ACCUINSIGHT+

Tech Education

Digital Tech. 센터, AI/Data 그룹
강사 - 김병태매니저
2023.05

SK 주식회사

SK 주식회사

CONTENTS

분석 플랫폼 필요성

Data-Driven 의사결정을 위한 환경은 개인용 분석도구 혹은 오픈소스의 단순 조합만으로는 불충분하며, 다양한 사용자가 손쉽게 접근/활용할 수 있는 분석플랫폼이 필요합니다.

AI 도입 및 확산 저조 원인

01 오픈 소스 활용의 한계

- 고객사 만의 특화된 고유 플랫폼 구축 필요
- 오픈소스의 고유 자산화 미흡
- 오픈소스 Software Built-in 기능 불편

02 일반 사용자 수용 실패

- 스크립트 위주의 모델 개발이 가능한 고급 사용자 중심
- 일반 사용자는 사용이 어려운 환경

03 전산 요청 및 운영 Workload 증가

- 분석가가 스스로 데이터 생성 어려움
  - Data 전처리 및 분석용 데이터 생성 등
- 이관된 모델을 운영하기 위한 Pipeline
  - Scheduler 도입의 필요

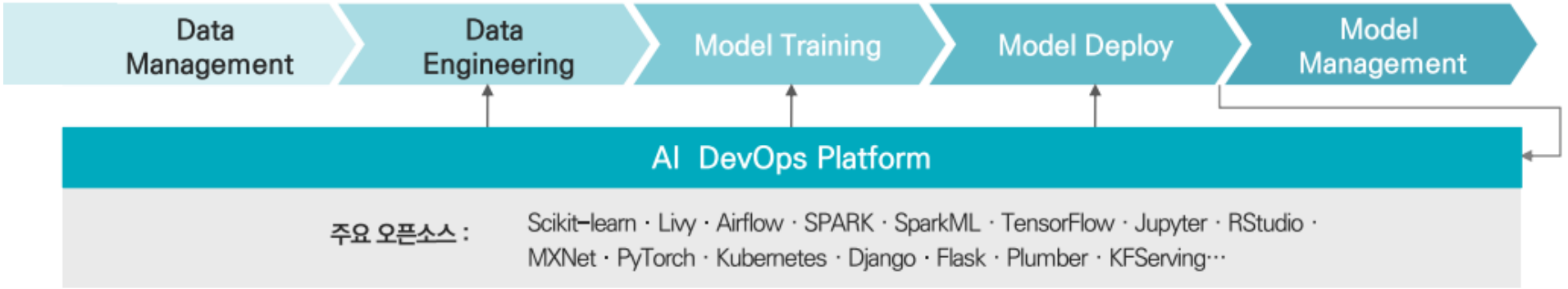
플랫폼 요구사항

Data-Driven 의사결정을 위한 환경

- ✓ **GUI 기반 개발 환경**
  Scripting 없이 모델링 수행
- ✓ **사용자 수준별 맞춤형 분석 환경**
  고급분석가, 일반분석가 별 맞춤 환경 제공
- ✓ **배포 / 운영 자동화**
  운영을 위한 별도의 전환 코딩 없음
- ✓ **모델 및 API의 Life Cycle 관리**
  모델 성능/API 트래픽 모니터링 등
- ✓ **분석가 Self Service**
  Data 처리 전 과정 Seamlessness 확보

AI DevOps Tech

데이터 수집/처리, 모델개발/배포 및 운영프로세스를 일원화한 AI DevOps Platform으로 발전하고 있으며, 적재적소에 맞는 오픈소스 활용 및 유기적인 결합이 중요해지고 있습니다.



Meta Data

- 메타 데이터 관리
- 데이터 파이프라인 구축 (Source → Data Lake)

Data Set

- 데이터 전처리
- 탐색적 데이터 분석
- Feature Engineering

ML Code

- ML Framework을 활용한 모델 학습
- 모델 Inference

API

- 추론을 위한 서비스 API 개발/배포
- 선정된 모델 운영 전환

Monitoring

- 배포된 추론 서비스 모니터링
- 실시간 모델 정확도 측정

통합 데이터 분석 플랫폼

AccuInsight+는 데이터의 수집부터 분석 및 활용까지 AI 기반 분석 활동을 담당하는 통합 분석 플랫폼입니다.

AI 기반 분석활동의 A to Z를 담당하는 통합 분석 Platform

데이터 수집/처리/분석 및 모델링/배포/모니터링/모델 관리 등

ACCUINSIGHT+

Data Source (On-Premise)	Data Pipeline (Data 수집 & 가공)	Model Pipeline (모델 개발 및 예측)	Model Development (학습 및 최적화)	Model Serving (배포)	Model Management (모니터링 및 관리)
• Object Storage • Database • Data Lake (HDFS)	• 데이터 수집 • 데이터 탐색/조회 • 데이터 가공 • 피처 엔지니어링	• 알고리즘 선택 • 학습 실행 • 모델 적용/스케줄링 • 예측	• Sandbox 생성 • 형상관리 • 모델 학습 • Auto ML • Multi Experiment	• 배포 • Auto Scaling	• Data Drift • 버전 관리 • Pipeline Builder

Task 01 — End To End 분석 및 모델링

- 데이터의 수집 및 처리부터 배포된 모델의 성능 관리까지 AI 기반 분석활동 개발 및 관리 지원

Task 02 — 맞춤형 분석 환경 제공

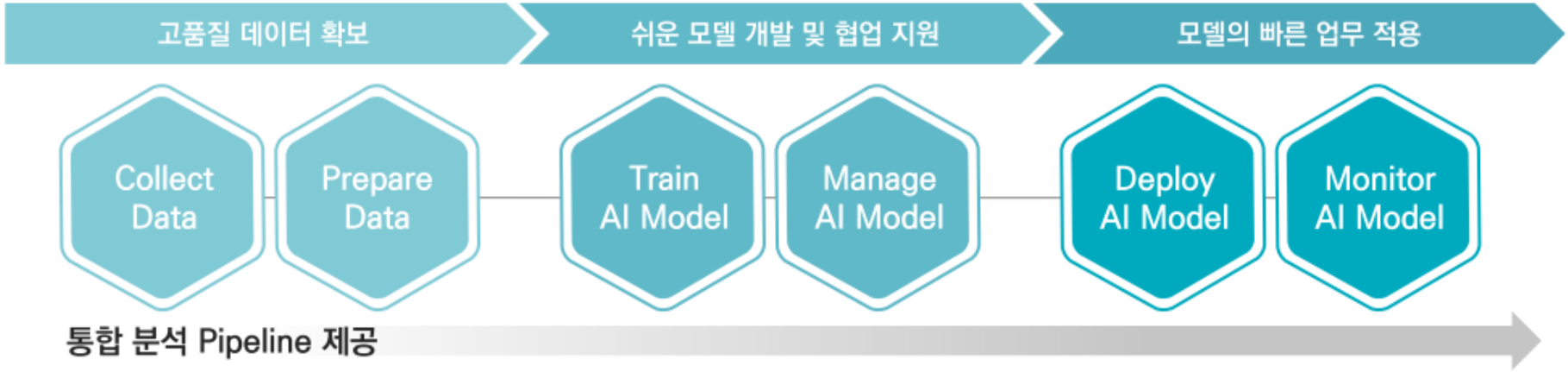
- 사용자 수준별 맞춤형 분석환경 제공 및 배포/운영 자동화
- 초급 분석가/현업을 위한 AutoML 제공
- 고급 분석가를 위한 DL Framework, IDE 제공

Task 03 — 모델 관리

- 모델 Life cycle 관리를 통한 AI 모델 관리 및 사용성 강화(Deployment/Monitoring)

통합 분석 Pipeline 제공

분석모델을 빨리 개발 및 학습하고, 운영 환경에 쉽게 적용하기 위한 분석 Pipeline의 중요성이 증가하고 있습니다.



Task 01

AI 모델 개발 및 적용 속도 향상

- No-Coding & Web UI 분석 환경
- 재사용 가능한 Workflow 기반의 모델 운영
- 분석에 필요한 고품질의 학습 데이터 생성

Task 02

CDS 분석 역량에 따른 맞춤형 지원

- AutoML을 이용한 One-Click 분석 자동화
- 업무 전문가를 위한 자동화된 모델 생성
- 고급 분석가를 위한 독립된 분석환경

Task 03

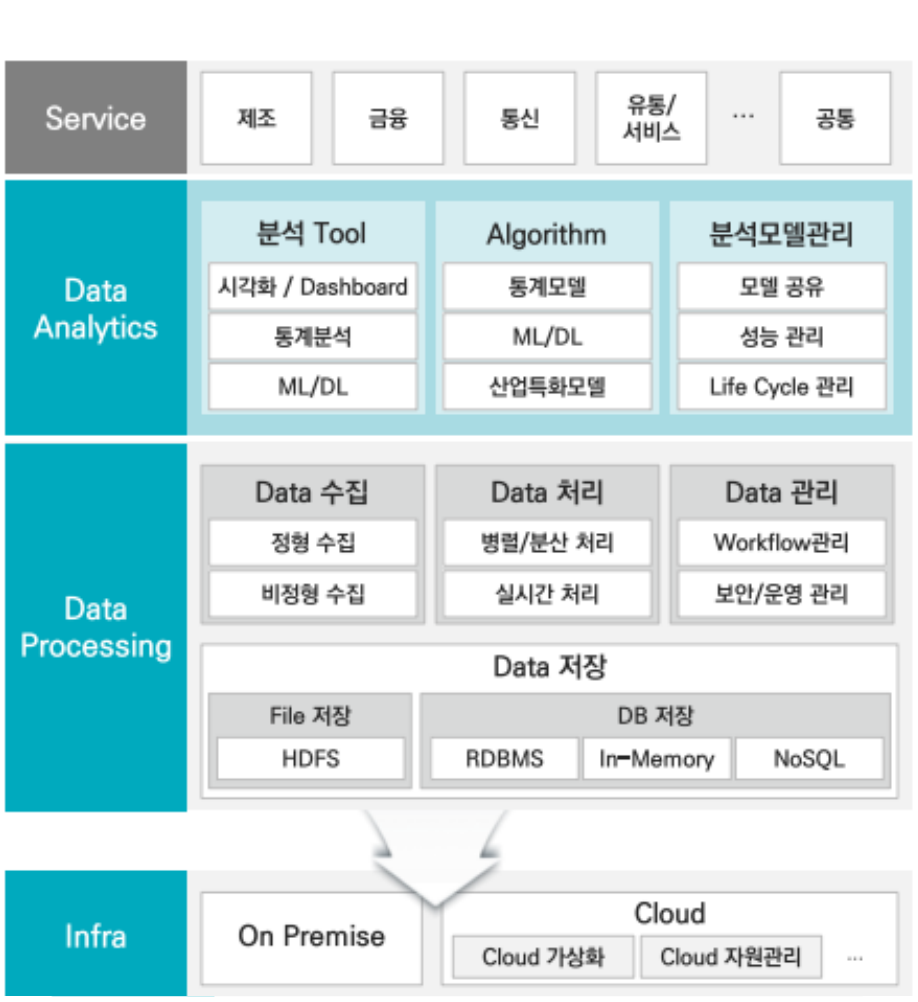
효율적인 AI 모델 관리

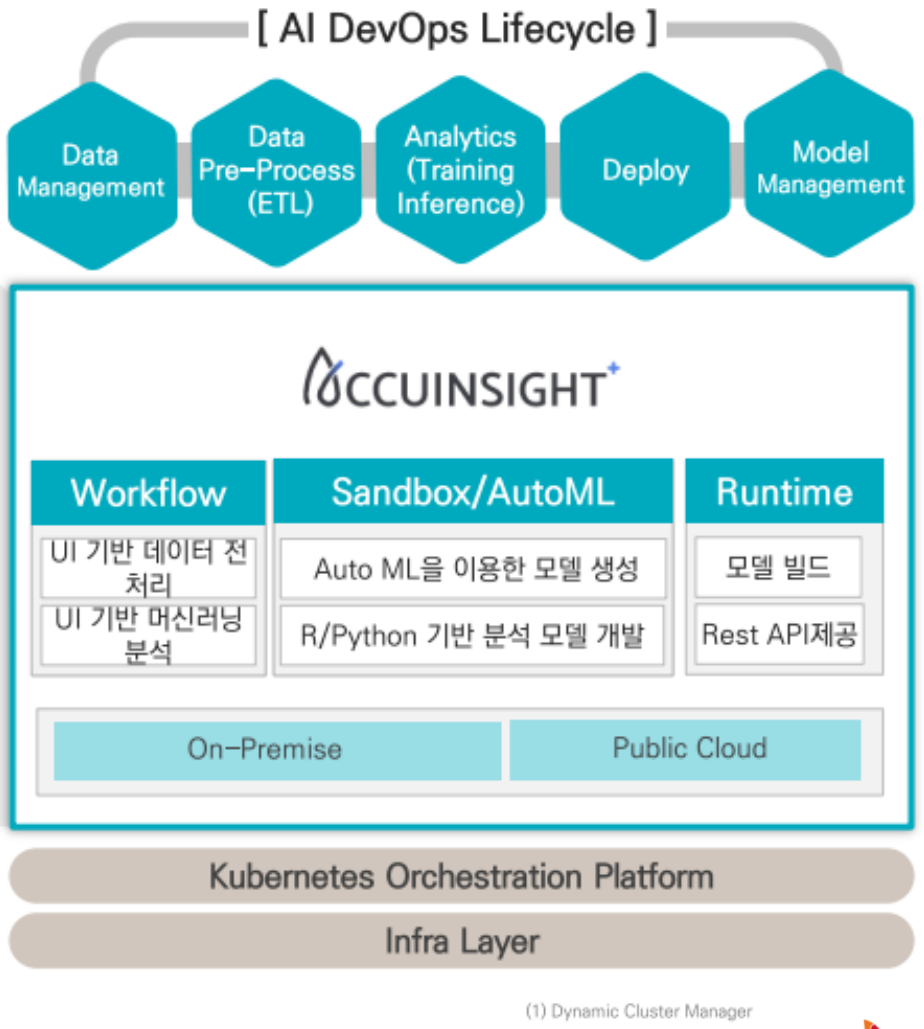
- 조직/분석가 별 분산된 모델의 통합관리
- AI Model의 전체 lifecycle 관리
- 분석가가 필요한 자원의 동적 할당 및 모델 재사용

* 참고 : MLOps Tooling Landscape V2 (https://huyenchip.com/2020/12/30/mlops-v2.html)

AccuInsight+ Achitecture

AccuInsight+는 Data 전처리 및 분석/운영의 Data Science 전 영역을 통합 지원합니다.

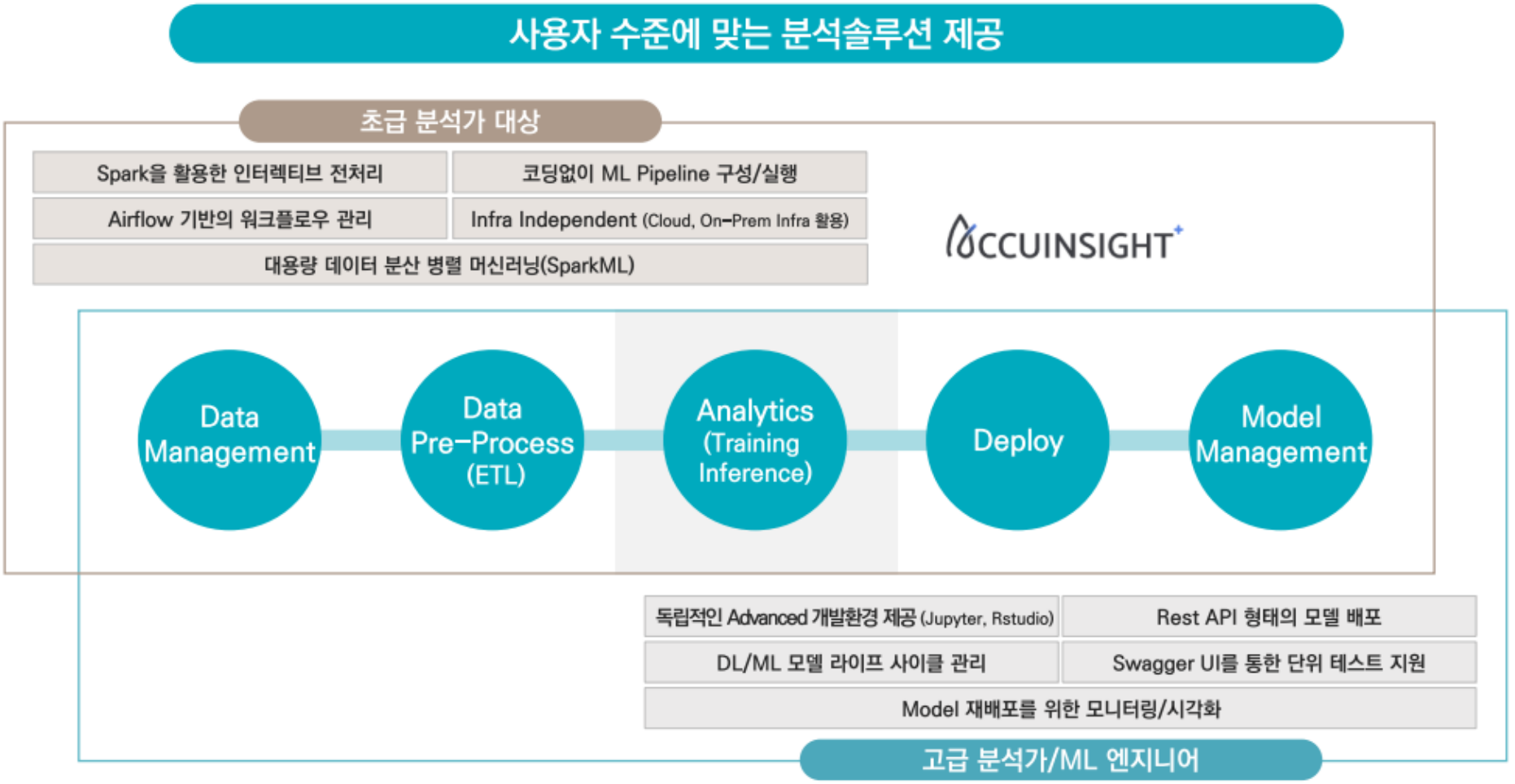




(1) Dynamic Cluster Manager

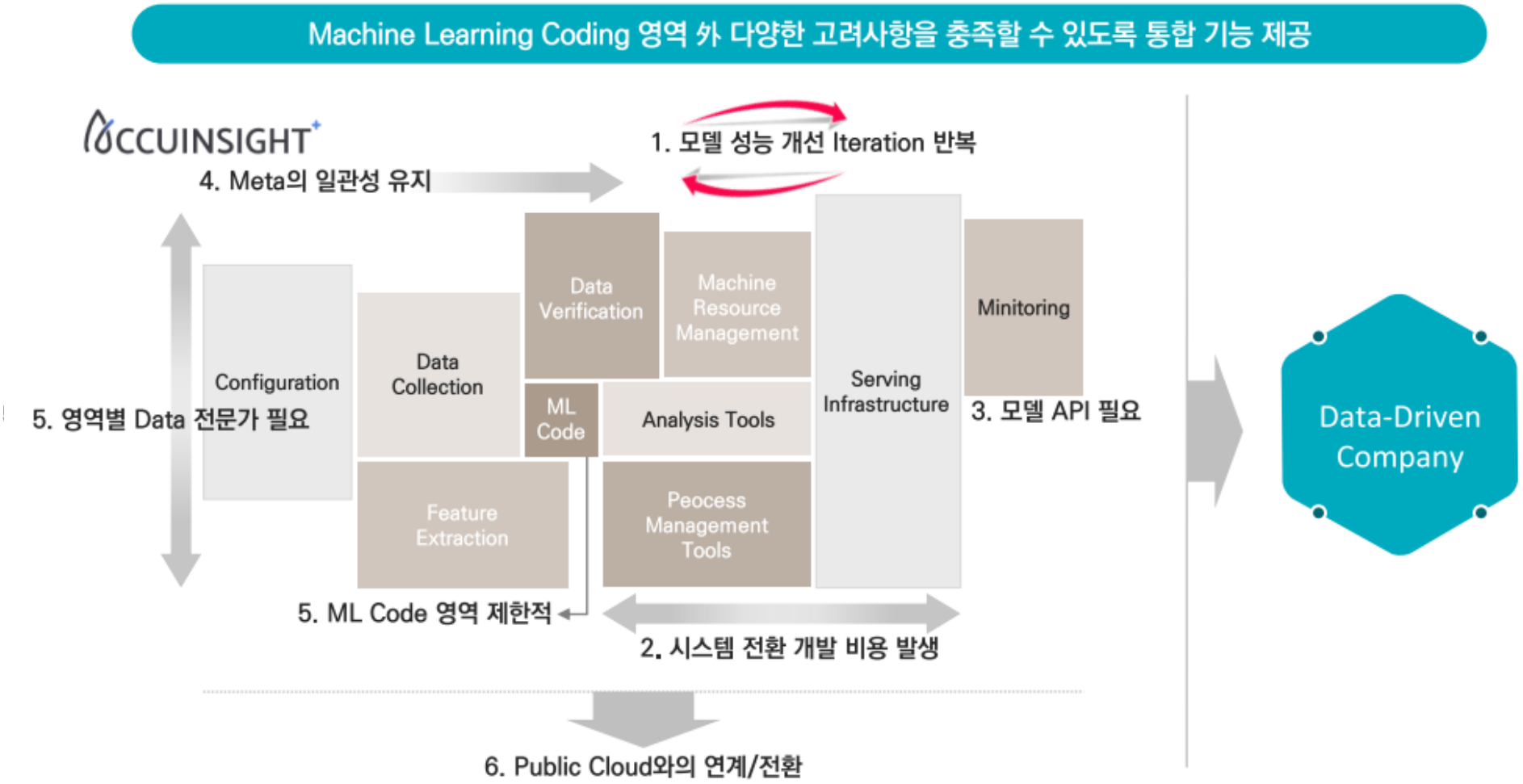
AccuInsight+ 기능 Coverage

초급분석가 및 현업을 위한 Drag & Drop UI 기반 데이터 전처리/머신러닝 분석 및 고급사용자를 위한 코딩 기반 분석모델 개발/운영 배포 모두 가능합니다.



AccuInsight+의 역할

모든 산업 도메인에 적용가능한 공통 플랫폼으로, 분석모델 개발 및 다양한 Data Science 기능 제공으로 고객사의 Data-Driven 비즈니스 전환을 효과적으로 지원합니다.



AccuInsight+의 차별화 point

데이터 분석에 필요한 다양한 협업, 자동화, 커스터마이징 솔루션을 인프라 환경에 구애 없이 활용 가능합니다.

Collaboration

고급-초급 분석가
모델 공유 및 협업 촉진

분석 모델 배포 시
엔지니어링 단순화

공동작업을 위한
Workspace 제공

Automation

분석모델 개발을 위한
최적 하이퍼 파라미터추천

분석 실험에 적용된
코드/모델 자동 이력관리

데이터 전처리과정
자동화

Customization

데이터 특성에 따른
파이프라인 최적화 지원

운영환경에 적합하도록
솔루션 성능 최적화

고객사 상황에 맞게
커스터마이징

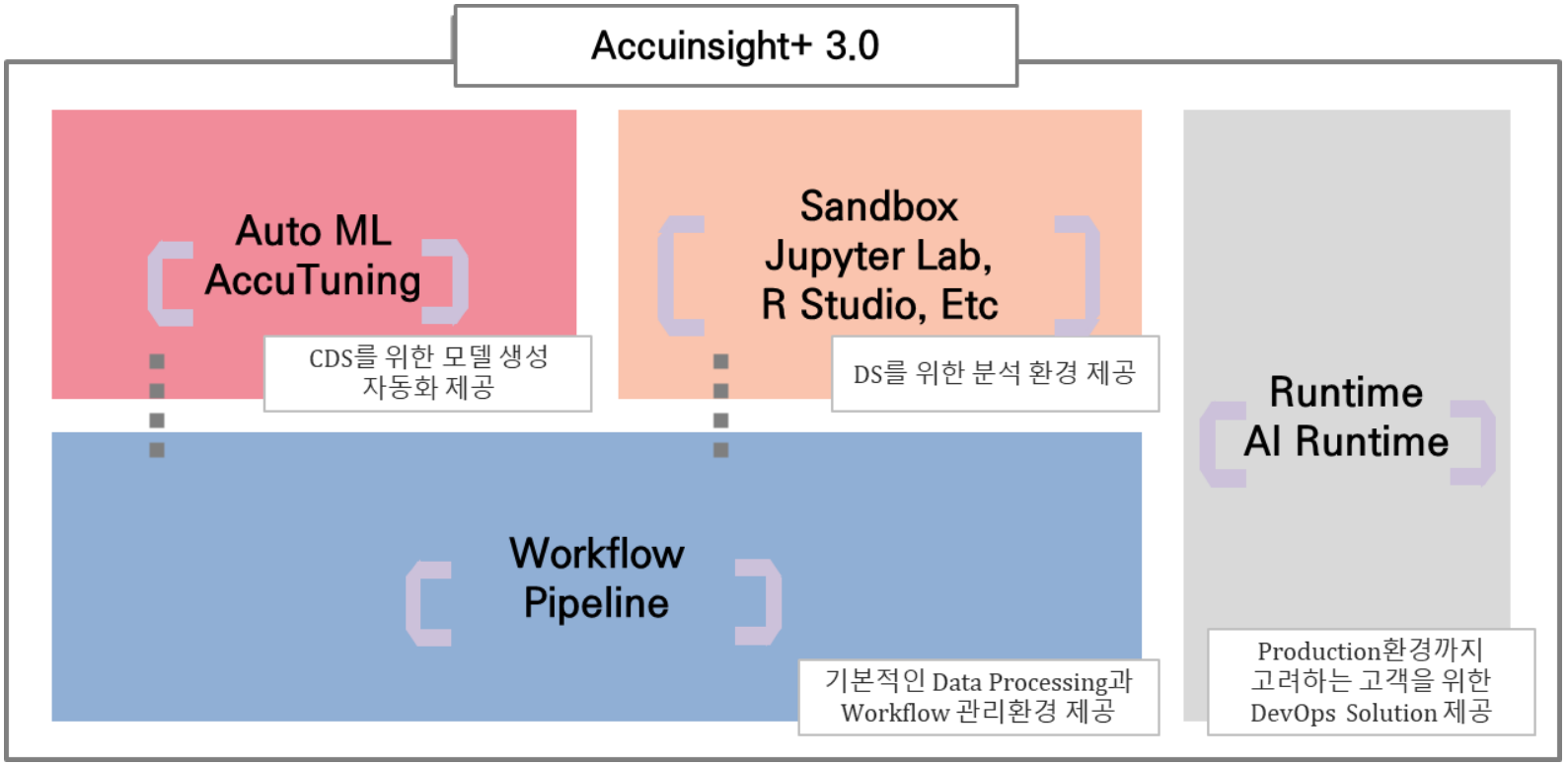
Any Infrastructure

다양한 Public Cloud 지원
(AWS, GCP, AZURE)

On-Prem. / Private Cloud
인프라 설치 가능

AccuInsight+의 3.0 주요 변경 사항

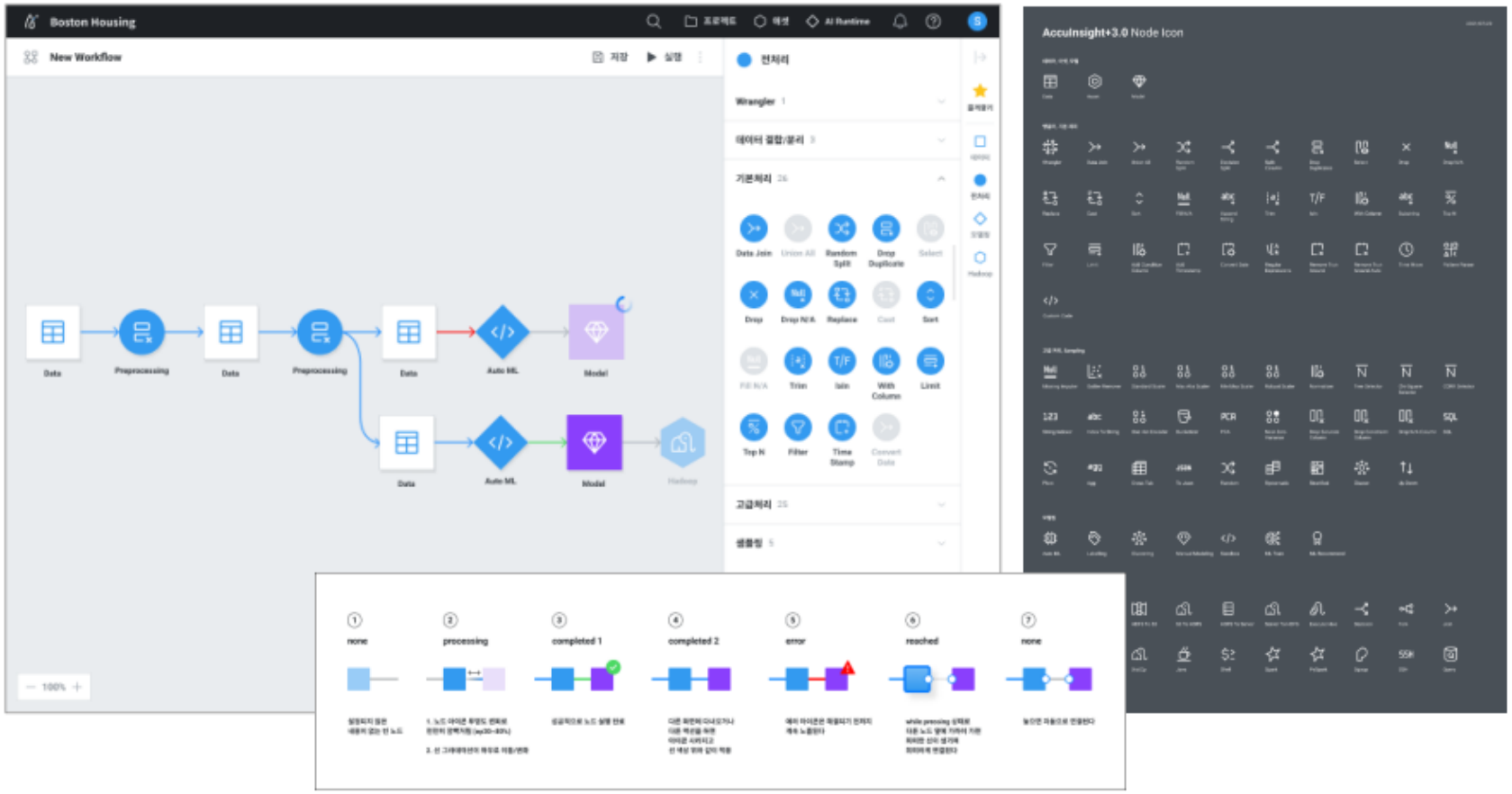
사내 주요 AI솔루션, 데이터분석 플랫폼을 하나의 제품으로 통합하였습니다.
Accuinsight+2.0에서 Pipeline, Modeler는 Workflow, Sandbox로 명칭이 변경됩니다.



- Workflow기반으로 AutoML, Sandbox가 노드 단위 기능으로 제공되며, 추후 개별 분리 Product로 분리가능하게 개선 예정 (*2022년도)

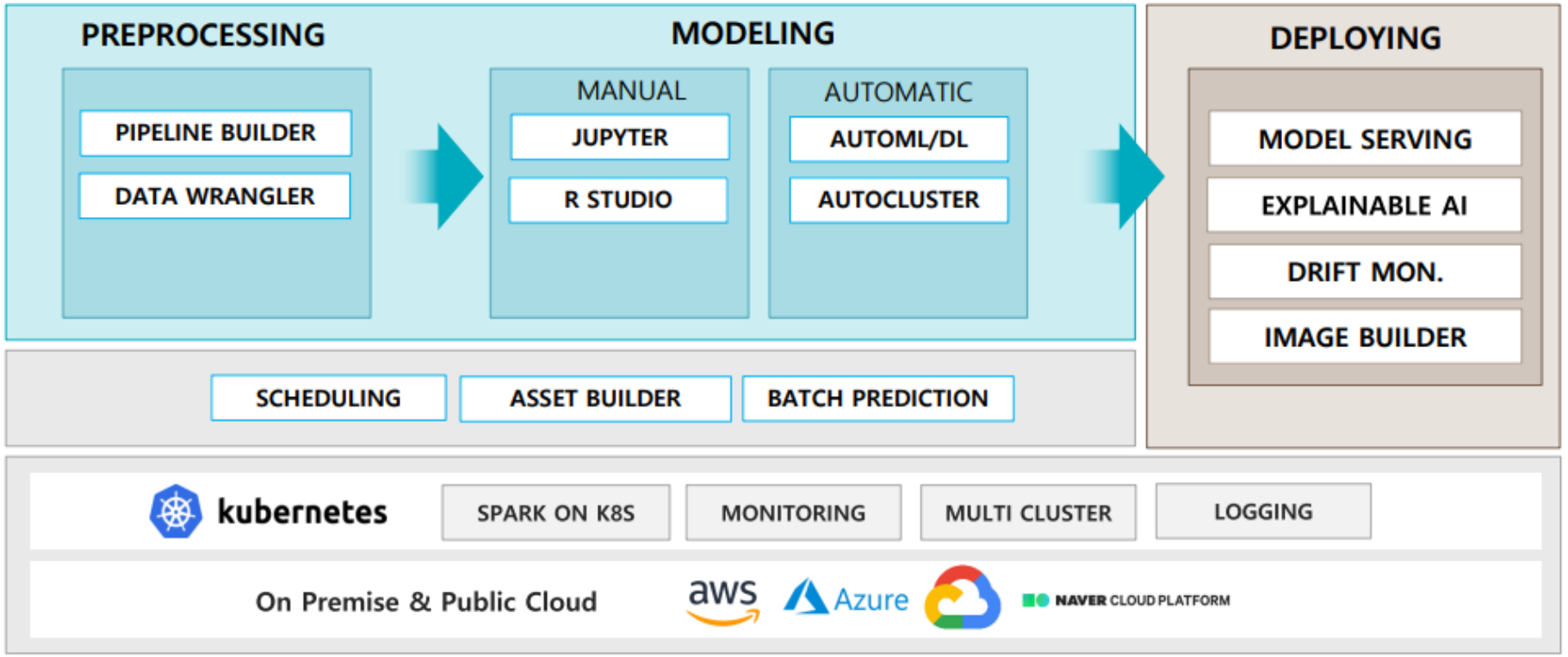
AccuInsight+의 3.0 주요 변경 사항

데이터분석에 필요한 사용자 경험을 최우선으로, 솔루션 전체 UI/UX, Frontend를 새롭게 설계하고 개발하였습니다.



AccuInsight+의 3.0 주요 변경 사항

멀티 클라우드 기반의 K8S기반의 동적으로 자원 사용이 가능한 구조로 전체 아키텍처를 개선하였습니다.



SK 주식회사

Ⅲ AccuInsight 3 실습

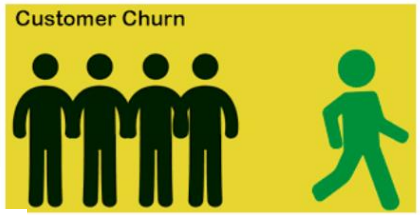
1. UI기반 ML 서비스 만들기
   - AutoML을 활용한 정형 데이터 모델 생성 후 런타임 배포 서빙

2. Code기반 ML 서비스 만들기
   - Sandbox 정형 데이터 분류 모델 생성 및 런타임 배포 실습
   - Sandbox 비정형(이미지) 분석 모델 개발 및 런타임 배포 실습
   - Sandbox 비지도 학습 : 영화 추천 모델 개발 및 런타임 배포 실습
   - Sandbox 비지도 군집화(클러스터링) 모델 개발 및 런타임 연계 실습

UI기반 ML 서비스 만들기 – AutoML을 활용한 정형 데이터 모델 생성 후 런타임 배포 서빙

1 실습 시나리오 1 : AutoML - 은행고객이탈 데이터 이진 분류 - 어떤 데이터를 갖고 있는 고객이 탈퇴를 할까?

- 주제 : 뱅킹 고객 이탈예측 (분류 문제)
- 종속 변수(Target) : 이탈 여부
- 독립 변수(Feature) : 연령, 성별, 지역구분, 신용점수, 잔액 등
- 데이터 특이사항 :
  - Drop 전처리 노드를 사용하여 CustomeId 삭제
  - Wrangler 를 이용 - name , baseYear , joinyear 를 삭제 / 파생변수생성 / balance의 빈 값을 0으로 대체
  - 'AutoML' 에서 Feature Engineering 처리



1. 기본 노드 전처리
- feature 삭제 (CustomeId)

2. Wrangler 전처리
- 컬럼삭제
- 파생변수생성
- 빈값 채우기

3. Auto ML 실행
- lightGBM 추가 체크

4. 예측용 워크플로우 예측
Predict.New Workflow1 생성 및 예측

[모델 배포]

5. 배포자원설정
CPU - 4
memory - 8

6. 서빙시작
서비스 대시보드 상태 확인

7. 예측테스트
- 예측테스트
- 예측이력
- 배치 예측

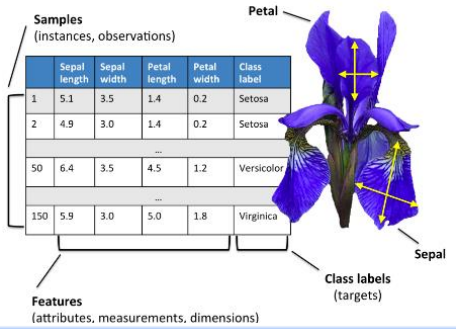
8. 모델 드리프트
모델 드리프트 등록 및 결과 확인

9. 카나리 배포
모델 프로모션

Code기반 ML 서비스 만들기 – Sandbox Jupyter Lab 정형 데이터 모델 생성 후 런타임 배포 실습

2 실습 시나리오 2 : Sandbox 모델링 - iris 데이터 멀티 분류

- 주제 : 붓꽃의 품종 예측 (분류 문제)
- 종속 변수(Target) : 품종(setosa, versicolor, virginica)
- 독립 변수(Feature) : 꽃받침 가로길이, 꽃받침 세로길이, 꽃잎 가로길이, 꽃잎 세로길이
- 데이터 특이사항 :
  - 변수에 결측치 존재하지 않음
  - 독립 변수에 수치형 변수 Normalization 필요
  - 독립 변수에 범주형 변수가 존재 하지 않음



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

1. Sandbox 생성 및 데이터 가져오기
- Accuedu-XX
- iris.csv

2. 전처리 노드 연결
DropDuplicate
Dropna

3 .노드 연결
Dataset노드와 Sandbox 연결

4. 템플릿실행
접속 및 템플릿 실행

5. sample실행
- Sample 실행
- Experiment 모델 성능 비교

6. 모델추출 및 배포
모델 추출 및 런타임 배포

7. 배포자원설정
CPU - 4
memory - 8

8. 서빙시작
서비스 대시보드 상태 확인

9. 예측테스트
- 예측테스트
- 예측이력
- 배치 예측

10. 모델 드리프트
모델 드리프트 등록 및 결과 확인

11. 카나리 배포
모델 프로모션

12. Rest api 호출
온라인 예측 실행

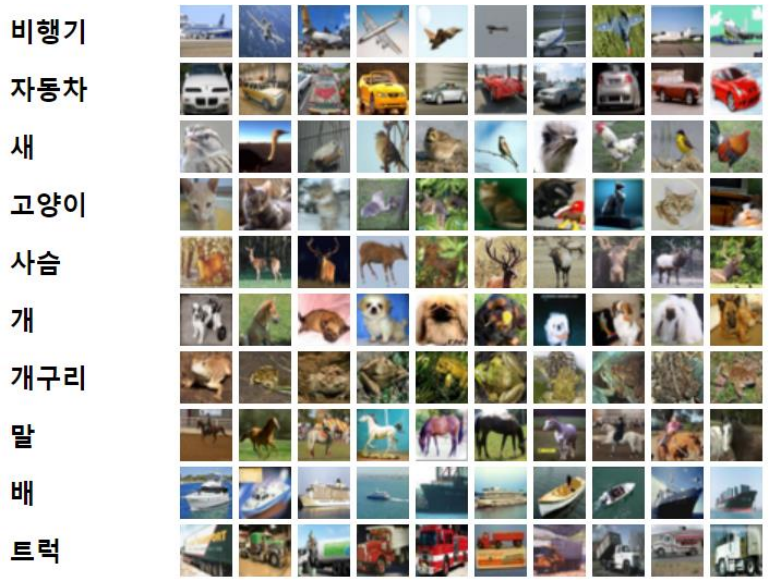
Code기반 ML 서비스 만들기 – Sandbox Jupyter Lab 비정형(이미지) 분석 모델 개발 및 런타임 배포 실습

3 실습 시나리오 3 : Sandbox 비정형(이미지) 분석모델 개발 및 배포 실습

1. 워크플로우 생성	2. 분석환경생성	3 .신규모듈 생성
Sandbox 바로가기	Sandbox 분석 환경 설정 진행 및 접속	신규 모듈 생성

4. 기존모듈 수정	5. 전체코드 실행	6. 모델추출 및 배포
기존 모듈 수정	cifar10-tf-cls-example.ipynb 실행	모델 추출 및 런타임 배포

7. 배포자원 설정	8. 서빙시작	9. Rest api 호출
CPU - 4 memory - 8	온라인 서비스 시작	온라인 예측 실행



- 주제 : CIFAR-10 dataset에 대한 이미지 분류를 Keras를 사용한 CNN(Convolution Neural Network) 구현
- 10 종류의 이미지와 정답 레이블이 들어있음
- 목표 : CNN을 통해 이 10가지 종류의 이미지들을 적절하게 분류

Code기반 ML 서비스 만들기 – Sandbox Jupyter Lab 비지도 학습 : 영화 추천 모델 개발 및 런타임 배포 실습

4 실습 시나리오 4 : Sandbox 비지도 영화 추천 모델 개발 및 배포 실습

1. 워크플로우 생성	2. 분석환경생성	3 .신규모듈 생성
Sandbox 바로가기	Sandbox 분석 환경 설정 진행 및 접속	신규 모듈 생성

4. 기존모듈 수정	5. 전체코드 실행	6. 모델추출 및 배포
기존 모듈 수정	movielens-recommender.ipynb 실행	모델 추출 및 런타임 배포

7. 배포자원 설정	8. 서빙시작	9. Rest api 호출
CPU - 4 memory - 8	온라인 서비스 시작 및 예측테스트	온라인 예측 실행

- 주제 : Movielens 데이터를 이용하여 추천시스템만들기
- '당신만을 위한 추천'과 같은 데이터를 기반으로 한 개인화 서비스를 제공
- 목표 : 추천 시스템에서 인기 있는 알고리즘을 활용해서 사용자의 평점 정보를 활용해 미시청 영화를 추천

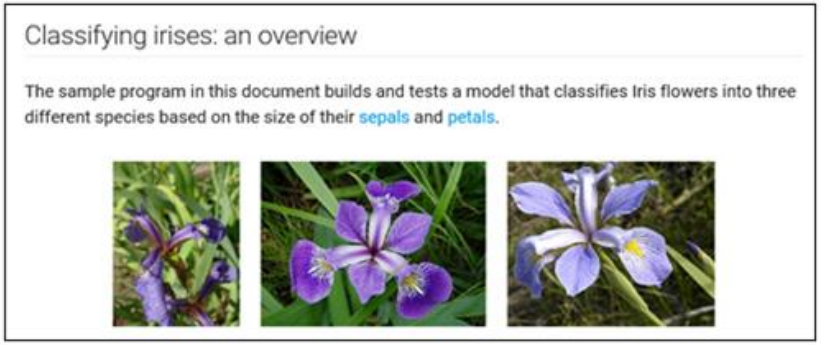
Code기반 ML 서비스 만들기 – Sandbox 비지도 군집화(클러스터링) 모델 개발 및 런타임 연계 실습

5 실습 시나리오 5 : Sandbox 비지도 iris 군집화 모델 개발 및 배포 실습

1. 워크플로우 생성	2. 분석환경생성	3 .신규모듈 생성
Sandbox 바로가기	Sandbox 분석 환경 설정 진행 및 접속	신규 모듈 생성

4. 기존모듈 수정	5. 전체코드 실행	6. 모델추출 및 배포
기존 모듈 수정	iris-clustering-example.ipy thon 실행	모델 추출 및 런타임 배포

7. 배포자원 설정	8. 서빙시작	9. Rest api 호출
CPU - 4 memory - 8	온라인 서비스 시작 및 예측테스트	온라인 예측 실행



- 문제 : 이 문제는 꽃 특징을 기반으로 아이리스 꽃 집합을 여러 그룹으로 나누는 것에 관한 것

- 상세 : Iris 데이터는 총 3개의 클래스로 구성되어 있으며 각 클래스에는 50개의 데이터가 각각 있습니다. 각 클래스에 있는 데이터는 꽃받침(sepal)과 꽃잎(petal)의 길이(length)와 너비(width)로 되어 있습니다.

- 목표 : clustering을 할 때 가장 중한 파라미터 중 하나인 k를 찾고 특정으로부터 데이터 집합의 구조를 파악하고 데이터 인스턴스가 이 구조에 어떻게 맞는지 예측

감사합니다.