AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

aws INNOVATE
ONLINE CONFERENCE

分会场四：人工智能与机器学习

Amazon SageMaker 平台让企业构建 AI 能力 So Easy

李南山，AWS 解决方案架构师

人工智能与机器学习正在影响各行各业

媒体与娱乐

- 内容生成
- 促销与市场活动
- 版权保护
- 内容分类与标签
- 字幕自动生成
- …

健康与医疗

- 药物发现与探索
- 医学影像自动识别
- 辅助医疗
- 自动诊疗
- …

个性化推荐

- 在线电商
- 产品推荐
- 信用评级
- 广告与搜索相关
- 个性化
- …

金融服务交易

- 服务管理与推荐
- 交易算法
- 舆情与新闻分析
- 潜在用户发现与推荐
- …

客户体验

- 语音服务
- 聊天机器人
- 呼叫中心优化
- 个人服务中心管理
- …

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

众多客户在 AWS 上构建机器学习应用

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

我们的愿景

AI @ AWS

让每位开发者和数据工作者可以方便的使用机器学习。

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS 机器学习堆栈

AI 服务接口

计算机视觉

AMAZON REKOGNITION IMAGE

AMAZON REKOGNITION VIDEO

语音

AMAZON POLLY

AMAZON TRANSCRIBE

语言

AMAZON TRANSLATE

AMAZON COMPREHEND

对话机器人

AMAZON LEX

ML 平台服务

AMAZON SAGEMAKER

ML 框架 & 基础设施

Frameworks

TensorFlow　mxnet　PYTORCH　Chainer　HOROVOD

Interfaces

GLUON　Keras

Infrastructure

AMAZON EC2 P3 Instances　AMAZON EC2 C5 Instances　FPGAs

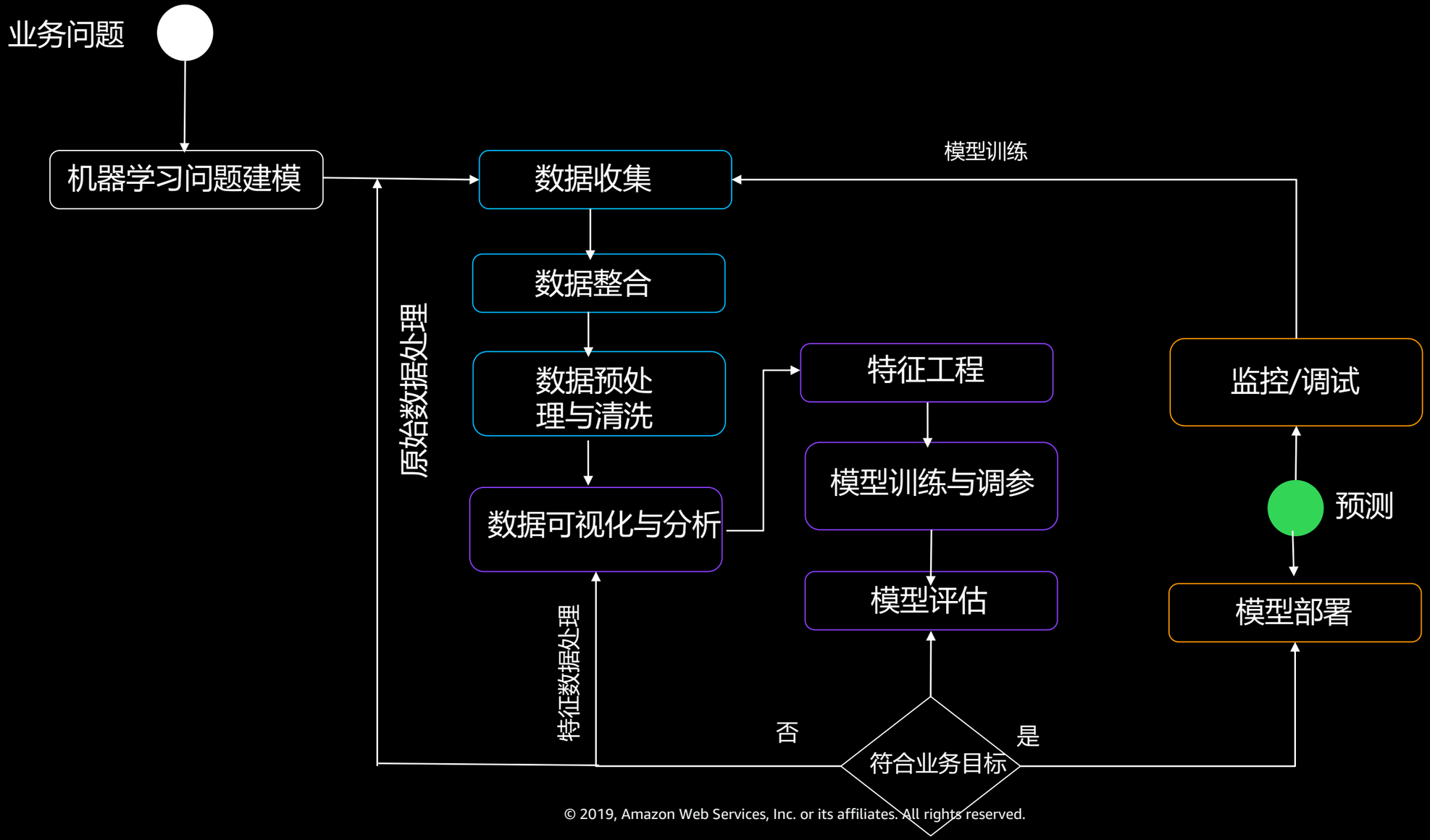
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker 功能与架构

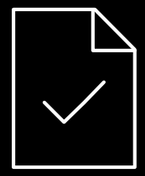
aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

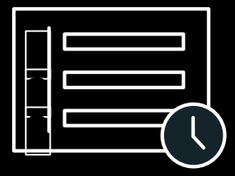
机器学习的流程



对于开发人员机器学习技术复杂



准备训练数据



选择&优化机器学习算法

搭建和管理训练环境

训练&优化模型

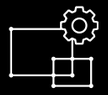
在生产环境部署模型

扩展&管理生产环境

Amazon SageMaker

1 构建

快速&精准打标签



内置高性能机器学习算法&notebooks

- Notebook 实例
- 从你的设备调用 APIs

2 训练

一键训练和调优

模型优化

- 全托管的
- 分布式
- 高性能 I/O

3 部署

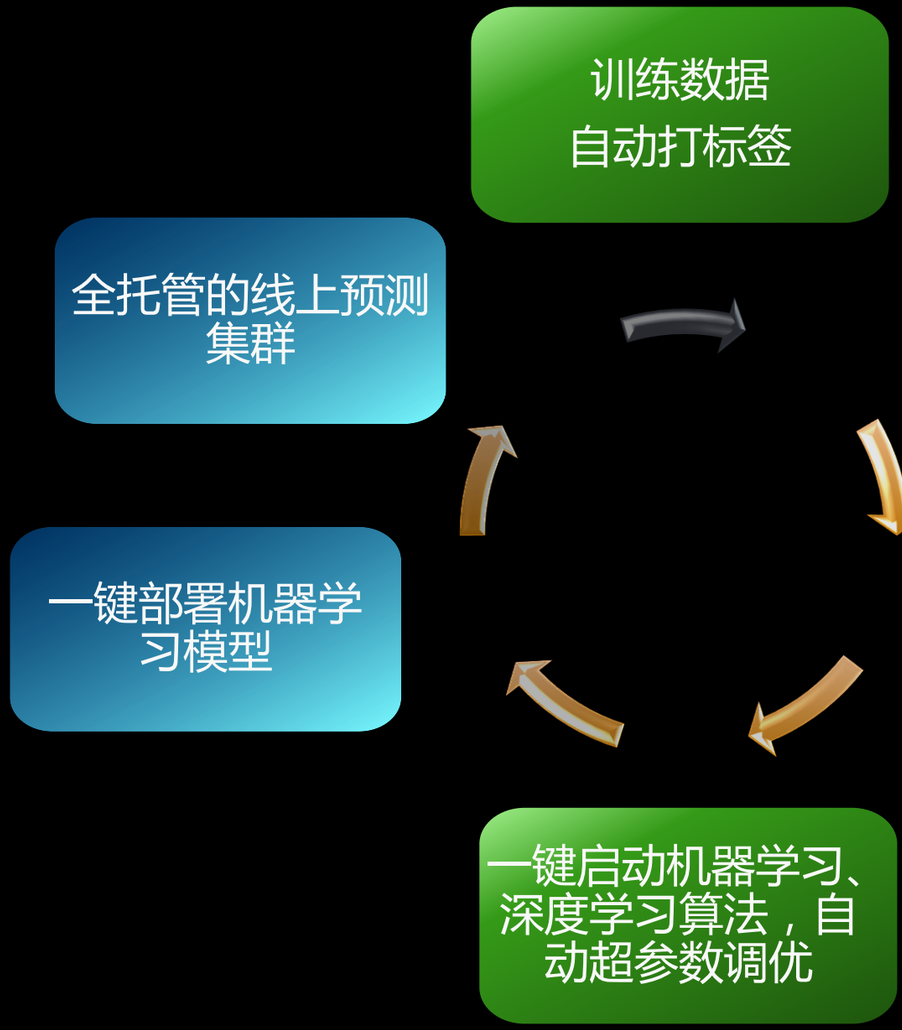
一键部署

全托管的自动扩展&弹性推理预测

- 实时终端节点分析
- 批量分析
- AWS IoT Greengrass
- AWS DeepLens

TensorFlow mxnet PYTORCH Chainer HOROVOD GLUON Keras

Amazon SageMaker 端到端机器学习平台



部署模型

构建模型

mxnet

TensorFlow

Train

Scikit-learn

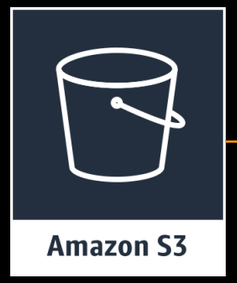
PYTORCH

GLUON

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker GroundTruth

使用自动标签 降低 标签成本高达 70%



原始资料

学习模组
Active
Learning
model

>80% 可信度

自动注释

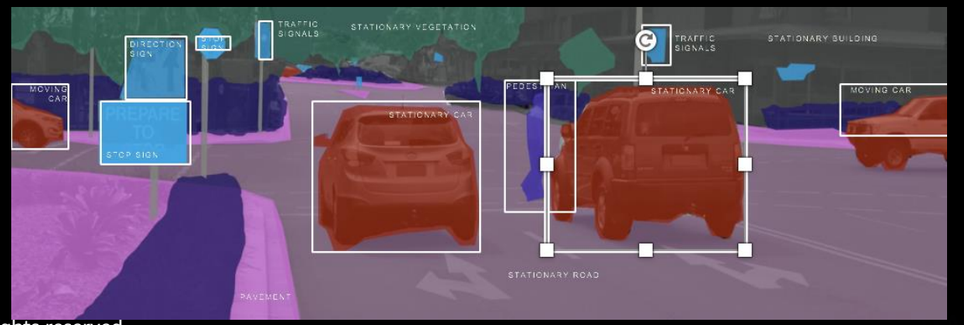
<80% 可信度

人工注释

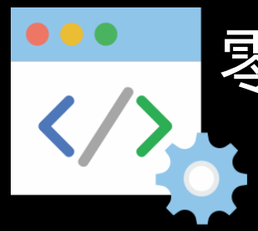
User

第三方业者
自行处里
简易建好的流程

训练资料

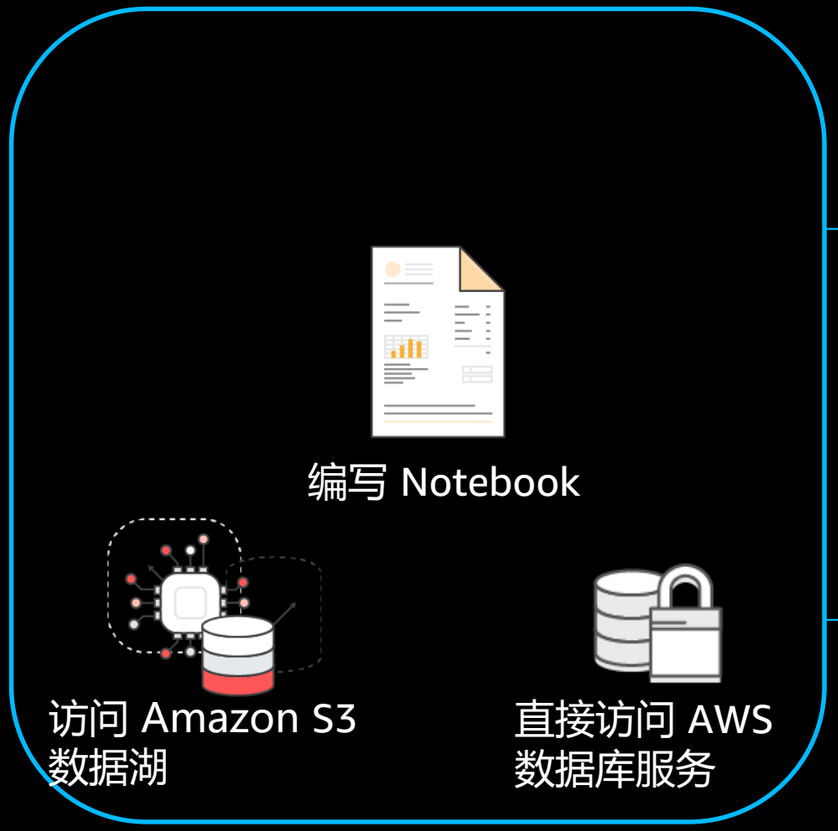


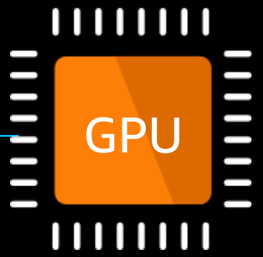
Amazon SageMaker NoteBook



零配置进行数据分析与探索

NoteBook 计算环境
预装 Conda





只需要添加数据

- 推荐与个性化算法
- 欺诈分析
- 预测
- 图片分类
- 客户趋势分析
- 市场与商品推广
- 日志处理与分析
- 语音文字转化
- 更多...

Amazon SageMaker：优化的算法

流式数据注入

更快的训练过程

基于大数据量的可靠训练模型与算法

十几种经典机器学习算法

Amazon SageMaker 内置算法

有监督学习	无监督学习
Linear learner 线性分类器: 回归, 分类	K-means 聚类
Factorization machines 因子分解机: 回归, 分类, 推荐	Principal component analysis (PCA) : 主成分分析，降维
K-nearest neighbors K 近邻: 非参数化回归, 分类	Random Cut Forest 随机森林砍伐: 异常检测
XGBoost: 回归, 分类, 排序	IP Insights: IP 异常检测
Image classification 图像分类: 深度学习 (ResNet)	Neural Topic Model 神经网络主题模型: 主题分类
Object detection 对象检测: 深度学习 (VGG or ResNet)	Latent Dirichlet allocation 隐含狄利克雷分布 ：主题分类(大多数)
Sequence to sequence: 机器翻译, 语音转文字	BlazingText: 基于 GPU 的 Word2Vec 和文本分类
DeepAR: 基于时序数据的预测 (RNN)	Object2Vec: 通用神经网络嵌入算法，学习高维对象的低维密集嵌入，文本-文本，标签-序列
Semantic Segmentation语义分割: 计算机视觉物体分割	

持续增加中

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

机器学习算法选择

分类

- Linear Learner
- XGBoost
- Factorization Machines
- SVMs (Spark, BYO)

回归

- Linear Learner
- XGBoost
- Factorization Machines
- SVMs (Spark, BYO)

推荐

- Factorization Machines
- Collaborative Filtering (Spark)
- Matrix Factorization (BYO)

聚类

- K-Means
- GMMs (BYO)
- DBScan (BYO)

预测

- DeepAR
- Linear Learner
- XGBoost
- Prophet (BYO)
- ARIMA (BYO)
- EST (BYO)

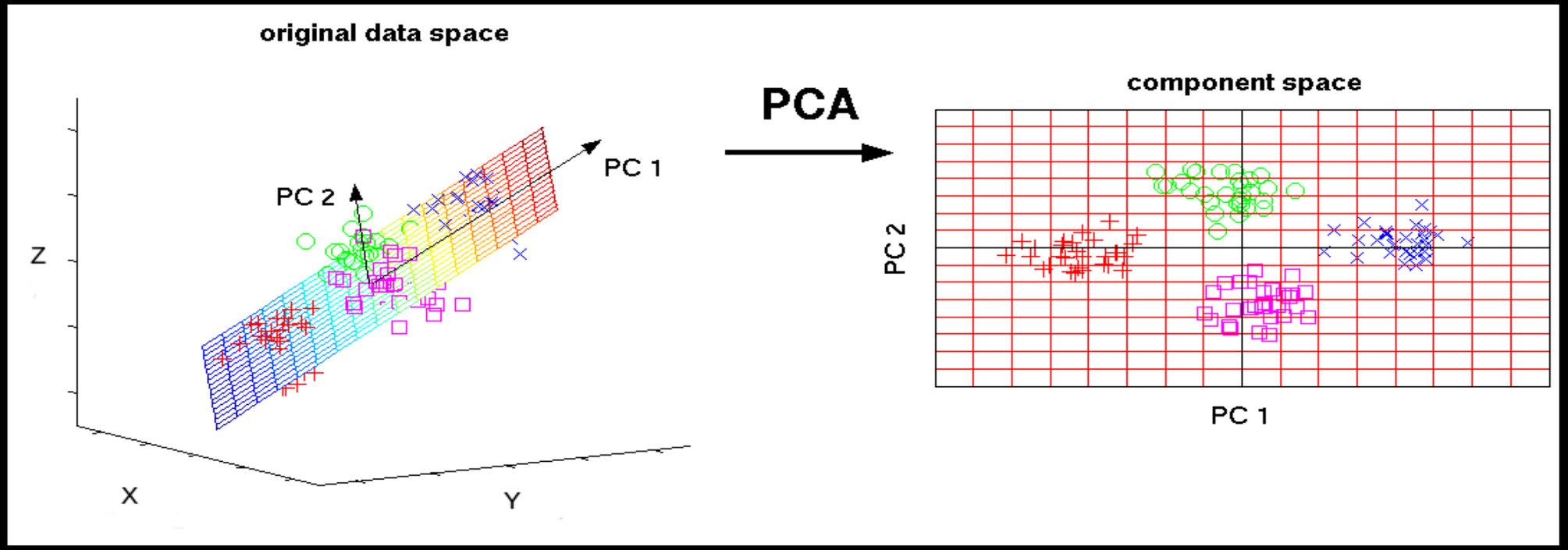
降维/异常检测

- PCA
- Random Cut Forest (Kinesis Analytics)
- t-SNE (BYO)
- Manifold Learning (BYO)
- Autoencoders (BYO)
- SVMs (Spark, BYO)

Amazon SageMaker 机器学习算法介绍（部分）

主成分分析 (PCA)

- 数据降维（降低特征的数量）
- 将特征映射到具体的成分

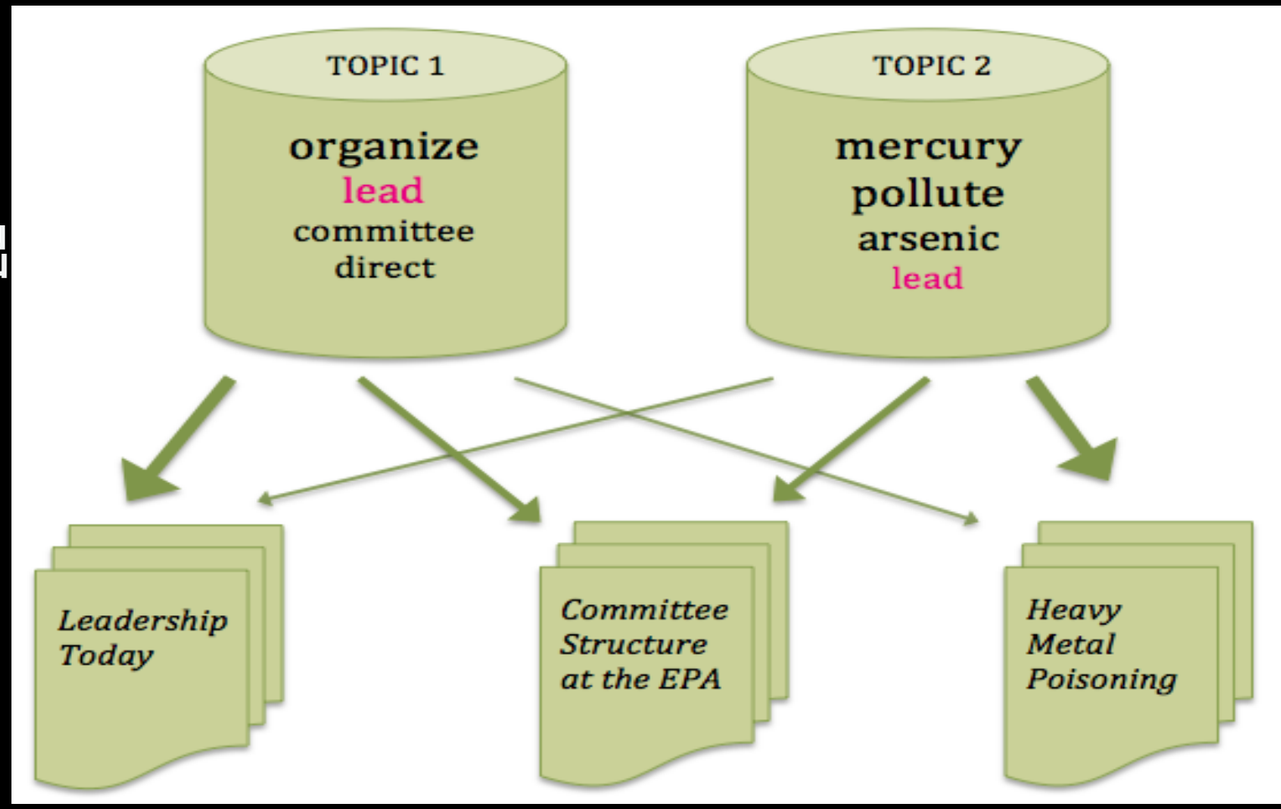


主成分分析 (PCA)

- 使用场景
  - 数据压缩
  - 图像处理
  - 探索性数据分析
  - 高维数据模式识别
  - 金融，生物信息学，心理学，数据挖掘
- PCA 支持 GPU 和 CPU 计算

隐含狄利克雷分布 (LDA)

- 在文本语料库中，发现文档中的主题
  - 每次输入都是一份文档
  - 特征是每个单词是否存在（或出现个数）
  - 对文档的分类为该文档的主题
- 主题通过对每个文档中出现的单词的概率分布进行机器学习
- 每个文档最终被描述为一些主题的集合



https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

隐含狄利克雷分布 (LDA)

- 使用场景
  - 使用相似性和相关性对文档进行分类和组织
  - 文档摘要
  - 从大型数据集中发现潜在的语义主题，无论是文本、图像、还是音乐记录
- LDA 当前支持单实例 CPU 训练

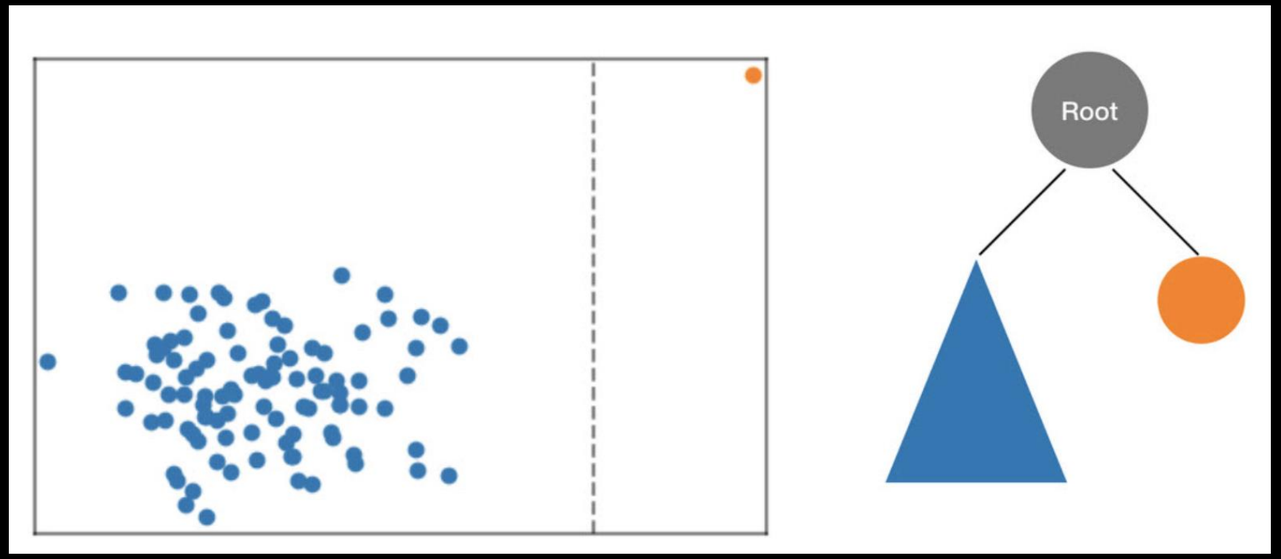
Neural Topic Modelling (NTM)

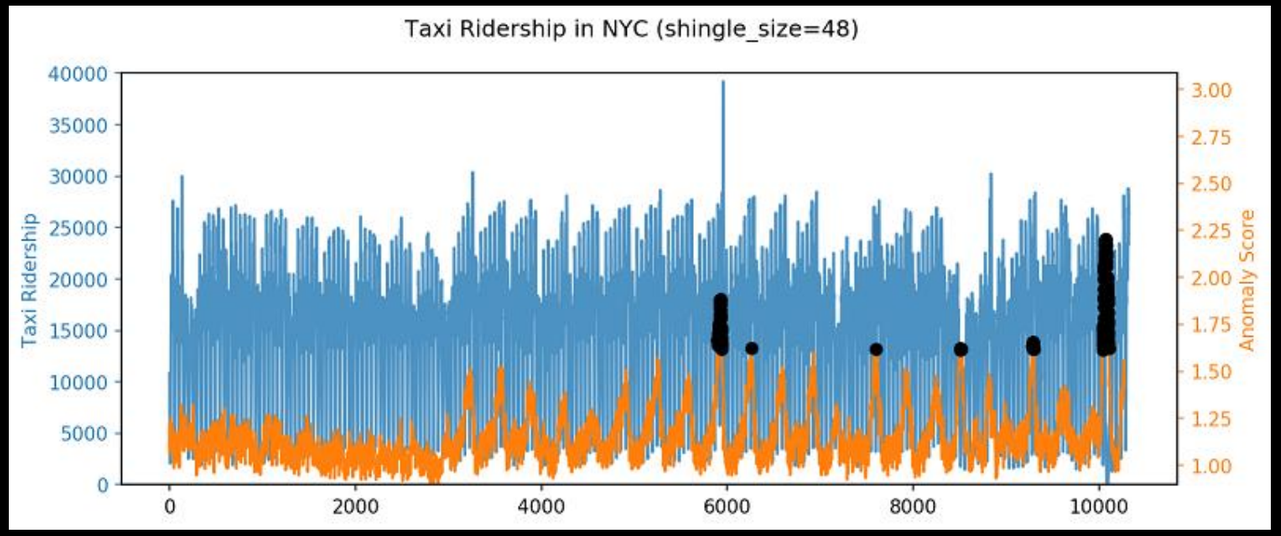
- 在文本语料库中，发现文档中的主题
- NTM 支持 GPU 和 CPU 实例类型
- LDA vs NTM
    - 两种不同的算法会在同一数据集上产生不同的结果
    - NTM 通常具有较低的混淆度
    - LDA 在少数主题上训练非常快，但不像 NTM 那样扩展到更多主题

结论：如果使用场景中需要判断很多主题和更好的“合适度（fit）”，则使用 NTM，否则使用 LDA。

RCF 随机森林砍伐

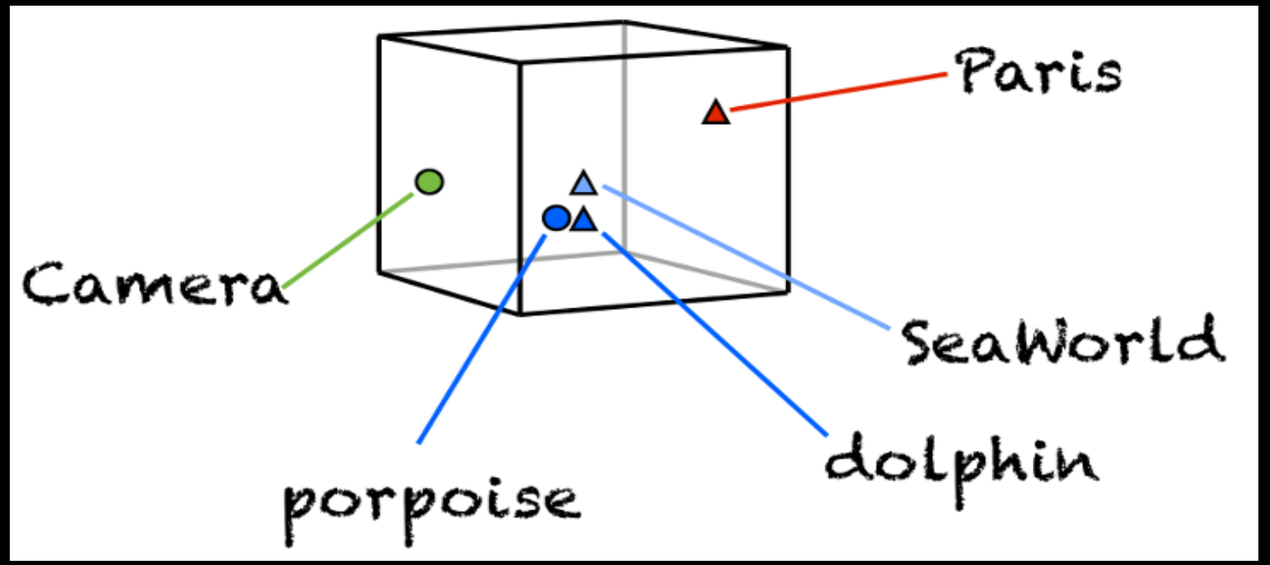
- 计算数据的边界，选择随机维 ，然后通过该维随机确定“砍伐”的位置，从而将这些数据组织到树中。
- 给每个数据点一个异常分数值，可用于时间序列数据
- 用于识别潜在欺诈行为，网络攻击，服务过期等异常现象
- RCF 支持 GPU 和 CPU 实例类型





BlazingText

- 生成 Word2Vec
- 生成文档中各单词的矢量表示
- 获取其中的意义，单词和上下文之间的语义关系

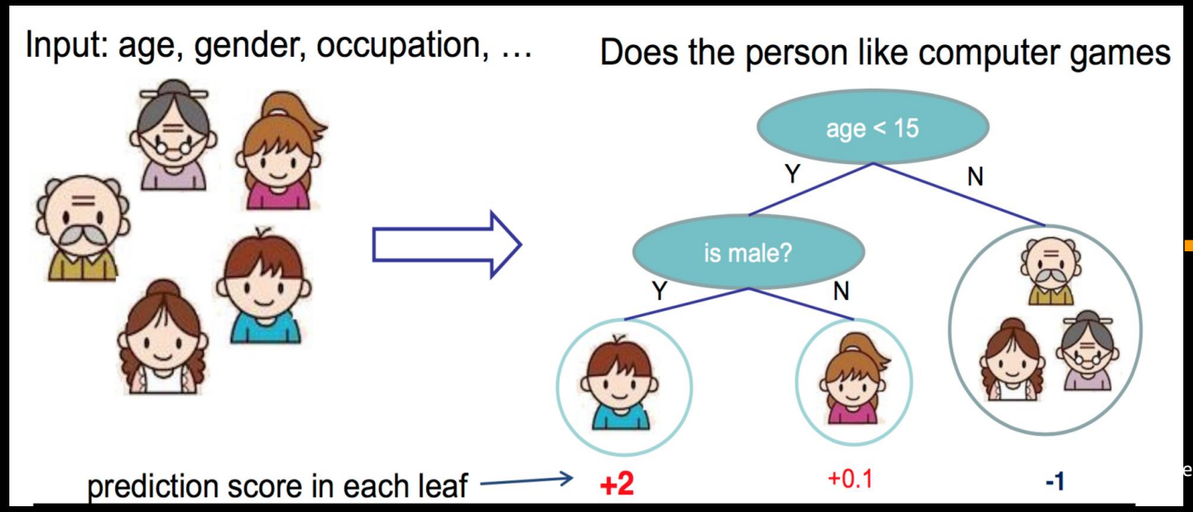


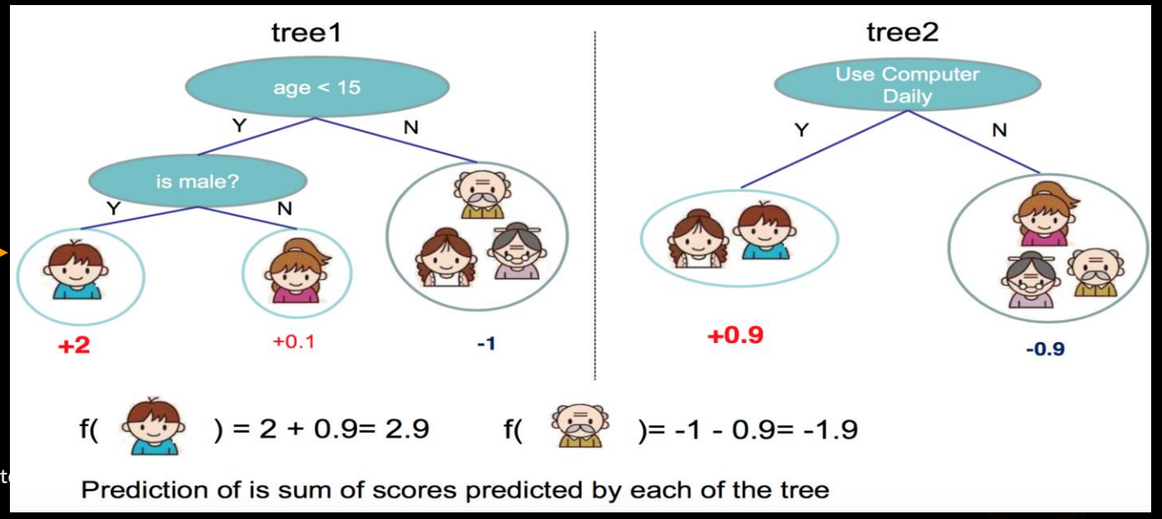
BlazingText

- 用于自然语言处理 (NLP)
  - 情绪分析
    - 更好的了解客户
    - 确定产品趋势
  - 机器翻译
    - 为网站提供多语言支持
  - 命名实体识别
    - 从文字中获取组织与主要参与者信息
- BlazingText 支持单个 CPU 实例和单个 GPU 实例

XGBoost

- Extreme Gradient Boosting
  - 基于 Gradient Boosting 决策树算法（GBDT）
  - 通过组合一组更简单，更弱的模型，把它们的预测结果相加来预测目标变量
  - 分类，回归，排行
  - 支持 CPU





因子分解机 (Factorization Machines)

- 线性回归的衍生
- 每个特征的个体权重 (weight) 与 k 维向量 (vector) 的比较
- 支持 CPU 和 GPU

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
User 1	0	3	0	3	0	0
User 2	4	0	0	2	0	0
User 3	0	0	3	0	0	5
User 4	0	0	0	0	3	0
User 5	4	0	0	4	0	0

因子分解机何时使用

图像分类

- 将图像分类为多个类别中的一类
- ResNet
  - 非常深的网络（默认为152层）
- 两种使用模式
  - 全量学习 (从随机参数开始训练，需要大量数据，结果准确)
  - 迁移学习 (利用公开成熟的模型，替换最后的一层或几层全联通层，不需要很多数据也能训练)

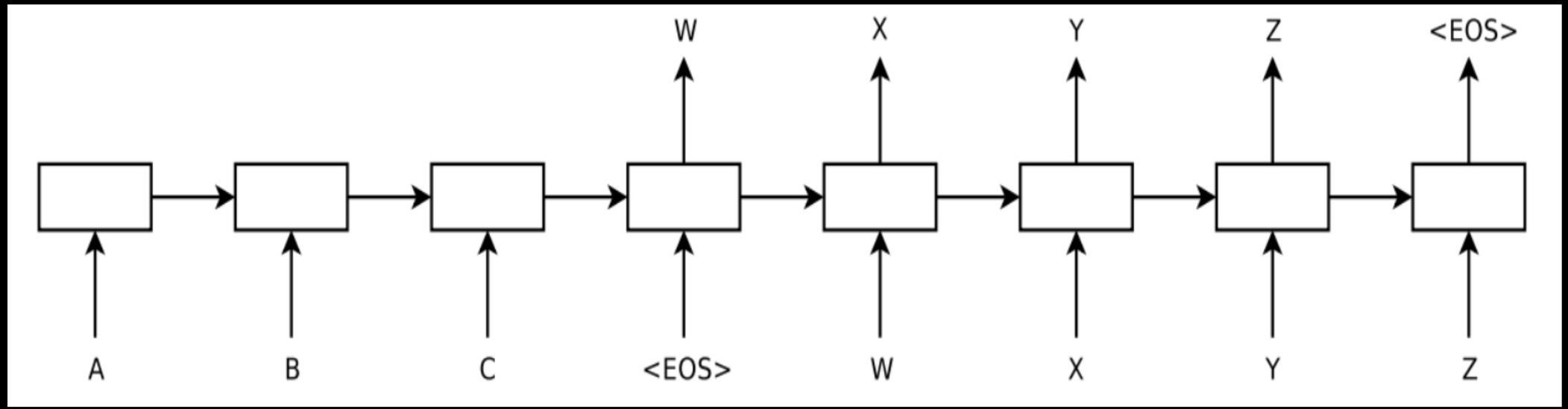
Amazon SageMaker 图像训练示例

- Mxnet 框架
- ResNet 深度卷积神经网络
- caltech 数据集

Image-classification-fulltraining-elastic-inference.ipynb
Image-classification-fulltraining-highlevel-neo.ipynb
Image-classification-fulltraining-highlevel.ipynb
Image-classification-fulltraining.ipynb
Image-classification-incremental-training-highlevel.ipynb
Image-classification-lst-format-highlevel.ipynb
Image-classification-lst-format.ipynb
Image-classification-multilabel-lst.ipynb

Sequence to Sequence（seq2seq)

- 输入一个序列并获得另一个序列作为输出。
- 编码器和解码器



Sequence to Sequence（seq2seq)

- 应用场景
  - 机器翻译
    - 以一种语言输入一个句子，并预测该句子在另一种语言中的含义
  - 文字摘要
    - 输入较长的单词串，并通过作为摘要的较短的单词串输出
  - 语音转文字
    - 输入一段音频，通过转化输出相应的文字
- 仅支持 GPU

DeepAR

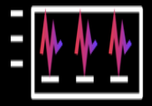
- 时间序列预测
- 亚马逊内部使用的算法
- 训练一组相关的时间序列，以获得更多的见解和更高的预测能力
- 最小化特征引擎
- 预测
  - 值（销量为 x）
  - 概率（出售金额在 x 和 y 之间的概率 z）

DeepAR 何时使用

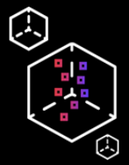
- 预测
  - 产品需求量
    - 供应链优化
  - 服务器负载
  - 网页请求
- 支持 CPU 和 GPU

Amazon SageMaker 强化学习

为开发人员和数据科学家准备的强化学习平台



全托管强化学习算法



通过 OpenGym 进行二维和三维仿真环境

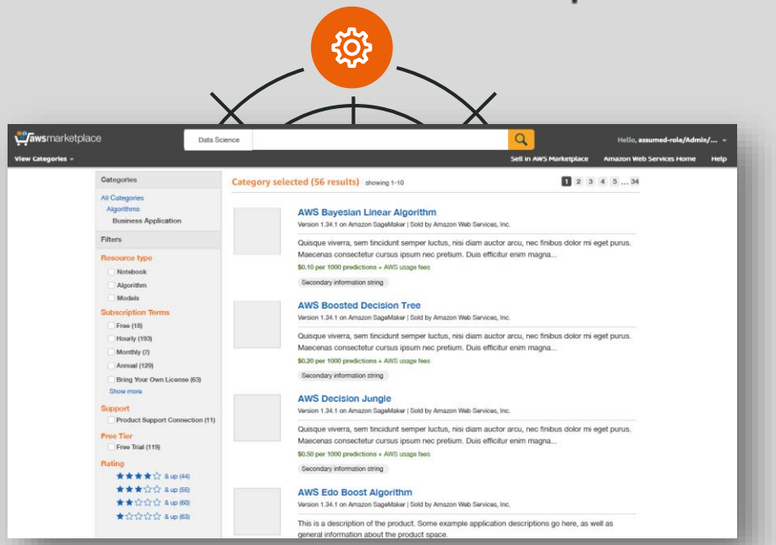
TensorFlow, MXNet, Intel Coach & Ray RL

使用 Amazon Sumerian 和 AWS RoboMaker 模拟环境

样例 notebooks 和教程

AWS 机器学习应用市场

aws marketplace



探索超过 150 种可信的，精准的机器学习算法和模型

53 个机器学习分类

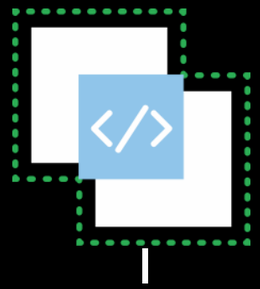
14 个行业

30+ APN 合作伙伴

9 家预览客户

兼具效率与灵活性的分布式托管训练

– 安全



机器学习训练服务

获取训练数据

训练代码

保存模型数据

保存预测镜像

Amazon ECR

- Matrix Factorization
- Regression
- Principal Component Analysis
- K-Means Clustering
- Gradient Boosted Trees
- 和更多算法!

AWS 提供的算法

mxnet

TensorFlow™

利用您已有的算法脚本

APACHE Spark

IM Estimators in Apache Spark

利用Docker封装您定制化的算法

CPU GPU

HPO

全托管服务

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

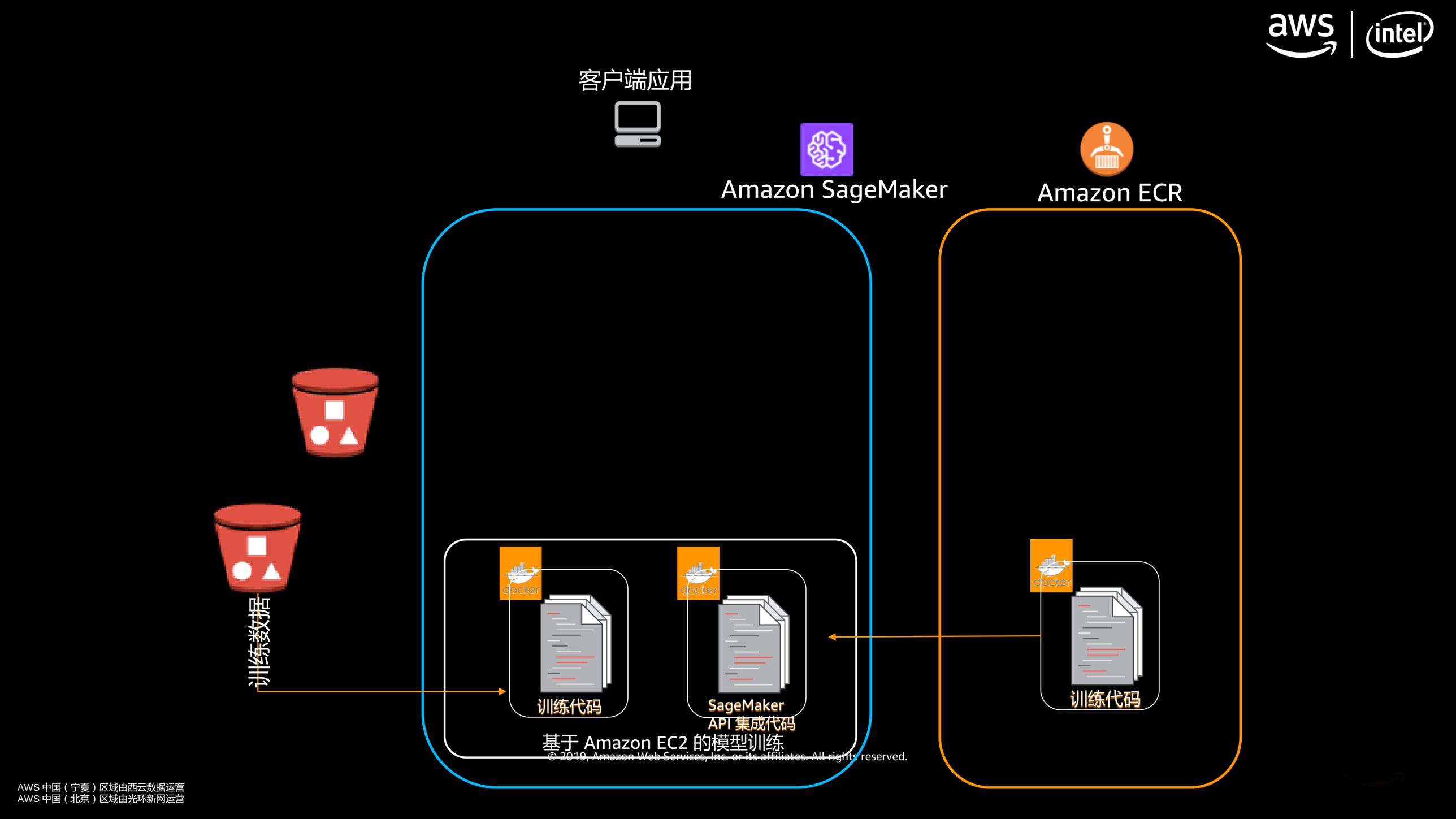
客户端应用

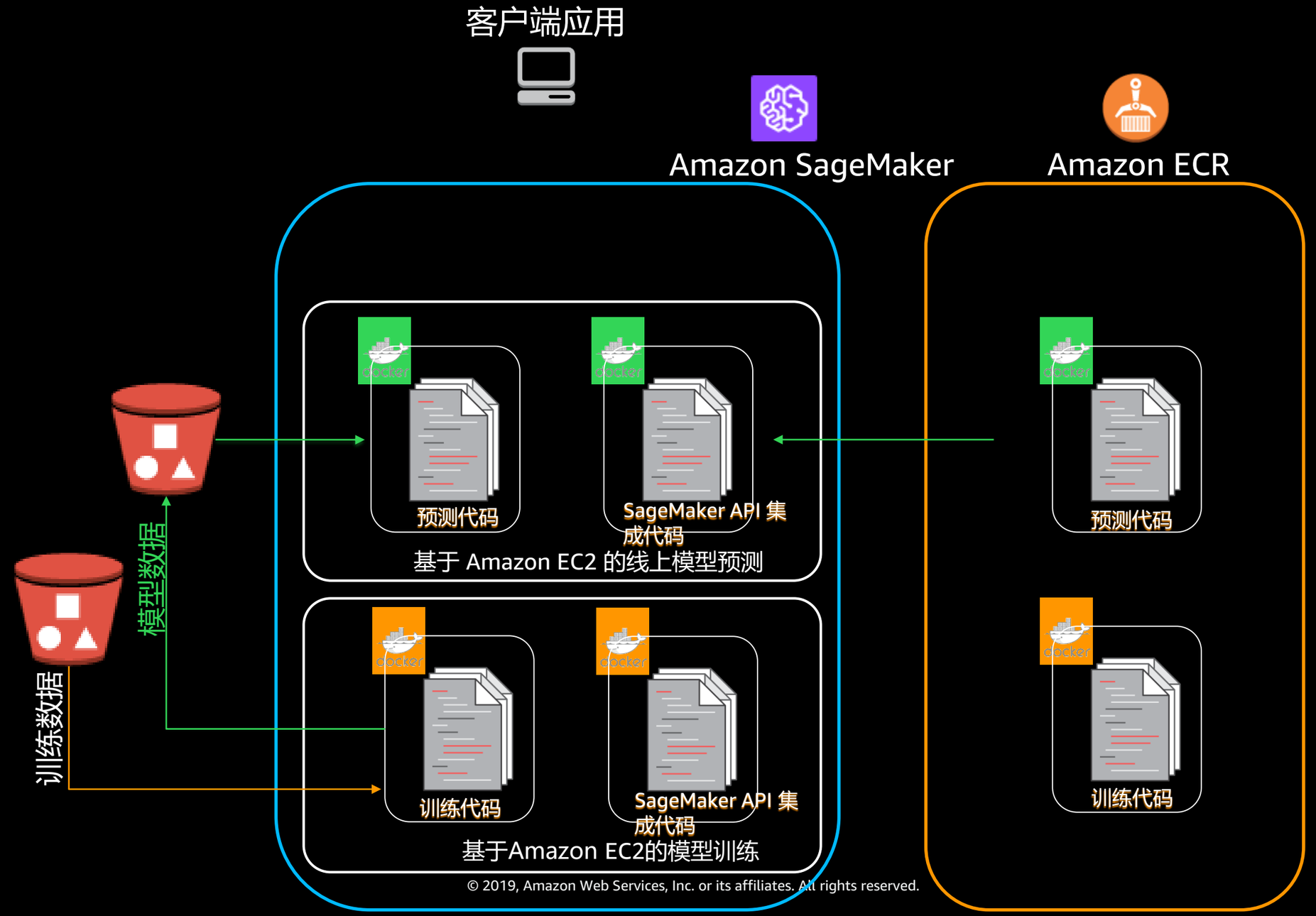
Amazon SageMaker

Amazon ECR

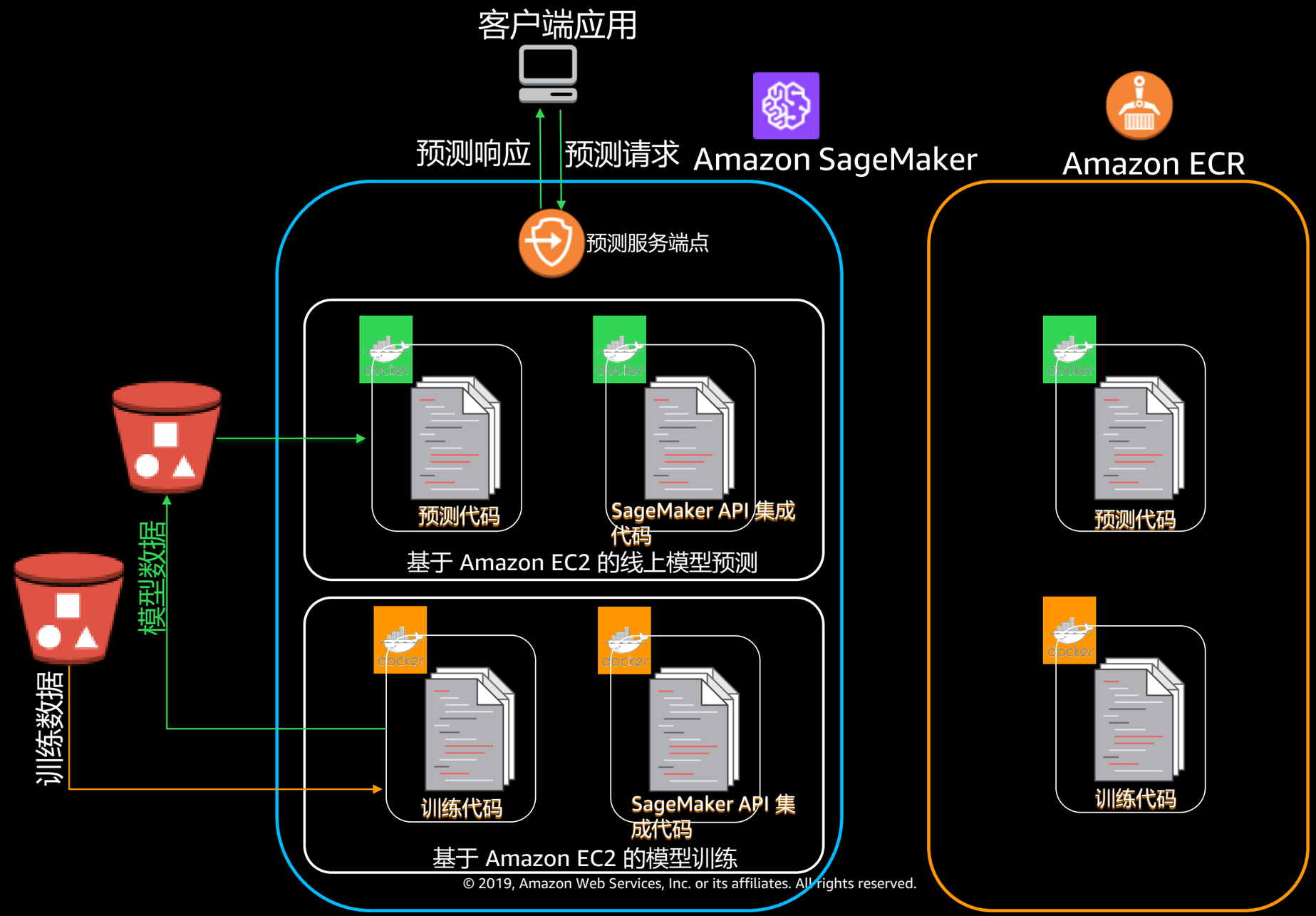
基于 Amazon EC2 的模型训练

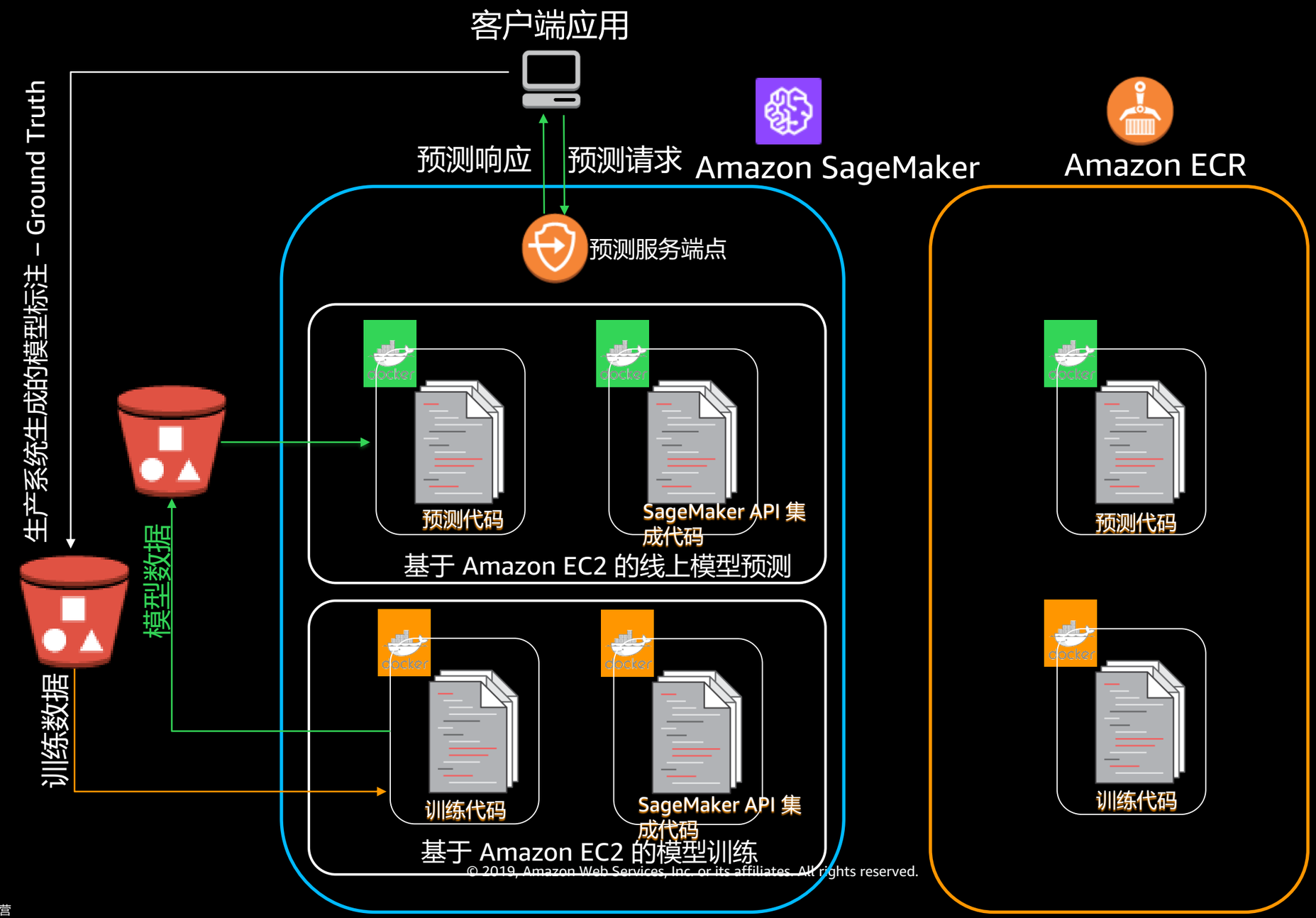
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营



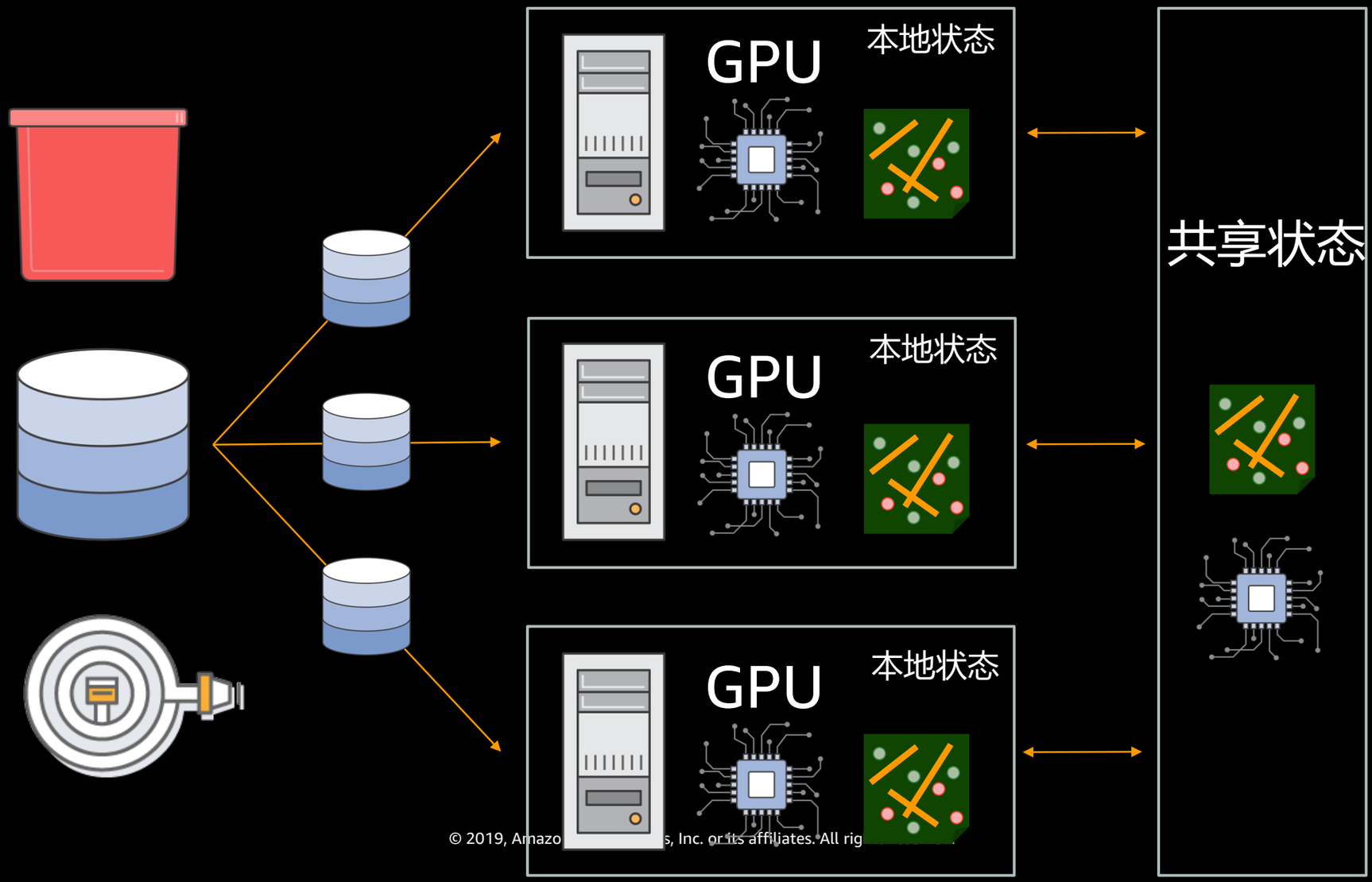


AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

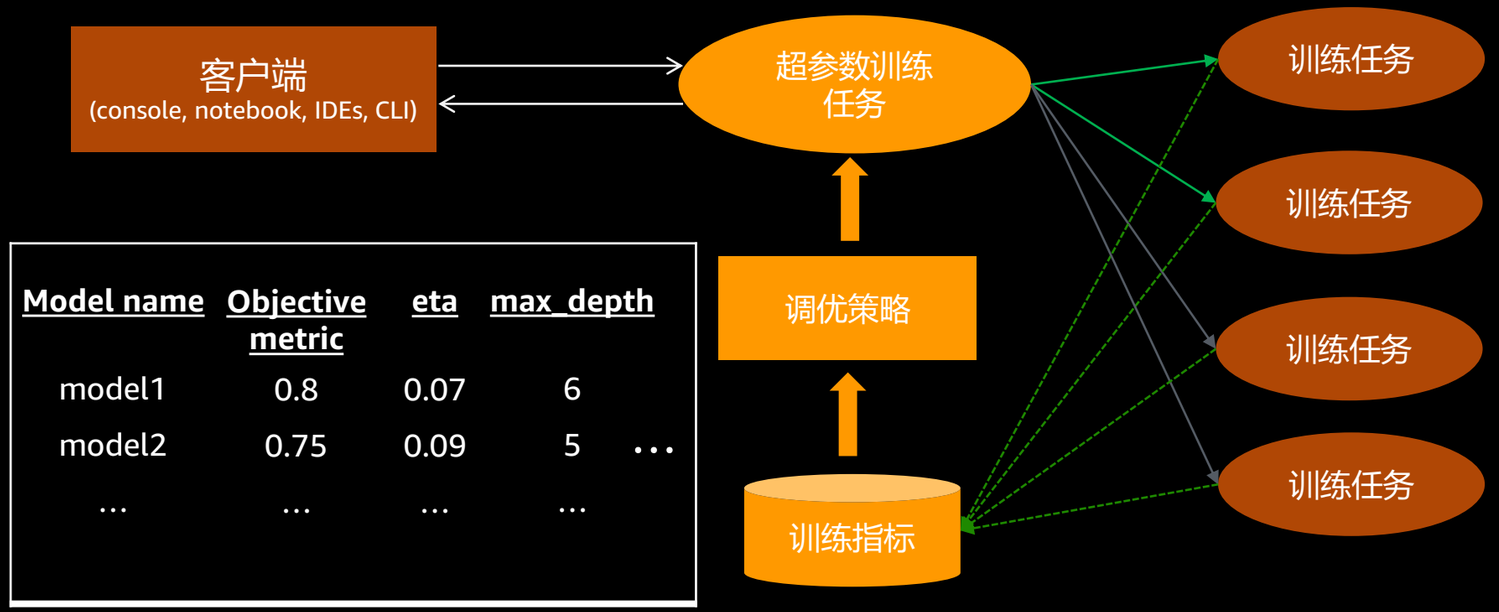




分布式训练架构实现



模型自动调优




```
max_count = 100
for count in xrange(max_count):
      reg = 10**uniform(-5, 5)
      lr = 10**uniform(-3, -6)

      trainer = ClassifierTrainer()
      model = init_two_layer_model(32*32*3, 50, 10) # input size, hidden size, number of classes
      trainer = ClassifierTrainer()
      best_model_local, stats = trainer.train(X_train, y_train, X_val, y_val,
                                              model, two_layer_net,
                                              num_epochs=5, reg=reg,
                                              update='momentum', learning_rate_decay=0.9,
                                              sample_batches = True, batch_size = 100,
                                              learning rate=lr, verbose=False)
```

note it's best to optimize in log space!

```
val_acc: 0.412000, lr: 1.405206e-04, reg: 4.793564e-01, (1 / 100)
val_acc: 0.214000, lr: 7.231888e-06, reg: 2.321281e-04, (2 / 100)
val_acc: 0.208000, lr: 2.119571e-06, reg: 8.011857e+01, (3 / 100)
val_acc: 0.196000, lr: 1.551131e-05, reg: 4.374936e-05, (4 / 100)
val_acc: 0.079000, lr: 1.753300e-05, reg: 1.200424e+03, (5 / 100)
val_acc: 0.223000, lr: 4.215128e-05, reg: 4.196174e+01, (6 / 100)
val_acc: 0.441000, lr: 1.750259e-04, reg: 2.110807e-04, (7 / 100)
val_acc: 0.241000, lr: 6.749231e-05, reg: 4.226413e+01, (8 / 100)
val_acc: 0.482000, lr: 4.296863e-04, reg: 6.642555e-01, (9 / 100)
val_acc: 0.079000, lr: 5.401602e-06, reg: 1.599828e+04, (10 / 100)
val_acc: 0.154000, lr: 1.618508e-06, reg: 4.925252e-01, (11 / 100)
```

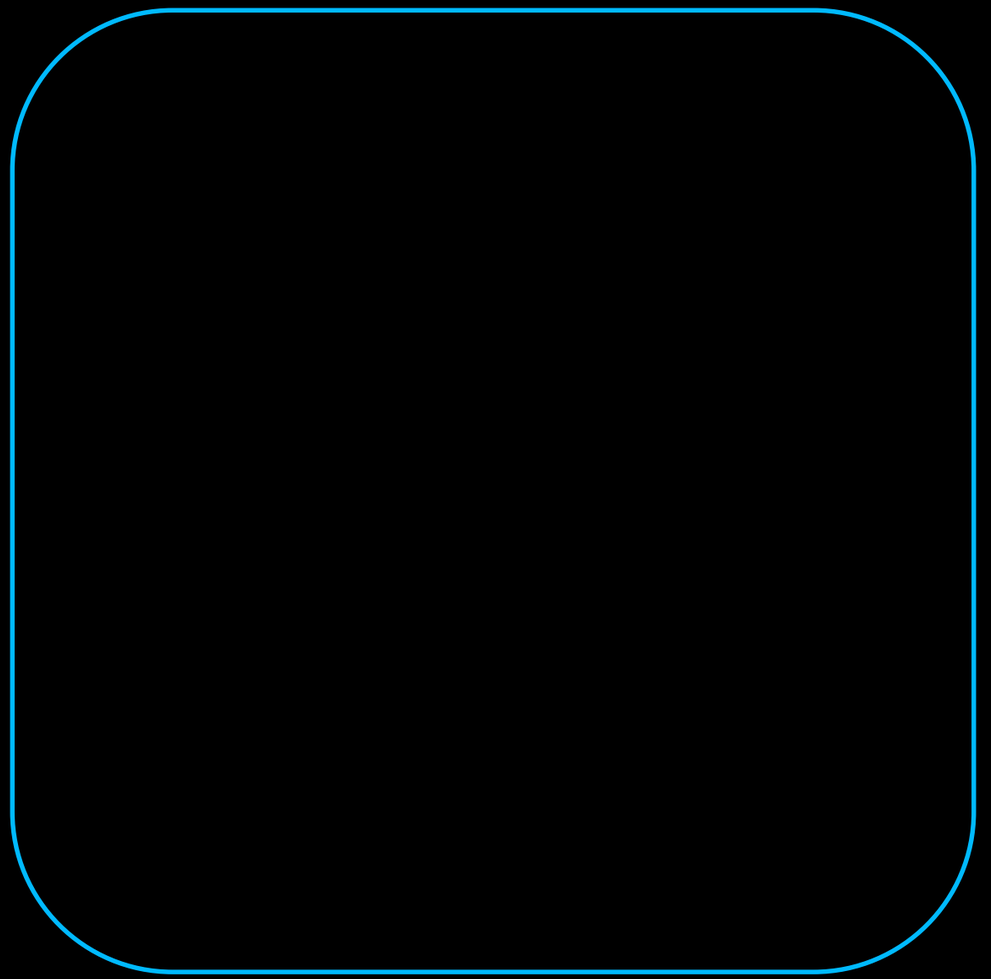
nice

轻松部署训练模型

机器学习线上部署服务

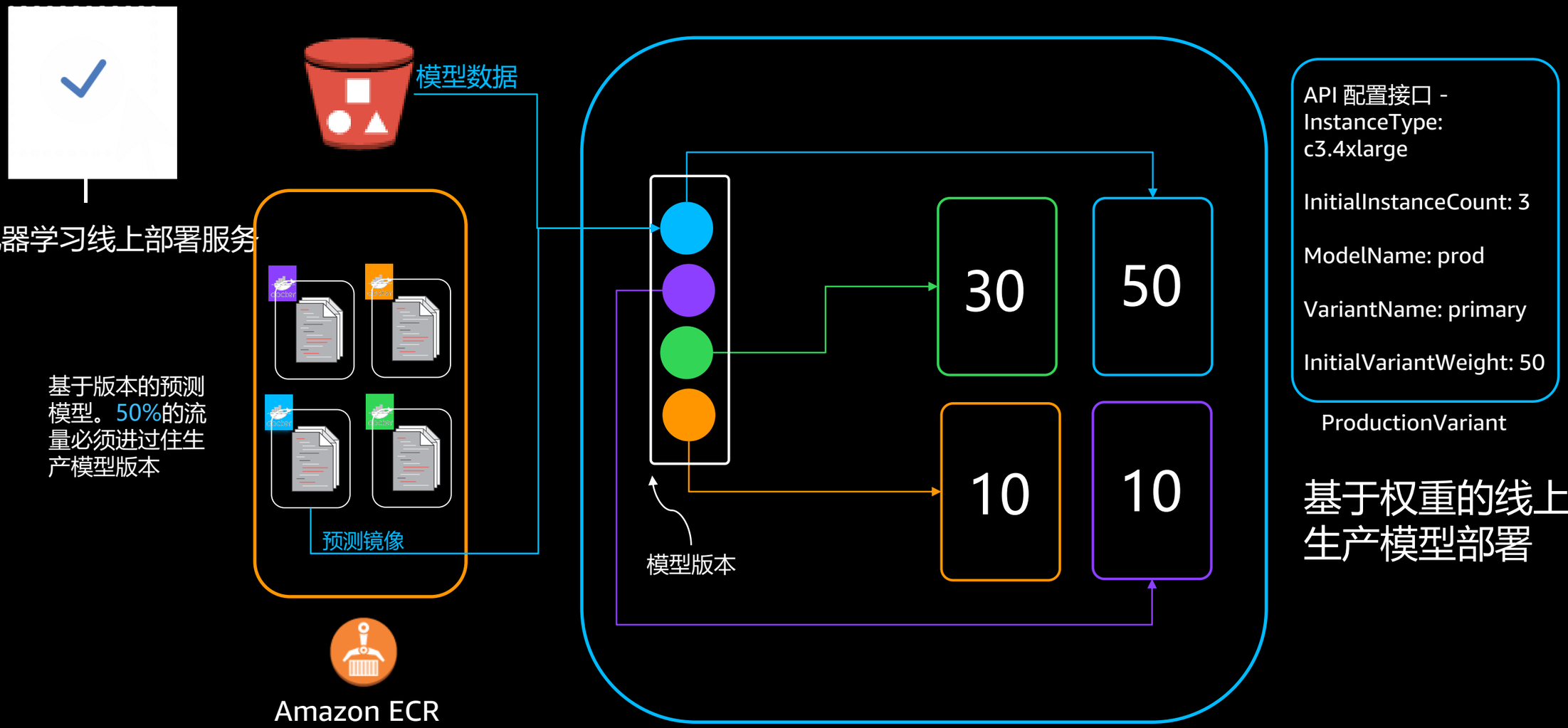
基于版本的预测模型。
50%的流量必须进过
住生产模型版本

Amazon ECR



Amazon SageMaker

轻松部署训练模型



Amazon SageMaker

轻松部署训练模型

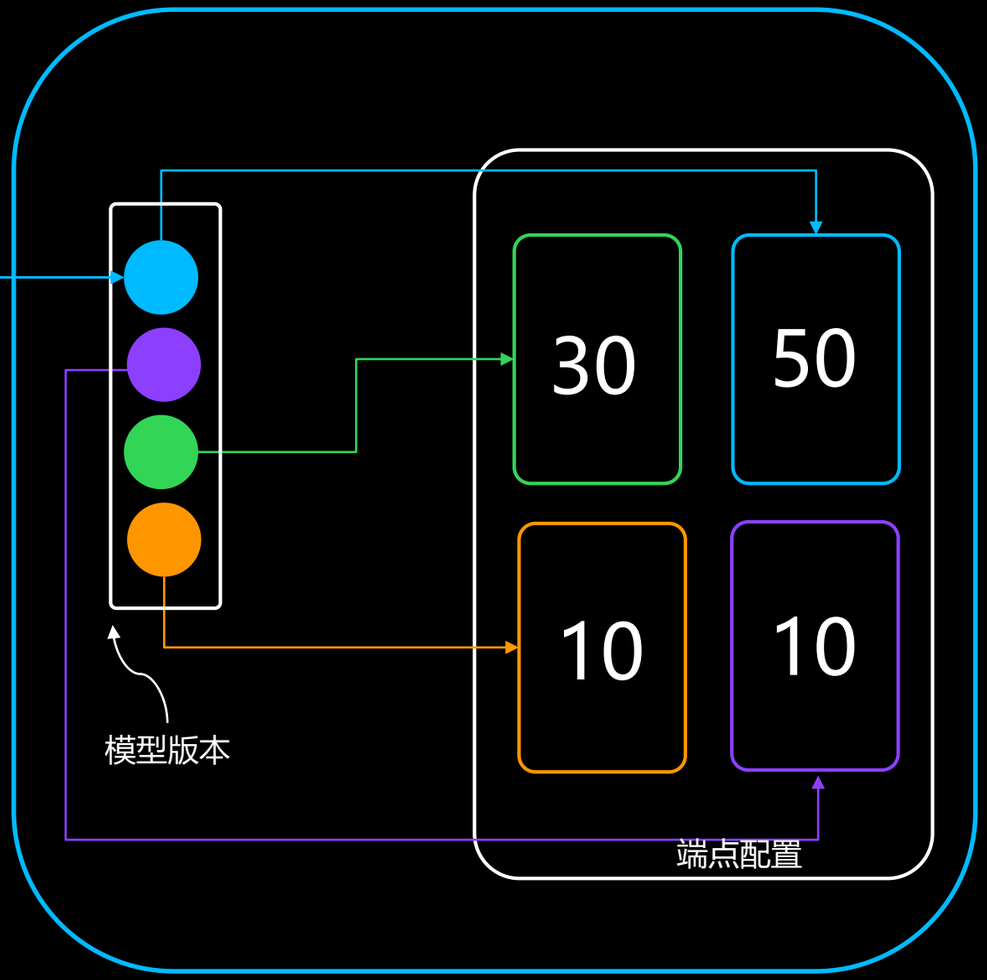
机器学习线上部署服务

模型数据

预测镜像

基于版本的预测模型。50%的流量必须进过住生产模型版本

Amazon ECR



API 配置接口 -
InstanceType:
c3.4xlarge

InitialInstanceCount: 3

ModelName: prod

VariantName: primary

InitialVariantWeight:
50

ProductionVariant

创建端点配置

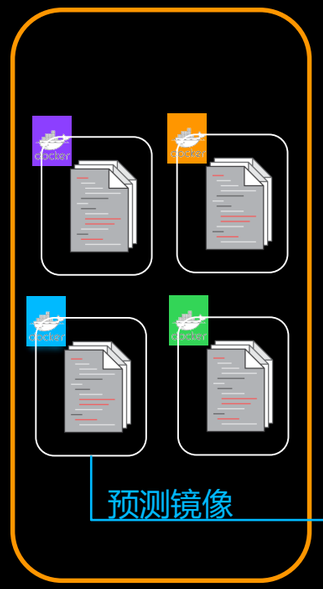
Amazon SageMaker

轻松部署训练模型

机器学习线上部署服务

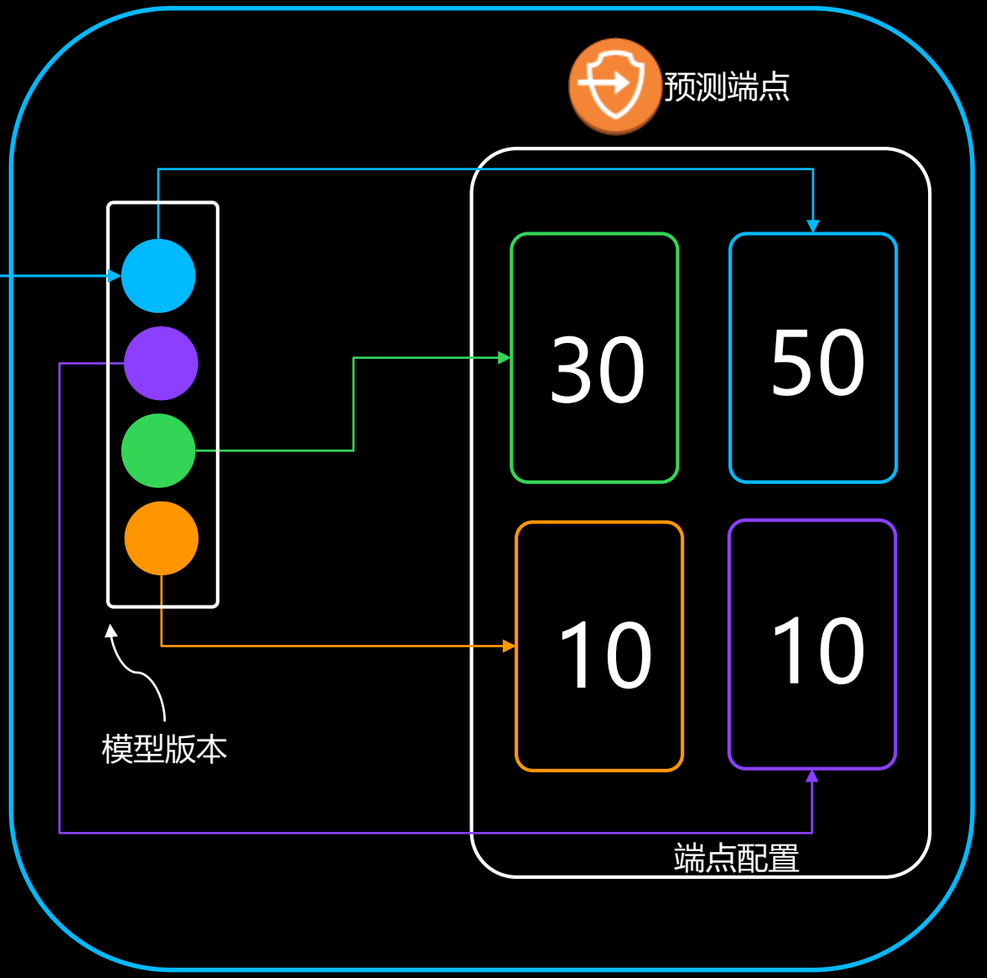
基于版本的预测模型。50%的流量必须进过住生产模型版本

模型数据



预测镜像

Amazon ECR



API 配置接口 -
InstanceType: c3.4xlarge

InitialInstanceCount: 3

ModelName: prod

VariantName: primary

InitialVariantWeight: 50

ProductionVariant

根据端点配置
创建服务端点

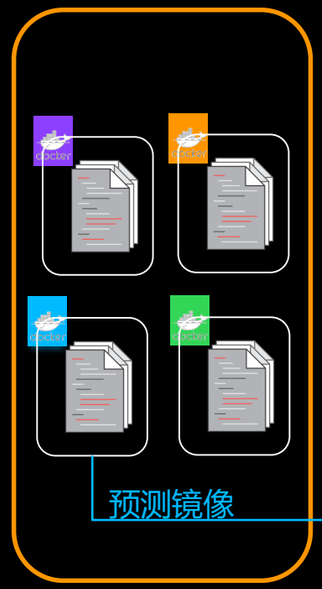
Amazon SageMaker

轻松部署训练模型

机器学习线上部署服务

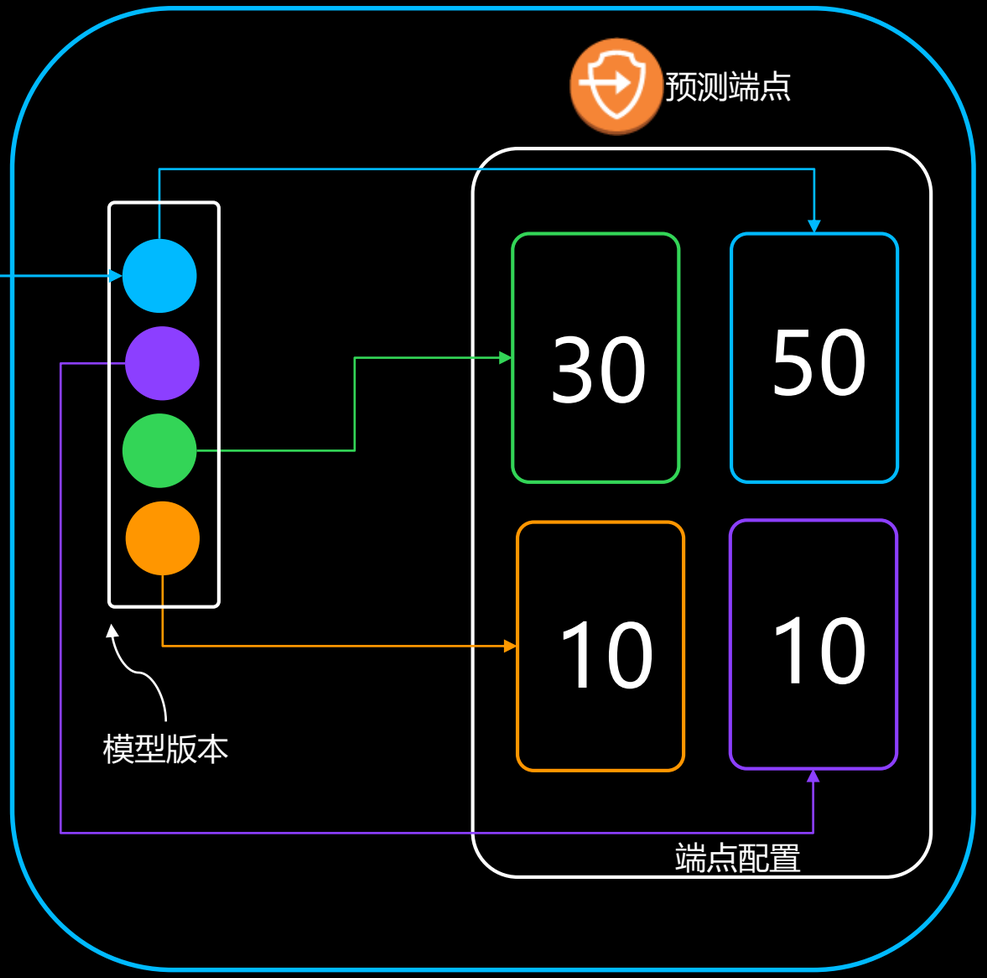
基于版本的预测模型。50% 流量必须进过住生产模型版本

模型数据



预测镜像

Amazon ECR



API 配置接口 -
InstanceType:
c3.4xlarge

InitialInstanceCount: 3

ModelName: prod

VariantName: primary

InitialVariantWeight:
50

ProductionVariant

一键启动

mxnet

TensorFlow™

亚马逊提供的算法

Amazon SageMaker

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker Neo 训练一次，运行在任何地方

编译器

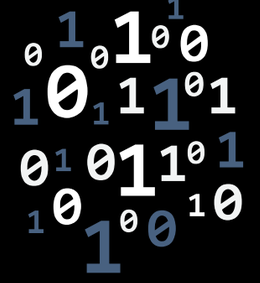
芯片处理器厂家提供
硬指令层面的优化

mxnet

TensorFlow

PyTorch

XGBoost



intel

NVIDIA

Qualcomm

XILINX

cadence

arm

运行环境

设备制造厂商将运行环境嵌入到物
联网边缘设备中

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

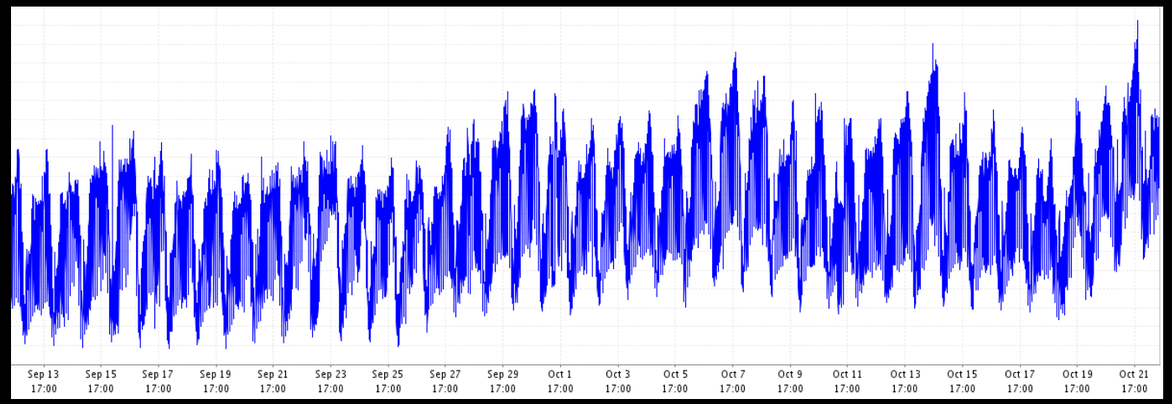
轻松部署训练模型

机器学习线上部署服务

Amazon SageMaker

- ✓ 弹性扩展的预测接口 APIs
- ✓ A/B 测试 (尽请期待更多功能)
- ✓ 低延迟/高吞吐
- ✓ 支持自带定制化模型
- ✓ Python SDK 开发包

自动扩展推理服务终端节点



Variant automatic scaling Learn more

Variant name	Instance type	Current instance count	Current weight
AllTraffic	ml.p2.xlarge	2	1

Minimum instance count: 2 - Maximum instance count: 5

IAM role
Amazon SageMaker uses the following service-linked role for automatic scaling. Learn more
AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint

Built-in scaling policy Learn more

Policy name
SageMakerEndpointInvocationScalingPolicy

Target metric: SageMakerVariantInvocationsPerInstance — Target value: 800

Scale in cool down (seconds) - *optional*: 120

Scale out cool down (seconds) - *optional*: 60

为业务构建机器学习定制化能力

AMAZON SAGEMAKER

降低成本

70%

使用 Ground Truth 大幅降低数字打标签成本

75%

使用弹性推理降低成本

提升性能

10x

更好的算法性能

2x

使用 Neo 优化模型提升预测性能

易于使用

一键部署

模型训练和部署

一次训练

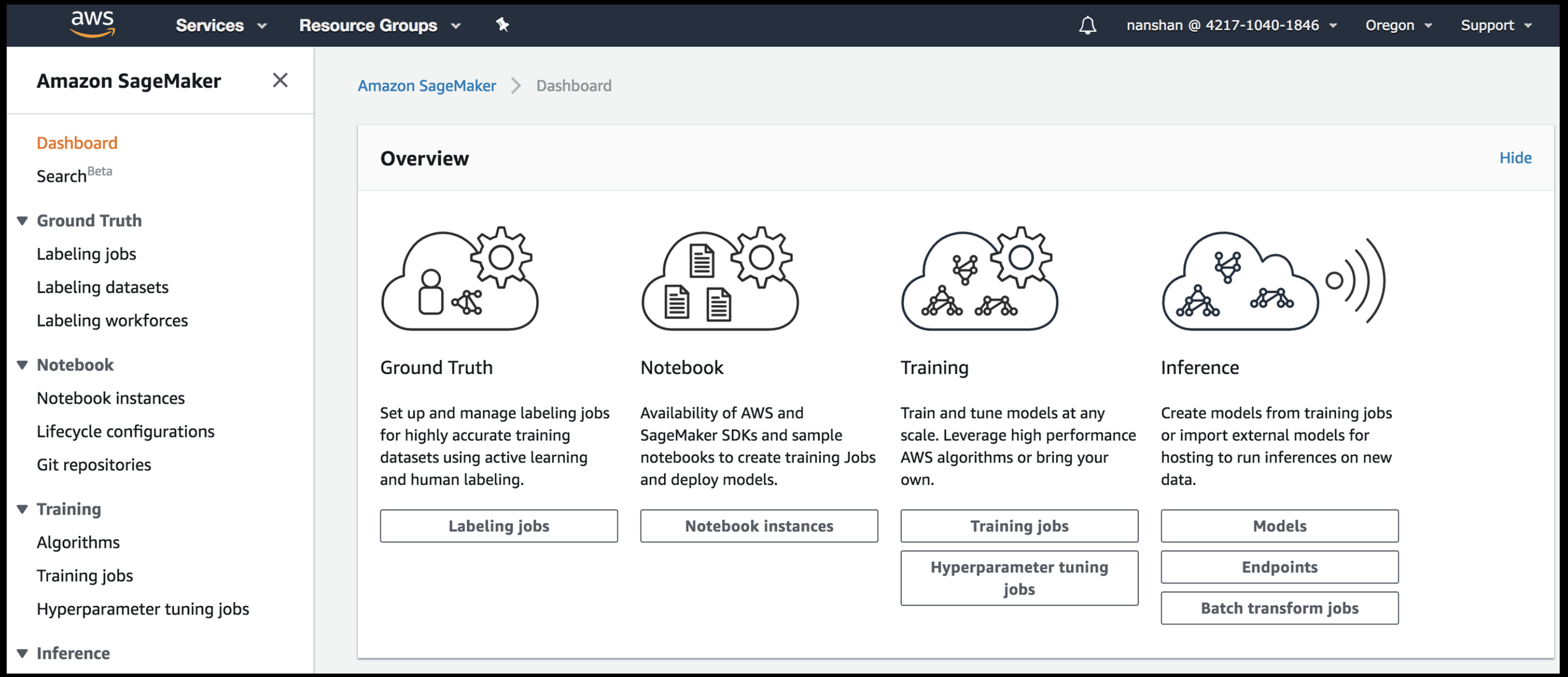
运行在任何地点

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker 的使用

aws INNOVATE
ONLINE CONFERENCE

Amazon SageMaker 开始使用



在命令行中启动训练任务

算法

输入数据

计算资源

```
profile=<your_profile>
arn_role=<your_arn_role>
training_image=382416733822.dkr.ecr.us-east-1.amazonaws.com/kmeans:1
training_job_name=clutering_text_documents_`date '+%Y_%m_%d_%H_%M_%S'`
aws --profile $profile \
    --region us-east-1 \
    sagemaker create-training-job \
    --training-job-name $training_job_name \
    --algorithm-specification TrainingImage=$training_image,TrainingInputMode=File \
    --hyper-parameters k=10,feature_dim=1024,mini_batch_size=1000 \
    --role-arn $arn_role \
    --input-data-config '{"ChannelName": "train", "DataSource": {"S3DataSource":{"S3DataType":
          "S3Prefix", "S3Uri": "s3://kmeans_demo/train", "S3DataDistributionType":
"ShardedByS3Key"}},   "CompressionType": "None", "RecordWrapperType": "None"}' \
    --output-data-config S3OutputPath=s3://training_output/$training_job_name
    --resource-config InstanceCount=2,InstanceType=ml.c4.8xlarge,VolumeSizeInGB=50 \
    --stopping-condition MaxRuntimeInSeconds=3600
```

在 Amazon EMR 中启动训练任务

计算资源

超参数设置

开始训练

部署模型

```
%spark
import com.amazonaws.services.sagemaker.sparksdk.KMeansSageMakerEstimator
import com.amazonaws.services.sagemaker.sparksdk.IAMRole

val roleArn = "<your_role>"

// kMeansSageMakerEstimator
val kMeansSageMakerEstimator = new KMeansSageMakerEstimator(
 sagemakerRole = IAMRole(roleArn),
 trainingInstanceType = "ml.p2.xlarge",
 trainingInstanceCount = 2,
 endpointInstanceType = "ml.c4.xlarge",
 endpointInitialInstanceCount = 1)
 .setK(10).setFeatureDim(784)

// train
val trainedKMeansModel = kMeansSageMakerEstimator.fit(trainingData)
```

READY

```
%spark
// inference. For simplicity, we're just transforming our training data.
val transformedData = trainedKMeansModel.transform(testData)
```

READY

在 Amazon SageMaker Notebook 中启动训练任务

计算资源

超参数

开始训练

```
import boto3
import sagemaker

sess = sagemaker.Session()

pca = sagemaker.estimator.Estimator(containers[boto3.Session().region_name],
                                    role,
                                    train_instance_count=1,
                                    train_instance_type='ml.c4.xlarge',
                                    output_path=output_location,
                                    sagemaker_session=sess)
pca.set_hyperparamters(feature_dim=50000,
                       num_components=10,
                       subtract_mean=True,
                       algorithm_mode='randomized',
                       mini_batch_size=200)

pca.fit({'train': s3_train_data})
```

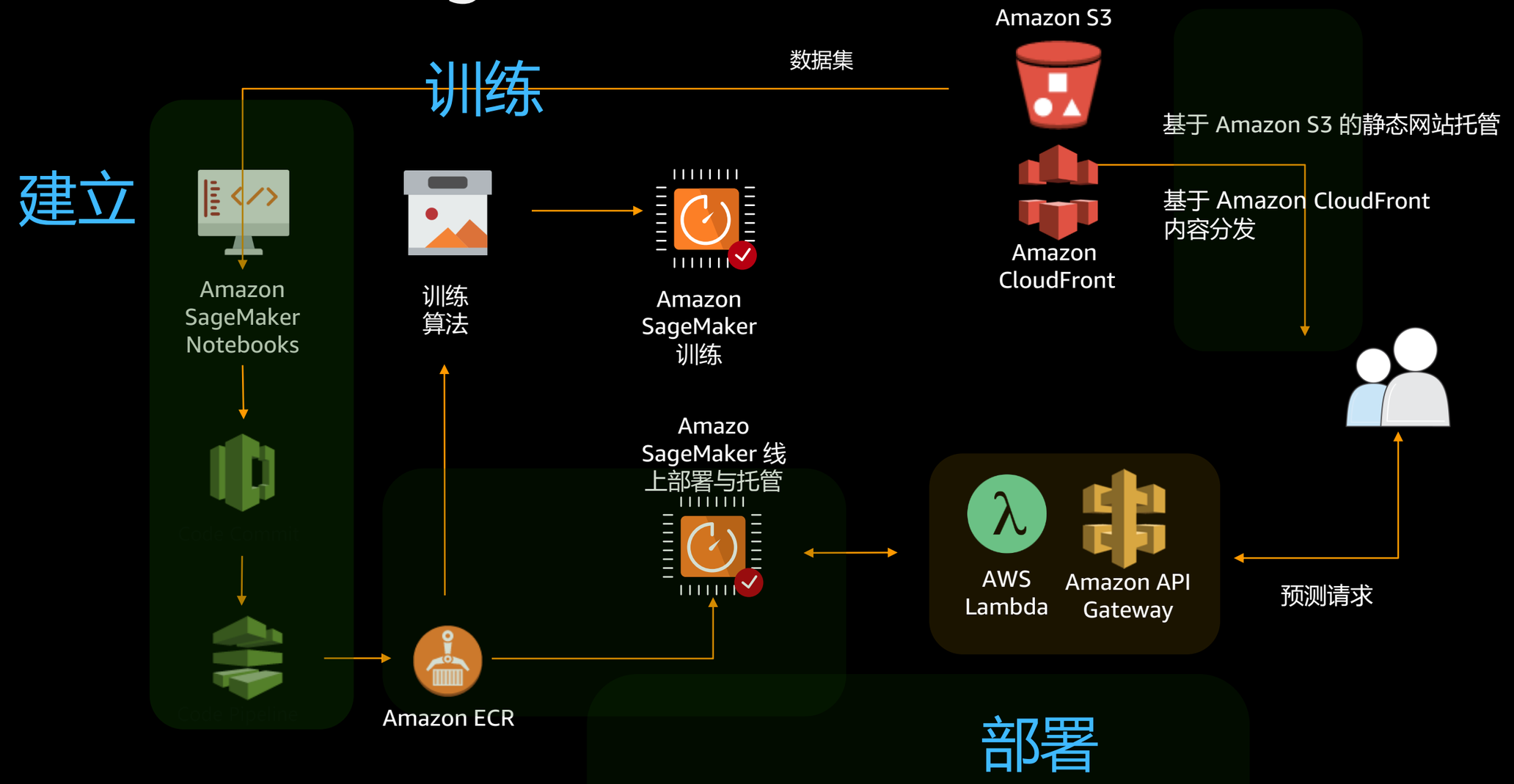
部署模型

```
pca_predictor = pca.deploy(initial_instance_count=1,
                           instance_type='ml.c4.xlarge')
```

Amazon SageMaker 帮助

- Amazon SageMaker 入门文档: https://aws.amazon.com/sagemaker/
- 使用 Amazon SageMaker SDK:
  - Python: https://github.com/aws/sagemaker-python-sdk
  - Spark: https://github.com/aws/sagemaker-spark
- Amazon SageMaker 样例: https://github.com/awslabs/amazon-sagemaker-examples
- 把你构建的应用告诉我们!

基于 Amazo SageMaker 端到端参考架构



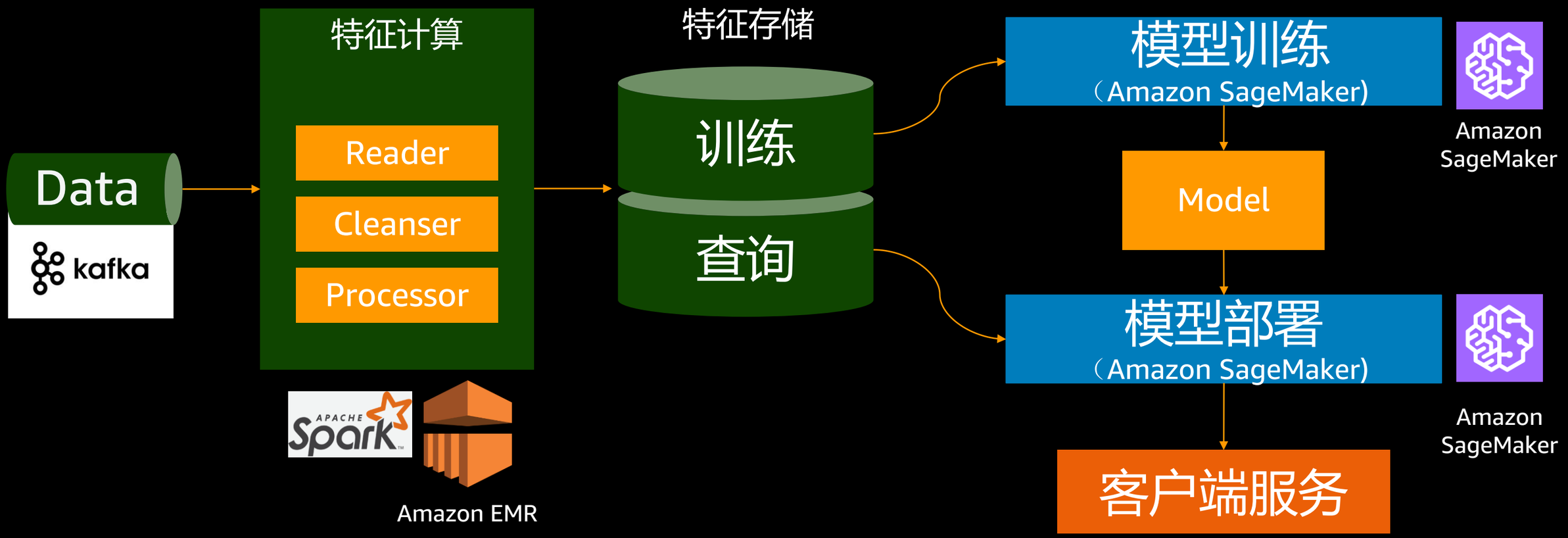
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon SageMaker 客户案例

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Intuit 构建准实时的欺诈检测



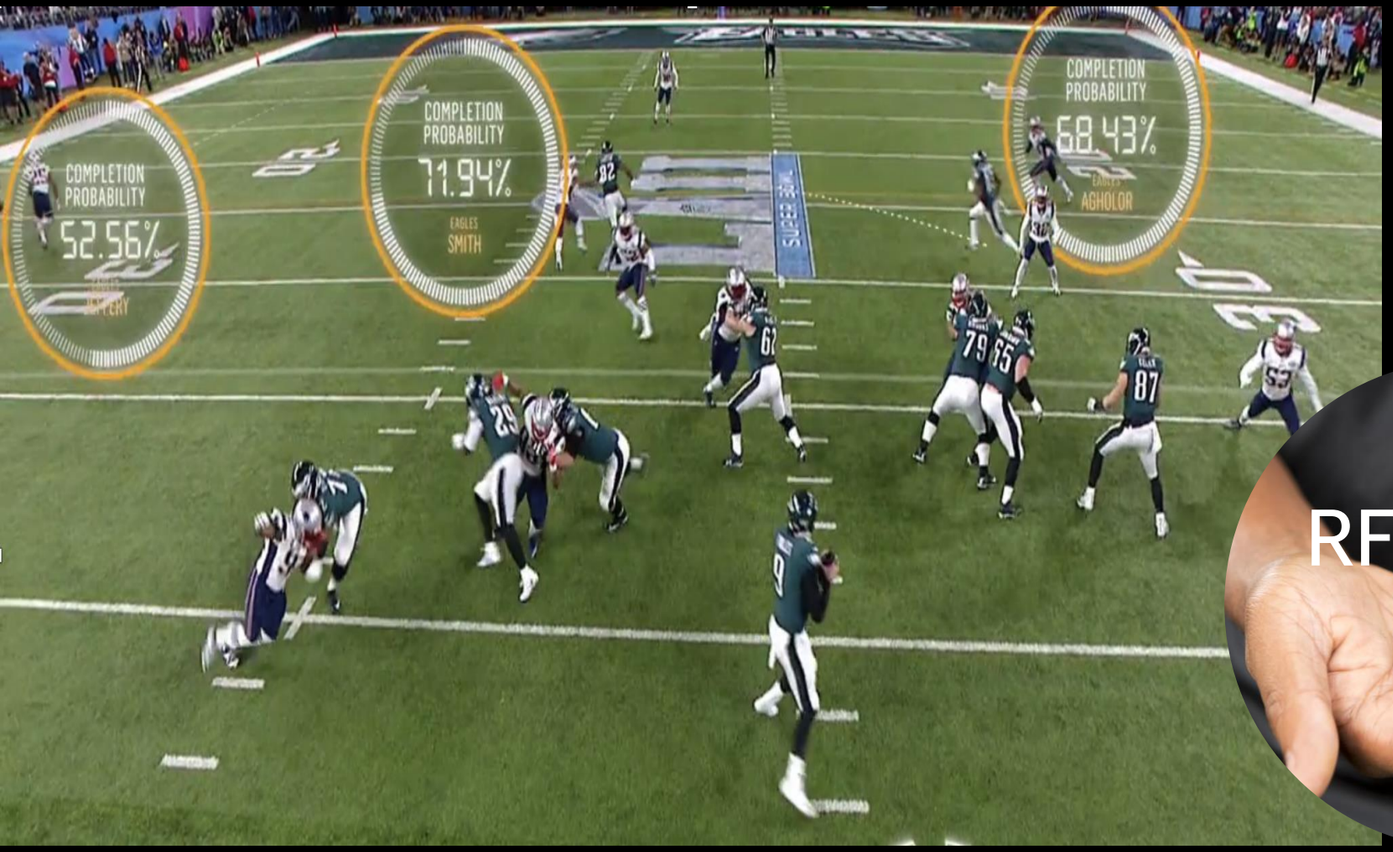
intuit

Intuit 使用 Amazon SageMaker

从前	现在
需要临时设置 和管理 notebook 环境	使用 Amazon SageMaker notebook 轻松完成数据探索工作
有限的模型部署选择	通过虚拟化手段 达成极强的灵活性
团队之间需要争抢计算资源	自动扩展的模型部署环境

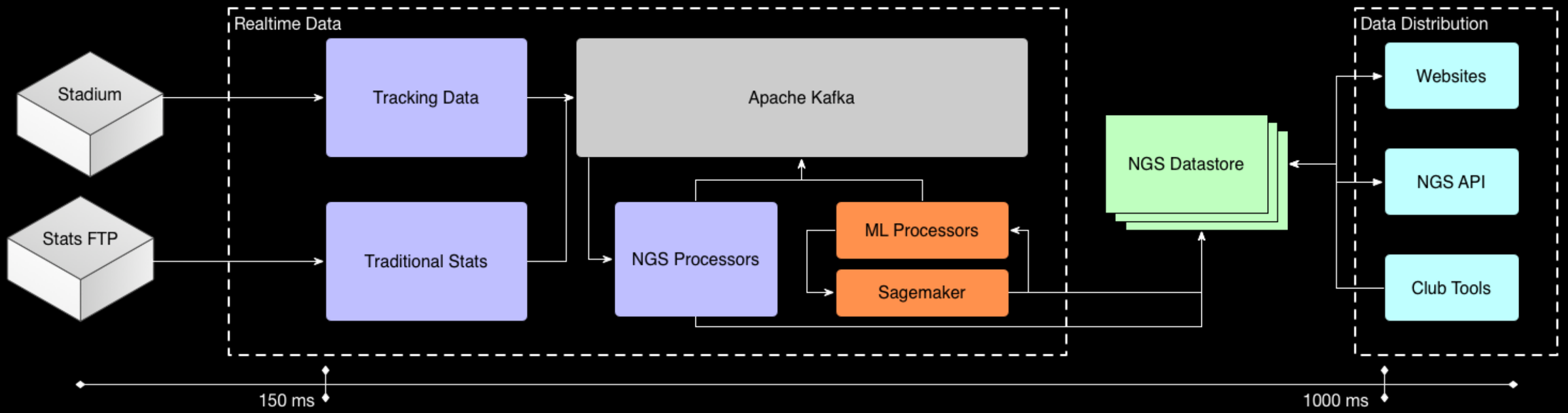
NFL 增强球迷观赏体验

NEXT GEN STATS



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

使用 Amazon SageMaker 的 Next Gen Stats 系统



感谢参加 AWS INNOVATE 2019 在线技术大会

我们希望您在这里找到感兴趣的内容！

也请帮助我们完成**投票打分**和**反馈问卷**。

欲获取关于 AWS 的更多信息和技术内容，可以通过以下方式找到我们：


微信公众号：AWSChina


新浪微博：https://www.weibo.com/amazonaws/


领英：https://www.linkedin.com/company/aws-china/


知乎：https://www.zhihu.com/org/aws-54/activities/


视频中心：http://aws.amazon.bokecc.com/


更多线上活动：https://aws.amazon.com/cn/about-aws/events/webinar/

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营