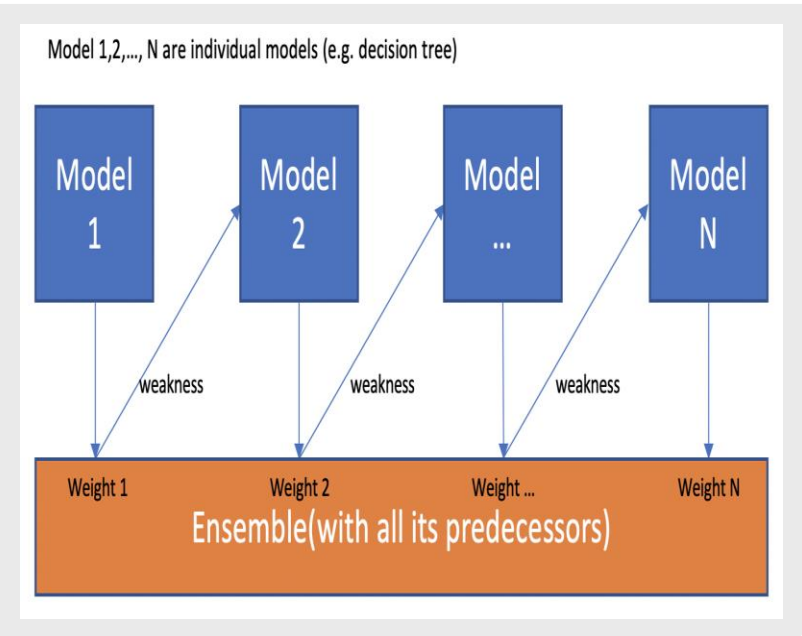
트리의 각 분기점(node)에 데이터셋의 특성을
하나씩 위치시키고, 각 분기점(node) 에서
임의의 조건식으로 가지를 나누면서 데이터를 구분

랜덤포레스트 알고리즘



훈련 데이터셋의 서브셋을
무작위로 구성하여
각기 다르게 학습시키는 방법

부스팅 알고리즘

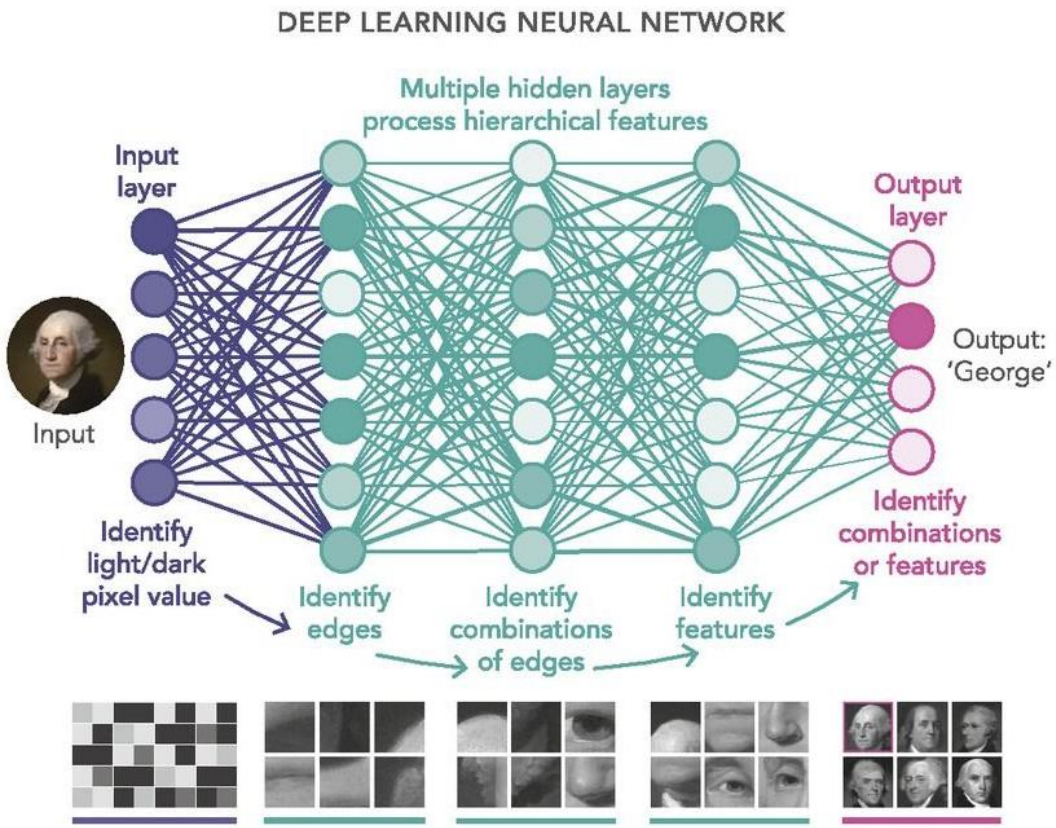


여러 개의 모델을 순차적으로 학습.
잘못 예측한 데이터에 대한 예측 오차를
줄일 수 있는 방향으로 모델을 계속 업데이트

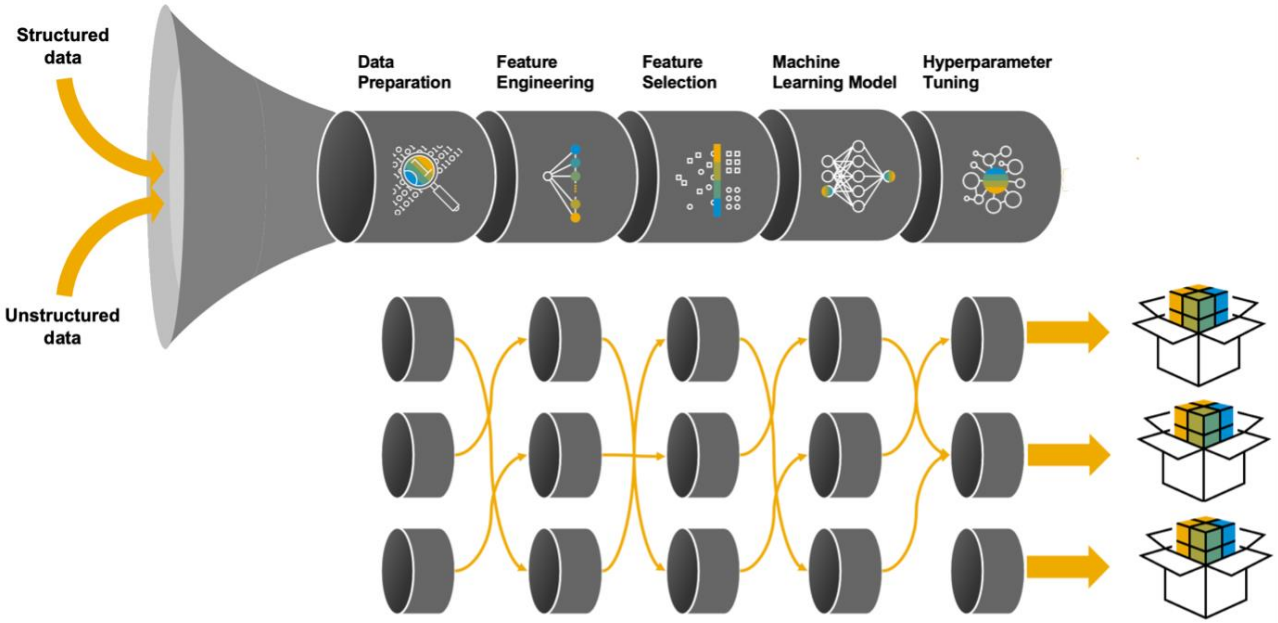
딥러닝 분석

인공뉴런과 링크들로 구성된 인공신경망이 복잡한 대용량 데이터셋의 패턴을 학습
비즈니스 사용자가 AutoML을 사용하여 고품질 모델을 자동으로 학습시킬 수 있음

인공 신경망

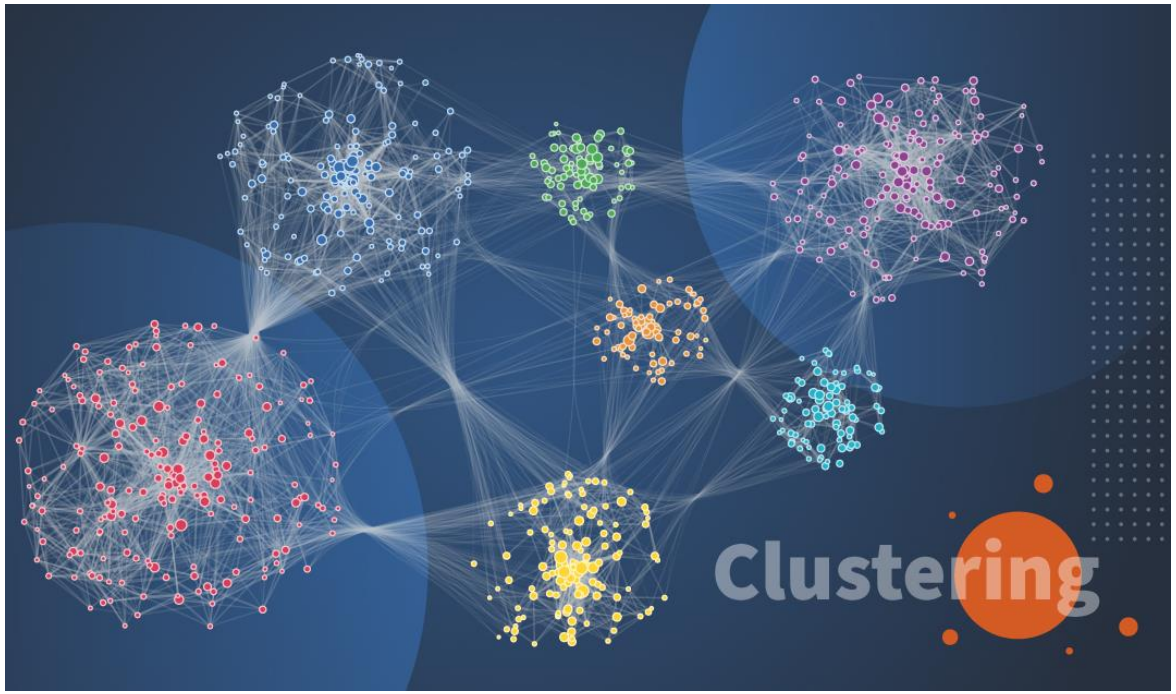


AutoML



군집 분석

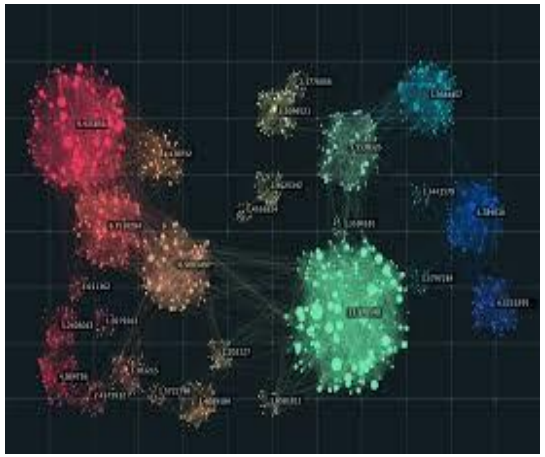
개체 간 유사성 을 측정하여 유사성에 따라 그룹화하는 방법으로 고차원의 데이터를 저차원으로 군집화



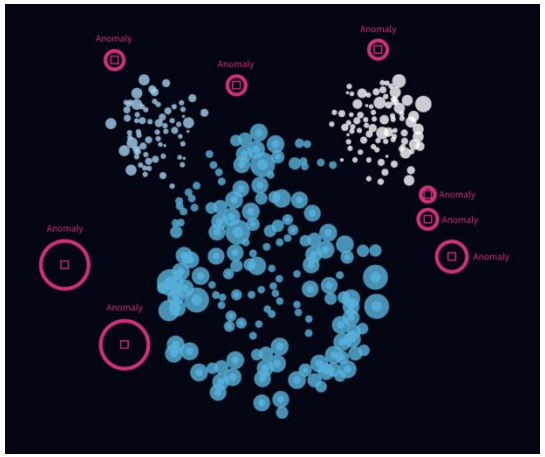
탐색적 데이터 분석

고객 세그멘테이션

데이터 시각화



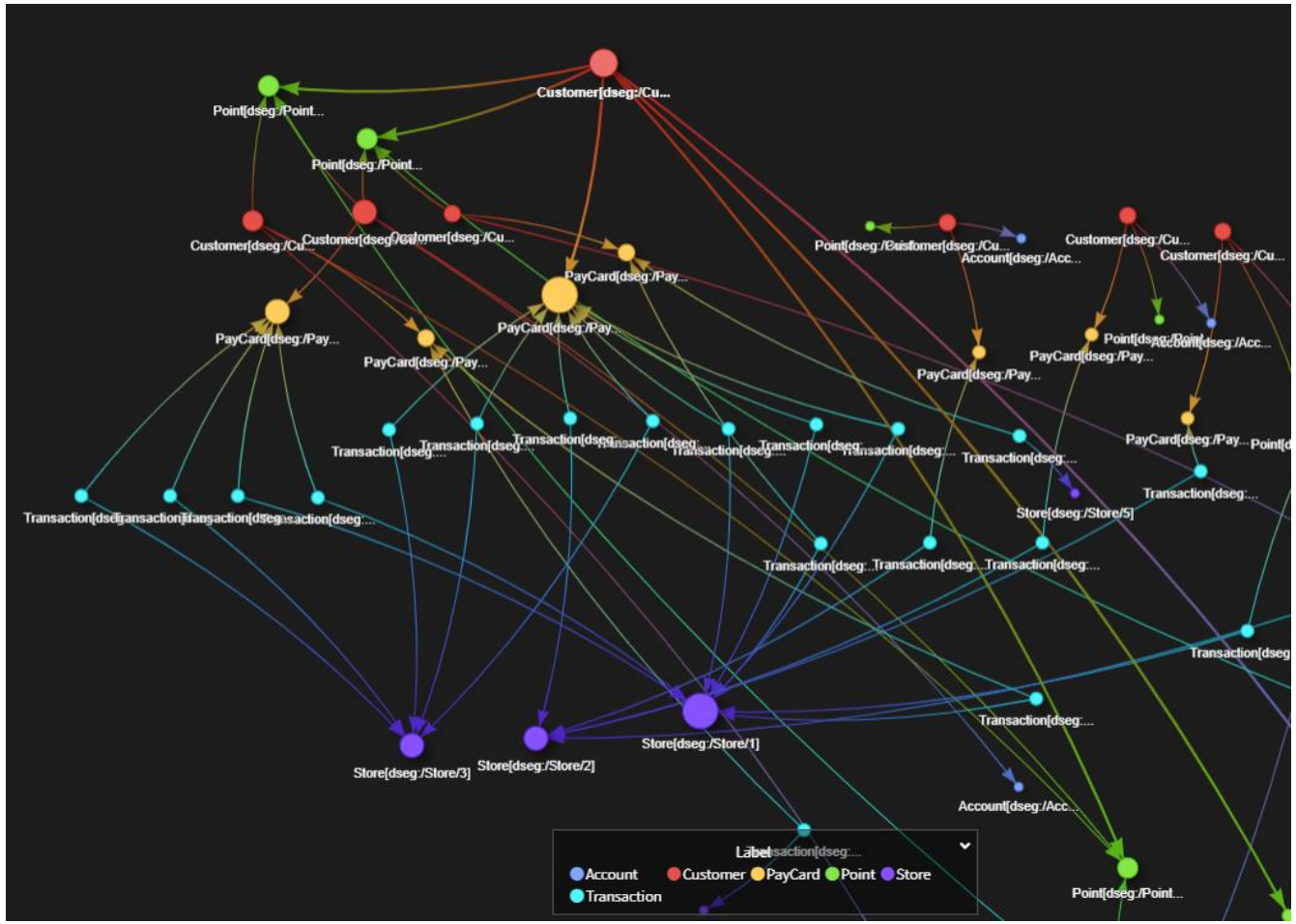
이상 탐지



그래프 분석

데이터 포인트를 노드로, 관계를 간선으로 사용하여 데이터를 그래프 형식으로 분석하는 방법
대량의 데이터 간 연결 관계를 손쉽게 분석할 수 있어 개인화 추천, 소셜 네트워크 분석, 경로 분석에 활용

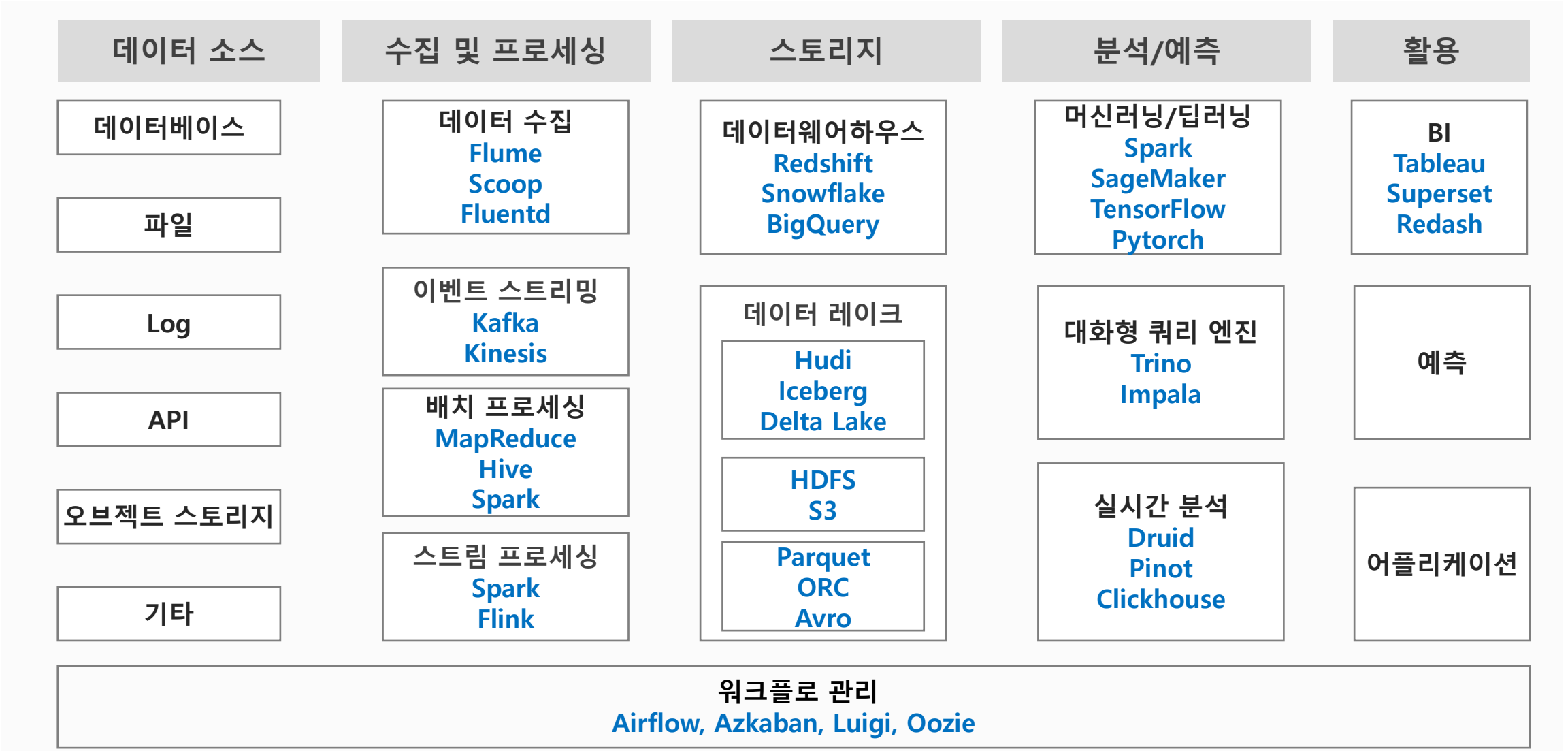
그래프 데이터



소셜 네트워크 분석

데이터 플랫폼 아키텍처

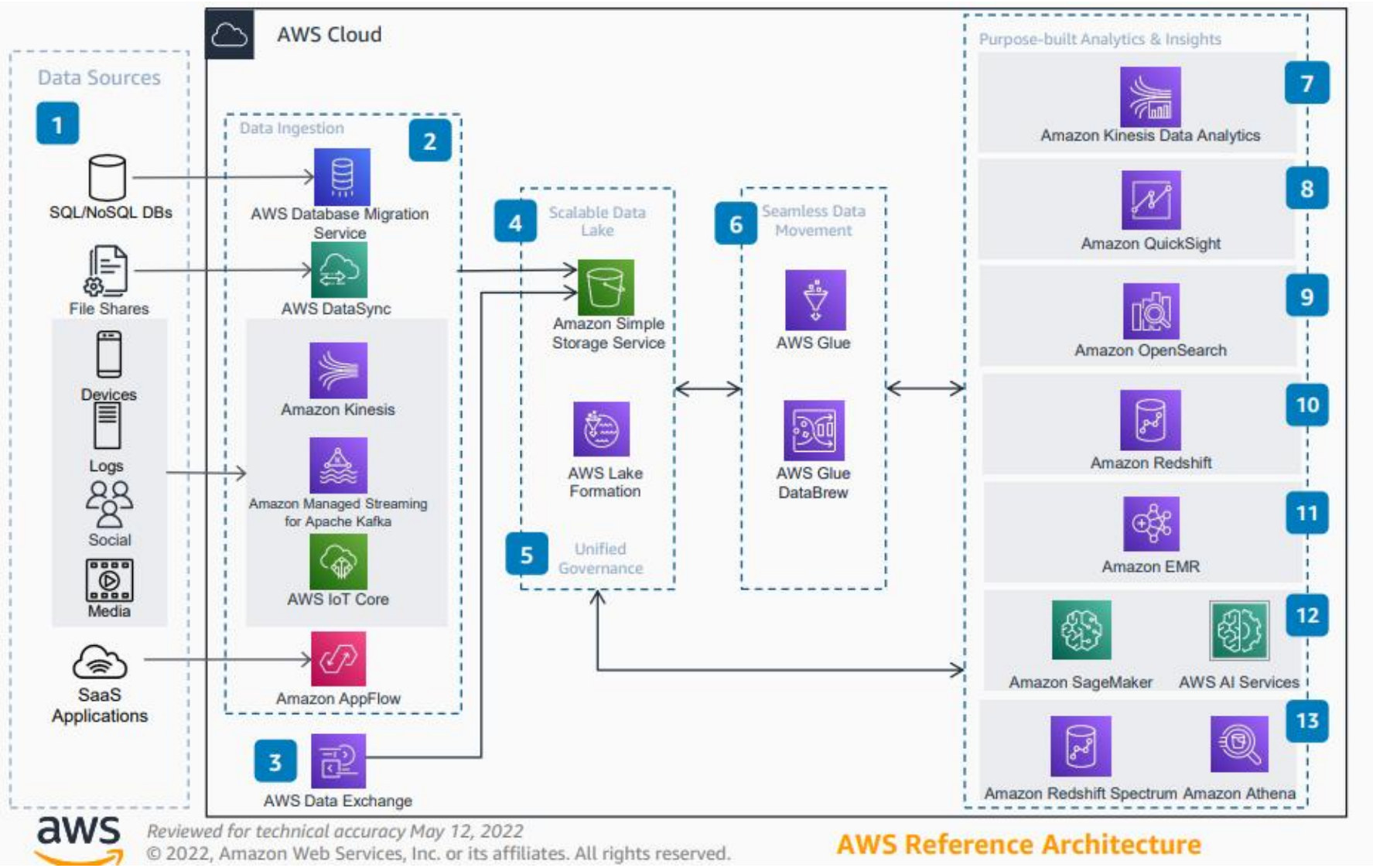
규모, 다양성, 속도 모든 측면에서 데이터 프로세싱이 가능하도록 데이터웨어하우스와 데이터레이크를 결합하여
어떤 유형의 데이터이든 비용 효과적인 방법으로 수집, 저장, 처리해서 분석/활용할 수 있는 아키텍처 구성 필요



데이터 전문가 역량

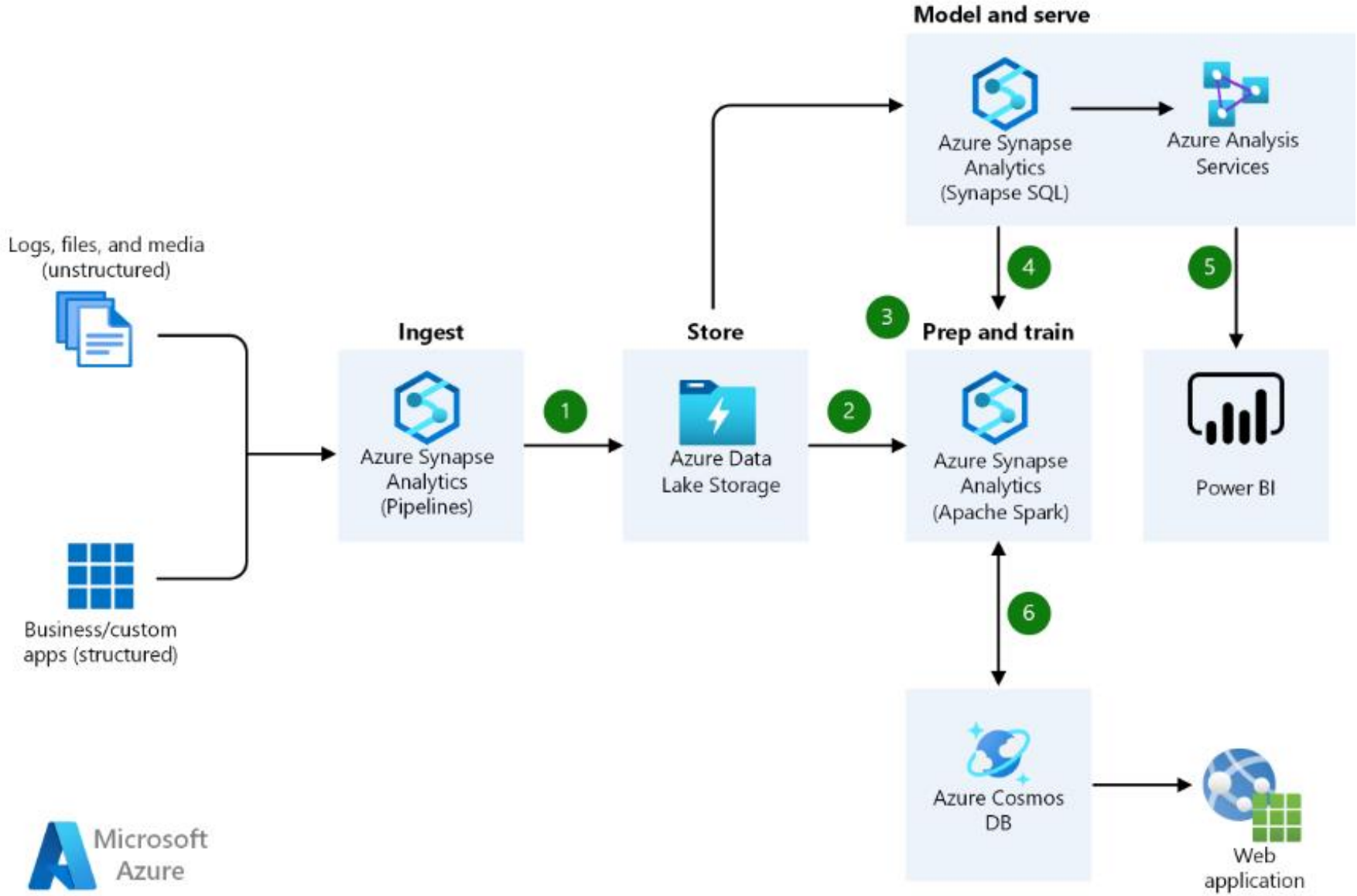
	데이터 분석가	데이터 엔지니어	데이터 사이언티스트
스킬셋	▪ SQL/데이터베이스 기본 지식 ▪ 데이터 웨어하우징 ▪ 스크립팅 및 통계 기술 ▪ 데이터 레포팅 & 시각화 ▪ 기본 프로그래밍 지식	▪ SQL/데이터베이스 심층 지식 ▪ 데이터 웨어하우징 및 ETL ▪ 하둡/스파크 기반 분석 ▪ 머신러닝 & 딥러닝 개념 ▪ 스크립팅 및 레포팅/시각화 ▪ 고급 프로그래밍 지식	▪ 통계 및 분석 기술 ▪ 데이터 마이닝 ▪ 하둡/스파크 기반 분석 ▪ 머신러닝 & 딥러닝 원리 ▪ 데이터 최적화 ▪ 의사결정 및 소프트 기술 ▪ 심층 프로그래밍 지식(Python/R)
역할과 책임	▪ 데이터 수집 및 전처리 ▪ 레포팅/시각화를 통한 데이터 표현 ▪ 통계 분석 및 데이터 해석 담당 ▪ 데이터 수집 및 유지 관리 보장 ▪ 통계 효율성 및 품질 최적화	▪ 아키텍처 개발, 테스트 및 유지 관리 ▪ 통계 모델 및 AI/ML 모델 배포 ▪ 다양한 ETL 작업을 위한 파이프라인 구축 ▪ 데이터 정확성과 유연성 보장	▪ 운영 모델 개발 담당 ▪ 머신러닝과 딥러닝을 활용한 데이터 분석 및 최적화 수행 ▪ 데이터 분석을 위한 전략 플래닝에 참여 ▪ 데이터 통합 및 ad-hoc 분석 수행 ▪ 이해관계자와 고객 사이의 gap 메우기

퍼블릭 클라우드 데이터플랫폼 아키텍처 – AWS



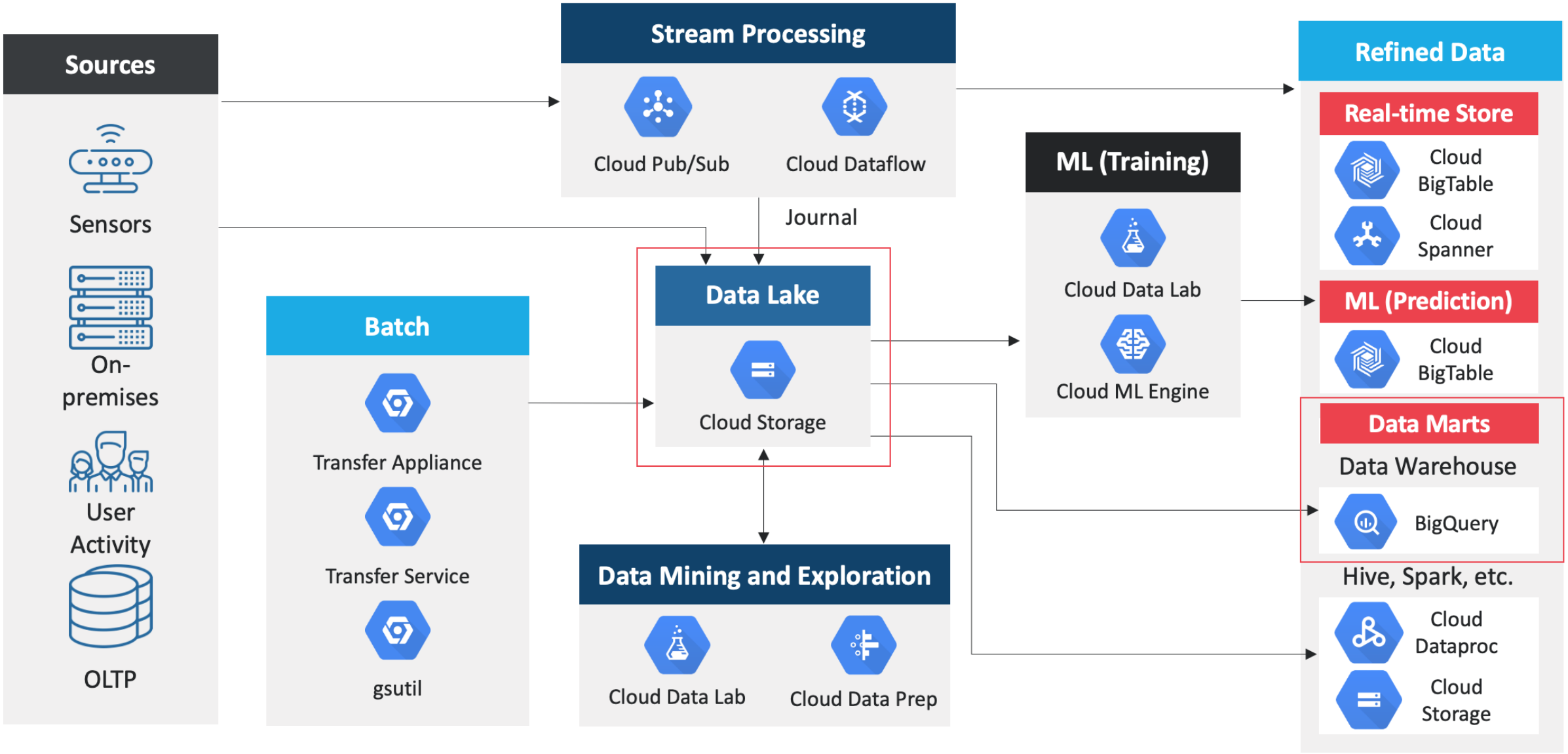
https://d1.awsstatic.com/architecture-diagrams/ArchitectureDiagrams/modern-data-analytics-using-lake-house-ra.pdf?anld_da6

퍼블릭 클라우드 데이터플랫폼 아키텍처 - Azure



https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/advanced-analytics-on-big-data

퍼블릭 클라우드 데이터플랫폼 아키텍처 - GCP



https://blog.miraclesoft.com/data-foundation-with-modernized-data-lake-data-warehouse/

Thank you

kgpark88@gmail.com