

数据中台框架与实践

吴信东^{1,2}, 应泽宇^{1,2}, 盛绍静^{1,2}, 蒋婷婷^{1,2}, 卜晨阳^{1,2}, 张赞^{1,2}

1. 大数据知识工程教育部重点实验室(合肥工业大学), 安徽 合肥 230009;

2. 合肥工业大学计算机与信息学院, 安徽 合肥 230009

摘要

数据中台将一个机构(企业、事业或政府部门)的数据作为战略资产进行管理,是从数据收集到处理应用的一套管理机制,以期提高数据质量,实现广泛的数据共享,最终实现数据价值最大化。给出数据中台的定义,提出数据中台参考技术框架,并分别对物理管理、逻辑管理、数据资产管理、数据服务和信息安全管理的技术进行了展开讨论。最后以华谱系统建设为例,介绍面向家谱大数据、结合HAO智能模型的数据中台实现——华谱数据中台。

关键词

数据中台; 数据资产; 数据治理; 数字化转型

中图分类号: TP311.13

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2023034

Data central-platform: architecture and practice

WU Xindong^{1,2}, YING Zeyu^{1,2}, SHENG Shaojing^{1,2}, JIANG Tingting^{1,2},
BU Chenyang^{1,2}, ZHANG Zan^{1,2}

1. Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Hefei 230009, China

2. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

Abstract

A data central platform positions the data of an entity-be it a corporate entity, institutional body, or governmental department-as a pivotal strategic asset. It's a management mechanism that spans from data collection to processing and application, aiming to improve data quality, achieve extensive data sharing, and ultimately maximize the value of the data. A definition for data central-platforms was provided, and a generic architecture was presented along with the core technologies and functions of physical management, logical management, data asset management, data services and information security management. Finally, taking the construction of Huapu system as an example, a realization of the data central platform, which is geared towards genealogical big data and integrated with the HAO intelligence model, was introduced-Huapu Central-Platform.

Key words

data central-platform, data asset, data governance, digital transformation

0 引言

随着移动互联网、物联网、云计算等信息技术的快速发展,人类社会进入了大数据时代。为了充分利用大数据时代海量数据蕴含的巨大价值,社会各机构(政府、企业等)纷纷进入数字化转型,引发了数字化转型方法的积极探索。但是数字化转型却伴随着许多困难,很多机构的数字化转型并不成功,而其中一个重要的原因是没有统一、可行的转型路径和相关技术方案。

通过中台实现数据化转型是一条可行的道路。2015年年底,阿里巴巴集团对外宣布全面启动阿里巴巴集团2018年中台战略^[1]。此后,腾讯、今日头条等企业开始了中台建设的摸索与实践。中台建设已经成为互联网企业寻求数字化转型的突破口。经过近5年时间的发展,中台已经取得相当成熟的推广和应用。数据中台行业的市场份额已经由开始商品化的2019年的38亿元在两年时间内增长到101亿元^[2],选择搭建数据中台的企业也越来越多,已经不限于互联网企业。不过数据中台面临着未有统一的中台概念和数据中台建设标准、规范以及评价指标的问题^[3],这使得很多有数字化转型需求的机构无法准确把握中台的意义而错失了中台建设的良机。

因此,为数据中台下一个贴切的定义,总结一套具有共性的数据中台建设方法很有必要。本文首先分析中台建设必要性和重要性,对国内外数据中台相关研究进行介绍,并给出数据中台的正式定义;其次给出基本技术框架(物理管理、逻辑管理、数据资产管理、数据服务、信息安全管理)并介绍相关内容;接着,以华谱数据中台建设为例,介绍华谱数据中台结合HAO

智能模型的总体架构和相关开发方案是如何成功处理海量家谱数据的;最后总结当前数据中台的挑战和产业发展前景。

1 研究背景

人类利用计算机技术管理数据到目前为止大体上经历了(手工)报表、报表系统、数据仓库系统、大数据平台等概念阶段^[4]。每个阶段人们都在寻找更加有效的数据利用方式,而数据中台是现阶段有效实现数据价值的解决方案之一。

目前国内对数据中台的研究趋于成熟,数据中台已在互联网、电力等多个行业广泛应用^[5]。阿里巴巴提出的OneData中台体系为中台建设的实践打下了坚实的基础^[6]。在基础能力平台的建设上,林鸿等^[7]提出使用分布式微服务技术架构的方案。在数据治理上,中国通信标准化协会发布《数据治理标准化白皮书》,确定了一系列数据治理的标准和实践方案。在数据服务方面,微服务框架Duboo提供了一系列可进行微服务治理的开源组件^[8]。

国外并没有提出数据中台这一概念,但在大数据技术上有成熟的发展。在数据治理上有ApacheAtlas这样在Hadoop生态系统上的元数据治理框架^[9],基础能力平台上有诸如Spring Cloud等微服务框架。

综上所述,在数字化转型的背景下,各行各业都在建设数据中台。但各行各业的场景千差万别,难以总结出行业间共同享有的业务共性,本文由此提出可能用于数据中台构建过程的技术框架。

2 数据中台概念

中台的概念是基于前后台架构的传统

模式提出的。前台指的是由各类用户终端系统组成的一个整体，后台指的是可以管理企业的核心资源（数据+计算）的系统^[10]。前后台基于特定的业务而成立，并仅为这条业务线服务，久而久之许多业务产生各种独立的前后台整体，形成“烟囱式开发”的形式。前后台交互模式如图1所示。

由于大数据时代业务需求具有快速响应和定制化的特点，单独设立后台定制化开发会产生重复开发、交付周期延长等弊端。同时前后台开发会导致“数据孤岛”，数据和数据能力彼此孤立、难以共享，造成信息化协作困难，大数据分析和知识图谱建设无法进行，数据价值不能体现，不能快速响应业务需求。为解决上述问题，中台的概念开始被提出。中台对全域数据进行数据治理，以共享数据服务的方式实现数据共享，通过搭建独立的中间平台，打通业务和数据环节，减少冗余，增加复用，快速响应用户需求，实现数据驱动的业务创新。

总体来说，数据中台是想要打通“数据孤岛”提高数据价值的策略，但目前还没有统一的定义^[11-12]。笔者在此提供一个从数据资产化角度的定义^[13]：数据中台将一个机构（企业、事业，或政府部门）的数据作为战略资产进行管理，是从数据收集到处理应用的一套管理机制，以期提高数据质量，实现广泛的数据共享，最终实现数据价值最大化。

从上述定义出发，可以得到数据中台必须满足的两点要求。其一，数据中台需要实现数据的全局管理。首先，数据中台管理全局数据是从数据收集到处理应用的全流程管理。此外，数据中台掌控的全局数据要为业务赋能、避免“数据孤岛”，需要提供广泛的高质量数据共享服务。因此，数据中台的管理要做到全数域管理、全时段管理、数据全平台共享。

其二，数据中台需要实现数据智能化。数据中台管理全局数据的最终目的是数据资产价值最大化，其主要方法是通过

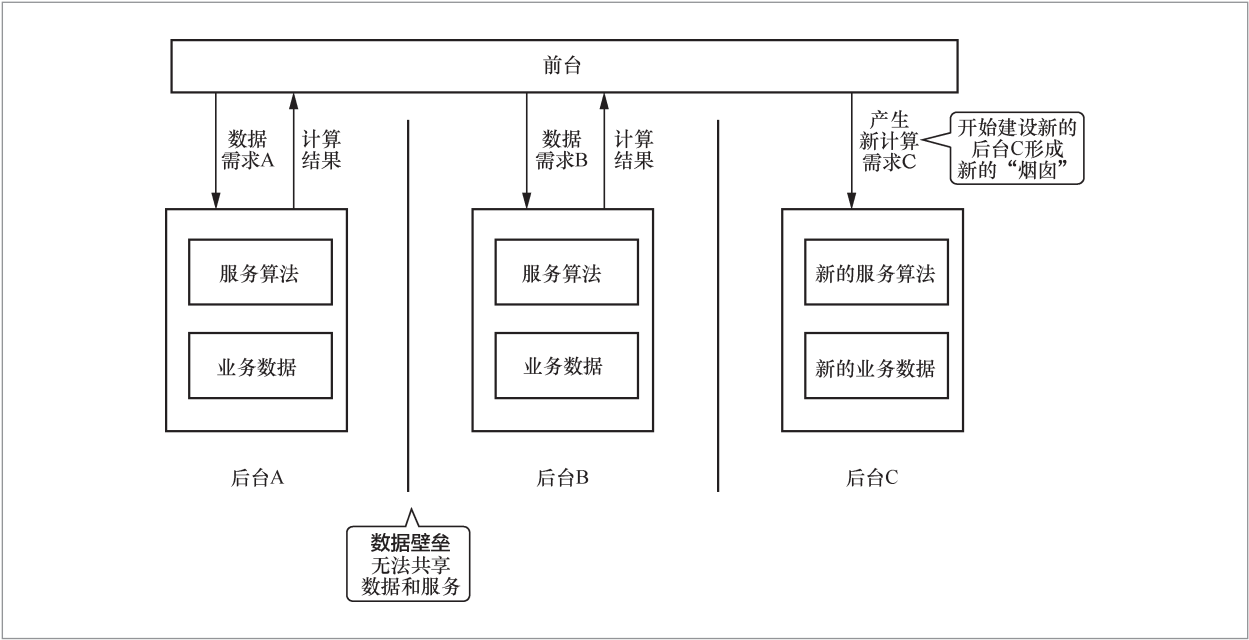


图1 前后台交互模式

数据挖掘等技术从数据中获取有价值的信息和新知,为业务和决策赋能。

综上所述,数据中台在整个机构的业务行为链中处于中心位置,具有核心重要性。各个业务前台和组织部门通过中台取得驱动业务、决策展开的数据或数据洞见,同时,前台在业务中产生的新数据也会源源不断地汇聚到中台中,形成交互闭环,驱动机构良性发展。中台交互模式如图2所示。另外,笔者为强调中台的中心化管理能力 and 核心重要性,没有使用Middle Platform等说明中台处在前后台中的中间环节的英文翻译,而是使用了Central Platform作为中台的英文翻译。

3 数据中台框架

数据中台的建设依托于建设机构的原信息化系统和具体业务需求进行^[4],因此数据中台的建设方案具有特殊性,很难有完全通用的数据中台标准建设架构。但是,数据中台的建设都基于相同的最终目

的——实现数据价值最大化(数据全局管理和数据智能化)。同时数据中台的建设与前沿的大数据技术息息相关,因此在数据中台的建设思路和技术选择上具有共性。本文结合大数据技术的发展,以笔者在对数据中台的实际建设中归纳出的7个数据中台的核心功能^[13](如图2所示)为基础,提出了数据中台的参考框架,如图3所示。

数据中台的参考框架是一个层次模型。物理管理解决整个系统数据存储、运算、共享等基础能力的实体机器搭建和软件操作平台搭建,并完成大数据采集和汇聚;在此基础上对存储数据进行逻辑管理,使得原始数据经过整治变成标准可操作的有统一数据模型管理的数据集;数据资产管理对标准化、可操作的全局数据进行价值管理和共享管理,并利用算法开发数据资产,使之智能化,释放资本价值;数据服务与业务相结合,通过统一的数据接口为业务提供数据服务,完成数据资产的实际利用;信息安全管理贯穿建设的始终,为数据的处理提供安全保障。

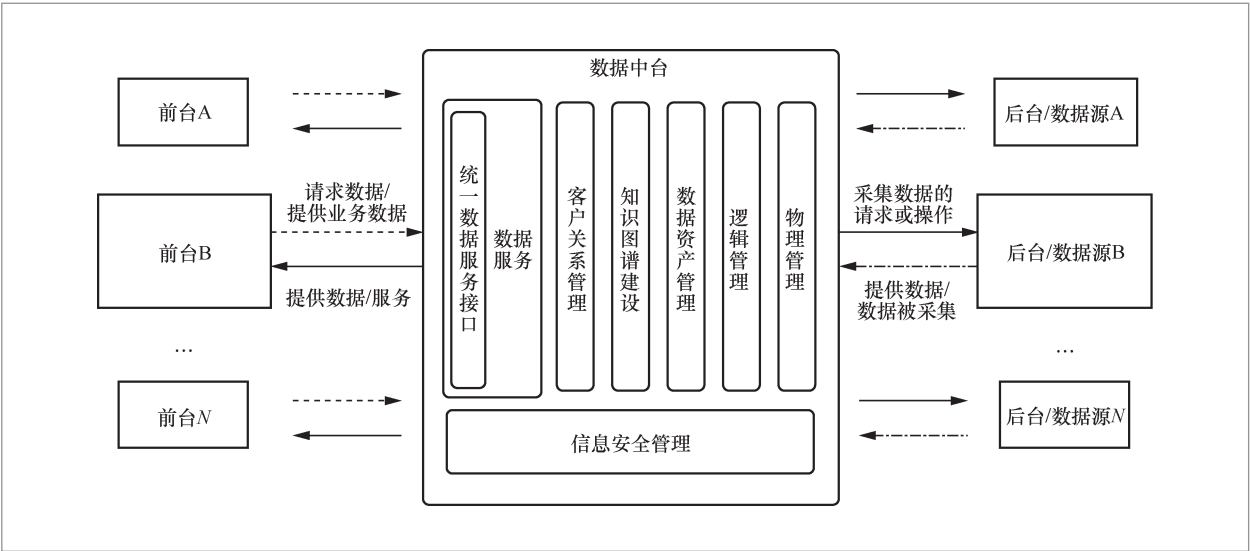


图2 中台交互模式

3.1 物理管理

物理管理包括物理工具支撑、基础能力平台、数据采集与汇聚。物理工具支撑是最底层的硬件集合，基础能力平台是对物理工具存储、计算能力的抽象、管理，数据采集与汇聚是基于基础能力平台实现的数据中台的初步功能。

(1) 物理工具支撑

随着机构规模的扩大、数据量的激增，机构对算力和存储等物理工具的要求自然进入了云计算的时代，机构需要部署其云服务平台以满足基本的算力和存储需求。此外，机构还需要根据自己的业务特性增派一些特殊实体资源，如物联网系统需要增派大量传感器。

云服务系统的架构部署有3种模式^[14]：公有云、私有云和混合云。公有云是第三方公司通过互联网连接提供给用户的云，如AWS、华为云等，采用这种架构能够减少硬件开销，但安全性得不到足够的保障；私有云是机构搭建基础设施且只供内部使用的云，这种架构安全性高，但是需要耗费硬件且运维成本高；混合云的架构结合了公有云和私有云的优势，将重要的数据服务建设在私有云上以求得安全稳定，把不重要的资源铺设在公有云上，减少硬件成本。

(2) 基础能力平台

基础能力平台是对底层复杂硬件资源的抽象，并提供统一数据存取、计算等基础能力的技术集合。单体架构模式^[15]难以满足云计算时代应用可弹性扩展、异构数据资源标准化管理、适应业务快速迭代等需求，因而云计算领域经过多年的发展积累已经形成了一套高可用的、弹性、可管理的被称为云原生(cloud native)技术的基础架构。

云原生架构是一种利用云计算优势

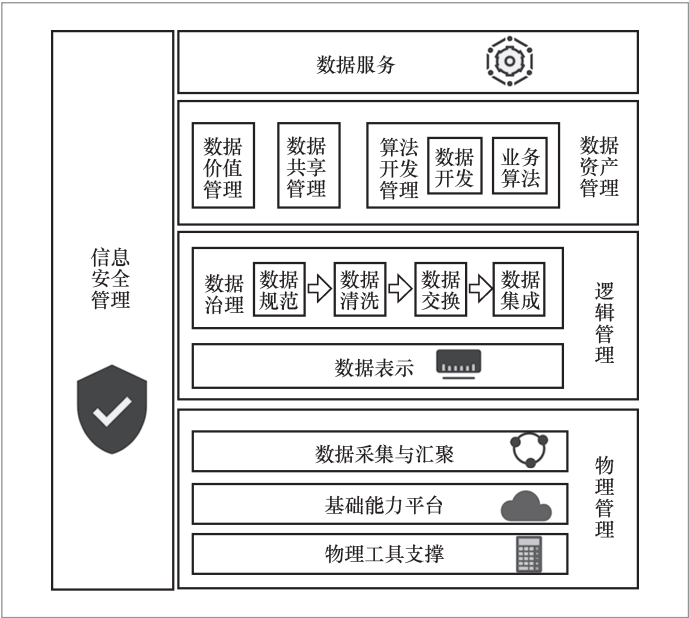


图3 数据中台的参考框架

来构建和运行应用程序的方法^[16]，它是一个技术和方法论的集合，包含4个要素：容器、微服务、DevOps、持续集成和持续交付(CI/CD)。这4个要素可以很好地满足中台建设的需求。首先，容器化开发使应用能够轻易地扩容到系统之中，容器化又具有相对封闭性保障数据的安全，通过微服务的接口方式使共享变得简单；其次，通过持续集成和持续交付技术能够极大地提高软件上线效率，满足了快速迭代的需求；再次，通过容器化和微服务方式开发的应用能够被当作组件，由云平台的自动化工具统一管理、实时监控，实现了标准化配置和管理；最后，云原生架构本身就是建立在云计算基础上的架构体系，对分布式架构具有优良的适应性。综上所述，数据中台的基础能力平台建设适合在云原生架构上进行。

CNCF(cloud native computing foundation)是Google、Red Hat、Microsoft等大型云计算厂商以及一些开源软件公司共同成立的云原生计算基金

会,它提供了云原生架构的路线图^[17],云原生架构建设步骤见表1。

云原生架构之内还需要选择合适的大数据计算能力。数据中台的大数据计算引擎可以分为即席查询、离线计算、分布式计算、流式计算4个平台^[18]。即席查询引擎需要对海量数据进行秒级的实时查询和计算,可以使用高性能和低时延的Impala;离线计算技术需要实现超大规模的批量计算,Hive是一款基于HDFS的MapReduce计算框架,对单节点的处理器利用率达到90%,是离线计算的选择之一;分布式计算平台可以选择Hadoop、Spark、Flink等;流式处理平台可以选择Storm、Spark Streaming等。

(3) 数据采集与汇聚

系统全域的原始数据都存放在各部门自身的业务系统中,需要经过数据采集将数据汇聚起来,构建一个打通所有原始数据域的数据湖(data lake)^[16]。

数据湖需要存储,可以根据数据的种类和结构类型选取适宜的存储工具。例如,日志数据和通用文件可以选择HDFS进行存储,HIVE存储关系型数据,采用图数据库存储具有关联性的大数据集等。

表1 云原生架构建设步骤

步骤	步骤名称	技术项目
1	容器化	Docker、Podman
2	CI/CD	Gitlab CI/CD
3	应用定义与编排	Kubernetes、HELM
4	监控&分析	Prometheus、fluentd
5	服务代理、服务发现和服 务网格	Envoy、CoreDNS
6	网络&策略	CNI、Open Policy Agent
7	分布式数据库与存储	Vitess、etcd
8	流与消息处理	gRPC、NATS
9	镜像私库&运行环境	containerd
10	软件分发	Notary、TUF

3.2 逻辑管理

逻辑管理要将在物理管理得到的原始数据经过处理加工,转变成可理解、可操作的具有统一语义和结构的数据资产。逻辑管理分为数据表示和数据治理两部分。

3.2.1 数据表示

数据中台需要集中管理海量多源异构的业务数据,因此需要根据不同数据源的数据特征,明确数据的结构、语义和标准等表示信息。数据表示主要包括数据标准管理和元数据管理。

数据标准是指保障数据的内外部使用和交换的一致性和准确性的规范性约束^[19]。一般数据标准会通过标准文件发布,但在中台建设中,由于各个“数据孤岛”间的独立性,各个业务系统人员对标准的理解难免产生认知偏差,难以保证标准的落实,因此数据中台要有一套由规范要求、流程制度、技术工具共同组成的管理体系确保数据治理各个阶段的数据的标准化以及标准的沉淀。数据标准管理包括数据接入标准、命名标准、数据格式标准、数据安全标准、资源管理标签等多个方面。数据中台数据标准管理可通过区块链^[20]、流程自动化等技术工具来保障。

元数据管理包含了一系列标准:数据格式、代码规范、数据隐私规则、数据表的命名原则等。其将数据资产用清晰直观的方式呈现,让数据资产真正被读懂。但由于“数据孤岛”问题,元数据自发产生多元化、非标准化的协调发展问题。为保障元数据的统一性,可以采用语义互操作、结构互操作、协议互操作等方法^[21]来解决相关问题。

3.2.2 数据治理

根据数据管理能力成熟度评估模型^[22],数据治理是指对数据进行处置、格式化和规范化的过程。由此可见,数据中台的数据治理是对数据中台中的机构全局数据进行处置、格式化和规范化的过程。数据治理的格式化、规范化过程在数据中台建设中的内涵可以理解为统一的数据规范和统一的数据建模及其管理的落地。数据中台的数据治理主要包括4个子功能的实现:数据规范、数据清洗、数据交换、数据集成。

(1) 数据规范

数据规范是指进入数据中台的数据(输入)和经过数据中台处理的数据(输出)都必须符合的规范^[16]。其通过一系列技术手段确保数据中台中的数据I/O符合数据标准。例如,通过对数据库属性值设置一系列约束(完整性约束、唯一性约束、空值规则等)实现对数据的检验。

(2) 数据清洗

数据清洗的功能是侦测数据集中的“脏数据”,并对“脏数据”进行清洗进而达到提高数据质量的目的。数据清洗可以分为属性错误清洗、不完整数据清洗以及相似重复记录的清洗^[23]。

属性错误清洗识别并清洗违反数据库原定义的完整性约束的冲突数据。识别方法有通过指定合法数据库实例模式的定性方法或在离群点检测的基础上采用统计方法识别冲突数据的定量方法。属性错误可以通过光滑噪声技术自动修正,或人工修正。

不完整数据清洗针对的是数据缺失现象。清洗工作可以通过忽略、全局变量填充的方法处理缺失数值,也可以通过统计和数值预测的方法,如中心度填充、最可

能值填充处理。

相似重复记录清洗的重点是识别出相同或不同数据集中两个实体是否指代同一实体,即实体对齐。实体对齐的基础方法是文本相似度度量,大致分为基于字符的(如编辑距离)、基于单词的(如 Jaccard 系数)、混合型(如 softTF-IDF)和基于语义(如 WordNet)的4种方法。相似重复数据集的清洗一般采用先排序后合并的思想,使用优先队列算法、近邻排序算法等实现。

(3) 数据交换

数据交换的功能是将原始数据转换为符合特定模式的目标数据,同时保证目标数据能正确反映原始数据的内容。数据交换给数据集成提供了基本条件。数据交换的实现一般分为以下两种方式^[23]。

- 协议式交换:指源系统和目标系统之间定义一个数据交换交互协议,遵循制定的协议,通过将一个系统数据库的数据移植到另一个系统的数据库来完成数据交换。

- 标准化交换:建立一个可供多方共享的方法作为统一的标准,实现跨平台应用程序之间的数据共享和交换。

在数据中台中需要灵活运用以上两种方法。如果几个数据源与中台的转换规则相似,那么可以使用标准化交换减少点对点的协议式交换开发成本;如果有的数据源比较特殊,无法复用标准化交换功能,那么可以单独做点对点的协议式交换开发。

协议式交换可与区块链技术结合。区块链技术具有公开透明、不易被篡改等优势,因此可以保障已经处理过的有价值数据进行流转、存储及追溯,从而提升数据交互效能。刘峰等^[24]基于此提出了一种面向双中台双链架构的内生性数据安全交互协议,取得了优良的实验结果。

(4) 数据集成

数据集成的目标是将多源异构数据汇

聚后的无序原始数据（可以是数据湖）通过标准化得到能够客观描述机构主要业务和功能时序状态的可理解、可操作的有序数据集。数据集成技术协调数据源之间不匹配问题^[25]，将异构、分布数据集成在一起，为用户提供统一视图，便于用户更加透明地访问数据源。数据集成的方式主要分为3种：数据复制、虚拟集成、基于本体建模的数据集成。

① 数据复制

数据复制方法是将用户可能用到的其他数据源的数据预先复制到统一的数据源中，用户使用时，仅需要访问单一的数据源或少量的数据源。数据复制方法使得异构、分布数据统一在一个数据库中，因此提高了解析查询效率；但数据复制需要一定的时间，因此数据的实时一致性难以保证。数据复制方法的常用方式是数据仓库方法^[26]。

② 虚拟集成

虚拟集成技术保持各数据源的分散状态，通过建立一个反映全局数据的由逻辑模型构成的逻辑视图达到间接掌握全局数据的效果。虚拟集成技术使用虚拟化技术实现逻辑模型到各分散数据源的数据控制。杜小勇等^[27]提出了一种基于中间模式的数据集成系统架构，中间模式对下层数据源管理各个分散数据源的逻辑封装，中间模式向上层数据应用提供统一的数据模式和数据访问的通用接口。当用户有查询请求时，中间模式按照元数据编译请求语句，分解成对各数据源的特定操作。

虚拟集成系统免去了海量数据汇聚的弊端，虚拟化服务和基于逻辑视图的操作对用户也非常友好。但如果异构的数据源繁多复杂，开发封装器和映射模式的代价将会非常大。

③ 基于本体建模的数据集成

集成同一领域的异构数据库有3个主

要问题：语义、语法和结构的异质性。基于本体的建模方法能描述领域概念术语，表达概念间的内在联系，实现不同概念之间的集成和转换，并保持语义上的一致性。同时能够通过本体的推理机制消除不同领域的重复定义，发现其中隐含的关系，能够充分解决上述3个问题^[28]。在具有跨专业跨流程的电网企业全类别全过程项目管理信息模型的信息描述中有成功应用^[29]。目前，基于本体建模的数据集成方法包括单本体方法、多本体方法和混合本体方法3种^[23]。

在数据中台的实际应用中，统一的本体模型不仅要结合组织绝大部分数据需求，还要提供兼容性和扩展性以满足特定的数据需求。以国家电网公司统一数据模型（SG-CIM）^[30]为例，其通过组织智能审定出企业的公共信息模型，保证统一性，同时其为分派部门提供了统一的物理模型基线版，保证底层结构的一致性，各部门按照本地需求可在模型的允许范围内扩展模型，这样从根本上保证了数据模型的一致性，又兼具了可扩展性。

3.3 数据资产管理

数据资产管理的任务是面向业务设置配套的管理体系以达到数据资产价值的释放。数据资产管理主要包括：数据价值管理、数据共享管理、算法开发管理。

3.3.1 数据价值管理

数据价值管理是对数据内在价值的度量，可以从数据投入成本和数据应用价值两方面来开展^[31]，即对数据资产的总投资收益率（return on investment, ROI）的衡量。数据价值管理目的是衡量数据的价值以便调整数据资产的投资结构，提高资产收益。其作用是挖掘数据资产价值潜力

和评估数据采集维护成本,将两者做综合评判以帮助机构合理调度整体资源,实现降本增效的目的,数据价值管理是数据资产管理的核心部分。数据投入成本主要包括硬件、能源和人力成本,这一部分成本投入较容易计算。而数据中台为机构提供集中化、智能化的数据服务,进而将数据转化为经济价值实现数据资产变现的过程是间接的、漫长的,故数据应用价值是难以即时获知的。

目前对数据资产总投资收益率的估计可以通过对数据进行统计和情感分析得到,用访问量、好评率等信息表现数据价值;或者通过回溯实际业务经验,归纳可靠的评估指标来界定数据的应用价值,如活性评估、数据质量评估、数据稀缺性评估、数据时效性评估、数据应用场景经济性评估等。

3.3.2 数据共享管理

数据中台提供的广泛数据共享会带来数据安全和资源占用等相关问题,如越权访问数据、大量访问导致服务器崩溃等。为了避免以上问题,维护核心功能和资产价值的顺利实现,需要一套预防、监控、分析数据共享的管理方案。

这套方案需要从共享的需求和实际问题入手,制订数据共享的条件、规范流程以及监管手段。例如,某数据服务在业务中共享量比较大,可以在管理机制上优先保障该数据服务的共享资源;又或者数据共享中涉及资源竞争,数据共享管理机制就要事先制订好涉及资源竞争时的共享策略。

3.3.3 算法开发管理

数据资产管理的核心作用是使经过采

集、治理后的数据通过算法被使用起来,服务于业务和产生商业洞见。根据使用数据的目的,算法可以分为业务型算法和数据开发型算法。

业务型算法是基于前台业务产生的一系列操作数据需求定制并由中台统一管理的算法。在中台框架下,业务数据需求会先搜索中台的数据服务体系中是否存在可复用的接口,如果直接调用,可免去二次开发的过程;如果不能直接调用,中台将开发任务派发给后台,让其按照中台的规定进行开发,后台接口开发完成后由中台统一管理。

数据开发型算法应由数据中台内部开发。数据开发型算法是基于数据中台管理全局数据的能力进行的对决策洞见、业务优化的探索。数据开发型算法会从全局数据中挖掘新知识,或者通过机器学习优化业务功能,以此达到数据驱动决策,数据赋能业务的资产价值。

数据中台进行全局大数据开发,首先要建立一个机器学习平台。机器学习平台建设基于数据中台的发展理念,要求对数据处理、特征工程、特征重要性分析、常见算法模型和一致性校验等核心功能进行组件化封装,同时要求开发按照统一的代码规范、协作机制、模型管理、上线流程进行,以便共享和复用,此外平台应该根据实际的算法技术特点和业务需求变化进行算法的升级。目前机器学习平台在市面上有不少产品,比如Google TFX、Facebook的FBLearner、阿里巴巴的PAI平台等^[32]。

3.4 数据服务

数据服务是数据中台的最后一个环节,也是数据资产发挥其价值的时刻。对全局数据进行能力抽象和统一管理是为了将数据价值和数据能力共享给各部门,各部门通过复用中台提供的数据服务给业务

赋能。数据中台管理数据服务有以下4点要求：

- 数据服务的形式是应用程序接口（application programming interface，API）且API要交由数据中台统一管理；
- API要避免重复建设；
- 保障数据获取及时、高效和稳定；
- 数据中台管理的API具有可扩展性。

数据服务可分为通用型服务和专用型服务。通用型服务是指能够被广泛复用的数据服务，一般指对数据中台管理的数据集的增删改查以及利用人工智能和数据挖掘技术开发的分析工具。专用型服务应对的是无法全部通过复用通用型服务解决的业务场景，基于特定场景的特征开发出的数据服务。专用型服务中具有共性的部分也可被分离出来作为通用型服务。

3.5 信息安全管理

数据中台构建在云上，与互联网天然连接，必须要做好对互联网的安全防范工作。全局的安全保障体系需要围绕 ISO 七层模型建立，对各个层级进行安全把控，保障系统正常使用^[33]。除了建设一般性的外部网络安全保障，在内部也需要依其自身特点构建相应的安全管理措施。此外，数据中台打通分散数据源的过程会带来不可预知的数据泄露风险。而且数据中台汇聚的数据是全局性的，一旦遭到威胁必将带来巨大的损失。

针对数据安全威胁，机构应当从数据安全和隐私保护两方面使用对应的安全管理技术手段。

(1) 数据安全

- 数据容灾备份：为避免因为灾害、停电、误删等意外性事件造成资产流失，数据需要备份。
- 数据权限：控制数据能够被哪些用

户做哪些操作。一般在数据建模阶段就要明确，同时要保证权限的动态调整。

- 非法操作警报和审计：要有详细的日志记录，用于实时审计或者事后审计，对敏感数据要建立实时的报警机制以防事态蔓延。

(2) 隐私保护

- 身份认证：证明用户身份，保证数据权限的实施。
- 数据脱敏：个人私密信息，如密码、身份证号、手机号、邮箱、地址等，应该采取加密存储、模糊化存储的方式。

4 华谱系统：数据中台的建设实践

家谱历史悠久，蕴含丰富的历史、经济和文化等信息，具有极高的社会和经济价值。与此同时，家谱数据是具有海量、多源、异构、自治等大数据特征^[34]的碎片化数据，通过信息化手段进行家谱数据系统建设和家谱知识挖掘、推理等面临许多难题。难题主要包括数据数字化、数据标准化、数据服务化3个方面。数据数字化是指将传统文本类型多样的家谱电子化并汇聚起来需要极大的成本；数据标准化是指多源家谱数据的不一致性使得数据融合、治理困难；数据服务化是指家谱修撰面向庞大且需求复杂的用户，家谱修撰功能不能单一，需要做到个性化、定制化。为提高家谱大数据的挖掘和分析利用，从2016年开始，笔者团队建设了一个面向所有华人姓氏的家谱系统——华谱系统。到目前为止，华谱系统已有超过1 867万条人物数据和721个姓氏。华谱系统主页如图4所示。

华谱系统为解决上述数据数字化、数据标准化、数据服务化问题，采用一套人类智能（HI）、人工智能（AI）和组织智能（OI）三者的交互和协同的HAO智能体

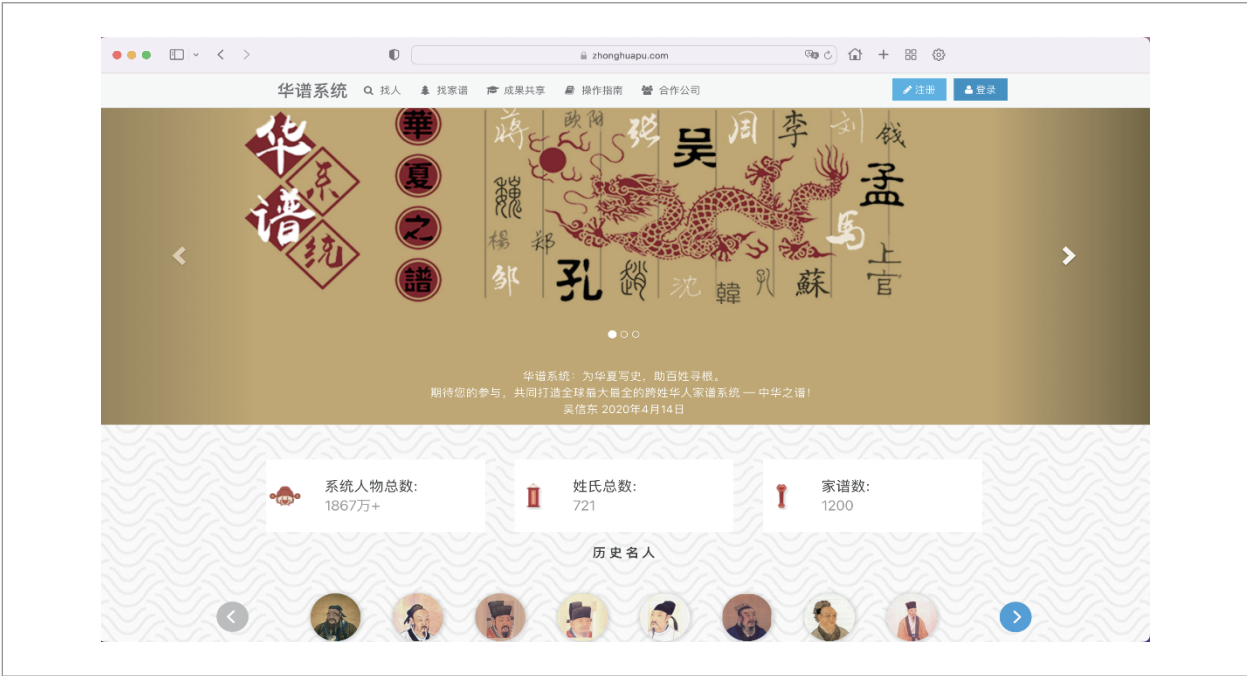


图 4 华谱系统主页

系^[35]，在数据中台的参考框架上进一步细化，结合家谱修建的具体场景打造了Huapu-CP框架进行数据中台建设。以下

将围绕Huapu-CP介绍现阶段华谱数据中台应对的问题和解决方法。Huapu-CP框架^[13]如图5所示。

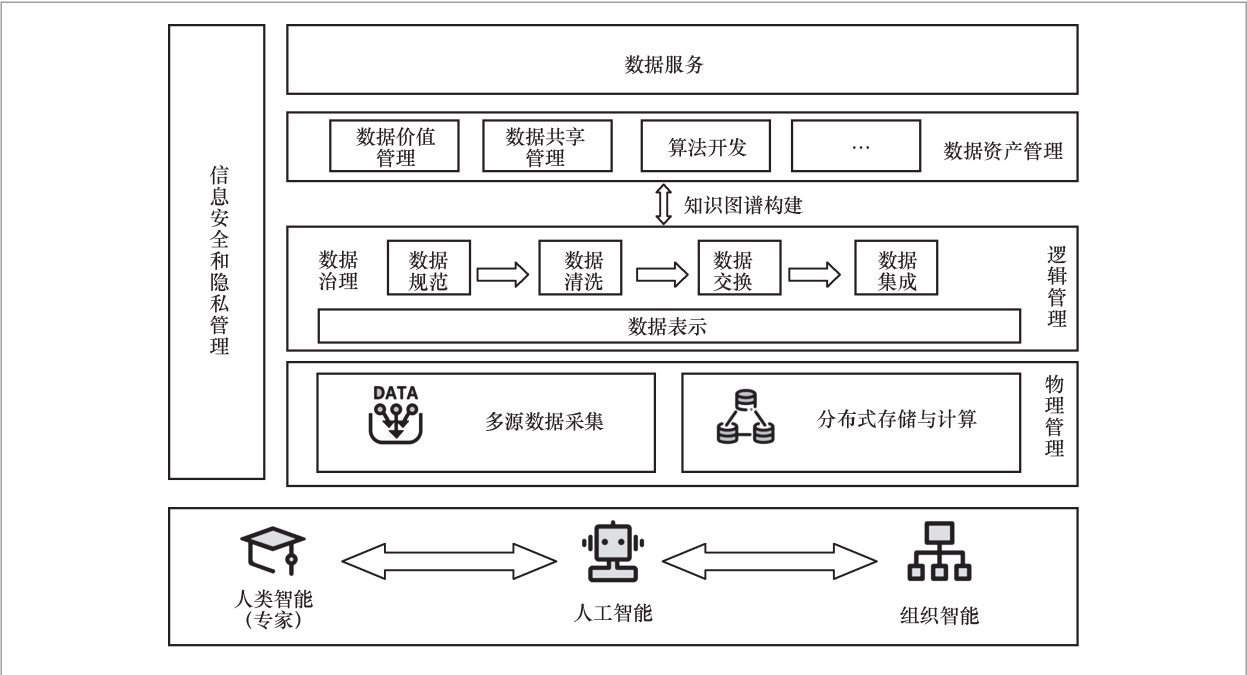


图 5 Huapu-CP 框架^[13]

4.1 华谱数据中台: 物理管理

在物理管理层, 华谱数据中台选用图数据库集群的方式将数据分布存储在云端, 实现高性能、弹性扩展和容灾的数据存储。在对比MapReduce与Spark用于大数据分析的优缺点^[36]后, 华谱数据中台选用Spark计算架构, 提高了系统的运行效率。在分布式查询方面, 华谱数据中台通过图划分算法, 将子图分布到不同机器上进行并行查询^[37], 分布式查询框架如图6所示。华谱数据中台的数据采集主要通过线下数据采集、互联网数据采集、线上行为采集和内部数据汇聚4种方式。其中互联网数据采集借助WebCollector^[38]工具。

4.2 华谱数据中台: 逻辑管理

逻辑管理层是数据中台建设的核心, 是数据资产化的主战场。其中逻辑管理包

括数据表示和数据治理两层。在数据表示层, 华谱数据中台基于HAO智能构建亲属关系模型, 并采用本体粒度划分技术, 分别以“家谱”“人物”“用户”为单元构建知识图谱。数据治理层分成了4个模块——数据规范、数据清洗、数据交换和数据集成, 4个模块依次进行。

华谱系统采用领域专家总结的一套计算机可理解的规则库的启发式字典方法和数据字典完成数据规范。数据清洗模块面对原始数据丢失、采集或录入信息有误的问题, 通过错误关系模式挖掘^[39]和属性自动填充机制解决。属性自动填充机制是将人类智能定义的属性自动填充规则转化为可执行程序的方法。错误关系模式挖掘是基于人类智能归纳总结定义的亲属关系图中的错误模式图, 通过子图匹配算法定位噪声, 然后将检测的噪声数据反馈给用户修正以实现数据清洗。数据交换模块采用协议式交换的方法, 制订了一套通用的必须遵守的交互协议, 打通了数据壁垒。团队利用实体对齐、冲突消解和数据融合等技术开发了碎片化家谱融合框架——FDF-HAO (结合 HAO 智能模型的碎片化数据融合框架)^[40], 对家谱数据进行集成, 实现碎片化家谱知识的融合。FDF-HAO框架深度融合了人类智能 (HI) 和组织智能 (OI) 中的专家知识和数据标准, 给出了一套新的面向家谱数据的数据抽取方法、无监督实体对齐算法、冲突解决机制和属性融合算法。笔者通过实验证明了在家谱数据上的有效性和高性能。FDF-HAO框架如图7所示。

4.3 华谱数据中台: 知识图谱构建

从海量异构的家谱数据中获取潜在知识是使数据资产价值最大化的关键, 因此构建知识图谱是数据中台数据智能化的关

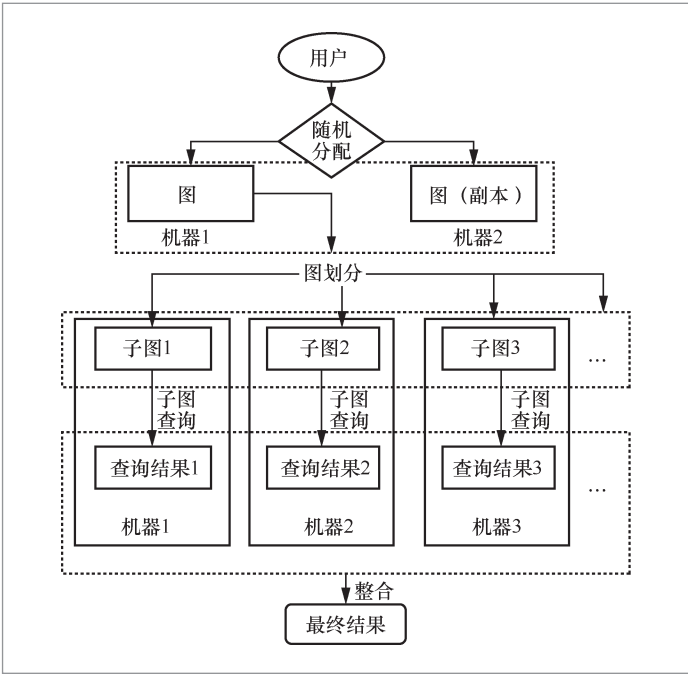


图6 分布式查询框架

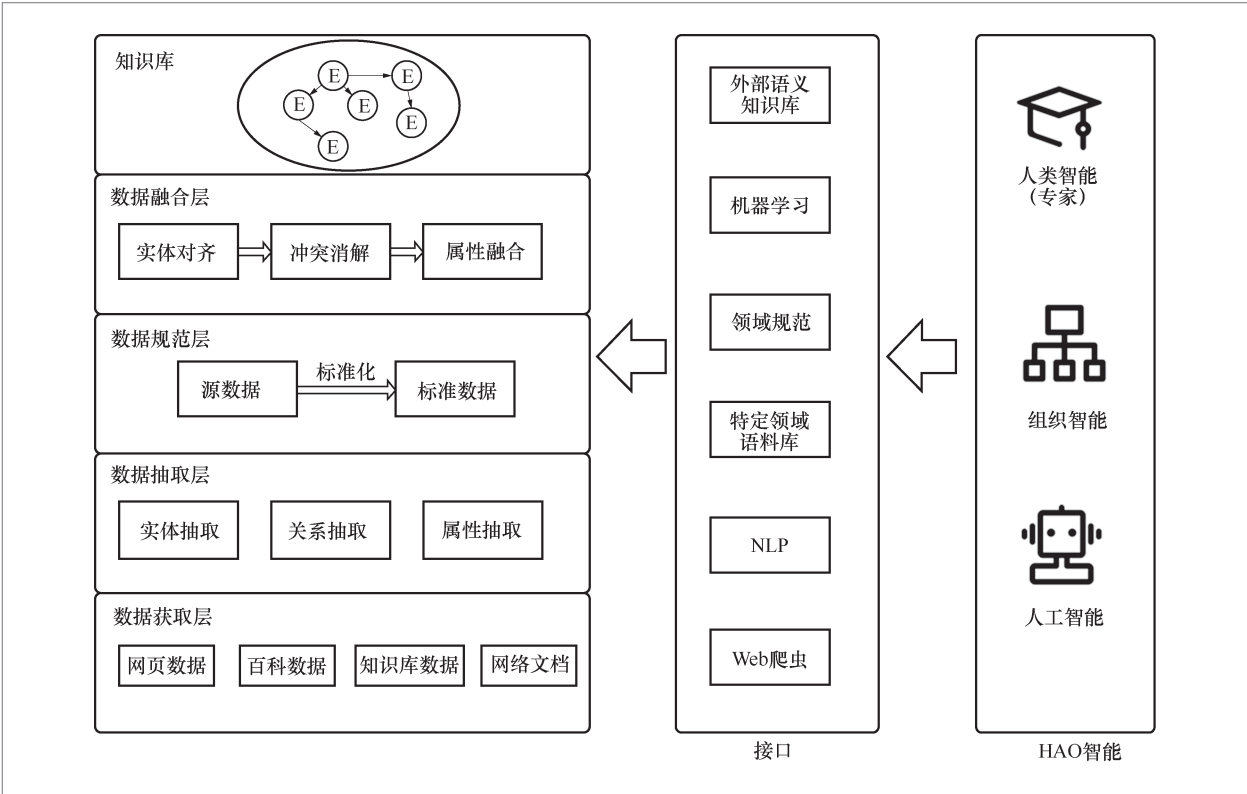


图 7 FDF-HAO 框架^[40]

键环节，是华谱数据中台数据资产管理的中心。但是由于知识图谱模型在谱系学领域还处于起步阶段，一般的家谱知识模型不适用于中国家谱的知识图谱构建，因此笔者团队基于中国家谱的特点构建了中国家谱知识图谱模型^[41]，并将其应用在华谱知识图谱（Huapu-KG）的构建中。

Huapu-KG采用本体粒度划分技术，划分了“家谱”“人物”“用户”3种家谱知识元素。在知识元素内部，本文对该粒度下本体属性的描述进行了静态属性和动态属性的区分，保证了本体信息的可变性。在家谱知识元素的基础上，通过组合家谱知识元素数据中台可得到能够表达复杂语义、提高知识描述能力的家谱知识单元，如“超点”“跨姓家谱”“群组”。

此外，由于中国家谱中存在大量专有

名词和生僻词汇，笔者在领域专家的帮助下，利用HAO智能和“简化复杂性”策略开发了亲属名词库。名词库解释了晦涩难懂的单词，提高了家谱知识的可读性。Huapu-KG示例^[41]如图8所示。

4.4 华谱数据中台：数据服务

华谱数据中台的建设是以数据应用为驱动的，通过开发数据服务，实现敏捷的应用开发，最终将数据使用起来，发挥数据资产的价值。目前华谱数据中台已经开发出家谱人物查询、家谱树展示、家谱自动分卷、谱系图打印和家谱打印等服务。

家谱人物查询服务提供了统一的人物查询入口，如图9所示，方便家谱人物的快速查找和后续操作。家谱树展示功能凭借

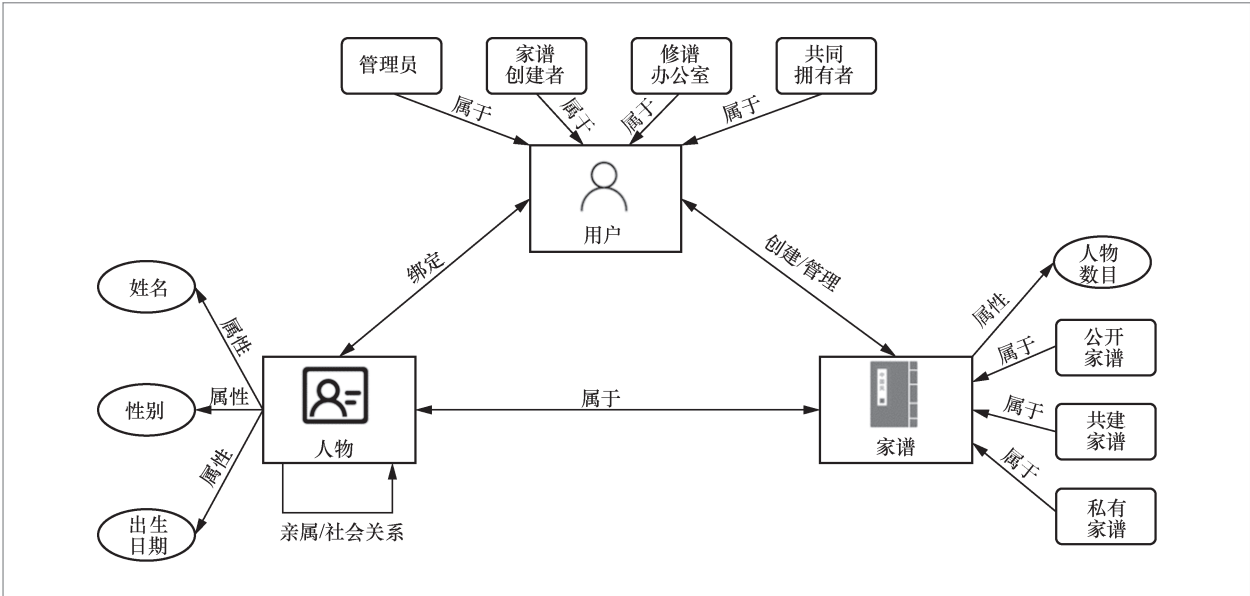


图 8 Huapu-KG 示例^[41]



图 9 家谱人物查询

华谱知识图谱，以树状图的形式向用户展示整个家谱结构，如图10所示，家族脉络清晰。家谱自动分卷服务适用于家谱立世人物较多的情况，系统会提供自动分卷功能，智能分配每卷人数。谱系图打印和家谱打印等服务可以离线文件的形式输出谱系

图便于修谱中的校对工作，如图11所示，也能够直接输出可打印的家谱文件。

4.5 华谱数据中台: 信息安全管理

在华谱数据中台的建设过程中，主要

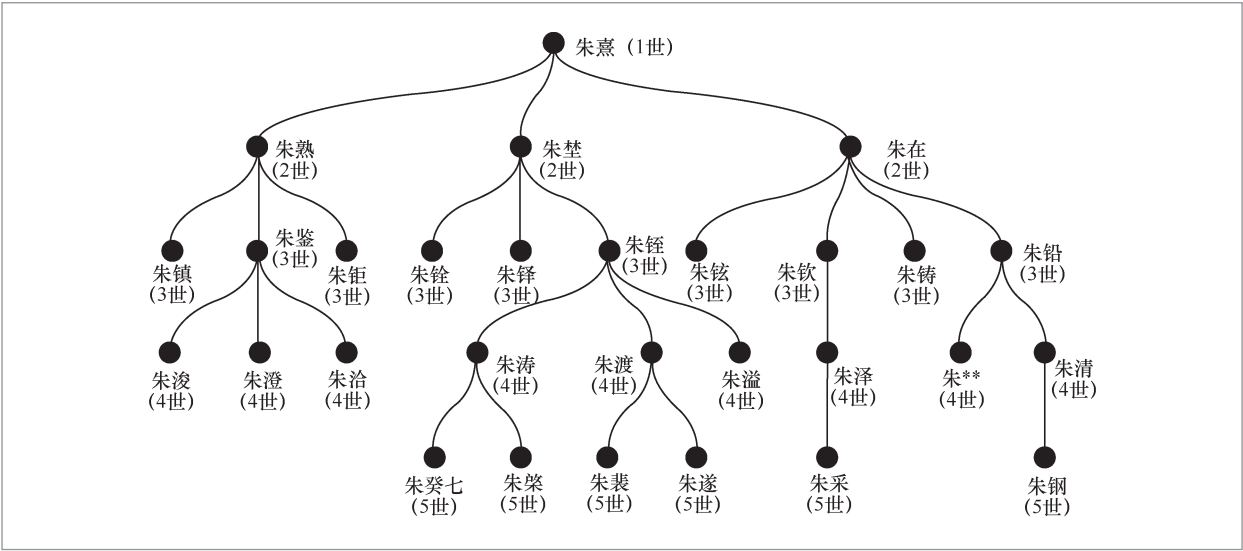
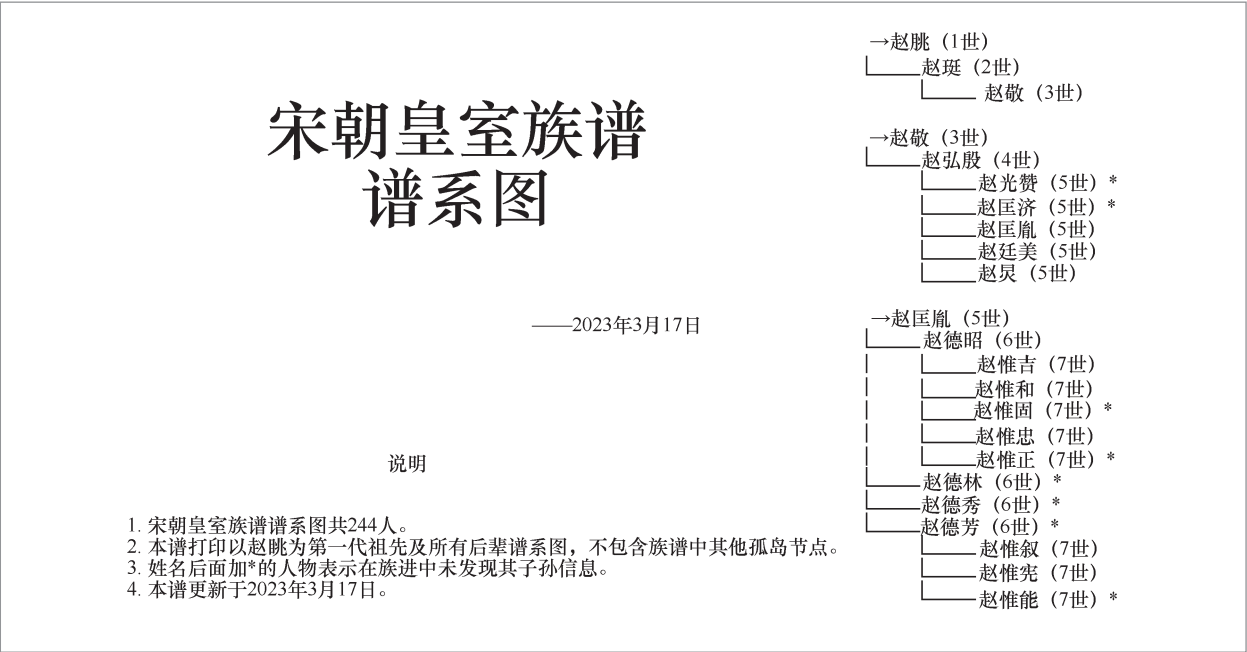


图 10 家谱树展示示例



权限也会动态变化。在这样复杂的权限变动中仅根据角色难以判断用户权限，需要进一步进行数据层面的权限判断。因此，如何管理好灵活变动的用户权限是一个重要的问题。

● 应用权限管理：华谱系统的应用都通过数据中台获取数据，但是不同系统能够获取的数据是不同的。限制数据中台对应用的数据获取权限保证数据安全，同时保证数据中台能够发挥最大限度的数据共享能力非常关键。

针对不一致数据的存储问题，华谱数据中台采用了基于超点的多源数据存储方案，如图12所示。超点是在保留原数据的情况下，将所有相同实体融合成一个新的节点。超点保留了各数据源的初始数据，做到全面保留信息以便对信息进行查缺补漏；通过人类智能校验，审查多条不一致数据是不是同一实体，保证了超点内信息的真实可靠，而且超点提供了信息溯源的依据，可以根据用户的需要和权限提供对

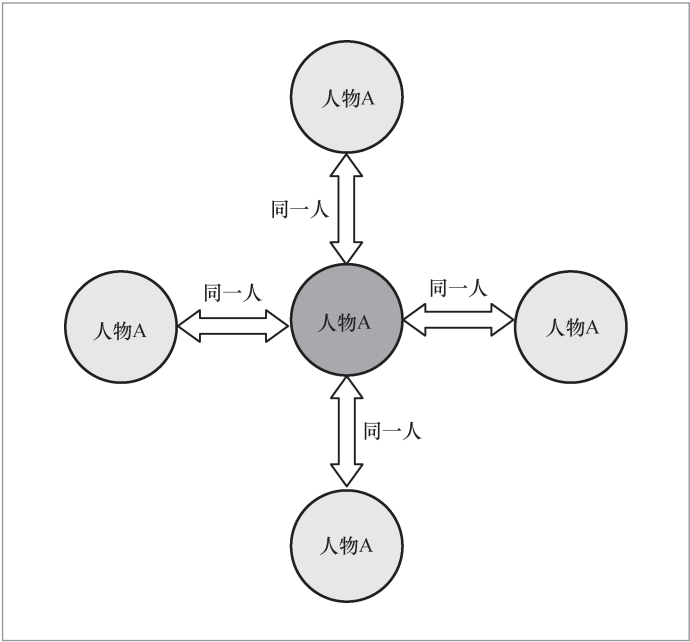


图 12 基于超点的多源数据存储方案

应实体最相关的信息。

针对用户权限的多重角色冲突问题和应用权限过高造成数据泄露的问题，笔者采用基于图数据库“粗细粒度结合”的权限管理方法和基于HAO模型权限管理的闭环架构，构建了统一的权限控制中心管理用户和应用权限^[13]。基于图数据库的“粗细粒度结合”的权限管理方法用粗粒度表示用户所拥有的角色，通过查询用户角色表和角色权限表实现，如果有权限则直接返回；否则，执行细粒度的权限查询。细粒度表示的是数据层面的权限管理，即某个具体的数据被设置的权限，权限控制通过查询图数据库中该数据被设定的权限边完成。“粗细粒度结合”的权限管理方法通过粗细粒度相结合的查询方法，既保证了访问控制管理的灵活性，又保证了用户权限的完备性，能够有效地解决同一用户多重角色冲突的问题。而且在细粒度查询中，图数据库用边的形式存储用户是否具有人物的修改权限，因此对细粒度数据查询的时间复杂度只有O(1)，具有良好的查询效率。“粗细粒度结合”的用户权限管理方法^[13]如图13所示。

基于HAO模型的权限管理闭环架构如图14所示，利用HI、AI、OI的协同作用，准确、灵活、可靠地分配多粒度下的权限。系统设计与维护人员发挥专家知识和组织通用标准决定角色定义、用户组划分等权限内容；AI基于日志进行用户分析，通过用户行为追踪、操作日志分析、用户画像等工作分析现有权限管理的不足之处，同时发掘操作行为较好的优质用户，数据分析结果交由系统设计维护人员重新定义权限；最后，系统管理员根据权限管理规定动态分配权限。

华谱系统中的应用或服务都通过数据中台的统一数据接口获取数据，用户通过请求应用或服务获取数据，为了实现对用

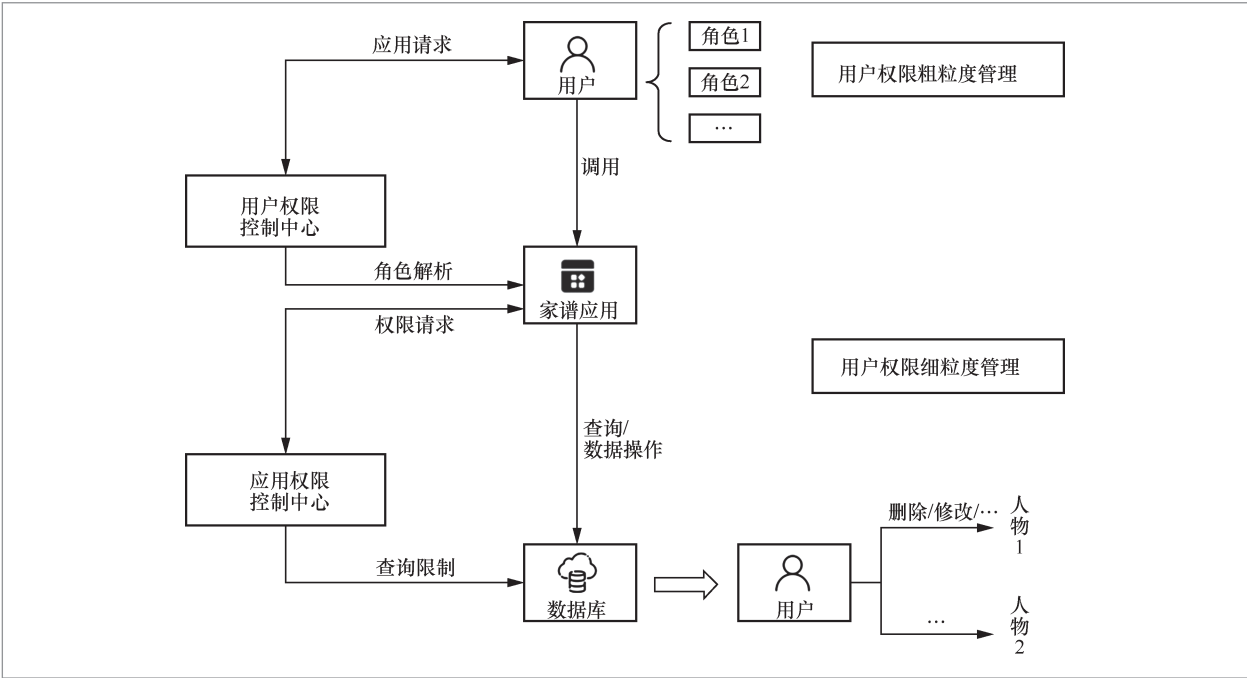


图 13 “粗细粒度结合”的用户权限管理方法 [13]

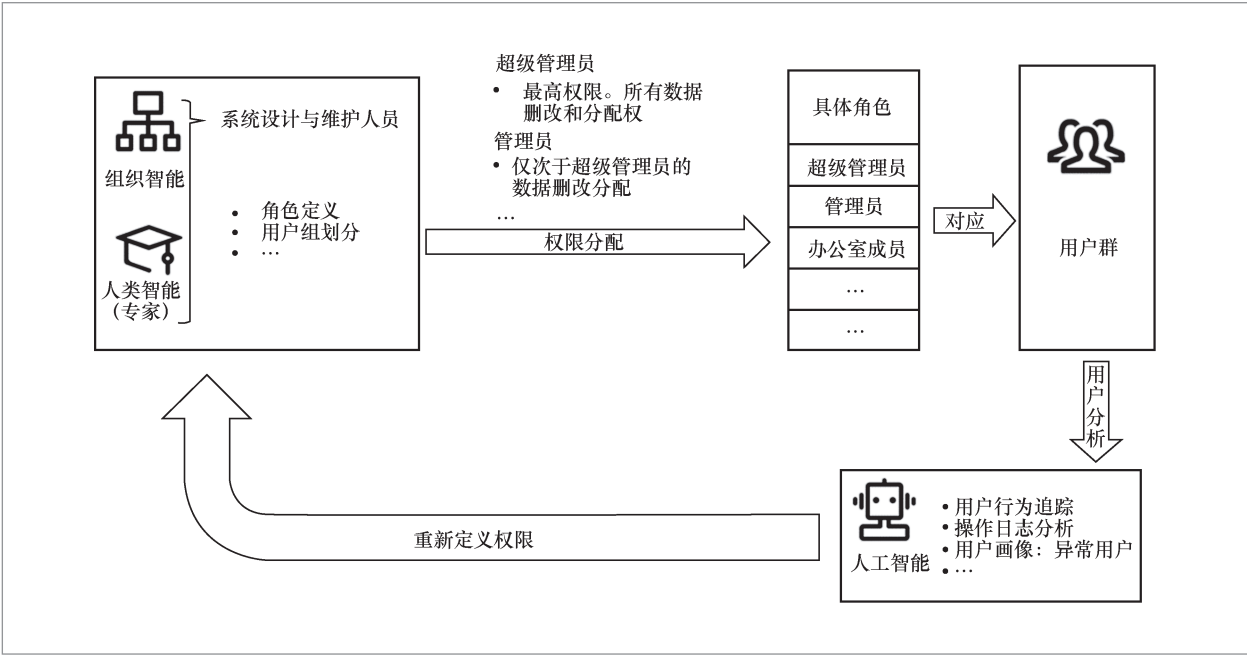


图 14 基于 HAO 模型的权限管理闭环架构 [13]

户权限、应用权限的统一管控，分别设置了用户权限控制中心和应用权限控制中心，

如图15所示，用户对应用/服务的请求会经过用户权限控制中心的验证，用户有权请

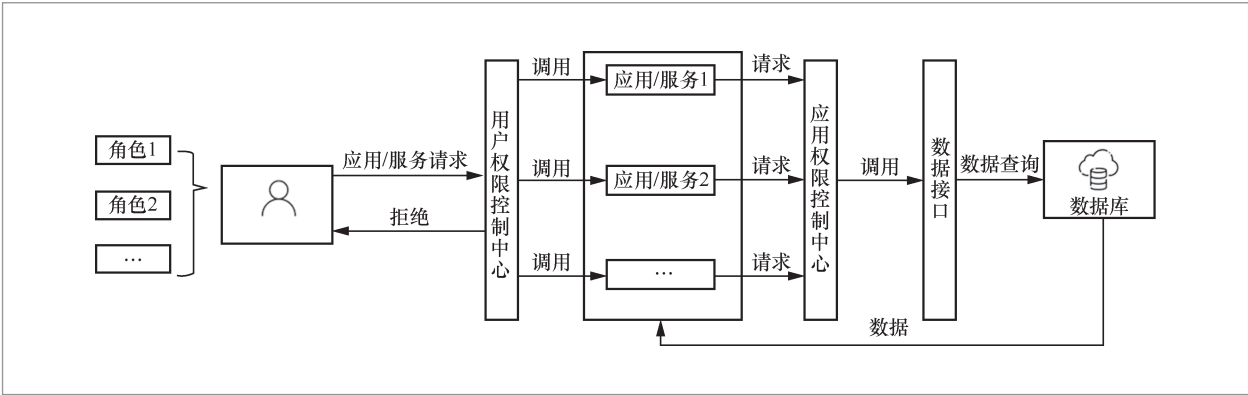


图 15 权限控制中心 [13]

求该服务，那么用户权限控制中心将调用请求应用/服务，如果验证不通过，用户请求将会被驳回。通过验证的用户请求调用的应用/服务将请求数据中台的数据接口，应用权限控制中心首先验证应用是否有使用该数据接口的权限，通过验证再由应用权限控制中心调用请求的数据接口，查询数据直接返回给应用。

5 挑战和前景展望

通过华谱数据中台的建设，笔者成功实践了基于HAO的家谱数据中台建设框架Huapu-CP。但在实践之中笔者也发现数据中台研究和开发还面临着以下挑战。

挑战1：广泛共享的理念和数据安全之间的矛盾。

广泛的数据共享是数据中台建设的初衷和目标之一。但要实现广泛的数据共享就需要提高分散数据源的数据开放程度，操作不慎可能会威胁到数据和隐私安全，得不偿失；如果过度重视数据和隐私安全导致“数据孤岛”则使数据中台的建设失去了意义。如何从实际业务环境中把握总台数据开放性的尺度是未来研究的重点和难点。

挑战2：难以将数据中台建设方案统一化。

虽然本文提出的中台框架的五大部分是绝大多数数据中台建设中要进行的，但却不能直接把这五大部分当成统一建设框架，原因有3个。其一，机构内原本的信息化建设程度不一，为了保持原系统的功能可持续同时接入统一的中台系统是难题；其二，机构内要整合的数据多源、异构，建设者必须要根据具体情况灵活使用技术、方法；其三，中台建设方使用中台的目的、业务场景多样，这就意味着必须根据实际业务需求调整中台的建设方向。

挑战3：数据中台建设的动态性、长期性。

数据中台的建设不是一朝一夕完成的。在大数据时代，业务需求会随着市场发生快速变化，建设者需要根据数据做出精细化管理，而组织数据、训练数据模型必须在实践中积累能力，中台沉淀这些能力为未来的业务和决策提供快速、高效的服务，因此数据中台是在动态中建设的，是一个长期性的工程。

挑战4：没有经验丰富的技术团队和成熟的检验工具、标准。

数据中台建设团队涉及业务、技术、管理部门之间的协作，单一的技术人才无法

胜任,需要更多的复合型人才。同时数据中台的建设缺少成熟的检验工具、标准,数据中台建设的优劣短期内可能仍旧停留在数据服务的效果这个单一评价指标上。

目前来看,数据中台依然有不错的前景,主要包括以下几个方面。

前景一:数据中台助力机构数字化转型将成为趋势。

对于海量数据的存储、管理和价值实现问题,数据中台展现了其可行性和有效性。这将促进越来越多的数字化转型机构选择建设数据中台。

前景二:数据中台产品逐渐具备标准化潜力。

随着近年来选择建设数据中台的领域越来越多,其中包括互联网、零售、物联网、政府部门、城市建设等,数据中台建设的总体经验越来越多,相信未来关于数据中台建设的理论一定会取得长足的进步。

前景三:促进机构内团队协作能力。

数据中台打破“数据孤岛”现象,通过数据将业务、技术、组织等相关人员联系在一起,加强了跨部门之间的交流,提高了团队协作能力,由技术上的数字化转型推动了组织结构上的平台化转型。

前景四:促进大数据、云计算及人工智能技术发展。

数据中台推动了海量数据在机构内的集成,为人工智能技术提供了数据资源。庞大的数据将会推动大数据治理技术的升级,海量的运算需求也会增加云计算的需要。

前景五:知识图谱技术对新一代数据中台技术的推动作用。

传统的二维表在知识表达上有很多局限性,而知识图谱则能更好地描述实体与关系,复杂的图结构更有利于探索数据之间的关联,获取知识。新一代数据中台技

术不仅要融合数据,还要存取知识、使用知识,因此知识图谱技术将对新一代数据中台的建设起到推动作用。

6 结束语

数据中台是数字化转型中的一个技术热点,目前数据中台面临理论不完善、概念不统一、建设方案差异大的问题。本文总结了数据中台的相关研究背景,叙述了数据中台的概念,接着分析了数据中台的架构方案差异和建设目标,给出了数据中台的建设框架,对建设框架中的物理管理、逻辑管理、数据资产管理、数据服务、信息安全管理做出介绍。然后以华谱系统的建设为例证明了基于 HAO 智能的家谱数据中台框架Huapu-CP的实际可行性。最后介绍了当前数据中台建设的挑战和前景。

参考文献:

- [1] 钟华. 企业IT架构转型之道: 阿里巴巴中台战略思想与架构实战[M]. 北京: 机械工业出版社, 2017.
ZHONG H. The way of enterprise IT architecture transformation: Alibaba's middle office strategic ideas and architecture practice[M]. Beijing: China Machine Press, 2017.
- [2] 艾瑞咨询. 2021年中国数据中台行业白皮书[R]. 2021.
iResearch. China data center industry white paper 2021[R]. 2021.
- [3] 刘长杰. 数据中台建设将把数字中国推向纵深: 与国务院发展研究中心研究员李广乾对话[J]. 中国发展观察, 2020(1): 86-90.
LIU C J. The construction of data center will push digital China deeper—dialogue with Li Guanggan, a researcher at the State Council Development

- Research Center[J]. China Development Observation, 2020(1): 86–90.
- [4] 缪翀莺, 谭华, 易学明. 数据中台的定位和架构分析[J]. 广东通信技术, 2019, 39(12): 57–62, 70.
- MIAO C Y, TAN H, YI X M. Positioning and architecture analysis of data center station[J]. Guangdong Communication Technology, 2019, 39(12): 57–62, 70.
- [5] 刘颖慧, 刘楠, 蔡一欣, 等. 数字化转型中不同企业的中台战略及架构设计[J]. 电信科学, 2020, 36(7): 126–135.
- LIU Y H, LIU N, CAI Y X, et al. Design of the strategy and architecture of different enterprises' middle platform[J]. Telecommunications Science, 2020, 36(7): 126–135.
- [6] 邓中华. 大数据大创新: 阿里巴巴云上数据中台之道[M]. 北京: 电子工业出版社, 2018.
- DENG Z H. Big data, big innovation: the way of Alibaba's data center on the cloud [M]. Beijing: Publishing House of Electronics Industry, 2018.
- [7] 林鸿, 方学民, 袁葆, 等. 电力物联网多渠道客户服务中台战略研究与设计[J]. 供用电, 2019, 36(6): 39–45, 66.
- LIN H, FANG X M, YUAN B, et al. Research and design of business platform strategy for multi-channel customer services in Internet of Things in electricity[J]. Distribution & Utilization, 2019, 36(6): 39–45, 66.
- [8] 苏萌, 贾喜顺, 杜晓梦, 等. 数据中台技术相关进展及发展趋势[J]. 数据与计算发展前沿, 2019, 1(1): 116–126.
- SU M, JIA X S, DU X M, et al. Research on the recent development and future trends of data mid-end technology[J]. Frontiers of Data&Computing, 2019, 1(1): 116–126.
- [9] ATLAS A. Data governance and metadata framework for Hadoop[J]. Data Governance and Metadata Framework for Hadoop, 2019.
- [10] 刘天. 中台产品经理宝典: 从业务建模到中台设计全攻略[M]. 北京: 电子工业出版社, 2020.
- Liu T. A guide for middle-office product managers: from business modeling to middle-office design[M]. Beijing: Publishing House of Electronics Industry, 2020.
- [11] 邓立君. 数据中台与大数据中心分析[J]. 电子世界, 2019(22): 85–86.
- DENG L J. Analysis of data center and big data center[J]. Electronics World, 2019(22): 85–86.
- [12] 陈新宇, 罗家鹰, 江威. 中台实践: 数字化转型方法论与解决方案[M]. 北京: 机械工业出版社, 2020.
- CHEN X Y, LUO J Y, JIANG W. Middle-platform practice: digital transformation methodology and solutions[M]. Beijing: China Machine Press, 2020.
- [13] 吴信东, 盛绍静, 蒋婷婷, 等. 从知识图谱到数据中台: 华谱系统[J]. 自动化学报, 2020, 46(10): 2045–2059.
- WU X D, SHENG S J, JIANG T T, et al. Huapu-CP: from knowledge graphs to a data central-platform[J]. Acta Automatica Sinica, 2020, 46(10): 2045–2059.
- [14] 王宇杰, 武宪杰, 傅子涛. 基于云计算的空管系统设计方法研究[J]. 电子技术与软件工程, 2021(17): 141–144.
- WANG Y J, WU X J, FU Z T. Research on the design method of air traffic control system based on cloud computing[J]. Electronic Technology & Software Engineering, 2021(17): 141–144.
- [15] 克里斯·理查森. 微服务架构设计模式[M]. 喻勇, 译. 北京: 机械工业出版社, 2019.
- RICHARDSON C. Microservices patterns: with examples in Java[M]. Translated by YU Y. Beijing: China Machine Press, 2019.
- [16] 彭锋, 宋文欣, 孙浩峰. 云原生数据中台: 架构、方法论与实践[M]. 北京: 机械工业出版社, 2021.
- PENG F, SONG W X, SUN H F. Cloud native data center: architecture,

- methodology and practice[M]. Beijing: China Machine Press, 2021.
- [17] CNCF. Cloud native landscape[Z]. 2021.
- [18] 用友云平台团队. 数字化中台[M]. 北京: 电子工业出版社, 2021.
- Yongyou Cloud Platform Team. Digital Middleware[M]. Beijing: Publishing House of Electronics Industry, 2021.
- [19] 中国信通院大数据技术标准推进委员会. 数据标准管理实践白皮书[R]. 2022.
- Big Data Technical Standards Promotion Committee of China ICT Academy. White paper on data standards management practices[R]. 2022.
- [20] 宋俊典, 戴炳荣, 蒋丽雯, 等. 基于区块链的数据治理协同方法[J]. 计算机应用, 2018, 38(9): 2500–2506.
- SONG J D, DAI B R, JIANG L W, et al. Data governance collaborative method based on blockchain[J]. Journal of Computer Applications, 2018, 38(9): 2500–2506.
- [21] 毕强, 朱亚玲. 元数据标准及其互操作研究[J]. 情报理论与实践, 2007, 30(5): 666–670.
- BI Q, ZHU Y L. Research on metadata standard and its interoperability[J]. Information Studies (Theory & Application), 2007, 30(5): 666–670.
- [22] 国家质量监督检验检疫总局, 中国国家标准化管理委员会. 数据管理能力成熟度评估模型: GB/T 36073—2018[S]. 2018.
- General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. Data management capability maturity assessment model: GB/T 36073—2018[S]. 2018.
- [23] 吴信东, 董丙冰, 堵新政, 等. 数据治理技术[J]. 软件学报, 2019, 30(9): 2830–2856.
- WU X D, DONG B B, DU X Z, et al. Data governance technology[J]. Journal of Software, 2019, 30(9): 2830–2856.
- [24] 刘峰, 杨杰, 李志斌, 等. 一种面向双中台双链架构的内生性数据安全交互协议研究[J]. 华东师范大学学报(自然科学版), 2020(5): 44–55.
- LIU F, YANG J, LI Z B, et al. Research on an endogenous data interaction protocol for the dual-middle platform and dual-chain architecture[J]. Journal of East China Normal University (Natural Science), 2020(5): 44–55.
- [25] DOAN A, HALEVY A, IVES Z. Data matching[M]//Principles of data integration. Amsterdam: Elsevier, 2012: 173–207.
- [26] CHAUDHURI S, DAYAL U. An overview of data warehousing and OLAP technology[J]. ACM SIGMOD Record, 1997, 26(1): 65–74.
- [27] 杜小勇, 陈跃国, 范举, 等. 数据整理: 大数据治理的关键技术[J]. 大数据, 2019, 5(3): 13–22.
- DU X Y, CHEN Y G, FAN J, et al. Data wrangling: a key technique of data governance[J]. Big Data Research, 2019, 5(3): 13–22.
- [28] ASFAND-E-YAR M, ALI R. Semantic integration of heterogeneous databases of same domain using ontology[J]. IEEE Access, 2020, 8: 77903–77919.
- [29] 吴强, 沙宇恒, 邓春宇, 等. 基于本体论的项目全景模型建模及其在电网项目管理中的应用[J]. 电力系统及其自动化学报, 2021, 33(11): 135–144, 150.
- WU Q, SHA Y H, DENG C Y, et al. Panoramic unified information modeling based on ontology and its application to power grid project management[J]. Proceedings of the CSU-EPSC, 2021, 33(11): 135–144, 150.
- [30] 万齐鸣, 王思宁, 何鑫. 数据中台SG-CIM模型应用方法[J]. 电信科学, 2020, 36(3): 136–143.
- WAN Q M, WANG S N, HE X. SG-CIM model application method in data middle platform[J]. Telecommunications Science, 2020, 36(3): 136–143.

- [31] 中国信通院大数据技术标准推进委员会. 数据资产管理实践白皮书4.0[R]. 2021.
Big Data Technology Standard Promotion Committee of China ICT Academy. Data asset management practice white paper 4.0[R]. 2021.
- [32] 詹盈. 算法与数据中台: 基于Google、Facebook与微博实践[M]. 北京: 电子工业出版社, 2020.
ZHAN Y. Algorithms and Data Center: based on Google, Facebook and Microblogging practice[M]. Beijing: Publishing House of Electronics Industry, 2020.
- [33] 李信鹏, 刘威, 杨智萍, 等. 电网企业数据中台方案研究[J]. 电力信息与通信技术, 2020, 18(2): 1-8.
LI X P, LIU W, YANG Z P, et al. Research on data mid-platform scheme of power grid enterprises[J]. Electric Power Information and Communication Technology, 2020, 18(2): 1-8.
- [34] 吴信东, 何进, 陆汝钤, 等. 从大数据到大知识: HACE+BigKE[J]. 自动化学报, 2016, 42(7): 965-982.
WU X D, HE J, LU R Q, et al. From big data to big knowledge: HACE + BigKE[J]. Acta Automatica Sinica, 2016, 42(7): 965-982.
- [35] WU M H, WU X D. On big wisdom[J]. Knowledge and Information Systems, 2019, 58(1): 1-8.
- [36] 吴信东, 嵇圣础. MapReduce与Spark用于大数据分析之比较[J]. 软件学报, 2018, 29(6): 1770-1791.
WU X D, JI S W. Comparative study on MapReduce and spark for big data analytics[J]. Journal of Software, 2018, 29(6): 1770-1791.
- [37] JI S W, BU C Y, LI L, et al. Local graph edge partitioning with a two-stage heuristic method[C]//Proceedings of 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2019: 228-237.
- [38] 吴共庆, 胡骏, 李莉, 等. 基于标签路径特征融合的在线Web新闻内容抽取[J]. 软件学报, 2016, 27(3): 714-735.
WU G Q, HU J, LI L, et al. Online web news extraction via tag path feature fusion[J]. Journal of Software, 2016, 27(3): 714-735.
- [39] SHENG S J, ZHOU P, WU X D. CEPV: a tree structure information extraction and visualization tool for big knowledge graph[C]//Proceedings of 2019 IEEE International Conference on Big Knowledge (ICBK). Piscataway: IEEE Press, 2019: 221-228.
- [40] 吴信东, 李娇, 周鹏, 等. 碎片化家谱数据的融合技术[J]. 软件学报, 2021, 32(9): 2816-2836.
WU X D, LI J, ZHOU P, et al. Fusion technique for fragmented genealogy data[J]. Journal of Software, 2021, 32(9): 2816-2836.
- [41] WU X D, JIANG T T, ZHU Y, et al. Knowledge graph for China's genealogy[J]. IEEE Transactions on Knowledge and Data Engineering, 2021: 1.

作者简介



吴信东(1963-),男,博士,合肥工业大学计算机与信息学院教授、博士生导师,主要研究方向为数据挖掘、大数据分析、知识工程。



应泽宇 (1999-), 男, 合肥工业大学计算机与信息学院硕士生, 主要研究方向为数据挖掘、用户画像。



盛绍静 (1994-), 女, 合肥工业大学计算机与信息学院博士生, 主要研究方向为数据挖掘、知识图谱、家谱划分。



蒋婷婷 (1995-), 女, 博士, 毕业于合肥工业大学计算机与信息学院, 主要研究方向为知识图谱、知识表示学习、实体对齐。



卜晨阳 (1992-), 男, 博士, 合肥工业大学计算机与信息学院副教授, 主要研究方向为知识图谱构建和多模态融合。



张赞 (1987-), 男, 博士, 合肥工业大学计算机与信息学院副教授, 主要研究方向为因果学习和弱监督学习。

收稿日期: 2022-09-07

基金项目: 国家自然科学基金资助项目 (No.62120106008, No.91746209); 中央高校基本科研业务费资助项目 (No. JZ2023HGTB0270); 教育部创新团队项目 (No.IRT17R3)

Foundation Items: The National Natural Science Foundation of China (No.62120106008, No.91746209), The Fundamental Research Funds for the Central Universities (No.JZ2023HGTB0270), The Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education (No.IRT17R3)