

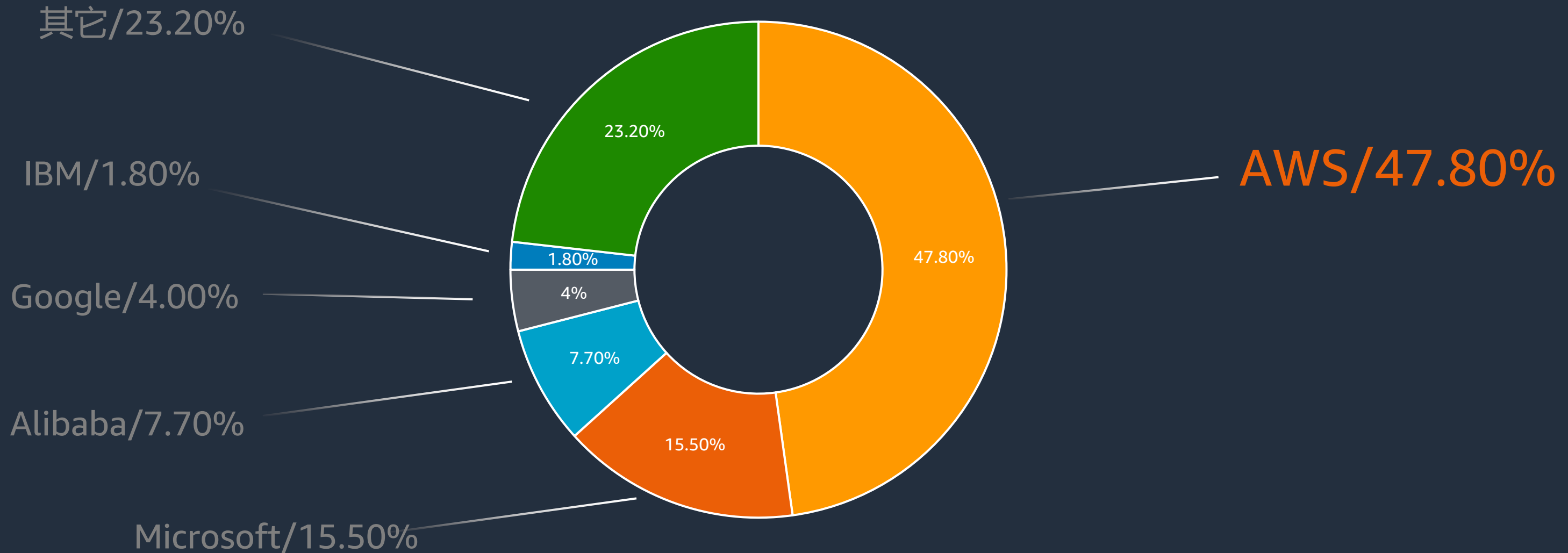


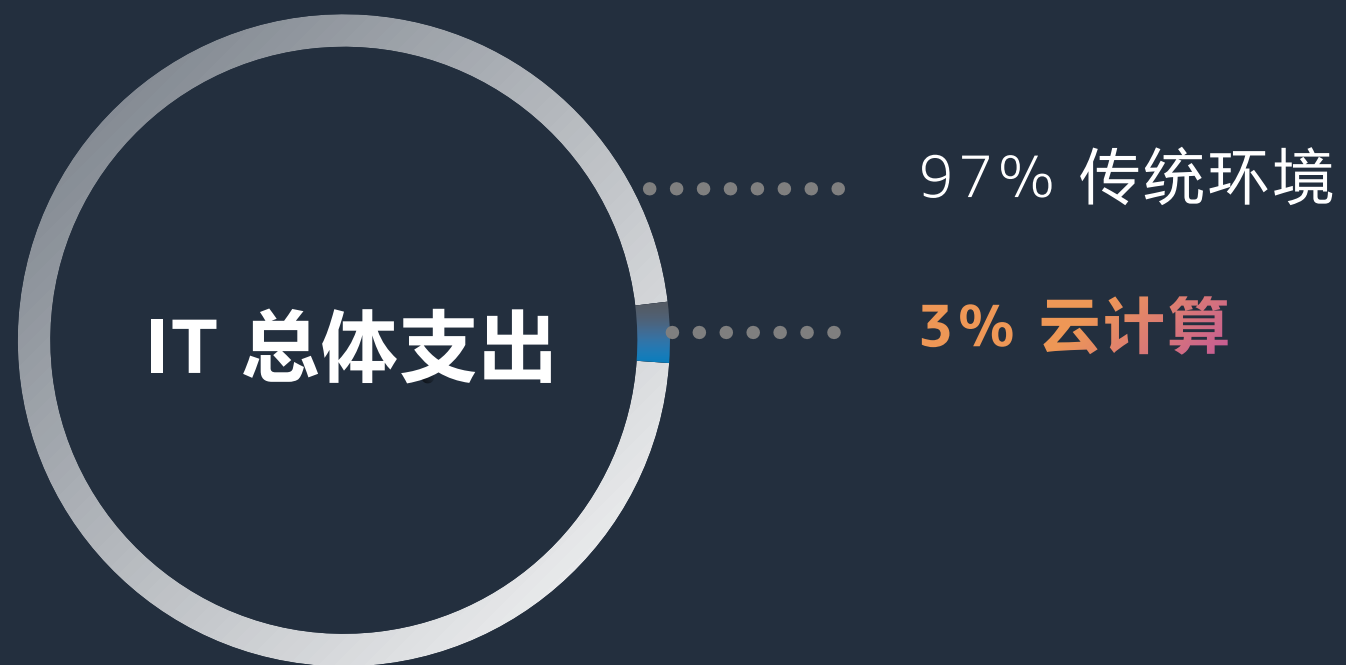
坐看云起，洞见未来

费良宏

Developer Advocate | Amazon Web Services

Gartner 全球 IaaS 市场份额





云计算还处于早期阶段

AWS re:Invent

DECEMBER 2 – 6, 2019 | LAS VEGAS, NEVADA



参会人数 65,000



4 场主题演讲



Monday Night Live



Andy Jassy



Global Partner Summit



Dr. Werner Vogels

3,000 余场技术分享



AWS
re:Invent **2019**

发布 **250+** 项
新的服务和特性

AWS
re:Invent

© 2019, A



re:Invent

Products

Solutions

Pricing

Documentation

Learn

Partner Network

AWS Marketplace

Customer Enablement

Events

Explore More



Contact Sales

Support

English

My Account

Create an AWS Account

AWS re:Invent 2019 Announcement Page

Watch the Keynote Addresses



AWS re:Invent 2019 - Keynote with Andy Jassy



AWS re:Invent 2019 - Keynote with Dr. Werner Vogels



Monday Night Live with Peter DeSantis



2019 re:Invent Global Partner Keynote with Doug Yeum



Announcing AWS DeepComposer with Jonathan Coulton

- Explore more recap videos on our [Re:Invent 2019 Youtube Channel](#) -

Featured Product Announcements

日程

- 计算
- 存储
- 数据库 & 数据分析
- 安全 & 网络
- 无服务器计算
- 基础设施 & 区域外的扩展
- 人工智能 & 机器学习

计算

介绍

Amazon EC2 Inf1 实例

云计算中速度最快、成本最低的机器学习推理

AWS Inferentia, 第一款 AWS 设计的机器学习芯片

与使用 GPU 支持的 G4 实例相比, 可以提高**3倍**的吞吐量并且每次推理的成本降低**40%**



语音识别



图像处理



欺诈检测



目标检测



自然语言处理



个性化

介绍

AWS Graviton2 处理器

为云计算上的工作负载提供最佳的价格/性能

Graviton1 处理器



主流云计算上的第一个基于 ARM 的处理器

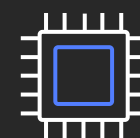


内置 64 位 ARM Neoverse 内核，采用 AWS 设计的 16 纳米芯片



多达 16 个 vCPU、10 Gbps 增强型网络、3.5 Gbps EBS 带宽

Graviton2 处理器



采用 64 位 ARM Neoverse 内核构建，采用 AWS 设计的 7 纳米硅工艺



高达 64 颗 VCPU, 20 Gbps 增强网络, 14 Gbps 的 EBS 带宽



7x 性能, 4x 计算核心, 以及 5x 更快的内存

介绍

基于 AWS Graviton2 的实例

对于通用任务、计算密集和内存密集型工作负载，其性价比提高了40%

M6g

专为：通用工作负载，如应用程序服务器、中型数据存储和微服务

实例存储选项：M6gd

现已 预览

C6g

专为：计算密集型应用程序（如 HPC、视频编码、游戏和仿真工作负载）构建

实例存储选项：C6gd

将于2020年发布

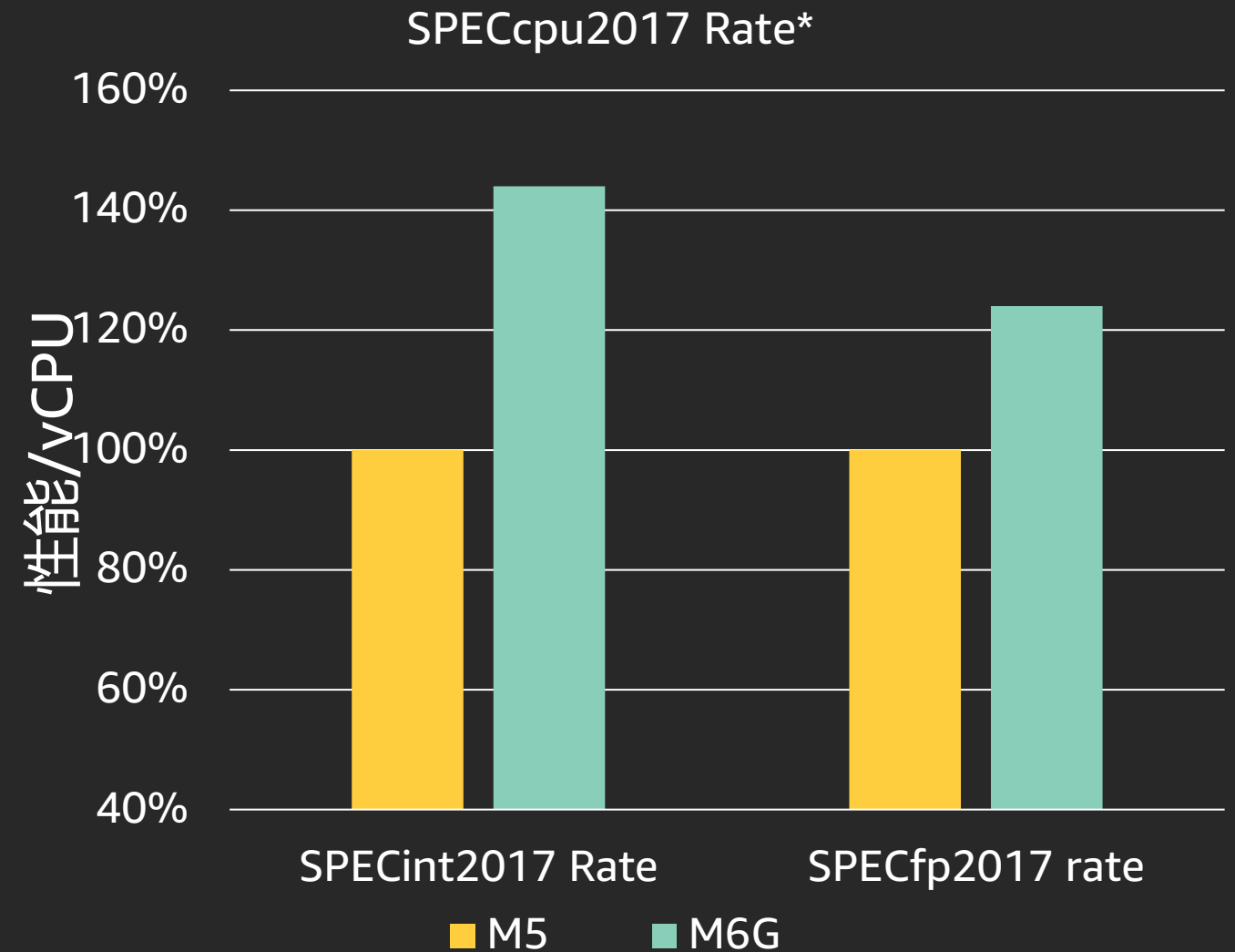
R6g

专为：内存密集型工作负载（如开源数据库或内存中缓存）构建

实例存储选项：R6gd

SPEC cpu2017

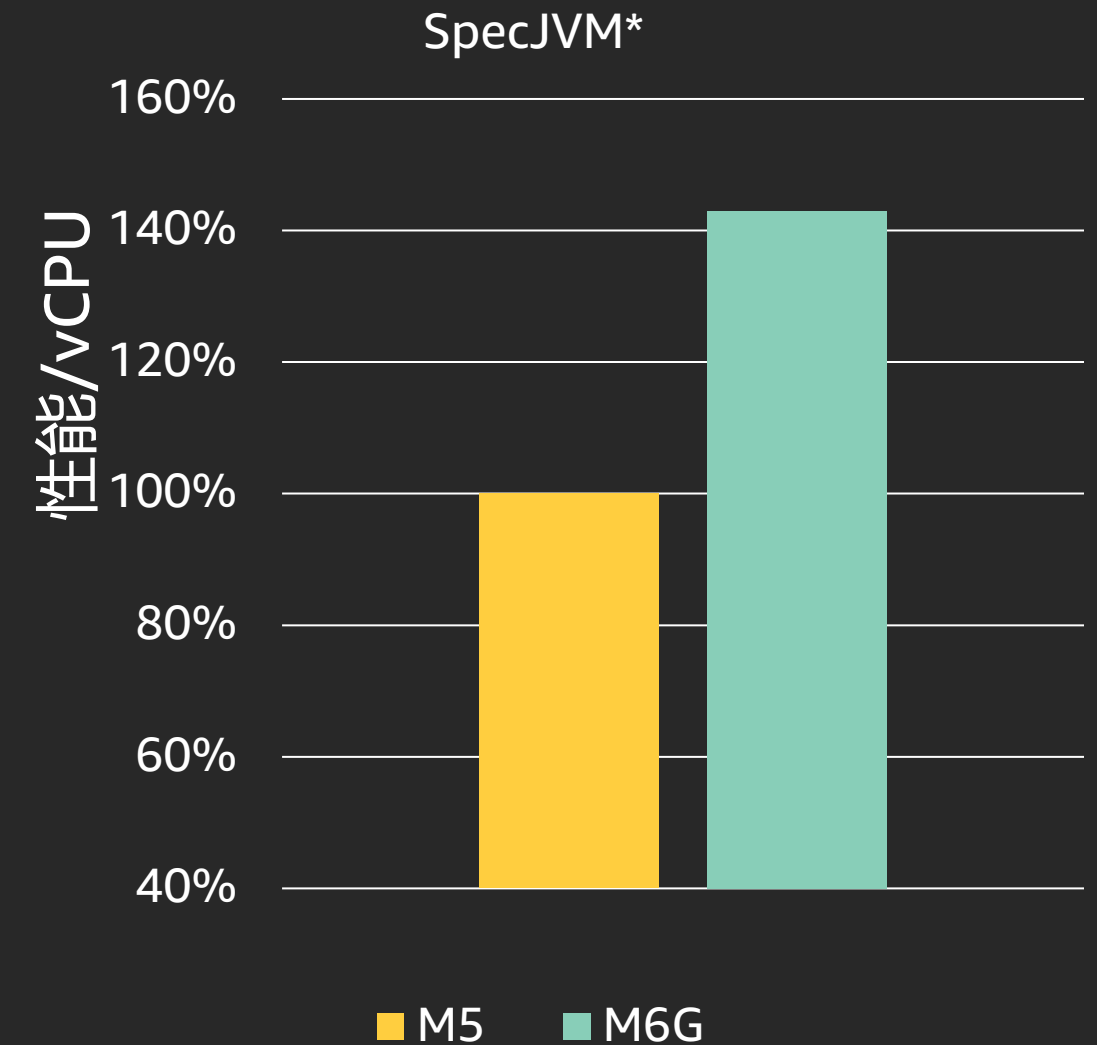
- 行业标准 CPU 密集型基准测试
- 并发地运行在所有的 VCPU 上
- 性能比较/vCPU



* All SPEC scores estimates, compiled with GCC9 -O3 -march=native, run on largest single socket size for each instance type tested.

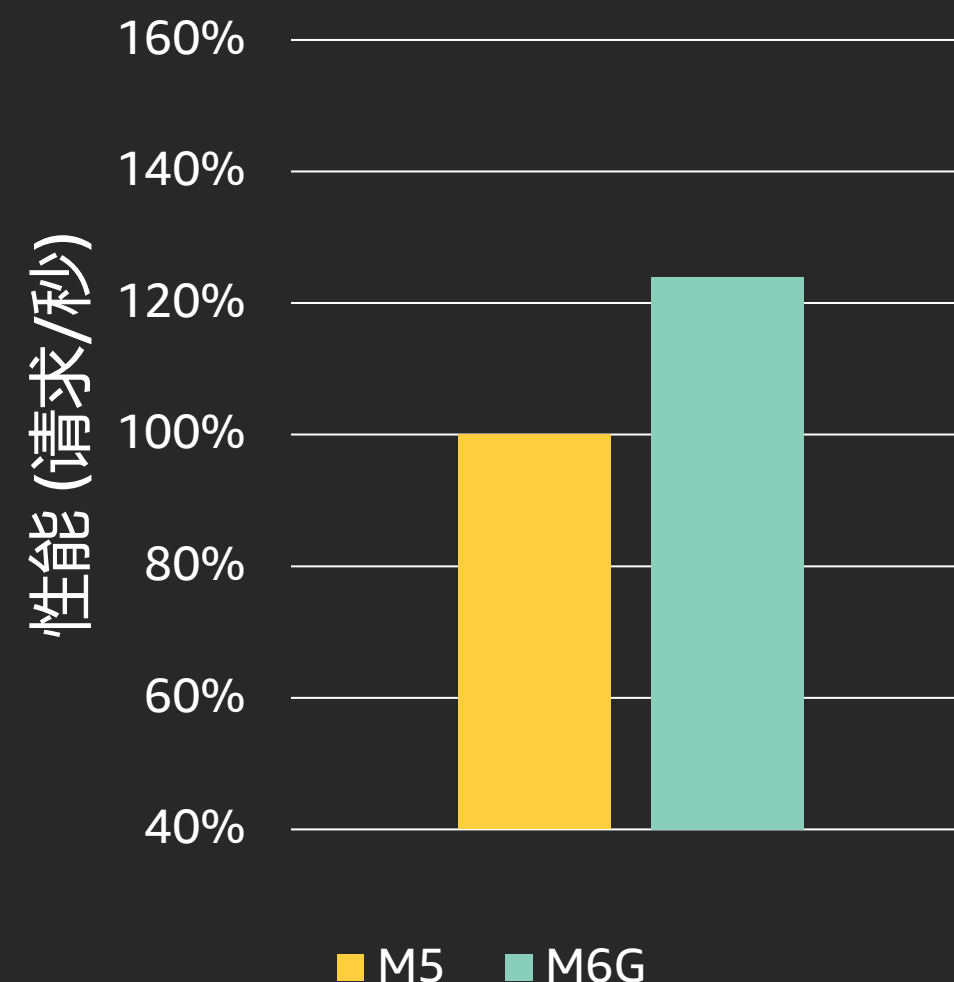
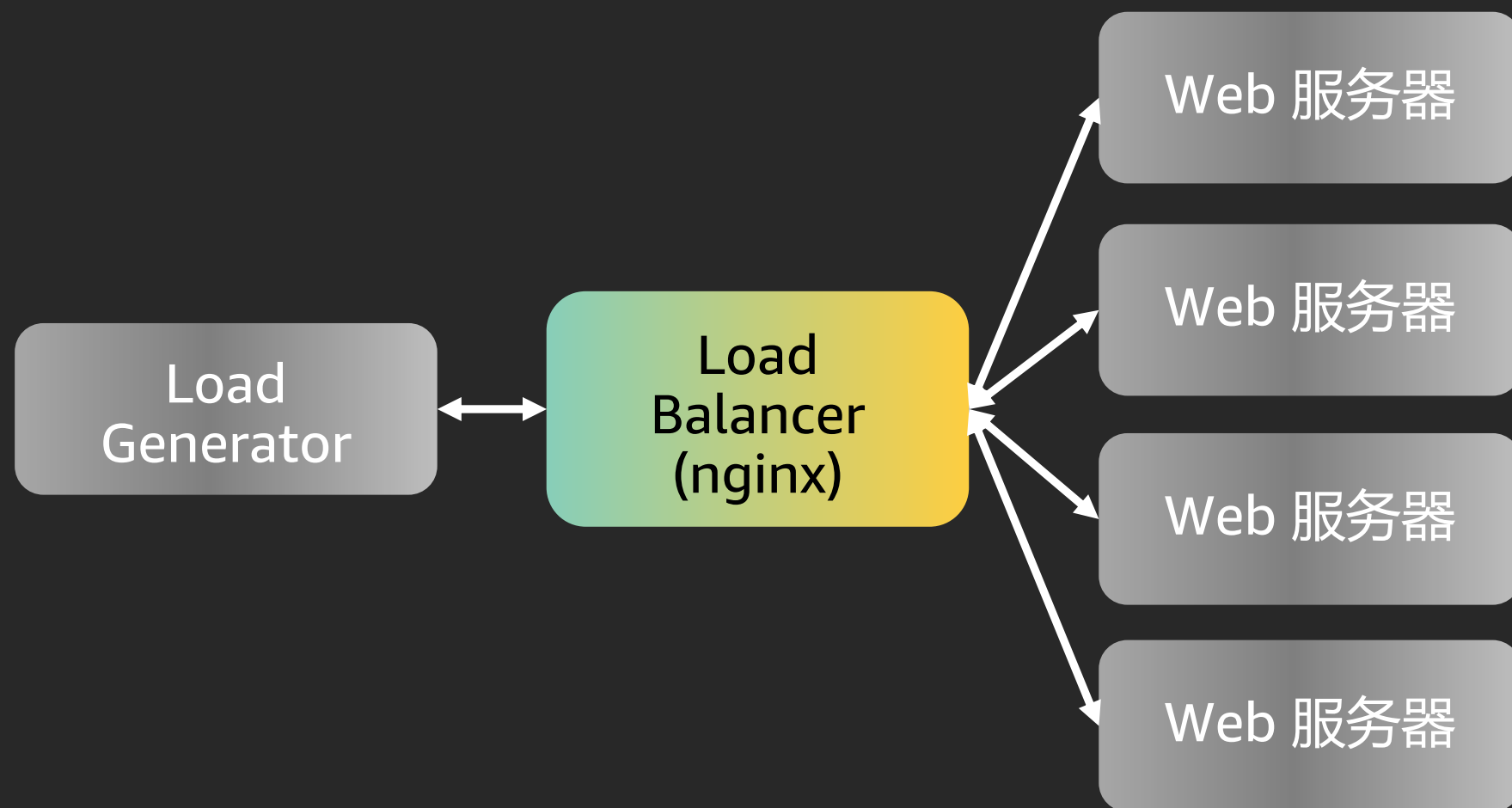
SPEC jvm2008

- Java VM 基准测试
-
- 并发地运行在所有的 VCPU 上
- 性能比较/vCPU



* All SPEC scores estimates, run with OpenJDK11 and skipping compiler* and startup.* tests
Tests run on largest single-socket instance size for each instance type tested.

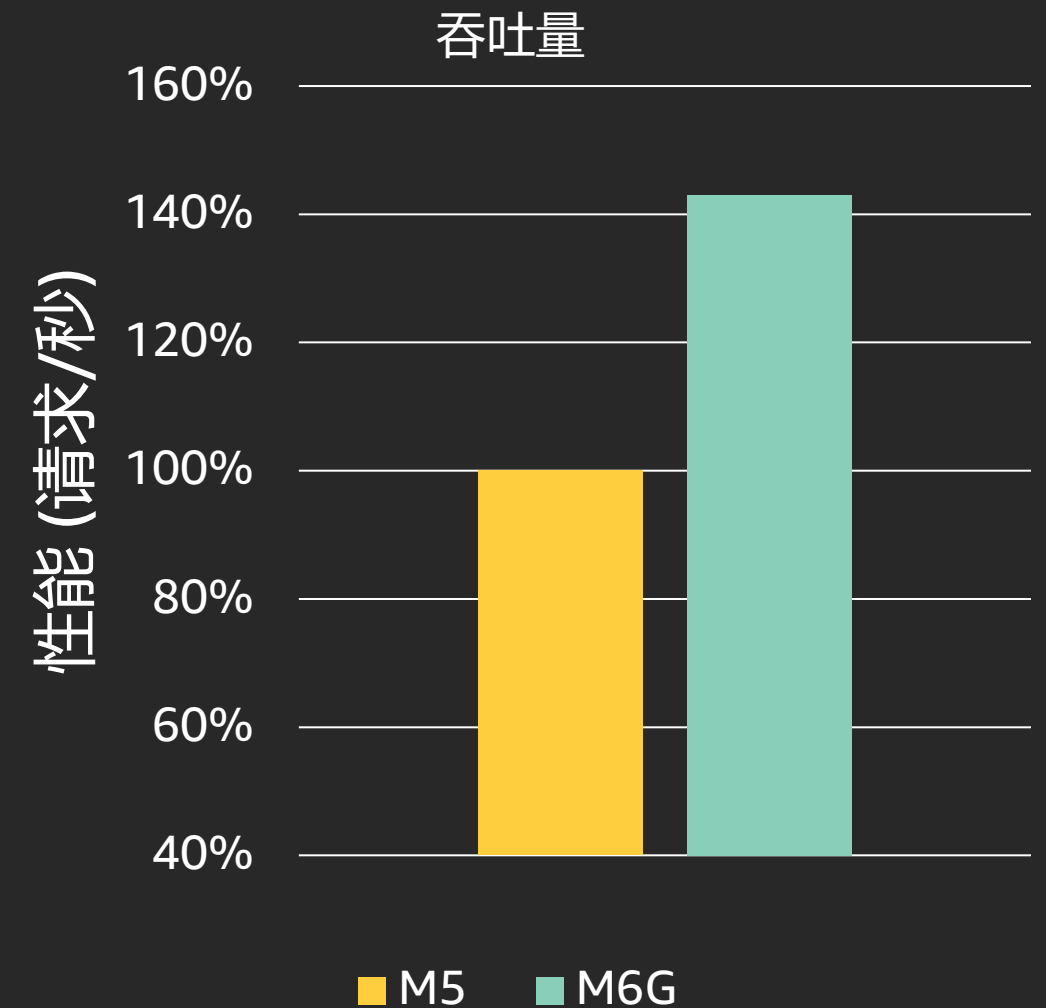
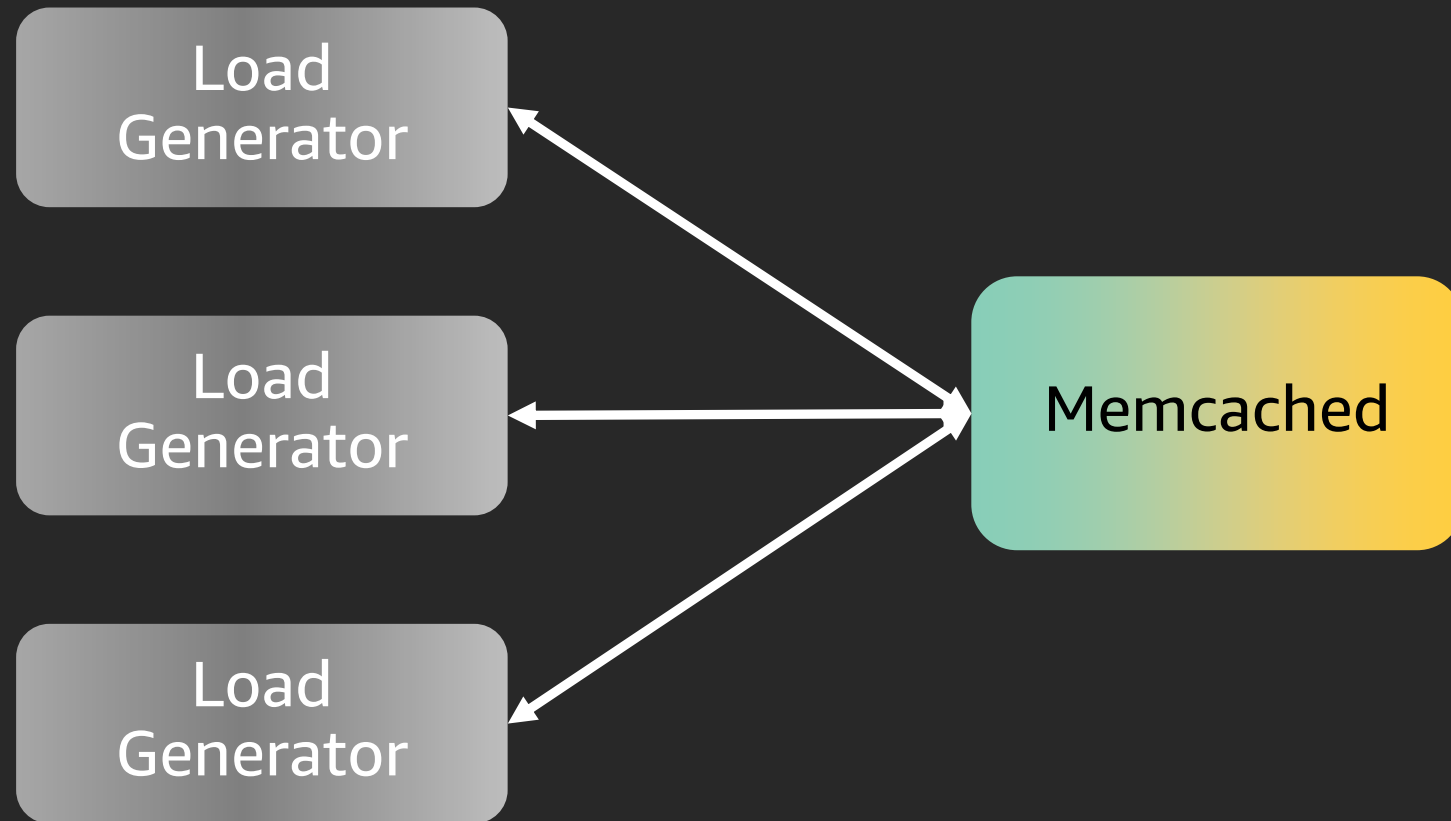
使用 Nginx 进行负载均衡



NGINX v1.15.9, 512 clients, 128 GET/POST payloads, all HTTPS connections, AES128-GCM-SHA256, OpenSSL 1.1.1, 4 target machines, all tests run on 4xl size; load generator c5.9xl; web servers c5.4xls; All servers run in a cluster placement group

Memcached

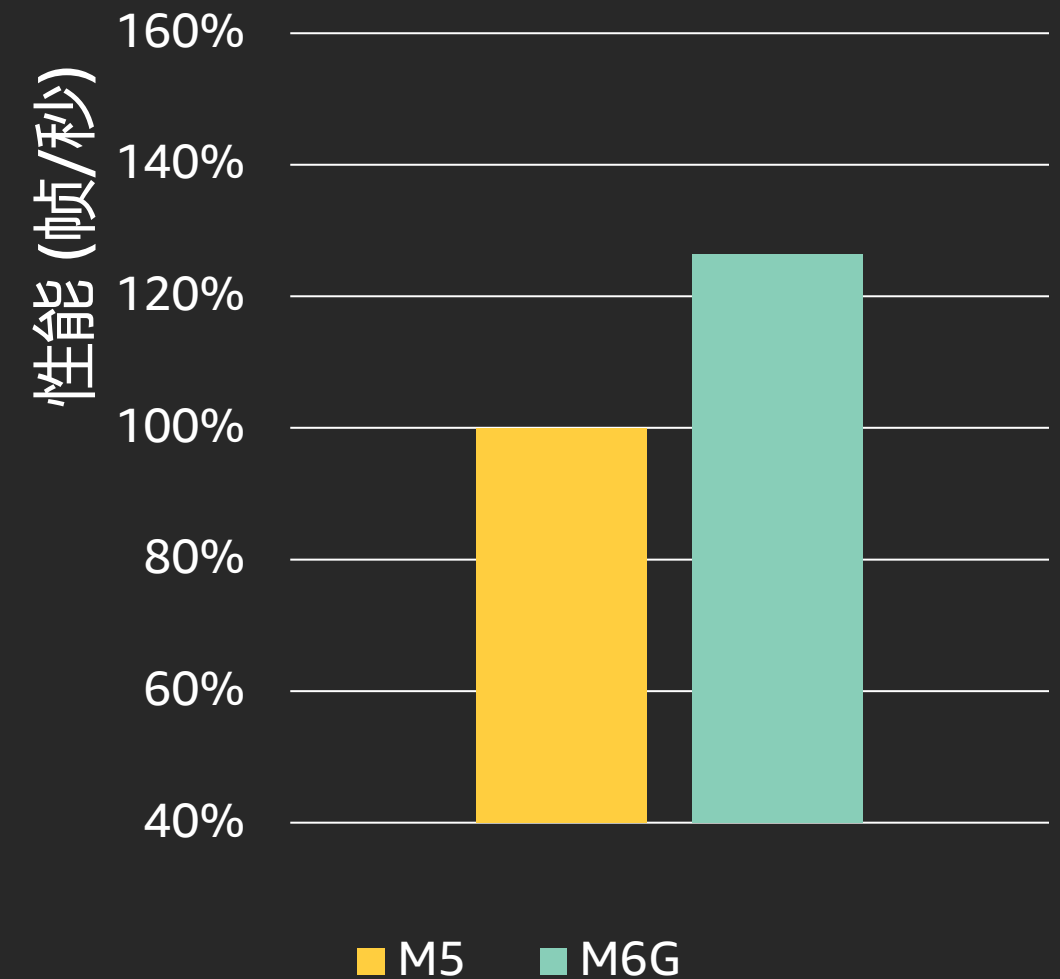
计算



Memcached v1.5.16, 16B keys, 128B values, 7.8M KV-pairs, 576 connections for load generation from 2x c5.9xlarge instances, 16 additional connections used to measure latency from 1 additional c5.9xlarge; each connection maintains 4096 outstanding requests; All servers in a cluster placement group

x264 标准的媒体转码

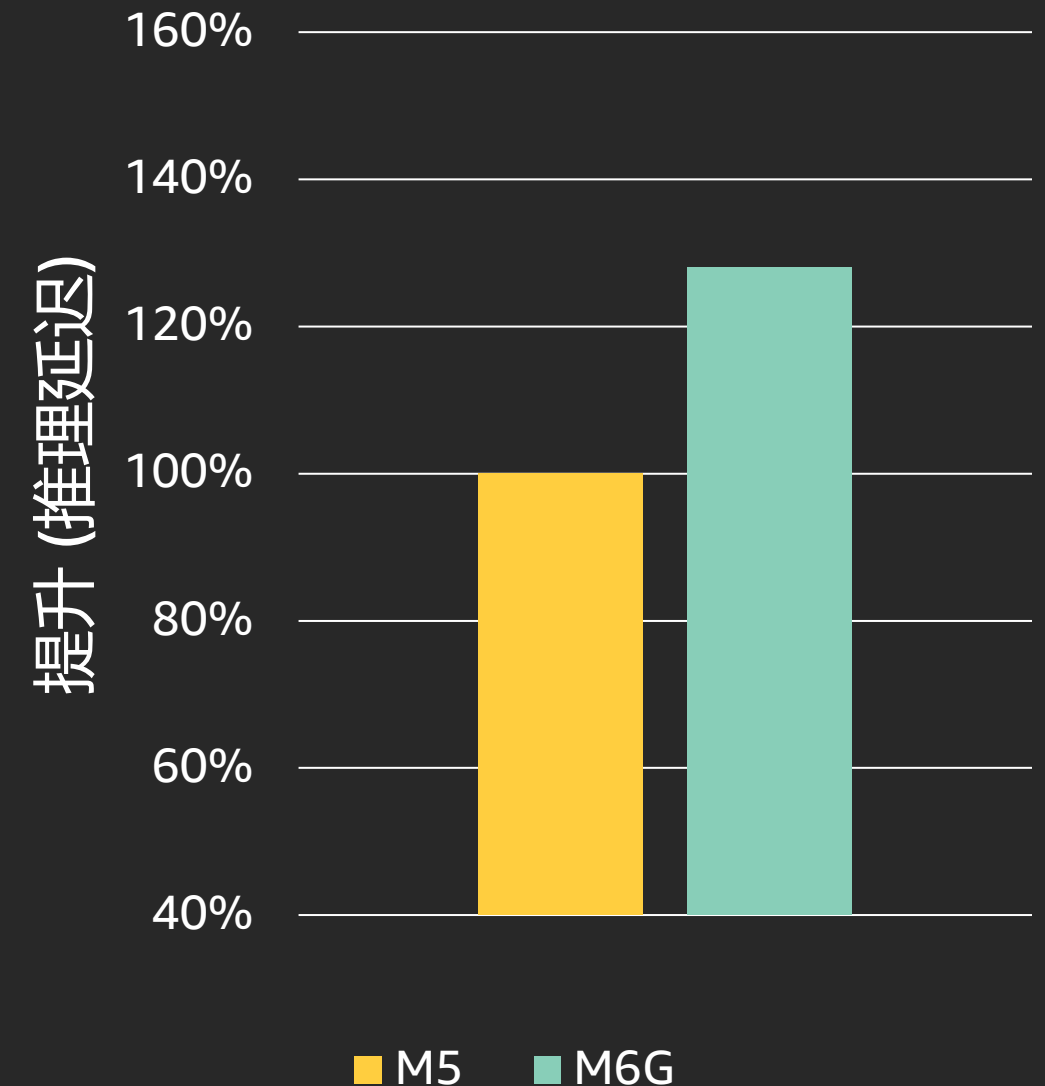
- 每天创建大量视频
- 转码减少了带宽，以提供和存储该视频
- 使用 libx264 编码器编码未压缩 1080p 到 H.264



X264 (videolan.org version 3759fcb7), 4xl instance size, medium preset, input uncompressed 1080p50, output encoded h264 1080p50

机器学习

- BERT: 最先进的(SOTA)文本编码器
 - 特征表示 (编码) 对于机器学习至关重要
 - 深度神经网络用于文本特征表示
- 支持 fp16 和 int8 来加速机器学习的负载
- M6g 性能优于 M5
 - 带 AVX-512 的 M5 仅限于 FP32
 - 借助 FP16 支持 M6g, M6g 性能更好, 适用于基于 CPU 的推理

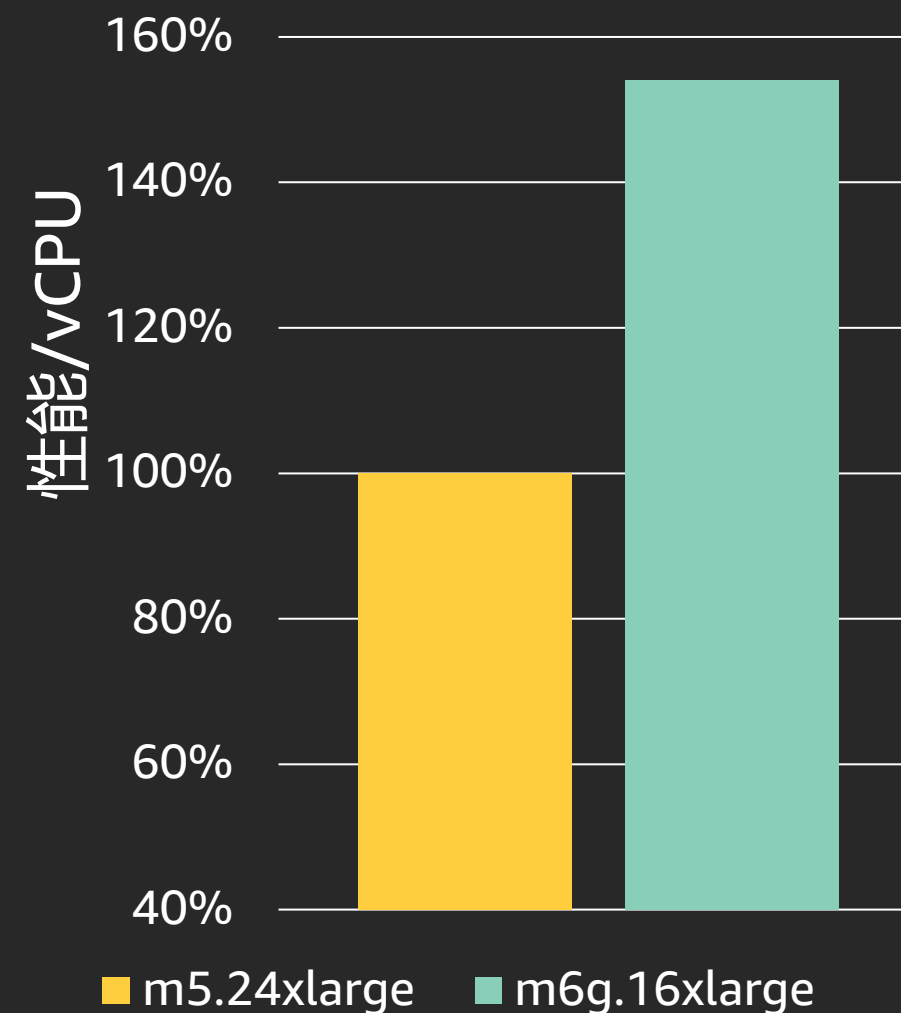


BERT classification using TVM and 64 length sequence on CPUs
Batch size of one; dedicated instances on xlarge size

EDA 性能 – Arm 与 Cadence Xcellium



- 芯片开发成本高昂
 - 初始设计就必须正确
- Arm 每月使用数百万小时的 CPU 时间来模拟其处理器设计
 - 需求根据项目的阶段而变化很大
 - Amazon EC2的完美用例
- 使用 Cadence Xcellium 模拟 Arm Cortex-A53
 - 在处理器的DPU RTL上进行570个验证模拟



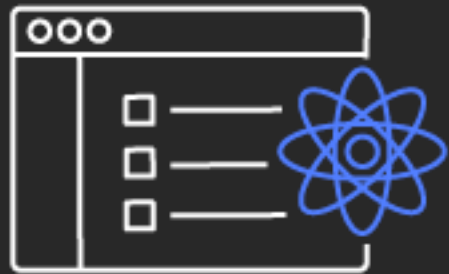
AWS Graviton的每个vCPU为EDA提供了明显更好的性能



介绍

Amazon Braket

完全托管的服务，使科学家和开发人员能够轻松地探索和实验**量子计算**



设计、测试以及运行量子算法的单一环境



D-Wave、IonQ、Rigetti 等提供的量子硬件实验



运行混合量子和经典算法

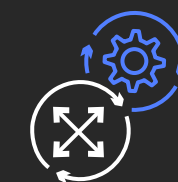
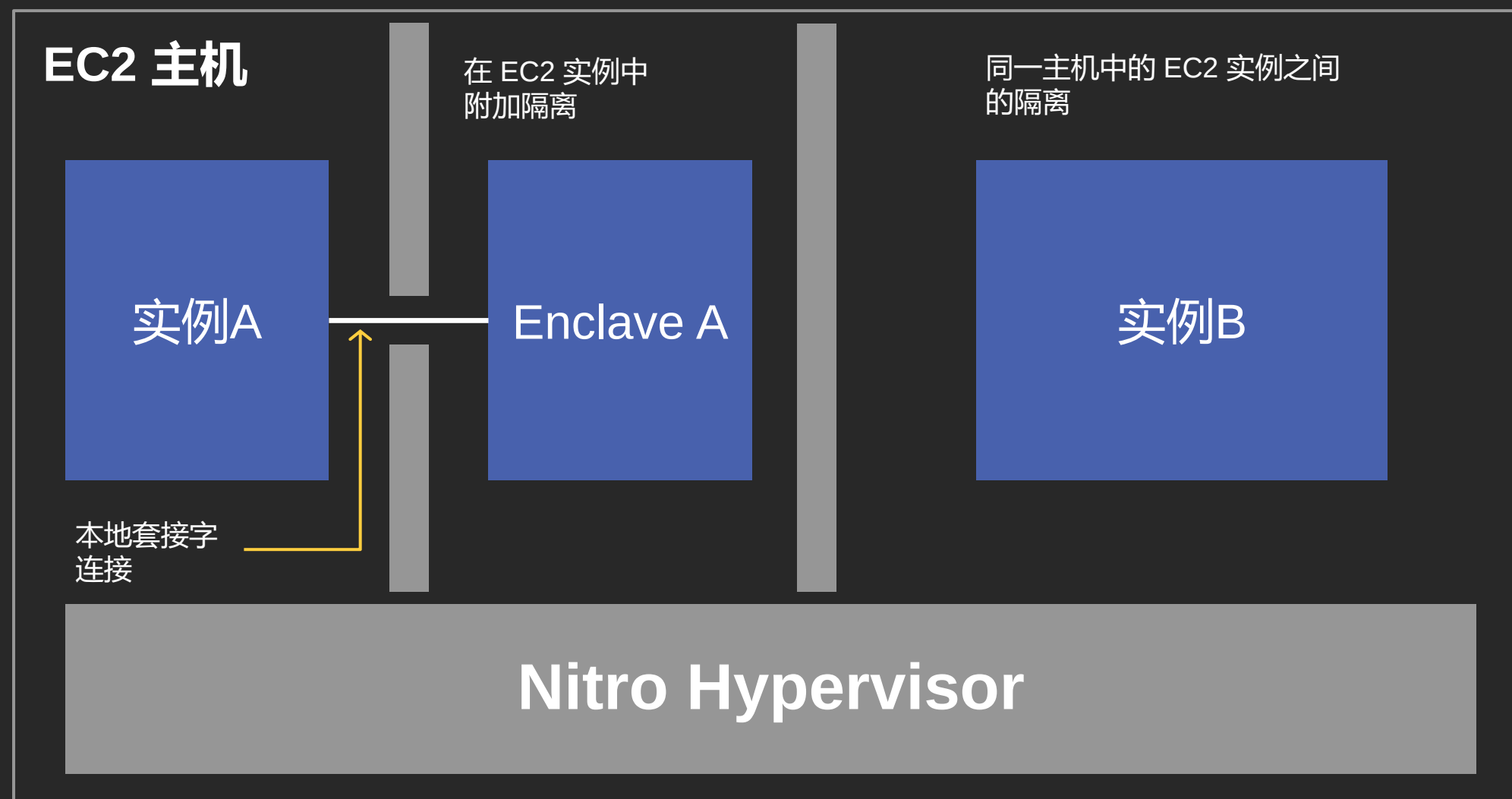


获取专家帮助

介绍

AWS Nitro Enclaves

创建额外的隔离，以进一步保护 EC2 实例中的高度敏感数据



安全隔离



灵活的资源分配



支持各种工作负载

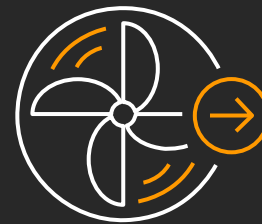
介绍

AWS Compute Optimizer

使用机器学习的推荐引擎确定最佳 EC2 实例和自动缩放组。与 AWS Organize 集成



轻松为工作负载确定最佳的 AWS 资源



在 M、C、R、T 和 X 实例系列中支持超过 140 个 EC2 实例类型选项



为 AWS 工作负载降低成本并提高性能

- Dashboard
- Recommendations per service
 - EC2 Instances
 - AutoScaling Groups

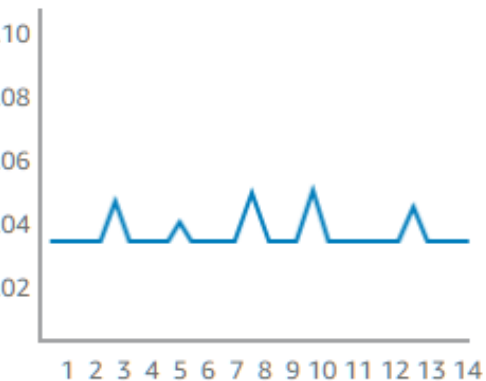
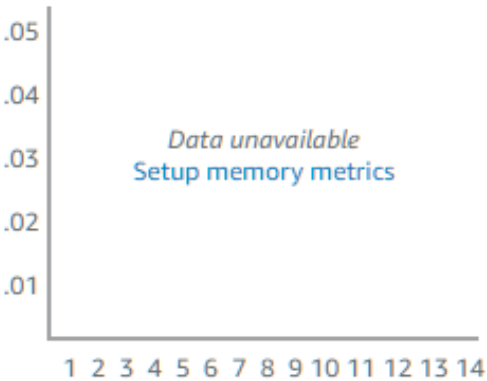
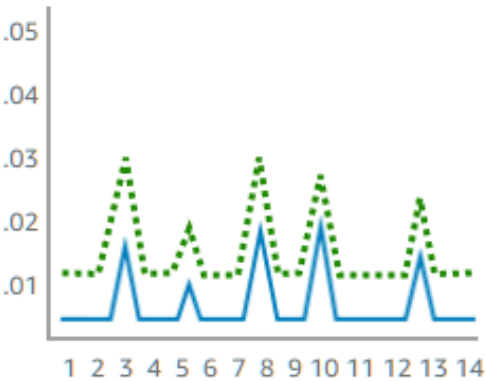
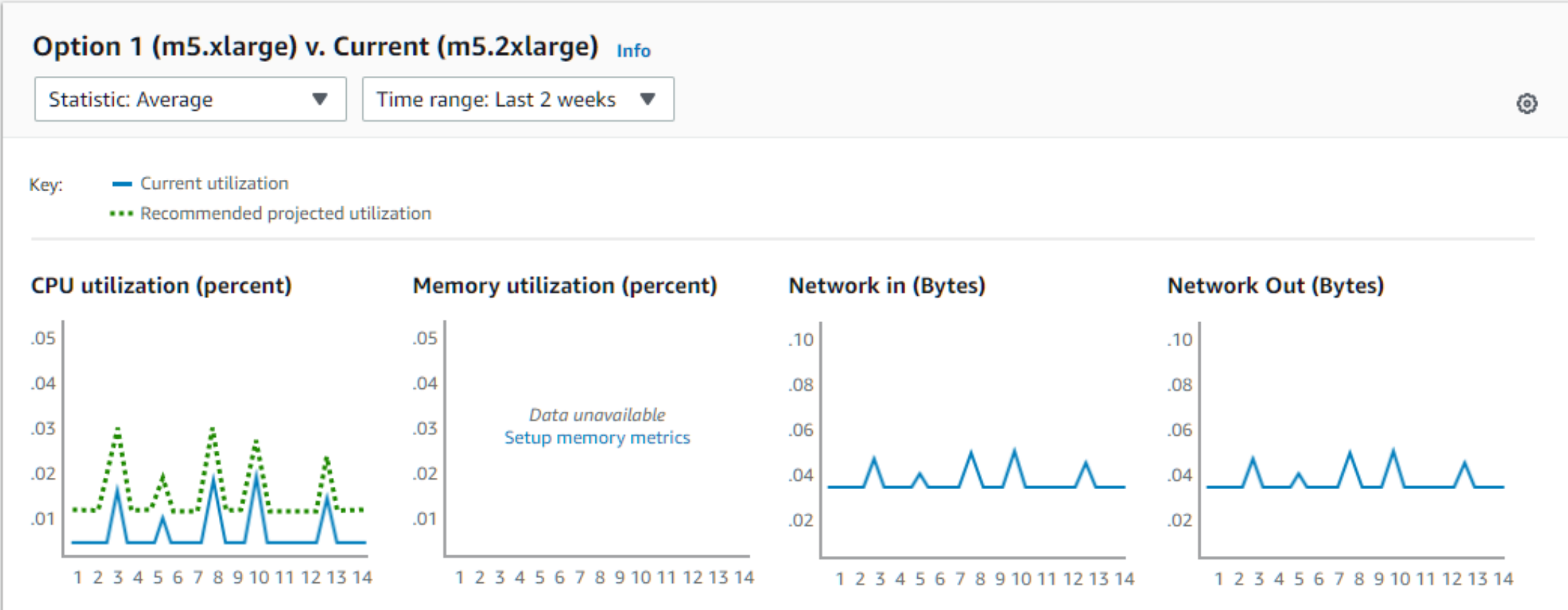
Recommendation detail: Portal1_DEV

View in EC2 console Export

Configuration comparison and selection

Select one of the recommended options to compare it against your current utilization metrics below.

Configuration	Instance type	On-Demand price	Price difference	Risk level	vCPUs	Memory	Storage	Network
Current (Portal2_P	m5.2xlarge	\$0.384 per hour			8	32GiB	EBS Only	Up to 10 Gbps
☒ Option 1	m5.xlarge	\$0.192 per hour	- \$0.192 per hour	Low	4	16GiB	EBS Only	Up to 10 Gbps
☐ Option 2	t3.xlarge	\$0.1664 per hour	- \$0.2176 per hour	Low	4	16GiB	EBS Only	Up to 10 Gbps
☐ Option 3	r5.large	\$0.126 per hour	- \$0.258 per hour	Medium	2	16GiB	EBS Only	Up to 10 Gbps



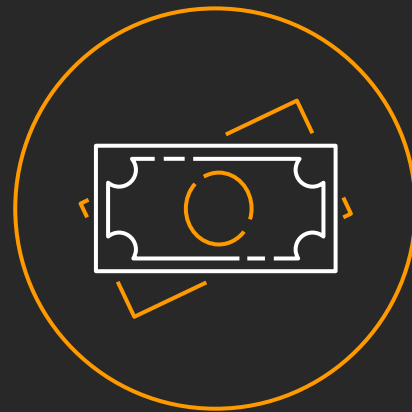
介绍

Savings Plans

通过灵活的定价模式简化采购，在 Amazon EC2/ECS 和 AWS Fargate 上节省成本。针对 AWS Lambda 即将推出！



自动接收更低的费率。易于与 AWS 成本管理器中的建议一起使用



显著节省
高达 72%



跨实例系列、大小、操作系统、租户或区域的灵活性

介绍

Amazon EKS for AWS Fargate

安全、可靠和大规模地运行 Kubernetes 容器，无需考虑底层服务器



简化了 AWS 上 Kubernetes
的部署、管理和扩展



默认情况下，每个 Pod 都
具有强大的安全隔离功能



专注于构建应用程序

新特性

加速 AWS 容器服务的势头

AWS Fargate Spot

Fargate 标准价格，闲置能力，
节省高达70%



Amazon ECS Capacity Providers

容器针对应用的计算



预览: Amazon ECS CLI 2.0

直接通过 CLI 在 ECS 上
部署和管理应用



ECS 群集自动 缩放

提高了可扩展性，降低了运
行容器的运营成本



介绍

EC2 Image Builder

更快、更轻松地构建、维护和共享安全操作系统映像。Linux 和 Windows 都具有组件和测试的自动化功能



对 AWS 和本地的简单
映像管理



开箱即用的安全性



提高 IT 工作效率
(无手动步骤)

新特性

AWS License Manager – 简化的 Windows & SQL Server BYOL

- 将符合条件的 Windows 和 SQL BYOL 许可证带到 AWS
- 利用现有许可投资来节省成本
- 自动化 EC2 专用主机的持续管理



Windows BYOL

Elasticity of EC2 for
Dedicated Hosts
with AWS License
Manager Integration
(New)

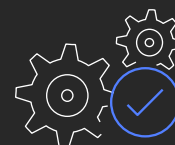
支持 Windows 服务器的 AWS 迁移计划

帮助客户升级旧版应用程序，在更新、受支持的 Windows Server 版本上运行，而无需任何代码更改

面向未来



与底层操作系统解耦



后续操作系统更新故障
风险低



支持所有操作系统版本

降低风险



改进了支持的新操作系
统的安全状态



隔离旧的运行时



遵从行业法规

经济高效



没有应用程序重构或重
新编码成本



无扩展支持成本



降低运营成本

存储

AWS Storage Day

有关整个存储产品组合的数据可用性、安全性、保护和成本优化的 15 项新功能

服务	公告
Amazon S3	复制时间控制
Amazon EBS	快速快照还原 (FSR)
Amazon FSx for Windows File Server	多可用区部署选项;数据重复数据;用户配额;编程共享管理;支持较小的文件系统;支持高可用性 SQL Server 部署;对传输中加密的细粒度控制
AWS Storage Gateway	VMware 的高可用性; 磁带和文件网关的性能提升
AWS DataSync	降价68%;任务计划;扩展到另外 5 个 AWS 区域
Amazon EFS	现已在所有商业 AWS 区域 (中国区域除外) 提供

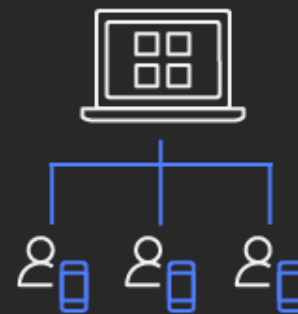
介绍

Amazon S3 Access Points

使用 Amazon S3 上的共享数据集简化应用程序的数据大规模访问管理。轻松为每个存储桶创建数百个接入点，每个接入点都有为每个应用程序自定义的唯一名称和权限



创建自定义访问非常简单
到同一存储桶上的共享数据集



添加接入点
随着应用程序集的增长



使用接入点轻松
限制对 VPC 的访问

介绍

EBS Direct APIs for Snapshots

一组简单的 API，提供对直接读取 EBS 快照数据的访问，使备份提供商能够以更低的成本实现 EBS 卷高达 **70% 的快速备份**

轻松跟踪 EBS 卷上的增量的变更的块，以实现：



备份时间缩短 70%



更精细的恢复点目标(RPO)



成本更低的备份

数据库 & 数据分析

介绍

Amazon Managed Apache Cassandra Service

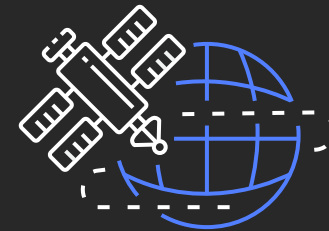
可扩展、高度可用且托管的 Apache Cassandra 数据库服务。运行与现在使用的应用程序代码和开发人员工具相同的应用程序代码和开发人员工具



与 Apache
Cassandra 兼容



无需管理服务
器



扩展的性能

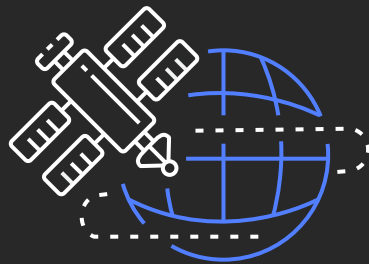


高度可用且安全

介绍

Amazon RDS Proxy

完全托管、高度可用的 Amazon RDS 数据库代理。池和共享连接，使应用程序更具可扩展性、对数据库故障的弹性以及安全性



具有数据库效率的可扩展应用程序



具有快速故障转移的弹性应用程序



通过数据保护保护应用程序

新特性

Amazon Aurora Machine Learning Integration

简单、优化且安全的 Aurora、SageMaker 和 Comprehend (预览)集成。使用 SQL 将基于机器学习的预测添加到数据库和应用程序，无需自定义集成、移动数据或机器学习体验



机器学习预测
的关系数据



与
SageMaker &
Comprehend 集成



熟悉的 SQL 语
言，无需机器
学习专业知识



低延迟、实时



安全与治理

介绍

UltraWarm for Amazon Elasticsearch Service

适用于 Amazon Elasticsearch Service 服务的低成本、可扩展的暖存储层。以现有存储层成本的 **1/10** 在单个群集中存储多达 10 PB 的数据



存储多达 10 PB
的日志数据



实现 90% 的成
本节约



运行交互式日志分析
和可视化

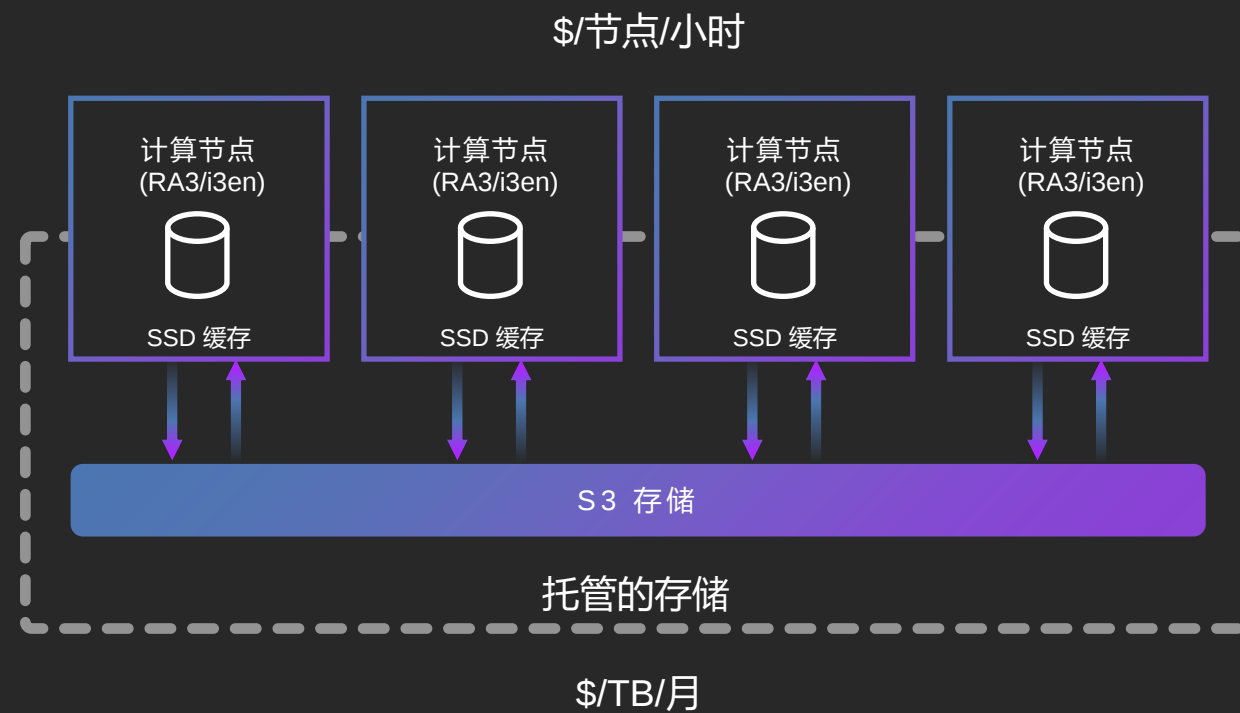


更高的性能和持久性

介绍

Amazon Redshift RA3 instances with Managed Storage

通过分别为计算和存储付费来优化数据仓库成本



提供现有云上数据仓库3倍的性能

同样定价的 DS2 实例(按需), 提供 2倍的性能和2倍的存储

自动扩展数据仓库存储容量

支持高达 8PB 的工作负载 (数据压缩)

介绍

AQUA (Advanced Query Accelerator) for Amazon Redshift

Redshift 的运行速度比其他任何云数据仓库快10倍，而不会增加成本



AQUA 将计算引入存储，因此数据不必来回移动

S3 上的高速缓存扩展到多个节点上并行处理数据

AWS 设计的处理器可加快数据压缩、加密和数据处理

100% 兼容当前版本的 Redshift

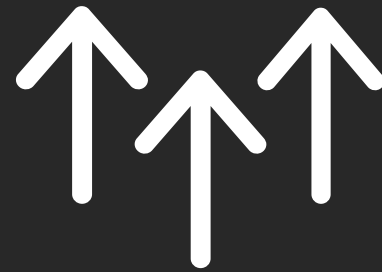
新特性

Amazon Redshift Federated Query

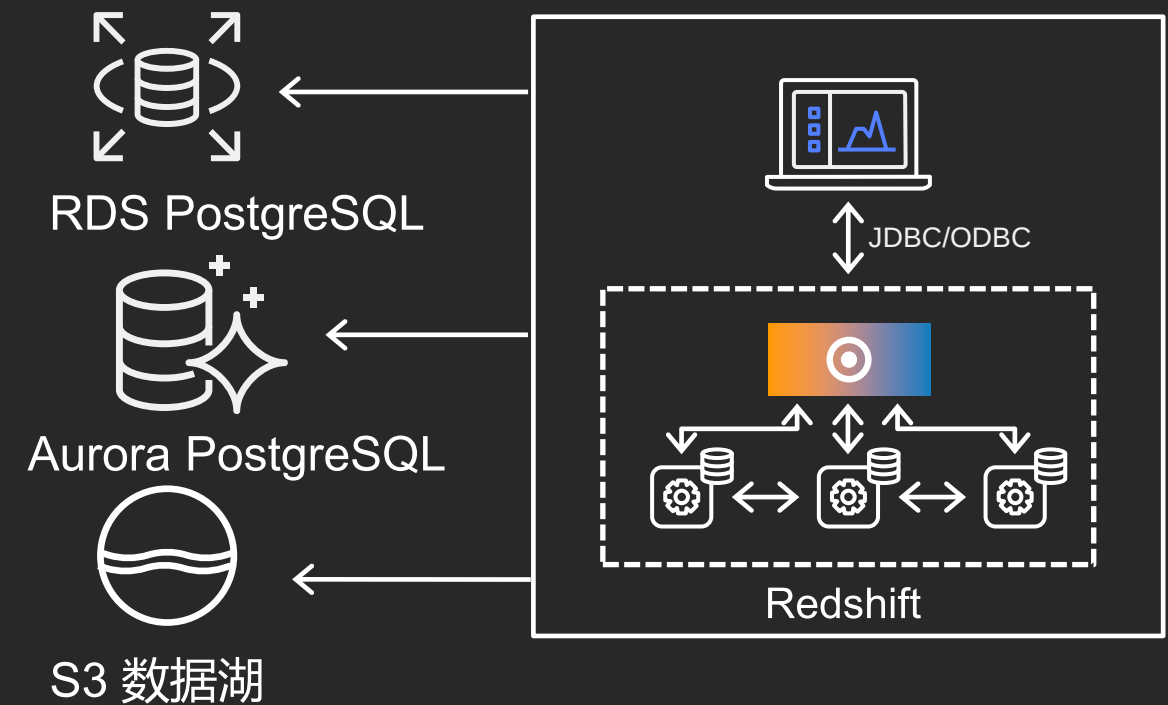
跨数据仓库、数据湖和操作型数据库进行数据分析



合并数据仓库和事务数据



从 Redshift 跨多个系统的查询



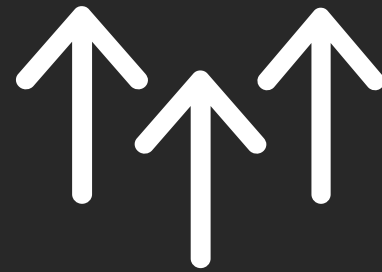
新特性

Amazon Athena Federated Query

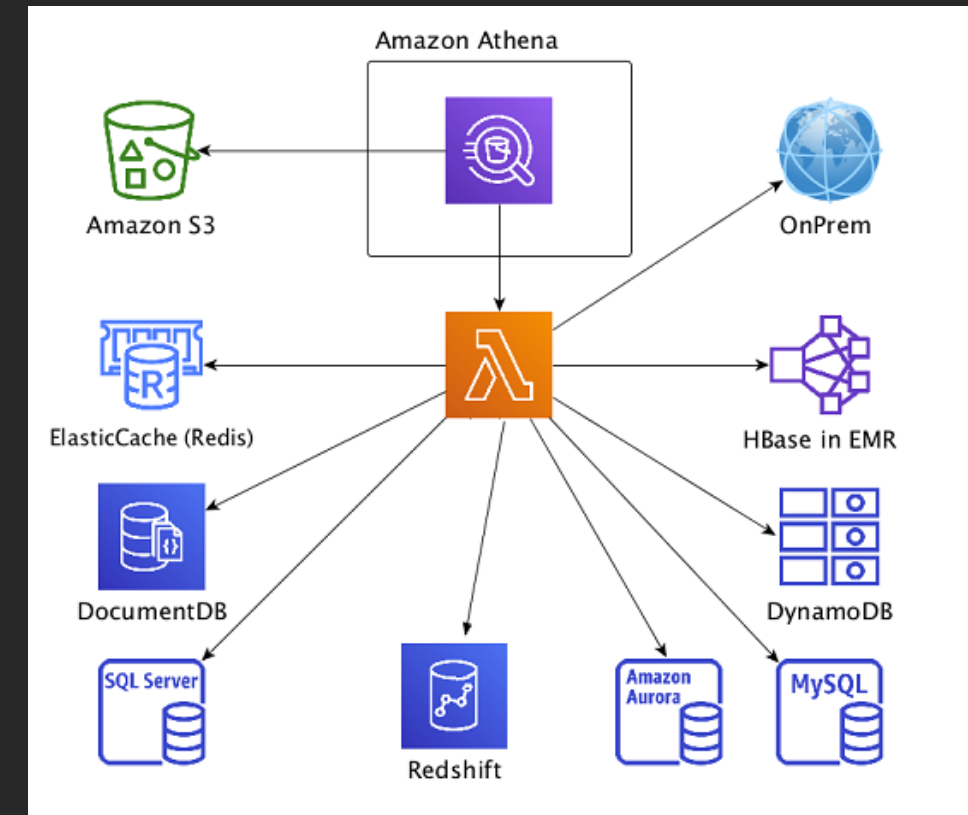
通过 AWS Lambda (SAR) 上运行的数据源连接器分析任何数据源的数据



将数据湖与任何其它
数据源合并



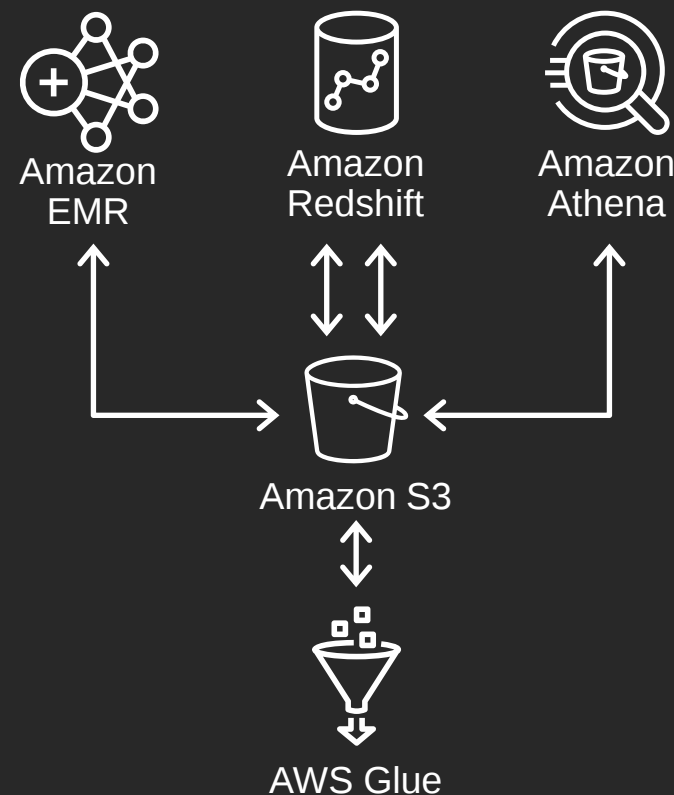
从 Athena 跨多个
系统的查询



新特性

Amazon Redshift Data Lake Export

没有其他数据仓库能够轻松地从所有数据中获得新的洞察力



使用开放的格式(如 Apache Parquet)将查询结果保存在 S3 数据湖中，以便进行优化分析



进一步分析来自其他分析服务的数据，如 Amazon EMR、Amazon Athena 和 Amazon SageMaker

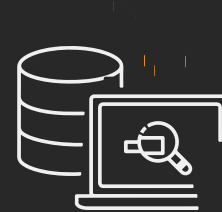
AWS Data Exchange

在云中轻松查找和订阅第三方数据

对于
订阅用户



在一个位置快速查找各种
数据

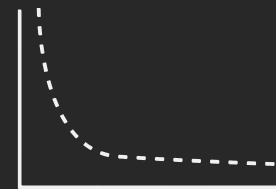


轻松分析数据



高效访问
第三方数据

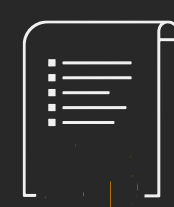
对于
数据供应商



覆盖数百万 AWS 客户



打包和发布数据产品的最简
单方法



内置安全性和合规性控制

安全 & 网络

介绍

AWS CloudTrail Insights

识别 AWS 账户中的异常（写入）活动



- 资源调配的意外峰值
- IAM 管理操作的突发性
- 定期维护活动的差距

- ✓ 节省筛选日志的时间
- ✓ 在问题影响到业务之前，提前做好准备

介绍

Amazon Detective

分析、调查和确定安全发现和可疑活动的根本原因。与 AWS Security Hub 集成



自动提取和组织数据到
一个图形模型



易于使用可视化，以更快和
有效的分析



随着新遥测数据可用而不断
更新

介绍

AWS IAM Access Analyzer

持续确保策略为资源提供预期的公共和跨账户访问，例如 Amazon S3 存储桶、AWS KMS 密钥以及 AWS Identity 以及 Access Management 角色



快速分析资源策略

分析公共或跨帐户访问的资源策略



持续监控和分析权限

分析新的或更新的资源策略，以帮助了解潜在的安全影响



提供最高级别的安全保证

使用自动推理（一种数学逻辑形式）来确定资源策略允许的所有可能访问路径

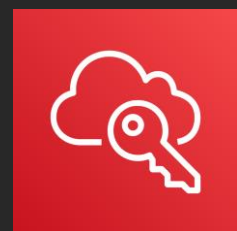
新特性

AWS Single Sign-On

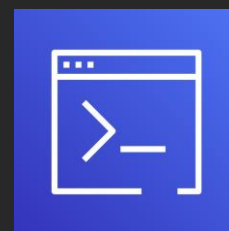
创建或使用现有身份（包括 Azure AD），并集中管理对多个 AWS 账户和业务应用程序的访问，以方便员工进行浏览器，命令行或移动式单点登录访问。



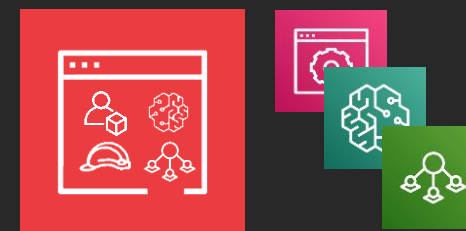
选择身份
认证源



集中管理访
问



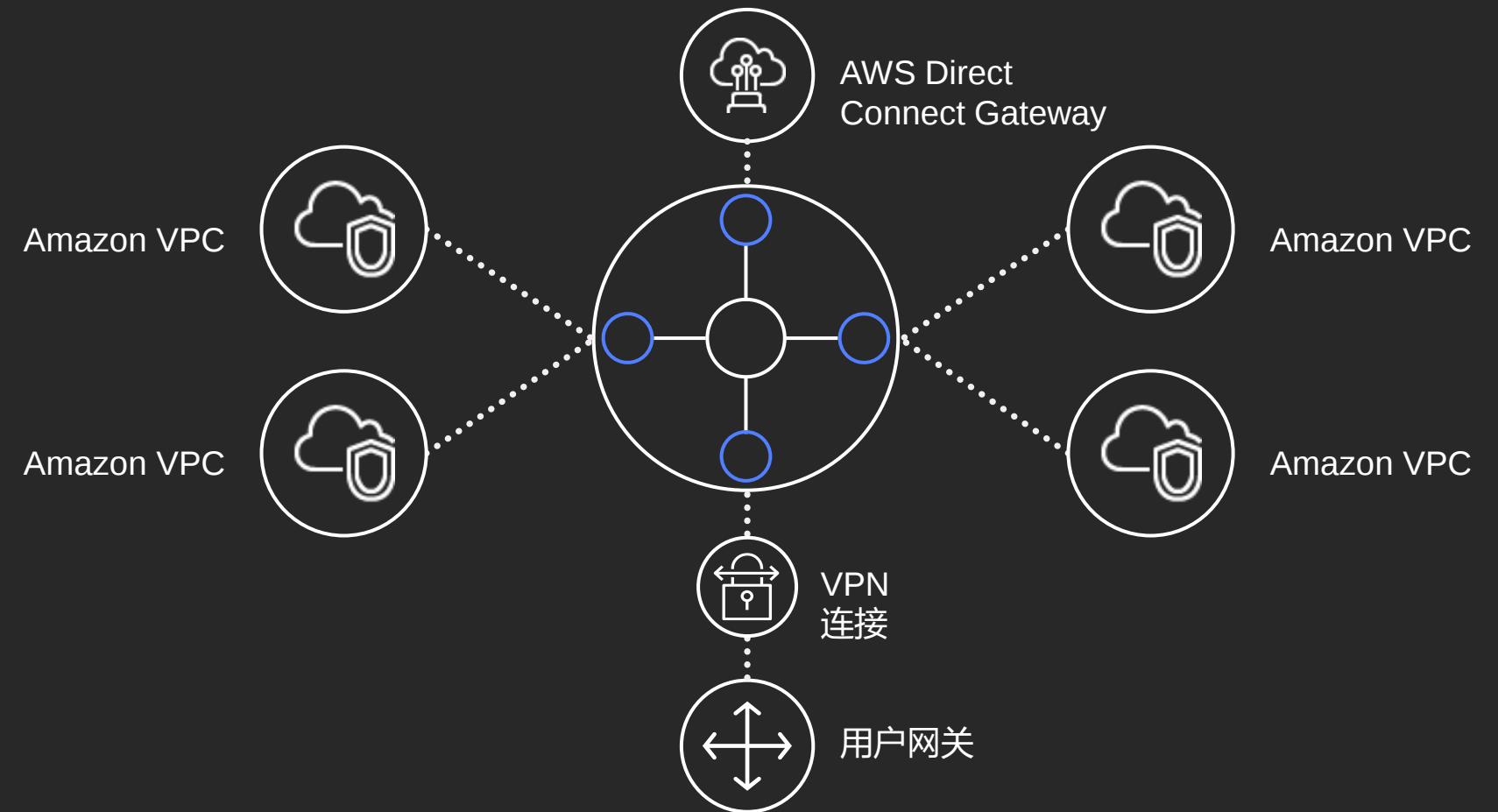
增加 CLI
安全性、生
产效率



浏览器和移动门户访
问帐户/角色/应用

AWS Transit Gateway

跨数千个 Amazon VPC、AWS 账户和本地网络扩展连接



新特性

AWS Transit Gateway Inter-Region Peering

通过跨多个 AWS 区域连接过境网关构建全球网络



在区域之间共享资源或复制数据以实现跨地理的冗余

区域间流量加密，无单点故障或带宽瓶颈

在一个 Hub（中心）和 Spoke（辐射）模型中扩展网络到多个 AWS 区域的数千个 VPC

新特性

AWS Accelerated Site-to-Site VPN

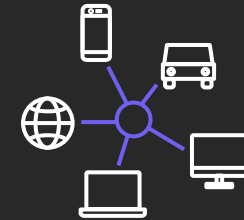
高可用性并提高了站点到站点 VPN 的性能



使用 AWS 全球网络安全地连接多个站点



提高 VPN 连接的性能



构建于 Global Accelerator 之上

介绍

AWS Transit Gateway Network Manager



全球网络的可见性和
健康状况



跨云和远程站点进行
管理

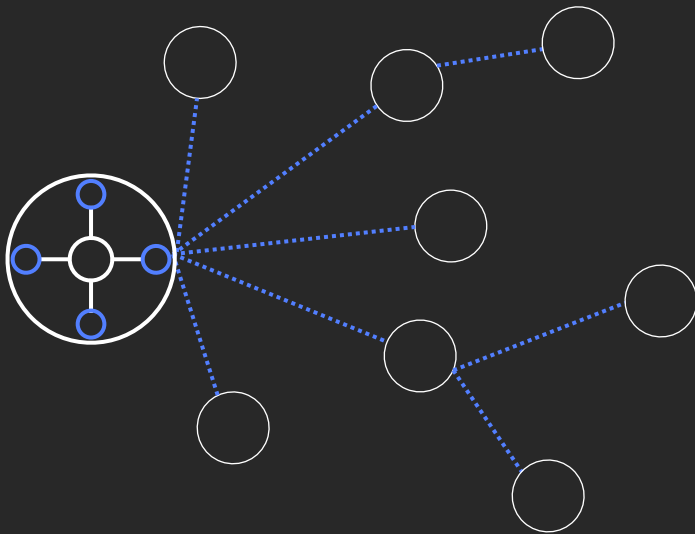


优化了与 AWS 全
球网络的远程站点
连接

新特性

Transit Gateway Multicast

在云中构建和部署多播(multicast)应用



根据需求向上或向下扩展多播应用

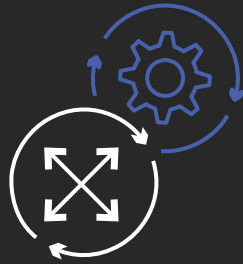
无需购买和维护自定义硬件以支持峰值负载

通过已验证的 API 精细控制谁可以产生和谁可以消耗多播流量

新特性

Amazon VPC Ingress Routing

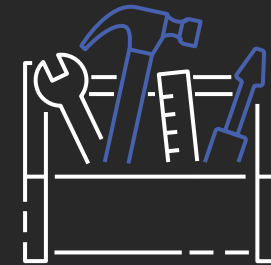
通过第三方或 AWS 服务路由入站和出站流量



通过一个设备传递所有
内联(inline)流量



内联流量检查可帮助流
量到达工作量之前对其
进行筛选和保护



通过第三方解决方案帮
助扩展功能

无服务器计算 (Serverless)

新特性

Provisioned Concurrency on AWS Lambda

- 保持功能的初始化和预热，确保启动时间保持在毫秒内
- 完全控制何时设置供应的并发性
- 无需更改代码即可在生产中的函数上预配并发性
- 可与 AWS Auto Scaling 相结合

介绍

HTTP APIs for Amazon API Gateway

与 REST API 相比，可实现高达 67% 的成本降低和 50% 的延迟降低。HTTP API 也比 REST API 更易于配置，使客户能够将更多时间集中在构建应用程序上



将应用成本降低高达 67%



将应用程序延迟减少多达
50%

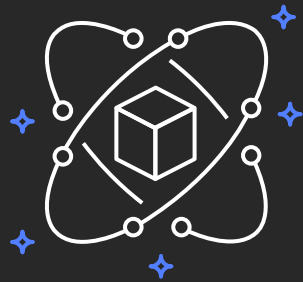


配置 HTTP API 比以前更快、
更快速

介绍

AWS Step Functions Express Workflows

以超过 100,000 个事件/秒的速率编排 AWS、数据库和消息传递服务，适用于高容量事件处理工作负载，如 IoT 数据引入、流数据处理和转换



更快：100K TPS



专为短时工作流程
而设计：<5分钟



大规模的成本效益

介绍

Amazon EventBridge Schema Registry

将事件结构（或模式）存储在共享的中央位置，以便更快，更轻松找到所需的事件。直接在 IDE 中生成代码绑定，以将事件表示为代码中的对象



模式发现会自动将模式添加到注册表（付费功能）



与 IDEs 集成：
VS Code、IntelliJ



生成 Java、Python
或 Typescript 的
代码绑定

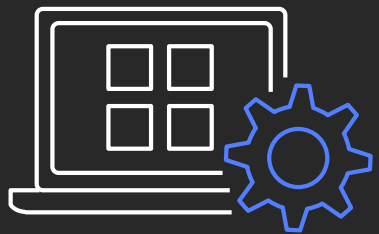


从15个 SaaS 合作伙伴，
自定义应用程序和
AWS 服务访问模式

介绍

Amplify for iOS & Android

开源库和工具链，使移动开发人员能够构建可扩展且安全的基于云的无服务器应用程序



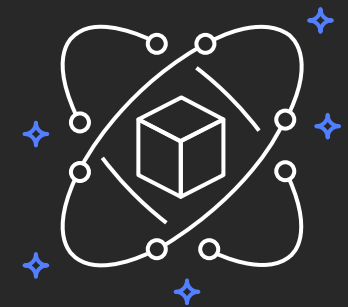
使用基于用例
的抽象加快开发速度



添加 Auth、分析、
AI/ML、API(GraphQL)、
DataStore 和存储



使用 GraphQL
从多个数据源获取
实时和离线数据



使用库和 SDK 访问
广泛的 AWS 服务

新特性

Amplify DataStore

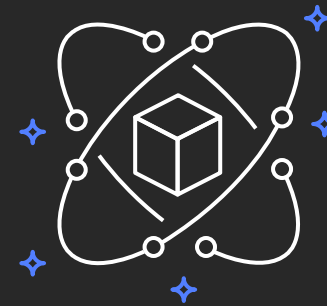
多平台 (iOS/Android/React Native/Web) 设备上持久存储引擎，可使用 GraphQL 在移动/ Web 应用和云之间自动同步数据



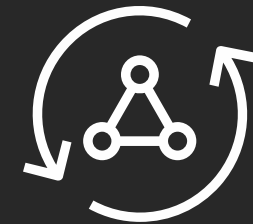
云中的自动版本控制，
冲突检测 and 解决



熟悉的和本地
优先编程模型



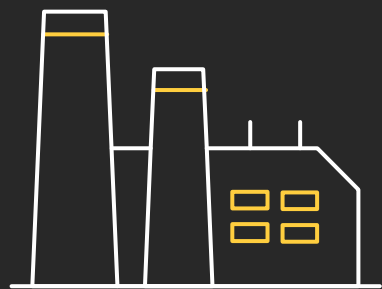
Data Delta Sync
(数据增量同步)
和自动合并



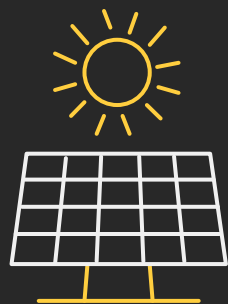
由 AppSync 和
GraphQL
提供支持

将 **AWS** 扩展到区域之外

客户在使用 AWS IoT 做什么



提高工业过程的性能和
生产效率



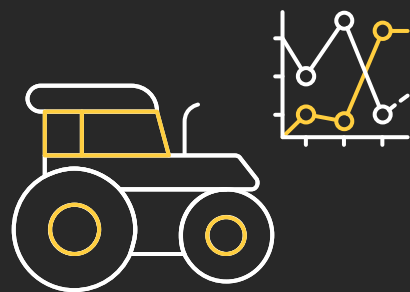
更高效地管理能源资源



远程监控患者健康与健康应用



借助互联和自动驾驶汽车改变
交通运输



以更高的效率种植更健康的
农作物



跟踪库存水平并管理仓
库运营



增强家庭，办公室和工厂车
间的安全性



在家庭，建筑物和城市构建
更智能的产品和用户体验



AWS IoT Day

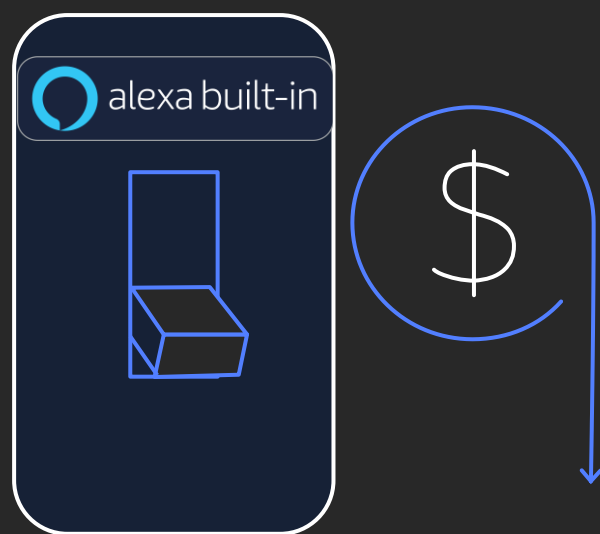
在整个 AWS IoT 上推出了 9 个新特性，其中包括在 IoT Day 上发布的 8 个新特性，可帮助客户轻松扩展其 IoT 部署

服务	新特性
AWS IoT Greengrass	支持容器的流管理器
AWS IoT Core	Fleet Provisioning 可配置的端点 可配置端点的自定义域 MQTT 连接的自定义授权者 Alexa Voice Service (AVS) 集成
AWS IoT Device Management	安全隧道
AWS IoT SiteWise	SiteWise 监视器

新特性

Alexa Voice Service (AVS) Integration for IoT Core

借助 Alexa 内置功能在诸如照明开关、恒温器和小型电器等新产品类别中快速，经济高效地投放市场



通过减少所需的计算和内存占用，将集成 Alexa Voice 的成本降低多达50%



在资源受限的设备（例如具有小于1MB 嵌入式 RAM 的 ARM'M' 类微控制器）上构建新的 Alexa内置产品类别



默认情况下，通过与 AVS Integration for IoT Core 配合使用的认证合作伙伴开发套件，可以加快上市时间

新特性

Container Support for AWS IoT Greengrass

将容器无缝部署到边缘设备



使用 AWS IoT Greengrass 将容器从云移动到边缘设备，而无需重写任何代码



使 Docker 和 AWS Lambda 组件能够在边缘无缝地协同运行



使用 AWS IoT Greengrass 机密管理器管理专用容器注册表的凭据

现已推出

AWS Outposts

完全托管的服务，可将 AWS 基础设施，AWS 服务，API 和工具扩展到几乎所有已连接的客户站点。跨本地和云环境的应用程序的真正一致的混合体验。低延迟或本地数据处理应用程序需求的理想选择



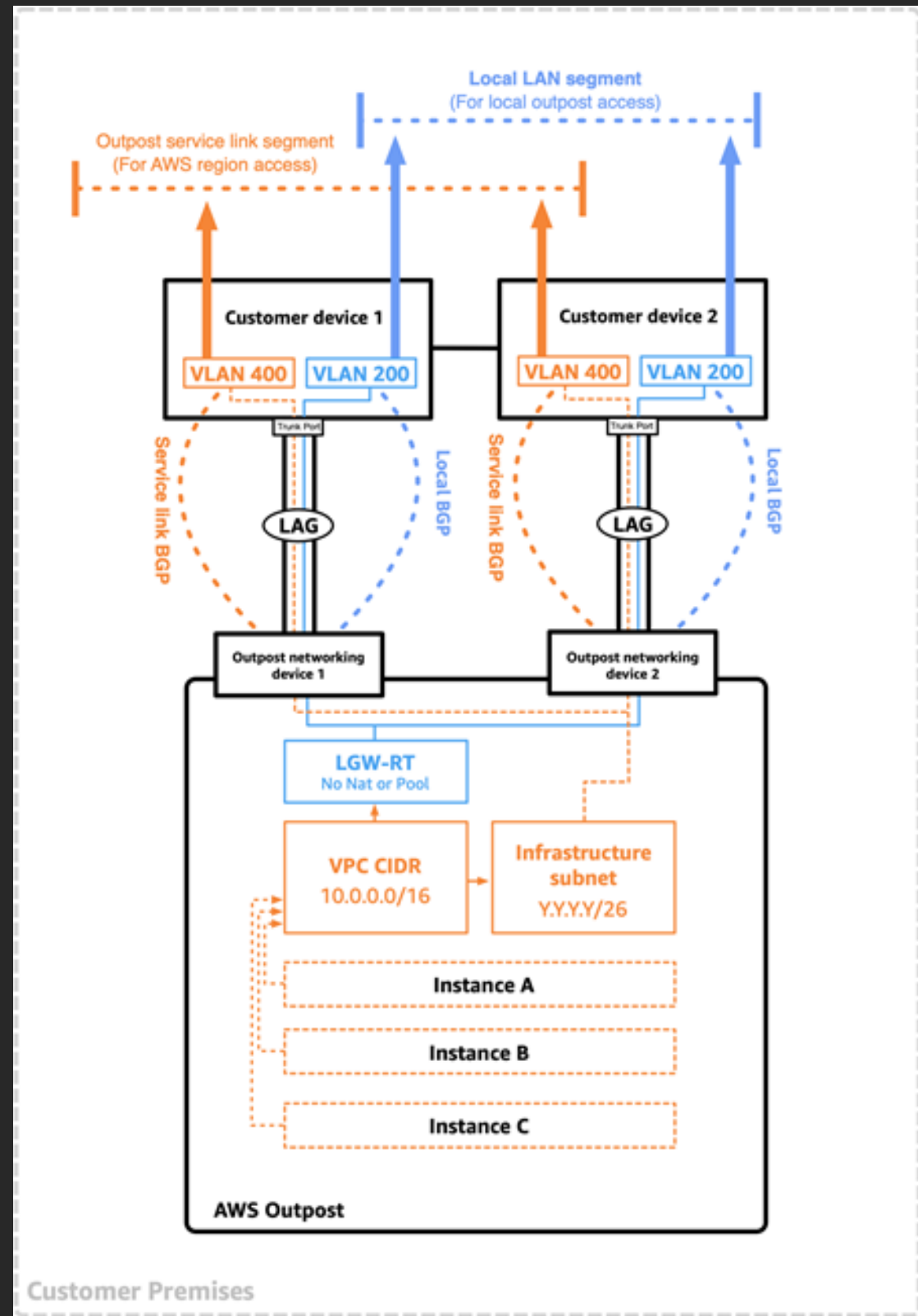
相同的 AWS 设计基础架构
与 AWS 区域数据中心一样
(基于 AWS Nitro 系统构建)
交付给客户设施

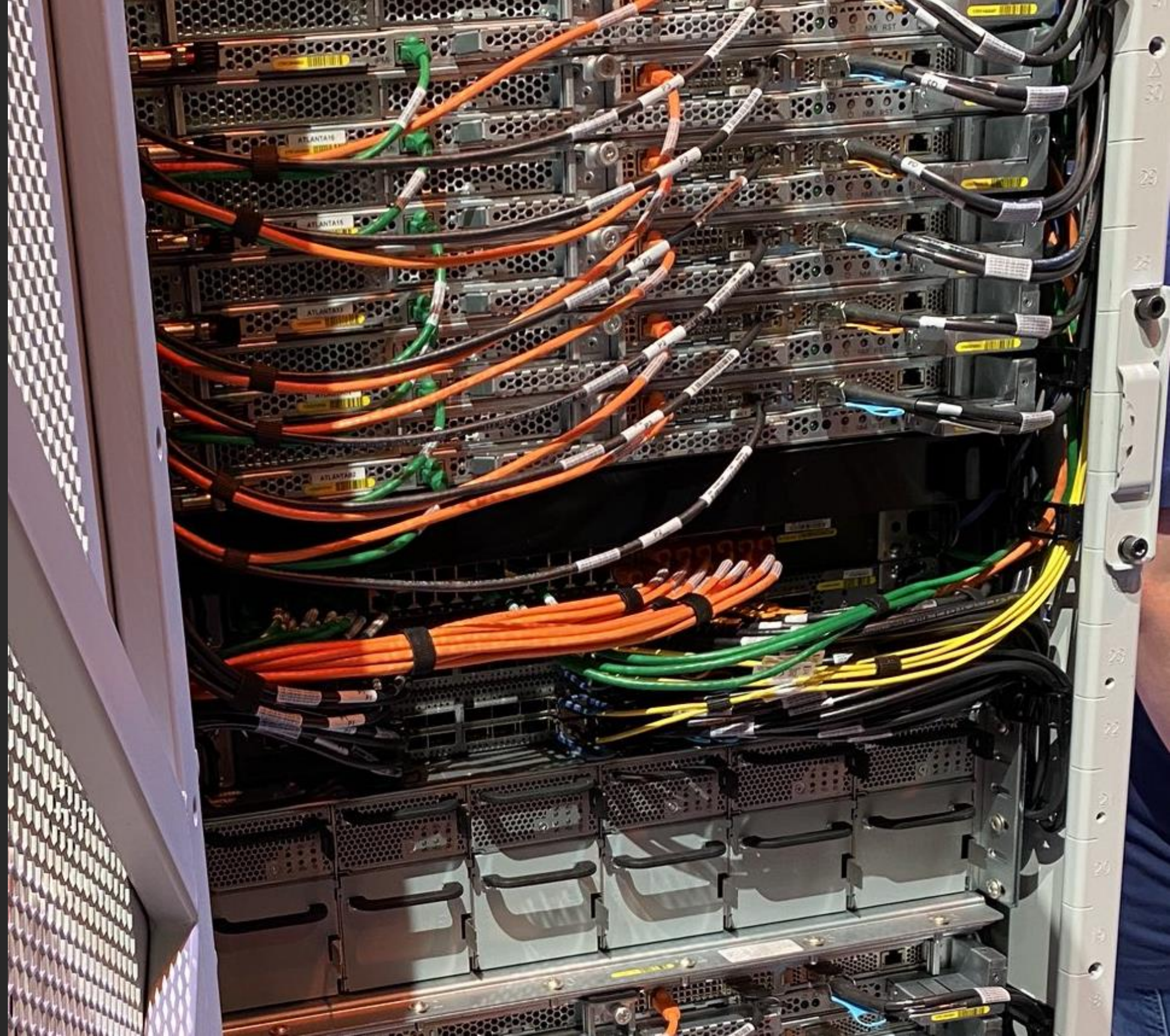


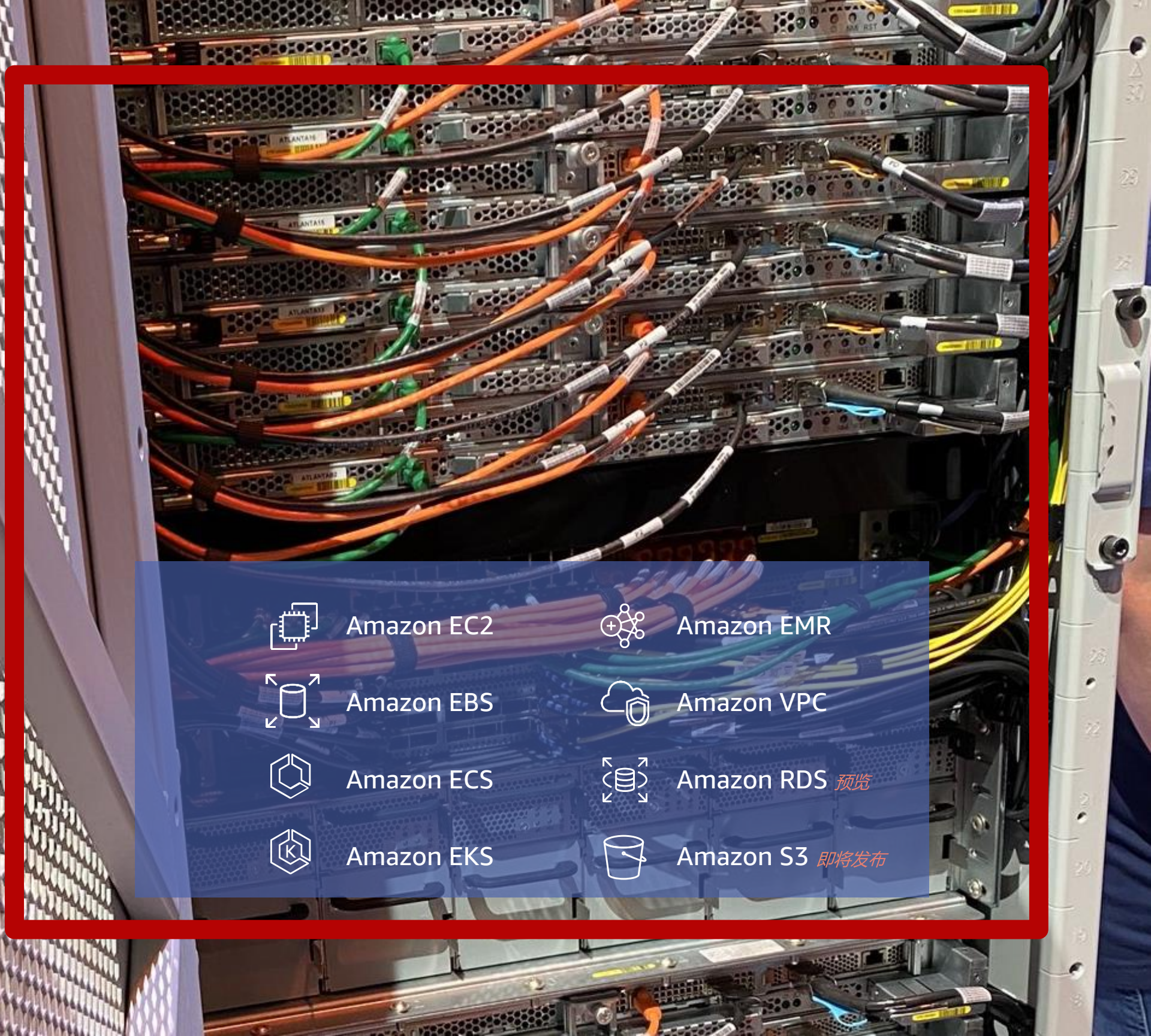
与 AWS 区域一样由 AWS 完全管
理，监视和操作



单一管理面板在云中提供
与 AWS 区域中相同的API和工具







Amazon EC2



Amazon EBS



Amazon ECS



Amazon EKS



Amazon EMR



Amazon VPC

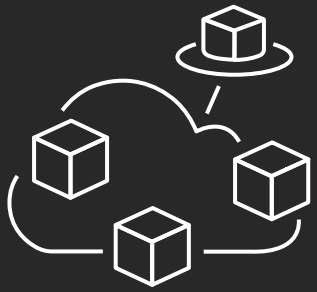


Amazon RDS 预览



Amazon S3 即将发布

AWS Outposts 上支持的服务 (除了 EC2 + EBS)



Amazon ECS,
EKS & App Mesh



Amazon EMR



Amazon RDS



Amazon S3
(2020年发布)

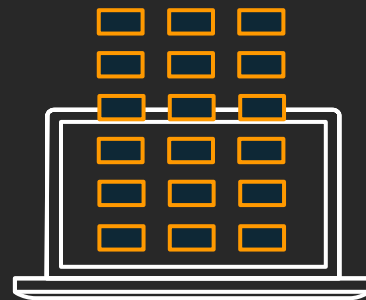
介绍

AWS Local Zones

将 AWS Cloud 扩展到更多位置并更接近最终用户，以支持超低延迟应用程序用例。使用熟悉的 AWS 服务和工具，并仅为所使用的资源付费



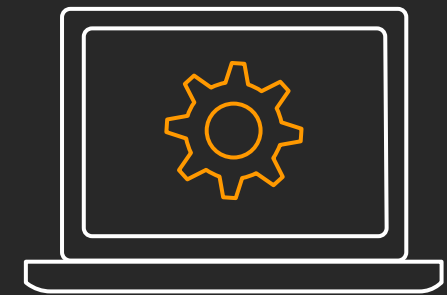
处于边缘位置的
AWS 基础设施



本地计算、存储、
数据库和其他服务



连接到 AWS 区域中
的服务



提供新的低延迟
应用

第一个要发布的 AWS Local Zones 将位于美国洛杉矶

介绍

AWS Wavelength

将 AWS 计算和存储嵌入电信提供商的 5G 网络中。使移动应用程序开发人员能够交付具有毫秒级延迟的应用程序，并只为使用的资源付费



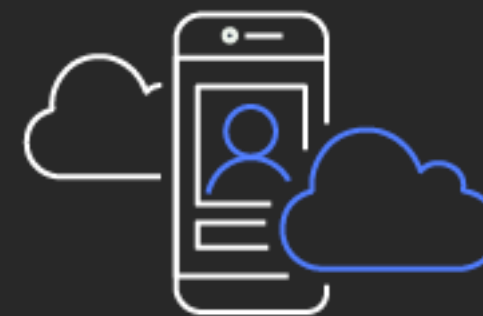
5G 网络中的
AWS 基础设施



本地计算，存储，
数据库和其他服务



连接到 AWS 区
域中的服务



提供新的移动应
用体验

介绍

AWS Wavelength

将 AWS 计算和存储嵌入电信运营商的 5G 网络中。使移动应用程序开发人员能够交付具有毫秒级延迟的应用程序，并只为所使用的资源付费



人工智能 & 机器学习

AWS 机器学习堆栈

人工智能服务

视觉


Amazon Rekognition

语音


Amazon Polly


Amazon Transcribe
+Medical^新


Amazon Comprehend
+Medical

文本


Amazon Translate


Amazon Textract

搜索^新


Amazon Kendra

会话机器人


Amazon Lex

个性化


Amazon Personalize

预测


Amazon Forecast

欺诈检测^新


Amazon Fraud Detector

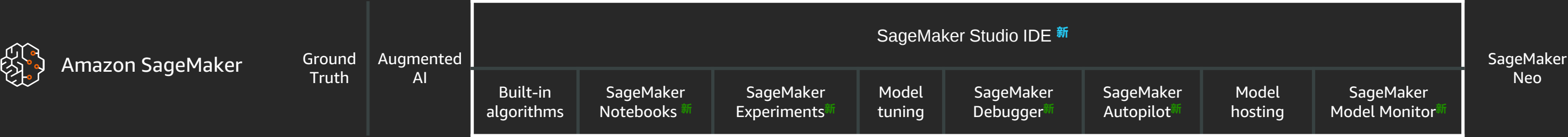
开发^新


Amazon CodeGuru

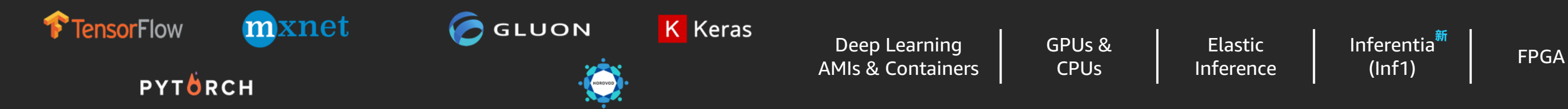
联系中心^新


Contact Lens
For Amazon Connect

机器学习服务



机器学习框架 + 基础架构



人工智能服务

re:Invent 前的发布亮点

<https://aws.amazon.com/about-aws/whats-new/machine-learning>

- Amazon **Comprehend**: 6 种新的语言
- Amazon **Translate**: 22 种新的语言
- Amazon **Transcribe**: 15 种新的语言, 用于文字转录
- Amazon **Lex**: SOC 和 HIPAA 合规性, 情感分析, Web& 移动与 Amazon Connect 的集成
- Amazon **Personalize**: 批量推荐
- Amazon **Forecast**: 使用任何分位数进行预测

随着AWS 区域的全面扩展!

Amazon Transcribe Medical

正式发布



准确

美国英语

Primary Care

听写转录

会话转录



易于使用

实时公共 API

自动标点符号

字级别的时间戳

字级别的置信度评分



实惠

即用即付模式

按转录使用情况收费

7.5 美分/分钟

60 分钟免费套餐

Transcribe Medical

USING THE [AMAZON TRANSCRIBE](#) WEBSOCKET API

Create an [AWS Account](#), attach the necessary [IAM policy](#), and e

Access ID:

.....

Secret Key:

.....

Session Token (if using MFA):

SESSION TOKEN

Press Start and speak into your mic

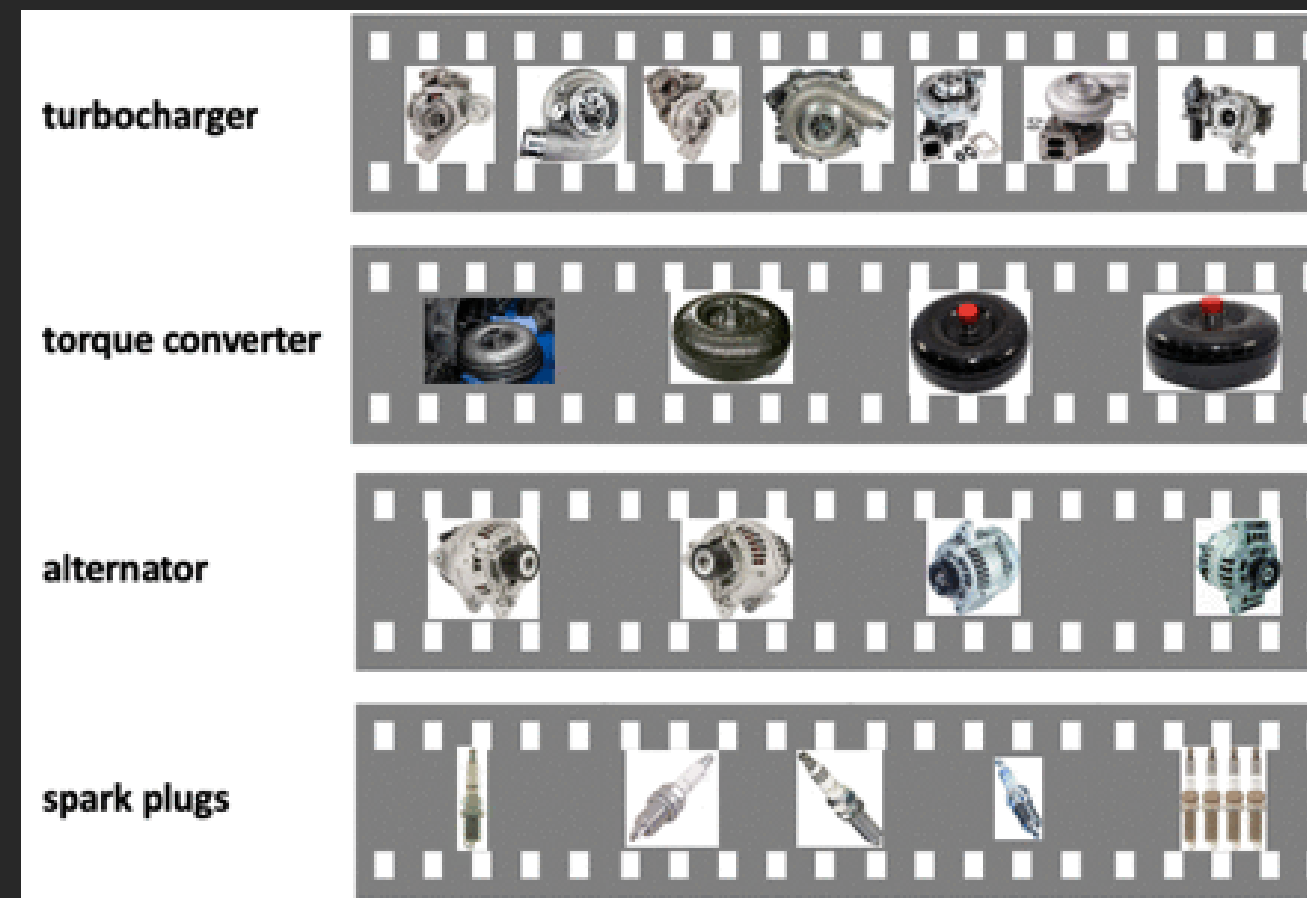
 Start  Stop Clear Transcript



```
The liver attachments to the adrenal kidney were d
UNREGISTERED
1 The liver attachments to the adrenal kidney were divided and the
liver was reflected superiorly. The vena cava was identified. The
main renal vein was identified. Coming superior to the main renal
vein, staying right on the vena cava, all small vessels were
clipped and then divided. Coming along the superior pole of the
kidney, the tumor was dissected free from top of the kidney with
clips and Bovie. The harmonic scalpel was utilized superiorly and
laterally. Posterior attachments were divided between clips and
once the whole adrenal was mobilized, the adrenal vein and one
large adrenal artery were noted, doubly clipped, and divided
Line 1, Column 371 Tab Size: 4 Plain Text
```

Amazon Rekognition

- 导入带有 Amazon SageMaker Ground Truth 标签的图像.....
 - 或根据文件夹结构自动标记图像
- 在完全托管的基础架构上训练模型
 - 拆分数数据集以进行训练和验证
 - 在训练结束时查看精度，召回率和F1分数
- 选择模型
- 使用常用的 Rekognition API



客户需求



+

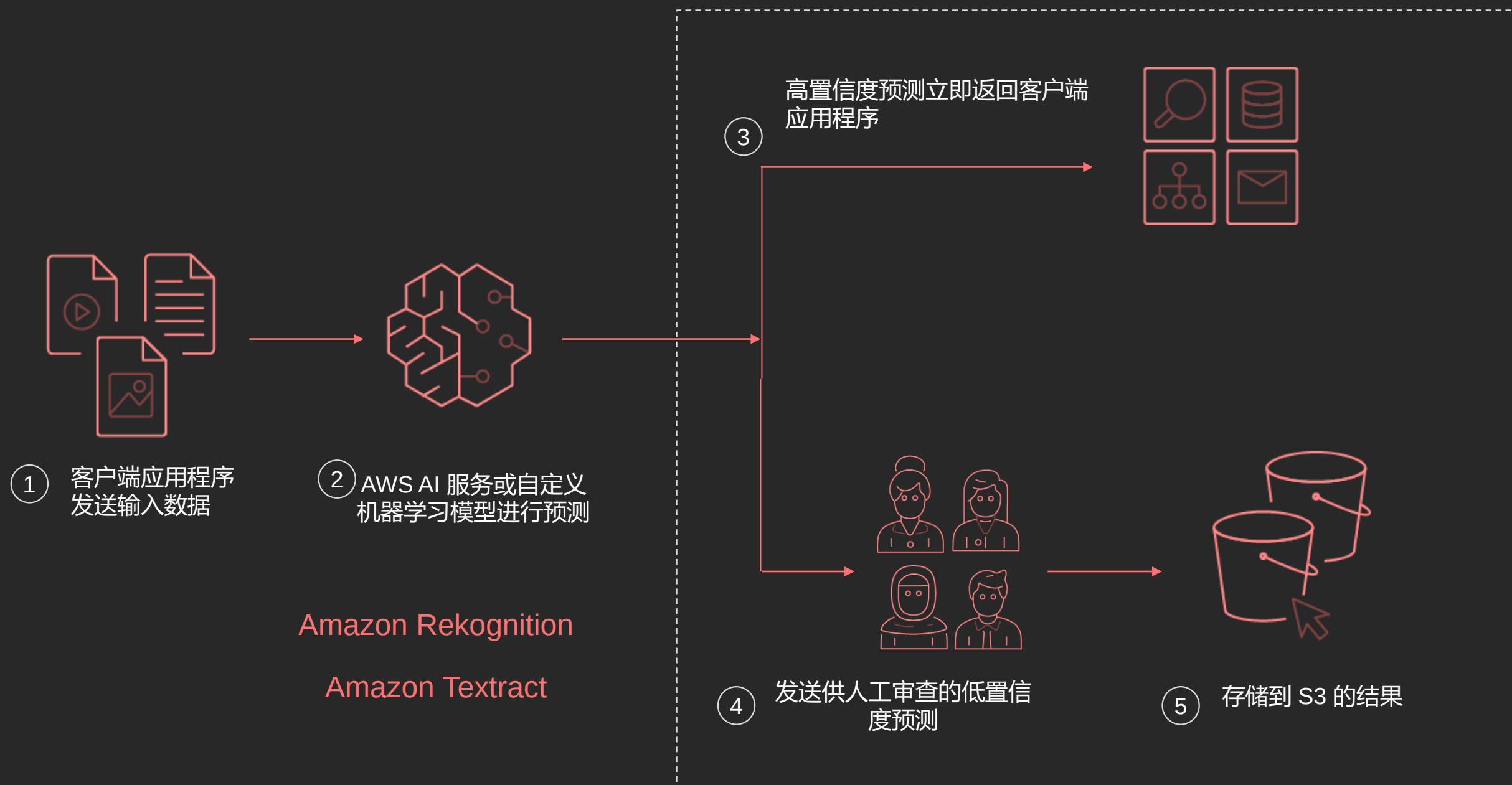


机器学习与人类协同工作

Amazon Augmented AI (A2I)

A2I 使您可以轻松地在机器学习工作流程中实施人工
审查，以提高复杂决策的准确性、速度和规模

Amazon Augmented AI 如何运作



人工审核工作



Amazon Mechanical Turk

由 Amazon Mechanical Turk 提供支持的全球超过500,000个独立承包商的按需全天候24x7劳动力



私有

自己派遣的一组员工，包括自己的员工或承包商，用于处理需要保留在组织内的数据



供应商

可通过 AWS Marketplace 获得的专门提供数据标签服务的第三方供应商的精选列表

Instructions



[View full instructions](#)

[View tool guide](#)

In this task, you will need to select the category that best describes the image.

Soccer



Swimming



Which sport is played in the image?



Select an option

Soccer	1
Swimming	2
Football	3
Basketball	4



Zoom



Fit image



Move

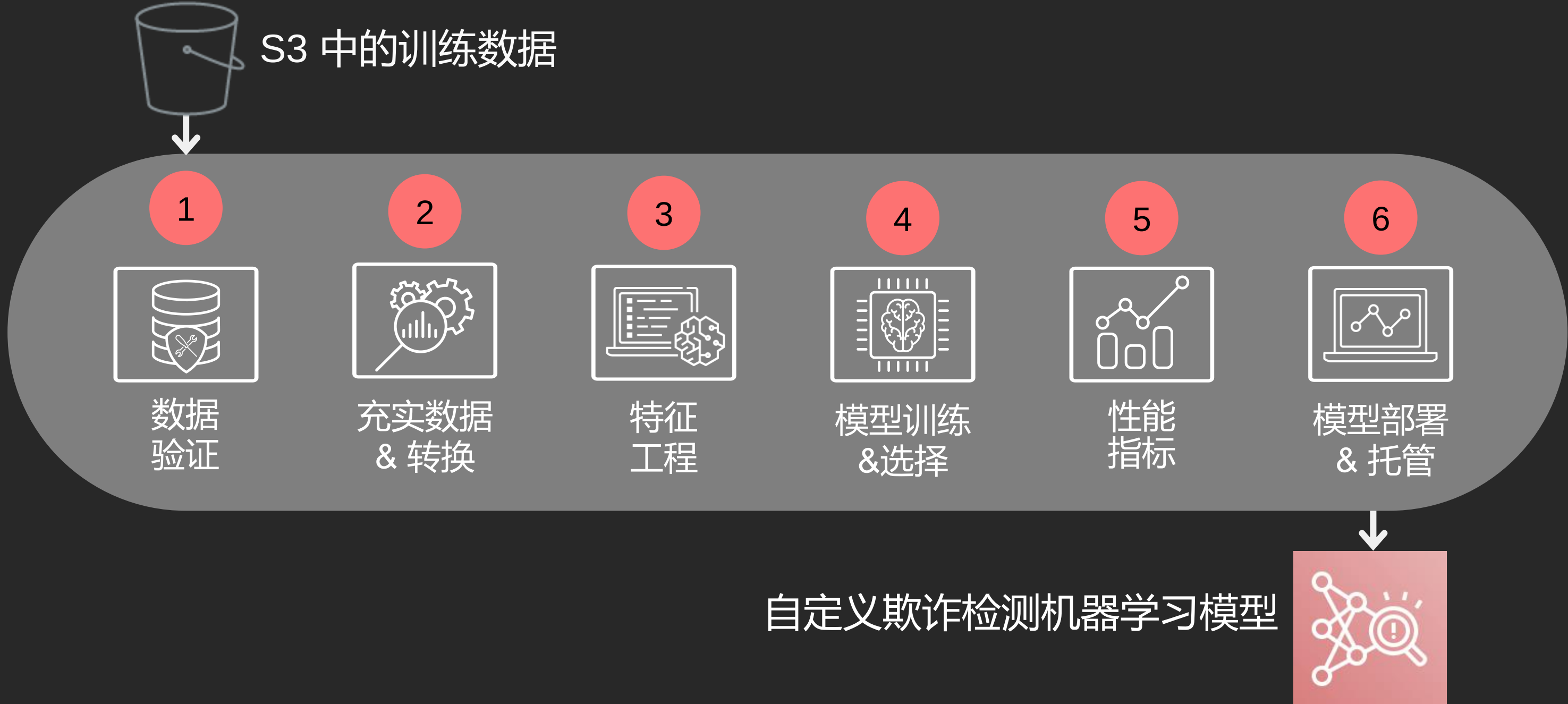
Submit

Amazon Fraud Detector

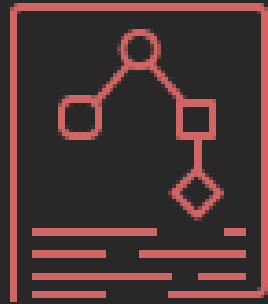
欺诈检测服务，可让企业轻松使用机器学习来实时，大规模地检测在线欺诈



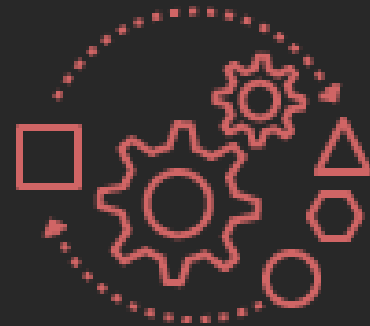
Amazon Fraud Detector – 自动化模型构建



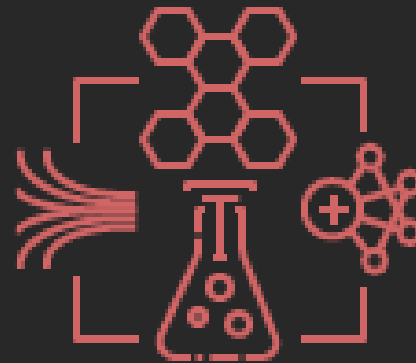
Amazon Fraud Detector – 主要特点



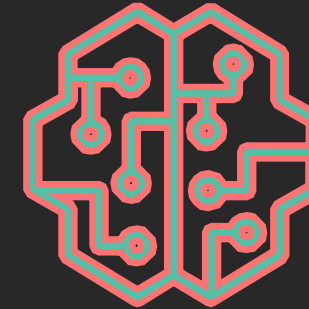
预构建的欺诈检测模型模板



自动创建自定义欺诈检测模型



模型借鉴了过去Amazon的经验



Amazon SageMaker 集成



一个界面可回顾过去的评估和检测逻辑

Filter

Past 24 hours

Property

Search property

Operator

Select operator

Value

Input value

Example: registration_detector

Add filter

Filter results

Results (21)

Find versions

< 1 >

Event ID	Detector name	Version	Outcomes	Evaluation timestamp
2959be0c-278f-4f3a-ae2a-11791b1d2e5b	new_account_risk	2.0	verify	2019-12-02T17:50:04.525Z
2dc6ed25-f29f-4294-83cd-d3d3d32a2c07	guest_checkout	4.0	fraud	2019-12-02T17:07:51.649Z
3b57c61a-8ee2-4754-815b-9d27be04c9ed	try_before_you_buy	3.0	NO_MATCH	2019-12-02T00:33:11.756Z
54424d1a-fc82-4abd-8968-6739bcaef303	try_before_you_buy	3.0	fraud	2019-12-02T00:33:40.894Z
56951daf-7e3b-447e-846c-08c3059b3e96	guest_checkout	4.0	fraud	2019-12-02T00:35:06.902Z
603a6301-46de-4f2a-bcb7-c5b2367afa34	new_account_risk	2.0	NO_MATCH	2019-12-02T17:13:31.249Z
68080587-b6c9-4103-8fc0-4f16ef76722a	new_account_risk	2.0	NO_MATCH	2019-12-02T17:49:55.868Z
82a8f124-af2c-4d59-a0a6-563c0df67666	guest_checkout	4.0	NO_MATCH	2019-12-02T17:07:27.098Z
88ebb30b-64e7-4a69-b40e-6991360fbd78	try_before_you_buy	3.0	fraud	2019-12-02T03:37:30.611Z
8c06bb88-c340-4480-8aa0-9218ddee534b	new_account_risk	2.0	NO_MATCH	2019-12-01T23:55:03.247Z
adc175a2-1c30-43c0-8986-fd4627d94c5e	guest_checkout	4.0	approve	2019-12-02T17:09:01.490Z
b18fa843-3010-4e11-9bf0-fd0713094ee4	try_before_you_buy	3.0	verify	2019-12-02T17:10:02.223Z
b3444b95-749a-45d5-a3ce-c6c681ad67b9	guest_checkout	4.0	NO_MATCH	2019-12-02T17:10:19.820Z
ba704eb0-e1a5-4d42-b956-3cb393d970a4	guest_checkout	4.0	NO_MATCH	2019-12-02T03:44:30.677Z
c6a70869-a4db-481b-99f0-a1174b876daf	new_account_risk	2.0	verify	2019-12-02T17:09:16.928Z
e93dd3ec-d659-4ff2-b34d-8d1d14197e32	new_account_risk	2.0	verify	2019-12-02T19:09:51.887Z

Add rules

Use existing rule

Define a rule

Rules are made of conditions and actions. If the con

Name

suspicious_customer

Rule names must be 1-64 lowercase characters (0-9

Description

Customers whose phone number and IP address are associated with the same account number.

Expression

Using Amazon Fraud Detector's simplified expression scores. To reference these variables, type "\$" to start for help.

```
1 $my_model_score > 500 and
2 $phone_number_country != $ip_add
```

Rule expressions must evaluate to true or false. When

► **Expression quick reference guide**

Outcomes

What should happen when the condition above is m

Choose one or more outcomes...

review ✕

[Add another rule](#)

Contact Lens For Amazon Connect

轻松使用机器学习的强大功能，提高客户体验质量，而无需任何技术专业知识

内置自动
呼叫转录



直接在 Amazon
Connect 中查看整个通
话记录

增强的
联系人搜索



根据语音和客户
情绪过滤感兴趣的来电

自动的联系人
分类



确定呼叫类型，例
如脚本合规性，
竞争性提及
和取消。

主题检测



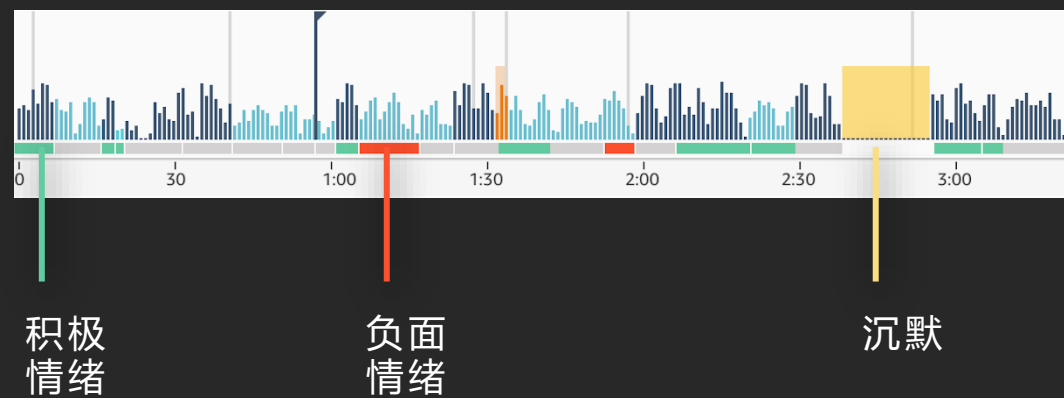
呈现重复
基于客户
反馈的问题

实时情绪
仪表板
和告警



客户现场通话
体验不佳的
快速识别

Contact Lens for Amazon Connect: 针对 Amazon Connect 的联系分析中心



只需单击"连接"即可激活

自动转录和分析客户电话，
并转录您以前录制的呼叫

通过扬声器检测获取全文转录 |

比现有解决方案精确 2 倍

分析情绪、长时间的沉默，以及座席和客户相互交谈的时间

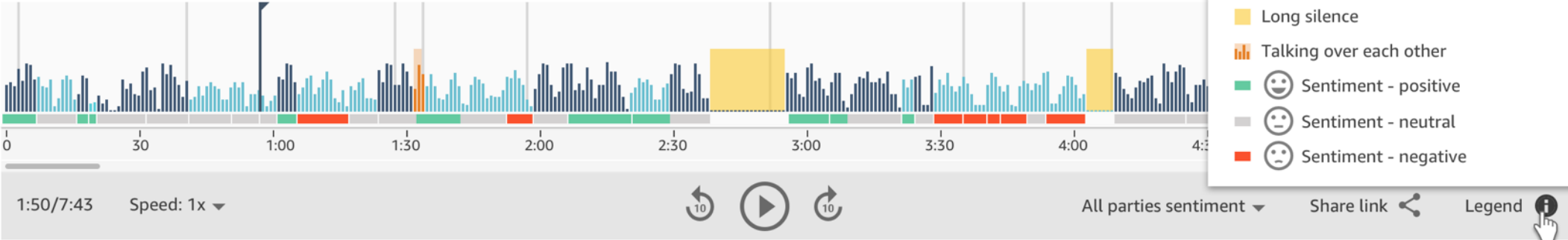
搜索所有转录关键字、短语和分析条件

随着呼叫的进行，近乎实时地查看成绩单

Categories

- Improper greeting
- Cancel account request
- Account verification
- Opology
- Upsell
- Competitor mention
- Difficulty hearing

Recording and transcript



Transcript

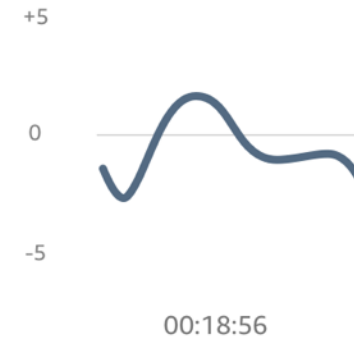
-  **Agent** 00:01
Hello. Thanks for calling [busniess name]. My name is Han Solo. Let's start with your date of birth, and please verify your address for me as well.
-  **Customer** 00:08
Hi. My birthday is December 22, 1981 and my address is 221A Baker Street.
I'll be moving soon, so I need to change that later.
-  **Agent** 00:18
Well, happy birthday.
-  **Customer** 00:20
Thanks
-  **Agent** 00:22
Hmm
I have a hard time finding your account. Do you happen to have your account ID?
Or tell me the last 4 digits of your social security number.
-  **Customer** 00:40
My account ID is 12345678
And the last four digist of my social is 1234
Actually that's the wrong number. Let me see.
It's 4567
-  **Agent** 01:03
Found ya. How may I help you today?



Contact Trace Record

Contact summary

Contact ID 1po9ie0-7fdc-2588-9erc-cd2iuy510987q
Contact Start and end time Dec 22, 18, 3:01:00 - 3:19:56 pm
Contact duration 00:18:56
Customer number 123-456-7890
Agent soloh
Queue customer account
Channel Voice



Customer sentiment trend



Customer sentiment



Totla talk time 00:18:56

Contact details

Contact analysis

Categories

Improper greeting

Cancel account request

Account verification

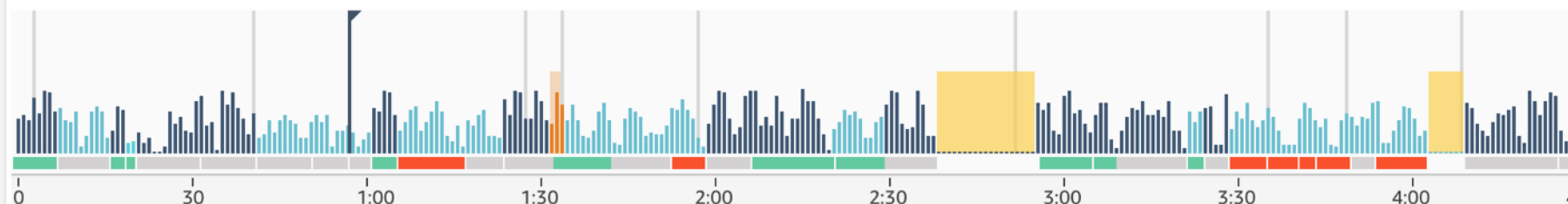
Opology

Upsell

Competitor mention

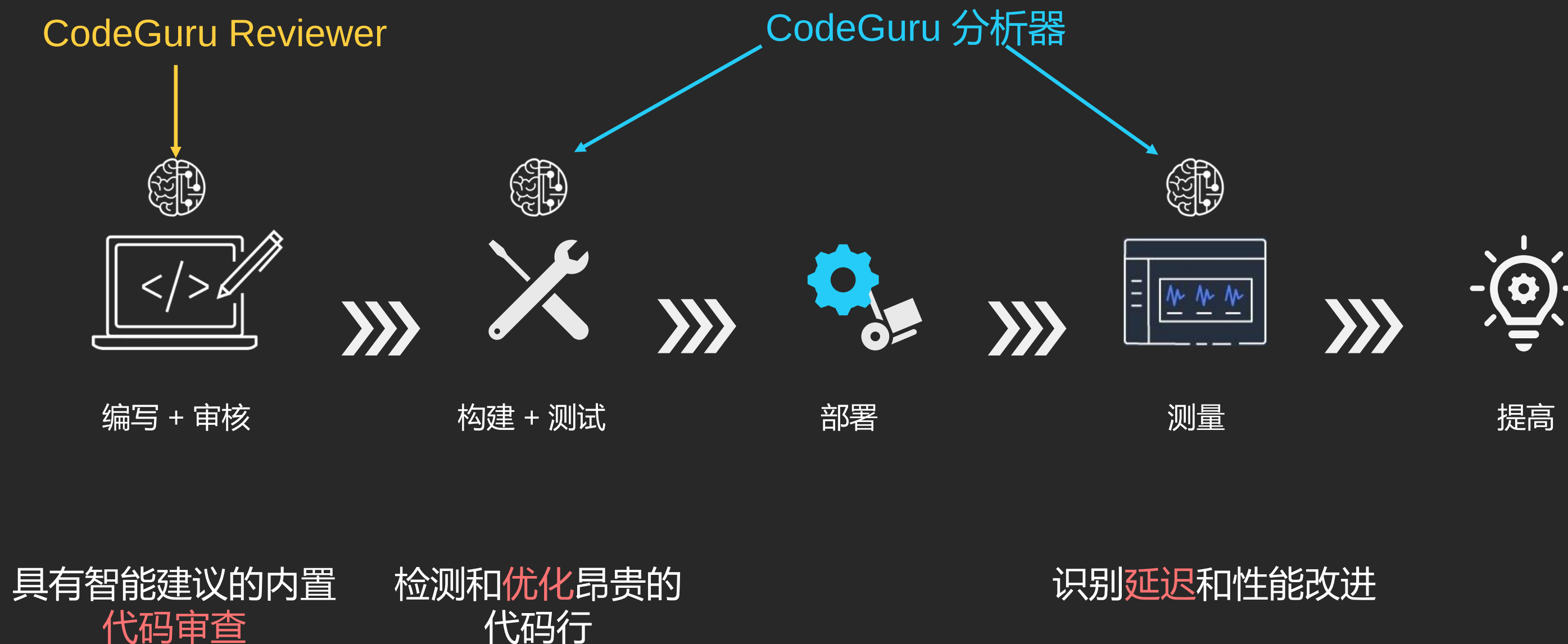
Difficulty hearing

Recording and transcript



- Agent
- Customer
- Long silence
- Talking over each other
- Sentiment - positive
- Sentiment - neutral
- Sentiment - negative

AWS CodeGuru



CodeGuru Reviewer: 如何工作

用户提供源代码作为输入

提取语义特征/模式

机器学习算法识别类似的
代码以进行比较

客户将建议视为
“Pull Request” 反馈

Java
AWS CodeCommit
Github



输入:
源码

特征提取

机器学习

输出:
建议

CodeGuru 示例 – Looping vs Waiting

代码

```
do {
    DescribeTableResult describe = ddbClient.describeTable(new DescribeTableRequest().withTableName(tableName));
    String status = describe.getTable().getTableStatus();
    if (TableStatus.ACTIVE.toString().equals(status)) {
        return describe.getTable();
    }

    if (TableStatus.DELETING.toString().equals(status)) {
        throw new ResourceInUseException("Table is " + status + ", and waiting for it to become ACTIVE is not useful.");
    }
    Thread.sleep(10 * 1000);
    elapsedMs = System.currentTimeMillis() - startTimeMs;
} while (elapsedMs / 1000.0 < waitTimeSeconds);
throw new ResourceInUseException("Table did not become ACTIVE after ");
```

建议

This code appears to be waiting for a resource before it runs. You could use the waiters feature to help improve efficiency. Consider using `TableExists`, `TableNotExists`. For more information, see <https://aws.amazon.com/blogs/developer/waiters-in-the-aws-sdk-for-java/>

开发者反馈

We should use waiters instead - will help remove a lot of this code.

CodeGuru Profiler – 示例

✔ Repeated regex compilation during String manipulation

Recommendations report

Definition

Version

11th Jan 2019 latest
v3 (Profiler) ▼

Description

The `java.lang.String` utility methods which accept regular expressions as `String` values are inefficient when used in hot paths. They have to compile the regular expression for every use, which is expensive.

Resolution steps

Pre-compile the regular expression into a `Pattern`, and replace any of the following uses as shown:

Before (inefficient):

```
1. input.matches("foo.*")
2. input.replaceAll("foo.*", "replacement")
3. input.replaceFirst("foo.*")
4. input.split("foo.*")
5. input.replace("not-a-regex", "replacement")
```

After (efficient):

```
static Pattern MY_REGEX = Pattern.compile("foo.*")
static Pattern MY_NOT_REGEX = Pattern.compile("not-a-regex", Pattern.CASE_INSENSITIVE);

1. MY_REGEX.matcher(input).matches()
2. MY_REGEX.matcher(input).replaceAll("replacement")
3. MY_REGEX.matcher(input).replaceFirst("replacement")
4. MY_REGEX.split(input)
5. MY_NOT_REGEX.replaceAll(Matcher.quoteReplacement("replacement"))
```

Note for number 5, `quoteReplacement()` is only required if the replacement string contains backslashes. It is required when used by `Matcher.replaceAll()` and `String.replace()`.

Example fix: <https://code.amazon.com/reviews/CR-3726256>.

The code fix

```
- private static final String INVALID_METRIC_CHARACTERS = "[^A-Za-z0-9\\\\.:@_\\-\\/]";
```

```
+ private static final Pattern INVALID_METRIC_CHARACTERS = Pattern.compile("[^A-Za-z0-9\\\\.:@_\\-\\/]");
```

```
- return metricName.replaceAll(INVALID_METRIC_CHARACTERS, "");
```

```
+ return INVALID_METRIC_CHARACTERS.matcher(metricName).replaceAll("");
```

https://github.com/

Dev2 #2

Changes from all commits

File filter...

Jump to...

0 / 1 files viewed

176

scan.java

134

+

135

+

136

+

if (null == lastKnownStatus || lastKnownStatus.getLeft() < timeStamp) {

137

+

lastKnownStatus = new MutablePair<Long, String>(timeStamp, status);

138

+

latestStatusForTrackingNumber.put(trackingNumber, lastKnownStatus);

1 hour ago

Author

Recommendation generated by Amazon CodeGuru-Reviewer

Problem

You are using a `ConcurrentHashMap`, but your usage of `get()` and `put()` may not be thread-safe at lines: 111, 114, 136, and 138.

Two threads can perform this **same check at the same time** and one thread can **overwrite** the value written by the other thread.

Fix

Consider replacing `put()` with `putIfAbsent()` to help prevent accidental overwriting.

`putIfAbsent()` puts the value only if the `ConcurrentHashMap` does not contain the key---and therefore avoids overwriting the value written there by the other thread's `putIfAbsent()`.

More Info

`putIfAbsent()` returns `null` if the value did not exist and returns the value in the map if one already exists.

Example Code

Hadoop at GitHub

Reply...

提高 CPU 利用率

降低成本

aws

Services

Resource Groups

Use of Inefficient Cryptographic Implementations

Why is CodeGuru recommending this change?

Your profile spent **3.84%** of the **RUNNABLE** and **NATIVE** time on frames related to this issue, but we would expect less than **1%** of your profile.

The top matched frames were `DigestBase.engineUpdate`, `ECDHKeyAgreement.deriveKey` and 3 others.

Description

Both the default cryptographic libraries shipped with Java as well as those from [BouncyCastle](#) include a number of inefficient implementations.

Suggested resolution steps

Evaluate switching to the [Amazon Corretto Crypto Provider ACCP](#).

Your application does not need to be directly calling the cryptographic code for the fix to work. Simply installing the new crypto provider should be sufficient.

Where was this found?

Matched function name	Portion of profile
<code>sun.security.ec.ECDHKeyAgreement.deriveKey</code>	0.26 %
Matching stacks	in 3 stacks
<code>0.15 %</code> <code>0.05 %</code> <code>0.05 %</code>	
<code>sun.security.ec.ECKeyPairGenerator.generateKeyPair</code>	0.10 %
Matching stacks	in 2 stacks
<code>0.05 %</code> <code>0.05 %</code>	
<code>sun.security.provider.DigestBase.engineUpdate</code>	3.12 %
Matching stacks	in 14 stacks
<code>0.61 %</code> <code>0.61 %</code> <code>0.46 %</code> <code>0.41 %</code> <code>0.31 %</code> <code>0.15 %</code> <code>0.10 %</code> <code>0.10 %</code> <code>0.10 %</code> <code>0.05 %</code> 4 more...	

CodeGuru

Profiling groups

ColorProcessApp

Recommendation

Actions

2019-10-29 @ 13:00 - 14:00 GMT

Overview

CPU

com.company.ImageProcessor.start

89.97% of total

NATIVE: 22 minutes

TIMED: 7 minutes

WAITING: 7 minutes

Estimated active CPU cost: \$1320

Feedback

English (US)

© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Privacy Policy

Terms of Use

企业搜索 (Enterprise Search)

常见问题、新闻和政策...哦，我的天呐！



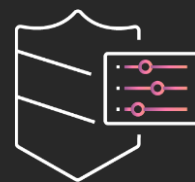
自由格式文本
多种格式



多个
数据孤岛



一个接一个的搜索“烟囱”
(按关键词)



复杂的权限
管理和安全

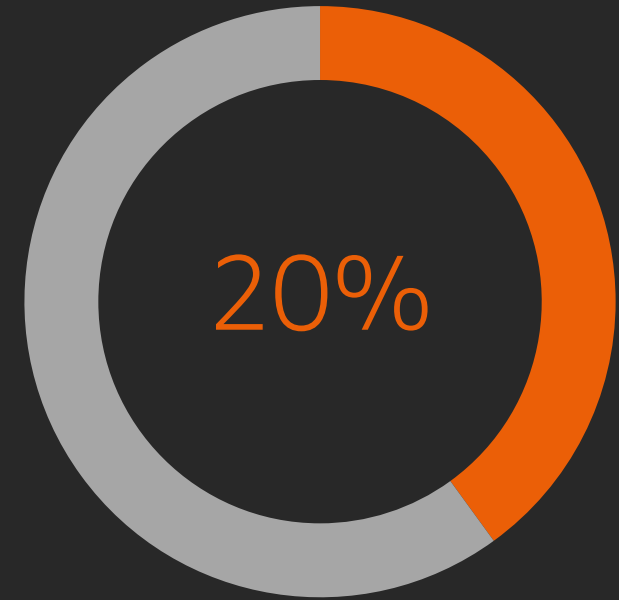


不友善的
关键字搜索

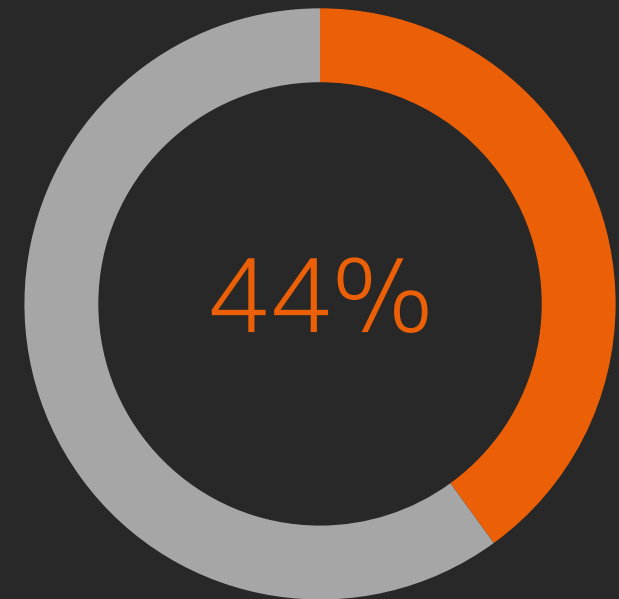


搜索结果
精度低

员工花费20%的时间来查找信息
—McKinsey



44%的时间里他们找不到工作
所需的信息
—IDC

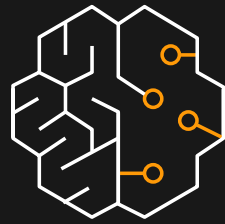


Amazon Kendra

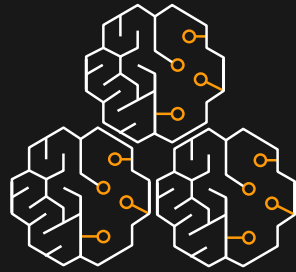
预览



自然语言
查询



自然语言理解
(NLU)与 机器
学习核心



领域专长



通过反馈的
增量学习



本机连接器
(S3、SharePoint、
文件服务器、
HTTP 等)

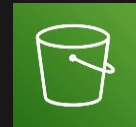


简单的 API
和控制台体验
代码样本

准确找到所需的内容

快速搜索、快速设置

Amazon Kendra connectors



Amazon S3



文件系统 (SMB)



Web 爬虫



数据库



SharePoint Online



SharePoint 2013,
2016, 2019



Box



Dropbox



Exchange



OneDrive



Google Drive



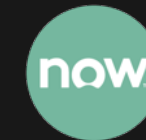
Salesforce



Confluence



Jira



ServiceNow



Zendesk



Jive

...2020年还将有更多

When is re:invent?



Kendra suggested answer

AWS re:Invent 2019

AWS **re:Invent** 2019 will be held **December 2-6, 2019** Las Vegas, NV. Download your calendar invites for upcoming AWS Training Talks (or listen to last week's recordings) below. Training Talks offer internal field reps a 60 mins deep dive into use cases, customer benefits, and a demo of new services... [\(More\)](#)

WEB | https://w.amazon.com/bin/view/AWS/reinvent_2019



What are Kendra suggested answers? [Info](#)

Frequently asked questions

[Where is re:Invent located?](#)



[How do I register for re:Invent?](#)



[When is re:Invent registration open?](#)



What are frequently asked questions? [Info](#)

1-10 of 65 Recommended documents

Sort: Relevancy ▼

APN re:Invent 2019

Encourage your APN Partners to join us December 2 at AWS Innovate, **re:Invent** Recap Edition. This is a free online conference providing a complete wrap up of AWS **re:Invent**... [\(More\)](#)

WEB | https://w.amazon.com/bin/view/APN_reinvent_2019/



re:Invent Live Stream Viewing Parties 2018

...Join us for **re:Invent** keynote viewing parties for our Amazon.com CDO organizations. We'll be providing breakfast, **re:Invent** water bottle giveaways, and a phone tool icon for your participation... [\(More\)](#)

WEB | <https://w.amazon.com/bin/view/reinventcampusevents/>



Deploy

Select the components you want on the page, copy the code snippets, then paste them into your search application.

Want it all? Copy all componens with our global code snippet.

[Copy all components](#)



Selected component

Kendra suggested answer

```
<div class="term-grid">
  <label>Lorem</label>
  <div class="definition">ipsum dolor sit amet
consectetur adipiscing</div>
  <div class="alternate">something</div>
</div>
.clearfix { display: inline-block; }
* html .clearfix { lorem: 1%; }
.clearfix { display: block; }
<div class="term-grid">
  <label>Lorem</label>
  <div class="definition">ipsum dolor sit amet
consectetur adipiscing</div>
  <div class="alternate">something</div>
</div>
.clearfix { display: inline-block; }
* html .clearfix { lorem: 1%; }
.clearfix { display: block; }
.clearfix { display: inline-block; }
* html .clearfix { lorem: 1%; }
```

[Copy code snippet](#)

Learn more

[Deploying Kendra in your web application.](#)

[Understanding API integration.](#)

Filter search results ▼

Search in

Everywhere (486)

Wiki (123)

Broadcast (265)

Sage (98)

Date range



Categories

☒ HR (23)

☐ IT (5)

☐ Marketing (5)

Authors

☐ miyingah (7)

☐ michealnb (4)

☐ angelak (4)

☐ masono (1)☐ kimfonta (1)☐ rwitaman (1)[Show more...](#)

amazon.com employee discount

Q

Kendra's suggested answer

New Hire Information

For more information, visit the Amazon US Benefits Inside Amazon page. **Employee Discount** Employees receive a **10% discount** (up to the value of \$100 annually) for selected products available on **Amazon.com**. **Employee discount** code numbers will be available once new hire documentation is complete... ([More](#))

WEB | https://w.amazon.com/bin/view/AmazonStudios/New_Hire_Information/



What is Kendra's suggested answer? [Info](#)

Frequently asked questions

How do I apply the amazon.com discount code?



How often can I use the amazon.com discount?



What items can I apply my employee discount?



What are frequently asked questions? [Info](#)

1-10 of 98 Recommended documents

Sort: Relevancy ▼

Employee Offer

With AT&T This wiki explains how Amazon employees can redeem their employee benefit...Note: Please refer to the **Employee Extras** page for current **discount** details. Some of the information below... [\(More\)](#)

WEB | <https://w.amazon.com/bin/view/key-benefits-offers>



机器学习服务

Amazon SageMaker 快速创新的步伐

50+

从去年开始
新增的功能

数据准备

SageMaker
Ground Truth

高精度的训练数据集，
并将标注成本降低
70%

内置算法

AWS ML
Marketplace

从合作伙伴获得数百种
新算法

一键训练

SageMaker
RL

率先添加强化学习算
法和模拟器

一键部署

Training
with Spot

比按需价格降低 90%
左右

训练和调优

SageMaker
Neo

以不到 1/10 的内存
占用量，速度提升 2
倍，且精度相同

数以万计的用户选择了Amazon Sagemaker



re:Invent 前的发布亮点

<https://aws.amazon.com/about-aws/whats-new/machine-learning>

- 在 Amazon Quicksight 中调用 Amazon SageMaker 模型
- 在 Amazon Aurora 中调用 Amazon SageMaker 模型
- 在同一 Amazon SageMaker 终端节点上部署多个模型

接下来机器学习挑战是什么？

使用 Kubernetes 进行机器学习 难于管理和扩展

在 Kubernetes 群集中构建和管理 机器学习的服务

+

使不同的开源库和框架以安全且可扩展的方式协同工作

+

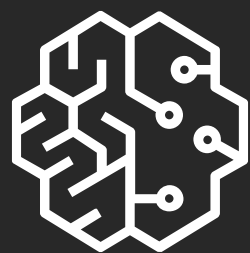
需要来自基础架构、数据科学和开发团队的时间和专业知识

=

需要一种更简单的方法来使用
Kubernetes 进行机器学习

适用于 Kubernetes 的 Amazon SageMaker Operators

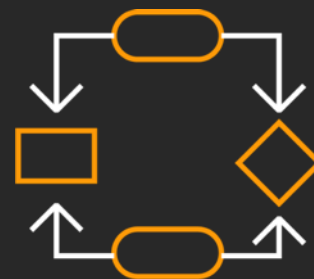
训练、优化和模型部署



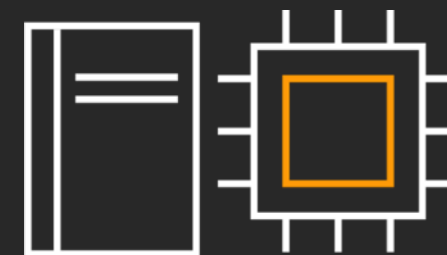
训练、优化和部署
SageMaker 的模型



在 Kubernetes 环境中
编排机器学习工作负载



创建 pipelines 和
Kubernetes 中的 workflow



完全托管的
SageMaker 基础设施



机器学习是迭代性的，涉及数十个工具和数百次迭代

不同阶段所需的多种工具
机器学习 workflow

+

缺乏集成化的体验

+

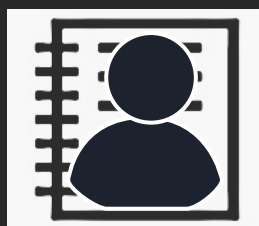
大量的迭代

=

繁琐、冗长的流程，导致生产力损失

Amazon SageMaker Studio

第一个用于机器学习的集成开发环境 (IDE)



大规模地项目合作

共享可扩展的 Notebooks
无需跟踪代码依赖性



轻松的实验管理

数千个实验的组织、
跟踪和比较



自动生成模型

无需编写代码即可获得具有完
全可见性和控制的准确模型



更高质量的
机器学习模型

自动调试错误、监控模型和
保持高质量



提高生产效率

在统一的可视化界面中编写、
构建、训练、部署和监控

xgboost_customer_churn.ipynr

conda_amazonei_mxnet_p27

- Have the predictor variable in the first column
- Not have a header row

But first, let's convert our categorical features into numeric features.

```
[ ]: model_data = pd.get_dummies(churn)
model_data = pd.concat([model_data['Churn?_True.'], model_data.drop(['Churn?_True.'], axis=1)])
```

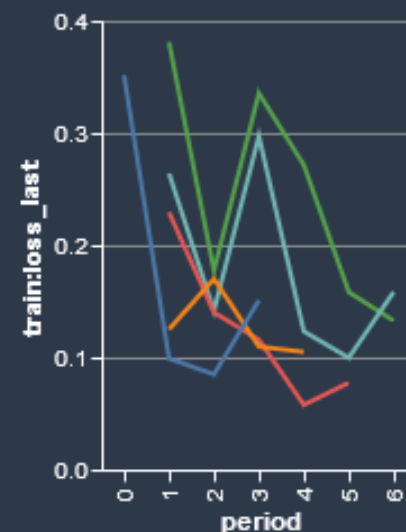
And now let's split the data into training, validation, and test sets. This will help prevent us from overfitting the model, and allow us to test the models accuracy on data it hasn't already seen.

```
[ ]: train_data, validation_data, test_data = np.split(model_data.sample(frac=1), [int(model_data.shape[0]*0.7), int(model_data.shape[0]*0.8)])
train_data.to_csv('train.csv', header=False, index=False)
validation_data.to_csv('validation.csv', header=False, index=False)
```

Now we'll upload these files to S3.

```
[ ]: boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'train.csv')).upload_file('train.csv')
boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'validation.csv')).upload_file('validation.csv')
```

Trial Component Chart



Trial Component List



TRIAL COMPONENTS

10 rows selected

Add chart

Deploy model



Status	Experiment	Type	Trial	Trial c
✓ Completed	customer-churn-predi...	Training job	Trial-3	Tr
✓ Completed	customer-churn-predi...	Training job	Trial-2	Tr
✓ Completed	customer-churn-predi...	Training job	Trial-1	Tr
✓ Completed	customer-churn-predi...	Training job	Trial-0	Tr



**数据科学和协作必须要
简单易用**

设置和管理资源

+

多个数据科学家进行协作

+

不同的数据科学项目有不
同的资源需求

=

**管理 Notebooks 和与多个数
据科学家进行协作非常复杂**

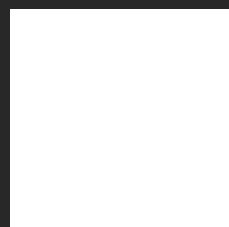
Amazon SageMaker Notebooks

快速启动的可共享的 Notebooks



通过单一登录
(SSO) 轻松访问

使用公司凭据在数秒内
访问 Notebooks



全面管理
和安全性

管理员可管理
存取和权限



无需显式设置

在不增加计算资源
的情况下启动 Notebooks



轻松协作

只需单击一下即可将
Notebooks 作为 URL 共享



灵活性

上调或下调计算资源



Launcher



my_notebook.ipynb



Markdown ▾



git

conda_pytorch_p36



The following shell code shows how to build the container image using `docker build` and push the container image to ECR using `docker push`. This code is also available as the shell script `container/build-and-push.sh`, which you can run as `build-and-push.sh pytorch-extending-our-containers-cifar10-example` to build the image `pytorch-extending-our-containers-cifar10-example`.

This code looks for an ECR repository in the account (the SageMaker notebook instance was created in). If the repository does not exist, it creates it. If the repository exists, it will need to retrieve ECR credentials to pull this package.

```
[ ]: %%sh
```

```
# The name of our algorithm
algorithm_name=pytorch-extending-our-containers-cifar10-example

cd container

account=$(aws sts get-caller-identity --query Account --output text)

# Get the region defined in the current configuration (default to us-west-2 if none defined)
region=$(aws configure get region)
region=${region:-us-west-2}

fullname="${account}.dkr.ecr.${region}.amazonaws.com/${algorithm_name}:latest"

# If the repository doesn't exist in ECR, create it.

aws ecr describe-repositories --repository-names "${algorithm_name}" > /dev/null 2>&1

if [ $? -ne 0 ]
then
    aws ecr create-repository --repository-name "${algorithm_name}" > /dev/null
fi

# Get the login command from ECR and execute it directly
$(aws ecr get-login --region ${region} --no-include-email)

# Get the login command from ECR in order to pull down the SageMaker PyTorch image
$(aws ecr get-login --registry-ids 520713654638 --region ${region} --no-include-email)
```

Create shareable snapshot

Share as

my_notebook.ipynb

☒ Include Git repo information☐ Include notebook output

Cancel

Create snapshot

Sharing notebook: MyNotebook.ipynb



数据处理与 模型评估涉及大量运营开销

为数据处理工作负载构建和扩展基础架构非常复杂

+

使用多种工具或服务意味着学习和实施新的 API

+

机器学习 workflows 中的所有步骤都需要增强安全性、身份验证和合规性

=

需要构建和管理工具以运行大型数据处理和模型评估工作负载

Amazon SageMaker Processing

用于数据处理和模型评估的分析任务



全面管理

实现集群的
分布式处理



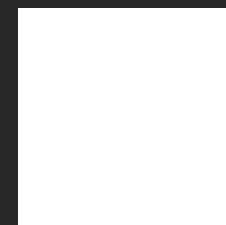
自定义处理

带上自己的脚本
进行特征工程



支持容器

使用 SageMaker 的内置
容器或自带容器



安全性和合规性

利用 SageMaker 的
安全性和合规性功能



自动创建 & 终止

资源是自动创建,
配置和终止的



管理试验和实验是件麻烦的事

数以百计的实验

+

每个实验数百个参数

+

比较和对照

=

非常繁琐且容易出错

Amazon SageMaker Experiments

一种组织、跟踪和评估训练实验的系统



大规模实验跟踪

跨实验和用户跟踪
参数和指标



组织实验

按团队，目标和假设
进行组织



通过可视化以获
得最佳结果

可视化并对实验进行比较



具有灵活性的
Python SDK 和 API

记录自定义指标 &
使用 API 跟踪模型



快速迭代

迭代和开发高质量模型

TRIAL COMPONENTS 5 rows selected. Select rows to toggle chart visibility.

Add Chart

	Experiment	Trial	Trial Component	Type	Training time	Actions
	customer-churn-predi...	Trial-9	Training-Run-9-aws-training-job	Training job	~6 minutes	Remove
	customer-churn-predi...	Trial-8	Training-Run-8-aws-training-job	Training job	~5 minutes	Remove
	customer-churn-predi...	Trial-7	Training-Run-7-aws-training-job	Training job	~6 minutes	Remove
	customer-churn-predi...	Trial-6	Training-Run-6-aws-training-job	Training job	~4 minutes	Remove
	customer-churn-predi...	Trial-5	Training-Run-5-aws-training-job	Training job	~4 minutes	Remove

1 CHART

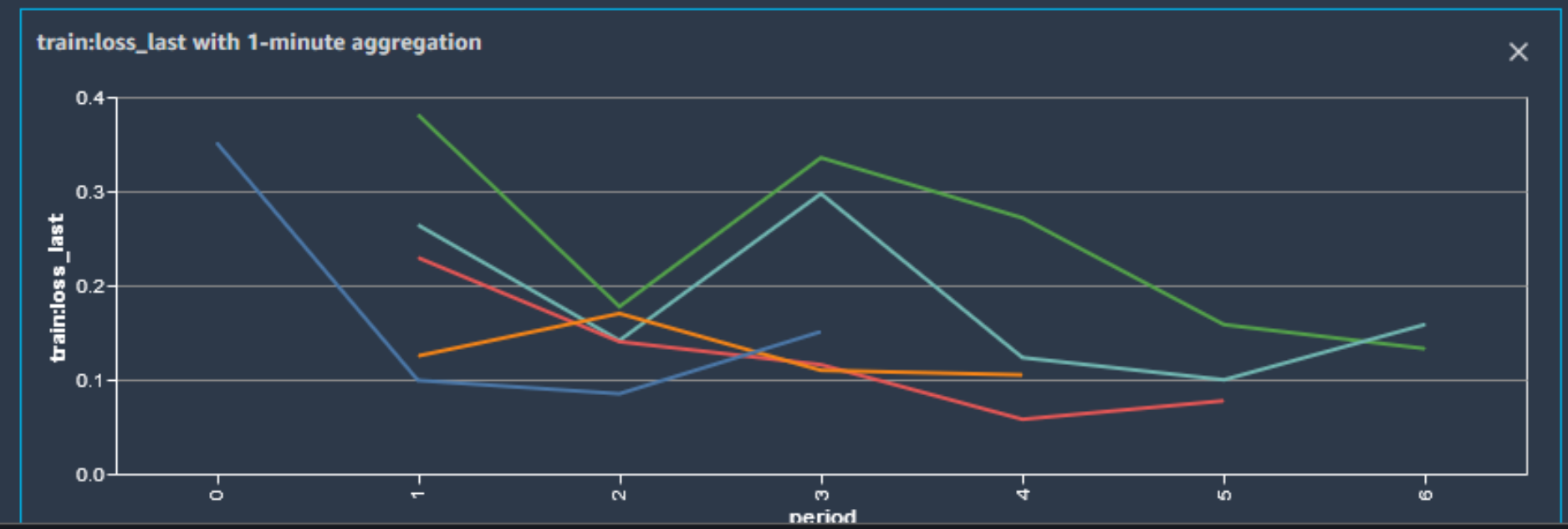


CHART PROPERTIES

Data type

- ☐ Time series
- ☒ Summary statistics

Chart type

- ☒ Histogram
- ☐ Line

X-axis dimension

- ☒ Epoch
- ☐ Time
- ☐ Periods from start

X-axis aggregation

- ☐ 1-minute
- ☒ 5-minute
- ☐ 60-minute

Y-axis

train:loss_last



深度学习的调试和分析 是让人痛苦的

大型神经网络
具有许多图层

+

许多连接

+

用于分析和调试的其它工具

=

检查，调试和分析“黑匣子”非常困难

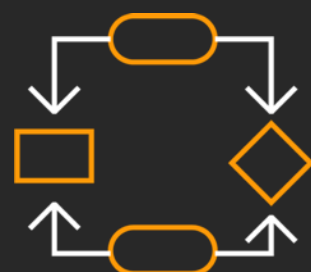
Amazon SageMaker Debugger

分析和调试，可解释性和警报生成



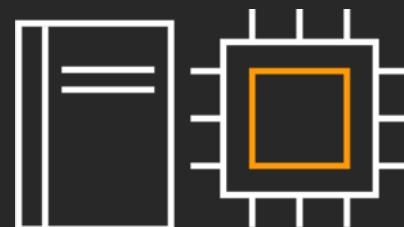
相关数据采集

自动捕获数据进行分析



自动数据分析

无需更改代码而分析和调试数据



自动错误检测

根据规则自动检测错误



通过警告提高
工作效率

根据警告采取纠正措施



可视化分析
和调试

从 SageMaker Studio
进行可视化分析和调试

SMDEbugger-CloudWatch-Loi X

conda_tensorflow_p36 git

Using SageMaker Rules

In this example we'll demonstrate how to use SageMaker rules to be evaluated against your training. You can find the list of SageMaker rules and the configurations best suited for using them here.

We specify a few rules that check for overfitting, decrease in loss across epochs and for saturated activations.

[8]: estimator = TensorFlow(
 role=sagemaker.get_execution_role(),
 base_job_name='mnist-tensorflow-example',
 train_instance_count=1,
 train_instance_type='ml.p3.2xlarge',
 image_name=cpu_training_image,
 entry_point=entrypoint_script,
 framework_version='1.15',
 py_version='py3',
 train_max_run=3600,
 script_mode=True,
 sagemaker_session=sess,
 ## New parameter
 rules = [Rule.sagemaker(rule_configs.vanishing_gradient()),
 Rule.custom(name='Overfitting', # used to identify the rule
 image_uri='759209512951.dkr.ecr.us-west-2.amazonaws.com,
 instance_type='ml.c4.xlarge', # instance type to run the
 source='my_custom_rule.py', # path to the rule source file
 rule_to_invoke='CustomGradientRule', # name of the class
 volume_size_in_gb=400, # EBS volume size required to be
 collections_to_save=[CollectionConfig(name='losses')], #
 rule_parameters={
 "threshold": "20.0" # this will be used to initialize
 })],
 hyperparameters = {'num_epochs' : 100 }
)

Note that Sagemaker-Debugger is only supported for py_version='py3' currently.

Let's start the training by calling fit() on the MXNet estimator

[9]: # After calling fit, SageMaker will spin off 1 training job and 1 rule job for you
The rule evaluation status(es) will be visible in the training logs
at regular intervals

estimator.fit(wait=False)

Result

Describe Trial Component X

Experiment: Unassigned
Trial: Unassigned

Trial stagesChartsMetricsParametersArtifactsAWS SettingsDebugger

mnist-tensorflow-example-2019-12-02-09-52-13-126-aws-training-job
Created 15 minutes ago

Status	Last modified	Rule name	Job ARN
Issues Found	4 minutes ago	VanishingGradient	arn:aws:sagemaker:us-west-2:3...
Issues Found	4 minutes ago	Overfitting	arn:aws:sagemaker:us-west-2:3...

Trial Component Chart X

TRIAL COMPONENTS 1 rows selected. Select rows to toggle chart visibility. Add Chart

Experiment	Trial	Trial Component	Type	Train
N/A	N/A	mnist-tensorflow-example-2019-12-02-09-52-13-126-aws-trainin...	Training job	~10

2 CHARTS

sparse_softmax_cross_entropy_loss/value:0_avg with 1-minute aggregation X

period	loss
8	38
9	38
10	38
11	36

trialComponentName
— mnist-tensorflow-example-2019-1...



**部署模型不是工作的结束，还需要在
生产中持续监视模型并迭代**

由于数据差异而导致的概
念漂移

+

模型性能可能因未知因素
而改变

+

持续监控模型性能和数据需要付出
大量精力和费用

=

模型监控既繁琐又关键

Amazon SageMaker Model Monitor

持续监控生产环境中的模型



自动数据
收集

数据会自动从端点收集



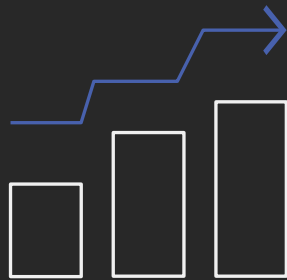
持续
监控

定义监控计划，根据预定义的基线检测质量变化



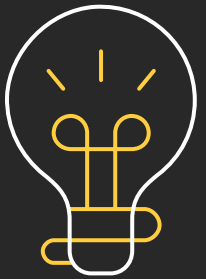
通过规则
实现灵活性

使用内置规则检测数据漂移或编写自己的规则进行自定义分析



数据可视化
分析

在 SageMaker Studio 中查看监控结果、数据统计信息和异常报告



CloudWatch
集成

根据 Amazon CloudWatch 警报自动执行纠正措施

ENDPOINTS	
Name	Endpoint status
AutoML-toms-super-notebo...	✓ InService
UC-DEMO-xgb-churn-pred-...	✓ InService
demo-xgboost-customer-ch...	✓ InService
mengyw-test-form-1	✓ InService
mengyw-test-capture	✓ InService
my-sagemaker-ap1	✓ InService
mengyw-test-vpc-3	✗ Failed
DEMO-xgb-churn-pred-mod...	✓ InService
mengyw-test-vpc-2	✓ InService
mengyw-test-vpc	✗ Failed

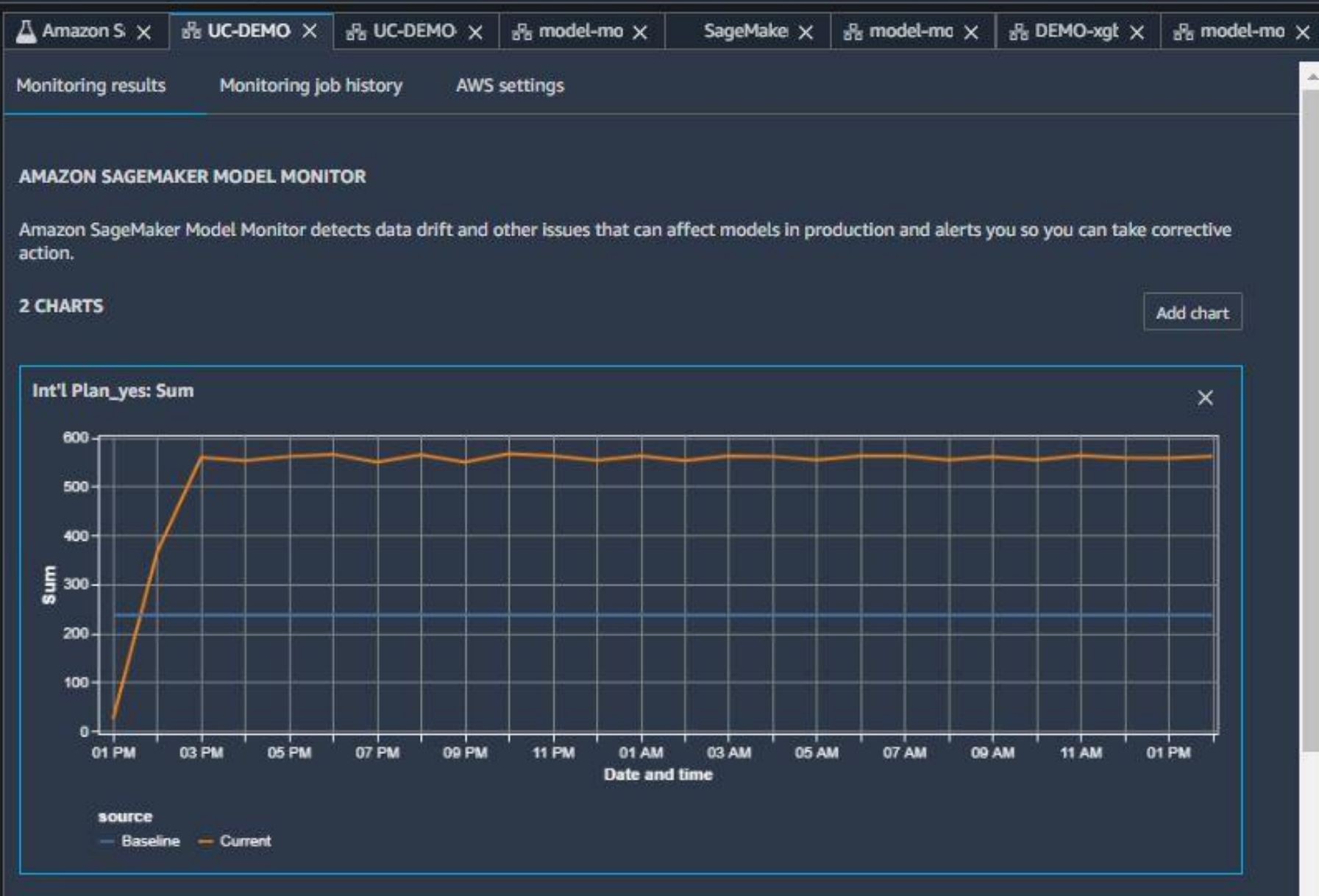


CHART PROPERTIES

Timeline

- ☒ 1 week
- ☐ 1 day
- ☐ 12 hours
- ☐ 6 hours
- ☐ 1 hour

Statistic

- ☐ Average
- ☐ SampleCount
- ☒ Sum
- ☐ Minimum
- ☐ Maximum

Feature

feature_data_Int'l Plan_yes ▼



**成功的机器学习需要复杂且难以发现的
算法、数据、参数组合**

大量探索性 =
迭代

+

需要广泛而完整的
机器学习领域的知识

+

缺乏可视化

=

**耗时，容易出错的过程，
即使对于机器学习专家也
是如此**

Amazon SageMaker Autopilot

具有完全可见性和控制功能的自动模型创建



快速启动

以表格形式提供数据
并指定目标预测



自动化
模型创建

通过特征工程和自动模型优化
获取机器学习模型



可视化 & 控制

在 Notebook 上运行源代码
以获得模型



建议与优化

获得排行榜并
继续改善模型

+

↑

↺

/ automl-preview /

Name	Last Modified
bank-additional	12 hours ago
model	12 hours ago
• sagemaker_auto...	3 hours ago

sagemaker_automl_direct_m; × my-sagemaker-autopilot × Deploy model ×

↺

EXPERIMENT: MY-SAGEMAKER-AUTOPILOT

Open candidate generation notebook Open data exploration notebook

Trials Job profile

TRIALS

0 row selected

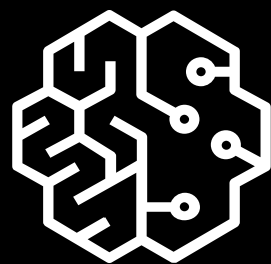
Deploy model

Trial name	Status	Start time	End time	▼	Objective
my-sagemaker-tuning-job-...	Completed	9 hours ago			0.9206119775772095
my-sagemaker-tuning-job-...	Completed	9 hours ago			0.9202479720115662
my-sagemaker-tuning-job-...	Completed	7 hours ago			0.9200050234794617
my-sagemaker-tuning-job-...	Completed	7 hours ago			0.9195190072059631
my-sagemaker-tuning-job-...	Completed	9 hours ago			0.9191550016403198
my-sagemaker-tuning-job-...	Completed	7 hours ago			0.9190340042114258
my-sagemaker-tuning-job-...	Completed	8 hours ago			0.9189119935035706
my-sagemaker-tuning-job-...	Completed	8 hours ago			0.9186699986457825
my-sagemaker-tuning-job-...	Completed	8 hours ago			0.9186699986457825

0 \$ 1

my-sagemaker-autopilot

快速构建，训练和部署机器学习模型



Amazon SageMaker

Ground Truth

Marketplace for ML

Algorithms & Frameworks

Collaborative notebooks

Distributed Training & Debugger

Tuning & Optimization

Experiments

Deployment, Monitoring, & Hosting

Reinforcement Learning

SageMaker Studio

SageMaker AutoPilot

AWS DeepRacer 的改进

- AWS DeepRacer **Evo**
 - 双目摄像头 (Stereo camera)
 - 激光雷达 (LIDAR sensor)
- 全新的 **比赛机会**
 - 创建自己的比赛
 - 目标检测 & 避让
 - 一对一的比赛



Mod your own vehicle

Mod specifications

The garage shows the DeepRacer vehicles that you can train models for. You can add vehicles by using the "build new vehicle button"

Sensor modification

Swap sensors to improve your DeepRacer's racing performance

☐ Front-facing camera

Single camera that captures the images with sizes of 160 x 120 in front of the agent at 15 fps. The camera has 120 wide angle lens. The images are converted into grey scale before being fed to the neural network

► Benefits of the front-facing camera

☒ Stereo cameras (right/left) sensor

Composed of two front-facing cameras, stereo cameras can generate depth information of the objects in front of the agent and thus be used to detect and avoid obstacles on the track. The cameras capture images with the same resolution and frequency. Images from both cameras are converted into grey scale, stacked and then fed into the neural network.

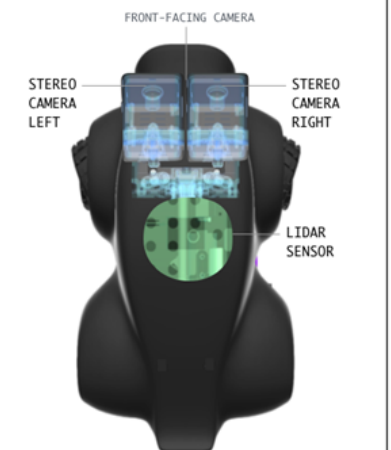
► Benefits of the stereo camera

Add-on sensors

☒ LIDAR sensor

LIDAR is a surveying method that measures a distance to a target by illuminating the target with laser light and measuring the reflected light with a sensor.

► How LIDAR works with autonomous driving





intel[®]
inside™

A close-up photograph of a white plastic component, likely a laptop bezel, featuring the Intel Inside logo. The logo consists of the word "intel" in a stylized font inside a circular swoosh, with "inside™" written below it. To the left of the logo, a small, bright blue LED light is visible. The background is dark and out of focus, showing some mechanical details and another blue light source.

AWS DeepComposer

- MIDI 键盘使用机器学习进行音乐生成
- 使用生成对抗网络 (GAN) 创作音乐
- 使用预先训练的模型，或训练您自己的模型





框架与基础设施

AWS Neuron SDK

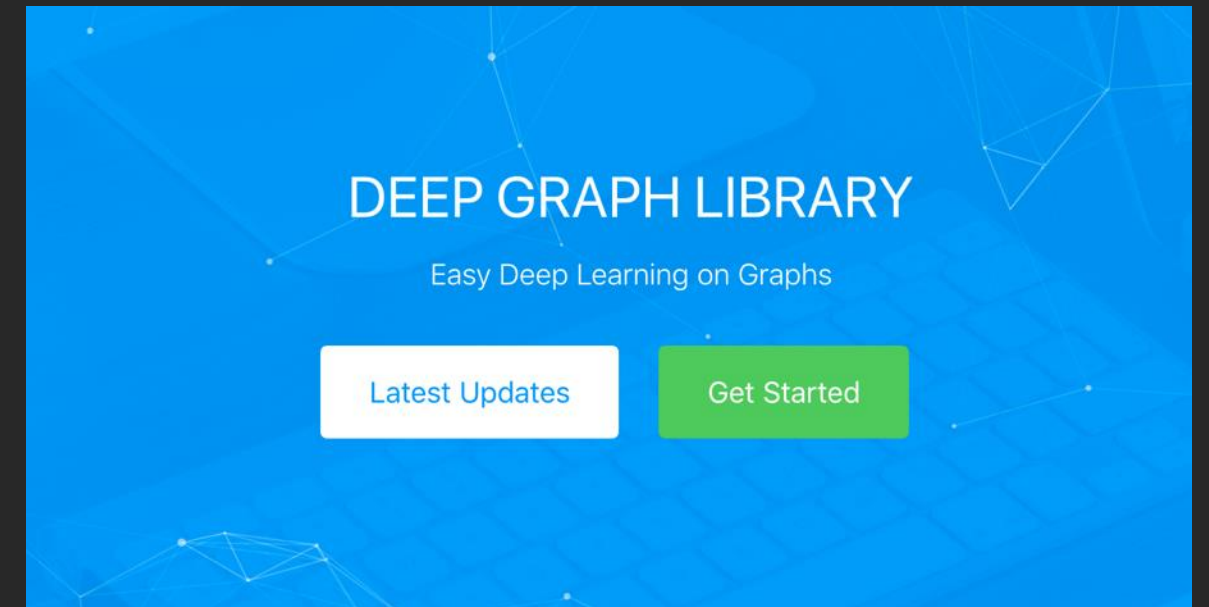
- 极低的以成本实现快速、低延迟的推理
 - 16位浮点 (FP16和 BF16) 上的64 teraOPS 和混合精度数据
 - 8位整数 (INT8) 数据上为128 teraOPS
- Neuron SDK: <https://github.com/aws/aws-neuron-sdk>
 - Deep Learning AMIs 与 Deep Learning Containers 已经包含
 - 支持 TensorFlow 与 Apache MXNet, PyTorch 即将支持

实例名称	Inferentia 芯片	vCPUs	RAM	EBS 带宽
inf1.xlarge	1	4	8 GiB	Up to 3.5 Gbps
inf1.2xlarge	1	8	16 GiB	Up to 3.5 Gbps
inf1.6xlarge	4	24	48 GiB	3.5 Gbps
inf1.24xlarge	16	96	192 GiB	14 Gbps

Deep Graph Library

<https://www.dgl.ai>

- Python 开源库，可帮助研究人员和科学家在其数据集上快速构建、训练和评估图神经网络
- 用例：推荐、社交网络、生命科学、网络安全等.
- Deep Learning Containers 中已经包含
 - 支持 PyTorch 与 Apache MXNet, TensorFlow 即将支持
- 可在 Amazon SageMaker 上进行训练



Deep Java Library

<https://www.djl.ai>

- Java 开源库，训练和部署模型
- 框架支持
 - 现已支持 Apache MXNet, 将会支持更多框架
- 训练自己的模型，或使用 Model Zoo 的预训练模型



Amazon 构建者库

<http://aws.amazon.com/builders-library>

“如果每个人都能快速高效地创新，并且事情运转地很顺利，世界将变得更美好，这一点毫无疑问。因此，我感到很开心能在 Amazon 构建者库中与大家分享我们学到的知识。”

– Charlie Bell, AWS 高级副总裁

架构

LEVEL 400



避免无法克服的队列积压

Author: David Yanacek

优先快速处理队列积压中的重要工作负载



架构

LEVEL 200



分布式系统相关挑战

Author: Jacob Gabrielson

介绍为分布式系统带来挑战性的系统属性



软件交付与运营

LEVEL 300



采用持续交付，加速交付进度

Author: Mark Mansour



软件交付与运营

LEVEL 400



确保部署期间安全回滚

Author: Sandeep Pokkuri



架构

LEVEL 300



使用可用区的静态稳定性

Authors: Becky Weiss, Mike Furr

设计使用多个可用区来实现高可用性和确保系统静态稳定的架构。



架构

LEVEL 300



避免在分布式系统中回退

Author: Jacob Gabrielson



架构

LEVEL 300



缓存挑战和策略

Authors: Matt Brinkley, Jas Chhabra

架构

LEVEL 300



分布式系统中的领导选举

Author: Marc Brooker

架构

LEVEL 200



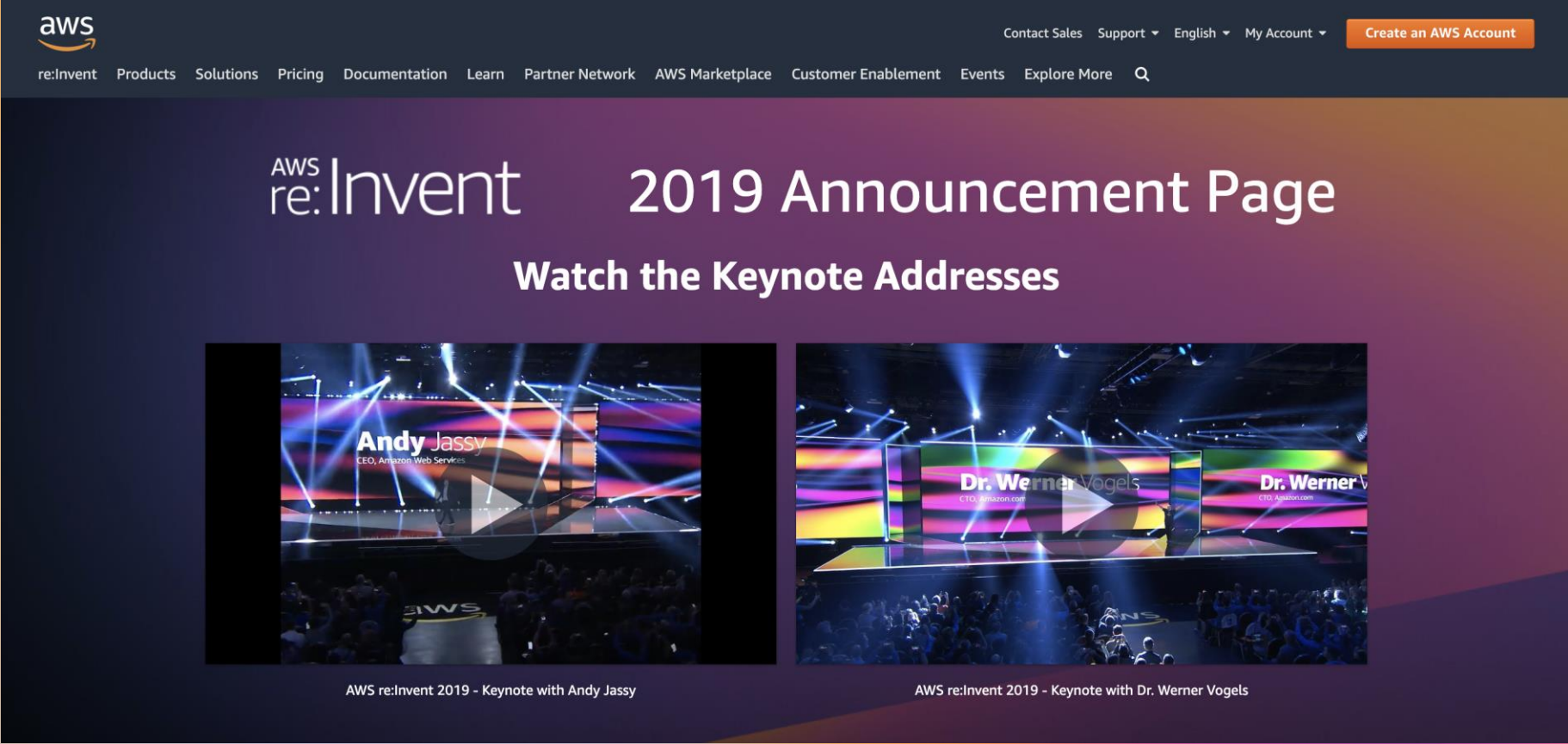
超时、重试和抖动回退

Author: Marc Brooker

re:Invent 2019, 最值得期待的五项发布

- 1. Graviton2 for EC2**
- 2. Amazon Kendra**
- 3. AWS CodeGuru**
- 4. Inf1 EC2 instances with Inferentia**
- 5. Amazon SageMaker Autopilot**

这里是所有新的发布!



The screenshot shows the AWS re:Invent 2019 Announcement Page. The header includes the AWS logo, navigation links (re:Invent, Products, Solutions, Pricing, Documentation, Learn, Partner Network, AWS Marketplace, Customer Enablement, Events, Explore More), and a search icon. A 'Create an AWS Account' button is in the top right. The main content area features the text 'AWS re:Invent 2019 Announcement Page' and 'Watch the Keynote Addresses'. Below this are two video thumbnails: 'AWS re:Invent 2019 - Keynote with Andy Jassy' and 'AWS re:Invent 2019 - Keynote with Dr. Werner Vogels'. Each thumbnail has a play button overlay.

aws

Contact Sales Support English My Account Create an AWS Account

re:Invent Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Customer Enablement Events Explore More

AWS re:Invent 2019 Announcement Page

Watch the Keynote Addresses

Andy Jassy
CEO, Amazon Web Services

Dr. Werner Vogels
CTO, Amazon.com

AWS re:Invent 2019 - Keynote with Andy Jassy

AWS re:Invent 2019 - Keynote with Dr. Werner Vogels

<https://aws.amazon.com/new/reinvent>

资源下载

AWS Events Content

Experience our event content at your fingertips. Explore, view, and download presentation decks from your favorite sessions and discover what's new. Learn from AWS experts, customers, and partners to continue your educational journey in the cloud.

Filter by:
[Clear filters](#)

▼ Event Series

- ☒ re:Invent
- ☐ Summits

▼ Session type

- ☐ 20 Minute Presentation
- ☐ Builders session
- ☐ Chalk talk
- ☐ Session
- ☐ Workshop

▼ Learning Level

- ☐ 200 - Intermediate
- ☐ 300 - Advanced
- ☐ 400 - Expert

▼ Topics

- ☐ Amazon.com
- ☐ Alexa
- ☐ Analytics

CHN202 LEVEL 200



AWS China Gateway: Power your business in China by...

The AWS China Gateway provides information that helps customers of all sizes get started in using AWS to grow their businesses in China. The AWS China (Beijing) Region and the AWS China (Ningxia) Region are located within mainland China. To provide the best experience for customers in China, as well as to comply with China's legal

re:Invent
19-Dec

WINDOWS & .NET

WIN407-R1 LEVEL 400



[REPEAT 1] Performance tune SQL Server using...

When you are running critical database workloads on SQL Server it's vital that SQL Server is configured for optimal performance. In this builders session, we outline an approach for how to perform deep inspection of the workloads running on SQL Server and tune configuration settings to match the workload's performance needs. We

re:Invent
19-Dec

WINDOWS & .NET

WIN406-R1 LEVEL 400



[REPEAT 1] Build multi-region SQL Server Always...

Multi-region SQL Server database configurations offer the benefits of both high availability and disaster recovery. Join us for this builders session, where you learn how to design and deploy SQL Server in a multi-region architecture on AWS and perform the configuration of a multi-region SQL Server Always On

re:Invent
19-Dec

<https://aws.amazon.com/cn/events/events-content/?awsf.filter-series=event-series%23reinvent>

视频回看



AWS Events

6390 位订阅者

首页

视频

播放列表

社区

频道

简介

AWS re:Invent 2019 - Daily Recap Episode 5...



AWS re:Invent 2019 - Daily Recap Episode 5 - Thursday

4,924次观看 · 2周前

Learn more about AWS re:Invent at <http://bit.ly/2s46SnA>. Join AWS Senior Solutions Architect, Rudy Chetty at AWS re:Invent in Las Vegas to explore the quirky side of this cloud computing conference. In the 5th and final episode of 2019, Rudy wraps up the week with the AWS Deep Racer Finals, Dr. Werner Vogels keynote, a visit to the AWS Certification lounge, and re:Play. Rudy finds out what customers have learned, what

了解详情

AWS re:Invent 2019 - Daily Recap

全部播放



2:57



1:56



1:57



2:00



1:56

AWS re:Invent 2019 - Daily Recap Episode 5 - Thursday

AWS Events

4924次观看 · 2周前

AWS re:Invent 2019 - Daily Recap Episode 4 - ...

Amazon Web Services

3454次观看 · 2周前

AWS re:Invent 2019 - Daily Recap Episode 3 - Tuesday

AWS Events

5031次观看 · 2周前

AWS re:Invent 2019 - Daily Recap Episode 2 - Monday

AWS Events

7146次观看 · 2周前

AWS re:Invent 2019 - Daily Recap Episode 1 - Sunday

AWS Events

1.4万次观看 · 2周前

AWS re:Invent 2019 Breakout Sessions | Games/GameTech

全部播放



55:55



55:37



1:00:09



44:11

AWS re:Invent 2019: How Rovio teaches Angry Birds t...

AWS Events

543次观看 · 2周前

AWS re:Invent 2019: Migrating the live Pokémon...

AWS Events

902次观看 · 2周前

AWS re:Invent 2019: How CAPCOM builds fun games...

AWS Events

427次观看 · 2周前

AWS re:Invent 2019: How Call of Duty uses ML to...

AWS Events

162次观看 · 1周前

<https://www.youtube.com/channel/UCdoadna9HFHsxXWhafhNvKw>



在这里让我们开始起步，
Go Build!



Thank you!

