



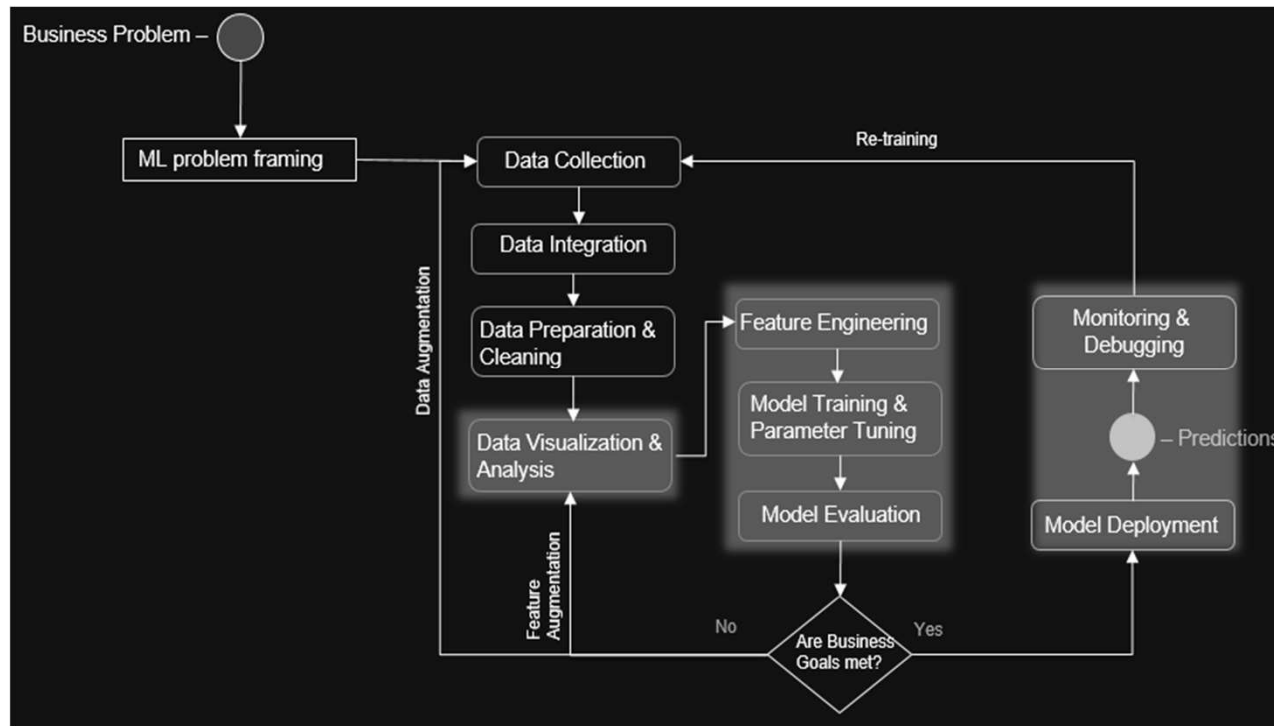
세종사이버대학교

01

ML 개발자 및 데이터 과학자를 위한 ML 서비스



1) 머신러닝(ML) 사이클



Amazon SageMaker
데이터사이언스에서
컴퓨터사이언스 제거로
데이터 중심 문제 해결에 집중

2) Amazon SageMaker 개요

데이터과학 및 ML워크플로우를 위한 완전 관리형 서비스

1

준비(Prepare)

- 훈련 데이터 수집 / 준비를 위한 데이터 처리 및 라벨링(SageMaker GroundTruth), 처리(Processing) 인스턴스

2) Amazon SageMaker 개요

 데이터과학 및 ML워크플로우를 위한 완전 관리형 서비스

1

준비(Prepare)

2

빌드(Build)

- 사전 빌드된 Jupyter Notebook 인스턴스,
다양한 SageMaker 내장 ML 알고리즘 선택 및 빌드



2) Amazon SageMaker 개요

데이터과학 및 ML워크플로우를 위한 완전 관리형 서비스

1

준비 (Prepare)

2

빌드 (Build)

3

훈련 (Train)

- 원클릭 데이터 훈련, 훈련 인스턴스의 손쉬운 확장, 분산 훈련 (Distributed Training) 지원, Hyper Parameter 최적화 (베이지언 최적화)를 통한 자동 모델 튜닝

2) Amazon SageMaker 개요



세종사이버대학교

데이터과학 및 ML워크플로우를 위한 완전 관리형 서비스

1

준비 (Prepare)

2

빌드 (Build)

3

훈련 (Train)

- 원클릭 데이터 훈련, 훈련 인스턴스의 손쉬운 확장, 분산 훈련 (Distributed Training) 지원, Hyper Parameter 최적화 (베이지언 최적화)를 통한 자동 모델 튜닝

- 훈련을 수행하기 위해 사전 설정해야 하는 값
- 세대 (Epoch)의 개수, 레이어 (Layer)의 개수, 학습율 (Learning Rate), 숨겨진 유닛 (Hidden Unit)의 개수 등이 대표적인 패러미터

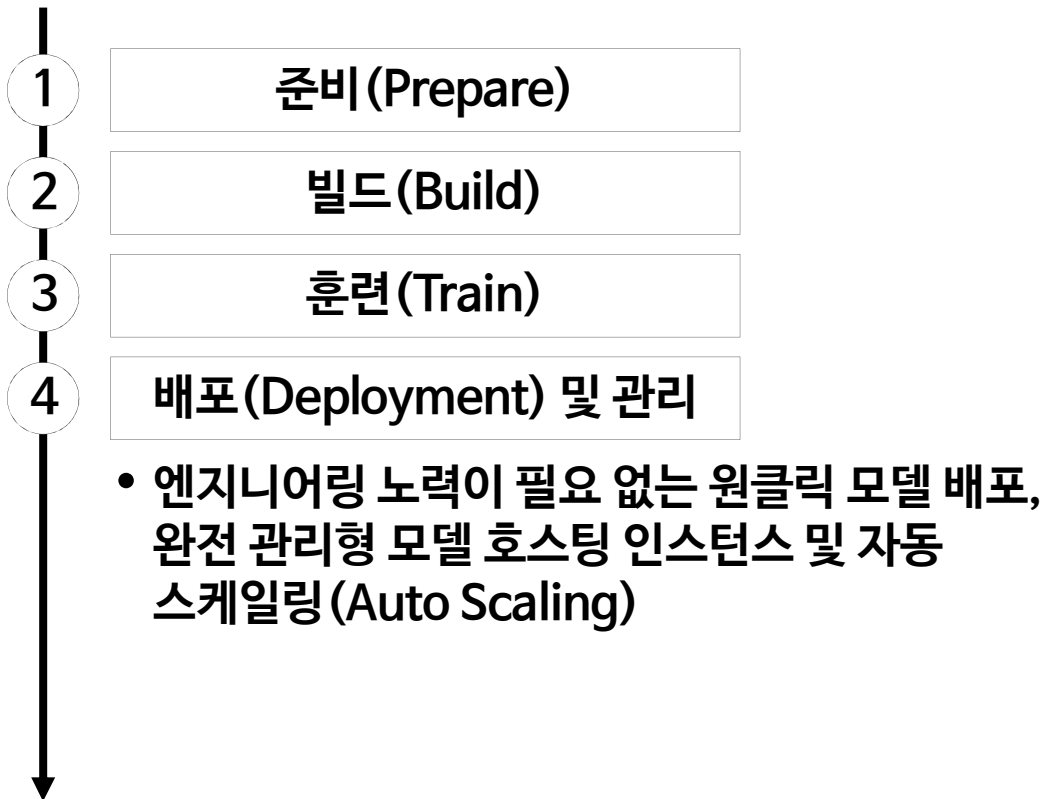


HPO (Hyper-parameter Optimization)

훈련을 수행하기 위해 사전에 설정해야 하는 값인 Hyper-parameter의 최적값을 탐색하는 문제로, 훈련 완료 모델의 일반화 성능을 최고 수준으로 높일 수 있도록 구성

2) Amazon SageMaker 개요

 데이터과학 및 ML워크플로우를 위한 완전 관리형 서비스



3) Amazon SageMaker 인스턴스

머신러닝 개발환경(예 데이터 탐색 및 시각화), 데이터처리환경, 훈련 환경, 추론 환경의 서로 다른 인프라(인스턴스) 요구 사항을 지원

노트북 인스턴스

모델 빌드를 위한 노트북 인스턴스
생성 후, 데이터 처리, 훈련, 모델 배포,
모델 테스트 및 확인 수행

처리(Processing) 인스턴스

데이터처리(예 원시데이터를
훈련데이터로 전환하기 위한 데이터
전처리 및 피처 엔지니어링 등) 및
모델 평가를 위한 분석 작업 지원

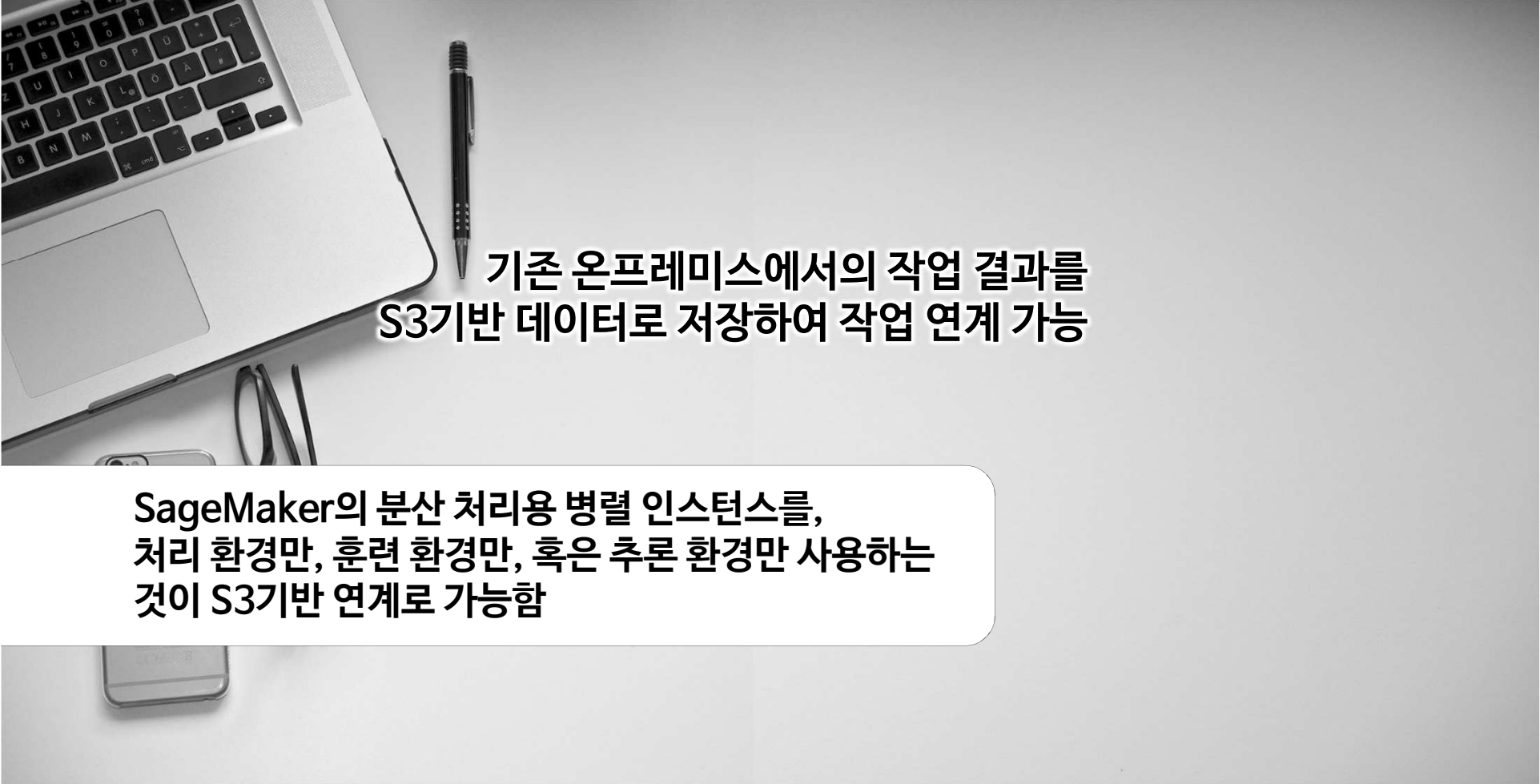
훈련 인스턴스

훈련용 데이터에 훈련 알고리즘을
적용하여 결과물인
모델 아티팩트를 S3에 저장

호스팅(추론) 인스턴스

훈련 결과인 모델을 사용자가
추론 결과를 요청할 수 있도록
엔드포인트로 구성

3) Amazon SageMaker 인스턴스

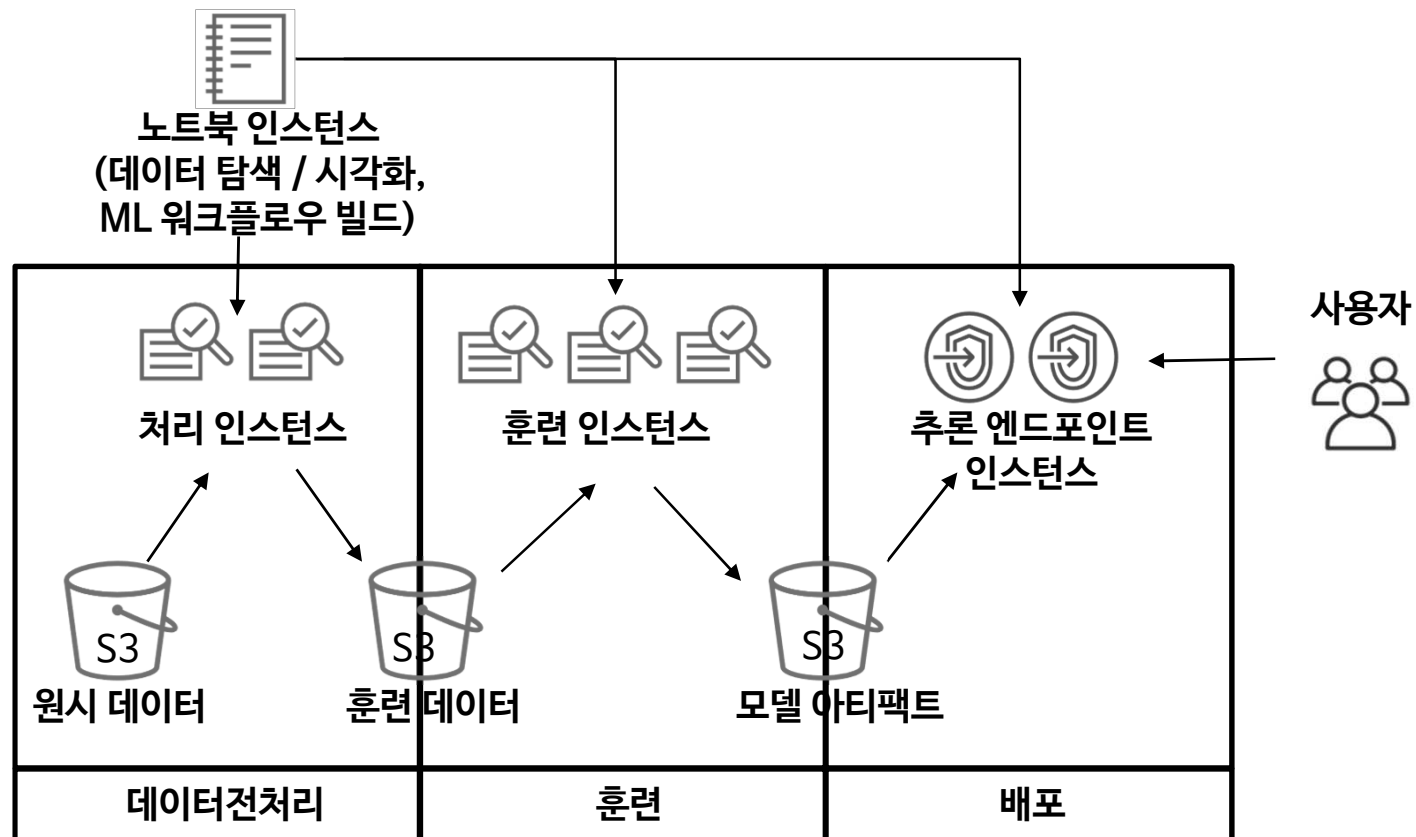


기존 온프레미스에서의 작업 결과를
S3기반 데이터로 저장하여 작업 연계 가능

SageMaker의 분산 처리용 병렬 인스턴스를,
처리 환경만, 훈련 환경만, 혹은 추론 환경만 사용하는
것이 S3기반 연계로 가능함

4) Amazon SageMaker 워크플로우

SageMaker 노트북 인스턴스에서 ML 워크플로우를 빌드



5) Amazon SageMaker 인스턴스 타입(유형)

 SageMaker는 다음과 같은 ML Compute 인스턴스를 제공

T 패밀리	범용 Burstable 성능 인스턴스, 노트북 인스턴스에 적합
M 패밀리	범용 인스턴스, 표준 CPU 대 메모리 비율
R 패밀리	메모리 최적화 인스턴스, 메모리 내에서 많은 데이터셋을 처리하는 워크로드에 빠른 성능을 제공하도록 설계
C 패밀리	컴퓨팅 최적화 인스턴스

5) Amazon SageMaker 인스턴스 타입(유형)

 SageMaker는 다음과 같은 ML Compute 인스턴스를 제공

P 패밀리	빠른 훈련을 위한 대규모 분산 훈련에 적합	} 하드웨어 가속기 혹은 Co- processor를 활용한, 가속화 컴퓨팅 제공 인스턴스
G 패밀리	가속화된 추론, 비용효율적인 작은 규모의 훈련에 적합	
Inf 패밀리	AWS가 설계 · 제작한 고성능 머신러닝 추론 칩인 AWS Inferentia칩 탑재 인스턴스	



Amazon Inf1 인스턴스는 AWS Inferentia 칩을
사용하여 구축됨

5) Amazon SageMaker 인스턴스 타입(유형)



세종사이버대학교

 SageMaker는 다음과 같은 ML Compute 인스턴스를 제공

EIA 패밀리

Amazon Elastic Inference용으로
사용되는 추론 가속화 인스턴스



Amazon Elastic Inference

EC2, SageMaker 인스턴스 혹은 ECS 태스크에
추가 부착되는 것으로 비용효율적인 추론 가속 환경
구성을 지원

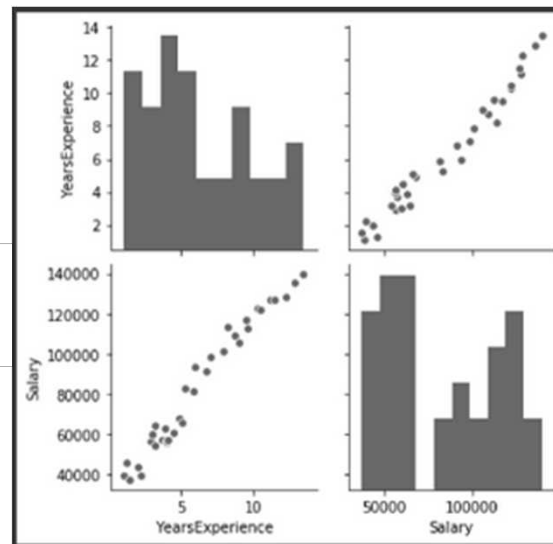
6) ML 라이프사이클 : 빌드(Build)

Amazon SageMaker 노트북 인스턴스 혹은 Amazon SageMaker Studio에서 모델 개발(Build)

예 데이터 탐색 및 시각화
대량의 데이터를 시각적으로
처리하면 머리로 쉽게 이해

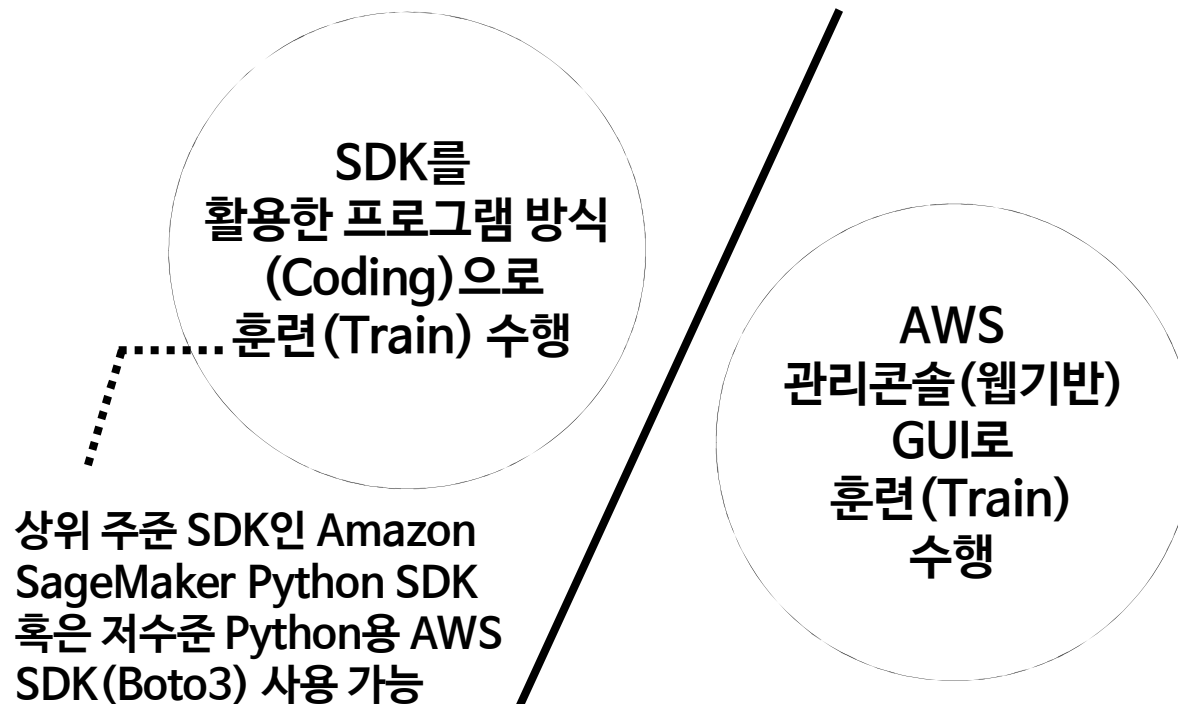
경력연수와 연봉은
양의 상관관계

영상보기



7) ML 라이프사이클 : 훈련(Train)

다음 두 가지 방법 중 하나로 훈련(Train) 수행 가능



7) ML 라이프사이클 : 훈련(Train)

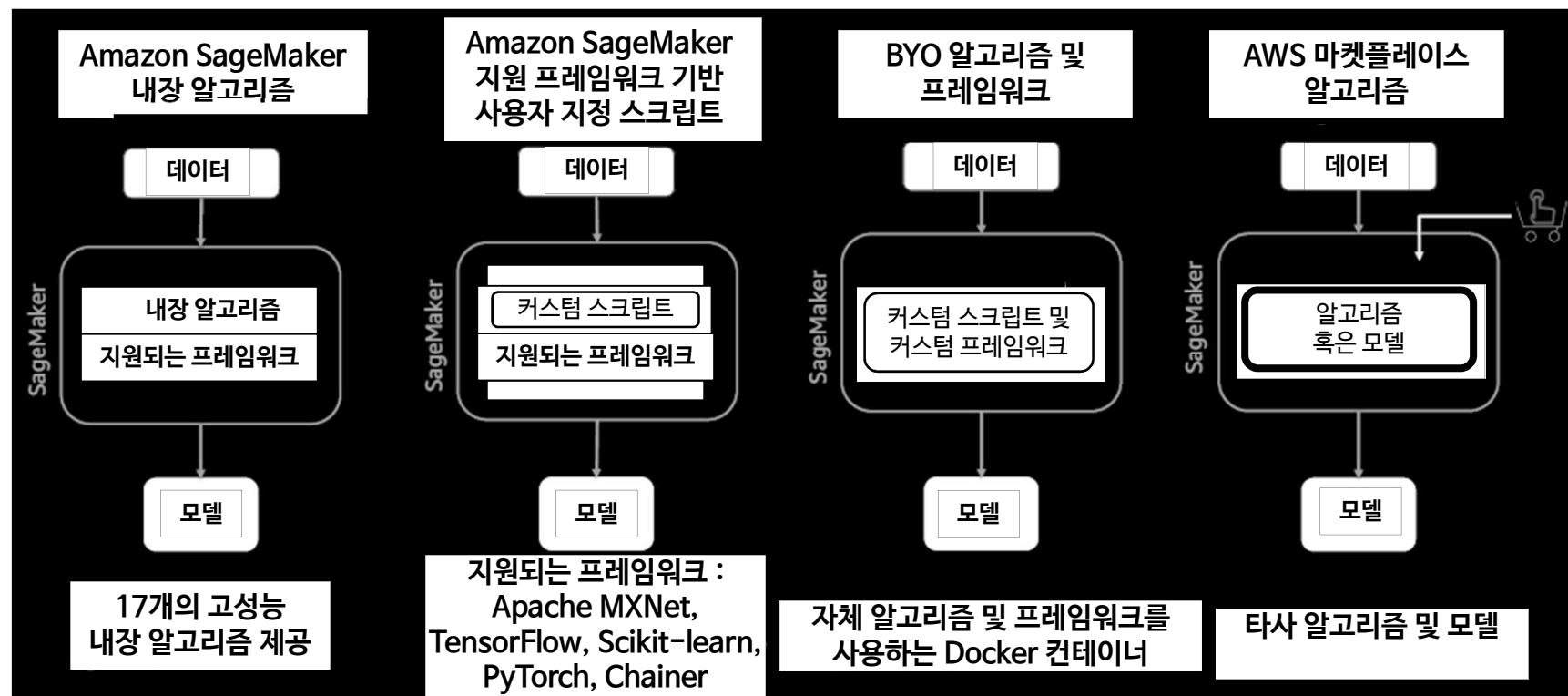
Amazon SageMaker 훈련 인스턴스에서의 훈련 알고리즘 지정 선택 옵션

추론 모델 생성을 위한 훈련 알고리즘을 지정 시
다음과 같은 네 가지 방법 중 하나로 가능

- SageMaker 내장 알고리즘 사용
(2020년 10월 기준 17가지 제공)
- SageMaker 지원 ML 프레임워크기반으로 사용자 스크립트 활용
- 자체 ML 프레임워크 기반 알고리즘 활용
(BYO 알고리즘 및 프레임워크)
- AWS Marketplace로 부터 알고리즘 받아오기(유료 혹은 무료)

7) ML 라이프사이클 : 훈련 (Train)

Amazon SageMaker 훈련 인스턴스에서의 훈련 알고리즘 지정 선택 옵션



7) ML 라이프사이클 : 훈련(Train)

Amazon SageMaker 내장 알고리즘 선택 - 어떤 문제를 해결하는가?

* = distributed training(분산 훈련), <> = incremental training(점진적 훈련)

분류(Classification) <ul style="list-style-type: none">Linear Learner*XGBoostKNN	컴퓨터비전 <ul style="list-style-type: none">Image Classification <>Object Detection <>Semantic Segmentation	주제모델화(Topic Modeling) <ul style="list-style-type: none">LDANTM
텍스트 처리 <ul style="list-style-type: none">BlazingText<ul style="list-style-type: none">SupervisedUnsupervised*	추천(Recommendation) <ul style="list-style-type: none">Factorization Machines*	예측(Forecasting) <ul style="list-style-type: none">DeepAR*
기계번역 <ul style="list-style-type: none">Seq2Seq*	이상탐지 <ul style="list-style-type: none">Random Cut Forests*IP Insight*s	군집화(Clustering) <ul style="list-style-type: none">Kmeans*
	회귀(Regression) <ul style="list-style-type: none">Linear LearnerXGBoostKNN	피처 축소(Feature reduction) <ul style="list-style-type: none">PCAObject2Vec

7) ML 라이프사이클 : 훈련(Train)

도커(Docker)기반 아키텍처 활용

도커컨테이너 저장소인 Amazon ECR내 등록된
훈련 이미지(Train Image)가 도커컨테이너에서 실행



지정된 S3 버킷 내 훈련 데이터를 로드 후 훈련 수행



훈련 작업(training job) 결과인 모델 아티팩트는
지정된 S3 버킷에 저장됨

7) ML 라이프사이클 : 훈련(Train)

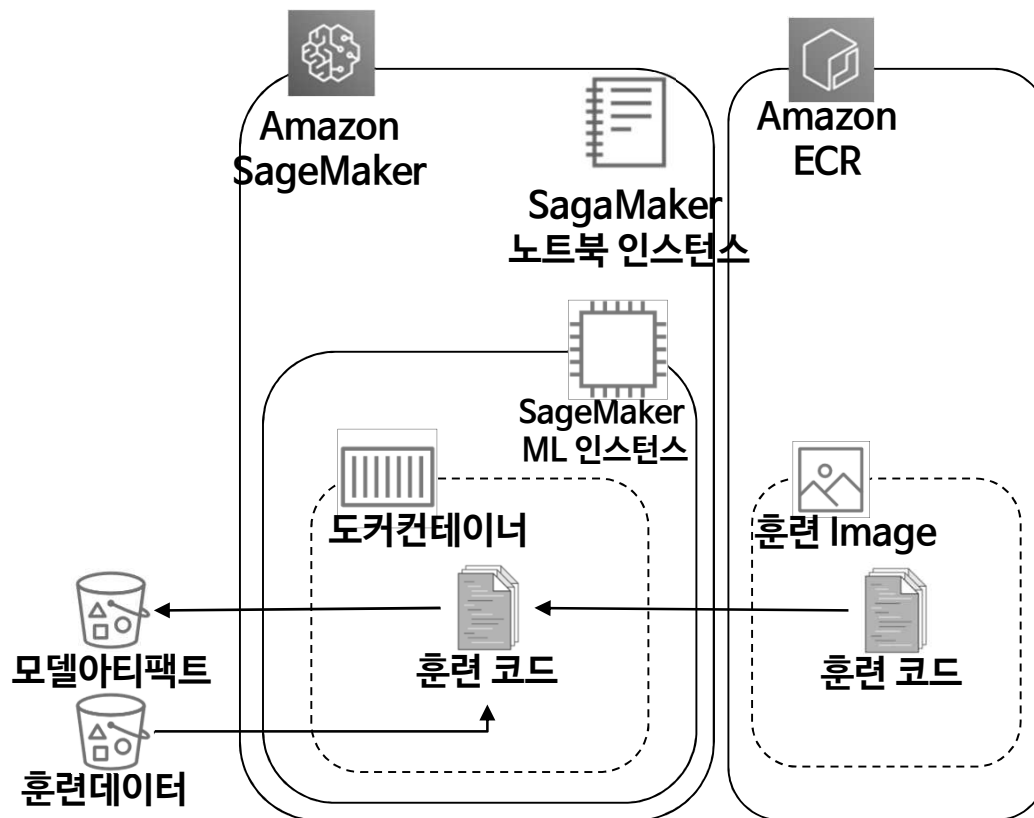


세종사이버대학교

도커(Docker)기반 아키텍처 활용

- a 훈련데이터셋 준비
- b 훈련코드 준비
- c 훈련환경 구성
- d 훈련 수행
- e 훈련이 완료된 모델 저장

영상보기



8) ML 라이프사이클 : 추론 (Inference)

도커(Docker)기반 아키텍처 활용

도커컨테이너 저장소인 Amazon ECR내 등록된 추론 이미지(Inference Image)가 도커컨테이너에서 실행



지정된 S3 버킷내 모델 아티팩트를 읽어 들여 호스팅

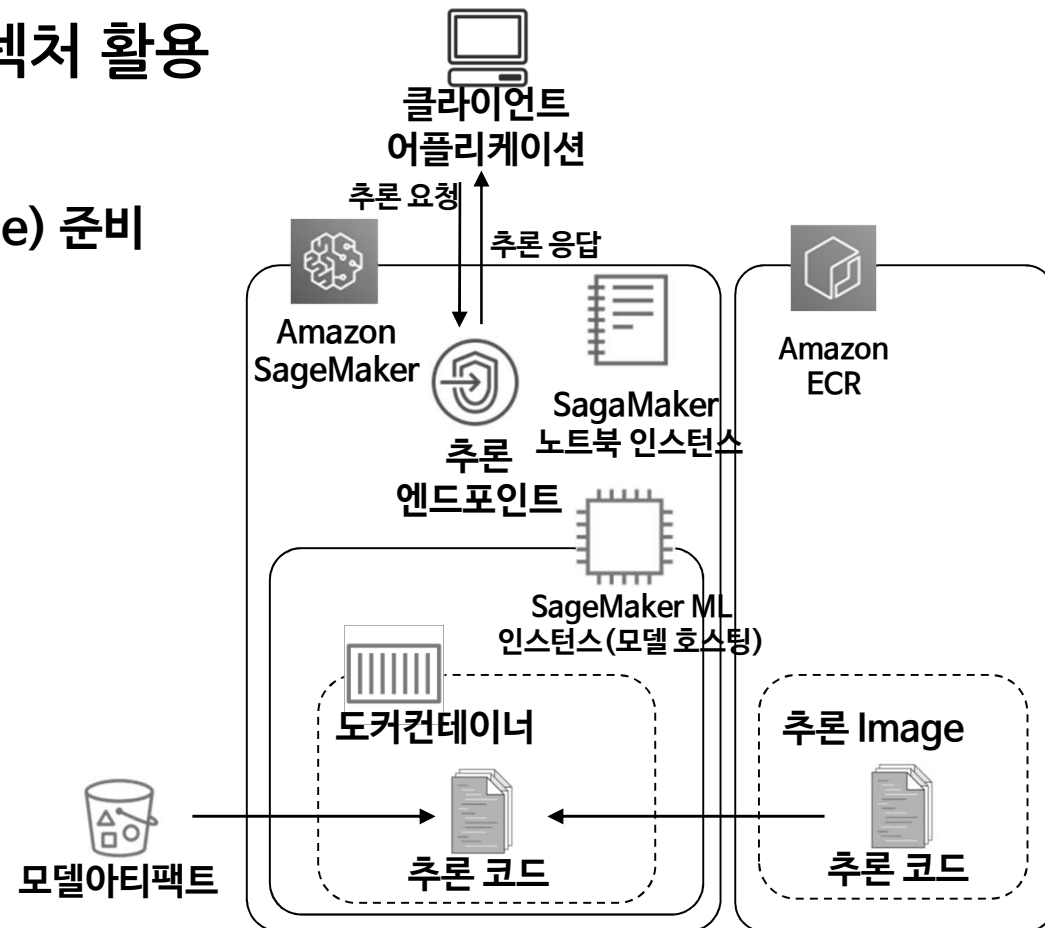


추론 엔드포인트를 통한 API 서비스를 제공

8) ML 라이프사이클 : 추론 (Inference)

도커(Docker)기반 아키텍처 활용

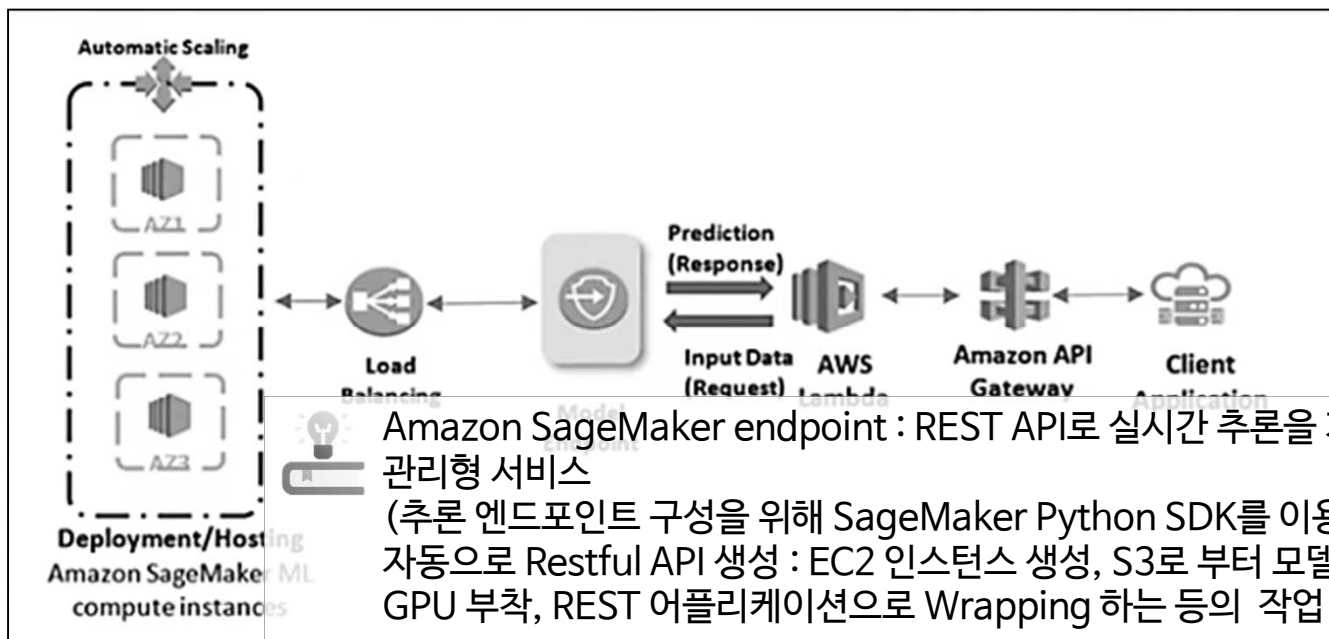
- ① 추론 코드 (Inference Code) 준비
- ② 추론 환경 구성
- ③ 모델 호스팅



8) ML 라이프사이클 : 추론 (Inference)

Amazon SageMaker Endpoint

“확장성 있는 운영 환경 구축을 위한 머신러닝 모델 배포”



8) ML 라이프사이클 : 추론 (Inference)

Amazon SageMaker Endpoint

온라인 추론 유즈케이스에 주로 사용되며, 이 경우 인터넷 기반으로 추론 요청에 대해 실시간으로 결과 제공

Sagemaker Python SDK로 Model Deploy 관련 1줄 Code 실행으로 관리형 엔드포인트가 생성

- SageMaker는 배포 요청된 모델에 대해 자동으로 Restful API를 생성
- 모델 엔드포인트는 RESTFUL API로, 다른 Restful API처럼 JSON으로 요청을 받아 JSON으로 결과 제공
- 일반적으로 API Gateway와 Lambda를 사용하여 온라인 추론 요청에 대응
 - 이 때 Lambda와 같이 사용하지 않을 때는 엔드포인트를 Off하는 것을 추천
- 온프레미스에서 머신러닝 모델 훈련 후 SageMaker에서 엔드포인트로 호스팅하는 것도 가능

8) ML 라이프사이클 : 추론 (Inference)

Amazon SageMaker Endpoint

Preprocessing 및 Post Processing을 위한 추론 파이프라인 활용

내부 구현 구조

- 엔드포인트 뒤에서 로드밸런서는 각 인스턴스에 대한 Health Check 수행하며 각 인스턴스는 요청에 대해 예측 결과를 제공
- 고가용성을 위해 다수 가용영역에 걸친 다수 EC2 인스턴스로 구성되며 각 EC2 인스턴스에는 요청에 대한 결과를 제공하는 웹서버와 모델아티팩트가 존재

9) Amazon Sagemaker 가격 모델 및 비용 최적화 방법



세종사이버대학교

가격 모델

사용한 만큼만 비용을 지불

- ML 컴퓨팅 파워에 대해 초단위로(최소 1분) 요금 지불

Amazon SageMaker 내 요금

- ML 인스턴스
- ML 스토리지
- 인스턴스에서의 데이터처리 비용

9) Amazon Sagemaker 가격 모델 및 비용 최적화 방법



세종사이버대학교

가격 모델

모델 훈련
(Training)
및 배치 변환
(Batch
Transform)
작업의 경우

Amazon SageMaker는 작업이
완료된 후 인스턴스를 자동으로 종료
작업 실행 시간에 대해서만 요금 청구

9) Amazon Sagemaker 가격 모델 및 비용 최적화 방법



세종사이버대학교

모델 훈련(Training)시 비용 최적화

작은 규모의 데이터셋에서도 성능이 잘 나오는 알고리즘을 선택

적합한 사이즈의 훈련 인스턴스 선택

- 작은 인스턴스로 시작하고 스케일아웃을 먼저 시도 후 스케일업 시도
- 관리형 스팟 훈련 (Managed Spot Training) 고려

<https://aws.amazon.com/ko/blogs/korea/managed-spot-training-save-up-to-90-on-your-amazon-sagemaker-training-jobs/>

9) Amazon Sagemaker 가격 모델 및 비용 최적화 방법



세종사이버대학교

모델 배포(Deployment)

모델 배포로 엔드포인트 생성시 해당 호스팅 즉 추론 인스턴스 생성됨

머신러닝 프로세스에서 실시간 추론(Inference)를 위한 모델은 24/7 동작으로 가장 많은 비용 소요

- 딥러닝 어플리케이션에서 추론은 총 운영 비용에서 최대 90% 차지

9) Amazon Sagemaker 가격 모델 및 비용 최적화 방법



세종사이버대학교

모델 배포(Deployment)

코드 변경 없이 낮은 가격으로 소량의 GPU 지원 추론 가속(Amazon Elastic Inference)을 연결하여 머신러닝 추론 비용 최대 75% 절감

- 일반적으로 CPU 인스턴스는 딥러닝 추론 수행에 속도가 느리며 이를 개선하기 위해 추론 수행 시 GPU 인스턴스를 사용하기에는 비용효율적이지 않음
- TensorFlow, Apache MXNet, PyTorch 및 ONNX(Open Neural Network Exchange) 모델을 지원하는 Amazon Elastic Inference 활용

9) Amazon Sagemaker 가격 모델 및 비용 최적화 방법



세종사이버대학교

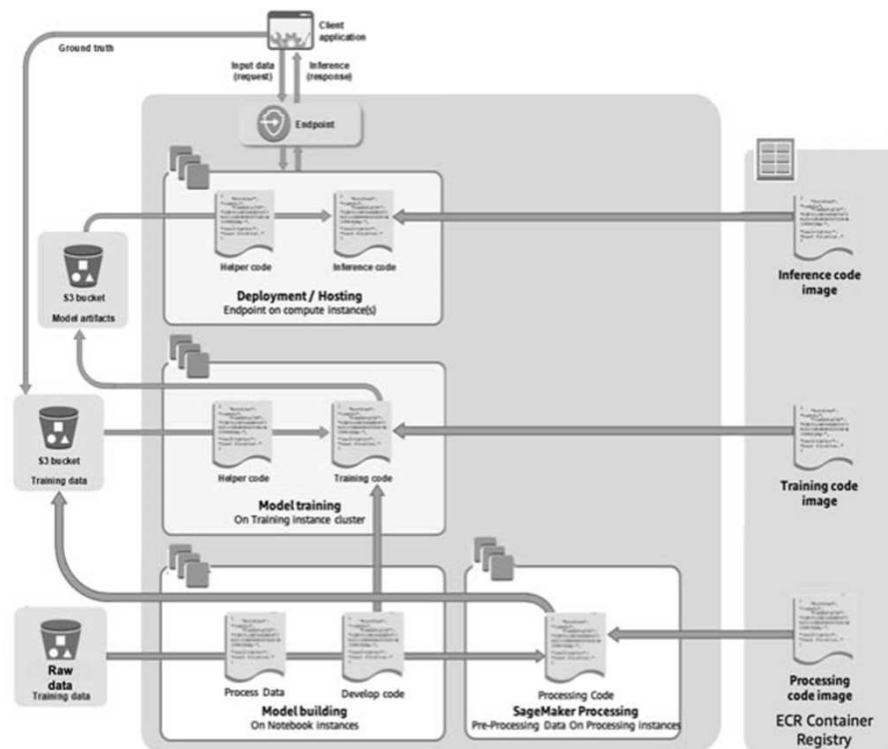
📍 빌드, 훈련, 배포 환경을 위해 서로 다른 컴퓨팅 자원 사용

▶ 도커(Docker) 기반 라이프사이클 관리

이후 실시간 추론을 위해
모델을 운영환경으로 배포
(혹은 배치 추론을 위해
Batch Transform을 활용)

↑
이후 모델을 훈련 및 튜닝

↑
모델 빌드 시작을 위해
원시 데이터 처리하여
훈련 데이터를 생성





02

ML 연구자를 위한 ML 프레임워크 및 인프라



1) 폭넓은 딥러닝 프레임워크 지원

딥러닝 컨테이너(AWS DL Containers)

딥러닝 프레임워크가 사전에 설치된 Docker 이미지

	DL 관련 소프트웨어 종속성과 버전 호환성 문제에 대한 부담 제거
	Tensorflow, Apache Mxnet, PyTorch 지원
	Amazon SageMaker, Amazon EKS, Amazon ECS 및 Amazon EC2 자체 관리형 Kubernetes에 배포 가능
	AWS DL 컨테이너는 Amazon ECR 및 마켓플레이스를 통해 무료로 제공

- 사용 자원에 대해서만 요금 지불

1) 폭넓은 딥러닝 프레임워크 지원

딥러닝 AMI

Amazon Linux 및 Ubuntu용으로 개발

TensorFlow, PyTorch, Apache MXNet, Chainer,
Microsoft Cognitive Toolkit, Gluon, Horovod 및 Keras가
사전 구성되어 제공됨



Horovod

우버(Uber)에서 만든 딥러닝 분산 훈련 프레임워크로, 멀티 GPU 및 분산 훈련을 손쉽게 최적화된 형태로 진행할 수 있게 해주는 프레임워크, Tensorflow, Keras, PyTorch, MxNet도 백엔드로 지원

2) 고성능,비용효율적, 확장성 있는 인프라스트럭처



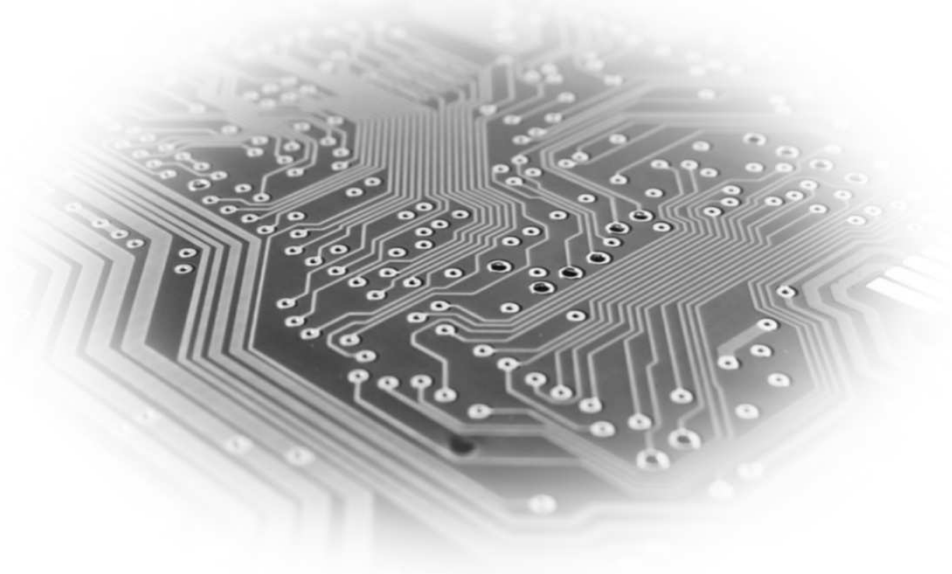
세종사이버대학교

AWS Inferencia

“

AWS가 설계 · 제작한
고성능 머신러닝 추론 칩

”



2) 고성능,비용효율적, 확장성 있는 인프라스트럭처



세종사이버대학교



Amazon Elastic Inference

TensorFlow, Apache MXNet, PyTorch 및
ONNX(Open Neural Network Exchange) 모델을 지원

EC2, SageMaker 인스턴스 혹은 ECS 태스크에
추가 부착되는 것으로 비용효율적 추론 가속 환경 구성

코드 변경 없이 낮은 가격으로 소량의 GPU 지원 추론 가속을
연결하여 ML 추론 비용 최대 75% 절감