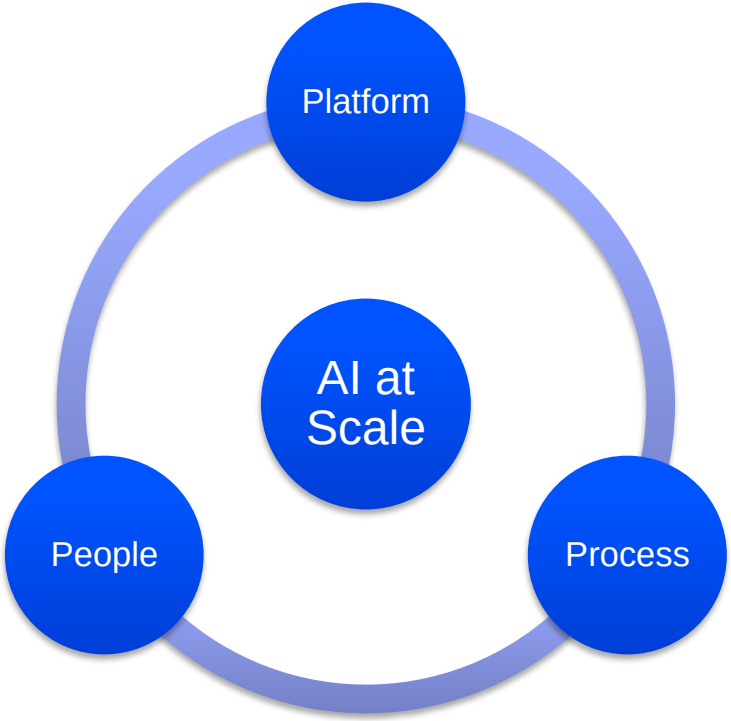
데이터 통합 관리 플랫폼으로 만나는 AI

Cloud Pak for Data



2021, Apr
Data & AI
JeongKwon Lee (jkwonl@kr.ibm.com)

시장 동향

*CIO는 **인프라구조에 대한 현대화** 필요하다고 생각합니다.*
cannot be achieved by lifting and shifting applications…

인프라구조 현대화를 시도하는 이유

Source: McKinsey expert interviews (N=52)

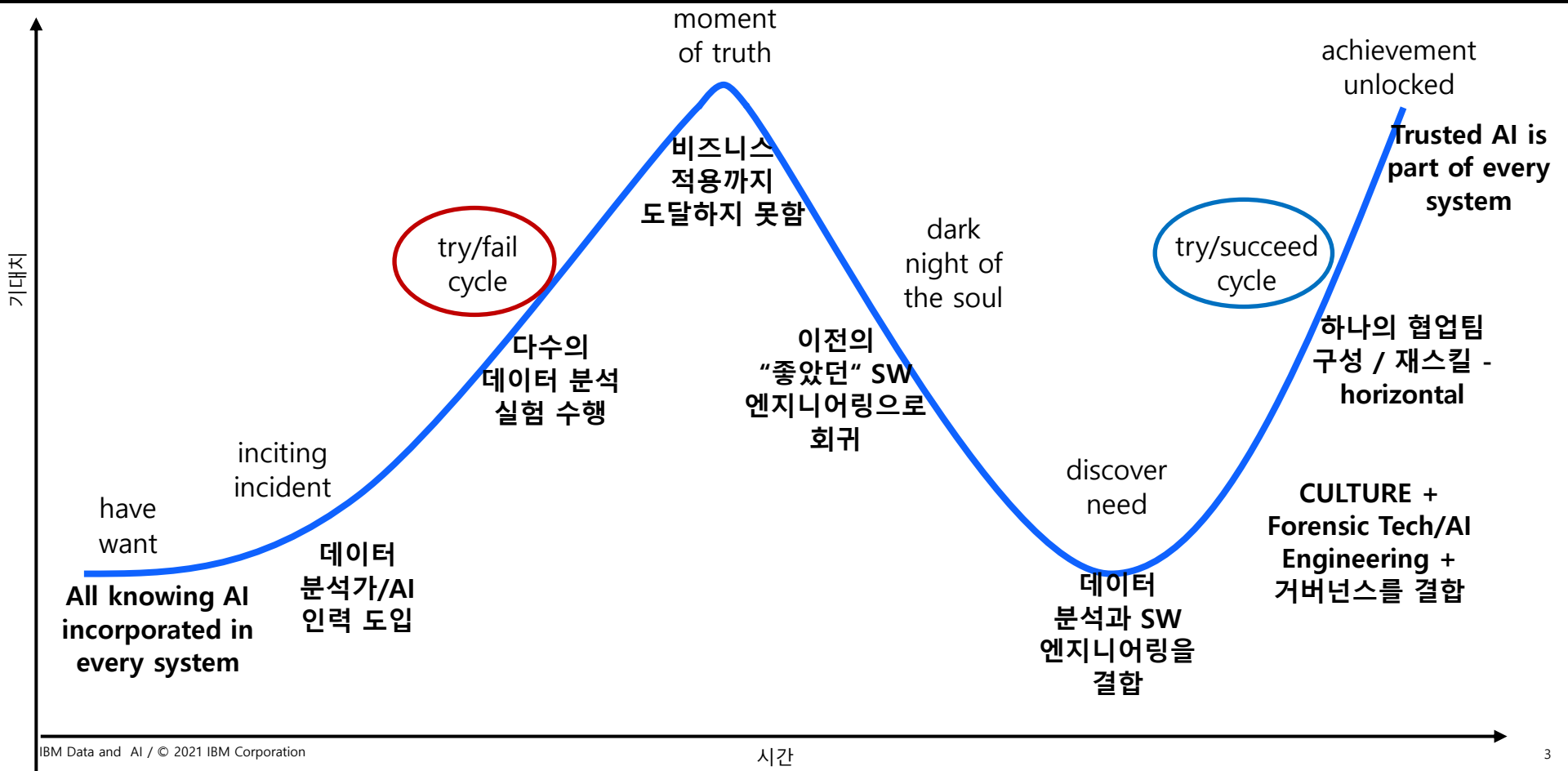
*CEO는 **디지털 혁신을** 최 우선 순위로 여기고 있습니다.*
*to jump start **growth**, **speed** time to market, and foster **innovation**…*

84%
의 글로벌 리더들은 **AI의 확산** 없이는 성장 목표를 달성할 수 없다고 말하고 있습니다.

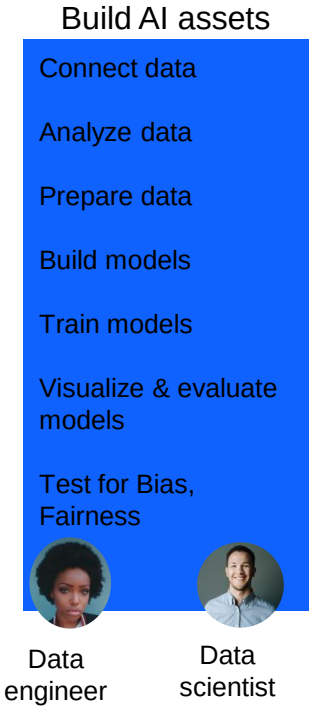
75%
의 글로벌 리더들은 **AI의 활용** 없이는 5년 내에 비즈니스 리스크에 직면한다고 생각하고 있습니다.

Source: Accenture, "AI – Built to Scale" Report (October 2019)

A typical AI story arc

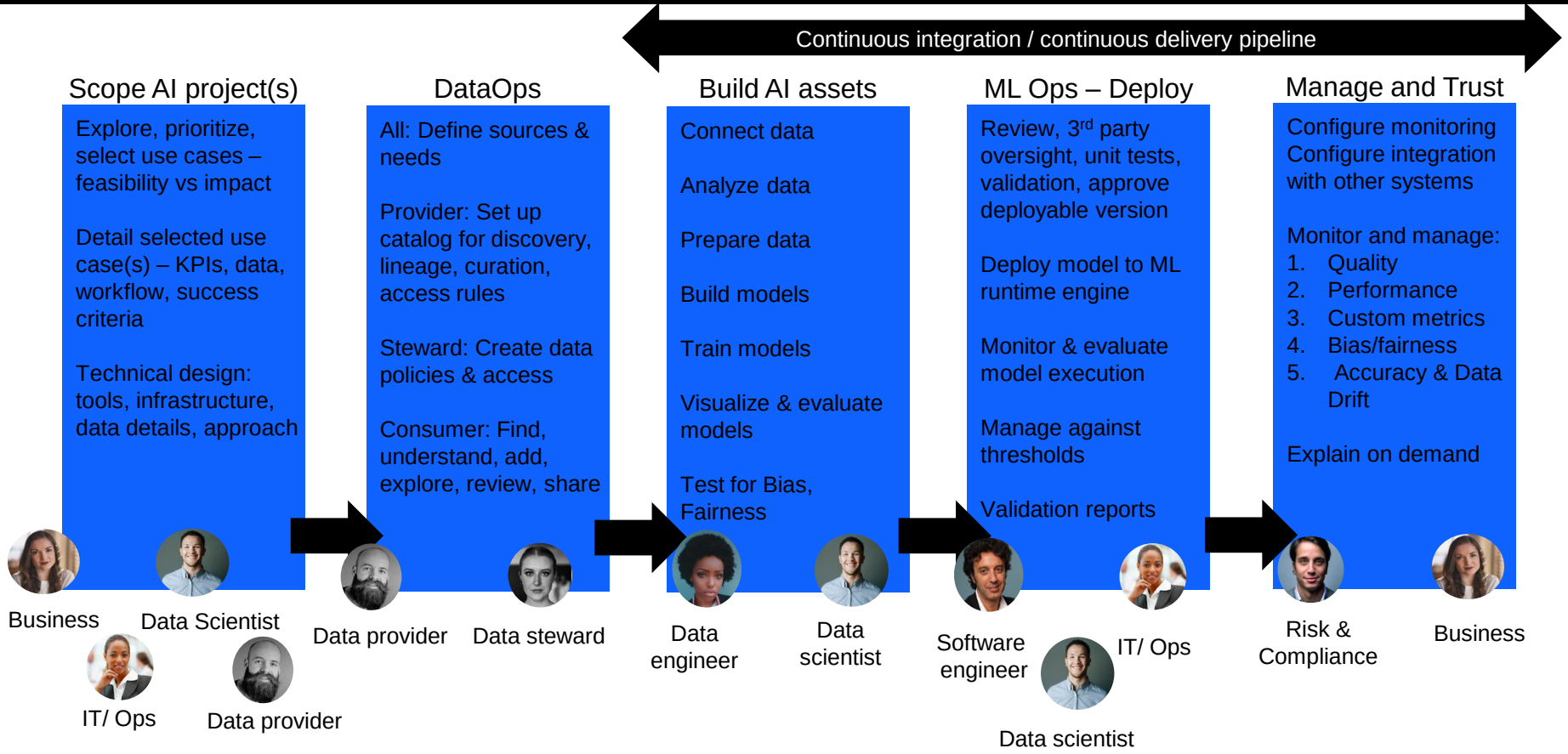


초기 단계의 AI/ML model 생성은 다음과 같이 보입니다...



But this is not the complete AI lifecycle!

AI 운영을 위한 단계들


Continuous integration / continuous delivery pipeline
Scope AI project(s)
Explore, prioritize, select use cases – feasibility vs impact
Detail selected use case(s) – KPIs, data, workflow, success criteria
Technical design: tools, infrastructure, data details, approach
DataOps
All: Define sources & needs
Provider: Set up catalog for discovery, lineage, curation, access rules
Steward: Create data policies & access
Consumer: Find, understand, add, explore, review, share
Build AI assets
Connect data
Analyze data
Prepare data
Build models
Train models
Visualize & evaluate models
Test for Bias, Fairness
ML Ops – Deploy
Review, 3rd party oversight, unit tests, validation, approve deployable version
Deploy model to ML runtime engine
Monitor & evaluate model execution
Manage against thresholds
Validation reports
Manage and Trust
Configure monitoring
Configure integration with other systems
Monitor and manage:
1. Quality
2. Performance
3. Custom metrics
4. Bias/fairness
5. Accuracy & Data Drift
Explain on demand
Business
Data Scientist
IT/ Ops
Data provider
Data provider
Data steward
Data engineer
Data scientist
Software engineer
IT/ Ops
Data scientist
Risk & Compliance
Business

There is no AI without IA

Artificial Intelligence

Information Architecture

" *No amount of AI algorithmic sophistication will overcome a lack of data [architecture]*

Data collection & preparation is the most time consuming and difficult part of AI "

MITSloan

Sources: 2018 MIT Sloan Report "Reshaping business with AI"

Data-AI **통합 플랫폼** 요구 사항

최근의 데이터와 AI는 상호 유기적인 관계로 빠르게 변화하는 비즈니스에 민첩하게 대응하기 위해서 다음과 같은 요구사항들을 필요로 하고 있습니다.

플랫폼 현대화

Cloud Native, Kubernetes

- 디지털 혁신을 위한 **빠른 provisioning 과 확장** 가능한 플랫폼
- 환경 변화 대응을 위한 **Hybrid Multi Cloud** 지원
- 다양한 요구 사항을 수렴할 수 있는 **컨테이너 기반 플랫폼**

데이터 관리

컴퓨팅과 스토리지 분리

- 정형/비정형을 포함한 분석에 필요한 **데이터의 Hub 역할을 위한 데이터 저장소** 구축
- 컴퓨팅과 스토리지 영역을 분리하여 업무 증가 및 워크로드에 따른 **유연한 확장성 제공**

데이터 거버넌스

Data Catalog, Self-Service

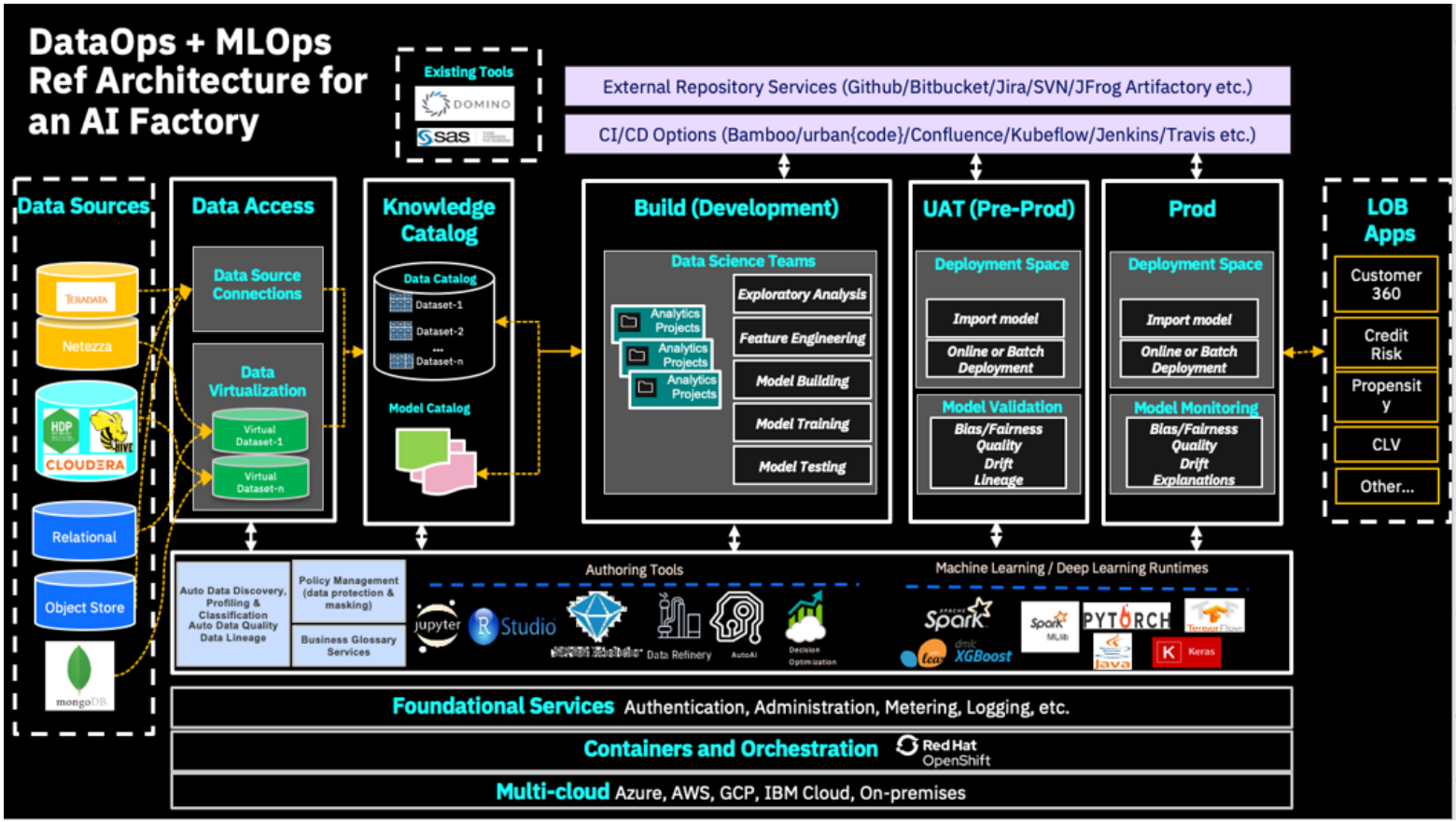
- 데이터 생산자와 소비자 간의 협업을 위한 **DataOps 지원**
- 비즈니스를 위한 데이터 준비를 위해 **거버넌스 체계**
- 셀프 서비스 분석을 위한 **데이터 디스커버리 환경**

AI 운영 관리

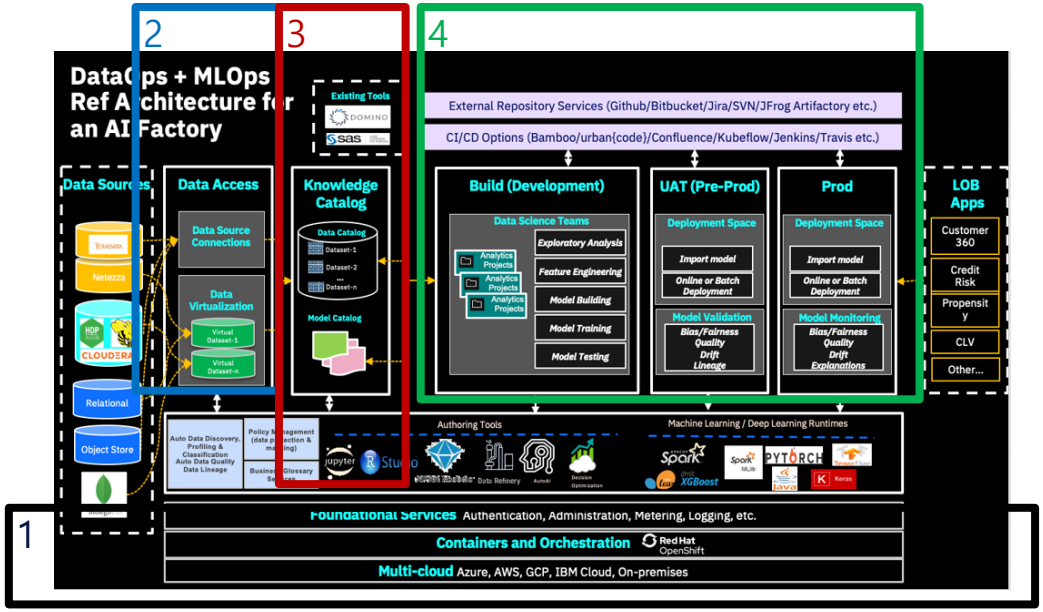
Operationalize Data Science & AI

- 분석 모델 생성, 관리, 거버닝을 포함하는 **AI 운영을 위한 기술** 도입
- 개방성과 최신 트렌드 반영을 위해 **오픈 AI 기술 반영**
- 분석 모델의 효율성을 높이기 위한 **Auto AI 기술** 활용

AI를 위한 Reference Architecture



AI를 위한 Reference Architecture > 아키텍처 구현 시 고려 사항



1 **플랫폼 현대화** : 컨테이너 기반 아키텍처를 통해 하이브리드 멀티 클라우드 지원하고 다양한 기능들을 통합하여 관리할 수 있는 통합 플랫폼 역량 필요

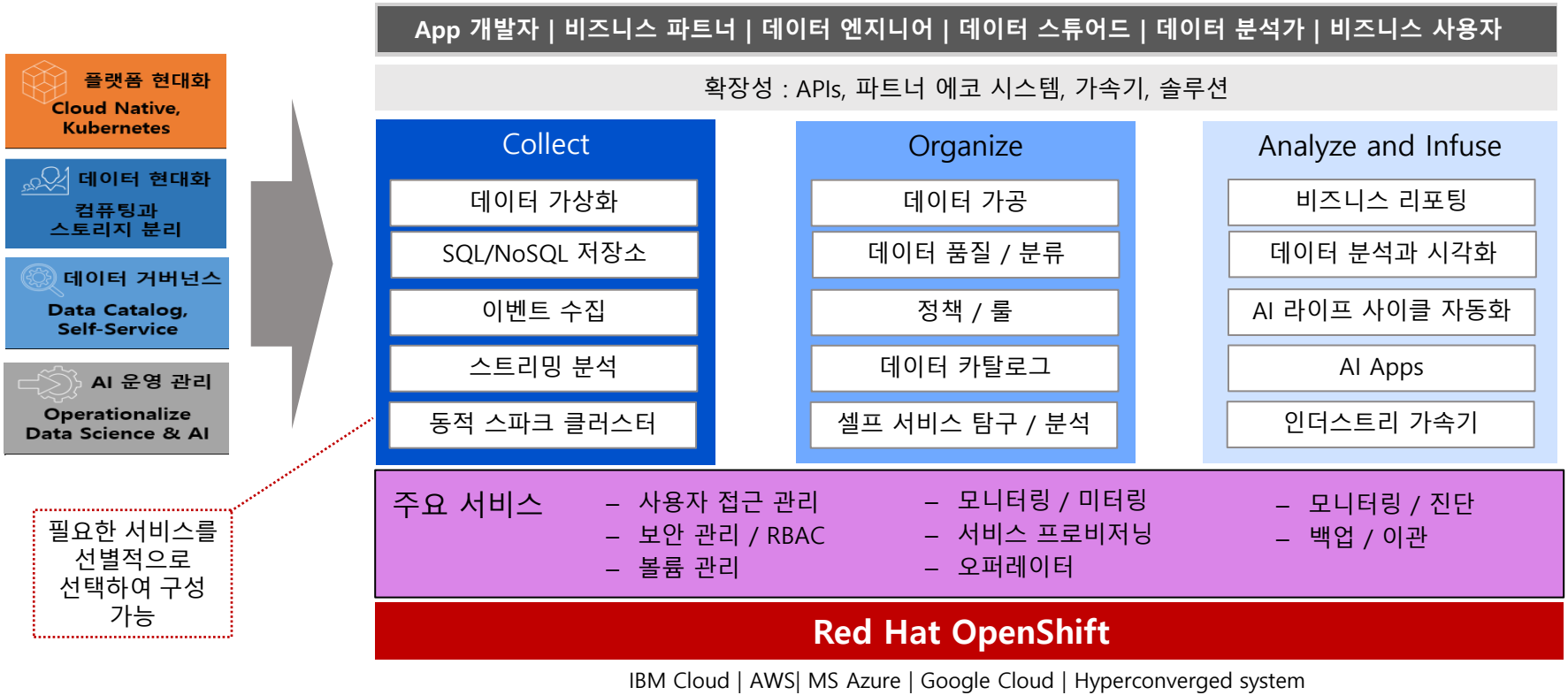
2 **데이터 현대화** : 컴퓨팅 영역과 스토리지 영역을 분리를 통해 유연성과 확장성을 확보 (고성능 데이터 마트를 위해서는 일체형 또는 on-premise 형태로 구축하는 것이 유리합니다.)

3 **데이터 거버넌스** : DataOps (메타 데이터 관리, 데이터 카탈로그, 데이터 리니지, 검색 등)를 통해 데이터의 품질 확보 및 거버넌스 체계 수립

4 **AI 운영 관리** : MLOps (ML/DL 라이프 사이클 관리 - 모델 생성, 배포, 운영, 모니터링)를 통해 AI 거버넌스 환경 구축

Cloud Pak for Data – **End-to-End Data-AI 플랫폼**

Cloud Pak for Data는 AI 플랫폼 구현에 필요한 여러 구성 요소들을 MSA 기반으로 구현하여 **필요한 서비스만 선별적으로 선택하여 구성**하는 것을 지원합니다.

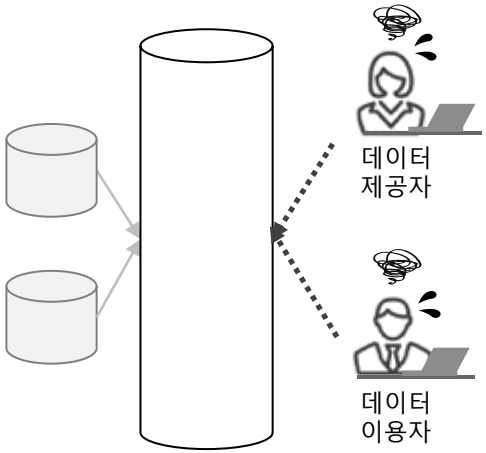


플랫폼 현대화 > 통합 플랫폼

지금까지의 분석 솔루션은 다양한 도구를 사용했기 때문에 데이터 제공자와 이용자는 별도의 노력과 시간이 필요했으나, IBM Cloud Pak for Data는 기존의 분석 솔루션 개념을 재정의하여 **단일 GUI에서 전체 분석 사이클을 지원하고 최적화합니다.**

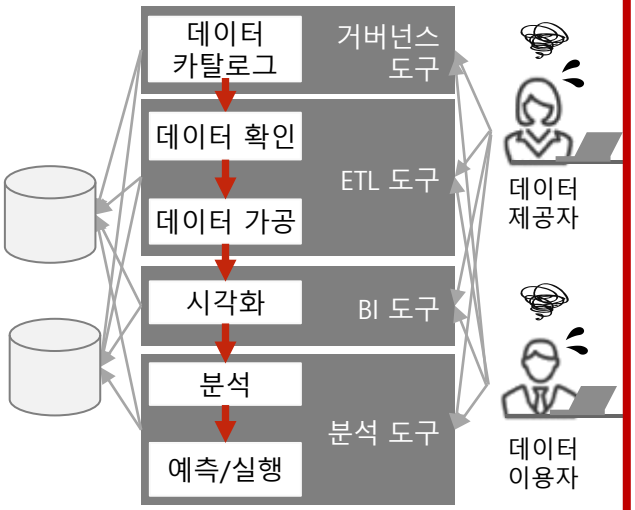
솔루션 도입 전

- 데이터의 전체를 복사하기 위한 대규모 기반 인프라를 조성
- 배치와 데이터 저장에 상당한 초기 비용이 소요
- 그럼에도 불구하고, 활용하기 힘듦



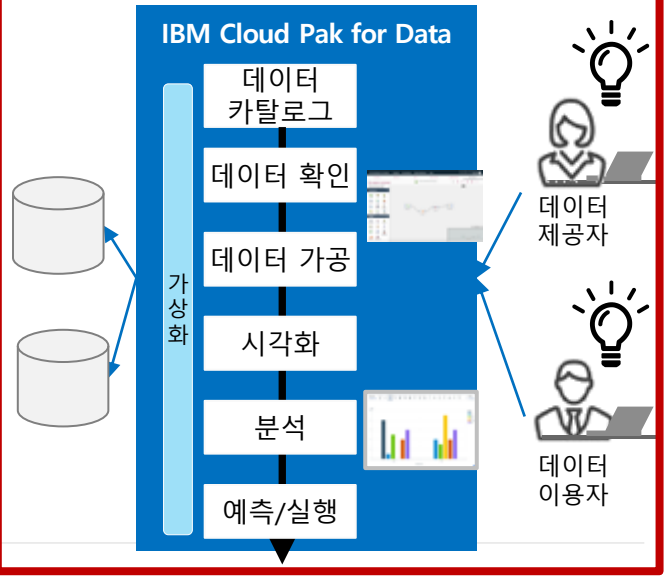
기존 개별 솔루션

- 다양한 도구를 적절하게 사용할 필요가 있음
- 각 도구의 학습이 필요



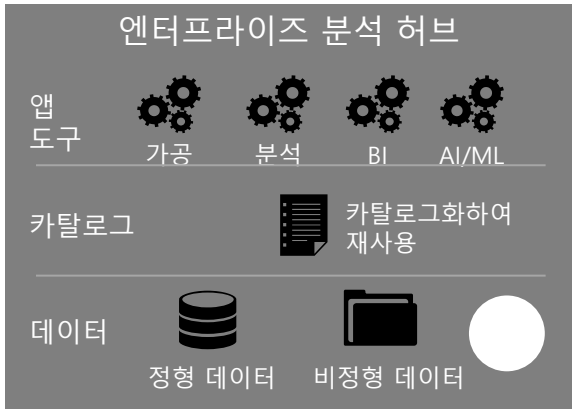
IBM Cloud Pak for Data

- 하나의 도구를 전체 사이클에서 사용
- 데이터 가공에서 분석까지 동일한 GUI
- 아이디어를 즉시 적용할 수 있는 장소에서 사용

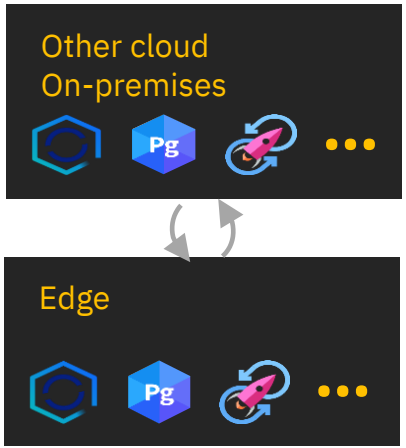


플랫폼 현대화 > anywhere, Distributed Analytics

IBM Cloud Pak for Data는 **특정 플랫폼 환경에 종속되지 않고**, 다양한 인프라에 데이터 자산을 활용할 수 있도록 멀티 클라우드 환경을 지원하고, 데이터 있는 곳에 분석이 수행되는 **Distributed Analytics을 지원하고 있습니다.**

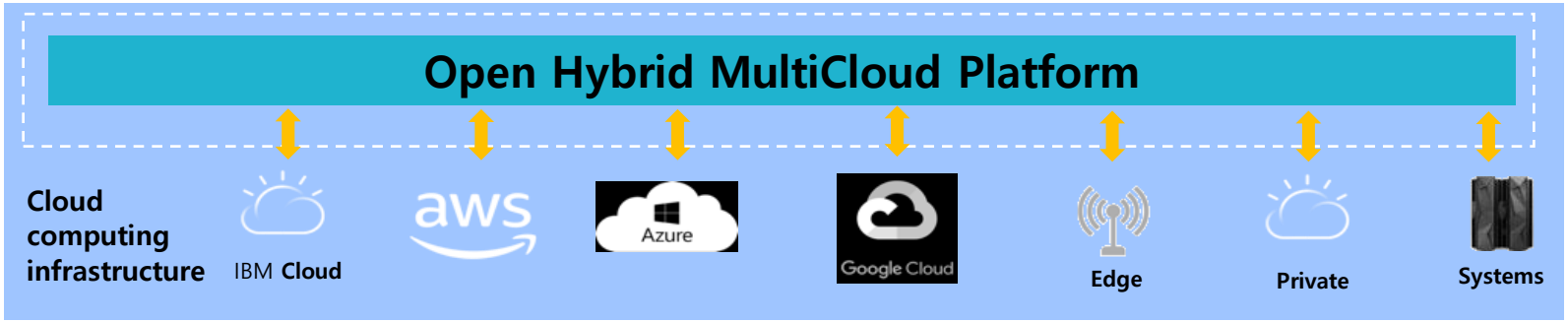


Distributed Analytics



IBM Cloud Satellite enables analytics consistency to siloed, disparate data sources

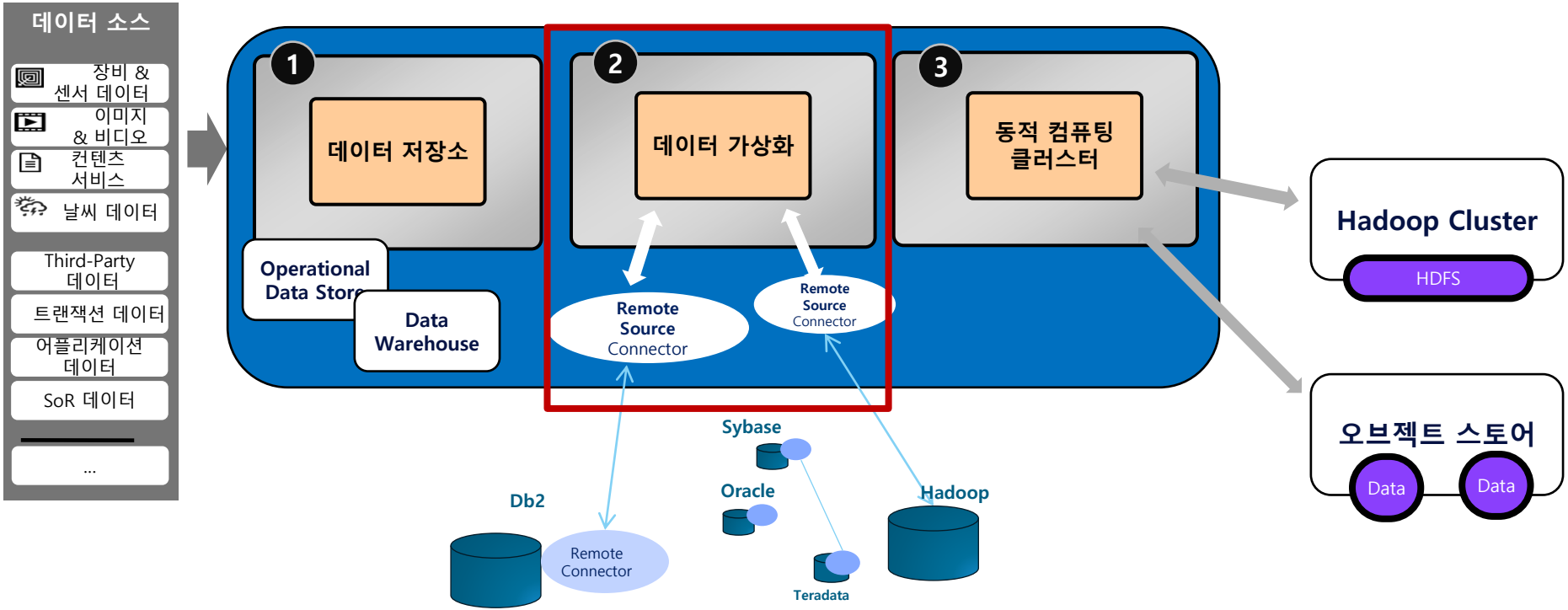
Streaming data – deploy local containerized, elastic analytics via IBM Cloud Satellite



데이터 현대화 > 다양한 데이터 처리 방안

다양한 데이터 유형과 소스 및 워크로드를 지원하기 위해서는

다양한 데이터 처리 방안 (1 : 데이터 저장소 구현, 2 : 데이터 가상화 구현, 3 : 동적 컴퓨팅 클러스터 환경)이 필요합니다.

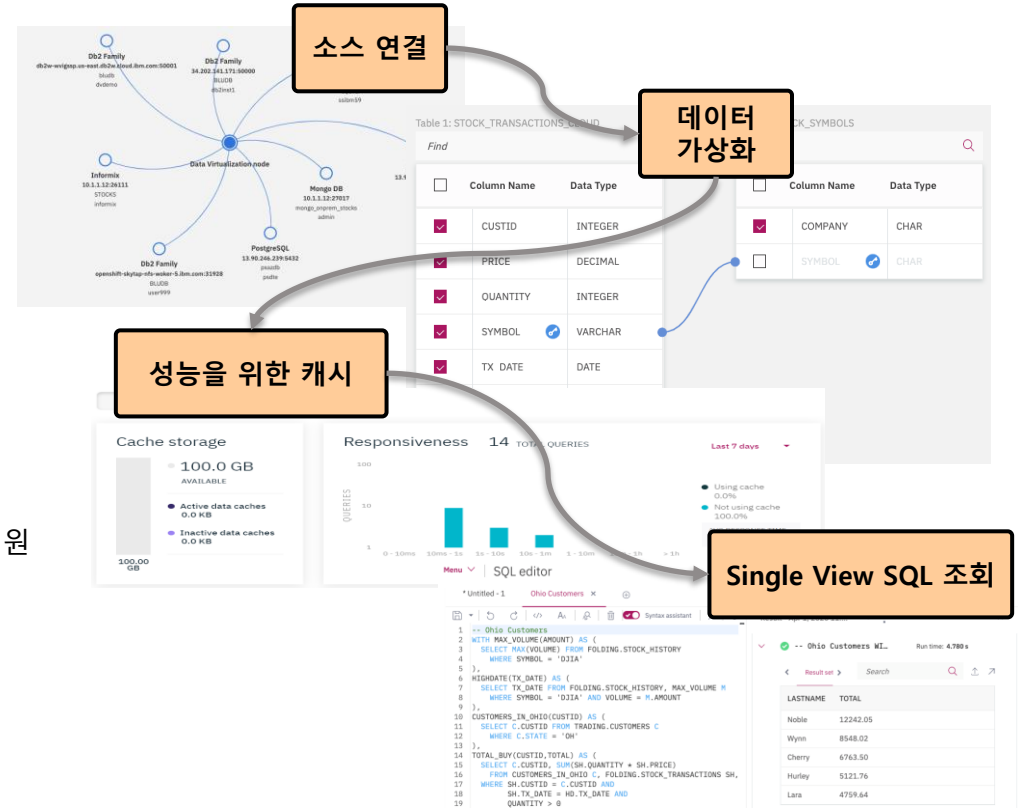


데이터 현대화 > 데이터 가상화

IBM Cloud Pak for Data는 기업 내/외부에 산재해 있는 다양한 데이터 소스들에 대해 데이터 가상화 기술을 통해 단일화된 Single Access Point을 제공하여 데이터 접근 가능성을 높여 줍니다.

가상화 장점

- 데이터 복제가 필요 없어 데이터 중복이 발생하지 않아 **스토리지 비용 절감**
- **Single Data View**를 제공하여 사용자들은 데이터 소스의 위치를 고려할 필요 없음
- **실시간 데이터 액세스가** 가능하여 빠른 인사이트 지원
- **캐싱 기능을** 통해 On-demand 가상 데이터 마트 구성 지원
- 데이터가 있는 곳으로 **작업을 Push Down하여** 기존 데이터 이동 방식 대비 빠른 성능



데이터 거버넌스 > DataOps의 필요성

기업은 하이브리드 통합 거버넌스 모델을 기반으로 신뢰성 있는 분석 기초를 세워 비즈니스 프로세스에 효율성과 투명성을 가져와 **정보 자산에 대한 통찰력을 높이고 규제 준수를 강화할** 수 있습니다.

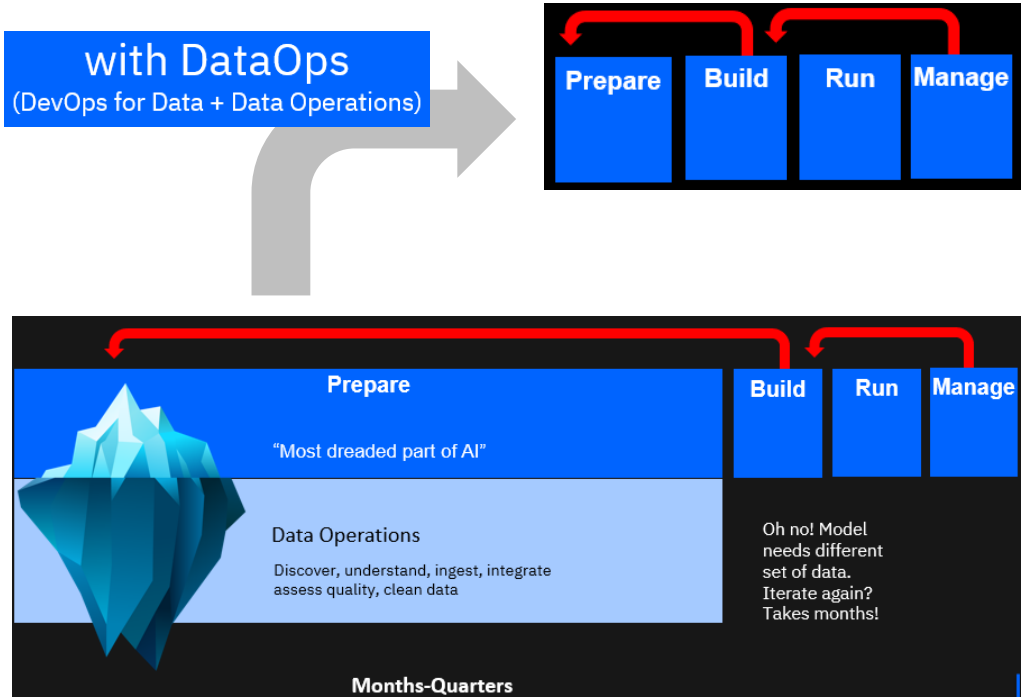
데이터 관련 기업의 Challenge

- ✓ 인사이트 및 예측 불가
- ✓ **Garbage성 데이터의 존재**
- ✓ 새로운 유형의 데이터 소싱의 어려움

- ✓ 고가의 데이터 소유 비용
- ✓ Data Lake의 부재
- ✓ **데이터 준비에 장시간 소요**
- ✓ **데이터 사일로 (Silos)**

- ✓ 방화벽 내/외부의 데이터
- ✓ **기존 데이터의 신뢰 및 품질 문제**
- ✓ 점점 더 많아지는 규제 및 준수 사항
- ✓ **개인 정보 보호**

with DataOps
(DevOps for Data + Data Operations)

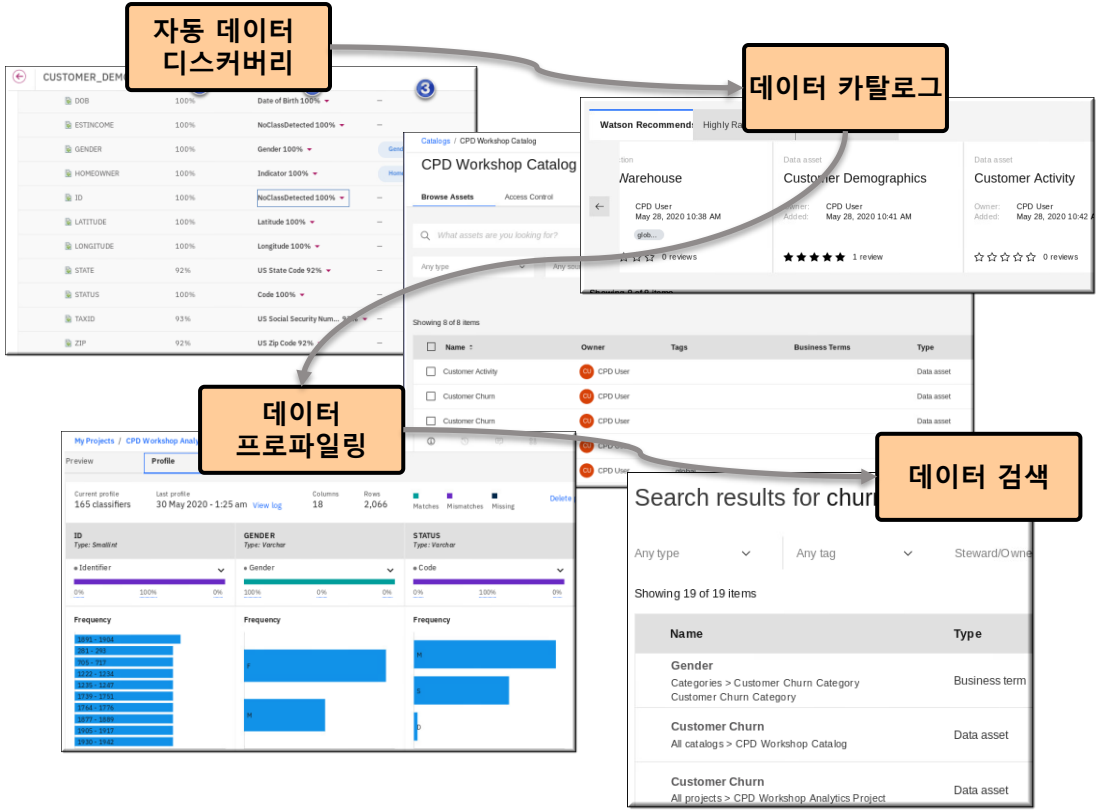


데이터 거버넌스 > 자동 디스커버리/카탈로그

IBM Cloud Pak for Data는 머신 러닝 기반하의 자동화 알고리즘을 통해 데이터의 품질/메타 정보/계보/카탈로그 등을 통해 데이터 소비자에게 신뢰 있는 데이터를 제공하여 분석의 신뢰도를 높여줍니다.

자동 디스커버리/카탈로그 장점

- 자동 디스커버리와 데이터 프로파일링을 통해 **데이터의 품질 확보**
- 자동 데이터 분류 기능을 통해 비즈니스 메타와 자동 매핑을 수행하여 **거버넌스 비용 감소**
- 데이터 소스와 사용자 간의 데이터 게이트 웨이 역할을 수행하여 **데이터 사용 추적 용이**
- 데이터 카탈로그와 데이터 검색 기능을 제공하여 **셀프 서비스 분석 환경 지원**



AI 운영 관리 > MLOps의 필요성

분석 모델의 개발에서 모델의 배포 및 관리, 모델의 성능 유지를 위한 모니터링까지 반복적인 모델 작업의 자동화를 위한 **AI 라이프 사이클에 대한 관리를 지원해야 합니다.**

Trusted AI를 위한 Challenge

Bias

학습 데이터와 AI 모델은 편향되어질 수 있습니다.

Quality

AI/ML 라이프 사이클 동안 잘 관리 되어져야 합니다.

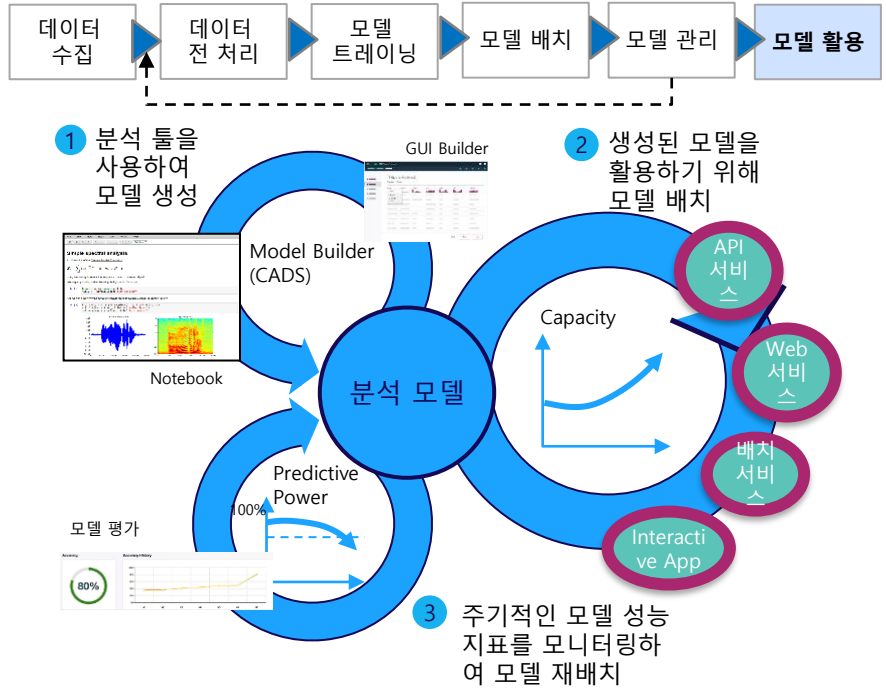
Drift

입력 데이터의 변화가 모델에 부정확한 결과를 야기할 수 있습니다.

Explainability

전통적인 통계 모델은 해석 및 설명이 단순합니다.

AI 라이프 사이클 관리 통합 스택 구현

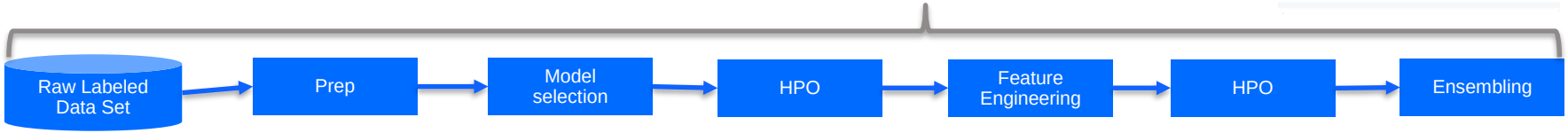


AI 운영 관리 > AutoAI

IBM Cloud Pak for Data의 AutoAI는 AI를 위한 AI 기능으로 수 주에서 수 개월이 소요되는 모델 분석 과정을 수 분 내에 수행할 수 있게 해 주어 개발 생산성을 향상시킵니다.

AutoAI 장점

- 자동 데이터 준비와 모델 개발을 통해 **개발 생산성 향상**
- 코딩 없이 몇 번의 클릭으로 시작할 수 있어 **스킬 갭을 보완**
- 모든 곳에 AI를 활용할 수 있어 **더 많은 Use Case를 발굴**
- 자동 Feature Engineering을 통해 **주요 변수 예측력 증대**
- 파이프라인의 후보 모델 비교를 통해 **최적의 모델 선정**



AI 운영 관리 > 모델 모니터링 – OpenScale

IBM Cloud Pak for Data의 Watson OpenScale은 Trusted AI를 위해 AI 모델 검증 및 모니터링 기능은 물론, payload 데이터를 기반으로 하여 더 나은 모델을 추천하는 기능도 지원하고 있습니다.

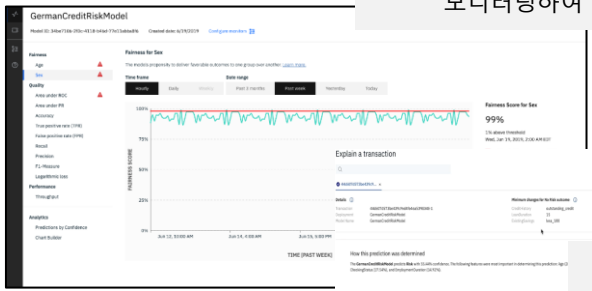
OpenScale 장점

- **Bias 편향성 감지** : Bias의 원인을 명확하게 밝히고, 이를 완화하기 위한 지표 및 데이터 제공
- **Explanability 설명성** : 개별 트랜잭션에 대한 이력 및 중요 속성에 대한 디테일을 포함하여 설명
- **Drift 이상현상 감지** : 시간이 지나면서 트레이닝 데이터와 실제 데이터 패턴과의 차이로 정확도가 떨어지는 것을 분석

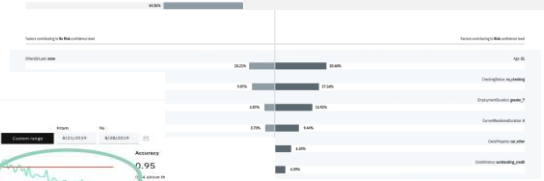
- 무의식적 Bias를 좀 더 잘 이해할 수 있게 됨

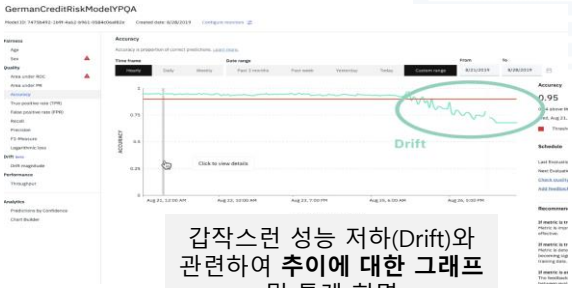
- **AI 도입 시 신뢰성을 제공**

특정 Feature에 대하여 Bias가 있는지 모니터링하여 경고 표시



개별 트랜잭션에 대하여 **모델 결과에 대한 설명** 화면





갑작스런 성능 저하(Drift)와 관련하여 **추이에 대한 그래프** 및 통계 화면

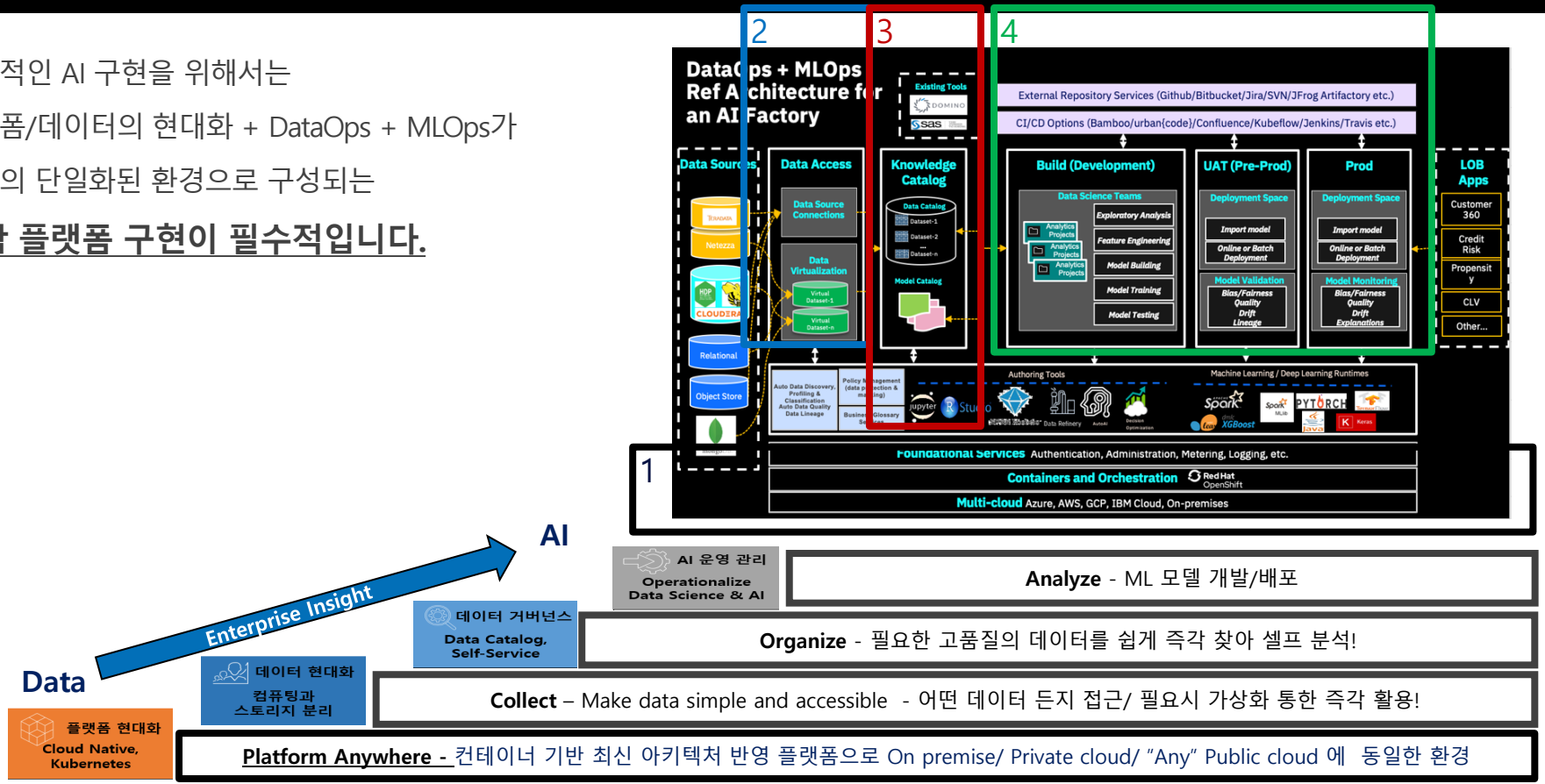
Summary > 통합 플랫폼 구현

성공적인 AI 구현을 위해서는

플랫폼/데이터의 현대화 + DataOps + MLOps가

하나의 단일화된 환경으로 구성되는

통합 플랫폼 구현이 필수적입니다.




IBM