
aws
INNOVATE
ONLINE CONFERENCE

aws INNOVATE
ONLINE CONFERENCE

分会场六：大数据分析

在 AWS 上构建数据湖及 AWS Lake Formation 介绍

袁春华，AWS 解决方案架构师

议程

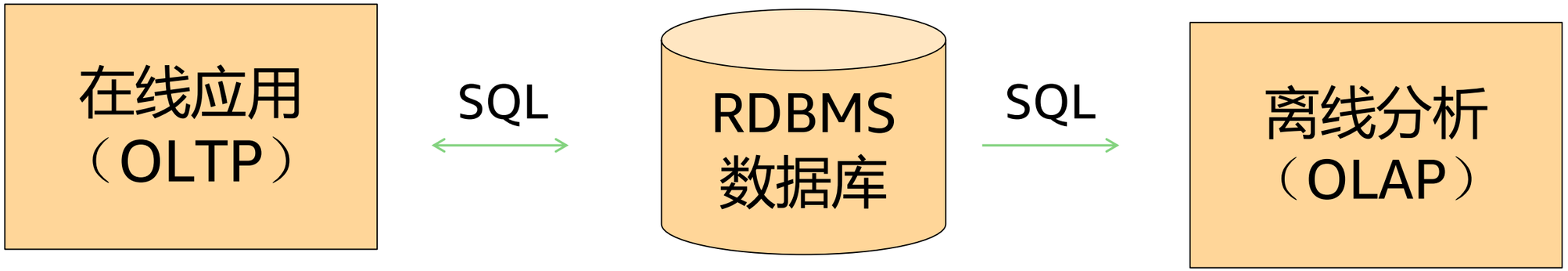
企业数据分析平台的演变

为什么需要构建企业数据湖

如何在 AWS 云上构建企业数据湖

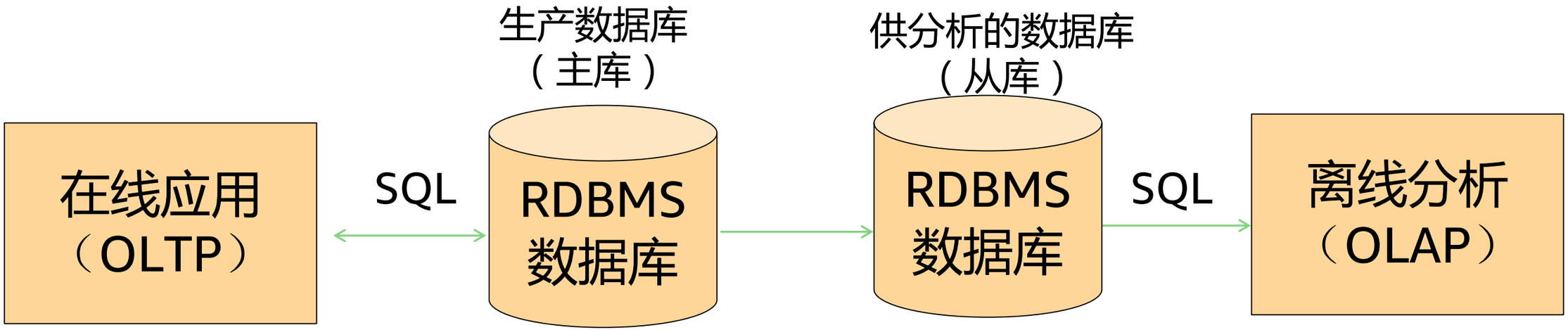
AWS Lake Formation 介绍

企业数据分析平台的演变



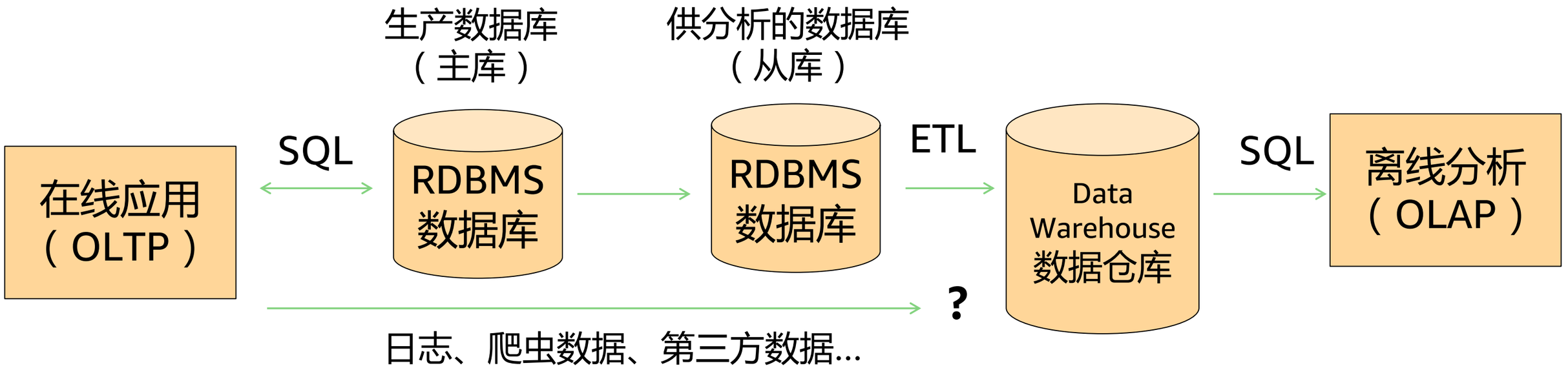
问题：离线分析 IO 大，影响生产业务

企业数据分析平台的演变



问题：RDBMS 为在线平台设计，不适合分析

企业数据分析平台的演变



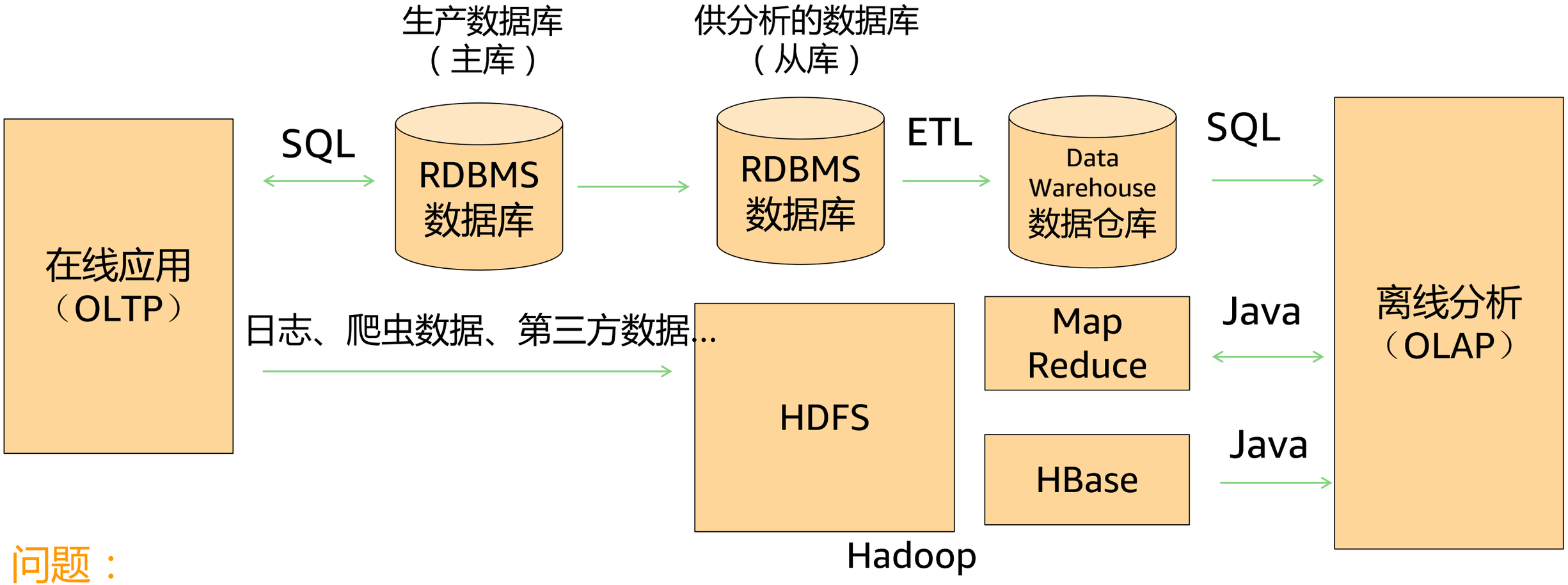
问题：

1）数据仓库的容量限制、性能瓶颈、成本

2）如何存储非结构化数据

3）“ Schema-on-Write”，不能直接存储非预期数据

企业数据分析平台的演变

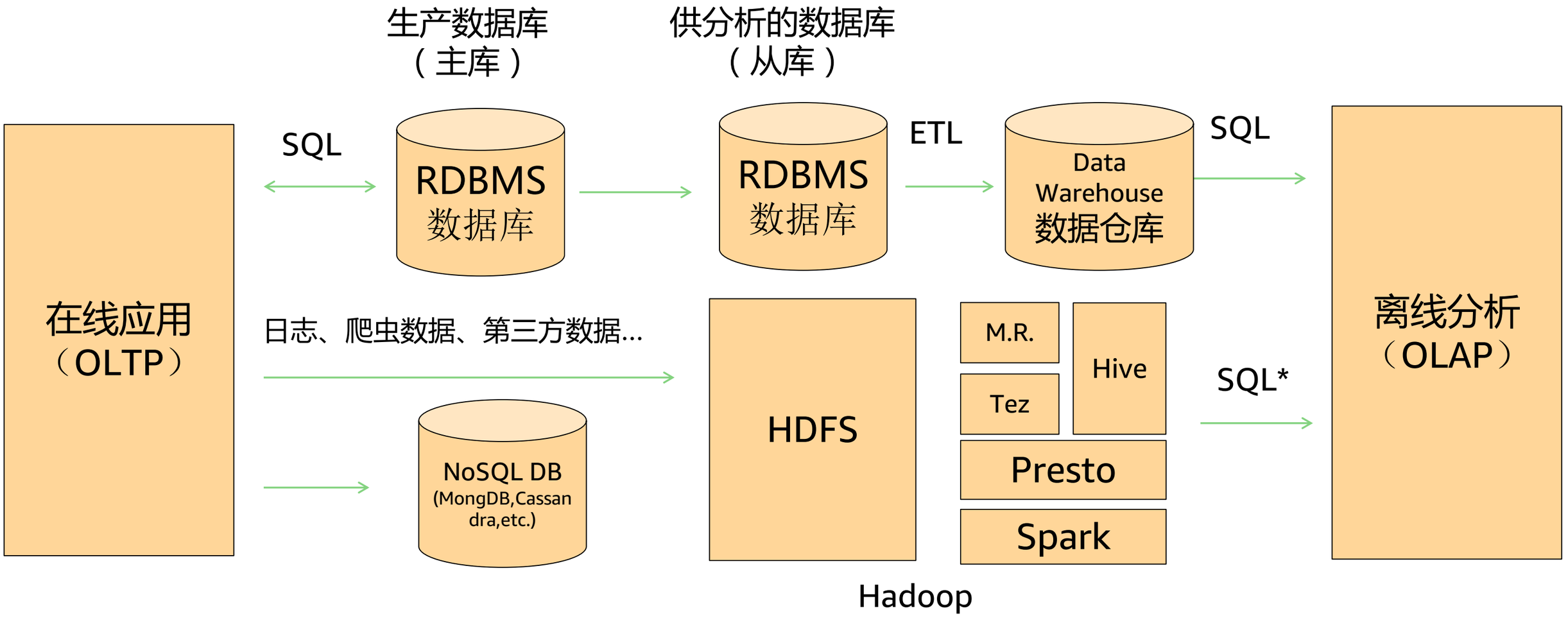


问题：

1）数据分析人员需要学习新的语言

2）Hadoop 需要新的数据工程师 (DBA、数据工程、数据科学三个角色就此逐渐分立)

企业数据分析平台的演变



议程

企业数据分析平台的演变

为什么需要构建企业数据湖

如何在 AWS 云上构建企业数据湖

AWS Lake Formation 介绍

数据资产时代

> 世界上最有价值的资源不再是石油，而是数据。

*Copyright: The Economist, 2017, David Parkins

实现数据变成资产的通用原则

停止丢弃数据

Stop throwing data away

让更多用户使用

Make it available to more users

多种数据处理技术

Arm users with more data processing technologies

数据规模呈爆发式增长

数据增长	保存时间	规模
>10 倍 平均每5年	**15** 年	**1,000** 倍

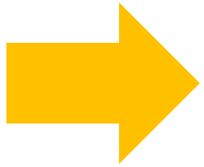
* IDC, Data Age 20215: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

技术发展日新月异

11 年以前 | 8 | 5 | 4

- Hadoop
- Elasticsearch
- Presto
- Spark

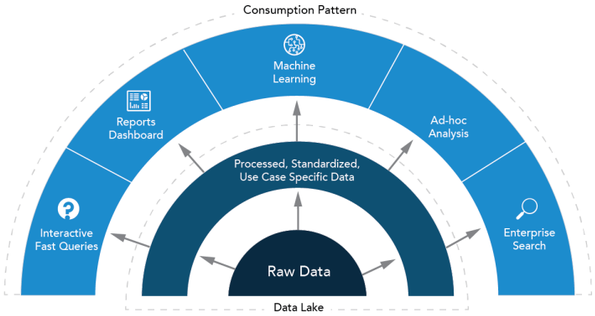
数据价值没有被充分挖掘

业务部门  IT部门

缺乏良好的企业内部协作

构建企业数据湖的目标

- 业务目标
  - 数字化经济，数据驱动业务
  - 增大企业运营效率
  - 预判发展趋势，加大企业竞争力
- 技术目标
  - 收集所有数据
  - 分析无处不在，采用多种技术
  - 自动化
  - 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

议程

企业数据分析平台的演变

为什么需要构建企业数据湖

如何在 AWS 云上构建企业数据湖

AWS Lake Formation 介绍

构建原则一：数据湖工作区划分

- **Raw Data Zone（原始数据区域）**：存放原始数据的区域，该区域敏感数据必须加密，标记化或以其他方式保护。
- **Trusted Data Zone（受信任区域）**：对原始数据区域中的数据执行数据质量、数据转换或其他处理后的数据存放区域，它也是中下游系统的“真实数据来源”，也就是说其下游系统会从该区域获取数据。
- **Discovery Sandbox Zone（分析沙盒）**：沙盒是数据湖不可或缺的一部分，因为它允许数据科学家和管理者创建特殊的探索性用例，而无需让IT部门参与或投入资金来创建合适的环境来测试数据。 数据可以从任何区域导入沙盒，也可以直接从源导入沙盒。这使公司能够探索某些变量如何影响业务成果，从而获得进一步的见解，以帮助做出业务管理决策。您可以将这些见解中的一些直接发送回原始区域，允许派生数据作为源数据，从而为数据科学家和分析师提供更多的工作空间。
- **Transient Zone（瞬时区域）**：用于在获取之前短暂保存数据，例如临时副本，流式或其他短期数据
- **Refined Zone（再处理区）**：操作和丰富的数据保存在此区域，这用于存储来自 Hive 或外部工具等的输出，这些工具将写入数据湖中。

Amazon Simple Storage Service (Amazon S3)

- AWS 的第一个云服务（2006年）
- 99.999999999% 数据持久性
- 不限对像格式
- 存储无上限
- 按使用付费，价格便宜
- 支持事件驱动的自动化
- 替换 HDFS，解耦计算与存储

AWS 上的数据仓库 - Amazon Redshift

- PB 级数据仓库
- 大规模并行处理 (MPP)
- 关系型数据仓库 (SQL)
- 高性能
- 低成本
- 管理简便、大幅扩容

领导者节点

统一的 SQL 访问端点
元数据存储
优化查询计划
协调查询执行

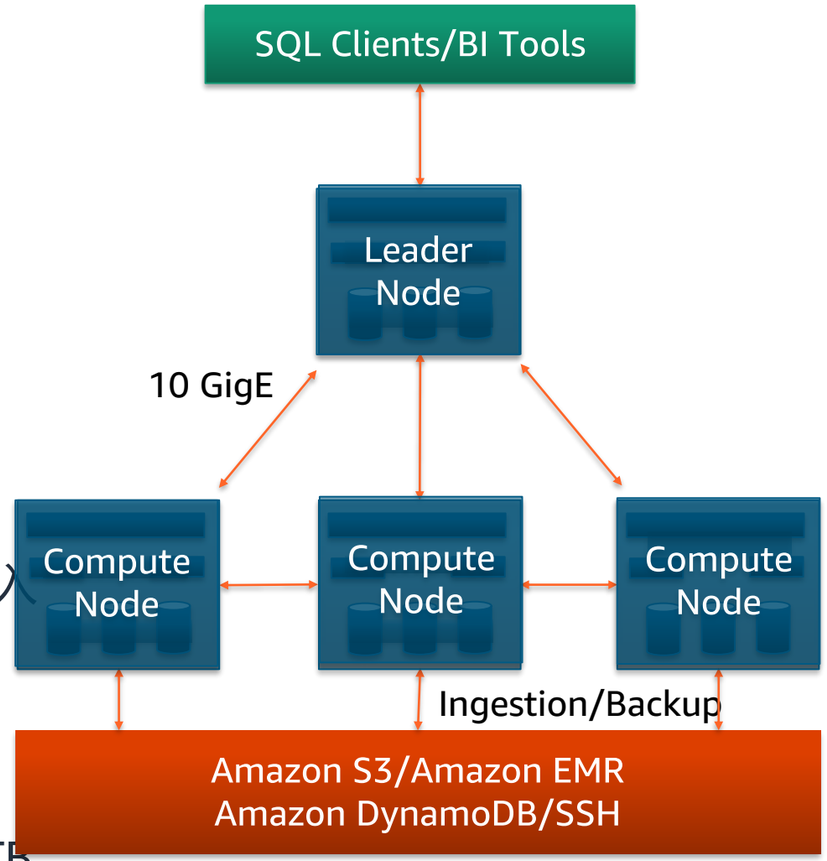
计算节点

本地列式存储
并行/分布式执行所有查询, 数据导入
备份, 恢复, 集群调整

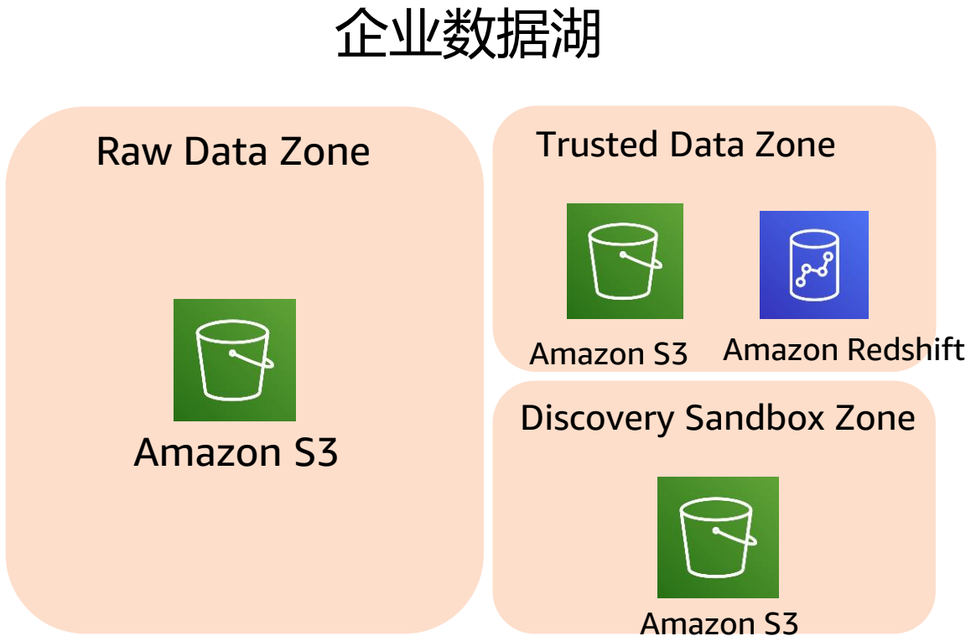
最小节点$0.25/小时, 最大至2 PB (压缩)

DC2: SSD; 容量从160 GB 到 326 TB

DS2: HDD; 容量从 2 TB 到 2 PB



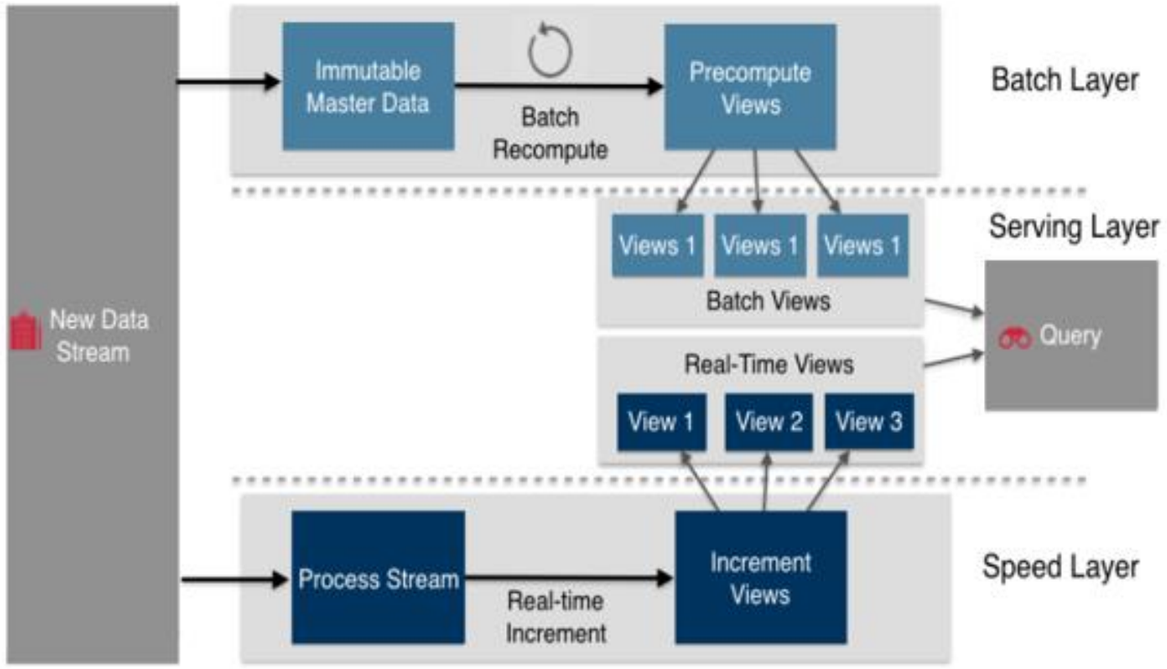
企业数据湖架构



构建原则二：遵循数据分析的 AWS Lambda 架构

数据分析处理分成三层:

- 批处理程序（batch laryer，非实时），比如午夜跑出来的报表，可以供第二天进行消费；
- 实时增量处理数据（speed layer），比如通过流计算工具进行的实时增量处理；
- Service layer，是对外提供服务的层，既可以访问 batch layer 或者 realtime layer，还可以整合两者数据然后对外提供服务。



AWS 上的实时流数据服务 - Amazon Kinesis

- Kinesis Streams
- Kinesis Firehose
- Kinesis Analytics
- Kinesis for video

实时

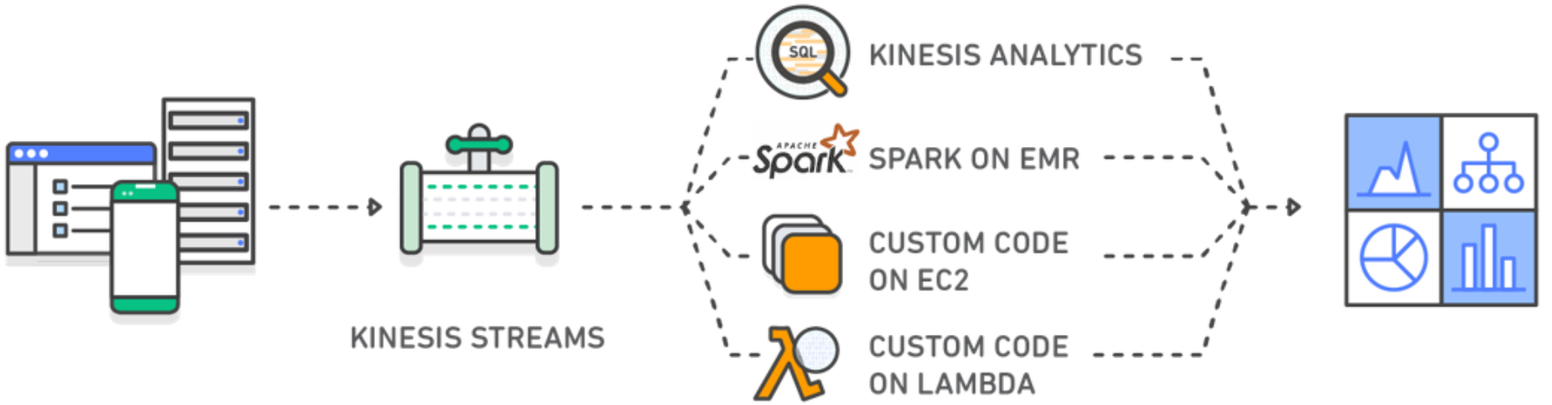
可以实时接收、缓冲和处理数据，从而可以在几秒或几分钟内得出分析结果

完全托管

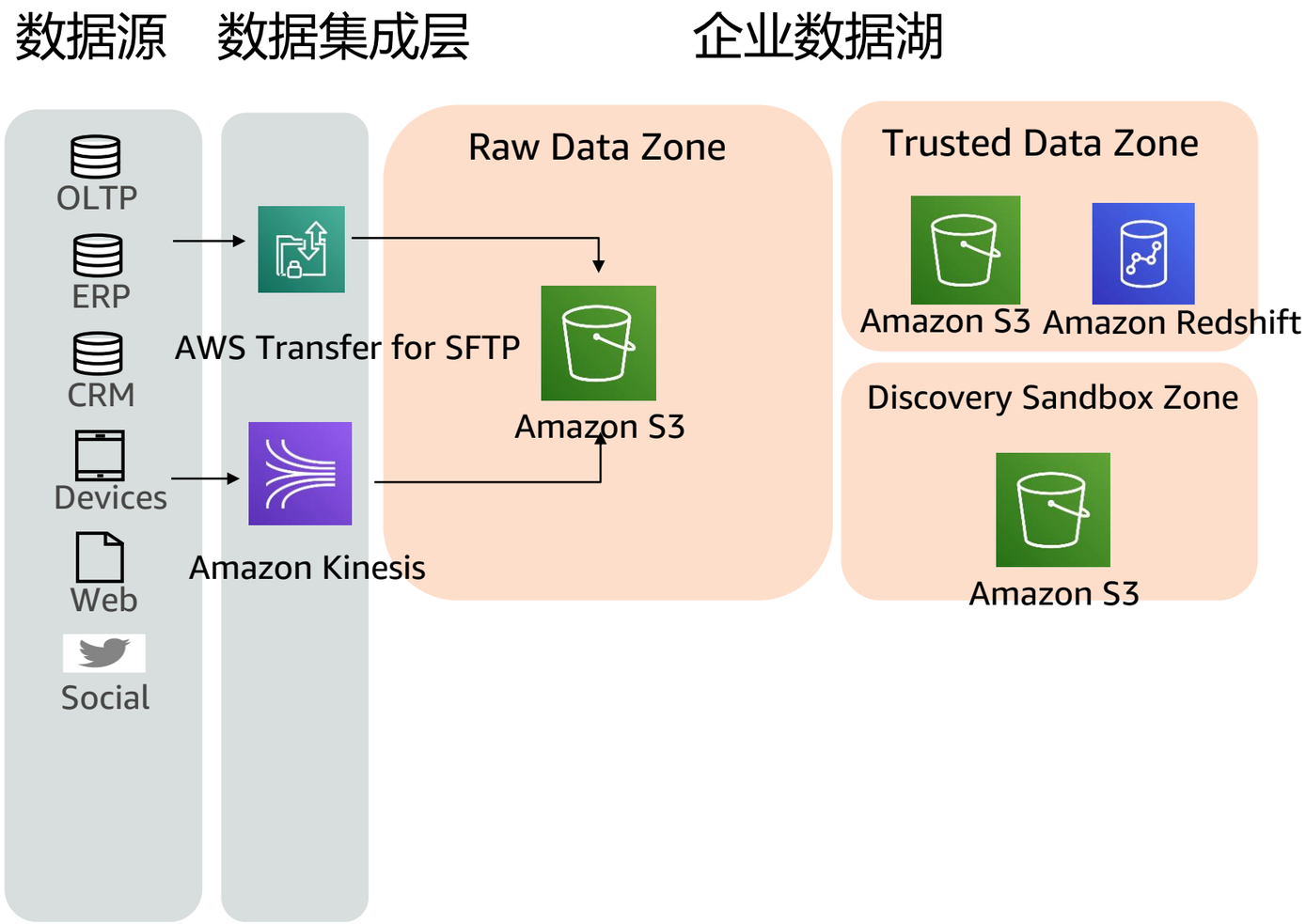
完全托管，不需要管理任何基础设计，冗余设计

可扩展

可同时处理来自几十万个来源的任意数量的流数据，延迟非常低

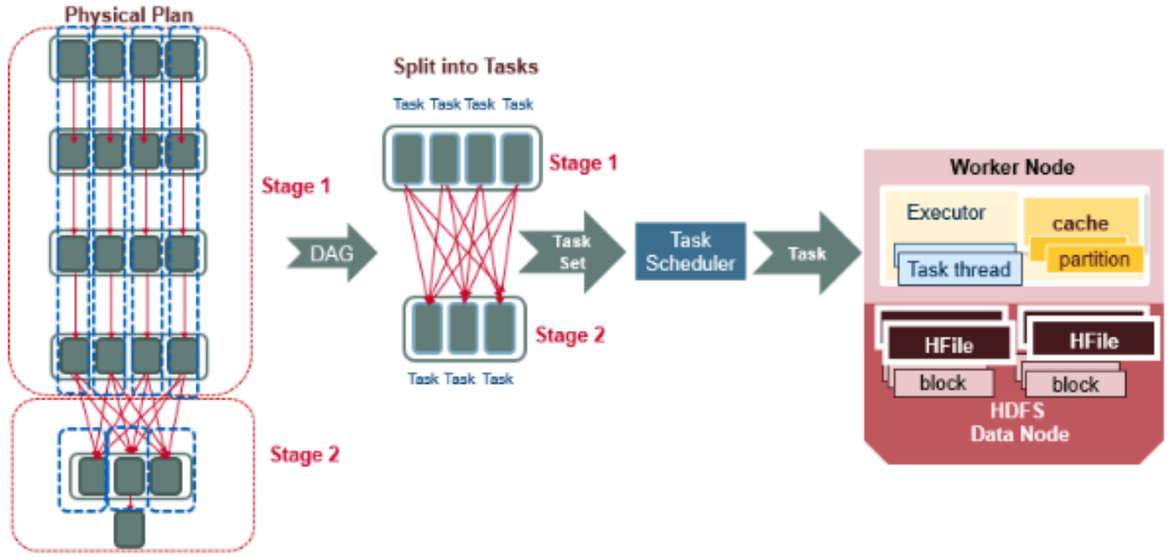


企业数据湖架构



构建原则三：统一的元数据管理

- 统一的元数据管理功能
- 自助式服务：
  - 内置数据转换和清洗函数或组件
  - 可视化作业编排
  - 自动化：基于时间、基于事件、基于策略的任务调度
    - 采用分布式的集成任务调度，并支持分钟、小时、日、周、月等多种时间调度周期，提升数据湖的数据集成效率
    - 多种控制策略：支持集成作业重试、作业依赖、人工重跑等多种作业控制策略，保障数据集成作业的 SLA



AWS Glue

AWS Glue

数据目录

- 兼容 Hive Metastore，并提供增强功能
- 爬虫自动提取元数据并创建表
- 与 Amazon Athena, Amazon Redshift Spectrum 集成

发现

任务编写

- 自动生成 ETL 代码
- 基于开源框架构建 – Python 和 Spark
- 面向开发人员 – 编辑, 调试, 共享

开发

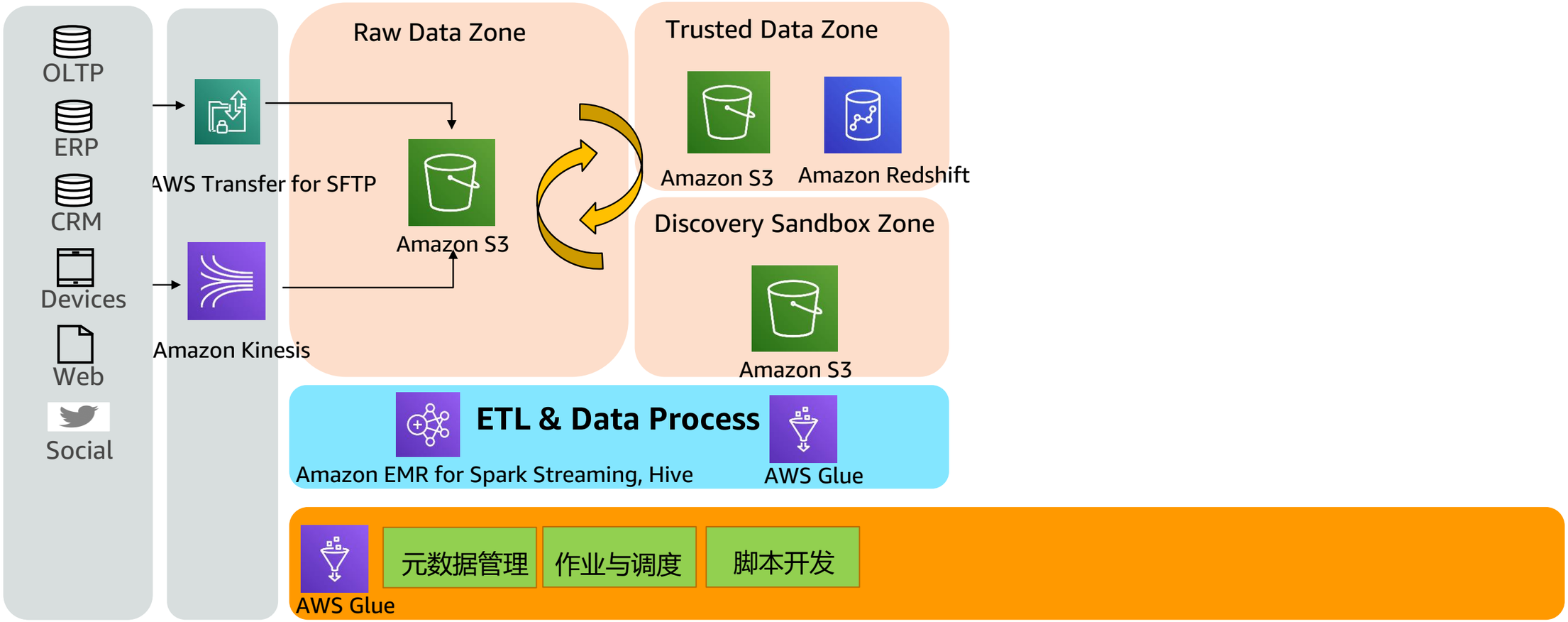
任务执行

- 在无服务器 Spark 平台上运行作业
- 提供灵活调度
- 处理依赖项解析，监视和警报

部署

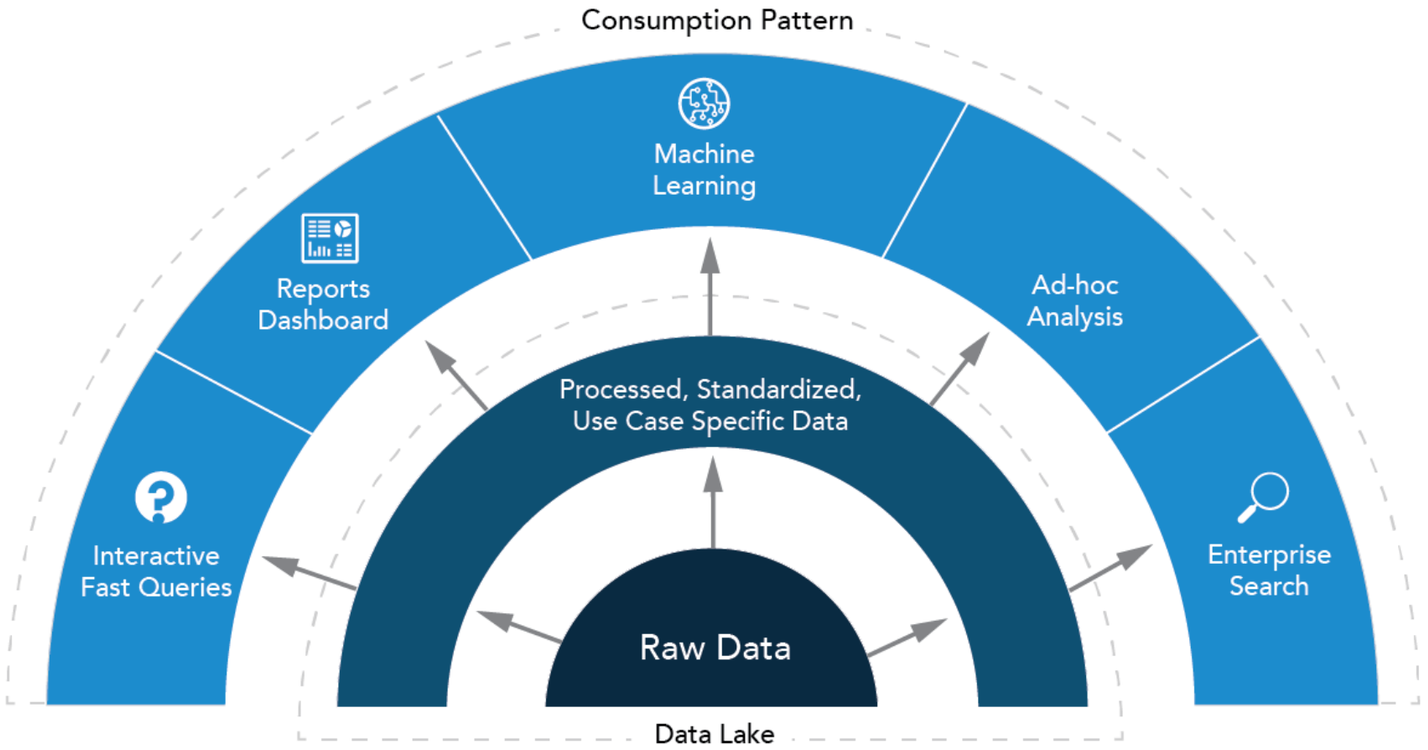
企业数据湖架构

数据源　数据集成层　　企业数据湖



构建原则四：满足企业各种场景的数据消费需求

- 即时查询
- BI 报表
- 数据探索
- 企业搜索
- 机器学习（ML & AI）



自助式服务

AWS 上的大数据分析引擎 - Amazon Athena

Amazon Athena

无服务器架构

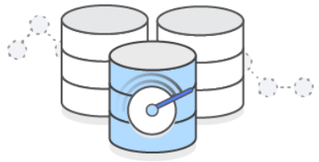
Athena 没有服务器。您可以快速查询数据，而无需设置和管理任何服务器或数据仓库, 无需 ETL

按查询量付费

借助 Amazon Athena ，您只需为您运行的查询付费

开放标准

Amazon Athena 使用支持 ANSI SQL 的 Presto, 并能处理各种标准数据格式, 非常适合用于进行快速的即席查询

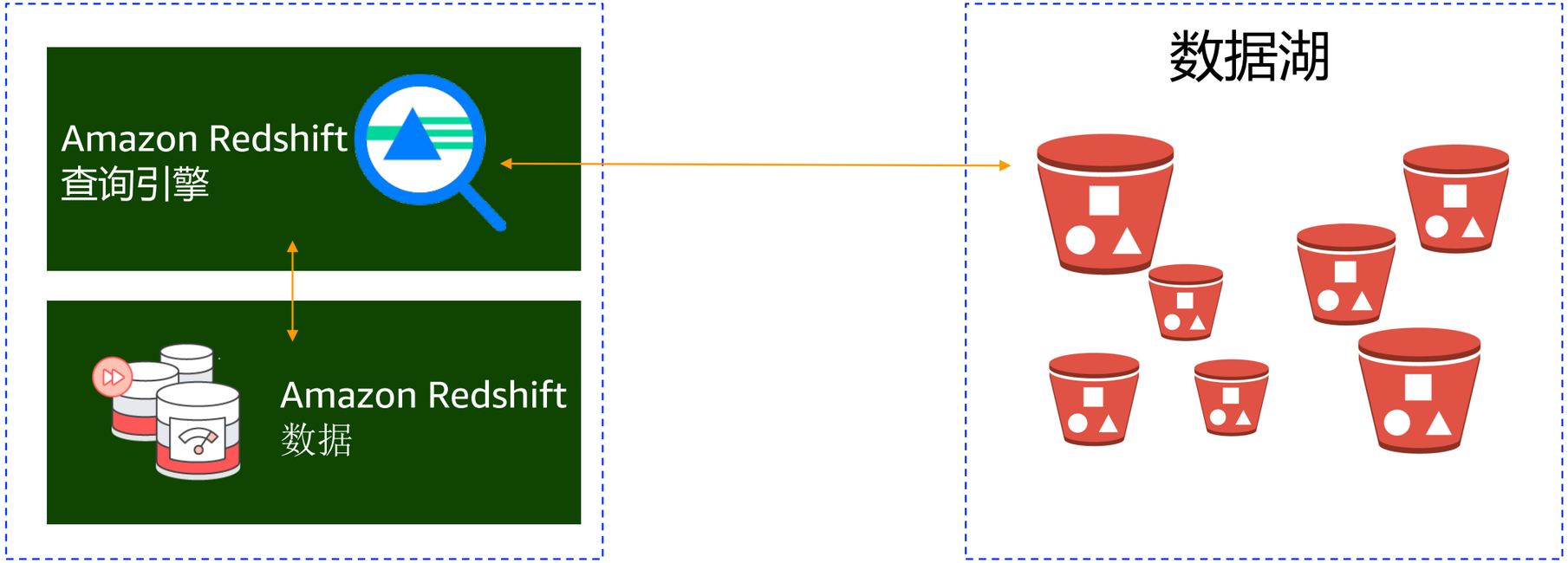


快速的交互式查询性能

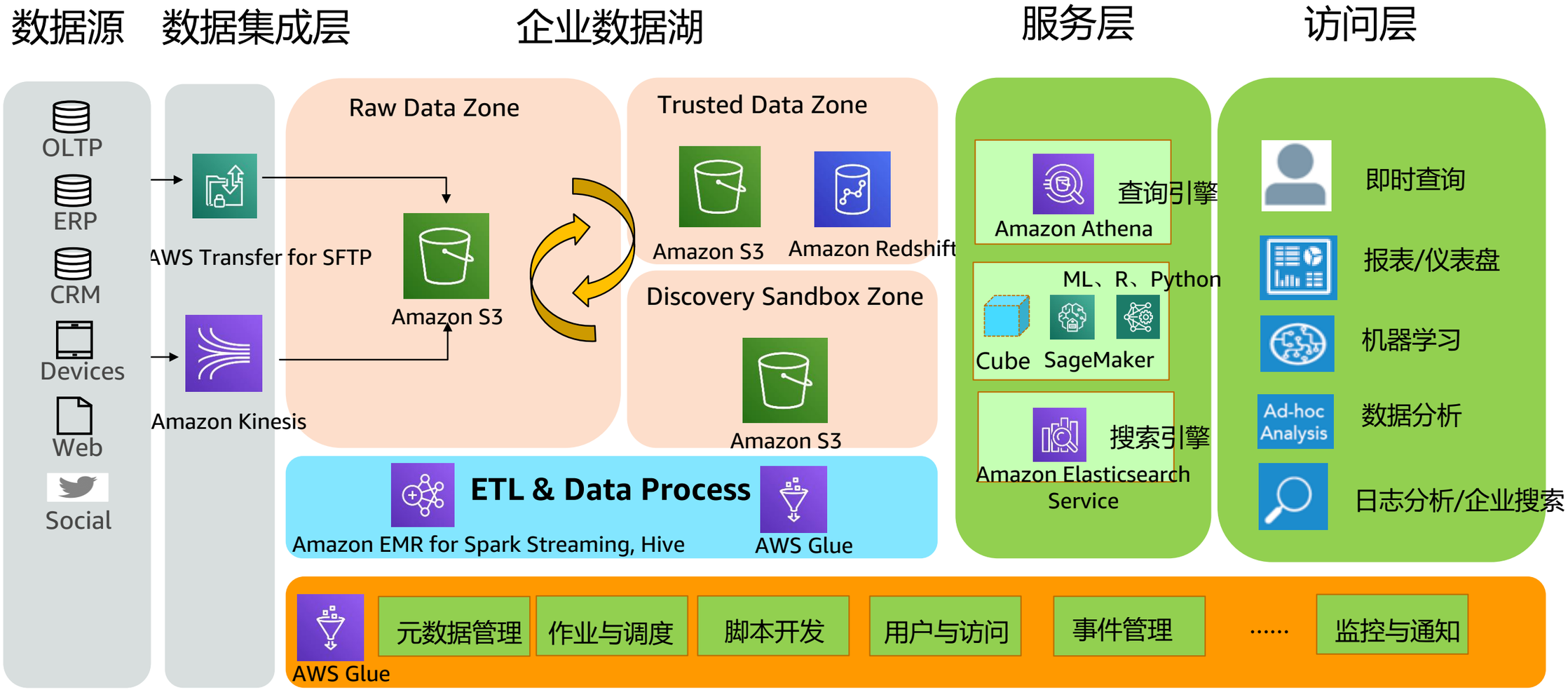
Amazon Athena 可提供快速的交互式查询性能，因此您无需担心自己没有足够的计算资源

Redshift Spectrum：将 Amazon Redshift 查询扩展到 Amazon S3

- 直接查询 Amazon S3 或跨 Amazon Redshift 和 Amazon S3 联接数据
- 支持CSV、JSON、ORC 和 Parquet 数据格式
- 单独扩展 Amazon Redshift 计算和存储



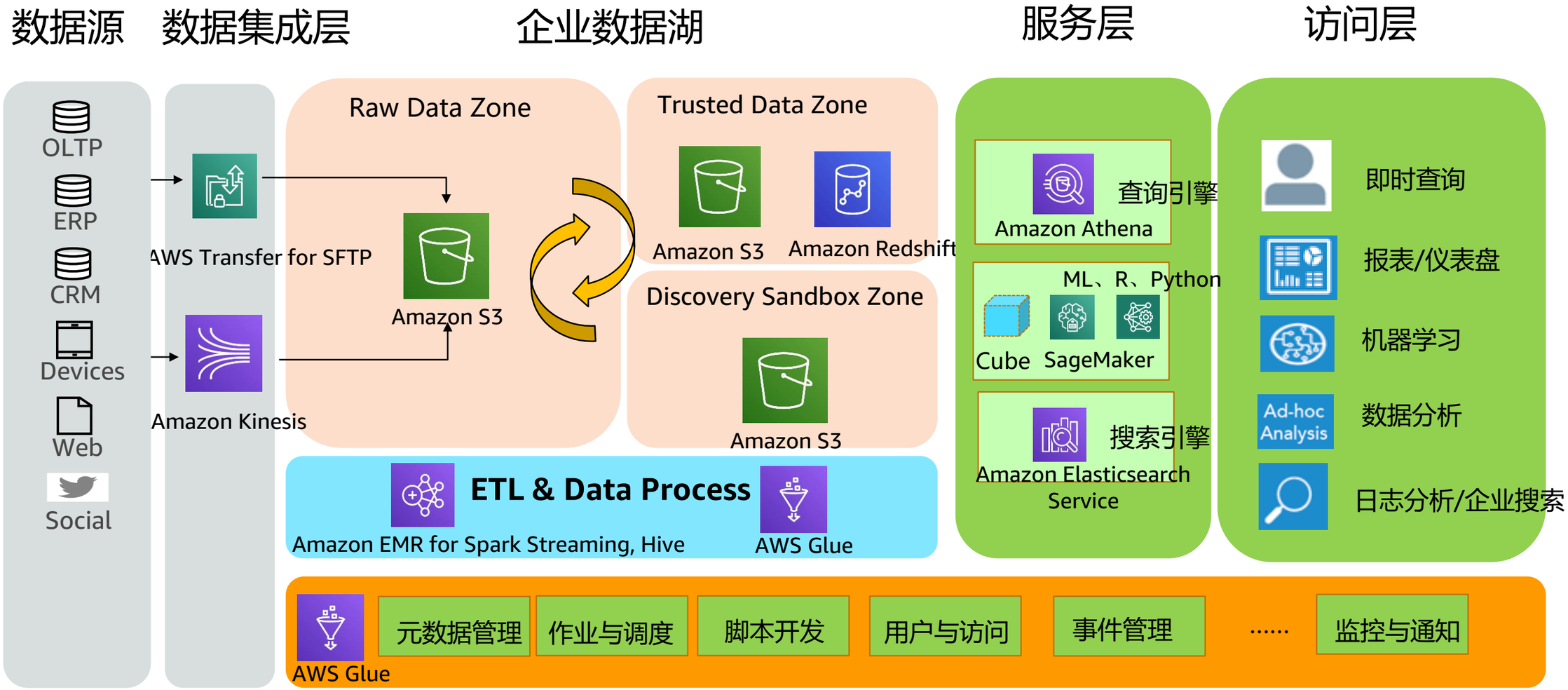
企业数据湖架构



构建原则四：统一的数据安全和访问策略

- 基于角色的管理访问控制（RBAC）
  - 用户及用户组隔离
  - 数据访问权限
- 数据保护、数据加密
- 安全审计
- 行业与企业数据安全标准
  - PCI DSS
  - ……
- 密钥管理、证书管理

企业数据湖架构



议程

企业数据分析平台的演变

为什么需要构建企业数据湖

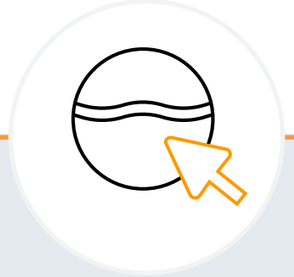
如何在 AWS 云上构建企业数据湖

AWS Lake Formation 介绍

AWS Lake Formation NEW!

在数天内构建安全的数据湖

快速构建数据湖
Build a data lake in days, not months



只需单击几下鼠标，即可构建和部署完全管理的数据湖

简化安全管理
Enforce security policies across multiple services

集中定义安全性、监管和审计策略，而不是按服务执行这些任务，然后跨其分析应用程序为您的用户实施这些策略。减少跨服务配置策略的工作量，并提供一致的实施和合规性。

轻松安全地自助访问数据
Combine different analytics approaches

统一数据目录，增强分析师和数据科学家的工作效率，使他们能够自助发现并安全地从单个目录访问所有数据

Blueprints / Data Importers

Templates

Blueprints 是数据摄取、转换、元数据（schema）和分区管理的模板. Blueprints 帮助客户快速、轻松地构建和维护一个数据湖。

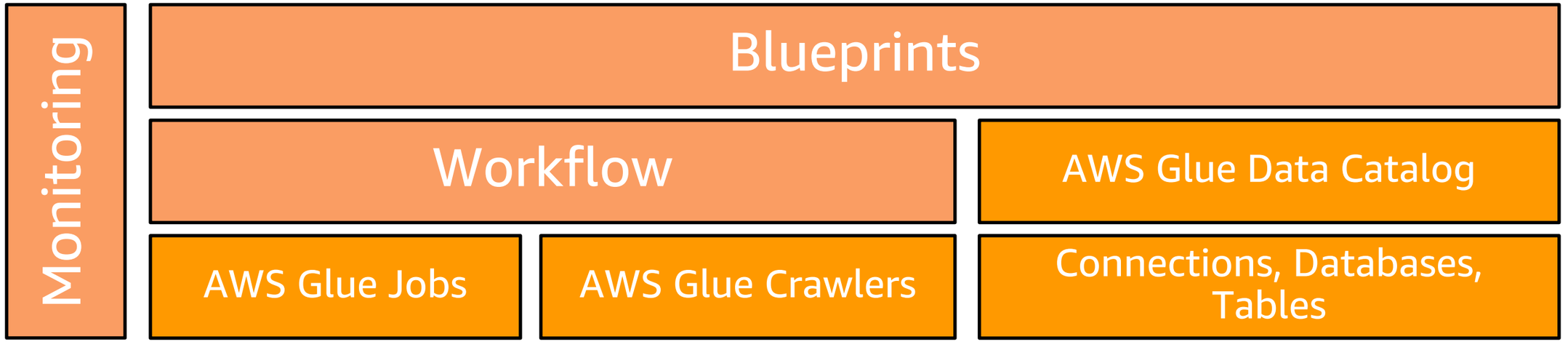
你

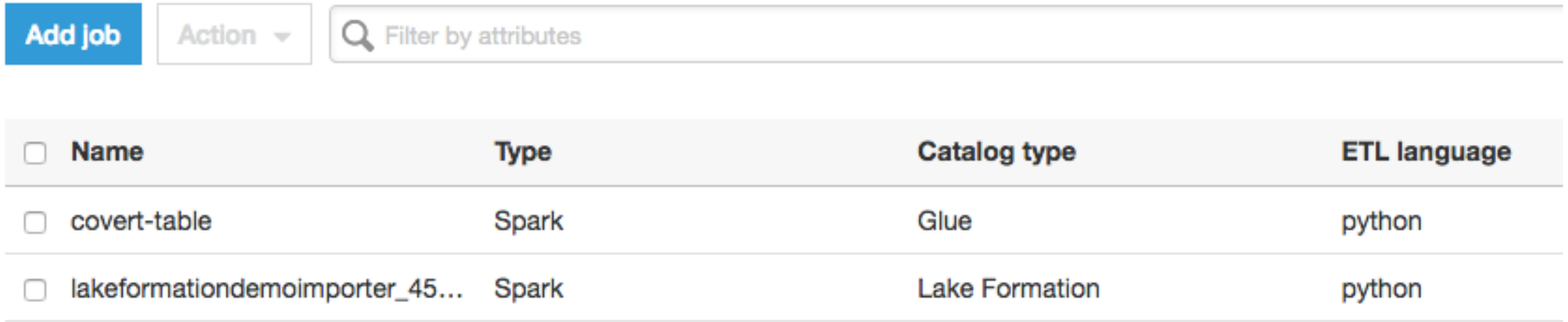
1. 数据源在哪里
2. 目标数据湖存储位置
3. 设定多久装载一次数据

Blueprints

1. 自动发现源数据 Schema
2. 自动进行源和目标表的数据转换
3. 自动数据分区
4. 作业状态跟踪
5. 灵活客户化 Blueprints 模板

Blueprints build on AWS Glue





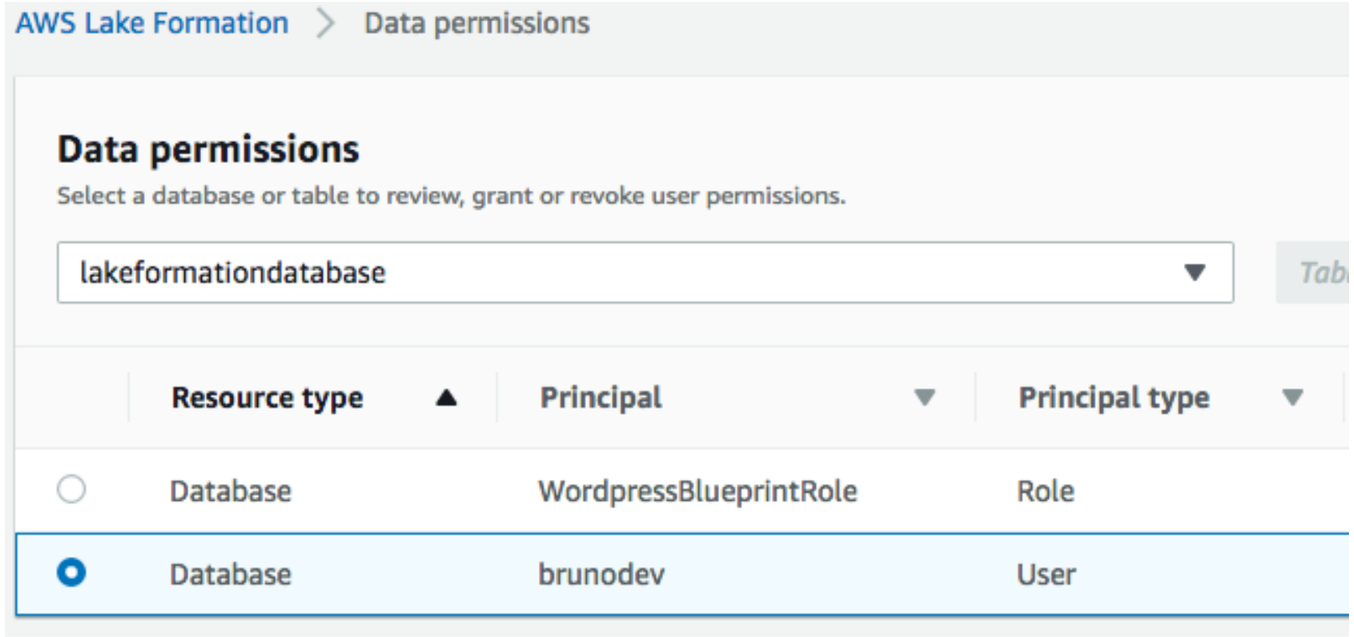
AWS Lake Formation 的安全实现

使用简单的授权和吊销权限

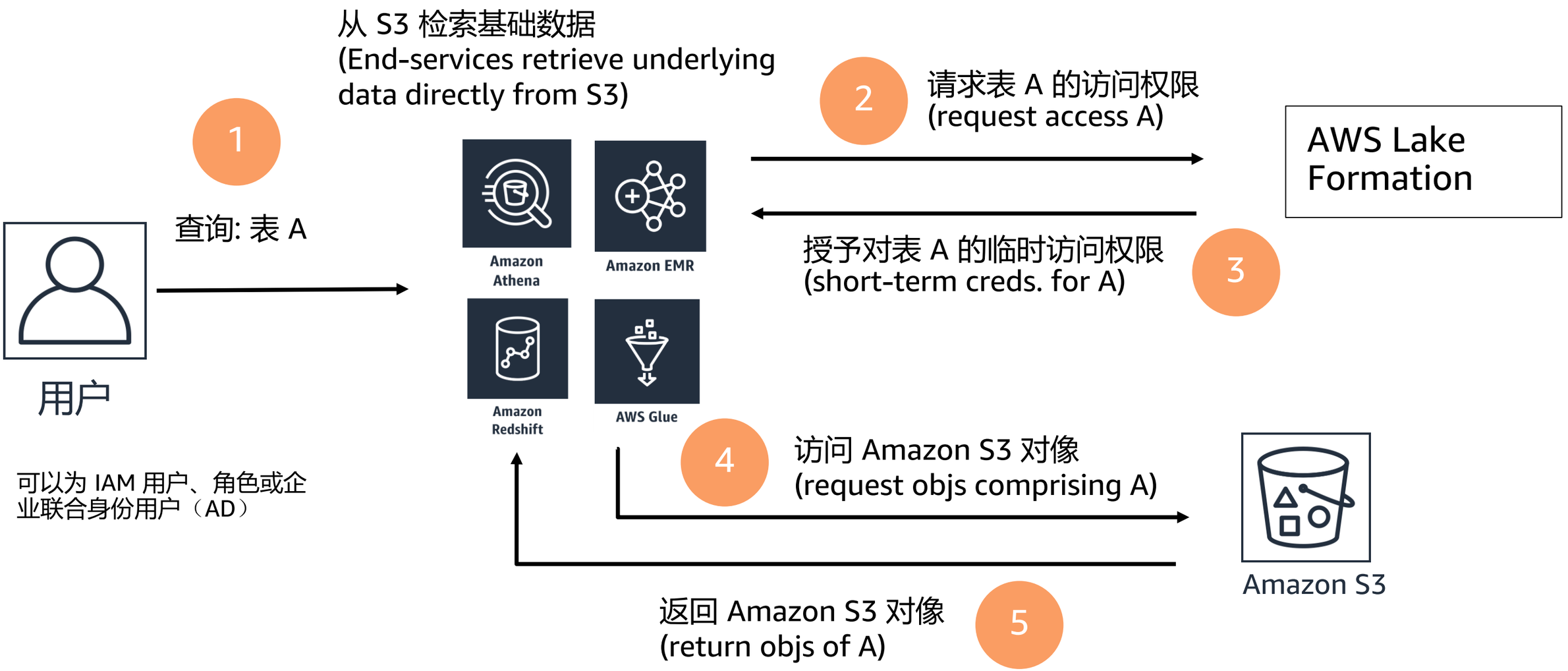
指定对表和列的权限，而不是对存储桶和对象的权限

轻松查看和管理用户权限

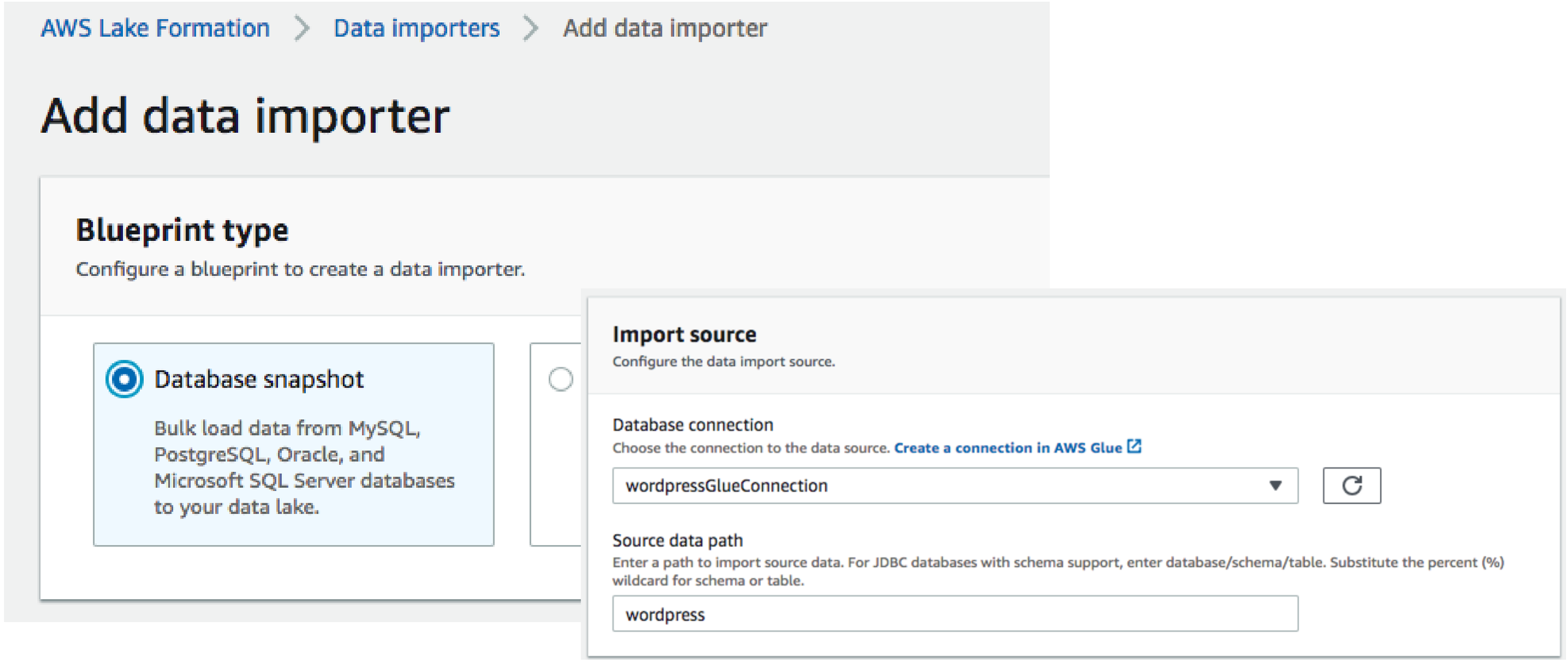
集中审计功能



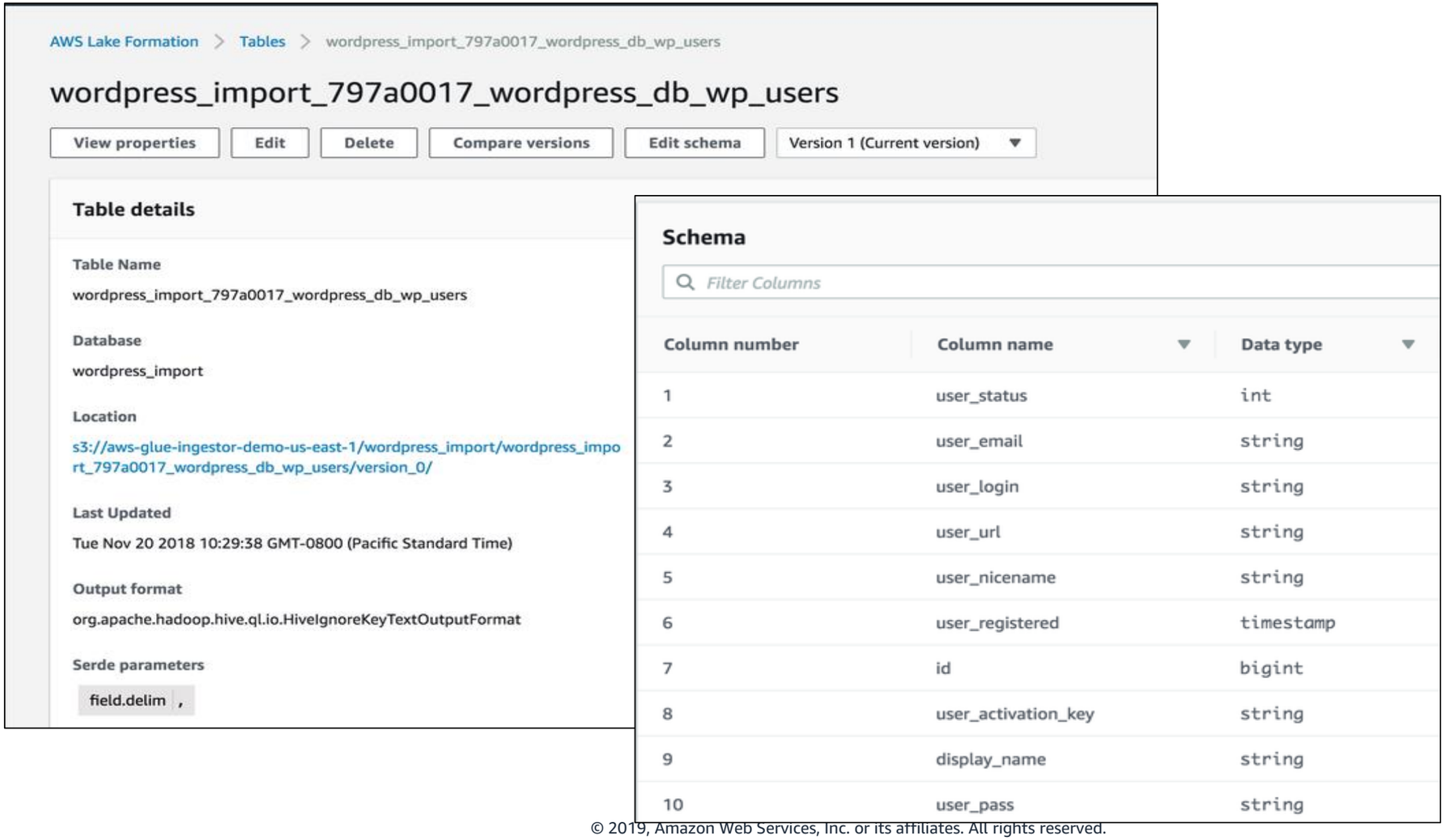
AWS Lake Formation 的安全实现



步骤 1：使用 data importer 导入数据

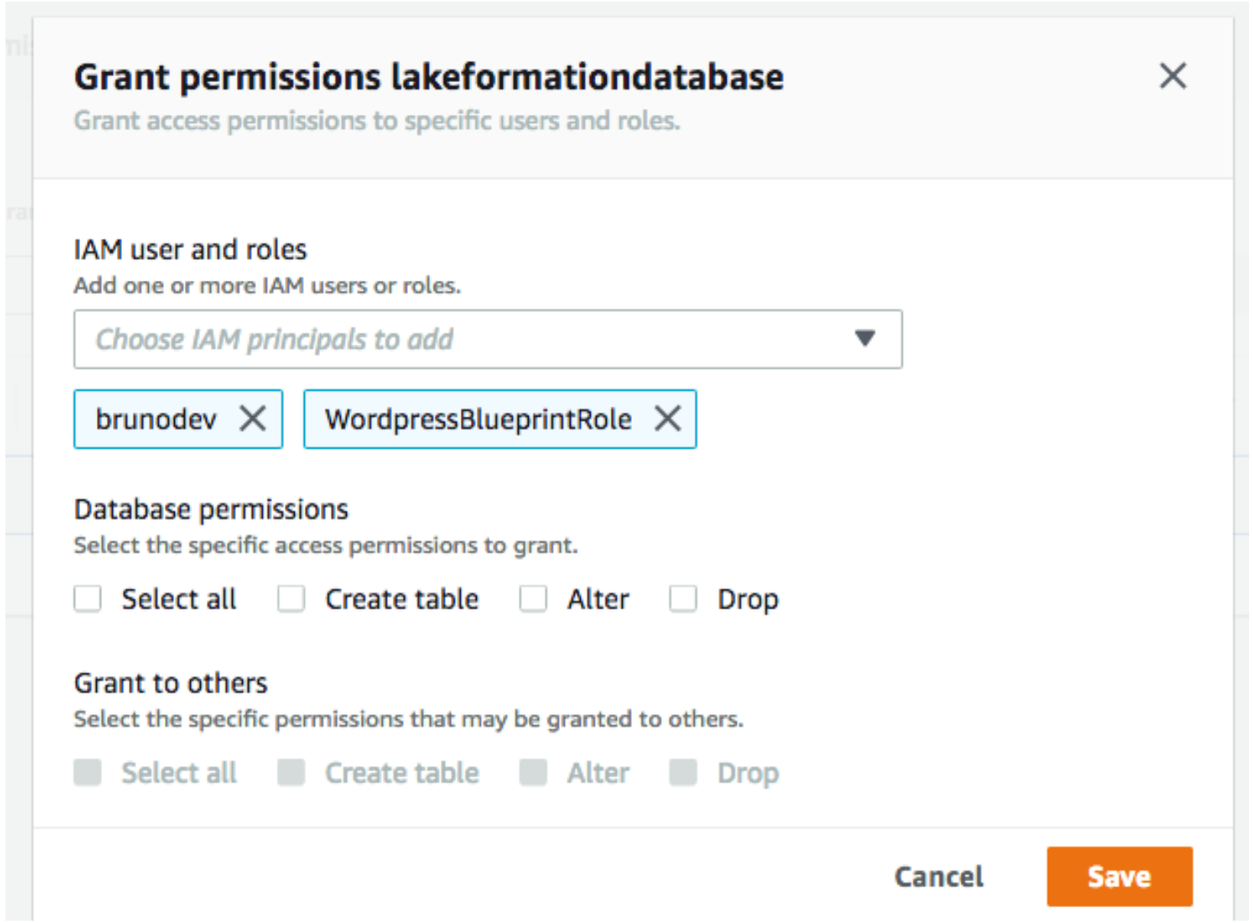


向数据湖导入数据

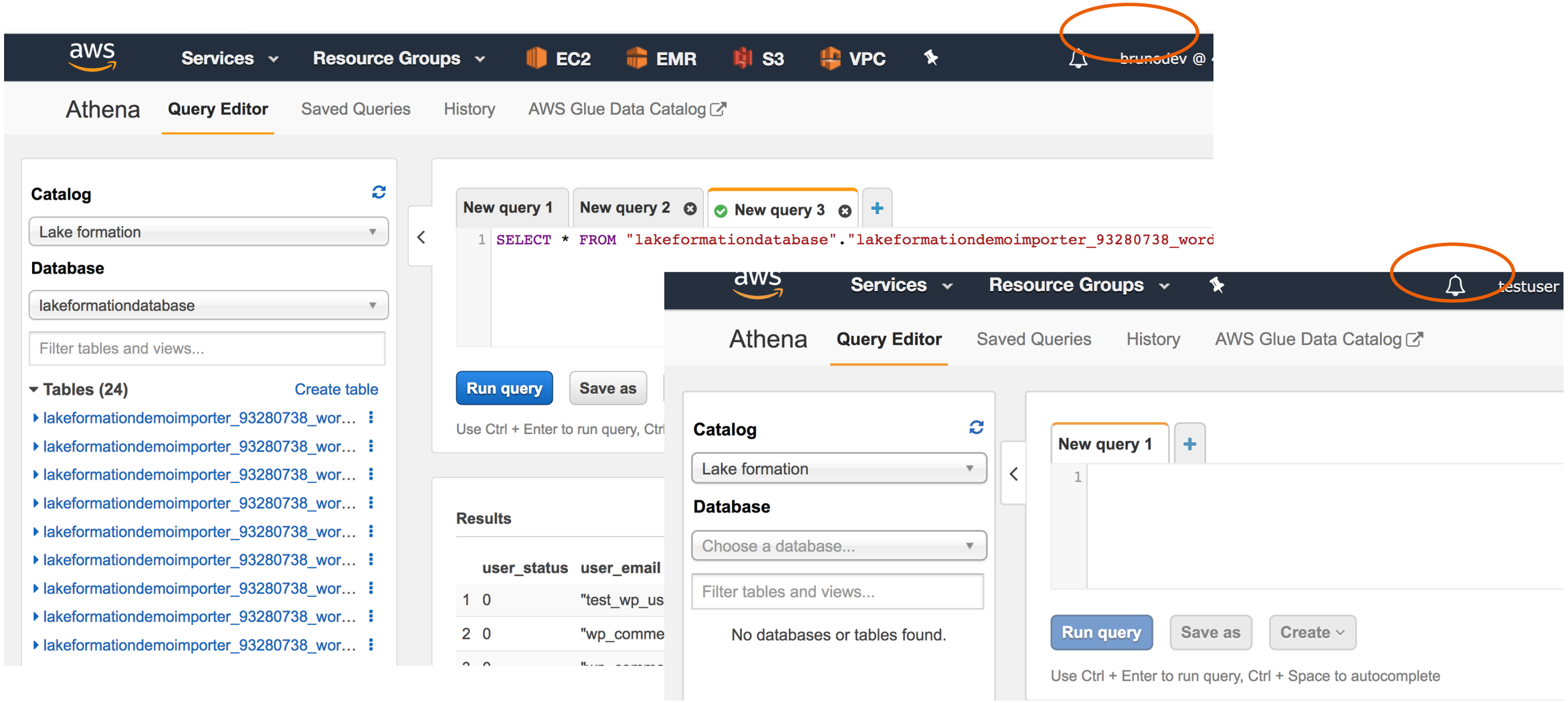


步骤 2：数据安全和访问权限设置

您可以针对某一张表进行访问授权，访问用户可以为 IAM 用户、IAM 角色，AD 用户或用户组。



步骤 3：查询数据（Amazon Athena）



感谢参加 AWS INNOVATE 2019 在线技术大会

我们希望您在这里找到感兴趣的内容！

也请帮助我们完成**投票打分**和**反馈问卷**。

欲获取关于 AWS 的更多信息和技术内容，可以通过以下方式找到我们：


微信公众号：AWSChina


新浪微博：https://www.weibo.com/amazonaws/


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


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


视频中心：http://aws.amazon.bokecc.com/


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

aws INNOVATE
ONLINE CONFERENCE