中国站 LIVE

全面解读 AWS 数据湖及数据分析的新服务和新趋势

张呈刚

AWS 高级解决方案架构师

内容概览

将数据转变为见解


Amazon Redshift
数据仓库


Amazon EMR
Hadoop + Spark


Amazon Athena
交互式分析


Amazon Kinesis
实时数据分析


Amazon Elasticsearch Service
运维分析

数据湖基础架构的现代化革新


Amazon S3/Glacier


AWS Lake Formation


AWS Glue

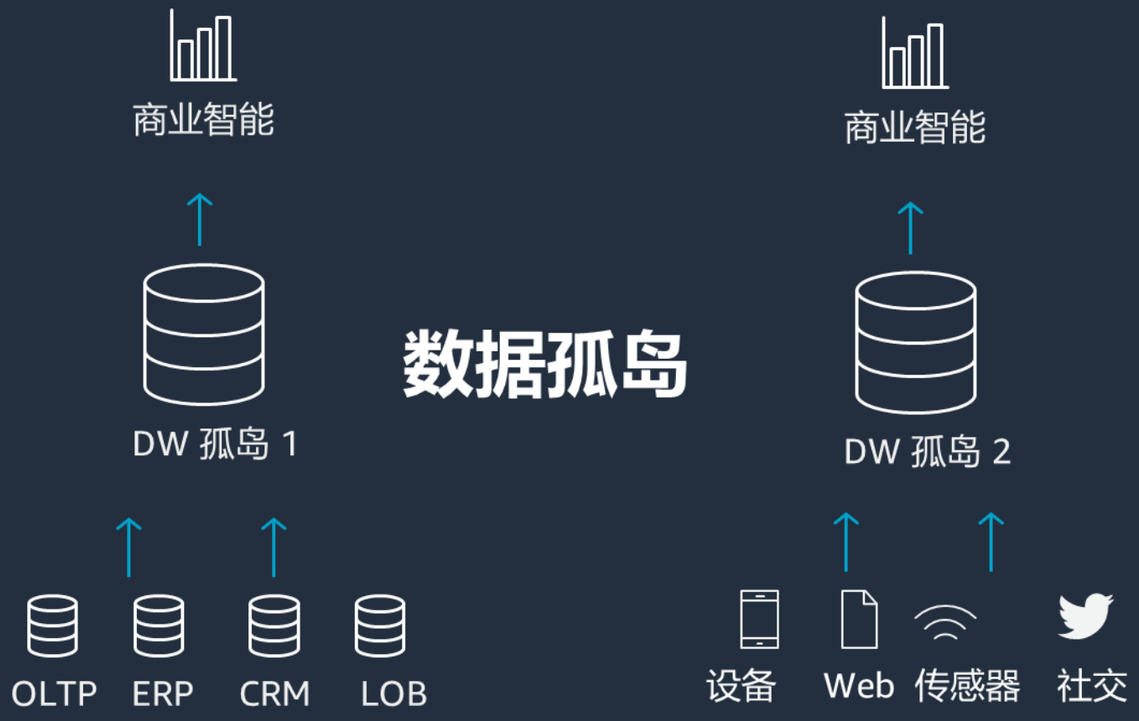
客户面临的全新境况

数据爆炸

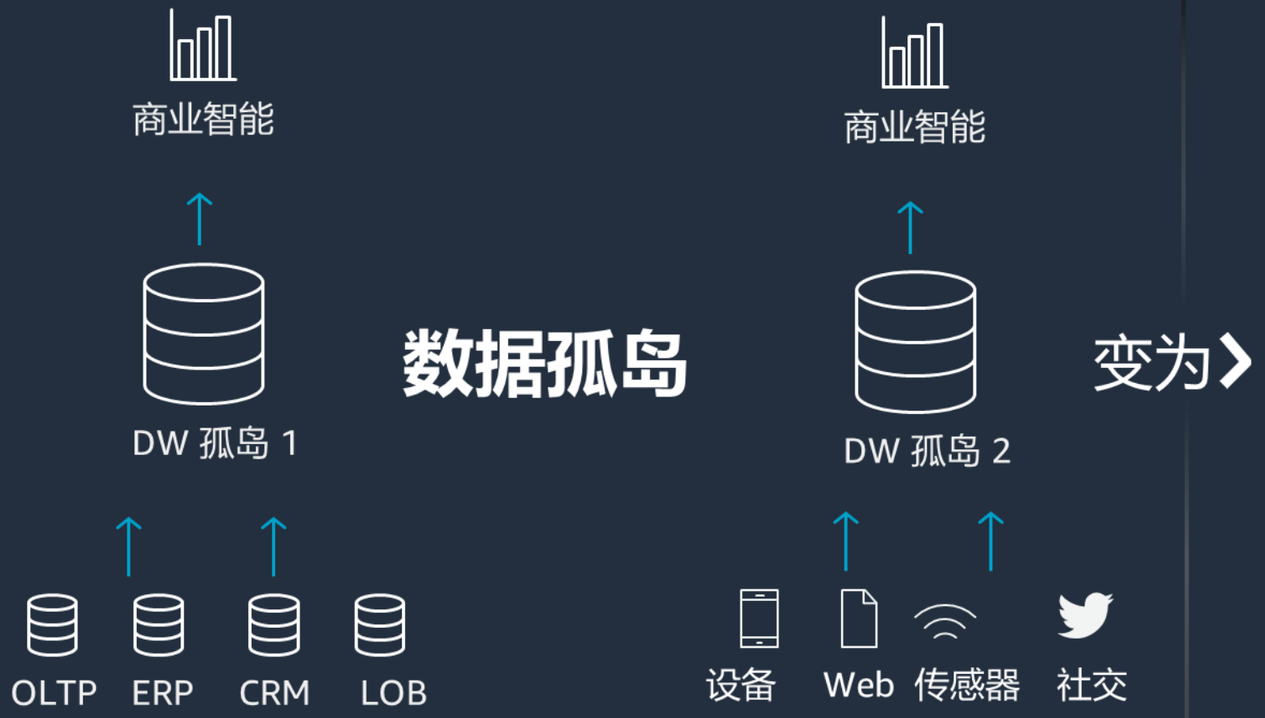
角色爆炸

通过实时数据更快速做出决策的全新需求

传统数据孤岛的聚合



传统数据孤岛的聚合


商业智能
DW 孤岛 1
OLTP
ERP
CRM
LOB
数据孤岛
商业智能
DW 孤岛 2
设备
Web
传感器
社交
变为


机器学习
数据湖
开放的格式
中央目录
数据仓库
BI + 分析

云数据湖赋能的未来

客户需要：

可扩展且具备成本效益的单一数据湖

使用自由选择且基于标准的数据格式

以丰富多样的方式分析自己的数据

Amazon S3 是构建云数据湖最流行的选择

无与伦比的**持久性**、**可用性**和**扩展性**

最易于使用且针对成本优化：智能分层

最丰富的数据摄取方式

最强大的**对象级别**控制

支持最广泛的分析工具

最佳**安全性**、**合规性**和**审计**功能

我们的产品组合

最全面、深入的产品组合，专为开发者量身打造

借助 AWS Lake Formation 在数天内构建数据湖
以最高速度从数据获得见解

更快速地移动、存储、分类、清洗数据

借助机器学习技术更快速地移动、存储、分类、清洗数据

跨越多种服务强制实施安全策略

跨越多种服务强制实施安全策略

获得并管理新见解

予力分析师和数据科学家获得并管理新见解

ETL 和数据目录：AWS Glue

简单、灵活、具备成本效益的 ETL

更省事

跨越不同 AWS 服务实现集成，支持：Amazon Aurora、Amazon RDS、Amazon Redshift、Amazon S3，以及 Amazon EC2 中 VPC 运行的常用数据库引擎

Serverless

Serverless：无需预配置或管理基础架构

更强大

自动生成代码，用于执行您的数据转换和加载过程

流式数据处理能力

新增！

借助 Amazon Kinesis Data Analytics 访问位于 Amazon VPC 中的资源
在 VPC 中从数据源读写数据，例如 Amazon Elasticsearch Service 集群、RDS 数据库以及 Redshift 数据仓库

Amazon MSK 发布支持 Prometheus 的 Open Monitoring
- 以极低延迟消费每一条 Apache Kafka 指标
- 通过 Prometheus 实现时间序列日志记录、警报和图表

在 AWS 上使用完全托管的服务同时运行 Apache Flink 和 Apache Kafka
- 使用 Kinesis Data Analytics 处理存储在 Amazon MSK 的流式数据
- 借助完全托管的服务，使用开源软件端到端运行流处理解决方案

数据交换：AWS Data Exchange (GA)

新增！

在云中轻松查找并订阅第三方数据

在一个位置快速找到多样化数据

>1,000 数据产品

>80 个数据提供商，包括 Dow Jones、Change Healthcare、Foursquare、Dun & Bradstreet、Thomson Reuters、Pitney Bowes、Lexis Nexis 以及 Deloitte

轻松分析数据

将数据下载或复制到 Amazon S3

使用现有数据进行组合、分析、建模

通过 EMR、Redshift、Athena 和 Glue 分析数据

高效访问第三方数据

简化数据访问方式，无需接收物理介质，无需管理 FTP 凭据，也无需集成不同 API

法务审批和协商工作量最小化

数据仓库：Amazon Redshift

最受欢迎和速度最快的云数据仓库

数据湖和 AWS 集成

跨越数据仓库、数据湖和运维数据库，分析 EB 量级数据

跨越多种分析服务查询数据

最佳性能，最具扩展性

AUQA 和 RA3：比其他云数据仓库快 10 倍

按需添加无限量计算容量，满足不受限制的并发访问需求

最安全且合规

AWS 级安全性（如 VPC、AWS KMS 加密、AWS CloudTrail）

获得所有主要认证，如 SOC、PCI、DSS、ISO、FedRAMP、HIPAA

最低成本

成本优化的工作负载，单独为计算和存储资源付费

成本仅为传统数据仓库的 1/10，$1000/TB/年

相比其他云数据仓库，价格最多可便宜 75%，成本可预测

Amazon Redshift 不断进行着快速创新

健壮的结果集缓存

支持更大数量的表 ~20000

针对 ORC、Parquet 提供 Copy 命令支持

IAM 角色链

弹性调整大小

组

Amazon Redshift Spectrum：日期格式、支持 scalar json 和 ION 文件格式、区域扩展、谓词过滤

自动分析

配合 Amazon CloudWatch 进行健康度和性能监视

自动确定表分配方式

CloudWatch 支持 WLM 队列

性能改进——哈希联接、vacuum、窗口函数、resize ops、聚合、控制台、union all、高效编译代码缓存

卸载至 CSV

自动 WLM

支持约 25 种查询监视规则（QMR）

AQUA

并发扩展

200+

过去 18 个月发布的新功能总数

DC1 迁移至 DC2

ROLLBACK 弹性处理

在 AWS 控制台中管理多方查询

自动分析表的增量变更

Spectrum 请求加速器

应用新的分布键

Amazon Redshift Spectrum：Parquet 和 ORC中的行组筛选，支持嵌套数据，增强的 VPC 路由，多分区

更快速的 Classic resize 及优化的数据传输协议

性能：联接中的 Bloom 筛选器，创建内部表的复杂查询，通信层

Amazon Redshift Spectrum：并发扩展

自动 WLM 及查询优先级

快照计划程序

性能：联接下推至子查询，混合工作负载临时表，rank 函数，联接中 null 的处理，单行插入

Advisor 针对分布键提供建议

Amazon Lake Formation 集成

Auto-Vacuum 排序，Auto-Analyze 和 Auto Table 排序

存储的过程

空间处理

列级访问控制及 AWS Lake Formation

RA3

区域间快照传输性能改进

联合查询

AZ64 压缩编码

重新设计的控制台

具体化视图

手工暂停和恢复

Amazon Redshift 联合查询（预览）

跨越数据仓库、数据湖和操作型数据库分析数据

新增！

通过 Amazon Redshift 跨越多种系统进行查询

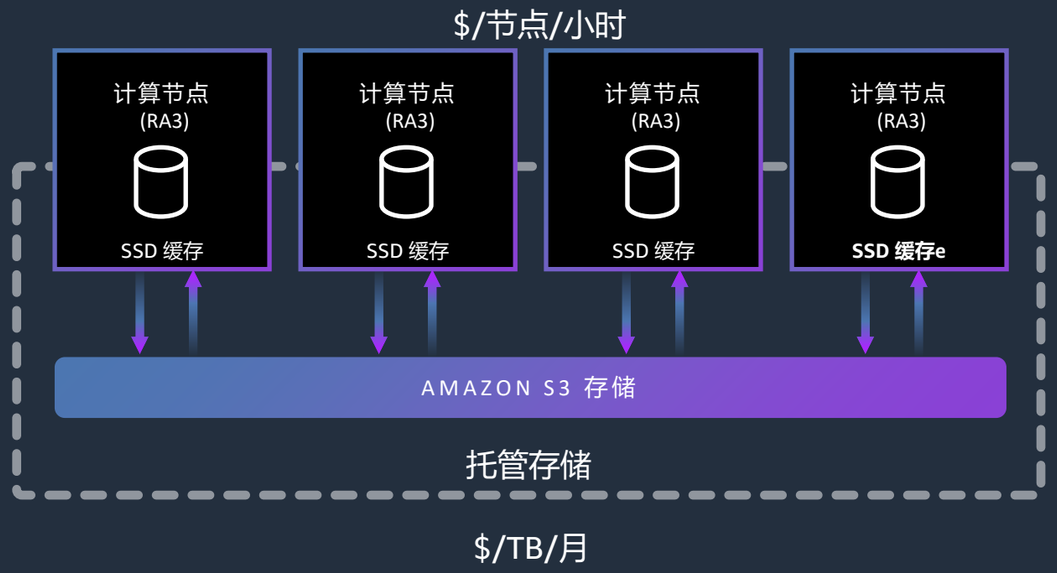
结合数据仓库与事务数据

兼容 Amazon RDS 和 Amazon Aurora (PostgreSQL)

新增！

使用 RA3 实例的 Amazon Redshift (GA)

通过单独为计算和存储容量付费优化你的数据仓库



可实现比现有云数据仓库高 3 倍的性能

DS2 客户可在迁移后以相同成本获得 2 倍性能提升和 2 倍存储容量

自动扩展数据仓库的存储容量

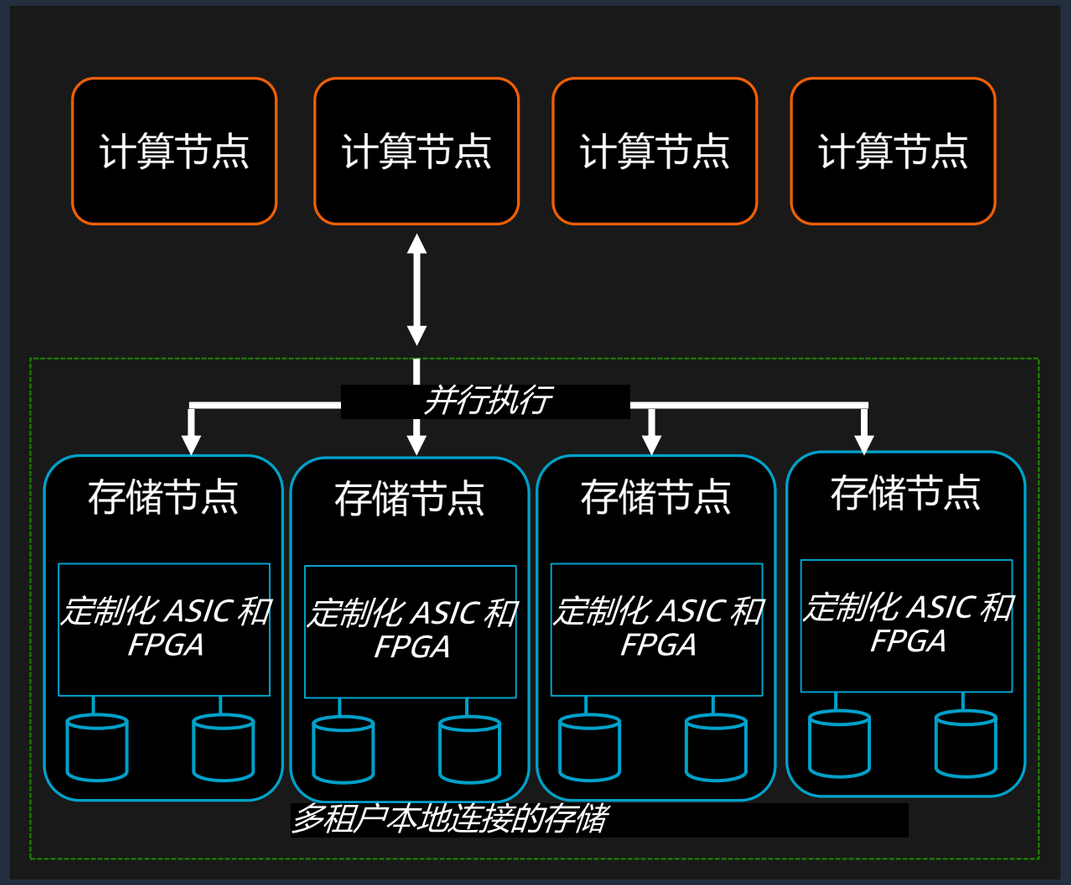
最大支持 8PB 工作负载（经压缩）

AQUA – Advanced Query Accelerator

不增加成本的前提下，让 Redshift 运行速度比其他云数据仓库**快** 10 **倍**

新增！

2020 年发布



AQUA 将计算能力带到了存储层，避免来回搬动数据

Amazon S3 基础上的高速缓存通过横向扩展可并行处理多节点数据

AWS 自行定制设计的分析处理器可加速数据压缩、加密和处理工作

100% 兼容当前版本的 Redshift

Apache Hudi for Amazon EMR (GA)
在 Amazon S3 中实现记录级别的插入、更新和删除

新增！

Apache Hudi 本身开源，并使用了开放的数据格式，借此数据湖将能够：

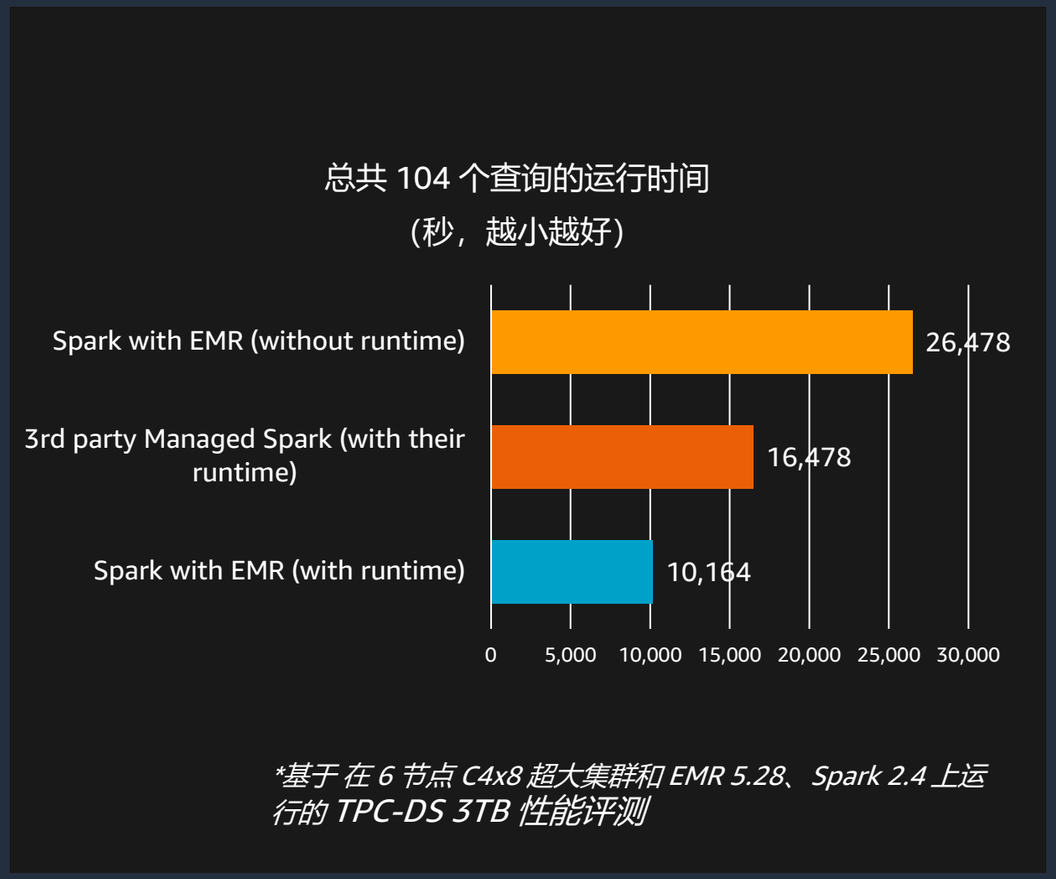
- 遵守数据隐私法案
- 使用实时数据流并更改数据捕获
- 承受延期抵达的数据
- 追踪变更历史并回滚

并包含了对 Spark、Hive 和 Presto 的支持

Spark for Amazon EMR 的性能改进

Apache Spark 针对性能优化的运行时，以 1/10 的成本实现 2.6 倍提速

新增！



基于在 6 节点 C4x8 超大集群和 EMR 5.28、Spark 2.4 上运行的 TPC-DS 3TB 性能评测

针对 Apache Spark 进行性能优化的运行时

最佳性能

- 相比不使用运行时的 Spark with Amazon EMR，**提速 2.6 倍**
- 相比第三方托管的 Spark（使用对方的运行时），**提速 1.6 倍**

最低价格

- 成本仅为第三方托管 Spark（使用对方的运行时）的 **1/10**

100% 兼容 Apache Spark API

Amazon Elasticsearch Service (Amazon ES)

完全托管、可扩展、安全

开源的 Amazon ES API、Kibana 和 Logstash

开源的 Amazon ES API

托管的 Kibana

与 Logstash 集成

完全托管

在几分钟内部署 Amazon ES 集群：简化硬件预配置，软件安装/补丁，故障恢复，备份和监视

可扩展、安全、合规

通过一个 API 调用或几次鼠标点击，扩大或缩小集群规模

通过 VPC 实现安全的网络隔离，加密存储后和传输中的数据

合规：HIPAA、PCI DSS 和 ISO

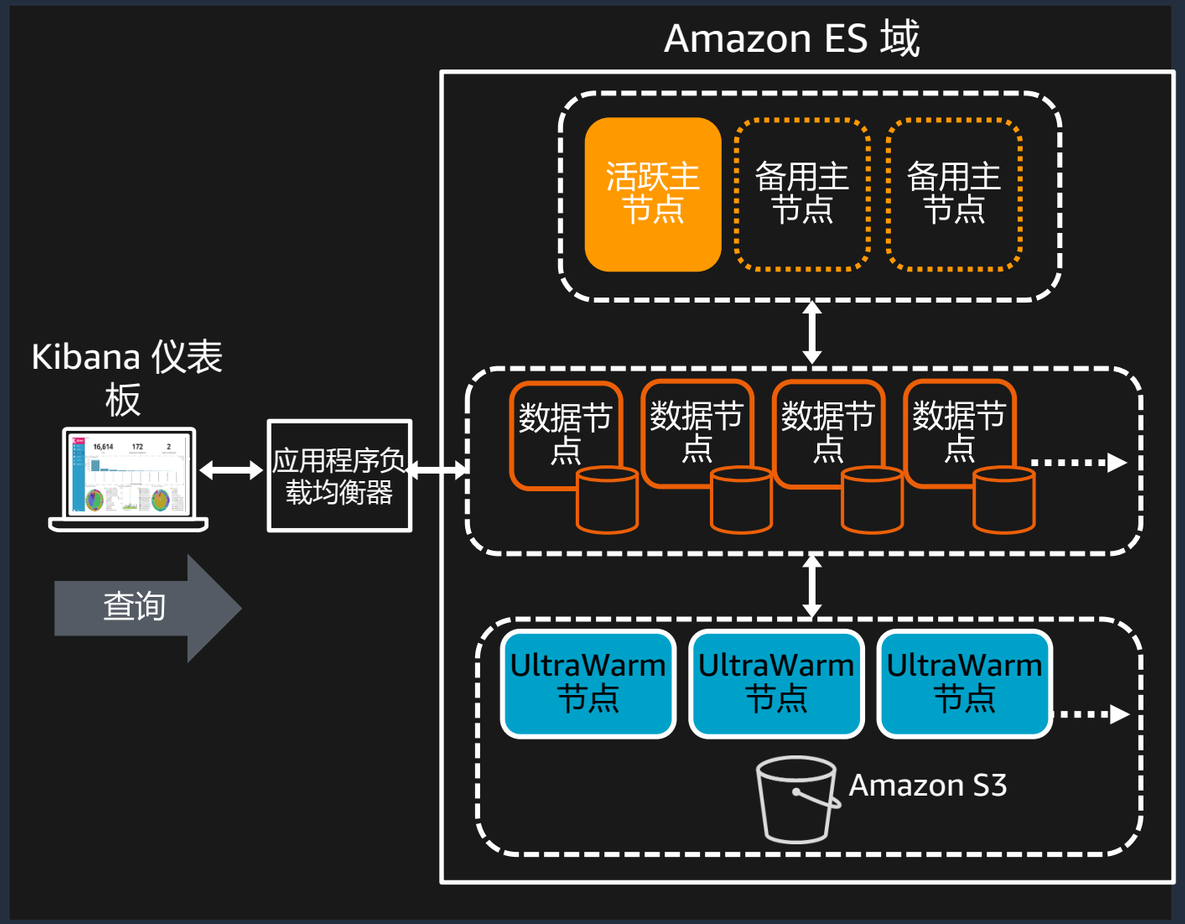
仅为实际用量付费

无前期费用或用量要求

自带关键功能：加密、VPC 支持、24x7 监视

UltraWarm for Amazon ES（预览）
面向 Amazon S3 的全新温存储层

新增！



成本可降低 90%

每个域通过扩展最大可支持 3 PB

分析多年积累的运维数据

对 Amazon ES 进行扩展

Amazon Athena

针对 Amazon S3 中的数据运行 SQL 查询

无需管理基础架构

按照查询数量付费

即时查询

无设置成本

指向 Amazon S3 即可开始查询

按查询付费

只为实际运行的查询付费

通过压缩，可将每个查询的成本节约 30–90%

开放

ANSI SQL

JDBC/ODBC 驱动程序

多种格式、压缩类型、复杂联接和数据类型

易用

Serverless：无需基础架构，无需管理

与 QuickSight 集成

Amazon Athena 联合查询（预览）
针对横跨多个数据存储的数据运行 SQL 查询

新增！

在云端或本地，针对关系型、非关系型、对象，或自定义数据源运行 SQL 查询

针对常用数据源提供开源连接器

可为自定义数据源构建连接器

连接器可在 AWS Lambda 中运行：无需管理服务器

Amazon QuickSight
首款根据每个会话付费且提供机器学习见解的 BI 服务

Serverless，云平台驱动的 BI 服务（无需管理服务器）

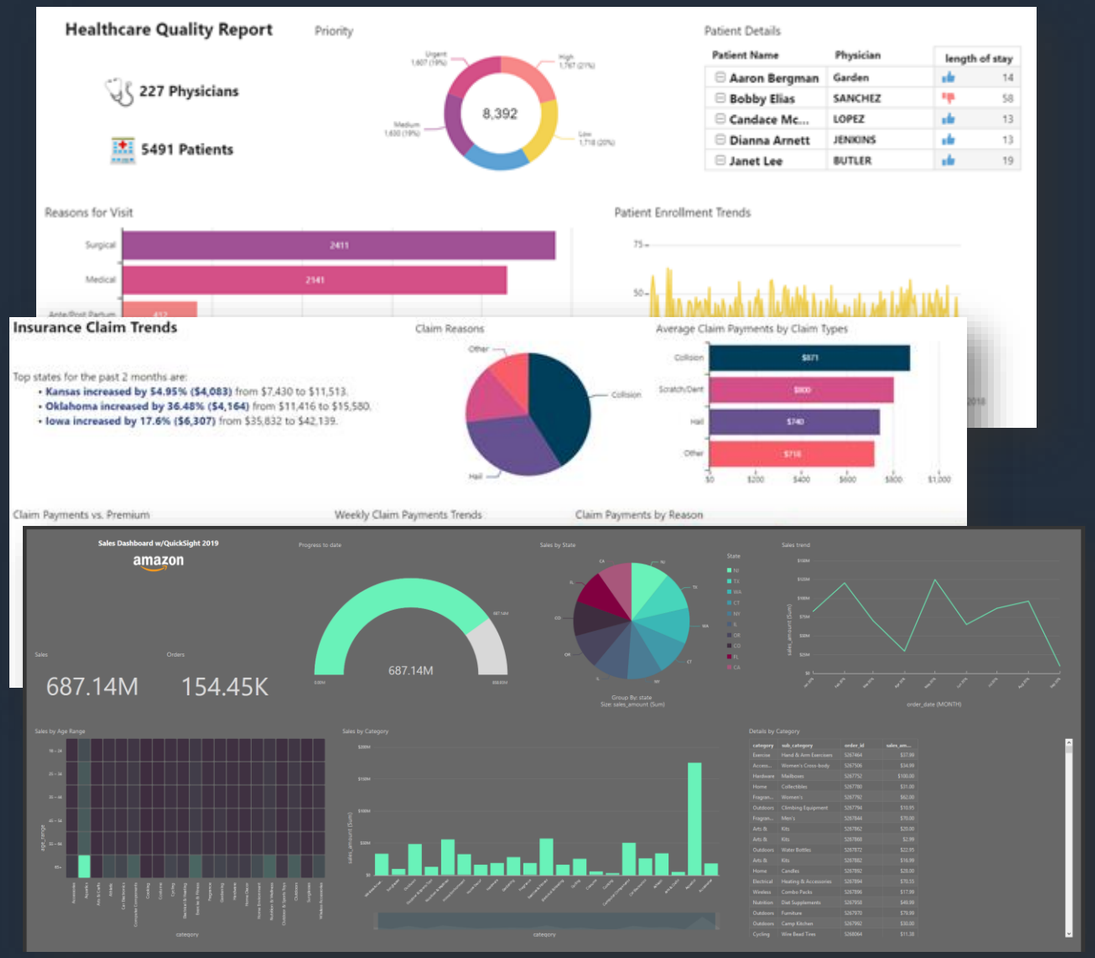
从十多位用户轻松扩展至数十万用户

只为实际使用的资源付费
- 读者：$0.30/30分钟会话，$5/用户/月封顶
- 作者：$18/月/作者

轻松集成于 Amazon S3、Amazon Athena、Amazon Redshift、Amazon RDS、Amazon Aurora 和 Amazon EMR

QuickSight 的 API 和定制能力

新增！



通过 API 部署并管理仪表板/数据

将应用程序 UI 与 QuickSight 主题保持一致

无需管理服务器，即可将仪表板嵌入应用

- 快速、始终如一的性能
- 按照会话付费

从数十人自动扩展至数千人

- 无需管理服务器
- 无需编写脚本

为何选择借助 AWS 从数据中获得见解?
利用所有数据，为所有用户最快速提供答案!

轻松构建大规模数据湖

- AWS Lake Formation
- Redshift 数据湖导出
- Redshift 联合查询
- Amazon Athena 联合查询
- Data Streaming for AWS Glue

最低成本提供最佳性能

- AQUA for Redshift
- RA3 for Redshift
- Redshift 具体化视图
- 适用于 Amazon S3 的 UltraWarm 存储层
- 针对 Spark in Amazon EMR 的性能改进

最全面且开放

- AWS Data Exchange
- Amazon EMR on AWS Outposts
- 适用于 Amazon EMR 的记录级插入/更新
- Amazon Athena 机器学习
- Amazon QuickSight 机器学习

最安全

- Amazon Westeros
- Amazon Macie
- AWS Lake Formation

Thank you!