

AWS数据湖及大数据服务助力 快消行业进行数字化转型

王辉 解决方案架构师

Hui Wang, Solution Architect, Amazon Web Services

2018年7月31日

July 31, 2018

课程内容

- 行业概述
- AWS数据湖解决方案
- 数据湖设计最佳实践
- 客户案例
- 实战演示

行业概述

行业概述

快消品，快速消费品（FMCG，Fast Moving Consumer Goods）的简称，是指那些使用寿命较短，消费速度较快的消费品。

快速消费品行业主要分为五个子行业：

1. 个人护理品行业，由口腔护理品、护发品、个人清洁品、化妆品、纸巾、鞋护理品和剃须用品等行业组成；
2. 家庭护理品行业，由以洗衣皂和合成清洁剂为主的织物清洁品以及以盘碟器皿清洁剂、地板清洁剂、洁厕剂、空气清新剂、杀虫剂、驱蚊器和磨光剂为主的家庭清洁剂等行业组成；
3. 品牌包装食品饮料行业，由健康饮料、软饮料、烘烤品、巧克力、冰淇淋、咖啡、肉菜水果加工品、乳品、瓶装水以及品牌米面糖等行业组成；
4. 烟酒行业，能够快速消化的物品。
5. 服装鞋帽。

主要的业务系统

- **SAP:** 核心系统，管理发货、退货、财务、采购数据。
- **DMS:** 经销商管理系统，管理货的流向到哪里，重点在于经销商协同。
- **POS:** 门店销售系统，门店销售数据。
- **TPM:** 营销费用管理系统，包括：渠道费用(TP)、消费者促销费用(CP)、广告费用(AP)。
- **SFA:** 销售能力自动化系统，管理销售人员的行为，重点在于客户开发。
- **EPC:** 驻店人员管理系统，收集 Offtake，会员数据。
- **WMS:** 库存管理系统，经销商库存数据。
- **CRM：** 客户关系管理系统，客户主数据。

业务维度模型

业务模块	业务流程	基础指标	时间	经销商 / 直供	渠道	产品	组织架构	库龄	门店	区域	竞品
销售	Sellin	数量，金额（含税，未税），达成率，毛利率	x	x		x				x	
	SellOut	数量，金额（含税，未税），达成率	x	x	x	x	x		x	x	
	OffTake	数量，金额（含税，未税）	x	x	x	x	x		x	x	
	总仓库存	数量，金额，DOH，库存新鲜度	x			x		x			
	经销商库存	数量，金额，DOH，库存新鲜度	x	x		x		x		x	
	经销商采购	数量，金额	x	x		x				x	
	门店库存	数量，金额，DOH	x	x	x	x	x	x	x	x	
客户活动	拜访	拜访次数，目标拜访次数	x	x	x		x		x	x	
	巡店	陈列面积（含竞品），巡店数量，货架占比	x	x	x		x		x	x	x
		分销，必卖单品满足率	x	x	x	x	x		x	x	
财务	活动费用预算	预算金额	x								
	活动费用核销	核销金额	x								
	经销商返利	返利金额	x	x						x	
	应收	应收金额	x	x							
市场	市场占有率	数量，金额，市场占有率	x		x	x				x	x

行业的“痛”

模式：

- B2B：生产商（B）到经销商（B）
- B2C：经销商（B）到消费者（C）
- B2B2C：生产商（B）到经销商/交易平台（B）到消费者（C）

痛点：

- 渠道割裂、缺乏商品有机链接
- 营销盲打，缺乏有效数据支持
- 假冒伪劣，缺乏产品溯源机制



行业的“变”

新零售：

- 全渠道
- 多样化场景
- 精细化运作
- 注重服务和体验
- 数据为核心



新诉求：

- 提高企业运行效率
- 减少商品流通环节
- 精确感知市场需求



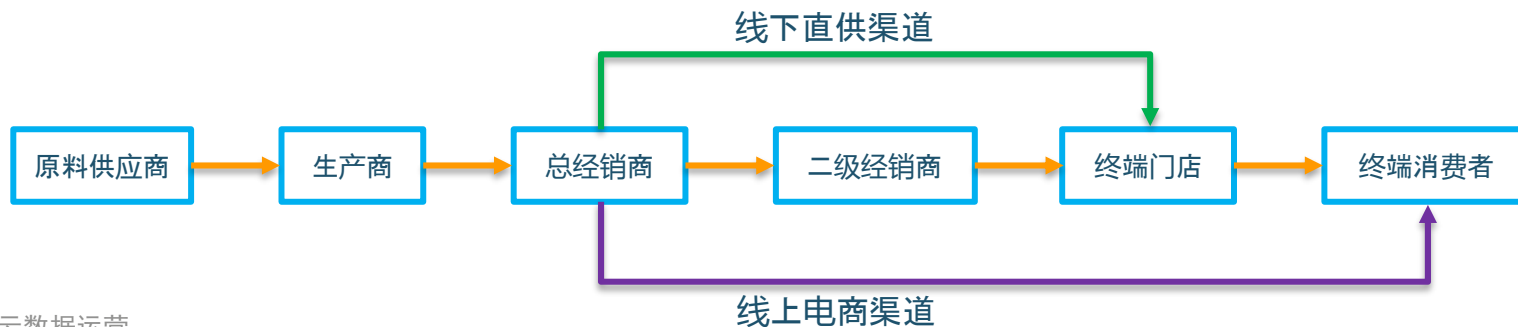
数字化转型：

- 利用互联网的思维模式对企业进行全面改造
- 运用数据化的手段提升渠道效率，降低渠道成本
- 打通线上线下数据通路，与流通环节建立直接链接



行业目标：

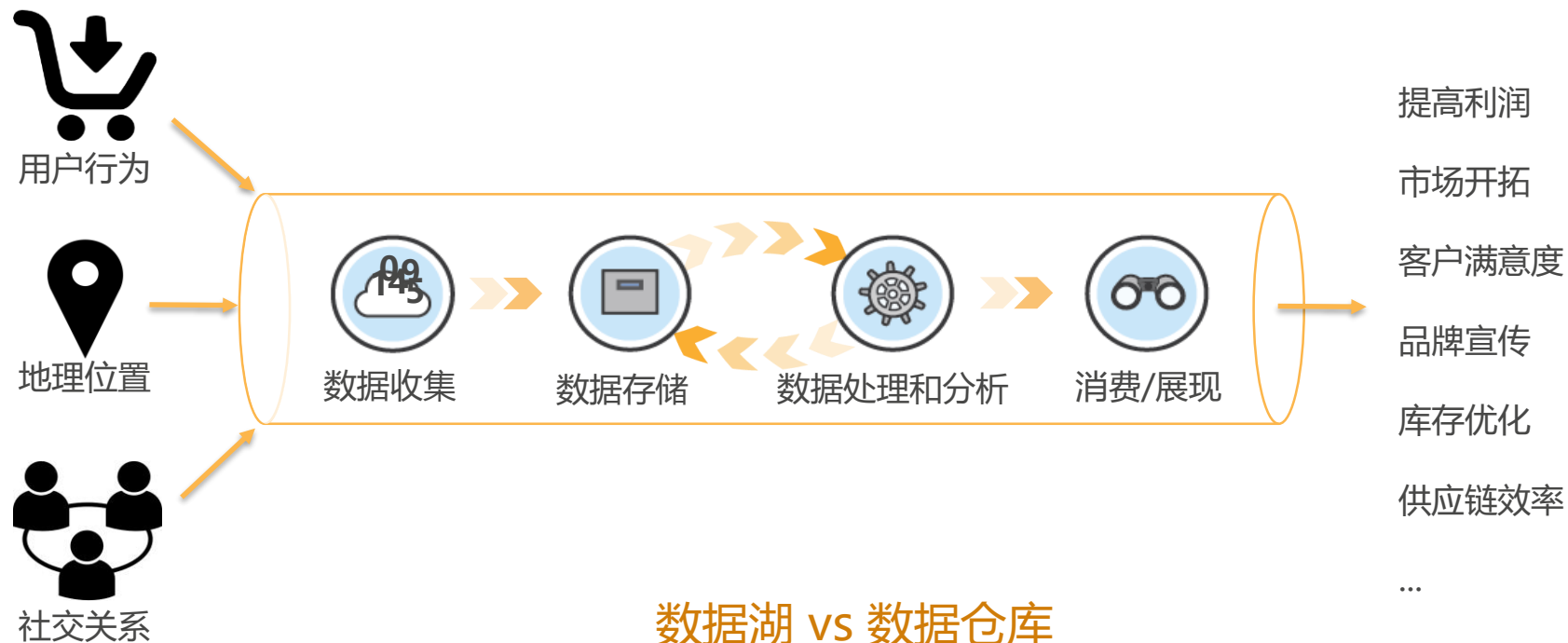
- 管理透明化
- 生产柔性化
- 营销自动化
- 流程数据化



AWS数据湖解决方案

数据湖-面向未来的大数据思想

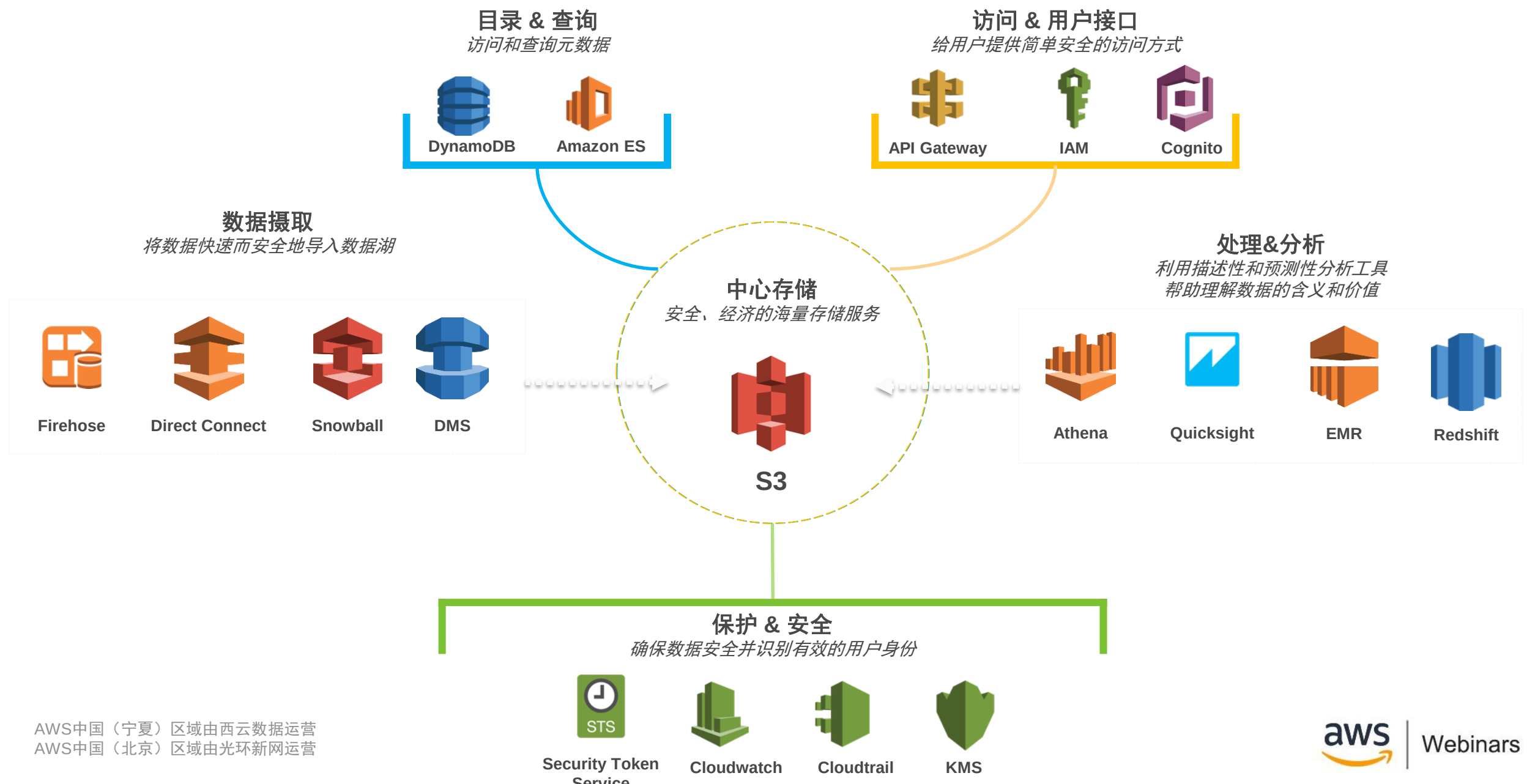
数据湖是大数据分析管道的重要组成部分，是存储了所有的源数据并可供所有人访问分析的**统一空间**



数据湖 vs 数据仓库

- 数据湖存原始数据，不要求统一的格式，可供任意目的的分析
- 数据仓库存处理后的数据，格式统一，预设明确的分析目的

AWS 数据湖 – Amazon S3



AWS 数据仓库 - Amazon Redshift

- **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

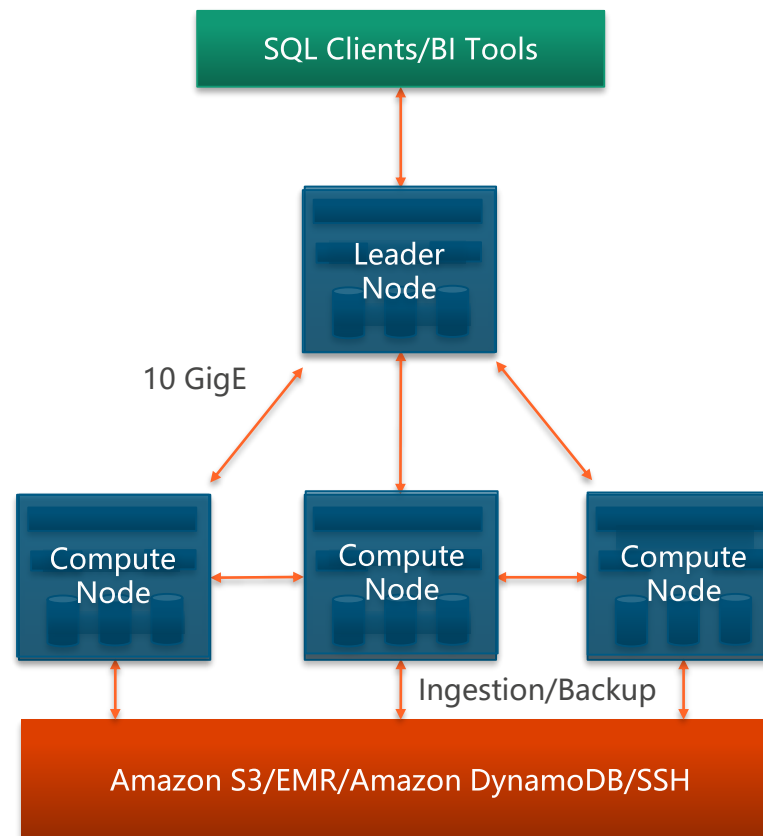
- **计算节点**

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

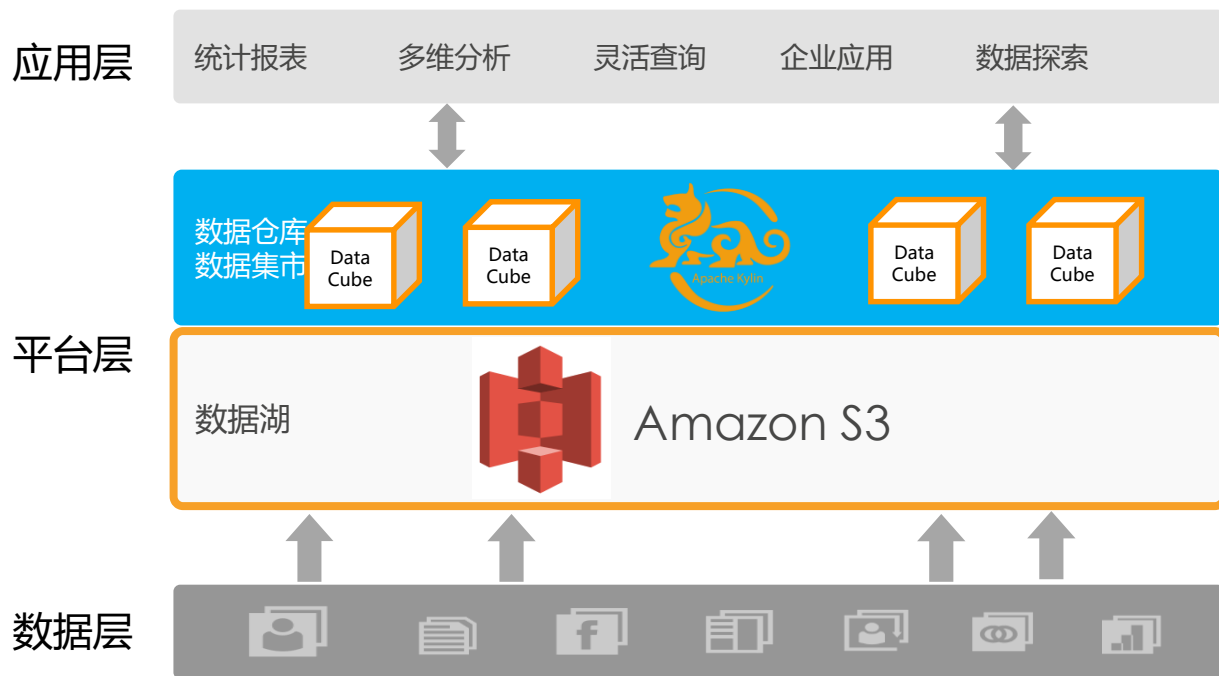
最小节点\$0.25/小时, 最大至2 PB (压缩)

DC1: SSD; 容量从160 GB 到 326 TB

DS2: HDD; 容量从 2 TB 到 2 PB

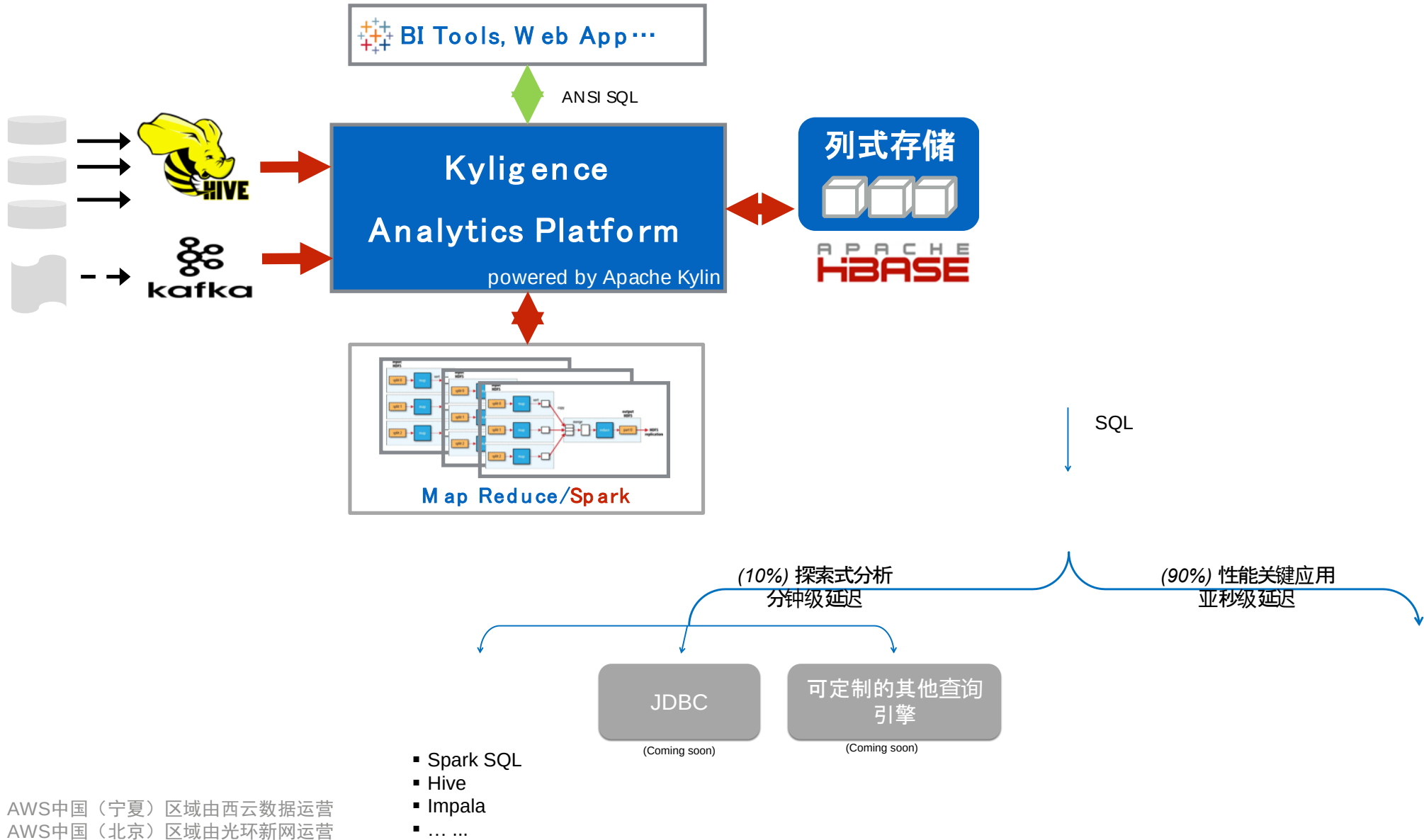


大数据之上多维分析系统 – Kylogence Analytics Platform (KAP)



- **高性能**
亚秒级交互式分析
- **高并发**
线性扩展
- **高可用**
KAP可自动恢复
- **易使用**
图形化开发
标准SQL访问
- **批量&实时**
同时支持批量及实时分析
- **Hadoop原生应用**
标准API交互
非侵入式部署

KAP 架构



- CUBE预计算
- 查询下压
- 数据可存于S3
- 明细数据查询
- 超高基维度查询
- SQL查询语言
- 融合查询
- 可视化展现
- 元数据驱动

KAP on AWS

集群

监控

支持

+ 新建集群

名称	状态	进度	创建时间	运行时间	描述	操作
kylligence-demo	RUNNING	100% 查看日志	2017-12-26 12:00:39 GMT+8	3 hr 53 min		停止 ...

现在, 您可以使用:

KAP

KyAnalyzer

Zeppelin

基础信息

操作日志

基础信息

集群类型: EMR

集群密钥: shaoxiongkey (ec2-user)

地区: us-east-1

S3 存储桶: awscreate

节点信息

核心节点: m4.xlarge * 1

修改

任务节点: m4.xlarge * 1

修改

SSH: ec2-34-203-190-213.compute-1.amazonaws.com

复制

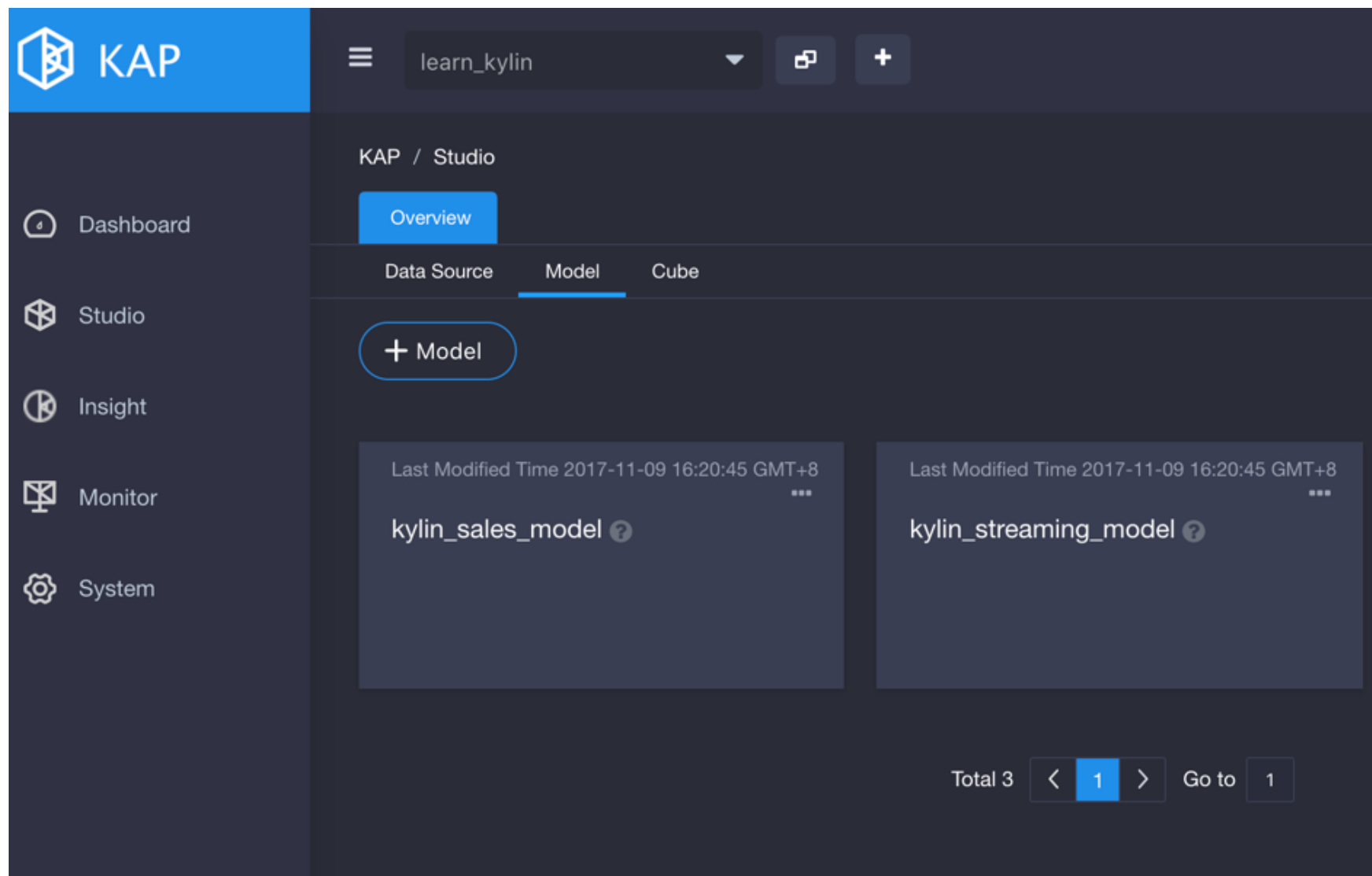
共 1 条

10 条/页

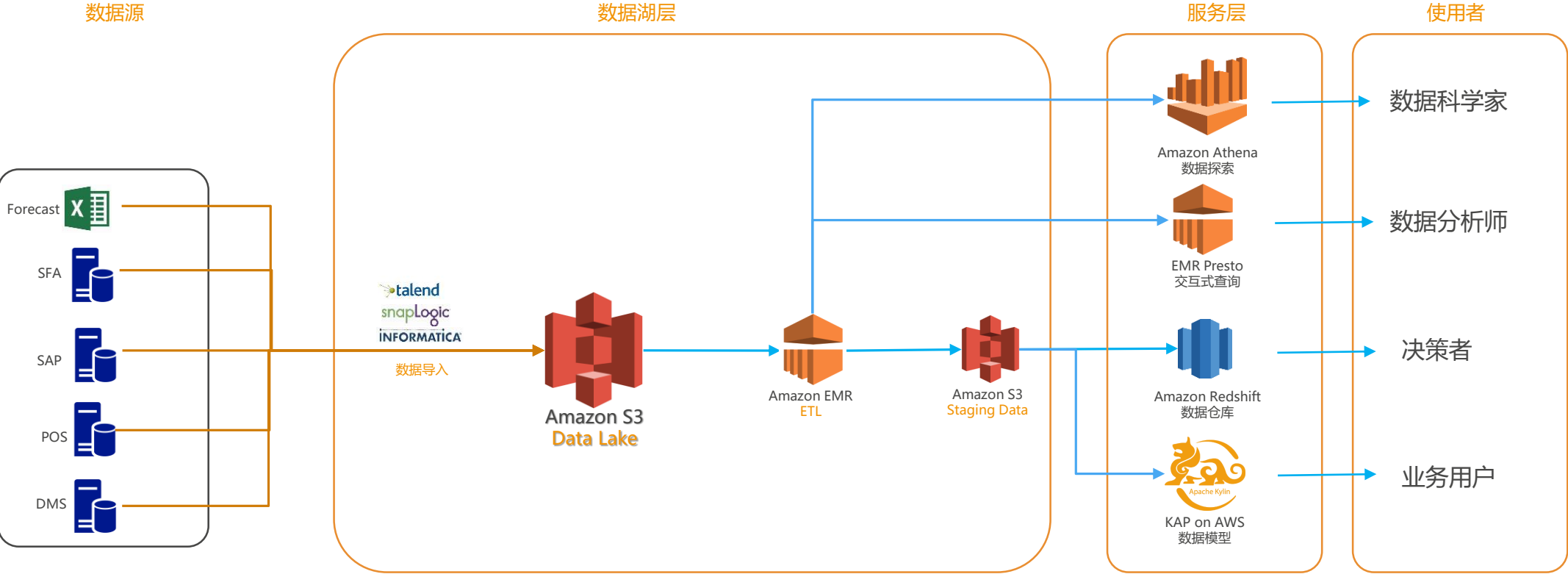
< 1 >

前往 1 页

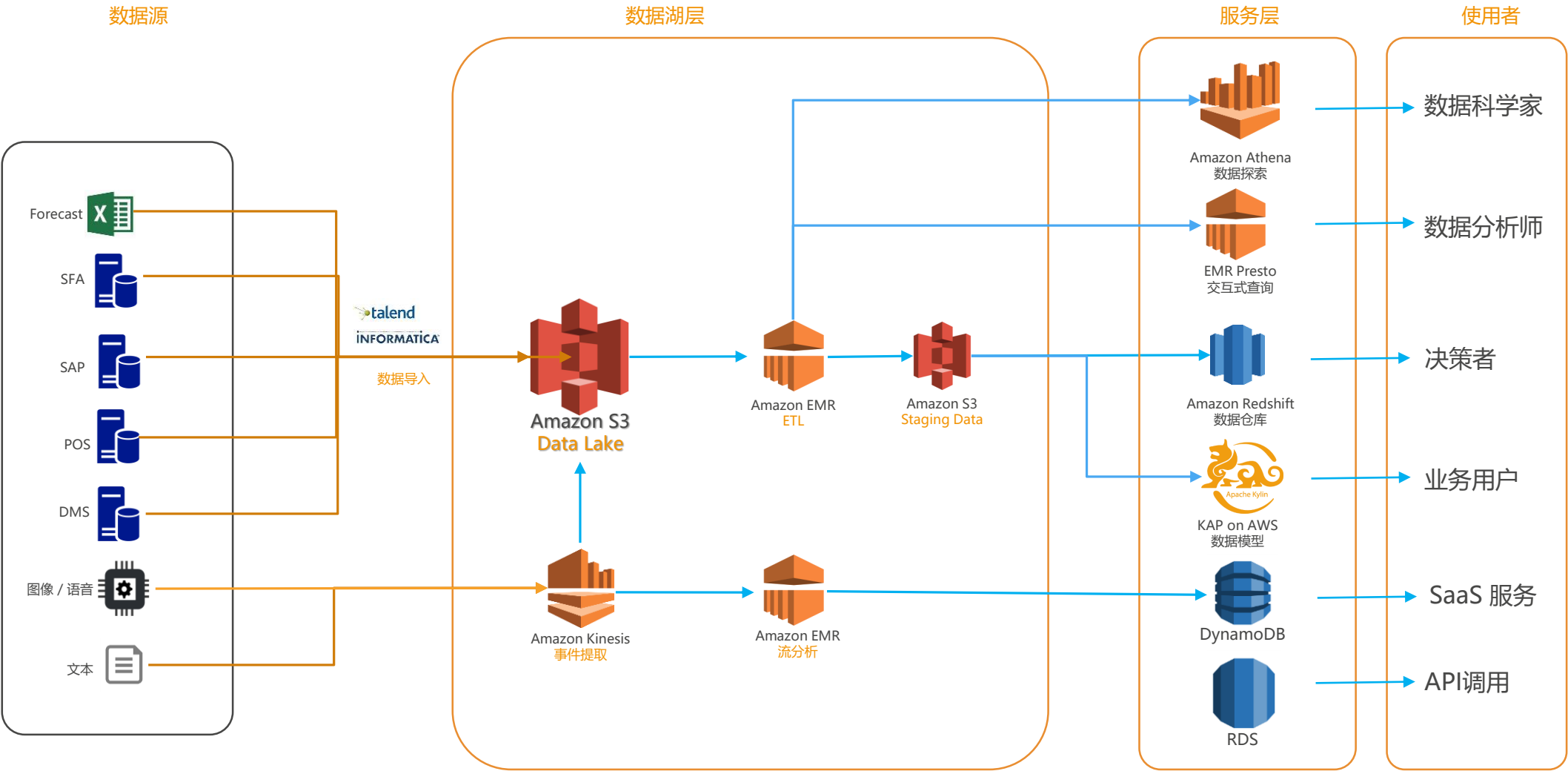
KAP on AWS



快消行业大数据平台架构（1）：数据湖



快消行业大数据平台架构（2）：实时数据分析



平台架构优势

单一数据源



- 通过建立数据湖，各业态的信息都可集中至统一的数据平台进行储存与分析，形成**Single Source of Truth**

丰富场景



- 传统数据平台不能有效支持**非结构化数据、海量实时数据分析**
- 未来将有**更多场景**需要利用数据驱动商业决策与运营

灵活扩展



- 随着业务发展，各类**数据必将井喷**，传统数据平台将无法支撑
- 数据湖利用**S3**为底层的**分布式处理**技术，可横向**无限扩展**来更好的支持企业未来业务发展

成本更优



- 数据湖的每TB储存价格仅为传统数仓的**20-40%**
- 随着**硬件价格走低**，大数据湖的**成本优势更加明显**

能力建设



- 大数据技术**是未来的趋势，现在利用好前沿技术有利于企业打造技术驱动的**差异化竞争能力**
- 大数据技术**不再是纸上谈兵**，逐渐将成为主流

数据湖设计最佳实践

数据分层设计



CUBE

数据模型层

- 多维分析模型1
- 多维分析模型2



Data Mart
Data Warehouse
Staging

数据仓库层

- 数据集市
- 数据仓库
- 数据加工



Amazon Glacier

Archived Data



Amazon S3
Data Lake

Raw Data
Staging Data
Result Data

数据湖层

- 原始数据
- 数据校验、清洗后的数据
- 计算、聚合后结果数据

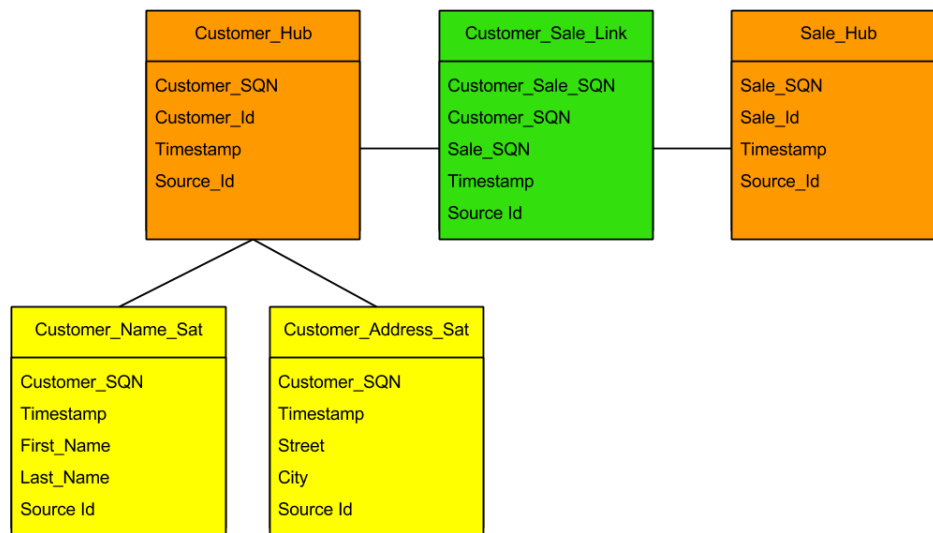
DW – Data Vault 设计

• 定义

Data Vault由Dan Linstedt在20世纪90年代提出 (<http://www.danlinstedt.com>)。是面向细节的，可追踪历史的，它是一组有连接关系的规范化的表的集合。这些表可以支持一个或多个业务功能，它是一种综合了第三范式 (3NF) 和星型模型优点的建模方法。其设计理念是要满足企业对灵活性、可扩展性、一致性和对需求的适应性要求，它是一种专为企业级数据仓库量身定制的建模方式。

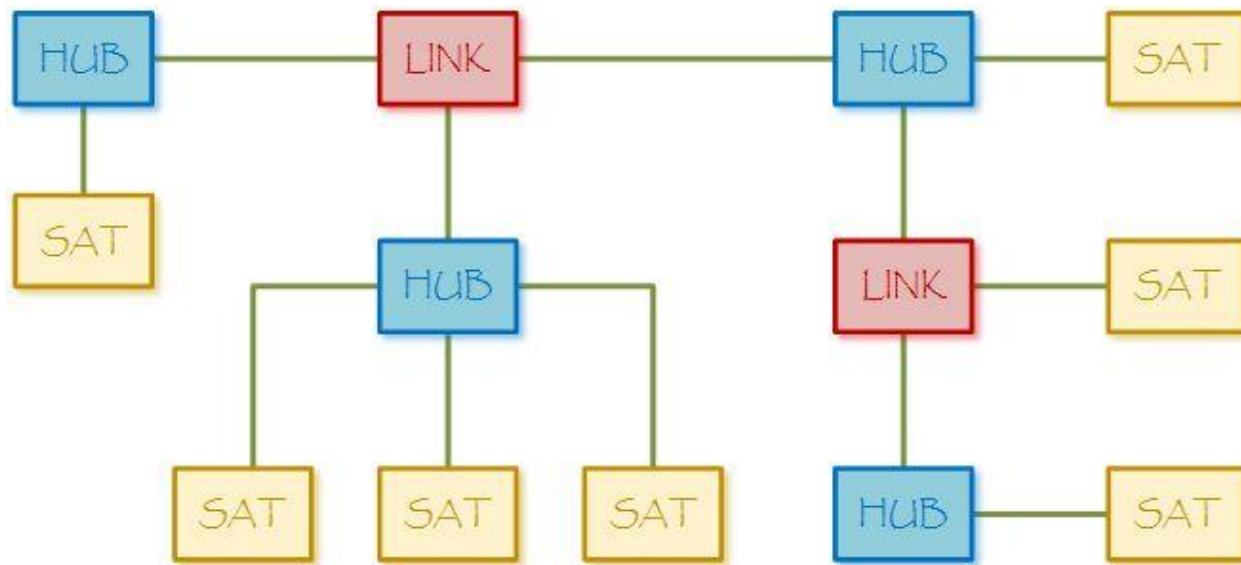
• 模型组成

- Hub: 独立数据域的核心属性（如：商品编码）
- Sat: 围绕Hub的附加属性（如：商品尺寸颜色）
- Link: Hub之间的关系（如：商品Hub与客户Hub构成了订单的关系）

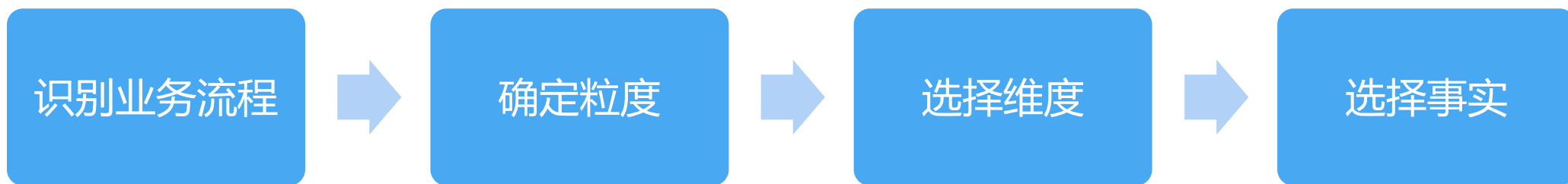


DW – Data Vault 设计

- 优势
 - 结构灵活，易于扩展，可以自由地增删
 - 可以实时加载数据
 - 可以方便地追溯历史
 - 由于不同表之间的解耦，ETL的任务可以并发执行，效率提升
 - 很方便地转换为维度模型



DM – Dimensional Model 设计



识别业务流程

Adventure Works Cycles: Enterprise Data Warehouse Bus Matrix Entries for Sales Department											
	Dimensions										
Business Process	Date	Product	Employee	Customer (Reseller)	Customer (Internet)	Sales Territory	Currency	Channel	Promotion	Call Reason	Facility
Orders Forecasting	X	X	X	X	X	X	X	X			
Orders	X	X	X	X	X	X	X	X	X		
Inventory	X	X	X								X
Call tracking	X	X	X	X	X	X				X	
Returns	X	X		X	X	X	X		X		X

确定粒度

业务流程的粒度及基础度量			Date	Product	Vendor	Contract Terms	Employee	Warehouse	Carrier
Business Processes	Atomic Granularity	Metrics							
Purchase Requisitions	1 row per requisition line	Requisition Quantity & Dollars	X	X	X	X	X		
Purchase Orders	1 row per PO line	PO Quantity & Dollars	X	X	X	X	X	X	X
Shipping Notifications	1 row per shipping notice line	Shipped Quantity	X	X	X		X	X	X
Warehouse Receipts	1 row per receipt line	Received Quantity	X	X	X		X	X	X
Vendor Invoices	1 row per invoice line	Invoice Quantity & Dollars	X	X	X	X	X	X	
Vendor Payments	1 row per payment	Invoice, Discount & Net Payment Dollars	X	X	X	X		X	



维度中粒度表示法

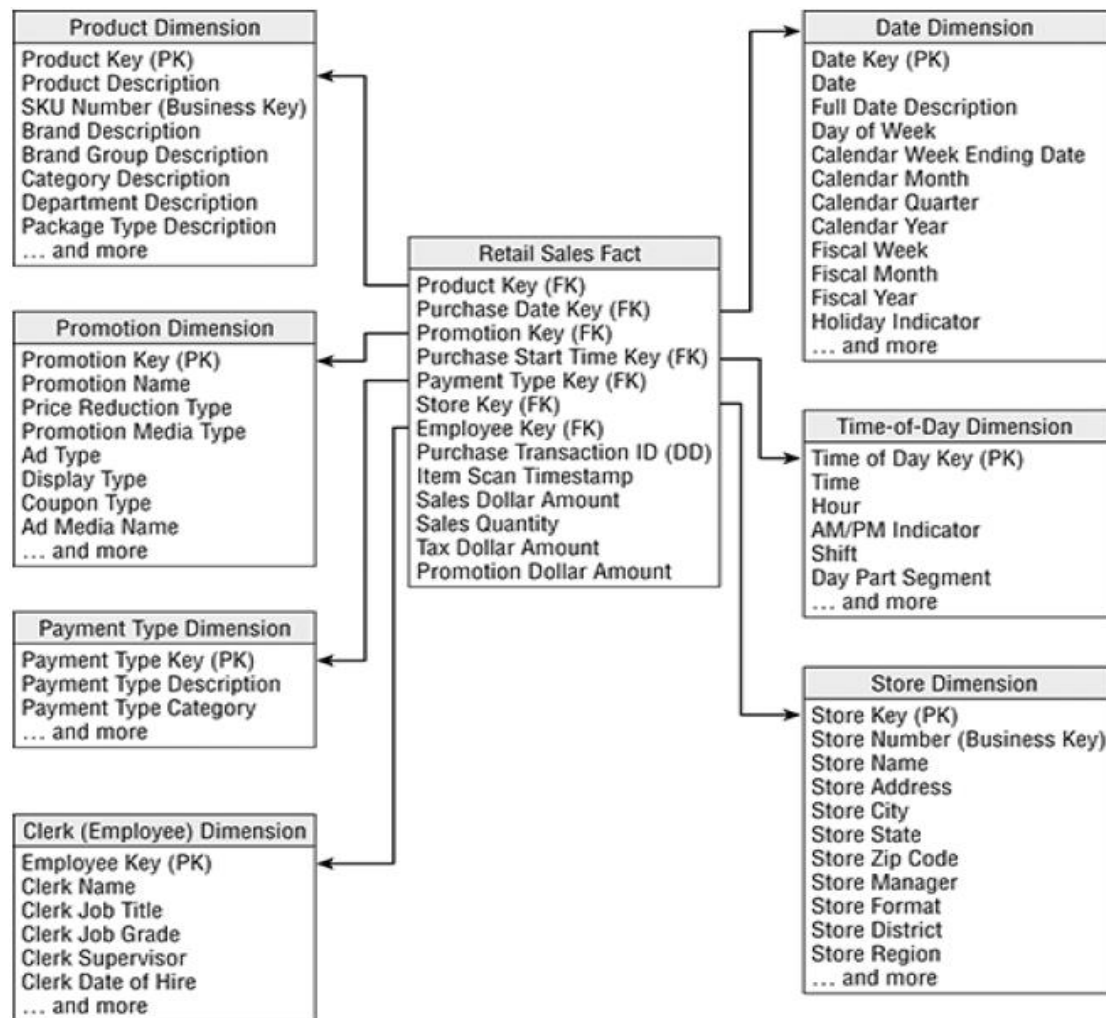
	Date	OR	Date	
			Day	Month
Issue Purchase Orders	X		X	
Receive Deliveries	X		X	
Inventory	X		X	
Retail Sales	X		X	
Retail Sales Forecast	X Month			X

为每个业务流程定义详细维度模型



根据维度模型详细定义进行数据仓库逻辑设计

- 维度表
- 事实表
- 星型架构



客户案例

欣和AWS数据湖项目



挑战

- 快速增长的业务规模需要更加高效、准确和灵活的数据分析能力
- 企业内部生产、营销、终端客户系统繁多，没有统一的数据分析平台，造成企业内部的信息壁垒
- 业务底层数据模型不统一，导致报表口径不一致，影响领导层决策

方案

- 设计数据湖架构和数据仓库应用
- 重新架构业务底层的数据模型
- 使用Kylin作为后端核心分析应用的数据引擎支持高并发近实时查询服务
- 设计数据生命周期管理方案减少资源费用

优点

- 将ERP，CRM，供应商门户，电商平台等数据源导入到Amazon S3服务，形成单一可信数据来源
- 优化的底层数据模型可以极大降低上层业务逻辑修改导致的大量重复开发工作
- 利用Amazon Redshift数据仓库的强大计算能力将每日报表的处理时间从数个小时减少到不到1小时
- 有效支持了欣和全国数十万各级渠道销售人员，形成快速的营销响应机制

在决定重构整个 IT 系统时，我们期望实现的目标是让技术平台快速匹配业务的需求，打造高安全性、高可用性且具备弹性伸缩能力的 IT 系统，让 IT 系统与业务系统实现最佳融合。AWS 云平台强大的大数据分析处理能力、对数据的全方位保护、开放性及丰富的合作伙伴解决方案是欣和选择 AWS 的主要原因。在AWS 的帮助下，欣和正在成为一家数字驱动型企业。

Company: 烟台欣和企业食品有限公司
Industry: FCMG
Country: 中国
Employees: >10000
Website: <http://shinho.com.cn/>

关于欣和:

烟台欣和企业食品有限公司创建于1992年，致力于为每个家庭提供高质量的调味和休闲食品。经过26年的不断发展，欣和已成为一家拥有十多家子公司的大型食品集团。所提供的产品包括原酿酱、顶级的高盐稀态酱油、美味方便的花生、辣椒制品、有机食品、味噌等，“六月鲜”、“葱伴侣”、“黄飞红”这些耳熟能详的品牌都出自欣和。产品远销世界100多个国家和地区，年销售额超过40亿元。

欣和CIO 国庆义

痛点

1. 内外部系统繁多，没有统一的数据分析平台，并造成企业内部的信息壁垒



2. 数据资产广泛但是数据质量有差异，缺乏统一标准



3. 数据处理能力有限，无法适应迅速增长的业务需求



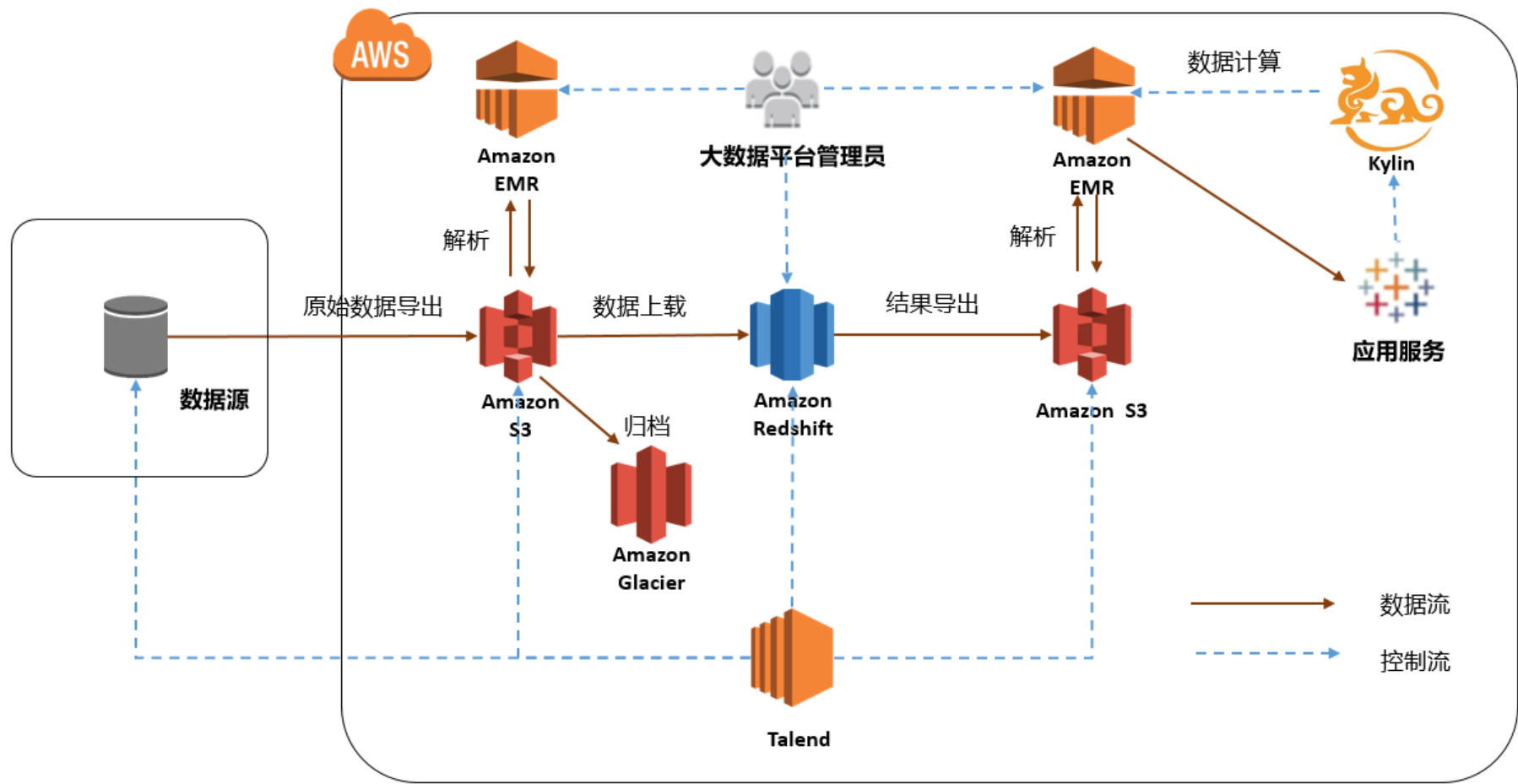
4. 底层数据模型不统一，导致最终报表口径不一致，影响领导层决策



5. 企业IT人员技能需要提升，业务理解力与技术实现力不能有机结合



项目实施——欣和数据分析平台架构



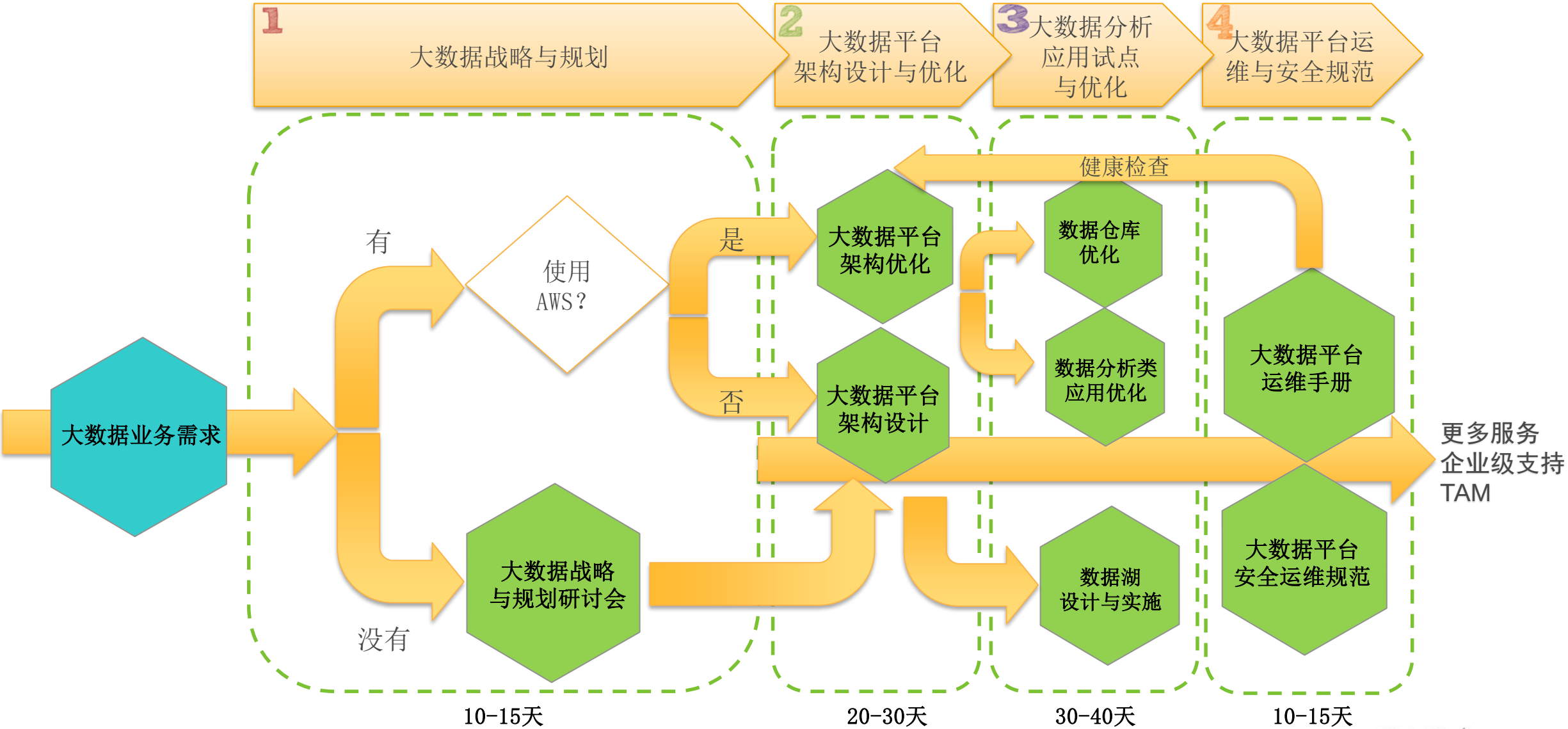
案例链接

<https://aws.amazon.com/cn/solutions/case-studies/shinho/>

<https://aws.amazon.com/cn/solutions/case-studies/china/>

实战演示

AWS大数据咨询标准服务内容



AWS专业服务团队的实践领域



咨询和IT转型

定义您的云计算方法 (从战略到变更管理再到技术), 并提供来自我们 IT 转型主管的专家指导。



安全与合规性

基于成熟的、经过验证的设计制订和实施安全策略和控制措施, 满足您内部和外部的合规要求。



迁移和自动化

加快迁移至 AWS 云的速度, 并降低实施风险。



运营集成

在您的组织内应用 DevOps 和敏捷实践、集成身份管理、自动化、资源监控等, 以实现团队的快速创新。



大数据、AI和IoT

利用适用于大数据、AI和IoT的 AWS 服务和专业知识发掘新洞见和商机, 以帮助您的架构、集成或部署数据采集、处理与分析平台。

Thank You!