



电信运营商数据湖技术实施策略

胡军军, 谢晓军, 石彦彬, 喻琦

(中国电信股份有限公司广州研究院, 广东 广州 510630)

摘要: 首先, 从分析电信运营商传统数据架构入手, 识别当前面临的数据质量不高、实时性不够、灵活性不足及存储和应用相互制约等问题。然后, 讨论引入数据湖技术的可行性, 明确数据湖的存储规模化和低成本、数据“原汁原味”和方便易用、应用按需建模等特点, 能够为电信运营商数据架构优化提供非常有益的参考。最后, 提出数据统一存储、统一标准、近源采集、与应用分离的电信运营商数据湖建设方案, 并针对数据湖数据分区、部署、入湖及应用数据动态加载、数据统一管理等相关要点给予明确的阐述。

关键词: 数据湖; 原生数据存储; 存储计算分离; 数据目录

中图分类号: TP399

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019029

Implementation strategy of telecom operators' data lake technology

HU Junjun, XIE Xiaojun, SHI Yanbin, YU Qi

Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: Firstly, the traditional data architecture of telecom operators was analyzed and the problems such as low quality, lack of real-time, lack of flexibility and mutual constraints of storage and application were identified. Then, the feasibility of introducing data lake technology was discussed, the storage scale and low cost of data lake, the originality of data and ease of use, and the application of on-demand modeling were clarified, which could provide great benefits for telecom operators' data architecture optimization. Finally, the data lake construction plan of unified storage, unified standards, near-source acquisition, and application separation was proposed, and key points such as data lake data partitioning, deployment, lake entry and application data dynamic loading, and unified data management were clearly stated.

Key words: data lake, native data storage, loose coupling between storage and computation, data catalog

1 引言

云计算、大数据、人工智能、物联网、移动互联网等技术的发展, 使得人们有能力以低成本

的形式记录、还原、映射和支撑商业世界的运行。数据可以随时在线, 成了连接云、网、端的核心; 数据也成为和货币、石油一样的资产, 是信息经济中的核心资源。以控制为出发点的 IT 时代, 正

收稿日期: 2018-10-08; 修回日期: 2019-01-22

在走向以激活生产力为目的的 DT 时代^[1]。

随着电信运营商网络智能化、业务生态化、运营智慧化步伐的加快, IT 系统的建设越来越强调提供面向客户的极致服务、面向业务的敏捷使能、面向网络的智慧运营能力。构建开放数字架构(如 TMFODA^[2]), 开放网络和业务能力(如 TMFOpenAPI^[3]), 实现网络即服务、支撑数字化转型是运营商的共识, 其中数据和服务开放是运营商提升竞争力的两大核心利器。

传统的电信运营商 IT 系统以功能性开发为主, 以数据分析应用为辅, 两套体系并行推进支撑企业运营和分析决策, 基本实现业务数据化。而如何加速推进以数据为中心的 IT 转型, 实现数据业务化^[1], 消除运营商内外部信息壁垒, 提供高质量、高可用、高实时性的数据, 驱动业务和应用快速创新, 有效支撑企业数字化转型和网络重构战略, 是运营商面临的重大挑战。

数据湖的概念是由 Pentaho 公司的 CTO Dixon^[4-5]率先提出的, 其定义为“未经处理和包装的原生状态水库, 不同源头的水体源源不断地流入数据湖, 为企业带来各种分析、探索的可能性”。数据湖推崇 Schema on Read 模式^[6], 强调数据无需加工整合, 可直接堆积在平台上, 最终由

使用者按照自己的需要进行数据处理。与传统的数据架构要求整合、面向主题、固定分层等特点不同, 数据湖为企业全员独立参与数据运营和应用创新提供了极大的灵活性, 并可优先确保数据的低时延、高质量和高可用, 给运营商数据架构优化提供了很好的参考思路。

2 电信运营商数据架构的现状与挑战

从数据的系统归属上看, 运营商数据可分为 MSS(管理支撑系统)的面向人、财、物管理类数据, BSS(业务支撑系统)的面向客户和产品的营销及客户服务数据, OSS(运营支撑系统)的面向产品和网络的功能及运营服务数据, 三者之间既相对松耦合, 又有着紧密的协作关系, BSS 和 OSS 的衔接点主要在产品及开通、排障服务, MSS 和 BSS、OSS 的衔接点主要在参与人和资源。从数据分类来看, 运营商的数据可分为作为企业核心的功能类实体数据、表示企业所有运营过程的活动类数据、体现内外部客户感知并围绕两大主线所产生的感知类指标数据以及与管理相关的人、财、物及流程数据^[7-8]。电信运营商数据范围示例如图 1 所示。

由于国内运营商以两级经营模式为主体, 系

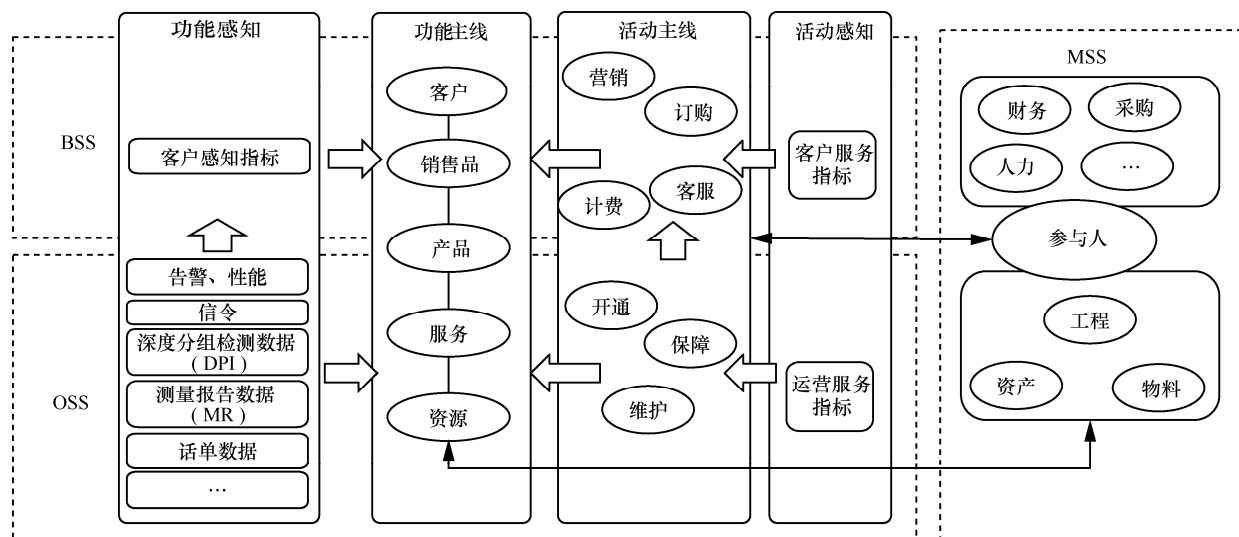


图 1 电信运营商数据范围示例



统的集约化建设程度相对较低,以分域(M/B/O)、分省建设为主,即便是同类系统的数据,因为分31个省市建设,各省市的业务管理模式、数据模型标准、主数据等千差万别,跨省、跨域、跨系统的模型标准统一非常困难,即便通过数据副本的模式进行整合汇聚,也存在转换不专业和数据失真等问题。同时,域与域之间虽是松耦合的,但因为使用者和建设者的不同,相互之间会冗余存储对方数据,而建模和主数据又不同,跨域之间数据的关联整合非常复杂,跨域、跨省的端到端应用困难。

运营商的数据还有一个显著的特点,就是与网络密切相关,网络运行数据和网络拓扑数据需要与网络保持实时一致,且数据量比较大,网络智能化后的实时数据应用需求也越来越多。通信网络是一张大网,即便引入云计算、虚拟化技术,依然有大量网络节点遍布31个省市,海量网络数据的实时采集、处理及应用也是运营商数据架构需要考虑的一个重要因素。

国内运营商目前都不同程度地建立了自己的企业级大数据平台,有的分总部/省两级部署,支撑两级数据分析,统一全网的架构、来源、算法、规则,总部数据轻度汇总,按需采集汇聚高价值

详单数据;有的采用1+N模式,建设总部和省互补协作平台,总部提供跨域数据和特定的大数据能力,作为N的省向总部提供本地化数据能力与自定义算法。电信运营商数据平台架构示例如图2所示。

不管采用哪种模式,都不同程度地存在其下属各专业公司、各部门根据各自需要,或在生产系统内构建含大数据技术的混搭数据架构,或建设域内自用的大数据平台,因此有很多数据未进入企业级大数据平台,或数据平台的应用未达到预期。其原因可归结为如下几点。

(1) 平台数据质量不高

平台数据来自于M/B/O的生产系统,而运营商分两级31省市建设的生产系统,不但数据模型、主数据标准不统一,业务管理模式的差异也很大。数据经过多次模型转换,存在严重失真的问题,且很难对数据质量问题追踪溯源。

(2) 平台数据不够实时

数据经过多级采集汇聚,处理环节多,采集周期长。网络相关海量数据跨省传输,占用大量带宽,数据时延较大。数据平台目前只能以支撑离线的决策分析为主,难以满足SDN/NFV/云网络及物联网等实时/准实时数据应用需求。

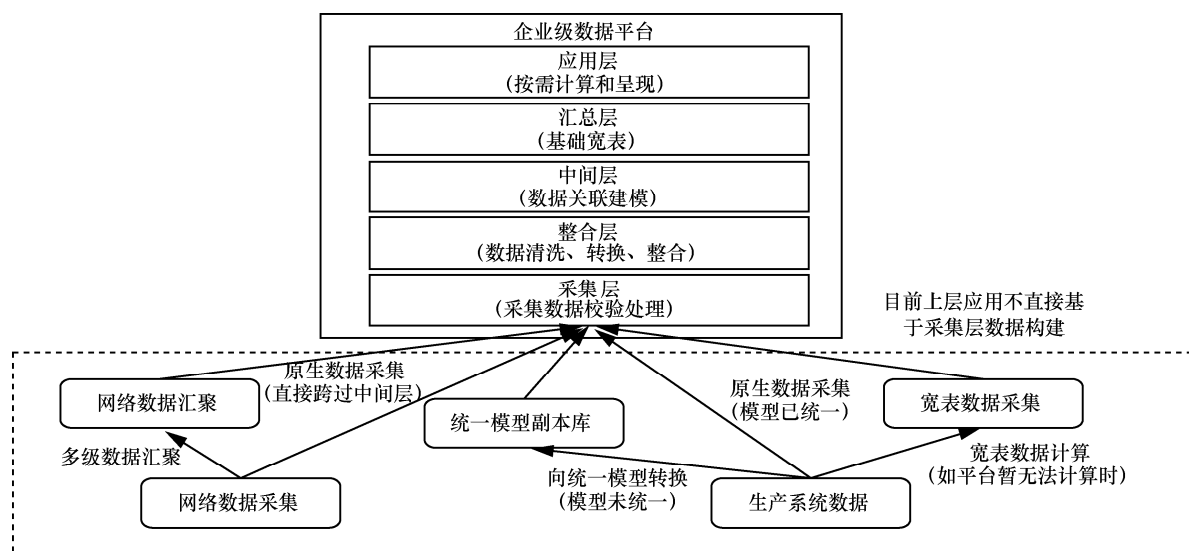


图2 电信运营商数据平台架构示例

(3) 平台的灵活性不足

数据平台的建设以存储计算一体化架构为主,平台与应用紧耦合,多基于公共数据平台和整合后的数据支撑应用创新。对于新的数据整合、数据计算分析技术引入、平台扩容支撑等需求响应不灵活,导致数据平台应用不足。

(4) 平台和应用互锁,形成恶性循环

企业级数据平台难以满足生产系统数据应用需求,生产系统就没有动力将自身数据和应用迁入数据平台,进而数据平台的数据质量和可用性越来越差。同时,还导致生产系统和各个大数据平台的数据重复采集、重复存储,且相互之间数据访问技术和管理壁垒严重,建设和维护成本大幅提高。

随着运营商网络智能化、业务生态化、运营智慧化战略的逐步落实以及 SDN/NFV/云/物联网等新网络的不断引入,其数据业务化^[1]的需求越来越强烈,不但要强化已有的精准营销和服务、精确管理、精益网运等分析决策类数据应用,还需要以数据贯通整个生产流程,注智 M/B/O 系统,构建开放的数字化架构,支撑企业的智慧运营,实现网络的集约、智能、自动化运维。虽然国内运营商均已走在以数据为中心的 IT 架构转型的路上,但依然面临着数据标准不统一、数据质量不高、跨域数据衔接整合困难、数据实时应用支撑不足、数据应用创新灵活性不够等问题。同时,其数据架构还需考虑运营商特有的遍布 31 省市的通信网络数据实时采集、处理和汇聚需求。

3 数据湖方案的价值及可行性分析

数据湖推崇存储原生数据,对不同结构的数据统一存储,使不同数据有一致的存储方式,在使用时方便连接,真正解决数据集成问题。数据湖的本质是一种数据管理的思路,利用低成本技术来捕捉、提炼和探索大规模、长期的原始数据存储的方法与技术。数据湖可存储任何种类的数据,高质量、高效率地存储数据,更快速、更廉价地处理数据,

将建模应用问题丢给最终开发者^[9]。

数据湖的方案应用可以带来如下几个显著的好处。

(1) 规模大、成本低

全企业海量数据统一存储,采用开源技术,基于低成本硬件资源,建立和维护成本相比数据仓库低一个数量级。

(2) 数据“原汁原味”

数据湖以原始形式保存数据,并在整个数据生命周期捕获对数据和上下文语义的更改,尤其便于进行合规性和内部审计。如果数据经历了转换、聚合和更新,将很难在需求出现时将数据拼凑在一起,而且几乎没有希望确定清晰出处。

(3) 数据方便易用

结构化、非结构化、半结构化的数据都是原样加载和存储,以后再进行转换,开发和保存成本低,产生和使用之间时延小。客户、供应商和数据运营者不需要数据拥有者提供太多帮助即可整合数据,消除了数据共享的内部政治或技术障碍。

(4) 应用按需建模

数据湖提供数据给灵活的、面向任务的结构化应用,详细的业务需求和艰苦的数据建模都不是数据湖的先决条件。数据湖给予最终用户最大的灵活度来处理数据,对于同一份原始数据,不同的用户可能有不同的理解。

目前,大部分运营商采用传统的以数据为中心的处理架构(存储计算一体化,如主流 MPP、Hive 和分布式计算厂商产品),好处是计算效率高、技术成熟,缺点也很明显,如灵活性不足,使得数据应用适用于少数人,这也制约了原生数据提供者向平台提供的积极性,进而导致数据的质量、数据的全面性都得不到很好的保障。引入数据湖概念的一个显著特点就是存储和计算松耦合,可采用以计算为中心的处理模式(存储与计算分离,如 Spark 技术及 AWS、阿里云等云服务提供商产品),使得运营商可以更加专注于数据



的存储和管理,存储和计算不用相互制约,从而优先确保数据的高质量、低时延、高可用,并为数据应用的快速构建提供了极大的灵活性。

综上所述,数据湖强调原生数据存储、事后绑定建模、统一数据管理,数据湖的建设不受限于数据应用的建模体系,提供了一个全新的数据管理思路。数据湖的理念可很好地实现规模化、低成本的原生数据存储,保存最“原汁原味”的实时有效数据,消除数据共享的政治和技术壁垒,规避数据存储和应用建模的相互制约,降低应用开发门槛,为运营商构建以数据为中心的 IT 架构提供了很好的参考,可以有效解决运营商数据架构面临的质量不高、预先建模、应用门槛高、成本昂贵等难题。

数据湖按照成熟度可划分为 4 个阶段:第一个阶段,应用程序独立建设,部分应用将数据提供给数据仓库,基于数据仓库构建分析应用;第二个阶段,数据湖和数据仓库并存,应用程序向数据湖提供副本数据,基于数据湖开发分析型应用,数据仓库和应用也可从数据湖提取数据;第三个阶段,新系统以数据湖为中心构建,应用通过数据湖交互彼此数据,数据湖成为数据架构的核心,数据仓库基于数据湖提供特定的应用需求,数据治理变得重要;第四个阶段,所有新的应用均基于数据湖构建,数据湖成为弹性的分布式平台,数据的治理和安全需持续加强,支撑企业的数据运营和分析能力^[10]。电信运营商目前普遍处于第二个阶段向第三个阶段演进的过程中,在构建数据技术方案方面具备较好的基础条件。

4 电信运营商数据湖建设思路及实施要点

调整现有分析型数据平台建设思路,将其数据与应用解耦,引入数据湖概念,强调原生数据入湖,并与全网生产系统模型和主数据标准化协同推进,兼顾层次化的传统数据架构和扁平化的数据湖架构的优点,Schema on Read 和 Schema on Write^[6]并存,统一支撑企业实时、准实时和离线

数据应用快速创新,是电信运营商实现以数据为中心 IT 架构转型的有效途径。

数据湖作为运营商数据存储和访问的唯一出口,成为所有 IT 系统共享的基础设施,统一存储全企业 IT 和网络数据,通过开放架构支撑智慧运营,并可作为 IT 系统集约化演进的纽带。

(1) 数据统一存储

统一存储 MSS、BSS、OSS 及网元平台的实时、历史、在线、离线数据,全网的原生数据只存储一份在逻辑统一的分布式数据湖内,原生数据与生产系统数据模型标准和主数据一致,新 IT 系统/网元平台的生产数据直接使用数据湖存储。

(2) 数据统一管理

所有入湖数据的目录、元数据、数据应用及数据质量、数据标准、数据安全必须统一管理。数据模型标准和主数据动态维护,数据质量集中治理,原生系统的数据问题溯源处理,生产系统建设者全程参与数据管理,责任权利保持一致。

(3) 数据统一标准

生产系统管理部门负责 31 省市系统模型和主数据的标准化;数据湖统一管理生产系统的数据模型及主数据;暂未进行标准化的生产系统数据模型,由对应系统的管理部门负责数据模型的转换和运营,协调推进生产系统数据标准进程。

(4) 数据近源采集

提供数据统一采集、实时订阅分发框架,支撑实时/准实时数据、离线数据的采集。各网元/平台数据采集能力以组件方式纳入数据湖,分专业采集、预处理加工,海量实时数可靠近网络近源部署前置采集模块。非网络类数据(如 BSS、MSS、OSS 流程等),初期以副本采集方式汇聚入湖,远期直接以服务交互方式入湖。

(5) 数据与应用分离

数据应用环境与数据存储环境分离,按应用计算的带宽需要就近部署。提供统一的服务化访问、小批量数据订阅、数据分析计算云平台

环境。基于云平台环境,应用开发者可自行整合数据、构建应用,数据存储、数据整合、平台组件、数据应用间相互解耦,建设的进程不会相互制约。

同时,建立全生命周期数据目录,统一标识各项数据,完善数据治理机制,管理数据湖数据的生产加工流程,对各项数据生成和使用过程进行跟踪记录,支撑数据的应用和溯源,是数据湖方案顺利实施的关键要素。并且还需要加强数据标准的全生命周期流程以及数据标准的元数据及数据质量问题收集、自动稽核、问题溯源、影响分析及跟踪处理等数据管理能力。可以采用爬虫的方式生成数据目录,在不影响数据所有者或用户的情况下自动生成,业界的工具如 Google 的 Goods^[11-12]、AWS Glue Data Catalog^[13]等;也可以由数据的所有者或用户主动地将数据变化写入数据目录。电信运营商数据湖架构示例如图3所示。

决定数据湖能否顺利实施的因素有很多,包括数据湖涵盖哪些数据及如何分区存储、数据湖如何分布式部署、纷繁复杂的现有IT系统数据如何入湖、数据和应用能否分离、数据湖与现有各

类数据平台的演进关系等。当然,更重要的是数据管理思维的转变,这是一切的基础。针对电信运营商数据湖的实施,提出如下4个方面的关键要点及建议。

(1) 要点1: 数据湖分区

数据湖逻辑上可划分为生产数据区、原生数据区、整合数据区、汇总数据区4个大的存储区域。数据湖的应用可基于PaaS平台按需使用各个区的数据,4个区的数据目录、元数据、数据加工处理流程及数据应用需要统一管理、维护和治理。

- 生产数据区。M/B/O系统生产数据的存储区域,涵盖实时交易型数据、实时/准实时网络采集数据等,可以是关系型和非关系型混搭的存储结构,各生产系统需要进行架构优化,数据与应用分层解耦,将数据存入生产数据区。
- 原生数据区。将各系统的生产数据直接写入数据湖原生数据区,以非关系型数据格式存储生产系统数据,方便各数据应用使用,生产数据和原生数据模型标准、主数据一致。原生数据区涵盖企业的任何内容,

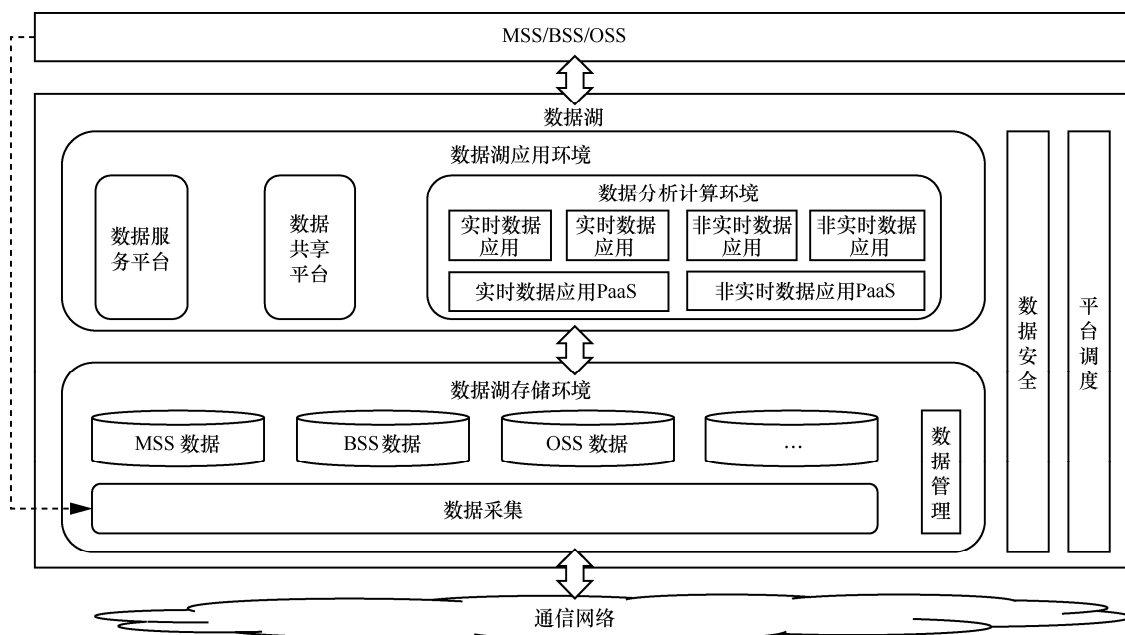


图3 电信运营商数据湖架构示例



无限接近企业各系统、部门的敏感信息。供数据湖科学家和技术人员访问使用。

- 整合数据区。存储按照数据分析需求建模加工后的公用数据。模型从生产/原生数据模型派生而来，被业务和 IT 部门熟知，可供企业各种应用程序使用。原生数据区中依然有很多数据或属性没有被真正理解，并未完全包含在这个数据区的模型中。
- 汇总数据区。存储按需求分析汇总的结果数据，一般可存储在关系型数据存储内，便于数据服务的快速加载呈现。

数据湖生产数据区和原生数据区作为最重要的数据分区，是数据湖内数据整合和汇总的源头数据，数据质量必须得到保障。另外，数据湖虽不鼓励应用特定模型，但也可划分特定数据区给私有应用使用，提供快速构建数据应用的途径，这些应用获取数据湖数据且具有数据处理能力，数据湖构建初期，可将已有业务应用数据导入数据湖特定数据区中。电信运营商数据湖数据分区示例如图 4 所示。

(2) 要点 2：数据湖部署

数据湖部署方案的设计需要考虑如下要素：现有 BSS/OSS 系统分省/总部两级建设和维护，源系

统模型属地管理；网络/平台数据量大，且贴近网络建设归属地，属地应用占比大；M/B/O 及网络/平台之间数据松耦合，主要通过企业主数据进行衔接。

数据湖原生数据区和生产数据区与数据源系统就近分布式部署（总部 1+省市 31 模式）。生产数据云节点由生产系统按需分区、分片部署，即支撑生产应用交易处理，也支撑实时网络数据采集和应用。原生数据云节点与生产数据云节点就近、集中部署，靠近数据归属地，数据实时从生产数据云节点写入原生数据云节点。原生数据云节点可再细分为核心数据区（如客户、销售品、产品、服务、资源、组织、人员等）、BSS 数据区、OSS 数据区、MSS 数据区、网络/平台数据区。

数据湖整合、汇总数据云节点采用 1+N 模式部署，统一管理、控制和调度节点环境，兼顾全网统一和个性化应用需求，数据科学家逐步探索和建模数据，开放数据应用。1+N 模式中的“1”支撑全网应用，“N”支撑省内应用，并作为创新基地，有条件、数据量大、应用丰富的省可选择建设 N 分区。分区节点内可按照应用范围（全局需求、特定需求）、地域归属（集团、省）、数据层次（整合、汇总）、数据分级（普通、密级）等进一步分区存储。电信运营商数据湖部署方案示例如图 5 所示。

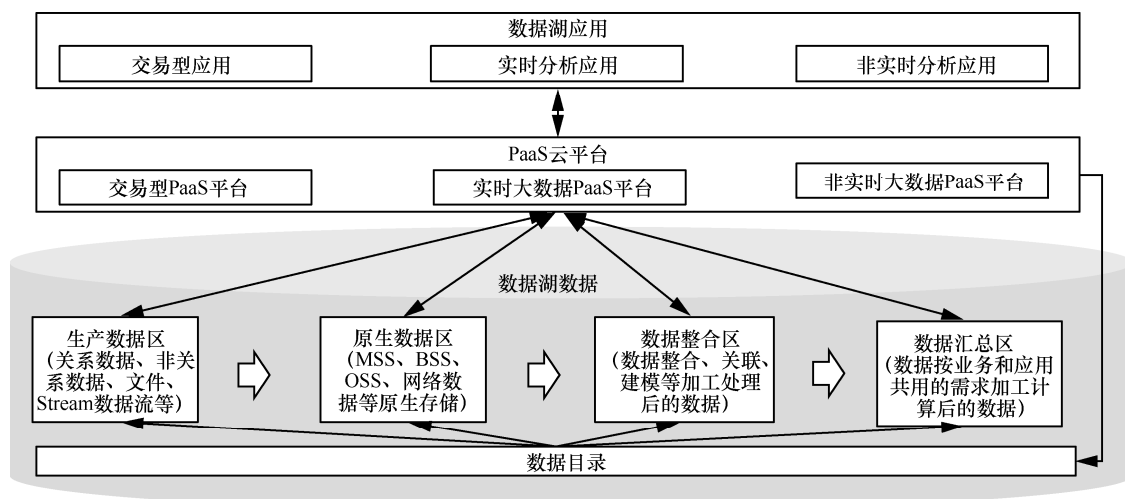


图 4 电信运营商数据湖数据分区示例

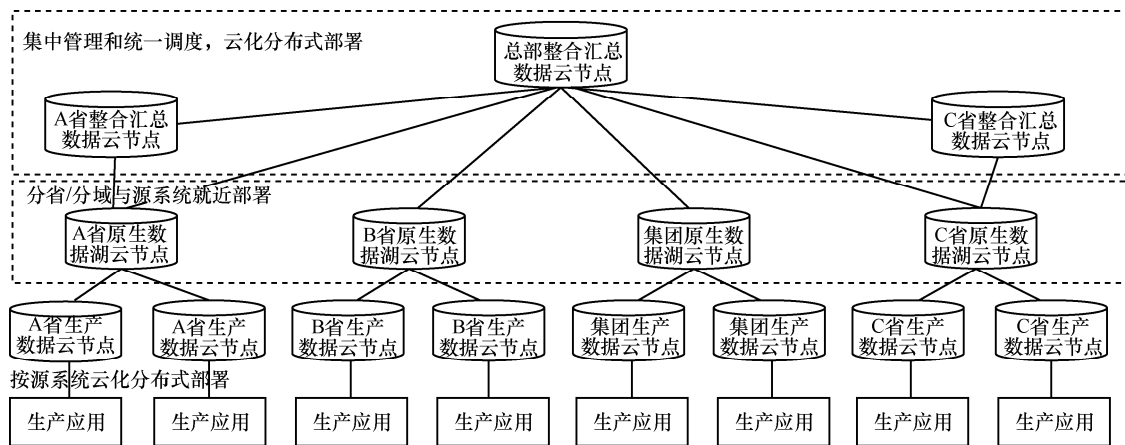


图5 电信运营商数据湖部署方案示例

(3) 要点3: IT 系统数据入湖

数据湖的建设不可能一蹴而就,需要根据运营商 IT 系统建设情况分别采用不同策略进行数据入湖演进。电信运营商 IT 系统入湖方案示例如图6所示。

- 方式一: 数据同步方式。适合交易型系统已存在、数据模型和主数据已全网统一的场景,生产数据直接同步写入原生数据区,如 BSS、MSS、传统 OSS。
- 方式二: 数据同步/转换方式。适合交易型系统已存在、数据模型和主数据并未全网统一的场景,如 BSS、MSS、传统 OSS。将非标准生产数据写入原生数据区,支撑省内整合汇总应用及集团标准的宽表需求;将非标准生产数据按全网统一标准转

换,提供给全网数据整合汇总及数据治理使用。

- 方式三: 数据正本方式。适合交易型系统新建模式,如新一代 OSS 资源、编排、告警等。正本数据写入生产数据区,统一模型和主数据标准,基于交易型 PaaS 平台完成应用;生产数据区数据直接写入原生数据区。
- 方式四: 采集入库方式。适合网络监控分析型系统新建模式,如新一代 OSS 的网络采集数据、资源拓扑、深度分组检测(DPI)数据等。数据采集文件、流数据等暂存在生产数据区;写入原生数据区后,生产数据区不再保留;统一原生数据模型和主数据标准,基于实时和非实时 PaaS 平台完成分析型应用。

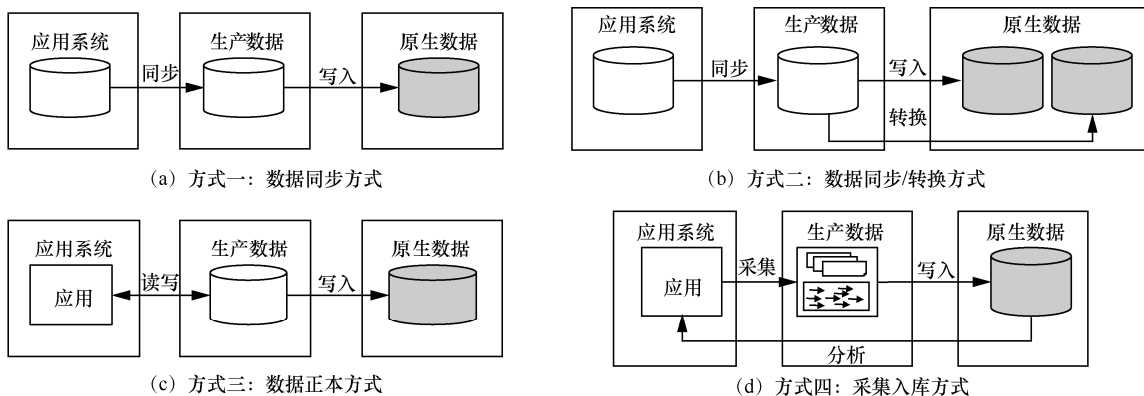


图6 电信运营商 IT 系统入湖方案示例



(4) 要点 4: 数据湖数据与应用分离

数据湖通过数据服务平台、数据共享平台及统一数据应用环境按需支持交易类、实时监控类、分析类应用。数据增、删、改、查服务统一部署在数据服务平台上,供交易类应用访问调用;通过订阅需要监控的数据,由数据共享平台将数据实时分发给监控类应用使用;数据的加工整合、分析应用、海量搜索、人工智能等应用均可部署在应用环境内,按需动态加载并临时存储数据,结果写回到数据湖存储环境,以服务方式启动任务和查询结果数据。其中,应用环境公共组件随着技术的更新不断叠加,逐渐平台化共享,暂时无法满足应用需求的可由应用在统一环境内部署组件及加载数据。

数据湖应用加载数据的方式可分为实时增量加载、准实时增量/全量加载、离线批量加载等,数据可按需全量或增量短期加载。对于应用和数据无法解耦的组件(如 Hive、MPP 等),按需复制数据,以空间换数据管理和应用的灵活性;对于应用和数据可以有效解耦的组件(如 Spark 等),可以按需动态、实时加载数据。应用组件逐渐由与数据紧耦合的组件向与数据松耦合的组件演进。

数据湖采用读写分离、应用计算与数据存储分离、关系数据与非关系数据存储并存的模式,并提供数据存储节点分布式部署、服务化访问及统一数据加载、共享及分发能力,降低数据湖数据存储访

问负载,提升数据的可用性及数据访问效率。由数据湖提供数据的统一迁移,包括主从库的复制、关系库到非关系库的数据转换等;提供统一的关系和非关系库数据访问及分布式数据路由以及数据共享开放和订阅分发管理框架,实现高效的数据访问;提供统一的数据应用环境管理,包括配额管理、数据访问权限管理、数据回写节点分配管理等,独立部署分析计算类应用,分析计算节点与数据湖数据存储节点分离;提供统一的分布式服务运行框架,基于服务调用实现交易类增、删、改、查应用的数据访问,避免直接操作数据。电信运营商数据湖应用方案示例如图 7 所示。

(5) 要点 5: 数据湖数据统一管理

数据湖的实施,需要实现模型和主数据标准的动态维护以及数据的集中治理,避免数据湖成为数据墓地。而数据来源众多,数据管理需要依赖于多方的密切合作以及数据标准管理、目录/元数据管理、应用/服务管理、质量等管理及海量数据探索分析等高效的管理工具。电信运营商数据湖管理体系示例如图 8 所示。

电信运营商数据涉及系统众多、关系复杂,没有任何一个独立的团队能够通晓所有的数据模型和关联关系,因此需要企业数据管理团队与专业数据管理团队分工合作,共同完成数据模型标准/主数据的管理及数据集中治理。建立横纵向一体化的数据管理体系,明确企业数据管理和原生数据部门职责分工,固化数据管理流程制度。

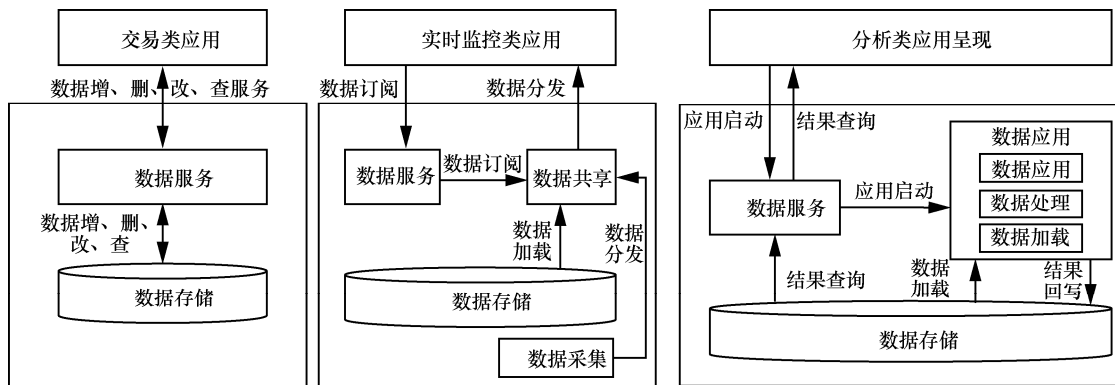


图 7 电信运营商数据湖应用方案示例

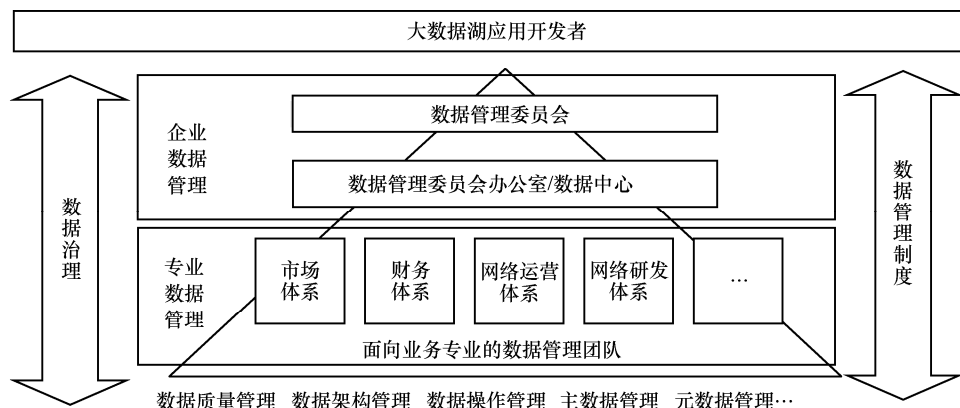


图8 电信运营商数据湖管理体系示例

企业数据管理团队负责统筹标准和主数据管理及数据治理工作，负责数据建模挖掘和跨专业数据治理协作，负责为业务部门和应用开发者提供数据建模和平台技术支持；专业数据管理团队负责建立专业数据的模型标准和管理主数据，识别数据问题及跟踪处理；数据湖应用开发者负责提出数据需求，按需整合和构建应用，反馈数据问题，评估数据变更影响。

另外，作为企业最核心的数据资产，其全生命周期的安全管理非常重要。需要针对数据采集、数据存储（生产数据、原生数据、整合数据、汇总数据）、数据应用、数据服务、数据分发共享等环节构建端到端的安全管控体系。对涉及用户行为特征及关键信息的敏感数据进行统一处理，脱敏后提供给应用使用；不管是敏感数据还是非敏感数据，所有数据的直接访问均在数据湖的管理范围内进行，具体措施包括数据应用环境、服务访问环境、共享分发环境、数据存储环境统一管控，需要经过统一的对象和属性等的鉴权才能访问数据，数据不出数据湖（即数据访问不出台），只能使用服务化方式或经过鉴权认证的数据共享分发方式进行数据访问。同时需要对大数据安全事件具备闭环管控能力，增强数据安全事件快速分析能力，提升安全事件发生后的应对处置效率。

5 结束语

数据湖统一汇集和管理各生产系统的运营数据，提供统一的存储和数据服务，各级数据应用和数据开发人员可以各取所需，充分发挥数据价值。在数据存储方面，数据湖按照生产系统的原生模型存储企业数据，不仅简化了数据的处理过程，保证数据真实可靠，同时能对具有时效性要求的实时应用提供更好的支撑；在分析应用方面，上层应用可以通过数据湖提供标准化接口获取数据，进行自定义的模型转换，能够更加快速地满足日益多样的分析需求；在数据治理方面，数据湖提供统一的目录管理、权限控制、编排调度、追踪溯源等功能，能够有效识别和管控各项数据及其处理过程，促进企业数据质量的持续提升。

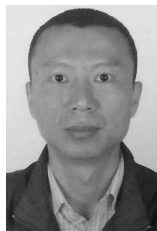
参考文献：

- [1] AliResearch. Activating productivity: model upgrading and paradigm transfer in DT era [R]. 2015.
- [2] TMF:TM_Forum_open_digital_architecture_whitepaper_V2[R]. 2018.
- [3] TMF:TMF630_REST_API_design_guidelines_3.0_R17.5.1[S]. 2018.
- [4] DIXON J. Pentaho, Hadoop and data lakes[Z]. 2015.
- [5] Wikipedia. Data lake[Z]. 2017.
- [6] PASQUA J. Schema-on-read vs schema-on-write[Z]. 2019.
- [7] TMF. GB922 information framework models R18.0.0[S]. 2018.
- [8] China Telecom. EDM 4.0[S]. 2011.

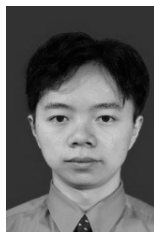


- [9] PWC. The enterprise data lake: better integration and deeper analytics[R]. 2014.
- [10] WILDER-JAMES E. The data lake dream[Z]. 2014.
- [11] HALEVY A, KORN F, NATALYA F. Managing Google's data lake: an overview of the goods system[Z]. 2016.
- [12] HALEVY A, KORN F, NOY N F, et al. Goods: organizing Google's datasets[Z]. 2016.
- [13] AWS. Populating the AWS glue data catalog[Z]. 2018.
- [14] 郭云峰, 蔡珩, 戈磊. 基于大数据网格的电信 IT 系统智能运维体系[J]. 电信科学, 2018, 34(6): 153-161.
GUO Y F, CAI H, GE L. Telecom intelligent operation system based on big data grid[J]. Telecommunications Science, 2018, 34(6): 153-161.

[作者简介]



胡军军（1969- ），男，中国电信股份有限公司广州研究院 IT 研发中心高级工程师、总经理，主要研究方向为电信支撑系统、云计算/大数据、移动互联网商业模式、智能可穿戴设备等。



谢晓军（1972- ），男，中国电信股份有限公司广州研究院 IT 研发中心高级工程师、技术总监，主要研究方向为 BSS/OSS 系统规划、设计、建设与运营的研究与支撑。



石彦彬（1980- ），男，中国电信股份有限公司广州研究院 IT 研发中心高级工程师、高级业务主管，主要研究方向为 OSS 系统规划、设计、建设与运营的研究与支撑。



喻琦（1985- ），男，中国电信股份有限公司广州研究院 IT 研发中心工程师，主要研究方向为 OSS 系统规划、设计、建设与运营的研究与支撑。