QCon 全球软件开发大会 INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2018

Amazon SageMaker

端到端的托管机器学习平台

演讲者 / AWS 王世帅

主办方 Geekbang. InfoQ

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE
全球技术领导力峰会

通往**年薪百万**的CTO的路上，
如何打造自己的技术**领导**力？

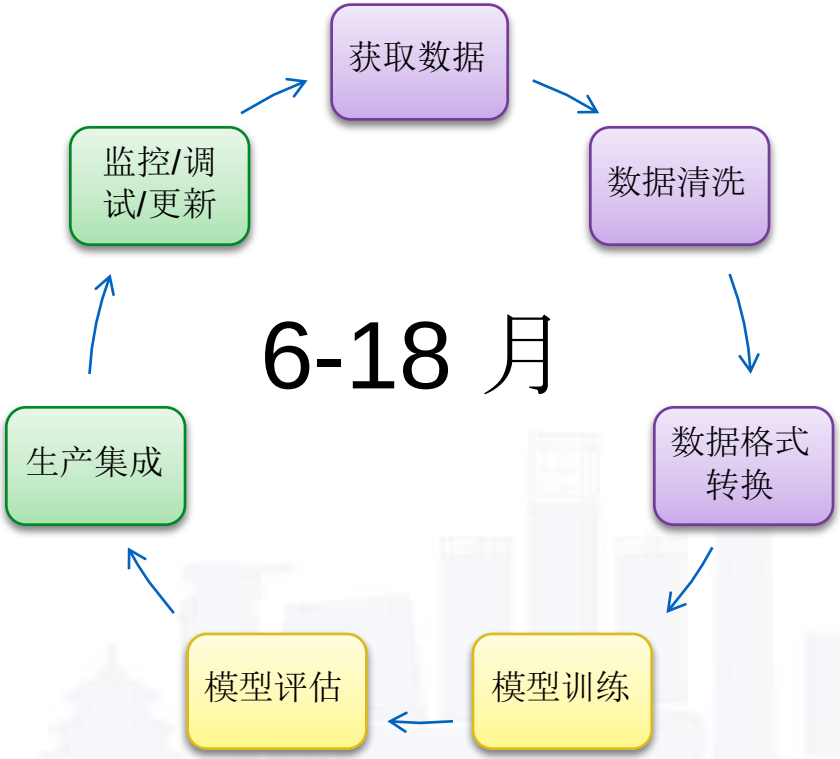
扫描二维码了解详情

目录

- 为什么我们要推出Amazon SageMaker？
- 什么是Amazon SageMaker?
- 如何使用Amazon SageMaker
- Amazon SageMaker案例分享
- Q&A

为什么我们推出
Amazon Sagemaker

应用机器学习艰难而漫长



亚马逊在人工智能领域的大量深度创新

1995

亚马逊自从成立以来一直在人工智能和机器学习领域进行大量投入，并且把我们的知识与能力与客户分享

2017

商品智能推荐

机器人与物流仓储

新产品

供应链管理

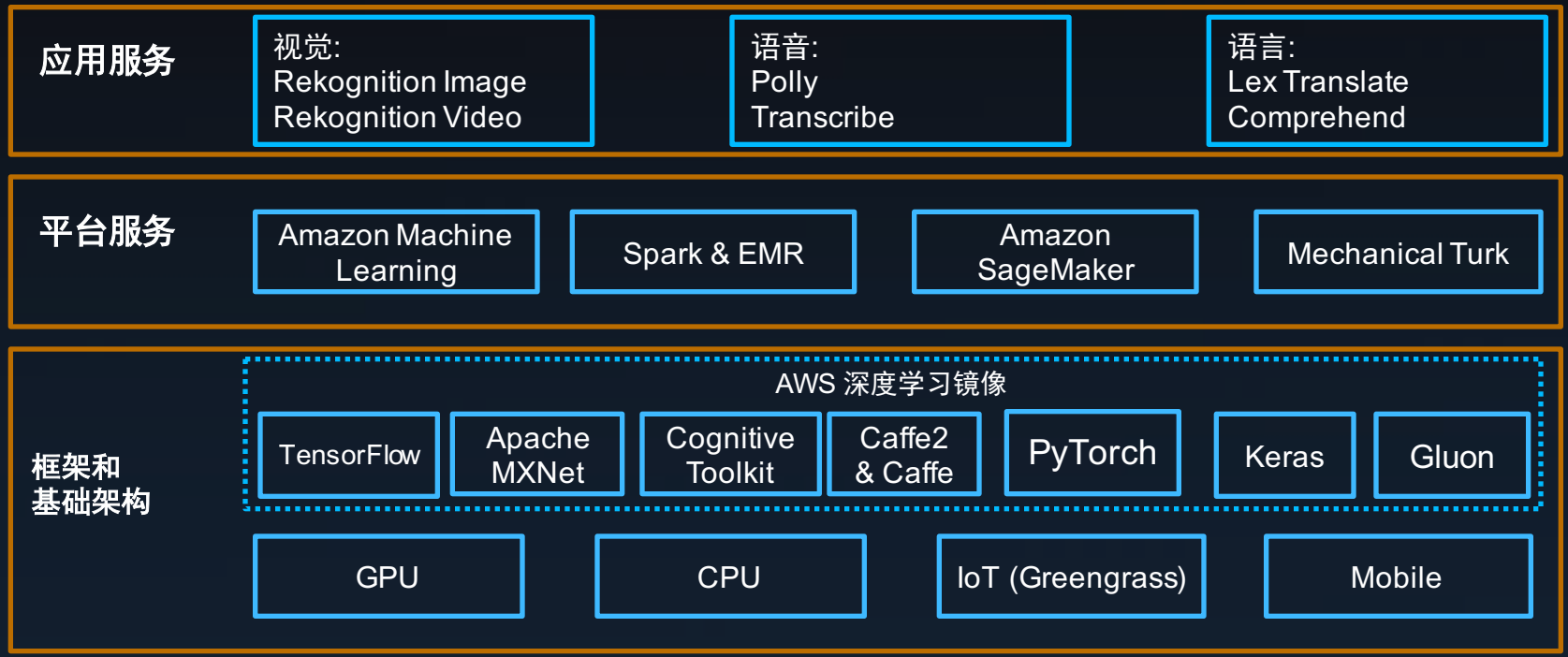
智慧呼叫中心

无人值守商店

我们的使命：

为每一个开发者和数据科学家打造机器学习平台

AWS上的机器学习技术堆栈



什么是Amazon Sagemaker？

Amazon SageMaker

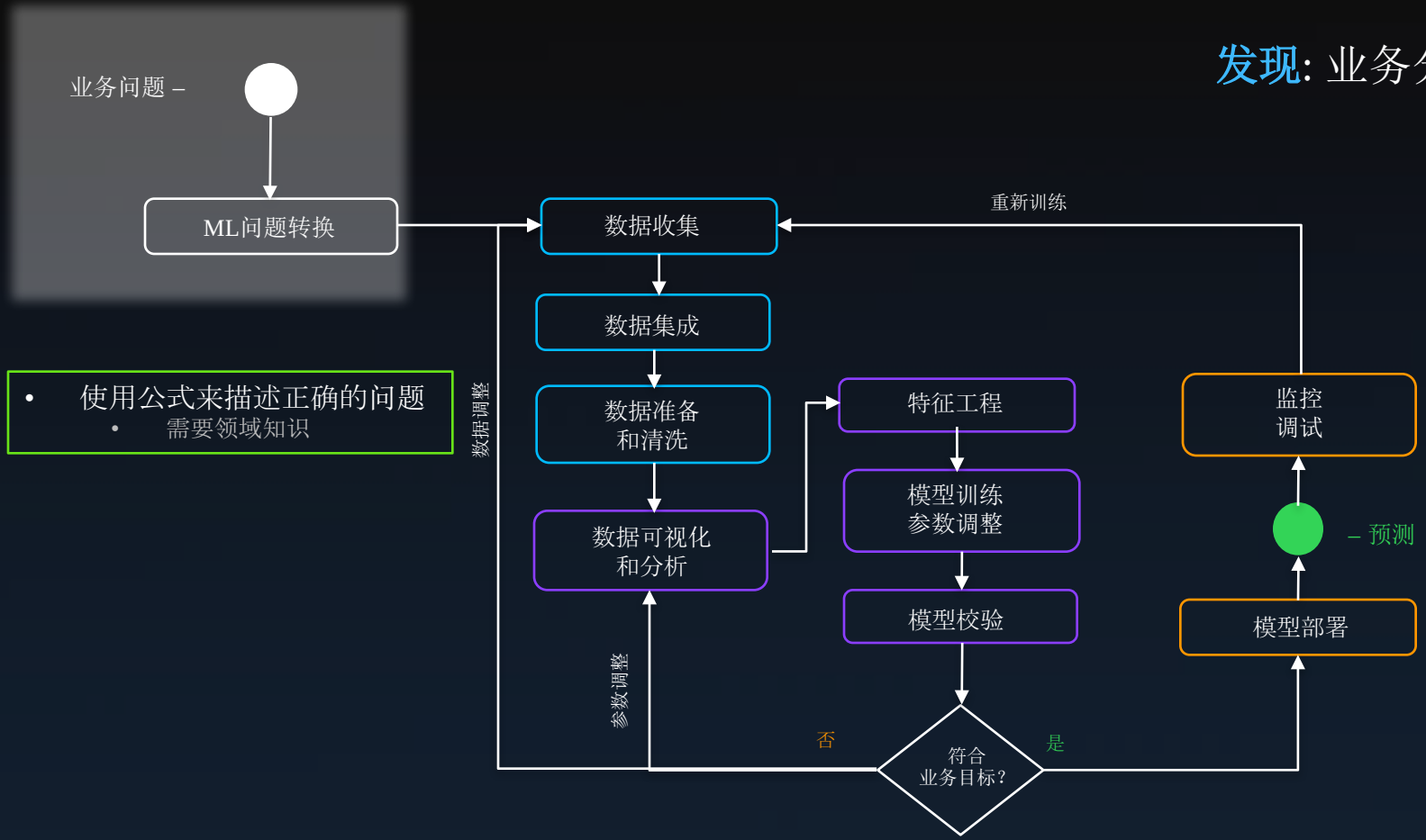
A **fully managed service** that enables **data scientists** and **developers** to quickly and easily **build** machine-learning based models **into production** smart applications.

一个**全托管服务**，可以帮助**数据科学家**和**开发者**快速而轻松地**构建**基于机器学习的模型的**生产环境**智能应用

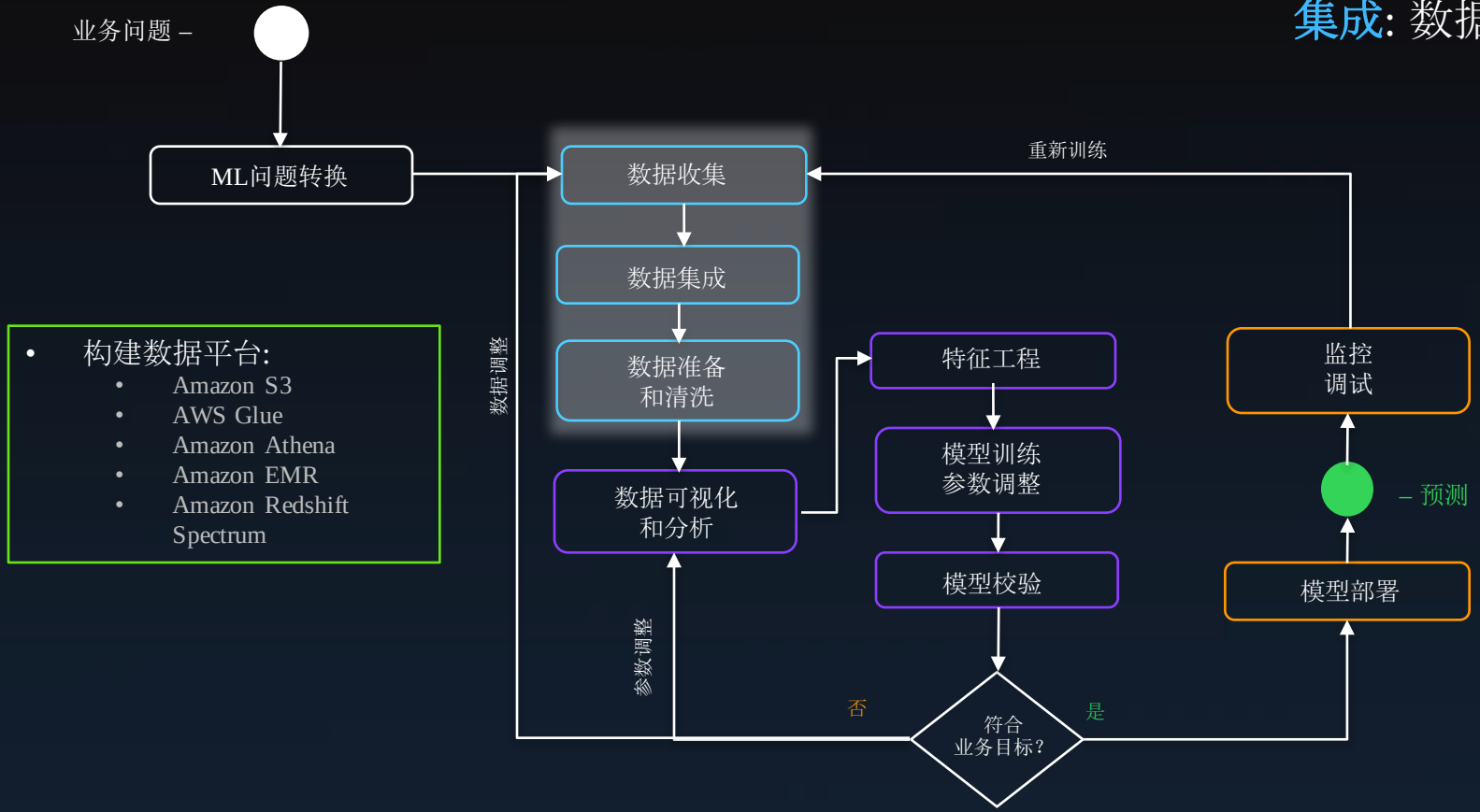
aws

机器学习流程

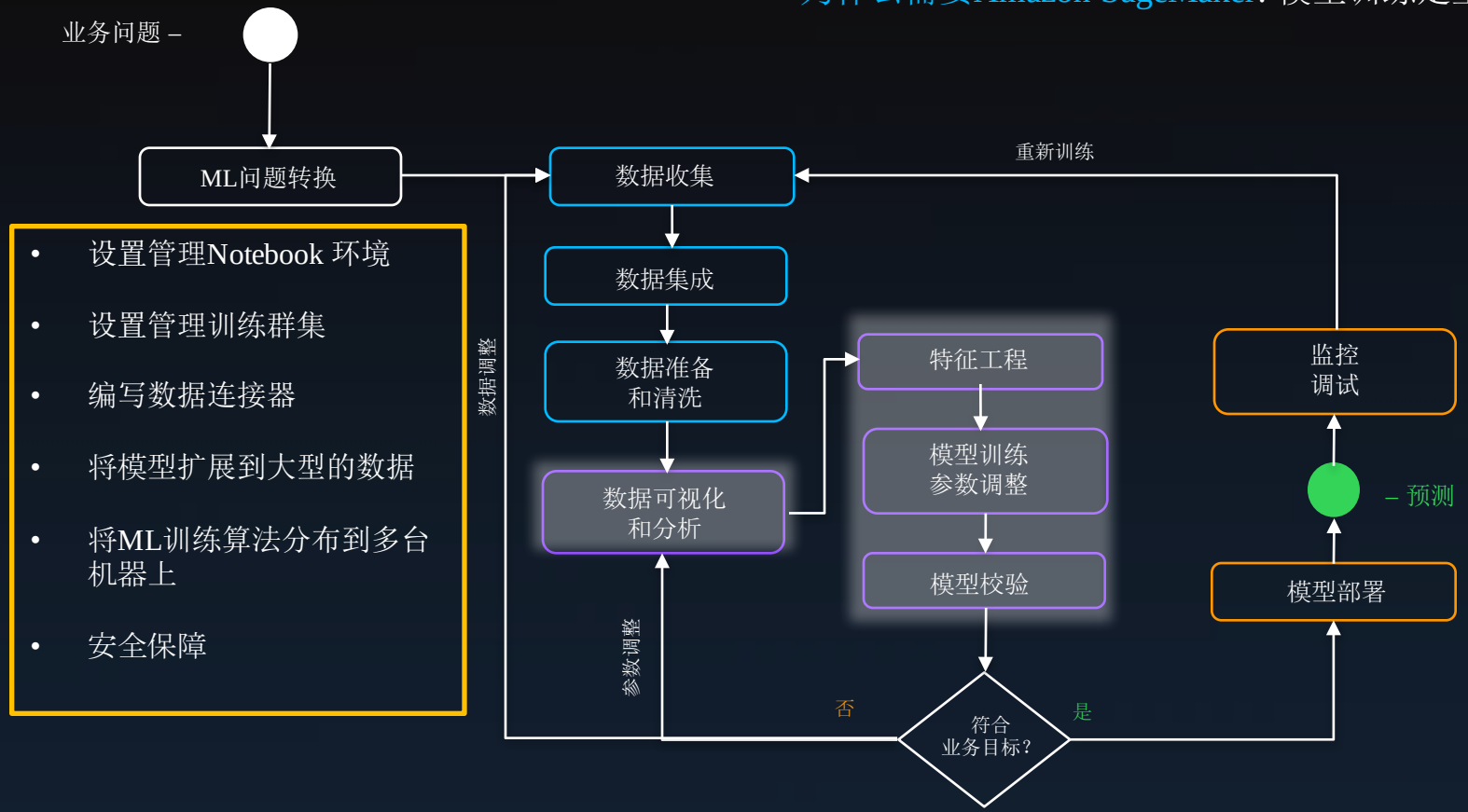
发现: 业务分析过程



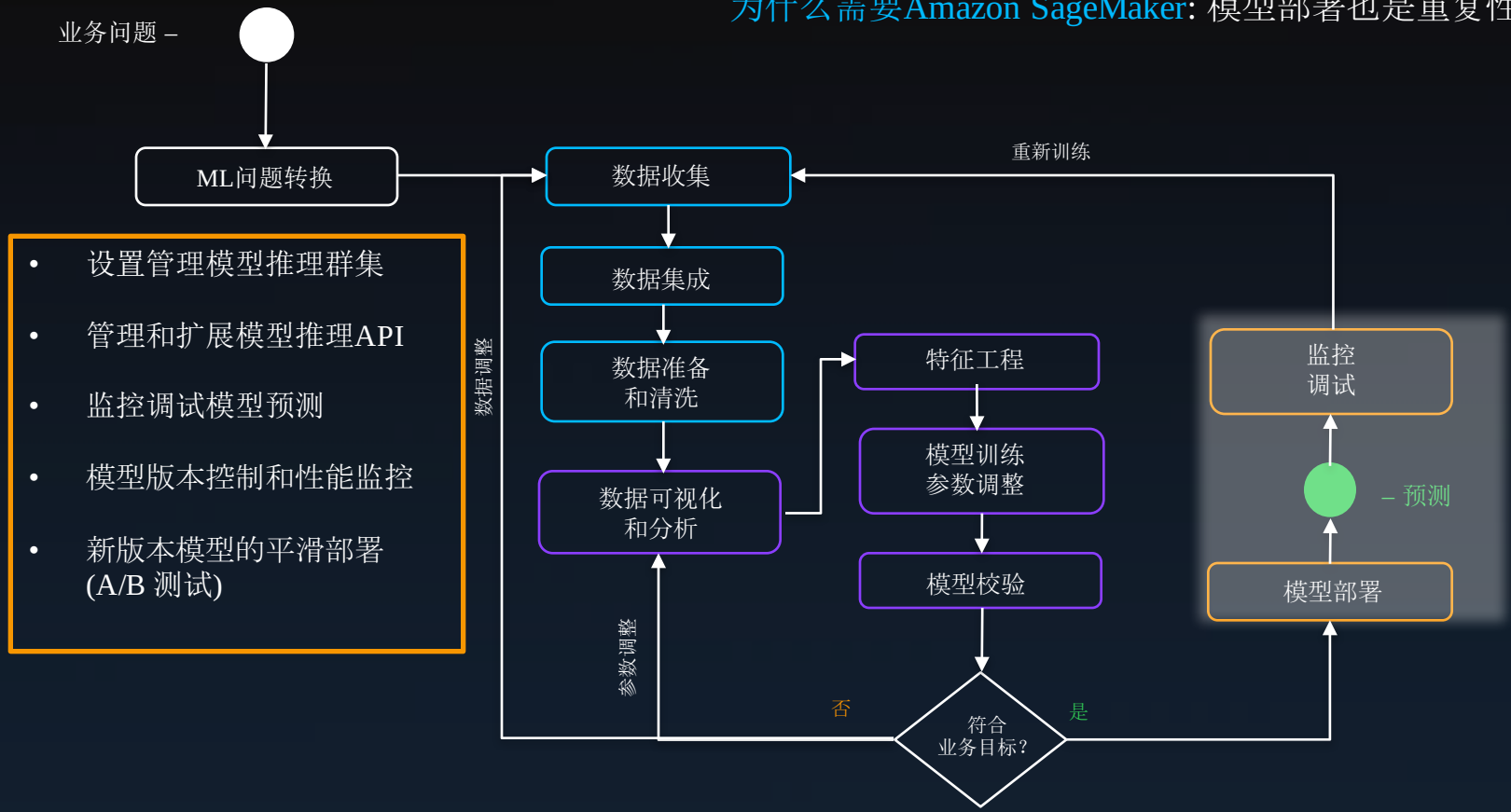
集成: 数据架构过程



为什么需要Amazon SageMaker: 模型训练是重复性的重体力工作



为什么需要Amazon SageMaker: 模型部署也是重复性的重体力工作



Amazon SageMaker

① 

I

Notebook 实例

② 

I

算法

③ 

I

模型训练服务

④ 

I

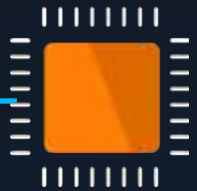
模型部署服务

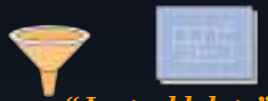
1

零配置 数据探索分析平台

I

Notebook实例





“Just add data”

- 推荐、个性化
- 欺诈检测
- 预测
- 图像分类
- 流失预测
- 市场活动、邮件的定位
- 日志处理，异常检测
- 语音文字转换
- 其它更多…

2

Amazon SageMaker: 10倍速 算法支持

I

算法

流式数据集
训练费用更低

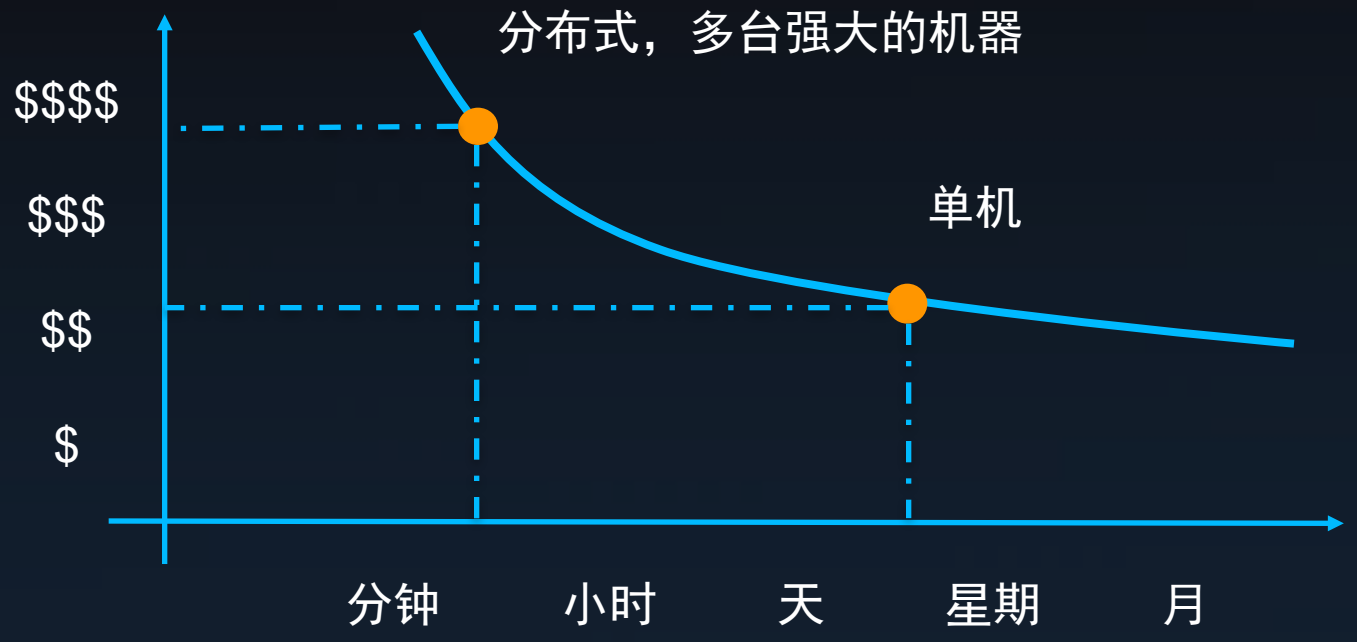
更快的训练速度

处理超大型数据集时
提供更高的可靠性

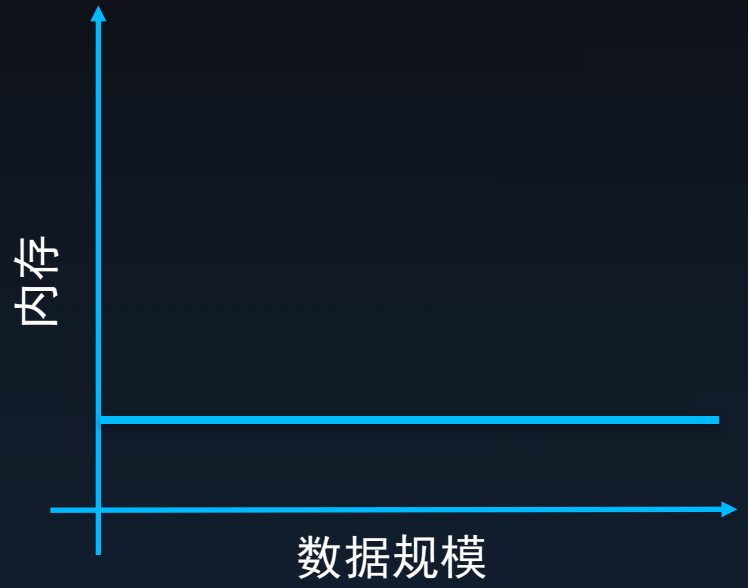
可以选择多种机器学
习算法

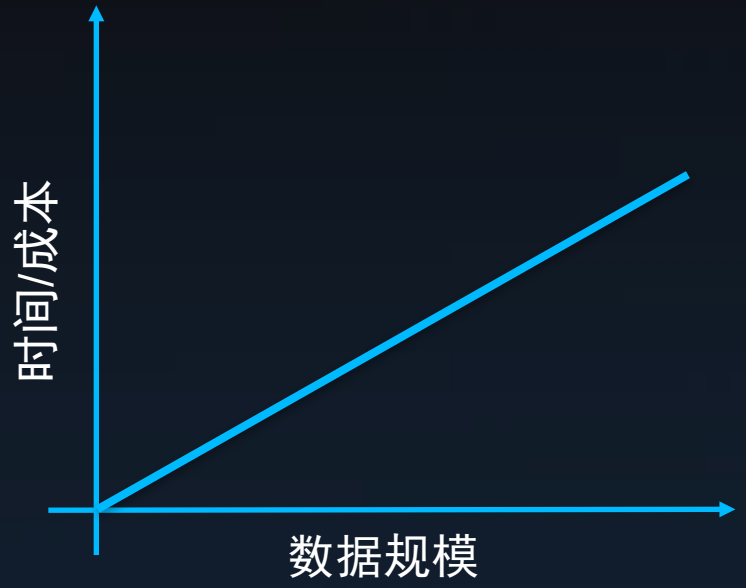
aws

费用 vs. 时间



流式数据





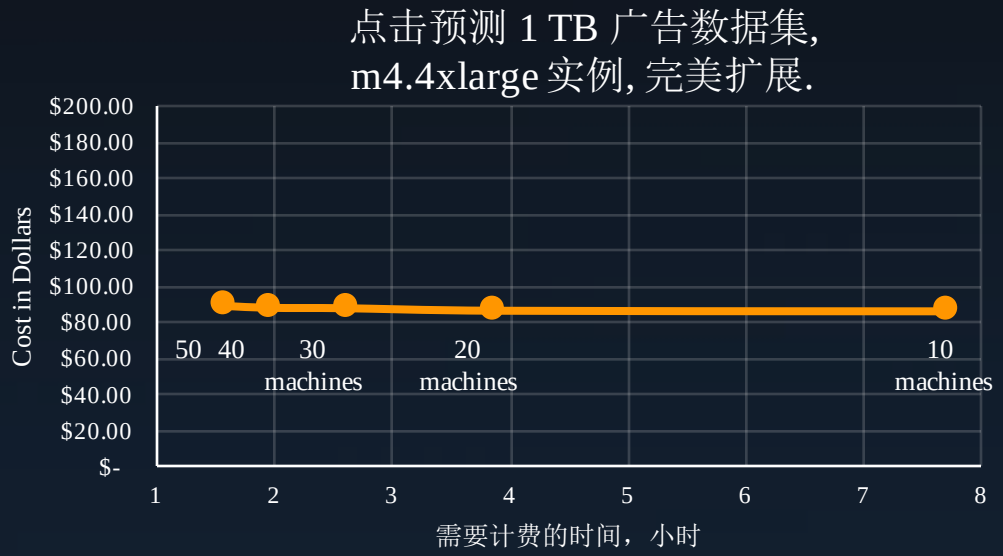
费用 vs. 时间

无限扩展的机器学习算法

因子分解机（Factorization Machines）

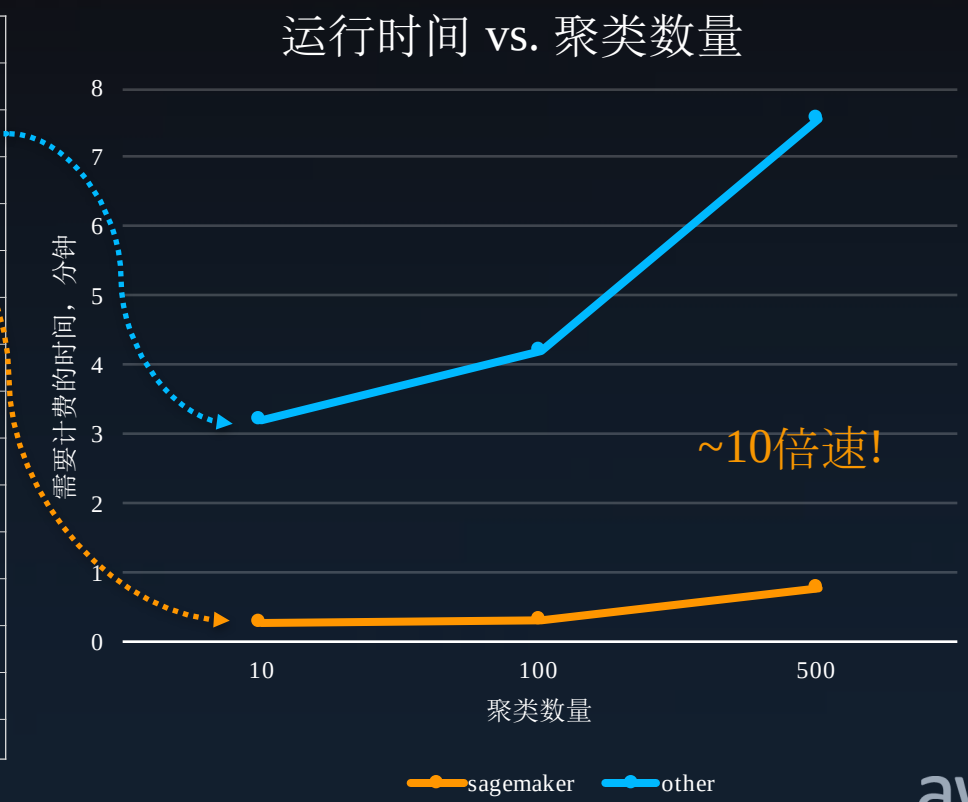
$$\tilde{y} = w_0 + \langle w_1, x \rangle + \sum_{i,j>i} x_i x_j \cdot \langle v_i, v_j \rangle$$

	Log_loss	F1 Score	**Seconds**
SageMaker	0.494	0.277	820
Other (10 Iter)	0.516	0.190	650
Other (20 Iter)	0.507	0.254	1300
Other (50 Iter)	0.481	0.313	3250

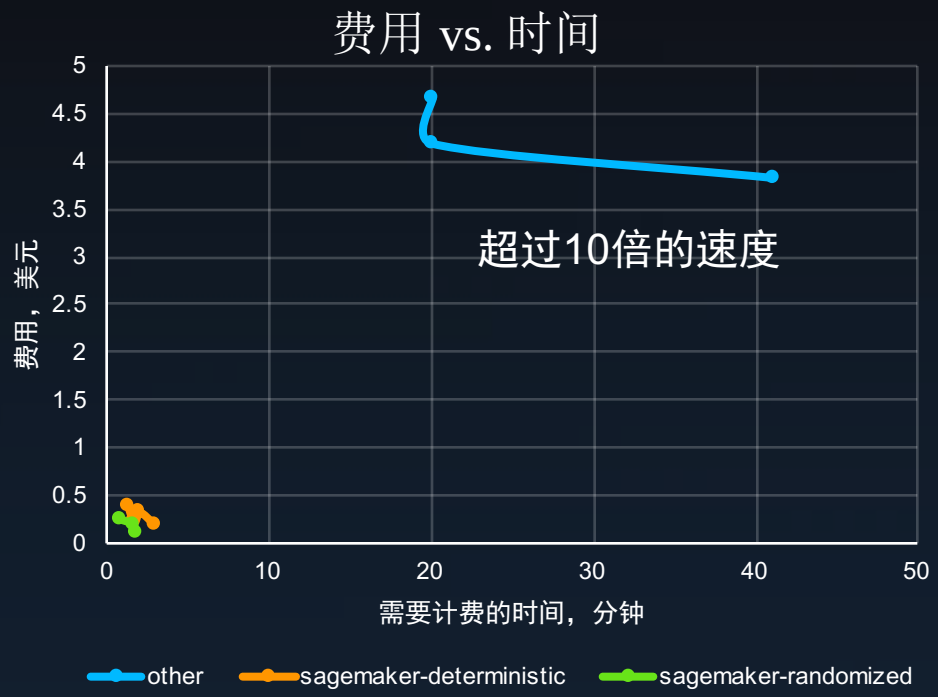


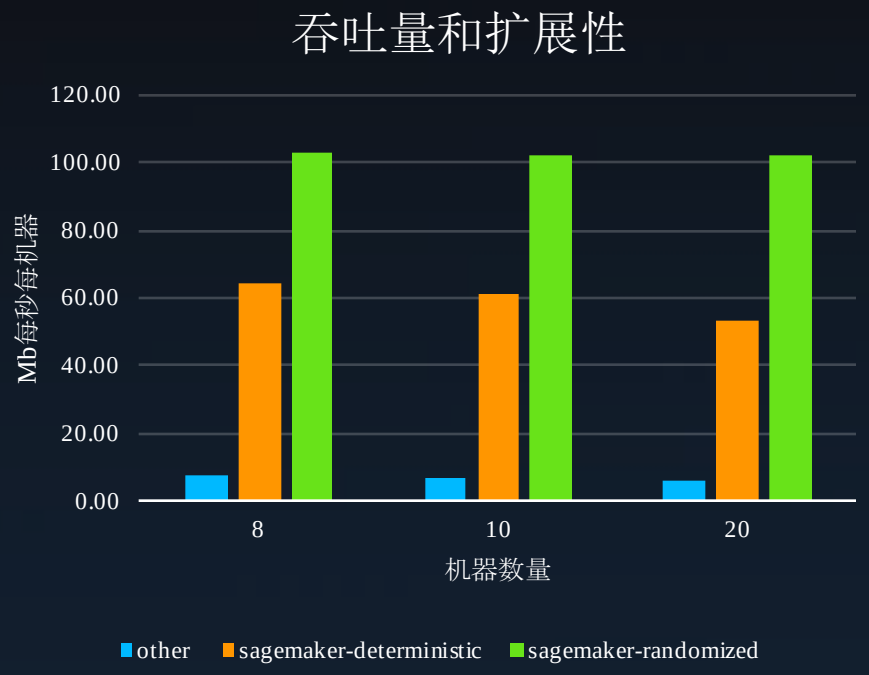
K-Means 聚类

	k	SageMaker	Other
Text 1.2GB	10	1.18E3	1.18E3
	100	1.00E3	9.77E2
	500	9.18.E2	9.03E2
Images 9GB	10	3.29E2	3.28E2
	100	2.72E2	2.71E2
	500	2.17E2	Failed
Videos 27GB	10	2.19E2	2.18E2
	100	2.03E2	2.02E2
	500	1.86E2	1.85E2
Advertising 127GB	10	1.72E7	Failed
	100	1.30E7	Failed
	500	1.03E7	Failed
Synthetic 1100GB	10	3.81E7	Failed
	100	3.51E7	Failed
	500	2.81E7	Failed

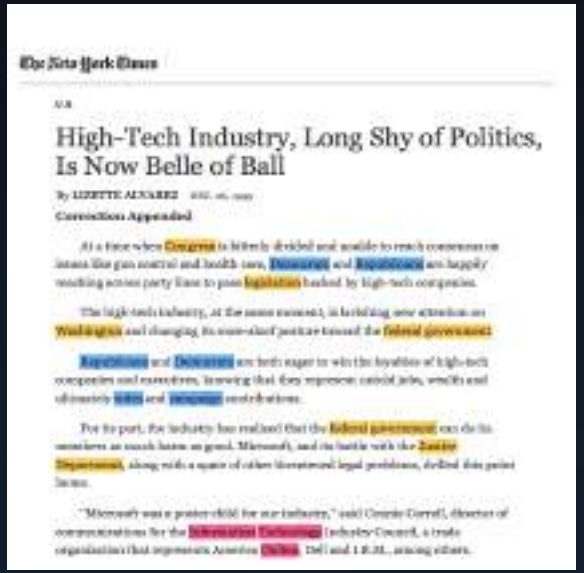


主成分分析 (PCA)





Spectral LDA



The New York Times

U.S.

High-Tech Industry, Long Shy of Politics, Is Now Belle of Ball

By LIZETTE ALVAREZ

Correction Appended

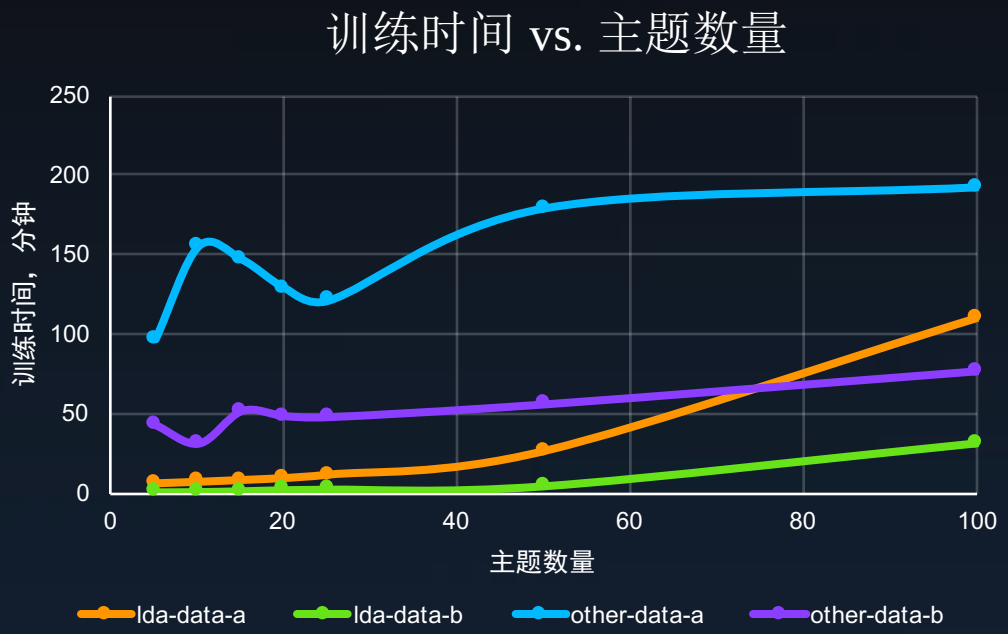
At a time when Congress is bitterly divided and unable to reach consensus on issues like gun control and health care, Democrats and Republicans are happily reaching across party lines to pass legislation backed by high-tech companies.

The high-tech industry, at the same moment, is lavishing new attention on Washington and changing its state-chief posture toward the federal government.

Republicans and Democrats are both eager to win the loyalties of high-tech companies and executives, knowing that they represent untold jobs, wealth and ultimately votes and campaign contributions.

For its part, the industry has realized that the federal government can do its members an much harm as good. Microsoft, and its battle with the Justice Department, along with a spate of other threatened legal problems, drilled this point home.

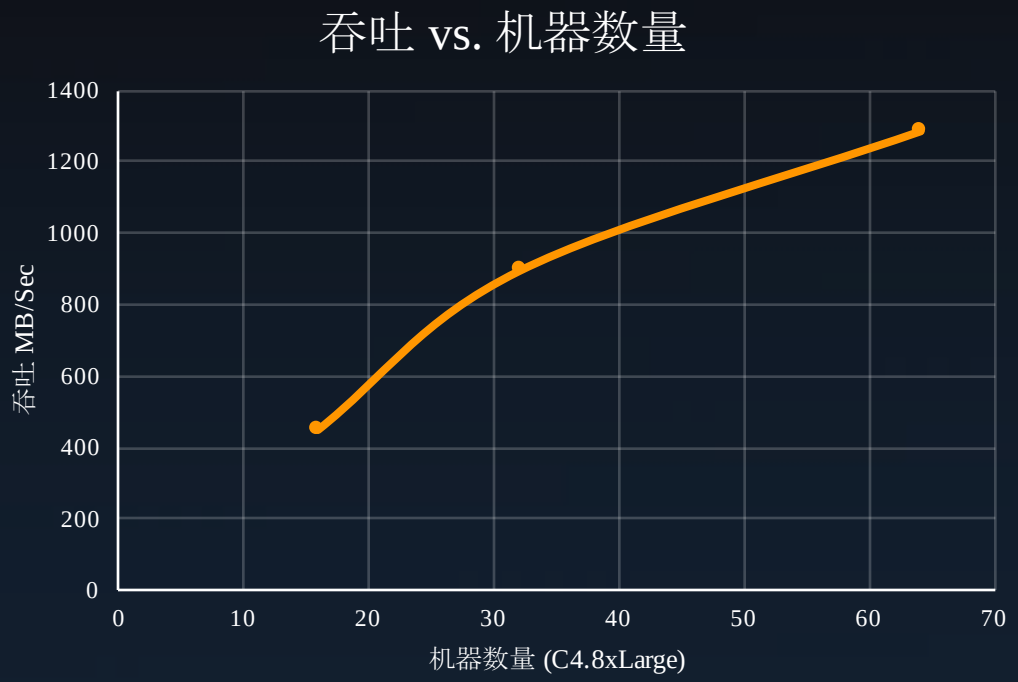
"Microsoft was a poster child for our industry," said Connie Correll, director of communications for the Information Technology Industry Council, a trade organization that represents Amazon, Dell, Dell and I.B.M., among others.



Boosted 决策树

XGBoost 是最常用的boosted决策树实现

现在在 Amazon SageMaker可以直接使用！

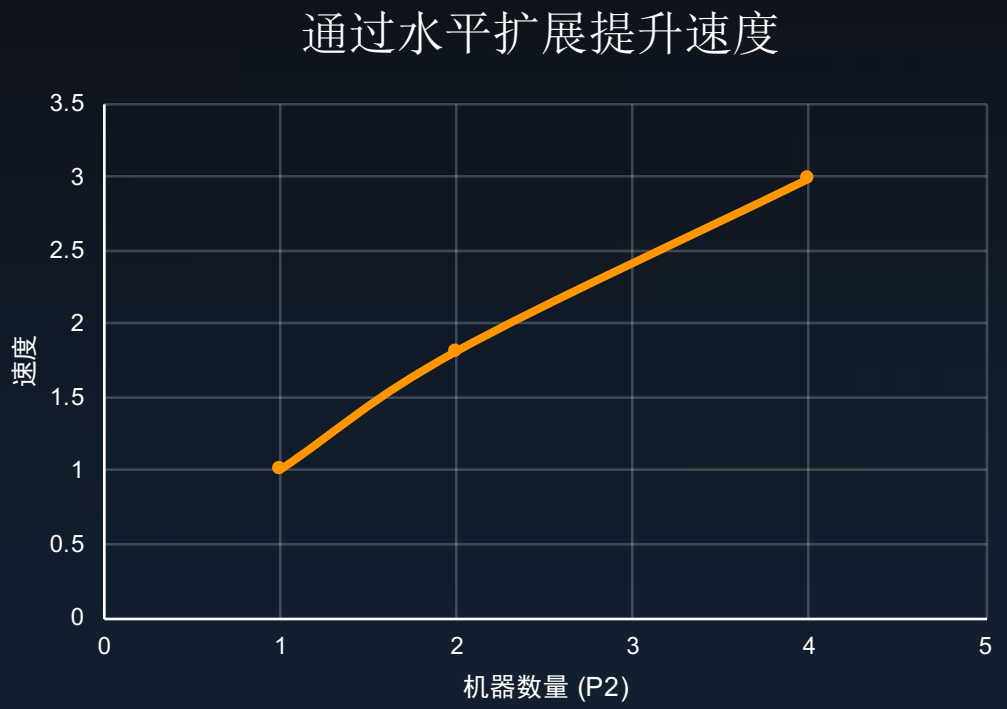


图像分类

通过MxNet 实现的ResNet.

未来会提供更多的网络结构，如 DenseNet， Inception

迁移学习: 从ImageNet上已经训练好的模型开始工作



3

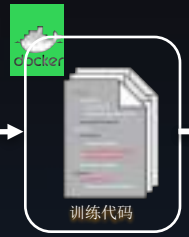
全托管的、灵活的 分布式训练

I

模型训练服务

安全保障

获取训练数据



训练代码

保存模型

保存推理镜像

Amazon ECR

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

AWS提供的算法

mxnet

TensorFlow™

自带脚本 (SM构建容器)

Apache Spark

集成

自带算法（自己通过容器构建）

CPU GPU

HPO

全托管

aws

4

轻松将模型部署到Amazon SageMaker

I

ML部署服务

- ✓ 自动扩展的推理API
- ✓ A/B测试
- ✓ 低延迟 & 高吞吐
- ✓ 使用自己的模型
- ✓ Python SDK

Amazon SageMaker

aws

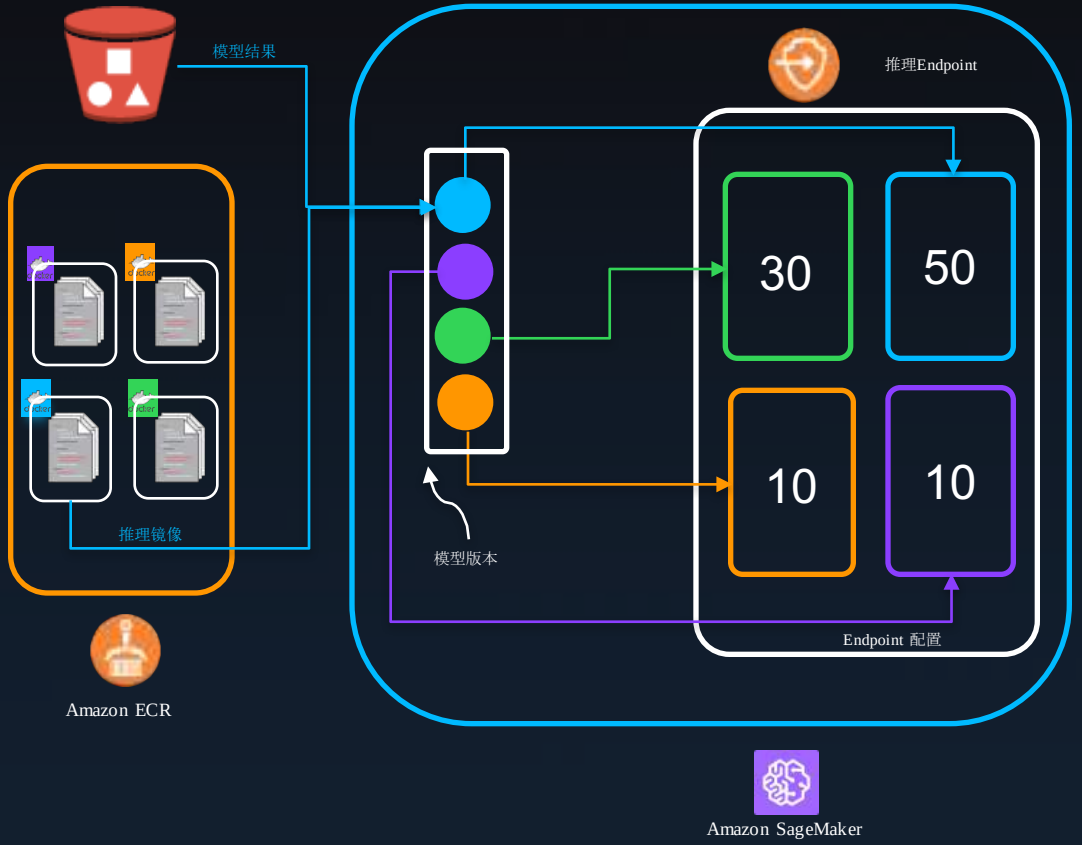
4

轻松将模型部署到Amazon SageMaker

I

模型部署服务

在推理容器中保存多个版本的镜像. Prod 是主要版本，支持50% 用户流量



实例类型: c3.4xlarge

初始实例数量: 3

模型名称: prod

版本名称: primary

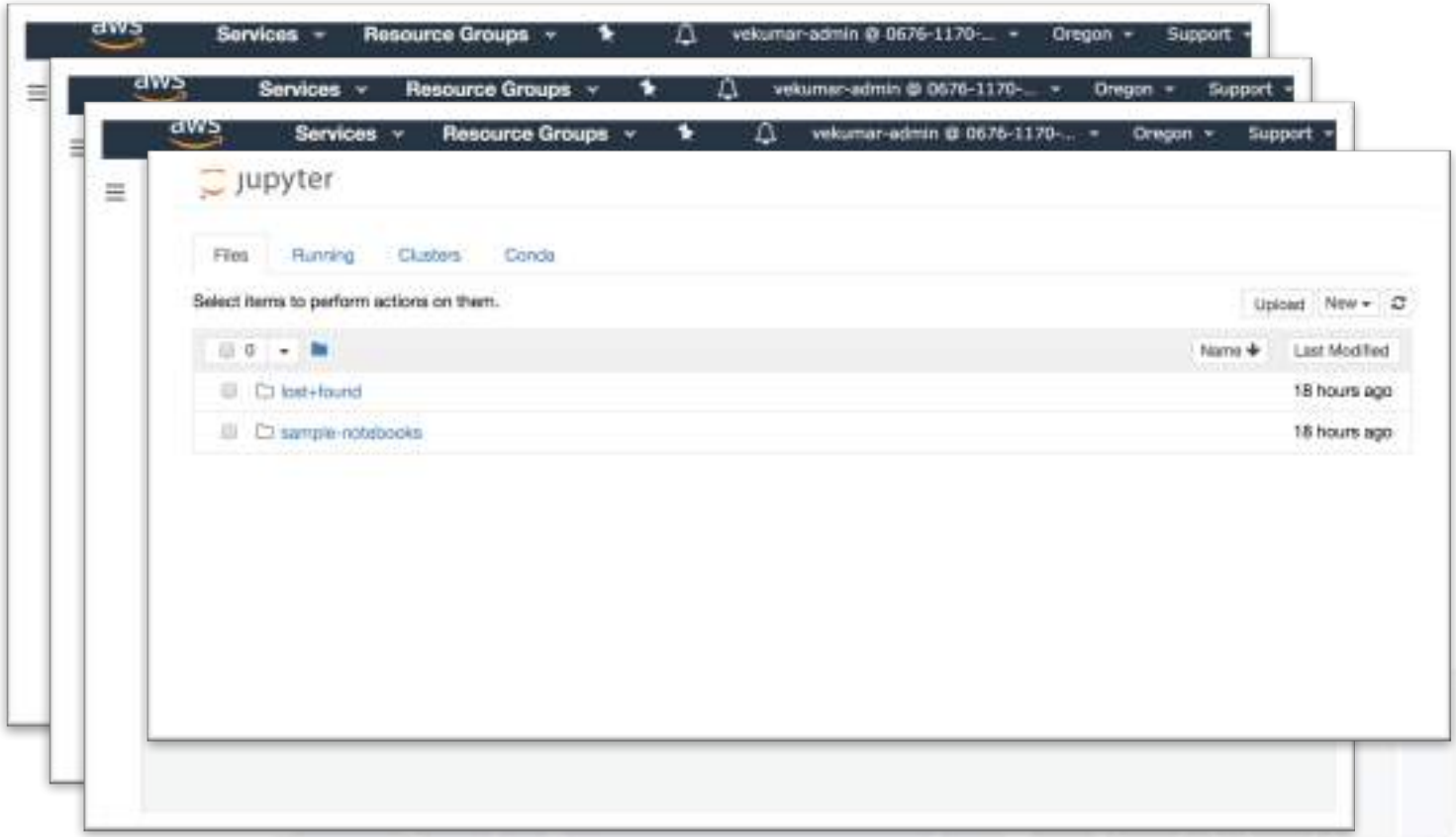
初始版本权重: 50

生产版本

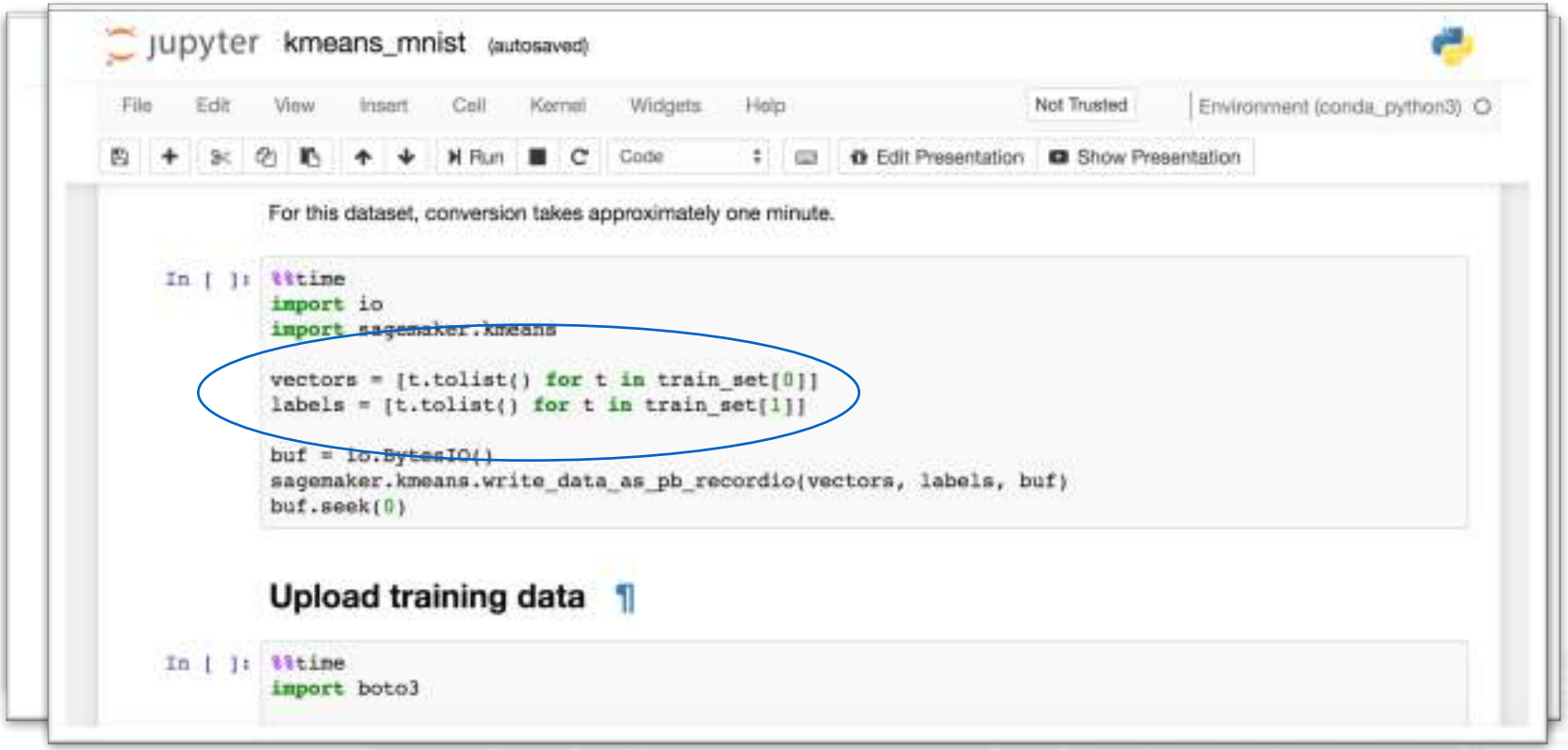
通过Endpoint配置 创建一个新的 Endpoint

如何使用Amazon SageMaker

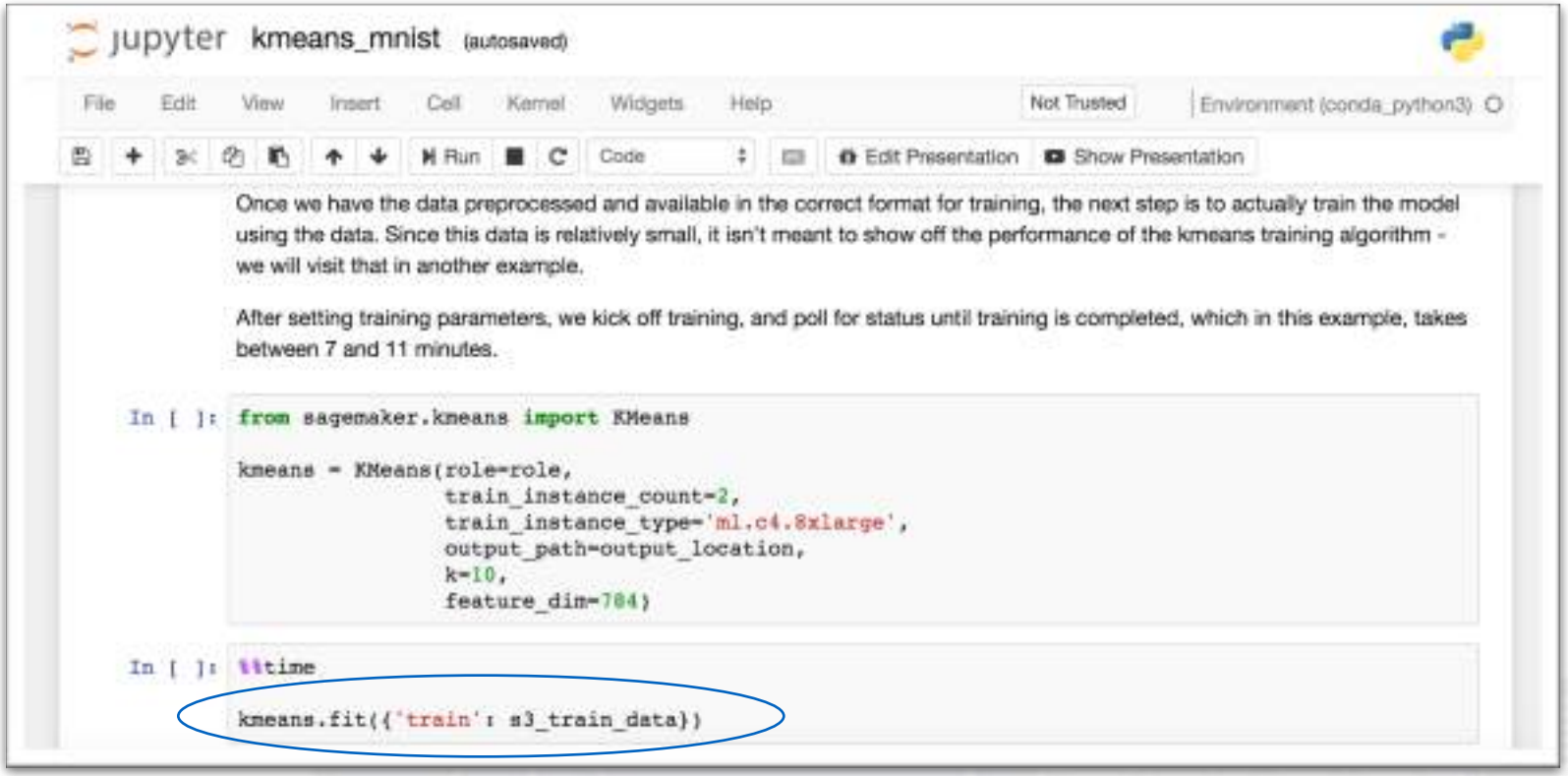
从Notebook样例开始



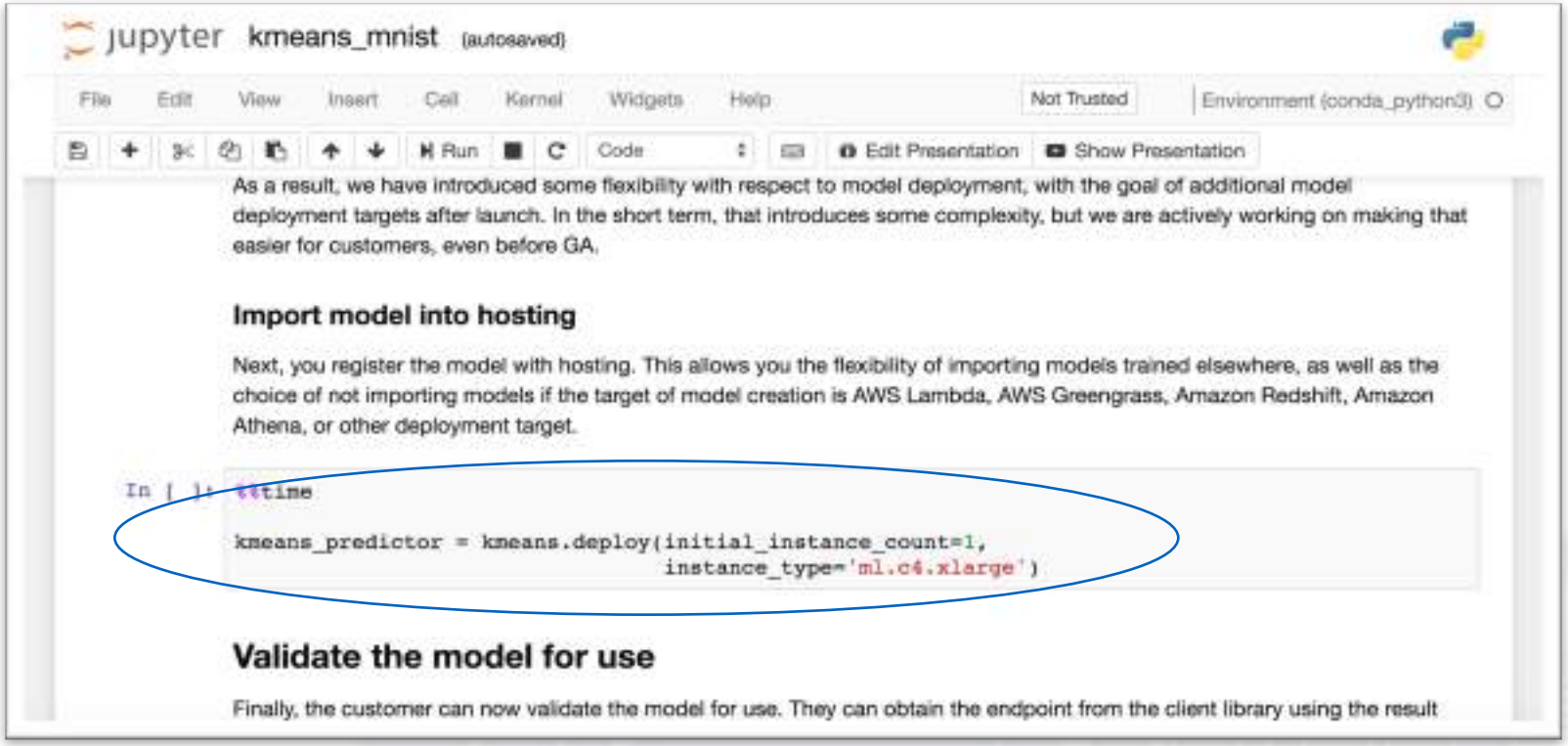
配置训练数据集


jupyter kmeans_mnist (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Environment (conda_python3)
Run
Code
Edit Presentation
Show Presentation
For this dataset, conversion takes approximately one minute.
In [ ]:
%%time
import io
import sagemaker.kmeans
vectors = [t.tolist() for t in train_set[0]]
labels = [t.tolist() for t in train_set[1]]
buf = io.BytesIO()
sagemaker.kmeans.write_data_as_pb_recordio(vectors, labels, buf)
buf.seek(0)
Upload training data ¶
In [ ]:
%%time
import boto3

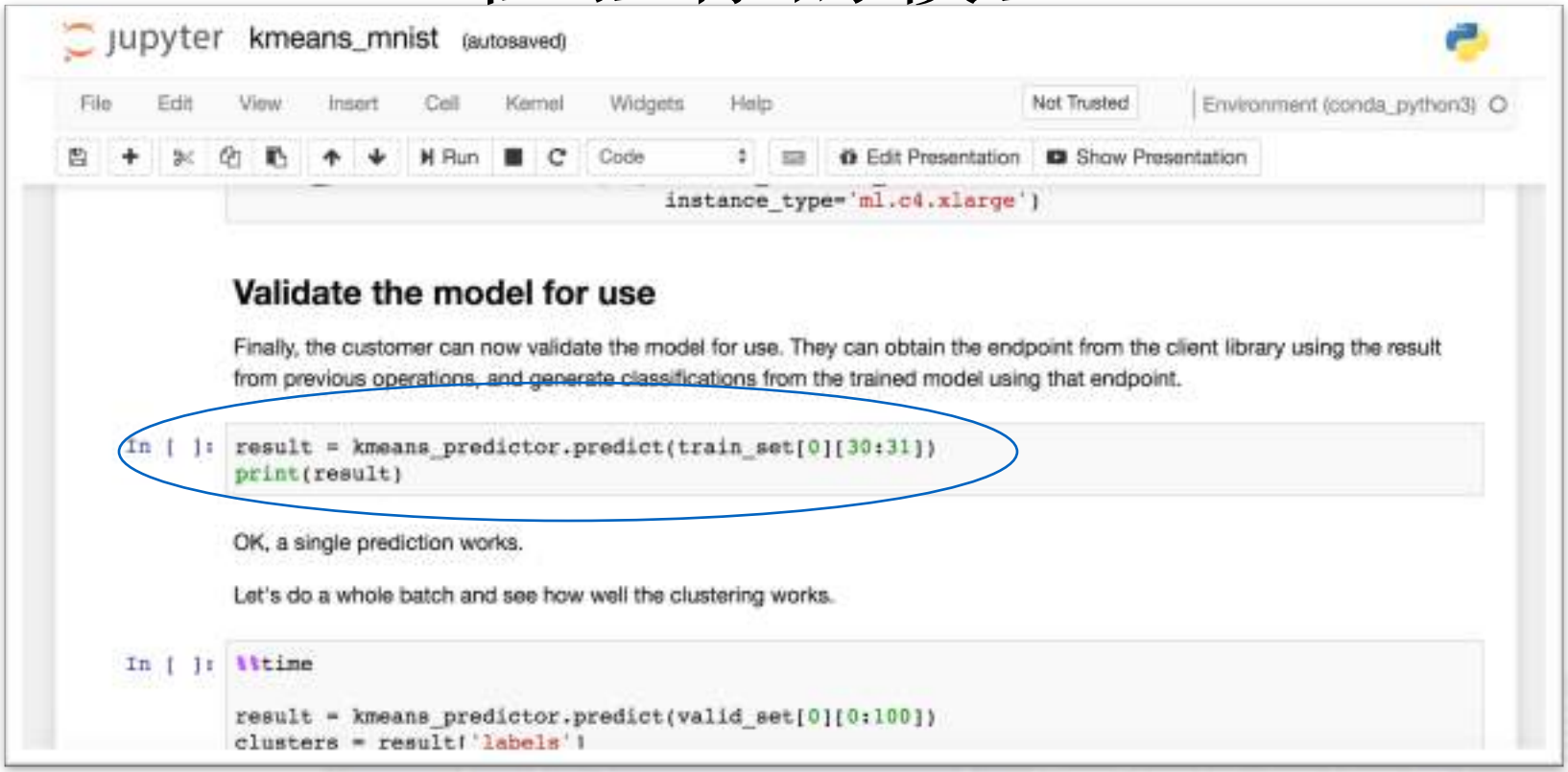
训练你的模型


jupyter kmeans_mnist (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Environment (conda_python3)
Run
Code
Edit Presentation
Show Presentation
Once we have the data preprocessed and available in the correct format for training, the next step is to actually train the model using the data. Since this data is relatively small, it isn't meant to show off the performance of the kmeans training algorithm - we will visit that in another example.
After setting training parameters, we kick off training, and poll for status until training is completed, which in this example, takes between 7 and 11 minutes.
In [ ]:
from sagemaker.kmeans import KMeans
kmeans = KMeans(role=role,
train_instance_count=2,
train_instance_type='ml.c4.8xlarge',
output_path=output_location,
k=10,
feature_dim=784)
In [ ]:
%%time
kmeans.fit({'train': s3_train_data})

部署你的模型


jupyter kmeans_mnist (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Environment (conda_python3)
Run
Code
Edit Presentation
Show Presentation
As a result, we have introduced some flexibility with respect to model deployment, with the goal of additional model deployment targets after launch. In the short term, that introduces some complexity, but we are actively working on making that easier for customers, even before GA.
Import model into hosting
Next, you register the model with hosting. This allows you the flexibility of importing models trained elsewhere, as well as the choice of not importing models if the target of model creation is AWS Lambda, AWS Greengrass, Amazon Redshift, Amazon Athena, or other deployment target.
In [ ]:
%%time
kmeans_predictor = kmeans.deploy(initial_instance_count=1,
instance_type='ml.c4.xlarge')
Validate the model for use
Finally, the customer can now validate the model for use. They can obtain the endpoint from the client library using the result

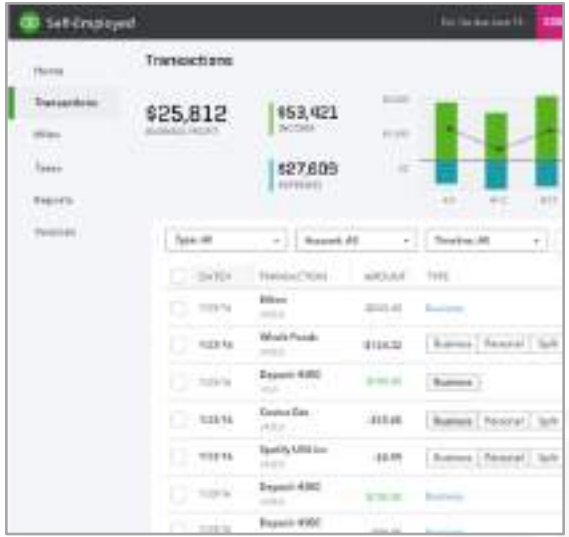
检验你的模型


jupyter kmeans_mnist (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Environment (conda_python3)
Run
Code
Edit Presentation
Show Presentation
instance_type='ml.c4.xlarge')
Validate the model for use
Finally, the customer can now validate the model for use. They can obtain the endpoint from the client library using the result from previous operations, and generate classifications from the trained model using that endpoint.
In [ ]:
result = kmeans_predictor.predict(train_set[0][30:31])
print(result)
OK, a single prediction works.
Let's do a whole batch and see how well the clustering works.
In [ ]:
%%time
result = kmeans_predictor.predict(valid_set[0][0:100])
clusters = result['labels']

Amazon SageMaker案例分享

AI and ML at Intuit

产品推荐



欺诈检测与防止

用户关怀与建议

intuit

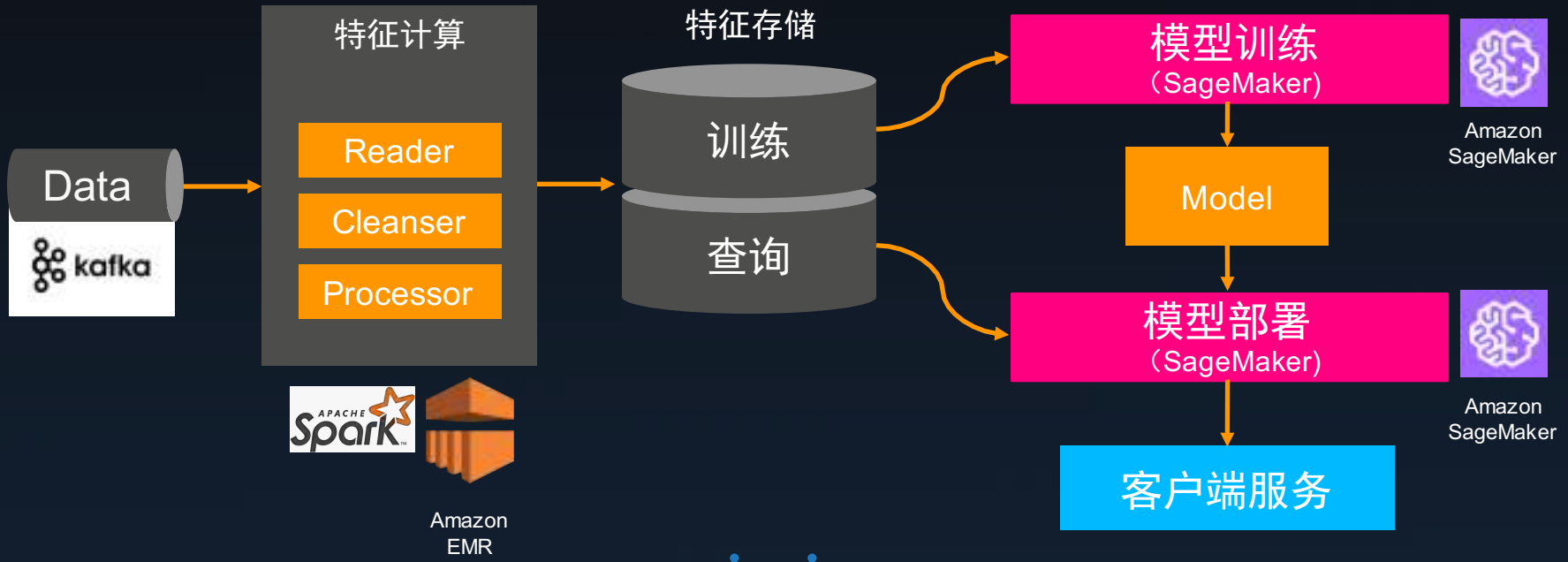
Intuit使用SageMaker获得的好处

From	To
需要临时设置 和管理notebook环境	使用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

intuit

aws

使用SageMaker在AWS上构建近实时的欺诈检测



Amazon SageMaker – 下一步动作

- 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
- SageMaker 样例: https://github.com/awslabs/amazon-sagemaker-examples
- 把你构建的应用告诉我们!

Amazon SageMaker

端到端的托管机器学习平台

!GO BUILD!

aws

极客邦科技
企业培训与咨询

Geekbang.

扫码关注

获取更多培训信息

关注QCon微信公众号，
获得更多干货！

Thanks!

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

Geekbang. InfoQ