DAMS

中国数据智能管理峰会

DATA & AI MANAGEMENT SUMMIT

基于数据湖构建云上的数据分析架构

演讲人：朱澜 AWS 资深架构师

企业应用数据的演进

数据察觉

数据监控

数据驱动

功能	Report	Monitor	Guide
能力	Static batch reports	Interactive dashboards, data warehouse	Data science, AI/ML, Data Lake
用户	Executives, department heads	Power users	Everyone
数据	Financial and operational data	Siloed data	All data

企业应用数据的现状和常见的功能障碍

发展趋势

随着数字化技术的成熟发 展，企业现在比以往任何时候都需要更好地处理自身拥有的数据，成为数据驱动型组织

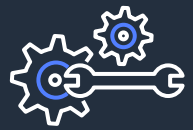
现实情况

- 企业越来越认识到数据的价值
- 企业在使用越来越多的复杂的技术捕获数据和处理数据。然而，
- 仍然有超过9成的数据没有被用到
- 有超过8成的企业被认为应用数据能力低

应用数据的能力赤字严重地限制了企业的发展并削弱了生存能力

孤立数据和被丢弃的数据

低保真数据



多样性的处理

散乱数据

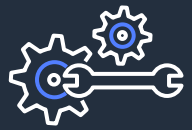
现代分析平台要具备的特点

访问想要的任何数据

- 数据驱动型决策需要访问众多不同类型数据
- 多种类型不同来源的数据转储到企业数据湖

对变化的响应性
（数据的访问速度决定了决策速度）

- 接近实时的处理和报告数据
- 即时响应上游数据源的变化
- 采用Amazon S3或Hadoop等大数据技术存储数据
- 采用流处理技术



提供交互式洞察的方式

- 正确的时间正确的工具以正确的形式提供数据
- 需要同时使用多种工具以满足不同用户需求
- 支持机器学习探索

智能嵌入业务流程

- 算法平台与业务平台集成
- 业务平台能集成外部数据源或API
- 能利用所有可用云服务帮助实现系统的现代化

本地数据平台难以支持现代化数据分析的要求

传统的数据技术链面临挑战，并且一直在艰难地试图适应企业规模的数据发展变化

无法扩展数据存储和处理时间

增加总体拥有成本以支持数据管理

数据模型变更延迟

数据分析部署时间长使洞察滞后

云数据平台帮助构建新型的数据洞察力和驱动力

围绕业务挑战、业务趋势和业务模式

快速洞察

- 更快、实时的洞察力
- 内部数据仓库无法实现的业务洞察力

创新

- 利用数据进行业务创新
- 设计未来的产品和服务

加强客户体验

- 处理数据多样性
- 移动、社交等多种数据来源

专注于卓越运营

- 利用云技术降低总体拥有成本
- 利用数据提供最高级别的卓越运营

确保法规遵从

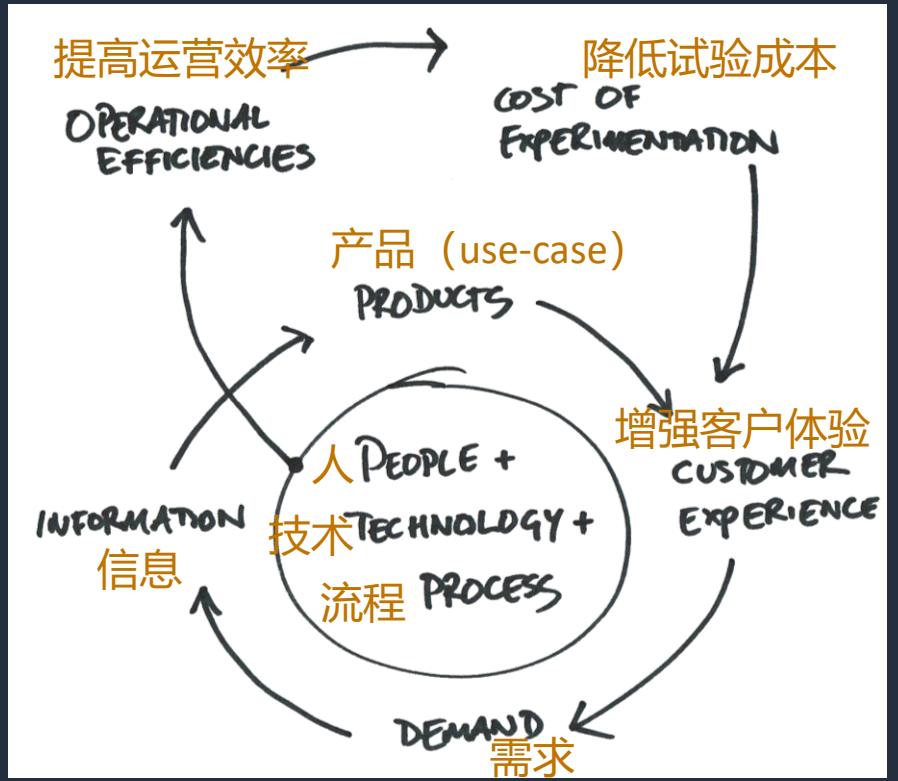
- 减少部署时间
- 网安法/等保2.0
- 行业合规要求

亚马逊对数据驱动型企业的理解

“一个将数据作为战略资产加以利用的组织，以驱动创新，并建立可行动的洞察力，从而为其客户、合作伙伴和员工提供增强体验”

关键词

- 资产：常被忽视、低估或误解
- 持续创新： 关键是持续，实现数据驱动的自我推动力
- 可行动的洞察力：能推动业务的洞察
- 客户体验：要增强体验就会产生新的特性和产品需求



亚马逊数据分析飞轮模型

持续创新的数据分析飞轮应用举例

传统数据平台的模式已成为数据分析能力的瓶颈

应用程序

- 集中的数据存储
- 集中的数据团队
- 企业BI能力中心

数据消费者

构建能支持敏捷业务的现代化数据分析平台

数据驱动的组织通过将责任扩大到边缘，将责任推广到数据的生产者和消费者身上，从而实现敏捷性

数据生产者

想要分享数据的团队

- 领域专业知识
- 数据所有权和治理
- 数据质量
- 元数据管理

数据湖平台

运行数据市场的团队

- 简化数据的注入
- 安全控制
- 构建并运行平台
- 企业数据集
- 培训和社区

数据消费者

想要使用数据的团队

- 执行业务优先级
- 业务分析开发
- 数据发现
- 数据管道的开发
- 创造新的洞察

现代化数据分析平台的需求

Data discovery, search, and collaboration

- Catalog and search
- Share data
- Dashboards
- Interactive Query
- SQL based pipelines

Support exploratory data analysis and ML

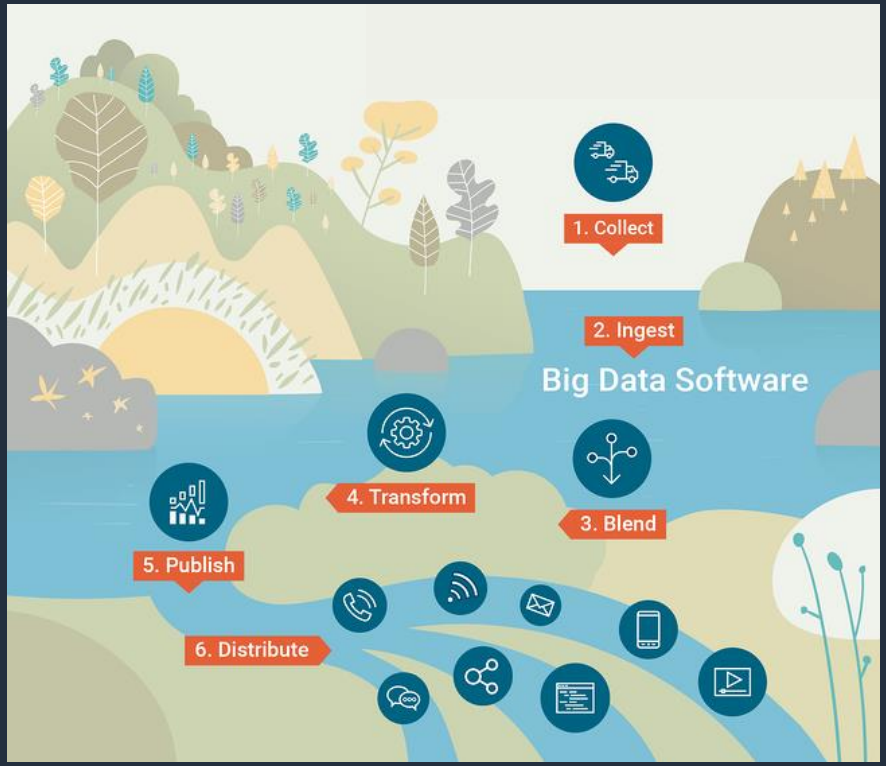
- Exploratory research
- Notebook automation
- Predictive analytics
- Operational analytics
- Pipeline scheduling

Data processing and platform frameworks

- Data ingestion
- Data transformation
- Data quality
- Code and infrastructure automation
- Security and management

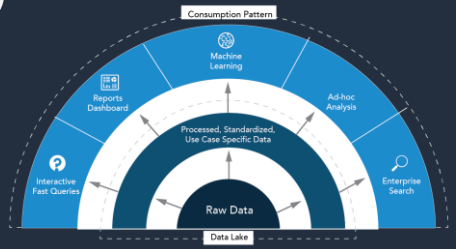
数据湖的定义

数据湖是一个集中式存储库，允许您以任意规模存储所有结构化和非结构化数据。您可以按原样存储数据（无需先对数据进行结构化处理），并运行不同类型的分析 – 从控制面板和可视化到大数据处理、实时分析和机器学习，以指导做出更好的决策。



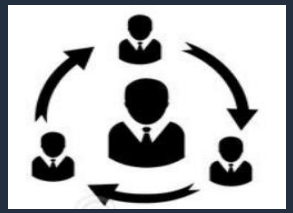
应用数据湖 实现企业数据变成资产

- 业务目标
  - 数字化经济，数据驱动业务
  - 提升企业运营效率
  - 预判发展趋势，提升企业竞争力
- 技术目标
  - 停止丢弃数据
  - 分析无处不在，采用多种技术
  - 自动化, API 化
  - 赋能给更多用户，建立数据探索能力
- 敏捷，自助式服务
- 协作，促进企业内部协作



建立数据探索能力

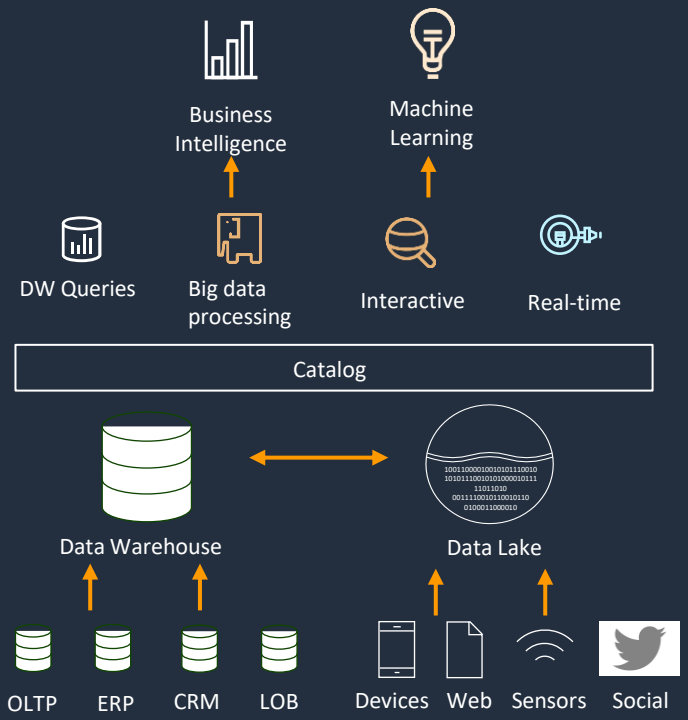
Reactive → Predictive



敏捷, 协作经济

被动式 → 自助式

基于亚马逊的数据湖构建云上大数据分析平台



数据湖提供:

存储关系型和非关系型数据

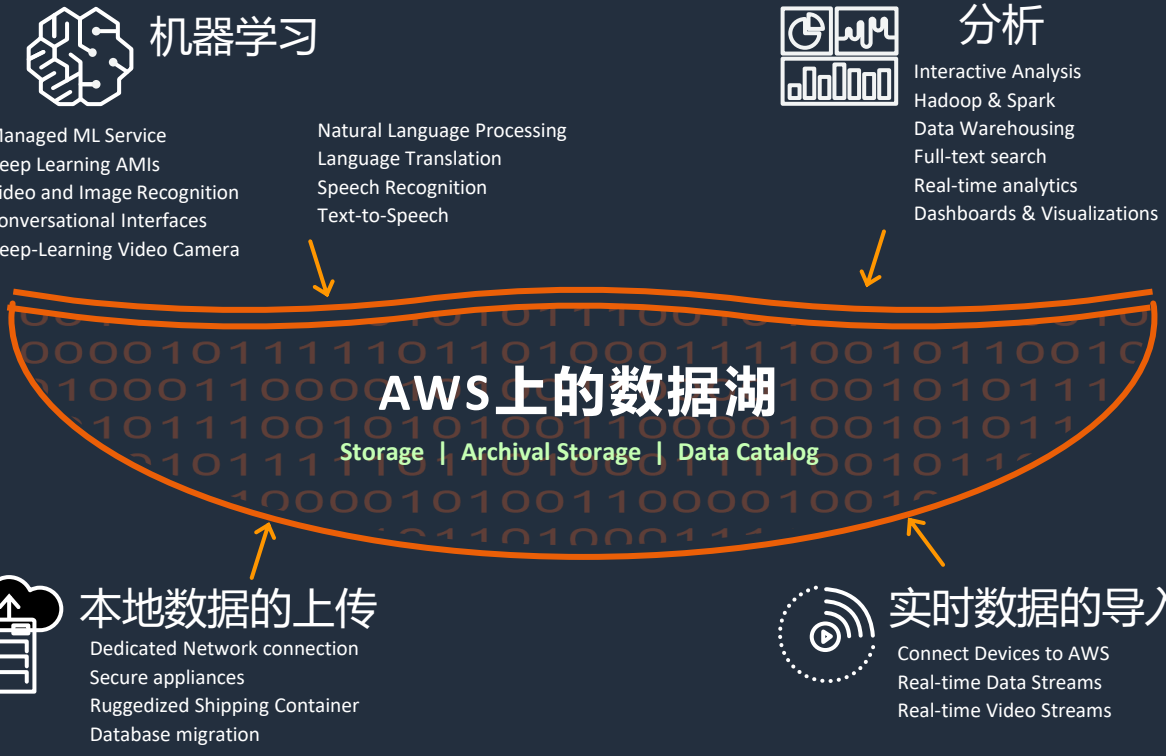
可扩展到EBs级别

众多的分析和机器学习工具

对数据进行加工而无须移动数据

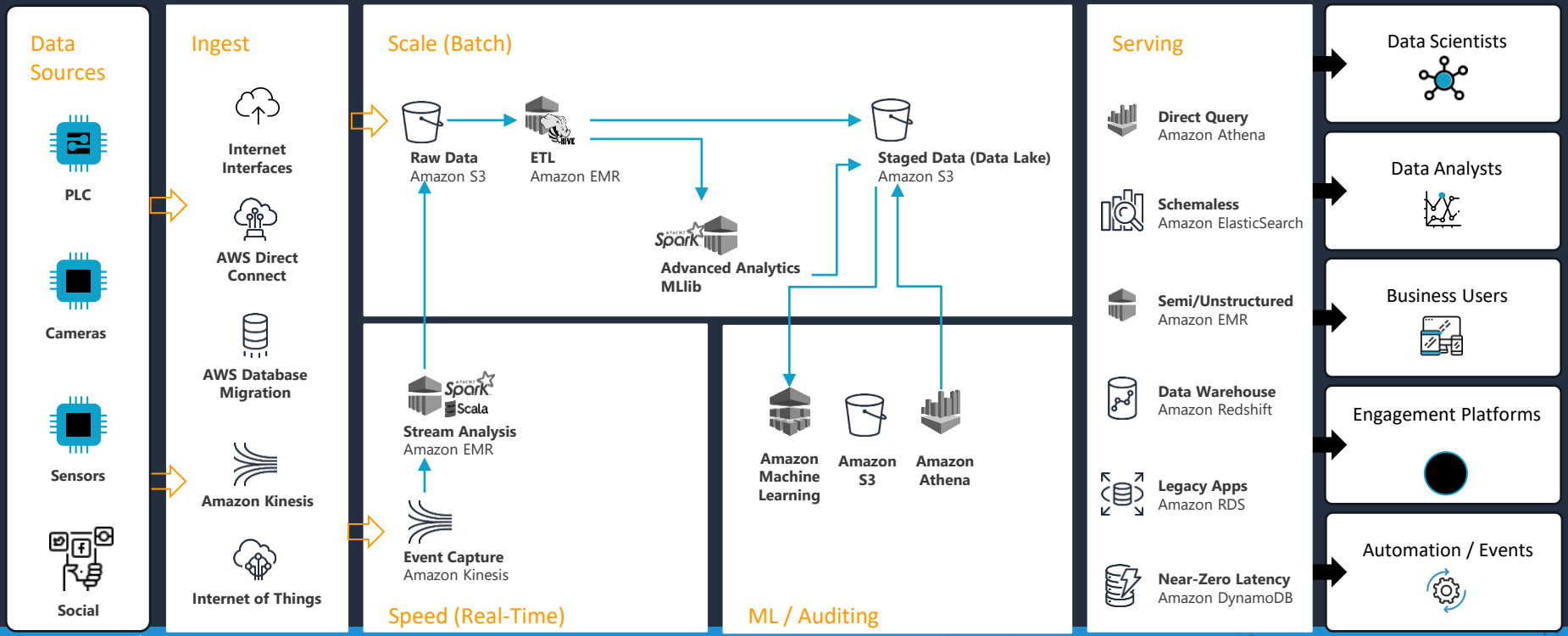
为低成本存储和分析而设计

AWS云上提供丰富的周边服务强化数据湖能力



基于AWS数据湖的现代数据架构

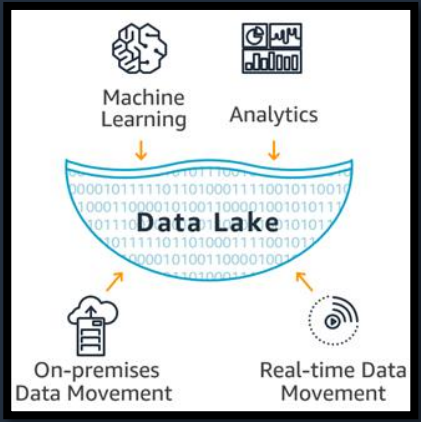
洞见增强业务应用和新的数字化服务




目录与搜索
访问和搜索元数据
DynamoDB
Amazon ES
访问和用户界面
为您的用户提供方便和安全的访问
API Gateway
IAM
Cognito
数据摄入
快速，安全的将数据存入S3
Firehose
Direct Connect
Snowball
DMS
中央储存
在S3中，安全经济高效的储存
S3
处理和分析
使用预测和规则分析来理解数据
Athena
Glue
Quicksight
EMR
Redshift
/ Spectrum
保护和安全
确保数据安全，并验证用户身份
STS
Security Token
Service
Cloudwatch
Cloudtrail
KMS

Amazon S3是AWS数据湖的核心

数据湖是非常适合部署在云中的工作负载，因为云提供高性能、可扩展性、可靠性、可用性、多种分析引擎以及规模经济带来的成本收益



高持久性

高达11个9的数据持久性

高可用性

99.99%数据可用性

高性能

- 并行吞吐
- 范围获取

易于使用

- 标准REST API
- AWS SDKs
- 写后读一致性
- 生命周期管理

无限扩容

- 按需储存，无需预估容量
- 储存与计算分离
- 无需承诺最小使用量

开放扩展

- 最受合作伙伴、供应商和AWS 产品支持
- Talend / Apache Camel
- Apache Nifi /Apahce Sqoop

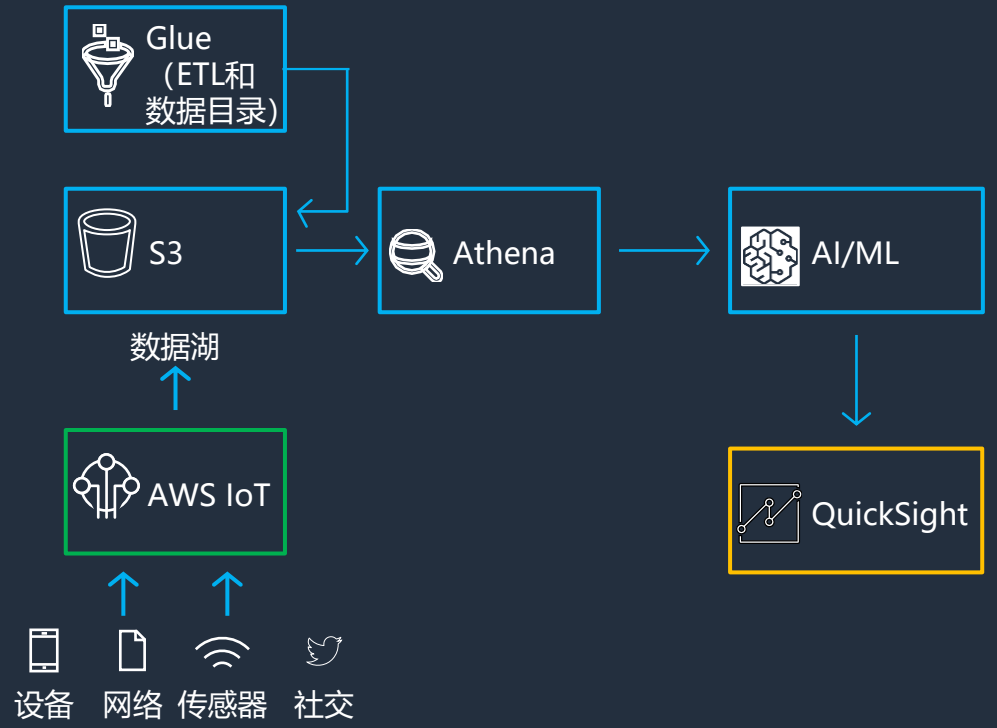
无服务技术使AWS数据湖实现按需响应和付费

无服务器。零基础架构。
零管理

无需为闲置资
源付费

根据使用情况自动
缩放资源

内置的可用性和容
错性



丰富的人工智能服务快速实现数据智能分析

AI 服务

视觉


REKOGNITION IMAGE


REKOGNITION VIDEO


TEXTRACT

语音


POLLY


TRANSCRIBE

语言


TRANSLATE


COMPREHEND &COMPREHEND MEDICAL

聊天机器人


LEX

预测


FORECAST

推荐


PERSONALIZE

ML 服务


AMAZON SAGEMAKER

构建

预先构建算法和笔记本

数据标记（GROUND TRUTH）

算法和模型（AWS MARKETPLACE，适用于机器学习）

训练

一键式模型训练和调整

优化（NEO）

强化学习

部署

一键式部署和托管

ML 框架和基础架构

框架

TensorFlow

mxnet

PYTORCH

接口

GLUON

Keras

基础架构


EC2 P3 & P3dn


EC2 G4


EC2 C5


FPGAs


GREENGRASS

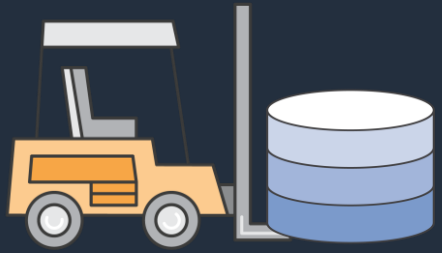

ELASTIC INFERENCE

数据湖的优势 – 所有数据在一个地方

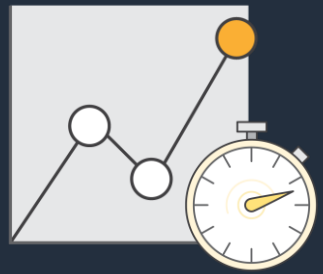
“我的数据储存在多个不同的地方，
那一份数据才是真实可信的呢？”

在一个集中的位置，
储存并分析来自所有来源的数据

数据湖的优势 – 快速提取

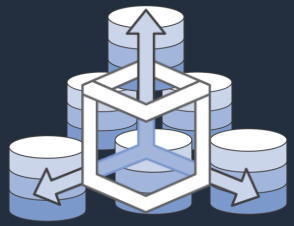


“如何快速从各种来源收集数据
并有效存储？”



快速提取数据，
而无需将其强制转换到范式中。

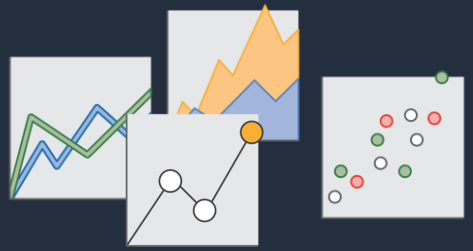
数据湖的优势 – 储存与计算分离



“如何扩展容量，
以应付持续增长的数据？”

将存储和计算分开，
可以根据需要缩放每个组件。

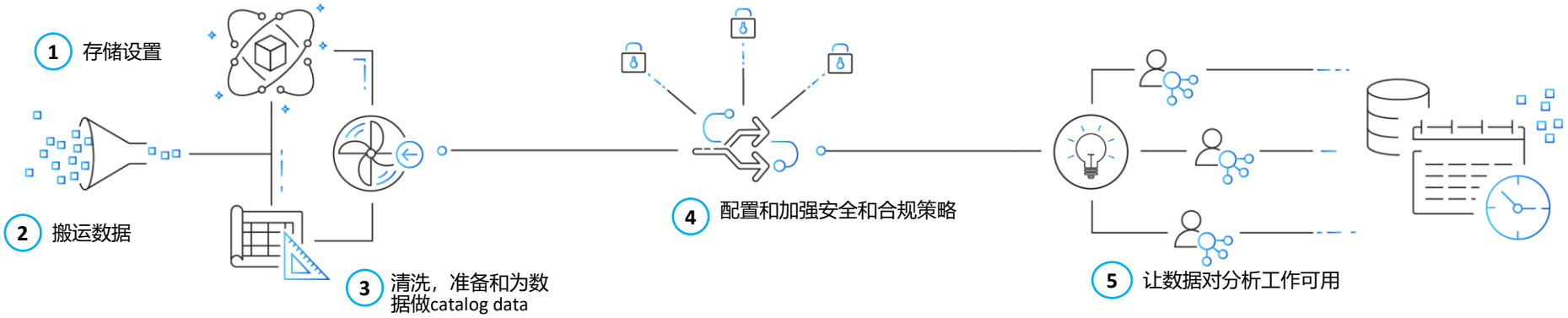
数据湖的优势 – 读取时范式化



“有没有办法将多个分析和处理框架应用于相同的数据？”

数据湖可以通过在读取时范式化来进行即时分析，而不是在写入时。

典型的构建数据湖的步骤



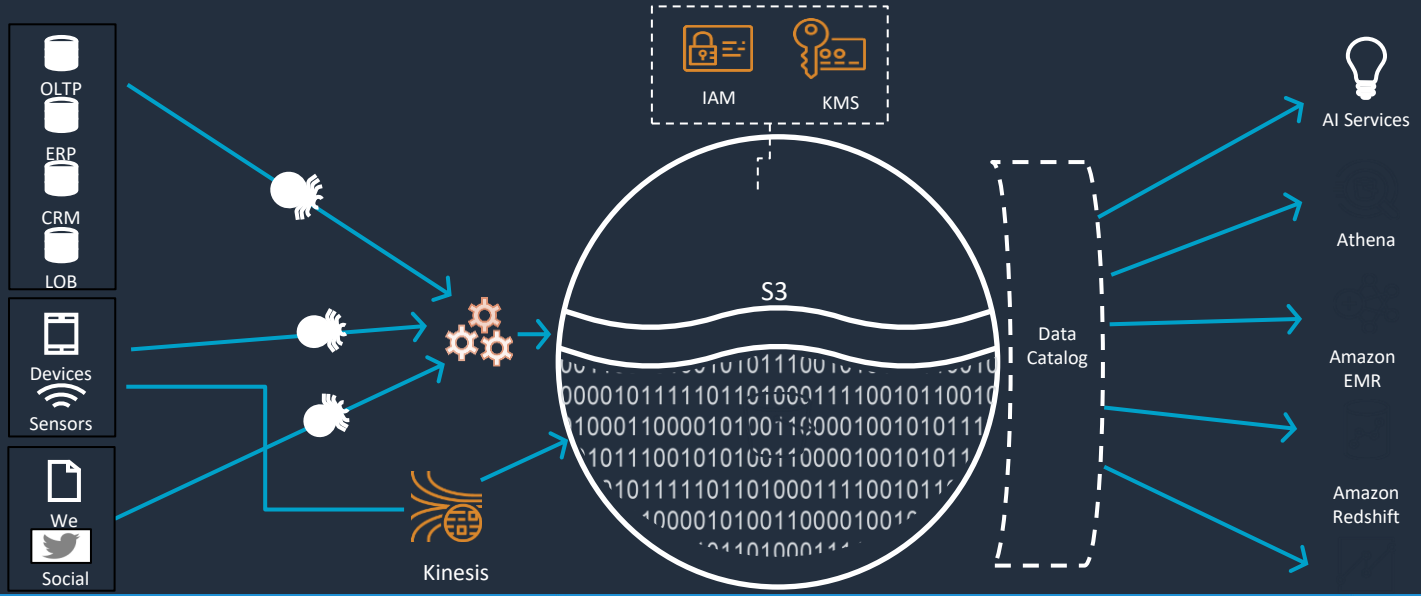
AWS Lake Formation - 在数日内构建安全的数据湖

Data Lakes and analytics on AWS

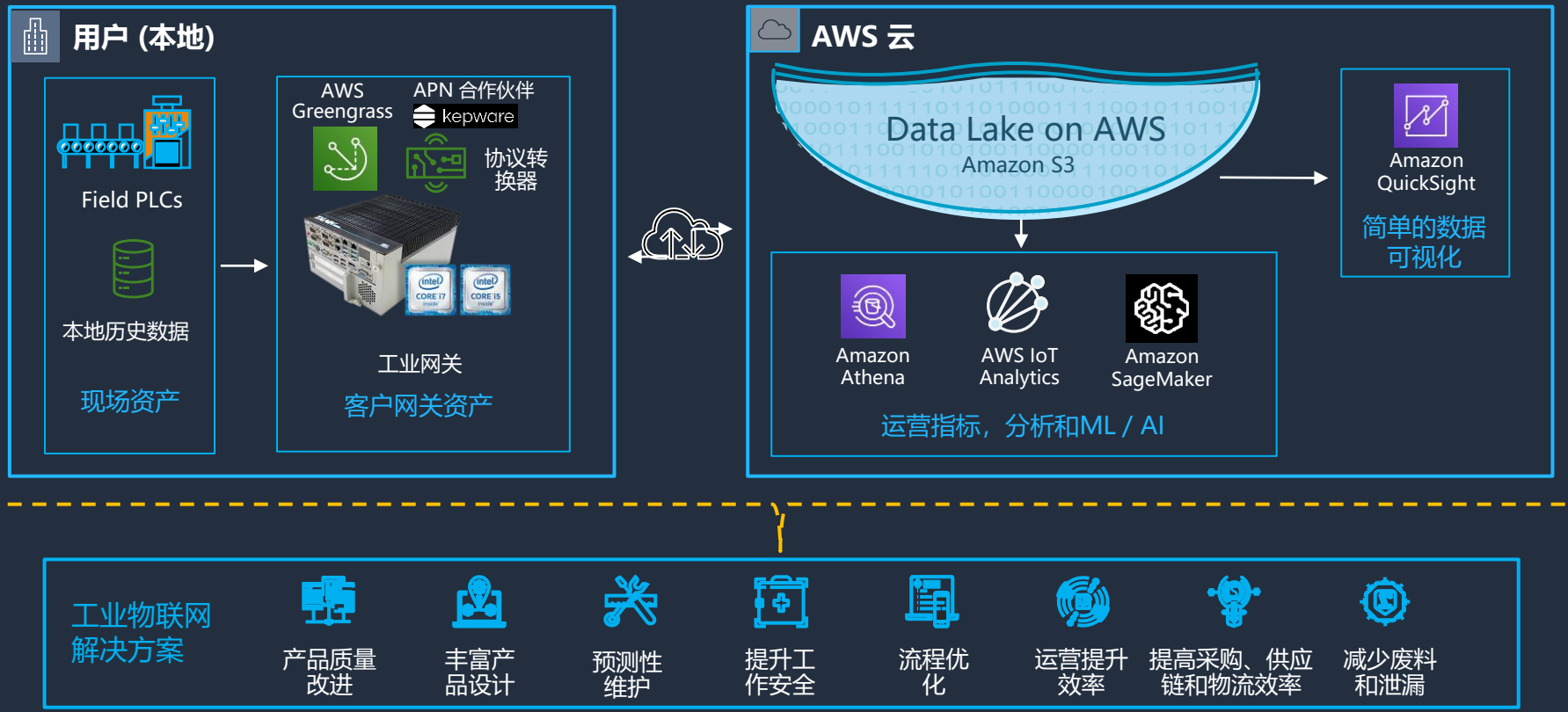
快速构建数据湖
- Identify, crawl, and catalog sources
- Ingest and clean data
- Transform into optimal formats

简化安全管理
- Enforce encryption
- Define access policies
- Implement audit login

轻松安全地自助访问数据
- Analysts discover all data available for analysis from a single data catalog
- Use multiple analytics tools over the same data



数据湖行业应用举例 - 工业数据湖



数据湖行业应用举例 - 零售数据湖

成本 / 价格
库存
需求
合约条款
供应链追踪

外部广告
成功案例
数字营销

数据湖

客户族群 Customer Cluster → 打包营销
预测客户购买行为 → 引导促销
搜索结果分析 → SEM/SEO 优化
销量分析 → 采购优化
库存分析 → 价格优化

数字营销

自动化采购

数据湖行业应用举例 - 游戏数据湖

游戏信息：道具、等级、杀敌
系统信息：开始、崩溃、帧率

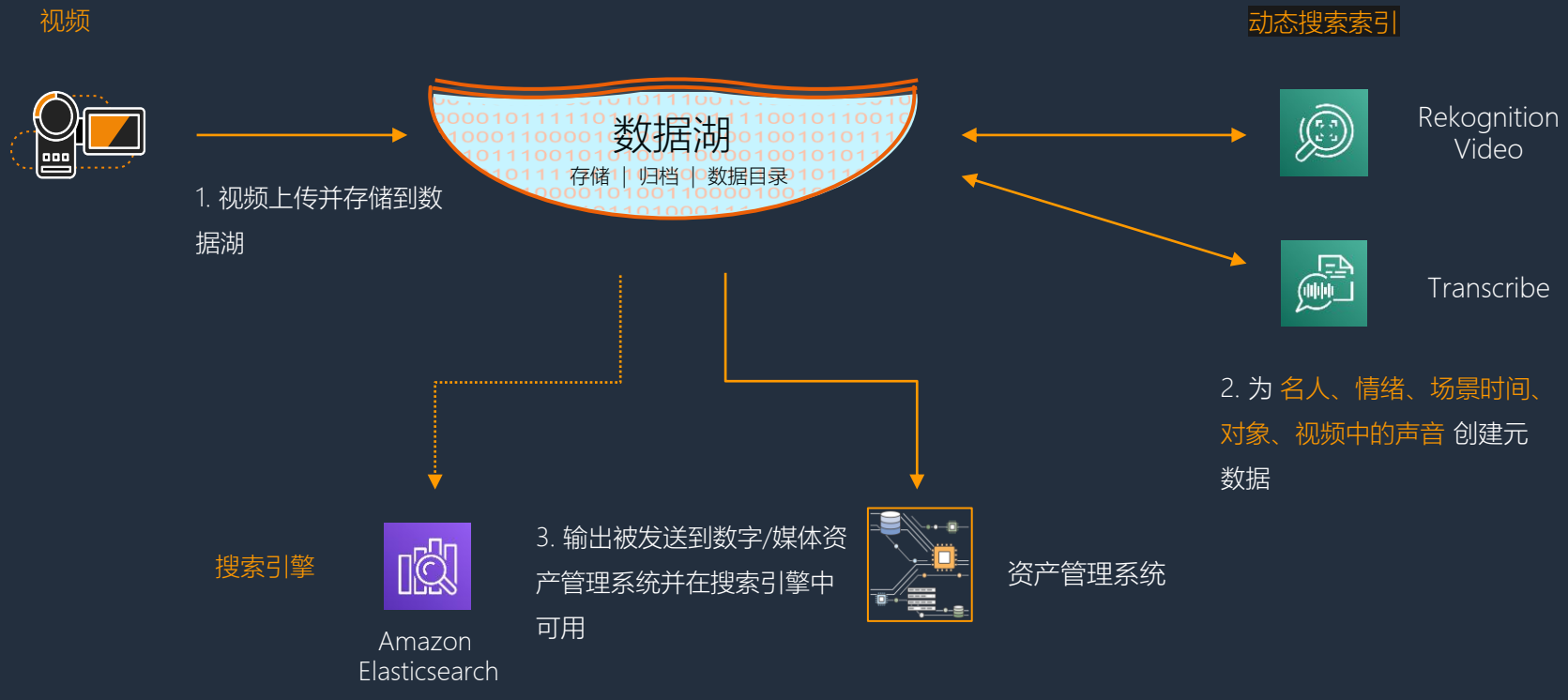
每分钟摄入数千万条事件数据

每月新增数个 PB 的数据存储量

能够执行提取、转换和加载 (ETL)

从海量数据中进行查询和分析

数据湖行业应用举例 -媒体数据湖



AWS数据湖在小红书的使用

小红书
找到国外的好东西

小红书是新一代社区电商，它将海外购物分享社区与跨境电商相结合，精准捕捉85后和90后的消费升级需求

小红书快速发展遇到的挑战

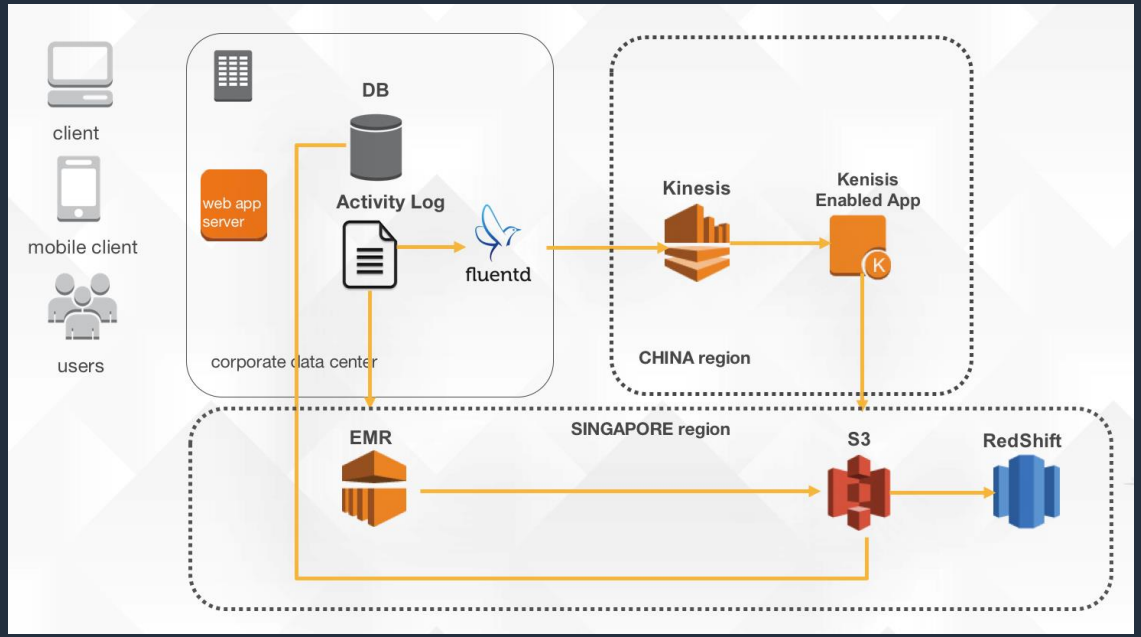
业务诉求

小红书的商业模式成功的关键是能够通过后台数据快速即时地了解到用户喜欢什么、在分享什么、点赞最多的是哪些，并通过对这些数据的分析，推测出哪些商品可能是爆款。

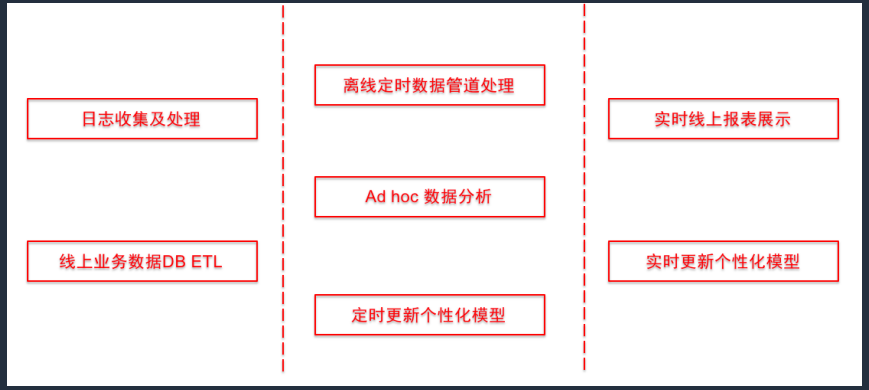
技术挑战

1. 用户从千万级到2亿多，数据从G级别到几百G，到T级别，再到现在的几十PB级别的数据；
2. 从原来只分析关系型数据，到现在还要分析各种非关系型数据；
3. 数据科学家，AI人员，BI分析业务人员的快速增长，各自分析需求不同；
4. 用户数量激增，分析的数据量也激增，如何降低DAU的成本；
5. 降低运维成本

小红书基于AWS数据湖的大数据分析平台



小红书基于AWS数据湖的数据分析成果



“借助Amazon EMR、Amazon Kinesis、Amazon RedShift云服务，我们成功地以几个人的小团队，在短时间内搭建起完整的数据处理系统，实现了高效的大数据分析。”

- 社区用户画像
- 电商销售追踪
- 拉新渠道效果
- 优惠券效果分析
- 推荐模型迭代
- ……
- 每周处理超过50个新需求

- 分析购物流程顺畅程度及网站产品分布合理与否
- 页面流量排名及场景转化率分析
- 站内搜索分析
- 客户为何离开页面分析
- DAU平均成本

- 每天更新超过200张报表

- 社区运营报表
- 电商运营报表
- 用户增长报表
- 财务报表
- 广告运营报表
- ……

推送
广告
薯券发放
AB测试
财务对帐单
个性化搜索和推荐
反欺诈
……

DAMS
中国数据智能管理峰会
DATA & AI MANAGEMENT SUMMIT

Q&A

DAMS

中国数据智能管理峰会

DATA & AI MANAGEMENT SUMMIT

THANK YOU!