

数据空间:一种新的数据组织和管理模式

范淑焕, 侯孟书

引用本文

范淑焕, 侯孟书. 数据空间:一种新的数据组织和管理模式[J]. 计算机科学, 2023, 50(5): 115-127.

FAN Shuhuan, HOU Mengshu. Dataspace:A New Data Organization and Management Mode[J]. Computer Science, 2023, 50(5): 115-127.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[生物标志物智能识别关键技术:环状RNA与疾病关联预测研究综述](#)

Key Technologies of Intelligent Identification of Biomarkers:Review of Research on Association Prediction Between Circular RNA and Disease

计算机科学, 2023, 50(4): 369-387. <https://doi.org/10.11896/jsjcx.220500114>

[医学知识图谱研究与应用综述](#)

Survey of Medical Knowledge Graph Research and Application

计算机科学, 2023, 50(3): 83-93. <https://doi.org/10.11896/jsjcx.220700241>

[基于水车模型的时序大数据快速存储](#)

Fast Storage System for Time-series Big Data Streams Based on Waterwheel Model

计算机科学, 2023, 50(1): 25-33. <https://doi.org/10.11896/jsjcx.220900045>

[基于人工蜂群算法的多维函数优化加速方法](#)

Acceleration Method for Multidimensional Function Optimization Based on Artificial Bee Colony Algorithm

计算机科学, 2022, 49(11A): 211200075-6. <https://doi.org/10.11896/jsjcx.211200075>

[基于申威众核处理器的Office口令恢复向量化研究](#)

Study on Office Password Recovery Vectorization Technology Based on Sunway Many-core Processor

计算机科学, 2022, 49(11A): 210900176-5. <https://doi.org/10.11896/jsjcx.210900176>

数据空间：一种新的数据组织和管理模式

范淑焕 侯孟书

电子科技大学计算机科学与工程学院 成都 611731

(fansh@uestc.edu.cn)

摘要 随着数字经济的快速发展,如何实现非可信环境下的多方数据融合,为跨组织场景的数据共享、数据分析以及数据服务寻找新途径,成为了社会数字化产业升级中面临的新问题。数据空间为解决这些问题带来了新思路。文中回顾了数据的组织和管理发展历程,指出在大数据背景下数据空间的系统研究具有紧迫性和重要性,分析了数据空间的内涵并进行了形式化描述,提出了基于数据空间的大数据平台架构,总结描述了3类经典的应用场景。围绕数据空间的构建工作,从数据建模、动态演变、数据查询处理、安全与隐私拓展方面分析了当前的关联研究问题和主要技术方法,简述了数据空间在不同领域的实现和应用情况。最后从多模态数据融合、高效的查询处理、数据的安全共享及基于数据空间的大数据平台构建分析方面展望了研究前景和挑战。

关键词 数据空间;大数据;数据共享;数据建模;动态演变;数据查询;安全与隐私

中图法分类号 TP311

Dataspace: A New Data Organization and Management Model

FAN Shuhuan and HOU Mengshu

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract With the rapid development of the digital economy, how to realize multi-party data fusion in an untrusted environment and find new ways for data sharing, data analysis and data services in cross-organizational scenarios has become a new problem in the upgrading of social digital industries. Dataspace brings new ideas to solve these problems. The development history of data organization and management is reviewed, and it points out that in the background of big data, systematic research on dataspace is urgent and important. The connotation of dataspace is analyzed and a formal description is given. A big data platform architecture based on dataspace is proposed, and three classic application scenarios are briefly described. Focusing on the construction of dataspace, it analyzes the current correlation research issues and main technical methods from data modeling, dynamic evolution, data query processing, security and privacy, and briefly describes the realization and application of dataspace in different fields. Finally, the research outlook and challenges are prospected from the perspective of multimodal data fusion, efficient query processing, safe data sharing, and the construction of a big data platform based on dataspace.

Keywords Dataspace, Big data, Data sharing, Data modeling, Dynamic evolution, Data query, Security and privacy

1 引言

当今社会,数据充斥着人们生产生活的方方面面,如个人生活数据包括学历数据、医疗数据、工作数据、保险数据、车辆数据、消费数据等,制造业企业中的数据包括产品设计数据、制造数据、销售数据、库存数据、维修数据等。这些数据都存储在物理隔离、跨组织的独立数据库系统中,它们由于在构建时期围绕各自独立的业务目标而形成了数据孤岛。近年来,数字驱动、场景融合已经成为社会发展的新趋势,也必将重构我们的社会生活,而这背后最核心的工作就是突破数据孤岛,实现跨部门、跨企业、跨行业数据的有效组织和管理。

如图1所示,数据组织和管理历程,经历了两个重要时期:

1)聚焦解决高效数据存取问题的数据库系统发展期;2)聚焦挖掘数据二次应用价值的数字化时期,特别是延续到现在的大数据时代。早期的数据库管理系统可以追溯到20世纪60年代,数据库中的数据是以树状或者网状结构进行组织管理的,实现了将数据的存储和访问功能从应用程序中分离出来。20世纪70年代,IBM的埃德加·科德首次提出了数据库关系模型的概念,其基础模型包括行存储、B树索引、基于代价的优化器以及ACID的事务属性。该模型实现了数据的物理存储和概念表示的分离,并提供了数据抽象和应用的灵活性。在之后十来年的时间,它支撑了几乎所有的数据库应用场景。后来,有学者提出“No one sizes fits all”^[1],因为市场上出现了大量的NoSQL数据库,如键值存储数据库、基于列或宽列的数据

到稿日期:2022-07-04 返修日期:2022-11-09

基金项目:国家重点研发计划(2019YFB1705601);国家自然科学基金(62072075)

This work was supported by the National Key R & D Program of China(2019YFB1705601) and National Natural Science Foundation of China(62072075).

通信作者:侯孟书(mshou@uestc.edu.cn)

库、图数据库等。它们不强调即时数据的一致性和结构化存储,而更侧重于半结构化数据存储、高性能、可用性、数据复制和可扩展性。这类数据库大多是分布式数据库或者分布式存储系统,如分布式文件系统(GFS 和 HDFS)。随着各类新型

数据库的快速应用,数据库系统的异构性和复杂性越来越凸显,且数据库管理系统中存储的数据量也呈指数级增长,形成了海量数据资源。如何对这些异构的数据库进行管理应用,从而让这些海量数据产生价值,是一个亟待解决的问题。

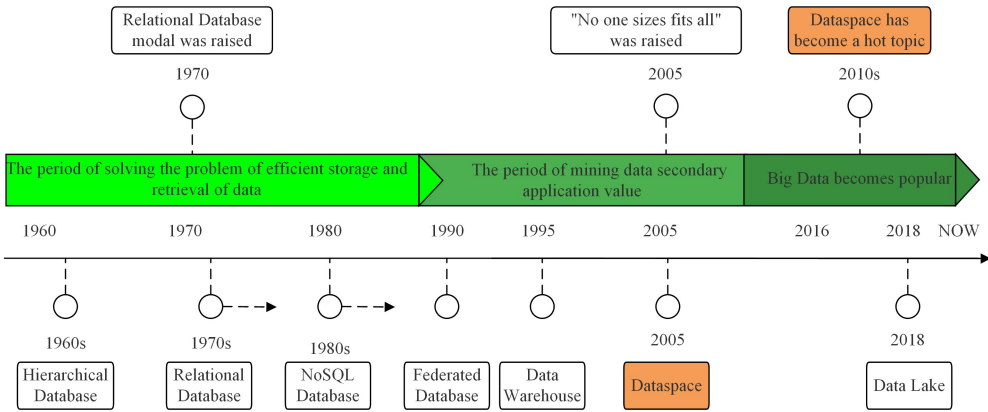


图 1 数据的组织与管理历史

Fig. 1 History of data organization and management

20 世纪 90 年代开始,人们已经关注到了对数据的二次价值挖掘。一般来说,有两种方式:一种是以数据仓库^[2]为代表的统一数据集成,将所有数据源中的数据通过 ETL 或 API 抽取集成到一个数据库中;而另一种方法数据联邦则无需将所有数据源数据都拷贝应用,其通过虚拟方式实现了数据的联邦应用^[3]。数据仓库对企业内部大量的应用系统、业务数据进行集成,通过关联分析,最终为企业决策提供支持,然而这种方式在前期及运行维护过程中都需要大量的投入成本。而数据联邦一般基于一致性查询语言从数据源获取数据,如联邦数据库系统^[4]、中间件^[5]、Polystores^[6]。近年来,类数据联邦的数据组织管理需求越来越多,在学术领域和工业领域也出现了一批数据联邦系统^[7]。除此之外,针对多源异构数据资产的管理,数据湖也提供了一种解决方案^[8],一方面提供了具有良好扩展性和可靠性的存储资源来满足数据的入湖需求,另一方面围绕数据处理提供了相关的工具,以完成数据的具体入湖和数据价值挖掘分析。

上述方案基本都围绕企业内部的数据组织管理,而跨企业甚至跨行业的数据共享亟待释放价值,数据空间提供了一种更通用的多源异构数据组织管理方式,它同时支持可信与非可信环境下的松散数据源组织管理,不关心数据原本的存储和数据结构,也不依赖于具体的某种算法和模型,使应用重点聚焦在数据的价值上。本文将数据空间与其他组织与管理模式进行了多维度对比,结果如表 1 所列,显然,数据空间是唯一支持跨企业、跨平台的数据对象的组织与管理的管

方式,它通过分布式的方式支撑大数据背景下的数据价值挖掘需求,同时相比传统的方法实现了一种更轻量级 pay-as-you-go 的数据融合,为大数据背景需求下的跨企业数据融合提供了新思路。2005 年,Franklin 等提出了数据空间的概念,并获得了学术界的关注^[9]。2008 年,文献^[10]指出,未来数据空间的研究包括 5 个方面:模式映射和匹配、索引融合、数据空间剖析、数据血缘和溯源、信息抽取和关键字查找。同年,Li 等阐述了数据空间概念及特性,从数据模型、数据集成、数据查询、数据更新、存储和索引、数据演化和系统实现等方面对当时的研究工作进行了梳理^[11]。2011 年,Singh 对数据空间的架构、模型、查询响应技术和索引进行了研究分析^[12]。近年来,关于数据空间的研究和应用越来越受到学术界和工业界的关注,解决了跨企业的数据融合、共享和分析,如在工业领域^[13-15]和医学领域^[16-17]等。同时,日益严格的隐私安全保护要求和海量的数据存量使得数据分散异质的多方之间不能随意共享数据,加剧了“数据孤岛问题”,而在这种跨企业数据孤岛现象和数据融合应用需求矛盾逐渐凸显的背景下,需要对数据空间相关技术进行更加深入、系统的研究。本文系统地研究梳理了 2005 年至今数据空间相关的研究成果,特别是近十年来的关联技术发展及领域应用,深入剖析了其在数据组织与管理中的内涵和问题场景,对其核心概念进行了形式化描述,通过数据建模、动态演变、数据的查询处理、安全及隐私分析了当前的关联技术研究方向及技术方法,简述了其领域应用,并对其研究前景和挑战进行了展望。

表 1 数据空间与其他模式的对比

Table 1 Dataspace vs other models

	Database	Federated database	Data warehouse	Data lake	Dataspace
Data Object	Certain Business Data	Inner-enterprise Cross-platform	Inner-enterprise Cross-platform	Inner-enterprise	Cross-enterprise Cross-platform
Data Storage	Centralized	Distributed	Centralized	Centralized	Distributed-centralized
Data type	A fixed data model	Multi-model	Relational	Multi-model	Multi-model
Data Schema	Schema-before	Schema-before	Schema-before	No	Schema-after
Data Sovereignty	All	Part	All	All	Part
Data Fusion Mode	No	No	Early one-time	No	Pay-as-you-go
Prophase integration cost	No	No	Big	No	Small

2 数据空间概述

2.1 定义及内涵

Franklin 于 2005 年对数据空间提出了最原始的定义^[9],数据空间不是一种数据集成方式,而更像是一种数据共存方法。它不在乎数据的集成程度,而是以一种 pay-as-you-go 的方式不断进化,对外提供服务。它更加聚焦在数据的价值应用上,在此基础上深入剖析数据空间的应用场景和目标,对数据空间进行拓展定义。

数据空间是一种通用的多源异构数据组织和管理模式,它实现了非可信环境下的多方数据的按需融合,为跨组织场景的数据共享、数据分析以及数据服务提供了新途径。

围绕上述定义,对相关概念展开了进一步界定。数据空间是与传统的以企业管理为主导的数据融合模式相对应的一种新模式,以数据价值联合挖掘为目标,力图实现跨组织的数据融合,为分散的大数据价值挖掘以及隐私数据挖掘提供了新方法。例如,为了更好地拓展某区医疗业务服务和科研支撑等应用,某单位拟构建区医疗大数据平台,覆盖区管辖下的十余所医院,很现实的问题就是如何让患者的隐私数据在不流出医院的前提下实现信息共享,这是一个经典的数据空间需求场景。显然,数据空间的参与方众多且不可控,其数据库的异构性强且数据价值挖掘目标明确,这些都导致数据空间的数据模型多样化,数据量巨大,数据传输频繁,数据分散存储。在此非可信环境下,跨组织的参与方彼此信任不足,既影响了信息的真实性和可靠性,也给数据空间构建带来了严峻的挑战。按需融合指在对外提供服务的过程中根据应用需求不断加强融合程度,在极端情况下,最终会达到传统前期集中集成的效果,而且在前期建设和持续维护的过程中大大降低了投入成本。跨组织指参与方数据归属权不再是传统的单个企业内部,而是分属于不同的组织,这些组织可能是制造业价值链上有合作的供需关系,也可能是有竞争关系的同类型车辆配件提供商,甚至是没有任何关系的两个独立组织,如车辆销售公司和保险公司。它们通过数据空间利用数据联合挖掘新的场景应用价值。数据融合指在数据的组织层实现数据的有效访问和管理,最终实现数据的共享和分析。

2.2 数据空间的形式化描述

文献[18]围绕数据空间建模的相关概念进行了定义,本文将在在此基础上对其进行拓展,并对数据空间支持系统(DSSP)进行描述,其具体定义如下。

定义 1(属性) 给定一个数据空间 ds ,定义属性为 $a=(ds,db,col,name)$,其中 ds 为数据空间名称, db 是其所属数据库名称, col 表示属性所在数据库中的数据集合,如关系型数据库中代表其所在的表名称,列存数据库中代表其所在的列族,文档数据库就是其所在的文档集合, $name$ 指属性名称。同时,定义 A^* 为 ds 中的属性全集。

定义 2(特征) 特征指一系列的属性集合,这些属性之间通过映射关联,定义为 $f=(name,a,A)$,其中 $name$ 指特征名称, a 是特征的代表属性, A 指所有和 a 代表同一概念的属性序列。

定义 3(数据空间) 一系列的特征集合组成了数据

空间, $ds=(f_1,f_2,\cdots,f_i,\cdots)$ 。

数据空间的创建、操作和管理,都要基于物理构建来完成,也就是数据空间支持系统(DSSP),其定义如下。

定义 4(数据空间支持系统) 数据空间支持系统实现对空间数据的组织、管理和应用,定义为 $dssp=(S,DS,manager,O,U,APP)$ 。 S 是构建空间的参与方数据源序列, $S=(s_1,s_2,s_3,\cdots)$,每个数据源表示为 $s=(name,A,D,dr)$, $name$ 是数据源名称, A' 是数据源的属性集合, $A=A_1'\cup A_2'\cup A_3'\cup\cdots$ 。 DS 是所有的数据空间序列, $DS=(ds_1,ds_2,ds_3,\cdots)$,不同的数据空间可能有特征交集,即 $ds_1\cap ds_2\neq\emptyset$ 。 $manager$ 指数据空间管理系统,一是实现参与方数据源的注册、接入、连接和维护管理;二是实现数据空间管理,包括空间创建人、空间有效期及空间的应用授权等;三是实现存储管理,包括缓存管理、元数据管理、算法模型等;四是实现安全管理,包括加密算法、安全策略、隐私策略等。 O 是系统管理员及其分层权限。 U 是系统的使用用户及其分层权限。 APP 是基于支持系统构建的应用序列及其与数据空间的授权关系,数据空间和应用的授权绑定是多对多的关系,即一个数据空间可能同时支撑了几个应用,一个应用也可能被多个数据空间同时支撑。

2.3 数据空间的特点

结合数据空间的定义,认为数据空间应该具有以下特点:

(1)不可控。由于数据空间中的每个参与方都有自己的主管组织、独立部署的硬件设备、独立构建的软件平台和独立运行的应用业务,因此具有很强的自主性。又因为参与方之间可能存在的竞争关系以及分散分布的客观环境,所以数据空间在一定程度上具有不可控性,如恶意参与方试图通过推理攻击来获得其他参与方的隐私数据,也可能通过投毒攻击来破坏空间数据。此外,参与方本身的实时负载情况、网络情况、可贡献的计算能力、数据质量都不可控。

(2)动态演变。指数据空间长时间处于动态变化的过程中,其演化来源包括两个方面:1)参与方数据源的动态性,可能有现有源的退出,也可能有新数据源的加入,对于各个源节点来说,也会有数据更新和数据库模式的变化;2)数据空间系统的动态性,数据空间的构建是一个 pay-as-you-go 的过程,随着空间的运行处于不断地演变过程中。

(3)数据属性多样。数据空间实现的是对数据的二次价值应用,因此在对数据的操作权限上,只有可读权限而无更新权限。同时,在数据空间的数据应用中,数据源对于用户来说都是透明的,数据更新时间、操作记录等维度对源数据来说也是不受关注的,然而这些信息对于数据空间中的数据组织、共享和分析来说却是非常重要的,无论是实现数据联合查询、数据追溯还是联合学习,丰富多样的数据属性维度都是一个必要条件。

2.4 基于数据空间的大数据平台架构

数据空间是在大数据背景下,为了充分挖掘数据应用价值而产生的一种数据组织和管理模式,图 2 给出了基于数据空间的大数据平台架构。从图中可以看到,整体的架构包括两个端。参与方端包含了来自不同企业、医院等领域单位、开放平台、政府部门等的数据源提供方,根据各自在角色中的

平台定位,提供了各种不同类型的数据信息,构成了平台的数据源层。中心端实现了基于数据空间的大数据平台,包含了丰富存储组件的存储层,其存储需要中心端存储的元数据、

不同类型的缓存数据、历史数据信息,并将这些数据输入核心层,同时也实现了缓存数据、热点数据、用户反馈等信息的写入。应用层则是基于核心层主要实现 3 类场景下的数据服务。

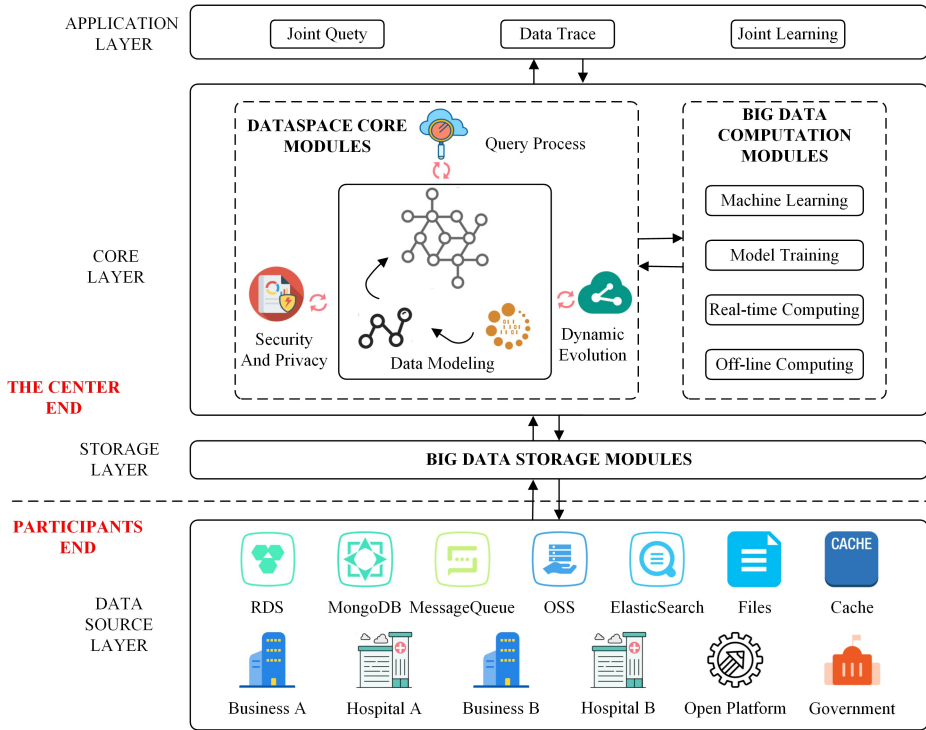


图 2 基于数据空间的大数据平台架构

Fig. 2 Big data platform architecture based on data space

(1)联合查询

联合查询即针对一个查询请求,查询多个跨组织的数据源,分别获取对应的结果信息,联合求解才是最终的查询响应结果。举一个最简单的例子,在某个区的医疗数据空间中,中心端接收到查询请求,“2022 年 5 月我区接诊的糖尿病病人总共多少人?”,中心端将查询请求进行分解并分发给不同的数据源端,如数据源端 A 医院收到的查询请求是“2022 年 5 月 A 医院接诊的糖尿病病人总共多少人?”,类似每个医院都给中心端返回自己医院的查询结果,最终这些结果在中心端求和就生成了最终的查询响应数据。联合查询的请求是一个查询语句,而联合求解方法可能是求和、求并集、求交集、比较等基本方法或者是基于它们的复杂操作,联合查询是数据空间最简单也是最高频的应用场景。

(2)数据溯源

数据溯源即实现数据的全生命周期记录,形成实体或者事件的全面可追溯。在数据管理中,形成元组的全面可追溯,即了解其所有变更记录,即哪个用户什么时间在哪个平台做了什么操作。在业务应用中,形成商品的全面可追溯,如一颗药从原料到生产、工艺、配方再到整个生产流程,既帮助用户了解了“一颗药的数字化之旅”,也适应了药企发展各个阶段的需求。在数据空间中,输入需要溯源的关键词,就可以在各个关联数据源获得其数据记录,形成全面可追溯,如在车辆产业链的数据空间中,输入某辆车的编号,即可以获得该车辆的全生命周期展示,包括其配件情况、销售商信息、发货仓库、物流订单、保养情况、配件维修和更换等。近年来,随着食品

安全、医药安全以及各领域行业的精细化管理的不断强化,此类应用场景越来越多。

(3)联合学习

联合学习即将学习目标分解成多个数据源学习子任务,分别在每个数据源独立学习,所有子任务的学习结果联合求解,才是原始学习目标的解。学习过程中多利用统计分析、机器学习等方法来挖掘数据规律,并进行分析预测。如制造业企业的数据空间中,结合多基地销售的实际情况,各个基地分别对海量的历史销售数据进行分析,最后通过联合求解实现月度销量的预测,预判企业车系的走势情况以及品牌市场关注度,这是一个经典的松散大数据分析需求场景。在医疗数据空间中,根据同一种罕见病的病患病例,各医院联合为罕见病群体患者画像,包括患者年龄段、性别、病情症状、病情发展、治疗方案及康复效果,由此可以实现针对某个罕见病患者的病情给出建议方案。又如保险公司为了制定针对司机群体的定制保单,就需要了解这些司机的车辆信息、驾驶信息、出险情况等,于是将对应的模型参数分发给各汽车运营商,分别在本地训练模型,最终返回司机驾驶习惯给保险公司,从而生成针对司机群体的保单模型,完成了针对司机群体保单模型的联合学习^[19],这些都是隐私数据的价值挖掘的联合学习场景。

以上都是基于核心层的平台功能来实现的,核心层为平台提供了丰富多样的数据计算组件工具,但并不是本文的研究重点,本文主要集中在数据空间的核心模块实现方法研究,它们全面系统地覆盖了数据空间构建过程中的基础关键问题,

包括:1)数据空间中的数据模型是如何构建的? 2)数据空间的动态演变是如何进行的,这些演变对系统产生了什么影响? 3)数据空间的查询是如何处理的? 4)数据空间中的安全和隐私问题有哪些?

3 数据空间研究进展

在深入剖析数据空间的基础上,本节将围绕上述提到的问题,分析近年来数据空间及其关联技术的研究进展,并关注其领域应用现状。

3.1 数据建模

数据空间中的数据建模是一个经典的多源异构数据的融合场景,这些异构的数据源可能是存储在不同数据库(如 PostgreSQL, MySQL, MongoDB)中的结构化数据、半结构化数据(如 JSON, XML)或者非结构化数据(如文本、图片、音频数据等)。不同于传统的数据融合,它们并没有被物理集中在一起来处理和使用,而是分布式地保存在原系统平台上。它们可能使用了不同的数据模型,也设置了不同的访问策略^[20],用独立数据库以不同的方式存储和进行逻辑表示。而这种场景下的跨域数据融合不能通过简单的模式映射和副本检测来实现,而是把不同的数据集提取的信息有机地整合在一起,从而感知这一区域的有效信息^[21],如图3所示。

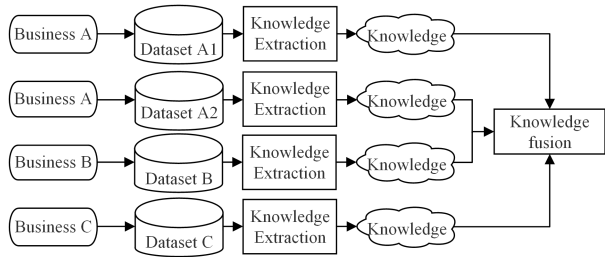


图3 跨组织的数据融合处理过程

Fig. 3 Data fusion process for cross-organization

已有的数据空间研究中,数据融合的目标是构建数据模型,从而支撑高效的数据存取。数据模型所描述的内容包括3部分,分别是数据结构、数据操作和数据约束。最早关于数据模型的研究都是围绕个人数据空间系统来展开的,如 iMeMex^[22]和 SEMEX^[23]。iMeMex 提出用 iDM 数据模型来管理用户个人信息,实现了对用户不同类型的数据信息的统一管理。此管理对象包括关系型数据、XML、文件内容、文件夹、邮件、数据流和 RSS 订阅等,但是这种数据模型只关注数据源的数据结构,即实现了对数据源的可访问,但未关注到数据源的语义及不同数据源之间的语义关系。以个人的邮件和文件为例,该数据模型只能确保可以访问到某封邮件和某个 ppt 文件,但是对这两个文件的内容是否描述的是同一个事件,它们之间又有何关联没有做深入探究,从而限制了对这类信息的价值挖掘。Elsayed 等提出了基于本体的数据模型,用来管理数据源之间的关系,允许它们的智能创建、表示和查找,同时应用 RDF 实现了对科学数据的组织,将不同数据源的元数据用三元组进行描述(实体、属性、值)^[24]。Yang 等提出了数据空间中实体、实体类和关系的定义,基于数据源之间

的数据语义关系构建通用的分层数据模型,包括实体数据图和实体模型图,分别用来捕捉实体之间的关系和实体类之间的关系。分层的数据模型为不同层的用户提供了对应的逻辑视图,这对于面向实体关系的语义关键字查询很有帮助,查询的结果将会是一系列实体或者实体对,可以帮助用户解析出了相对具像化的信息或者事实^[25]。

数据空间中的数据源关系挖掘是数据模型构建的难点。把每个数据源看作独立的数据集,数据源之间关系的挖掘过程就是独立数据集关系的发现过程,即数据集关系发现问题。以表格数据集发现为例,其核心是通过模式匹配找到数据集之间的语义对应关系,如检测两列数据是否可连接,具体包括属性匹配、值匹配、语义匹配、数据类型匹配、分布匹配和嵌入匹配等多维度^[26]。Koutras 等介绍了 REMA(Graph Embeddings-based Relational Schema Matching)这种语义模式匹配技术,它使用关系嵌入来捕获属性的语义相似性,同时实现了将数据库行、列和模式信息嵌入到可以揭示语义相似性的多维向量中^[27]。Benjamin 等提出了一种基于神经嵌入的新型端到端的模式匹配方法,通过使用不同级别的嵌入来表示整表或者单个属性^[28]。而在数据空间中,模式匹配的输出都是一组列表,如在文献^[29-30]中,使用不同的技术在不同数据源之间自动创建匹配模式和实体解析,预期的输出是一组潜在的匹配,确认匹配可以显著提高数据空间的性能。文献^[31]强调在这个阶段引入人的参与是非常必要的,来自用户的反馈可能会有很大的好处。Doan 等也认为,数据管理模型应该设计成允许在端到端数据管理过程的关键点上进行人工干预,然而用户是最稀缺的资源,不可能要求其确认所有的信息^[32]。因此,为了减少模式匹配中涉及到的人工工作,学者们开展了相关研究。文献^[33]提出了一种利用创新集的学习算法,用于重新排列模式匹配列表,并根据最佳匹配的排名改进功能,通过引入高质量的可选匹配集来辅助专家开展匹配工作,它也可以替代专家,直接在模式匹配阶段进行决策。Khine 等提出了多语言持久性的概念,即系统的不同部分用不同数据库来实现,即同一个系统中出现了不同的数据模型,从而解决了数据的多样性和复杂性问题^[34]。文献^[35]将多语言持久性的概念与数据建模过程相结合,在现有建模过程的侵入性尽可能小的原则下,通过拆分概念模型,来导出多个逻辑模型。多语言持久性的应用不可避免地会产生额外的开销,因此文献^[36]提出了一个可扩展的中间件 WA-RDF,它为多个 NoSQL 数据库提供了独特的网关,通过使用众所周知的语义网抽象来实现在文件数据库、key-value 数据库和图数据库中的 RDF 数据存储和查询。

3.2 动态演变

动态演变是数据空间的基本特征,也是重要特点, pay-as-you 是空间动态演变的基本原则,如图4所示。本小节将从数据空间初始化、数据空间更新及参与方数据源的自主演化3个阶段阐述其动态演变,围绕数据库模式和不确定性管理的已有研究工作,描述动态演变对空间数据的组织和管理带来的影响和挑战。

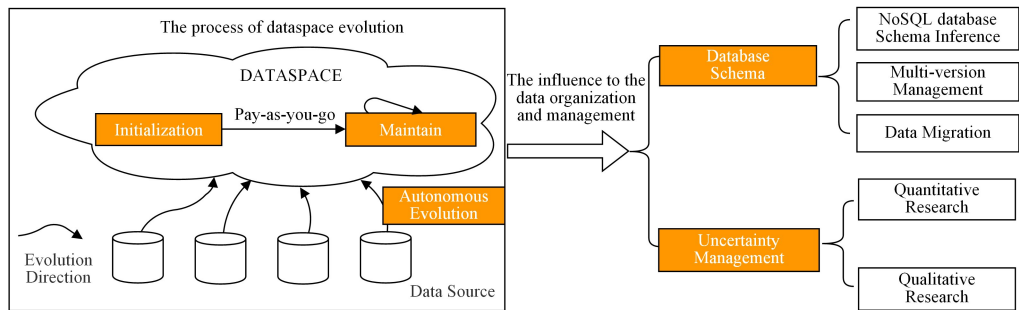


图 4 数据空间的动态演变的研究路径

Fig. 4 Research path of dynamic evolution of dataspace

数据空间的初始化工作研究最早是在 2008 年, Das 等实现了一种自动化构建 pay-as-you-go 的数据集成系统。这个系统构建的核心是为不确定性建模, 提出了概率中间模式的概念, 通过分析数据源模式自动构建系统, 并在多个领域的 50~800 个数据源上做了实验验证, 在无人干预的情况下获得了高质量响应^[37]。Jeffery 等认为^[31], 自动化的空间构建过程中产生的映射序列必须要用户反馈来确认, 而数据空间的大数据属性决定了不可能每一个映射匹配都去征求用户反馈。他们认为, 这些输出映射序列的排序就成为了关键问题, 每个序列相对于这个空间的收益大小成为了其排序的依据, 因此提出了完美信息价值的概念, 并给出了最终方法来对其进行量化。数据空间的更新是通过 pay-as-you-go 方式来完成, 即空间的数据融合和对信息的理解是在与用户的交互中不断地强化和完善的^[9]。Salles 等提出了 iTrails 这种声明式 pay-as-you-go 信息融合技术, 其核心是为检索引擎声明式地添加一个轻量“hints”, 从而不断增强松散的数据源融合效果, 也使得在数据空间获得更加完整的查询结果^[38]。而在数据空间初始化和更新的过程中, 核心的工作都是实现数据空间资源的分析, 即发现参与方数据源结构和不同数据源之间关系的过程, 很多场景下没有人拥有所有的参与方数据源知识, 甚至有的数据源没有任何人了解它的情况, 包括模式、实例、领域知识等^[39]。基于 pay-as-you-go 的演变思想, 如何持续通过空间资源分析, 覆盖尽可能多的用户检索和查询需求是数据空间演变的基本方向, Atzori 等通过对数据空间的数据及其之间的关系进行分析, 构建了一系列的组件和关系^[40]。2012 年, 文献^[41]提出了满足用户复杂信息需求的资源分析系统, 该系统的体系结构服务于两个目标: 1) 结合用户需求, 帮助用户从底层 XML 数据空间的数据源中选择有用的数据源; 2) 通过检测潜在的数据冲突来帮助用户验证所选数据源的一致性。而参与方数据源的自主演变主要包括以下两种场景: 1) 随着应用程序的代码自然演变, 如在敏捷开发中, 新的应用程序版本会频繁部署^[42], 新的功能需求对数据库数据服务提出新的需求; 2) 内部系统的优化需求, 如为了缩短查询响应时间, 系统会把频繁被访问的数据移动到固定存储区域, 为了降低云上数据备份的代价, 可能会把不常用的数据备份在线下进行物理存储^[43]。

数据空间的动态演变给数据空间的管理带来了问题和挑战, 从数据库模式和不确定性两个维度展开。数据库模式是对数据库中全体数据的逻辑结构和特征的描述。从数据空间

架构来看, 其模式分为两层, 一层是中心模式层, 即数据空间的全局模式, 用于响应面向数据空间的应用服务; 另一层是数据源模式层, 即参与方的本地模式, 既用于执行数据空间应用的子服务查询, 也负责响应本地应用查询。本地模式是构建空间模式的基础, 也是参与方数据源自主演变的直接呈现, 它可能是关系型, 也可能是非关系型。传统的关系型数据模式包括固定的数据结构, 而已经普遍应用在各场景中的 NoSQL 数据库并没有预定义的格式, 即它们的数据组织可以是无模式的, 这就给采用基于模式的数据库优化提出了挑战, 因此研究推断描述这些集合结构的模式方法具有重要意义。Veinhardt 等聚焦半结构化文档数据, 针对现有数据库模式推理解决方案进行了优化^[44]。JSON 作为一种流行的数据格式, 用于表示结构不规则或先验未知的数据, 其通常是海量且无模式的。文献^[45]对 5 种代表性 JSON 模式推理方法进行了比较分析, 发现在不同方向上在解决 NoSQL 数据库问题的同时尚存在缺点, 如果没有充分解决这些缺点, 则无法直接进行多模型数据的统一模式推理和建模。Baazizi 等分别从模式推理算法、参数调试、以用户为中心几个维度围绕海量 JSON 模式推断开展了系列研究^[46-49], Lbath 等围绕属性图模式推断开展了研究工作^[50]。

Hillenbrand 等就频繁模式演化问题引起的数据库中的版本化遗留数据给出了解决方案, 同时使用重复采样的概率蒙特卡罗方法, 来控制模式演化的复杂性^[42]。Curino 等提出了 PRISM(Panta Rhei Information Schema Manager)^[51], 用于支持数据库模式演变过程, 实现了信息的冗余验证, 通过查询重写或自动生成视图来实现遗留数据查询, 同时提供了演化预测功能。在此基础上, Curino 等^[52]于 2013 年提出了 PRISM++, 解决了模式演变过程中的完整性约束演变, 并通过结构模式改变和完整性约束演化实现了自动查询和更新重写。模式演变过程中多模式数据共存是一个普遍现象, 如何尽可能快地实现新旧模式的过渡, 得到了很多学者的研究。Herrmann 等认为, 这些工作不仅付出了很大的成本, 而且在很多的应用场景中有对多个版本模式共存的需求, 因此提出了 InVerD(Integrated Versioning of Databases)^[53], 实现了对多个模式版本管理的同时, 保证了数据的独立性, 并大大提升了数据的读写速度。模式演化引起的另外一个问题是数据迁移, 传统的数据迁移可能从一个数据库迁移到另外一个数据库, 也可能从一种数据模型迁移到另外一种数据模型。从数据空间的应用场景来看, 为了实现数据溯源、联合查询和联合

学习,迁移的方向是从数据源到数据空间。迁移操作可能是为了实现数据高速访问,将部分数据迁移到数据空间缓存,也可能是将历史的数据版本迁移到数据空间进行存储。根据不同的场景需求,选择不同的迁移方式来完成数据迁移,如即时迁移数据访问延时低,但是迁移成本高,懒惰迁移节约了迁移成本,但是响应延时会增加,而且很多时候会有多个不同版本的模式数据共存^[42]。预测迁移以期找到查询响应时间和迁移成本的平衡,Hillenbrand等提出了一种自适应数据迁移方法^[54],可以自动调整相关参数和策略,从而节约数据迁移成本。

动态演变加剧了数据空间的不确定性,管理的不确定性使系统在不了解数据的所有信息时(数据空间中的常见情况)能够以一种有原则的方式来进行工作^[10]。针对数据空间中管理不确定性的研究,主要有两种方法,一种是通过定量研究开展模式映射中的不确定性管理,另一种是定性研究,即利用用户反馈来降低不确定性。在对模式映射的不确定性管理研究中,Dong等提出了一种方法,通过引入概率模式映射的概念来管理数据集成中的不确定性,在这种概念中,每个生成的映射都附加了一个概率,这个概率是由一个概率质量函数导出的^[55]。Magnani等假设存在一组用概率注释的匹配,以表示一个用概率注释映射的数据集成过程,他们使用离散值来表示结构之间的语义关系^[56],而Christodoulou等在连续区间 $[0,1]$ 中分配了一定程度的正确性,同时描述了一种推导概率的系统方法^[57]。Van Keulen提出使用概率方法来处理数据清洗、映射和信息提取过程中的不确定性^[58]。

3.3 查询处理

随着数据量的不断增加,近十年出现了许多以大数据分析为目标的分布式SQL引擎。例如,Apache Hive^[59]为数据查询提供了类SQL接口,并为MapReduce问题提供了数据仓库解决方案;Spark SQL^[60]通过与Apache Spark结合,提供了多样的大规模数据分析能力;Presto^[61]是一个开源的分布式查询引擎,其目标是“SQL on Everything”,它提供了丰富的插件接口和连接器API,通过一个协调节点和多个工作节点实现了为多个异构数据源快速处理大量数据集。Presto为像Facebook这样的企业实现了利用一个SQL系统实现多源数据分析,并可简单查询多达1000个节点的存储系统中的海量数据。随着公有云的产生,出现了基于云的商业SQL产品。例如,Google BigQuery^[62]是一个基于云的全托管数据仓库;Snowflake^[63]具有支持多租户、事务性和弹性强的特点;Azure Synapse Analytics^[64]则实现了数据仓库和大数据工作负载两者功能的结合。通过研究分析发现,当前的分布式数据库查询引擎发展呈现以下特点:1)以实现大数据分析为核心目标;2)以数据联合为基本构建原则;3)以SQL或者类SQL语言为核心查询语言;4)支持半结构化、非结构化的大数据快速分析,融合云存储、机器学习技术开展研究,而这恰恰也是数据空间的构建目标。

数据空间的查询处理如图5所示,查询处理分为两层^[19],包括中心端和数据源端。中心端通过查询重写、查询优化等查询处理生成子查询,然后进行全局查找,如果找到匹配结果,则直接返回给用户;如果没有,则通过元数据连接

对应的参与方数据源,在数据源本地执行查找。各个数据源的返回结果在中心端联合求解,然后对结果排序,将最终的结果返回给用户。本节将从用户查询请求、查询优化和结果排序3个方面具体描述数据空间的查询处理。

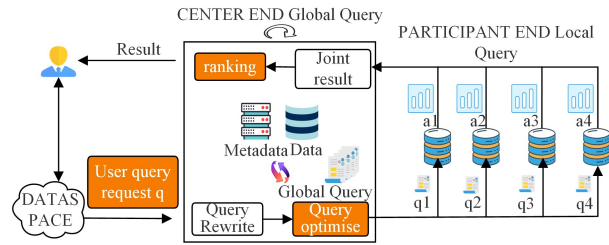


图5 数据空间的查询处理

Fig. 5 Query process in dataspace

根据不同阶段空间对数据间关系的处理能力,将用户的查询请求分为4种情况^[65],具体包括:1)关键字查询,不考虑数据源关系,根据关键字返回所有匹配的信息;2)关联查询,根据查询条件不仅返回匹配的实体本身,还返回关联实体;3)异构查询,即在数据空间中,不同的属性名可能代表同样的属性含义,它们之间的映射关系通过模式匹配发现,因此针对任何一个属性的查询,对应匹配的属性信息也同时返回;4)相似性查询,根据查询条件,通过语义查询优化等技术返回实体及与其相似的实体。从数据查询响应方式来看,包括3种响应模式:1)数据仓库模式,只能查询到中心端本地索引和副本;2)中间模式,中心端的本地索引被直接忽略,直接从参与方端执行查询响应;3)前两种的混合模式^[22],为了提高数据空间系统中查询的效率,用户可以基于查询框架通过扫描核心空间来有效地执行查询,而不是扫描整个数据空间^[66]。

查询优化是数据组织与管理研究中的一个重要方向,而数据空间中的动态演变、不可控特征,给查询优化带来了新的挑战。索引作为一种提高查询性能的基础方法,但需要支持对数据空间这类松耦合的数据集进行健壮的索引,从而有效地响应基于关键字和简单结构要求的信息查询^[66]。在索引的构建对象上,Song等提出了为异构数据建模^[67],通过研究数据库视图,将其物化以便查询重用,选择其中的一部分生成最小查询成本的最优计划,同时将数据中的元组进行分块来实现对top-k查询请求的快速响应。在索引的构建方式上,Dong等借鉴信息获取技术方法,描述了倒排索引技术^[68],而为数据空间面向用户提供混合数据的关键字搜索服务提供了方法支撑,同时也缩短了非结构化数据的查询处理时间。Lal等提出了使用改进的余弦相似度和向量空间模型的排序聚类方法^[69],应用向量空间模型、文档术语矩阵和TF-IDF权重来对异构数据进行索引和排序,为数据空间中非结构化数据的索引提供了更好的方法。在索引的性能优化上,Salles等提出将trails和分类压缩索引关联起来^[70],分类的压缩索引可以用于描述有少量边的图,从而为索引节约了时间和空间。2012年,Yang等^[71]提出了类似的方法,只是覆盖了更多的类别和图。另一个值得注意的工作是文献^[66]提出的两级索引策略。第一级是核心空间索引,第二级是全空间索引,分别用两种不同的策略来维护这两级索引,然而该方法的有效性还需要验证。类似地,目前学术界正在研究几种扩展倒排

列表的方法,但仍然需要进行实验来验证这些技术的实用性。另一个优化方向是对半结构化数据的快速分析方法的研究,由于开发人员更喜欢灵活性,因此 JSON 等半结构化数据在数据库应用场景中越来越流行,然而在现有的系统中,其因为缺乏固定模式而导致分析速度缓慢。Durner 等提出了 JSON 瓦片的概念^[72],用于实现对 JSON 数据的高性能分析。其背后的关键思想是自动检测对象集中的隐式公共结构,使用其信息推断键的类型,将频繁出现的键具体化为关系列,将不常用的键和异类(离群值)键以优化的二进制格式存储。它允许快速访问单个键,所有这些实现过程都是自动和透明的,可以在不牺牲格式灵活性的情况下对 JSON 数据进行快速分析,同时收集查询优化器统计信息,以便能够找到好的查询计划,使性能接近原生关系列存储的性能。近年来,一个重要的优化方向是运用机器学习技术提升优化器,Kassela 等实现了一种基于机器学习的多引擎查询优化器^[73],通过使用机器学习技术对所有的基础 SQL 操作构建代价模型,并用其估计查询执行代价,以此为依据配置统计信息。相比传统静态统计信息优化方式,此方法有效提升了优化器的效率。

数据空间的不可控使得数据的查询结果基本都是 best-effort 实现返回的。从长远来看,查询结果不是精确匹配的,因此需要满足以下 3 个要求:1)能够对不同类型的数据进行合适的展示;2)对不同的结果进行合理排序;3)支持用户进一步精确性地查询请求操作。因此开展基于查询结果的排序研究也成为了数据空间研究关注的方向。Lal 等^[74]从不同维度对各类排名算法进行了对比分析,包括主要技术应用、方法、输入参数、关联性、返回结果的质量、重要性和使用限制、搜索引擎和算法的时间复杂度等,同时详细深入地描述了数据空间中的排序问题及相关的挑战^[75]。2020 年,文献[69]提出了一种使用改进的余弦相似度和向量空间模型(VSM)的排序聚类方法。该方法的实现过程中,最匹配查询文档被首先显示,文档的排名是根据与数据空间上非结构化数据查询的相似程度进行的,其为非结构化数据结果排名提供了方法支撑。但是,目前此成果只针对空间内非结构化数据的查询,尚未拓展到其他结构化数据、流数据等数据类型。同时,Wang 等提出将概率查询应用在数据空间查询结果排序中^[65]。

不同数据源安全异构问题也是影响数据空间查询效率的一个重要问题。在数据空间中,每个数据源根据业务需求都配置了各自的安全设施和策略,如 Docker、加密等,而不同的设施和协议也反映了要实现数据获取的前期代价。在空间查询中,不同的查询请求会产生不同的数据查询计划,即确定了在哪些数据源如何实现目标数据的获取,这个过程的代价称为查询估计代价,同时也就确定了获取数据需要付出的安全代价,称为数据共享代价,因此可以将基于异构安全设置的查询评估看作一个双目标优化问题^[76],最小化数据共享代价和查询估计代价。

3.4 系统的实现与应用

基于数据空间构建的系统有自己的特色,组织多个跨组织的数据源并为用户提供一致的服务,需要通过大量的架构设计工作。早期的系统实现主要聚焦在个人数据管理上,

iMeMex^[77]和 SEMEX^[78]都是经典的个人数据空间管理系统。其中 iMeMex 包括数据源层、物理数据层、逻辑数据层和应用层,它通过一致的数据模型实现了对用户不同类型个人数据的统一表示,并实现了对个人全量全维度的数据浏览和查询。2008 年,Chung 等实现了企业内部的跨部门数据空间 COD(Cross-Organisation Dataspace)^[79],实现了灵活地管理数据源,并尽可能地让用户准确获取目标数据。这段时期的数据空间的系统能力,主要在数据源的管理和基于关键字的数据查询上,尚未开展深层的数据源语义挖掘。后来,Kui-cheu 等搭建了一个基于语义的异构数据管理系统^[80],提供了 3 个核心组件:同义词和推理字典、学习和索引、结果构建器和查询翻译器。字典组件部分,包含了已知的语法和语义,能帮助系统发现数据源之间的可能的关系,并处理不确定性和非一致性。学习组件部分能够通过用户活动提升字典的准确性。但是它们对于其他功能和数据空间的特征整合考虑得太简单。类似的,Jiang 等提出了资源空间模型^[81],它能够捕捉语义分类,同时在模型中实现坐标合并。它通过一种查询语言对所有组件实现访问,从而对模型和数据查询提供支撑。文献[82]则实现了一个分布式大型科学数据存储库,更强调通过数据全生命周期管理来实现数据组织。

数据空间如何在行业领域展开应用,一直是学者们关注的问题,关注度最高的是工业领域。Franklin 等最早提出数据空间在工业领域有广阔的应用空间^[9],近年来,与其有关的应用研究越来越多。Angrish 等认为,工业数据空间可以看作是跨企业的数据管理系统^[83]。Xiao 等尝试将基于本体的数据访问和知识图谱用于集成、存储、索引、语义查询和知识推理工业数据^[13]。GE 公司提出了语义驱动的框架 SemTK (Semantics Toolkit),实现了对异构数据的可靠存储和访问^[84]。框架通过访问多语言持久性大数据存储,允许跨多个存储平台(如 Hadoop、图形数据库和语义三重存储等大数据存储)存储数据,同时保持单个逻辑接口和点访问,从而为用户提供知识驱动的外观。为了让车间生产更智能可控,Li 等提出了 IDMW(Industrial Dataspacefor Machining Workshop)系统构建^[14],主要解决了 3 个方面的问题:1)对车间生产过程中不同阶段的数据进行管理;2)基于数据融合构建机器知识模式;3)机器操作控制研究中领域知识的再应用,实现了中心化架构-分布式数据存储的模式构建。

同样备受关注的应用领域是医学,不同组织机构间的数据共享可以加速医学诊断和生命医学研究。随着数据科学相关技术的不断发展,如何利用这些新技术、新方法加快医学数据的互联共享,从而挽救生命,也成为了学者们的一个重要研究方向。Nadal 等提出了数据空间系统 ODIN^[16],该系统因为被世界卫生组织忽视的热带病而提出,因为当时的卫生信息系统过于薄弱,无法确保高质量数据的收集、流动、验证、使用和传播,从而导致各国卫生部报告的官方数据往往不完整。ODIN 用于将现有官方数据与非政府组织、研究人员的数据源或其他来源(如药物警戒系统、各领土内病媒分布等)这些碎片化异构化的数据结合起来,以一种自底向上的方法,通过 pay-as-you-go 的方式对多源异构数据进行集成,并对用

户提供了多种形式的查询服务。针对医学用户的健康数据敏感的现实需求,Hernandez 等设计实现了 TIKD 数据空间^[17],称为可信知识数据空间。它基于关联数据技术和可信数据分享,实现了对用户个人隐私数据及个人日志这种敏感数据信息的共享,并基于这些数据实现了知识图谱集成。数据空间在其他领域的应用也有些探索,如 Howe 等设计实现了基于科技领域的数据属性浏览器 Quarry^[39],作为工具实现了对数据空间的资源分析。Elsayed 等实现了基于海量科技数据的管理系统^[85]。Borjigin 等研究了数据空间在数字图书管理上的应用^[86],实现了清华大学图书馆、国家图书馆、北大图书馆和国家学术和信息系统的数字图书空间构建。

尽管有很多数据空间原型系统正在开发中或被落地应用,但是没有一个是被认为是可以直接应用的可以达到最终用户期望的产品。数据空间在各个领域的应用仍有广阔的探索空间,这些探索是广泛的,包括系统、标准、技术开发和示范,它们不是独立进行的,而是相互促进的。

3.5 安全与隐私

近年来,公民个人数据隐私保护条例越来越多地进入各个国家,如欧盟的 GDPR(General Data Protection Regulation)以及加拿大的 PIPEDA(Personal Information Protection and Electronic Documents Act),它们都是通过明确的法律条文规定,确保用户的个人数据在企业平台安全存储,并授权给用户来决定数据如何共享和处理。而在我国,《中华人民共和国个人信息保护法》于 2021 年 11 月 1 日开始施行,其中明确提到,“个人对其个人信息的处理享有知情权、决定权,有权利或者拒绝他人对其个人的信息进行处理”,“个人有权向个人信息处理者查阅、复制、删除其个人信息”等。这些都对数据库的设计提出了更多的要求^[87],同时也加剧了“数据孤岛问题”。而在数据空间的研究中,我们更关注的是如何在实现隐私保护的前提下,实现数据的安全共享。

隐私计算是当前业务认可的保护数据隐私的有效技术方案,包括信息混淆脱敏方法、分布式计算、可信环境构建及多方安全计算。信息混淆脱敏方法常用于共享数据发布,在数据发布之前先进行加噪、删除、泛化等特殊处理。例如,差分隐私^[88]已经成为隐私数据发布的标准,攻击者无法从满足差分隐私要求发布的数据中推测出某个个体的数据。联邦学习^[89]是分布式计算的代表,可以实现让参与方数据不离开本地平台完成学习任务,只需要输出本地计算的结果,然而仍有中间结果泄露的风险。同时可以通过构建可信环境处理数据,通过多方安全计算,在数据加密之后对数据进行计算和查询处理,如同态加密允许在密文上进行运算,在不影响计算的前提下实现了数据的隐私保护^[90-91]。

在可信环境的构建研究中,Sun 等认为,数据空间中的数据不只是信息的携带者,而且包含了一定的隐私信息,而且不同的数据具有不同的隐私性,因此参照数据的价值和数据敏感度两个维度,对个人数据的敏感性定义了 9 个级别^[92],如表 2 所列。构建者可以选取合适的安全策略对不同层级的数据进行安全保护。基于此,他们设计了基于数据维度、商业维度和安全维度的认证数据空间系统,为构建通用的可信数据认证空间提供了参考。

表 2 个人数据安全分级及对应的安全策略

Table 2 Data classification and corresponding safety policies

Sensitivity	Value-Sensitivity	Safety Technology
Level 9	High-High	Data Encryption, Identity Authentication, Right Management, Content Signature, Data Backup, Log Management
Level 8	Medium-High	
Level 7	High-Medium	
Level 6	Medium-Medium	Data Encryption, Identity Authentication, Right Management, Log Management
Level 5	High-Low	
Level 4	Low-High	
Level 3	Medium-Low	Identity Authentication, Right Management, Log Management
Level 2	Low-Medium	
Level 1	Low-Low	

数据空间的分布式存储使得数据频繁传输成为另一个新问题,在已有的解决方案中,区块链已被应用在工业企业中^[93]。针对工业数据空间中存在的大量数据传输,区块链的安全性和透明度等特性使其适用于包含大量数据传输的场景^[94-95]。同时,对于目前的工业数据空间设计,需要有一个中央主机来提供数据分析和关系构建功能,这对数据传输能力和数据安全提出了至关重要的要求。在区块链的帮助下,可以在本地数据源上设置核心部分,减少数据容量和安全的压力。

4 研究展望

作为近年来的新概念,数据空间在保证参与方数据源应用的前提下,实现了跨组织多源异构数据的融合,围绕数据空间的构建与应用,未来仍有许多的开放方向有待探索。

4.1 多模态数据融合

数据空间中的数据量大、种类多、来源多样,它们呈现出了不同的模态,不能被直接处理,通过相互之间的支持、补充、修正,提供了更加准确的信息,而且对于信息的呈现有更强的鲁棒性。因此可以通过多模态的方式实现信息获取、环境感知和知识学习。而如何对这些多模态数据进行融合,是重要的研究课题。传统的数据融合包括数据融合、特征融合和决策融合 3 个阶段,根据融合阶段的不同可以构建不同的融合策略,而多模态融合需要突破数据异构特点,以及异构模态之间的关联关系。近年来,深度学习已经成为实现复杂数据价值挖掘的基础方法,特别是针对这种海量异构数据的无监督学习很有效,已经被广泛应用在图形图像特征提取等场景中。因此,深度学习也成为了多模态数据融合的一种支撑技术。多模态融合的研究有以下方向值得进一步深入探索:

- (1)针对不同模态的数据,特别是 JSON 和 XML 半结构化数据,对文本、图像、音频、视频等非结构化的数据进行分析理解本身就是一个难题。在非可信的海量数据环境中,如何寻找此类数据信息的隐层模式,实现数据的组织管理优化及快速分析?
- (2)如何突破不同模态数据的异质性特点,有效地识别和利用不同模态之间的语义相似性,实现多模态数据的准确表示?
- (3)多模态数据如何在不同的应用场景中进行协同和交互?联合查询中如何在异构的数据模态中获取相对准确的数据结果?联合学习中如何从不同模态的数据集中获得语义的补充和增强?

4.2 高效的查询处理

数据空间实现的是海量多源异构的数据组织、管理和应用,因此相比传统的数据管理系统,其在模式匹配、数据管理及查询优化等方面面临着挑战。如在制造业数据空间中,价值链上的数据呈指数级增长,使得建立精准的代价估算模型难度越来越大。同时,在求解多表连接路径这样的 NP-Hard 问题时,优化器通常使用经典的动态规划、启发式,甚至是穷举等算法,这些算法在过去数据量不大的情况下表现尚佳,但如今的应用是动辄上百个关系的大型复杂查询,传统算法难以在较短时间内选择出高效的查询计划。准确高效的查询处理,是支持数据空间运行的基本能力,然而因为数据空间的不可控、数据的多源多样及空间的动态演化特点,使得如何实现从数据空间中高效地获取准确数据信息成为了研究难点,未来的研究方向具体包括以下几个方面。

(1)高效的查询计划生成研究。针对不同类别场景甚至同类场景下的不同查询请求生成不同的查询计划,可以实现数据的准确快速获取。过程中需要综合考虑数据源的可信度、安全策略、稳定性、数据模型等因素,基于此设计数据空间查询计划构建框架,并提出方法来评估其有效性和高效性。

(2)数据查询优化方法研究。包括探索查询相似性、模式相似性、数据集相似性等方法的应用,如何利用机器学习提取数据空间特征,如何评价优化方法的有效性,都是值得探索的课题。

(3)查询结果的不确定性管理研究。包括如何尽量选择可信的数据源来降低结果的不确定性,如何通过用户介入获得尽可能准确的结果,如何应用合理的方法对查询结果的不确定性进行定量或者定性评价,都需要开展进一步的研究工作。

4.3 数据的安全共享

如何在满足数据安全、隐私保护和监管合规的前提下,突破数据孤岛,实现跨组织的数据融合,从而释放数据价值,是数据空间研究应用面临的重要挑战。而隐私计算是当前业务认可的保护数据隐私的有效技术方案,针对数据空间的具体场景,如何有效地实现数据的安全共享,仍然具有广阔的研究空间,具体包括:

(1)相比传统的数据发布,数据空间是个多方参与的多数据源发布场景,如何建立基于隐私保护的联合数据发布机制?

(2)针对数据空间中的联合学习场景,如何实现参与方数据源的安全加入与退出,如何防御联合学习参与方的主动攻击?

(3)针对数据空间不可信的特点,如何构建可信数据认证共享框架,实现数据空间的数据认证、处理和分析?

4.4 基于数据空间的大数据平台构建

大数据平台是数据空间运转的基础和载体,构建基于数据空间的大数据平台,是数据空间落地应用的必经途径,其主要面临的问题包括以下几个方面。

(1)灵活安全的数据源管理。数据源是数据空间的根本数据来源,做好数据源的灵活有序的接入和退出管理,是保证空间数据有效性的基础。当数据源加入时,提供基于 ETL 或者 API 等灵活多样的数据接入方式和有效的数据源可信

身份认证;在数据空间的对外服务过程中,实现高效的数据源计算资源和数据资源调度;当数据源退出时,实现非延时的数据源遗忘等,这些都是数据源管理中需要解决的问题。

(2)多模数据的存储与调度。数据空间的中心端的数据包含不同数据源的缓存数据、元数据、模型数据等多样异构的复杂数据,因此需要构建合适的存储策略,实现对这些多模数据的存储、组织和管理,基于此实现空间的信息管理中心,并对外提供面向服务的各类应用接口,这将大大增加数据空间的灵活性,提高空间可用性。随着数据库产品的不断发展,越来越多的数据库开始支持多模数据的存储,实现多模数据管理。同时分布式的中心架构也将有助于实现中心端对不同模态的数据进行存储组织,因此,如何配置多模数据存储组件,如何构建多模数据之间的交互和协同关系,为上层数据访问提供统一接口,如何针对多模数据设计分层存储策略,提升存取效率,这些都是多模数据存储与调度面临的技术挑战。

(3)多样的数据计算模式管理。近年来,随着大数据技术的快速发展,出现了一批大数据系统,如 Hadoop, Spark, Flink 等,它们都是根据不同的数据处理需求,提供不同的大数据计算模式。同时,机器学习为挖掘数据的潜在知识提供了新方法,出现了如 TensorFlow 等学术界和工业界都广泛使用的机器学习系统。而在数据空间大数据平台中,存在对批处理、流计算、机器学习等多样计算模式的共同需求,基于不同的应用场景,选取合适的计算策略,最终实现降低资源成本,提高系统运行效率,进而提升服务质量。这是一个有价值的研究课题。

(4)基于人机交互的不确定性管理。通过构建合适的人机管理策略,实现降低人力成本,提高空间数据结果的准确性。在数据空间的构建和运行过程中,存在大量的不确定性场景,实现对这些不确定性的管理是一个重要的研究问题,可以通过模式匹配、相似性分析、机器学习等方法来提高准确性。在此基础上,通过用户反馈来优化结果,如通过用户干预的方式来从查询结果列表中确定答案,通过用户参与来寻找最佳模式匹配。

结束语 数据空间作为一种新的数据组织和管理模式,突破了数据孤岛,实现了非可信环境下多方数据的按需融合,为跨组织场景的数据共享、数据分析以及数据服务提供了新途径。本文对数据空间的内涵进行了拓展定义,并对其核心概念进行了形式化描述,提出了基于数据空间的大数据平台架构,总结描述了经典的数据空间应用场景。围绕数据空间构建,分析了关联技术的研究方向及研究方法,简述了数据空间的领域实现与应用。最后,从多模态数据融合、高效的查询处理、数据安全共享和基于数据空间的大数据平台构建 4 个方面展望了研究前景及挑战。

参 考 文 献

[1] STONEBRAKER M, CETINTEMEL U. “One size fits all”: an idea whose time has come and gone [C] // 21st International Conference on Data Engineering (ICDE '05). IEEE, 2005: 2-11.

[2] INMON W H. What is a data warehouse [J]. Prism Tech Topic, 1995, 1(1): 1-5.

- [3] ELMASRI R, NAVATHE S B. Fundamentals of Database Systems[M]//Pearson Education, 2009:1104-1107.
- [4] SHETH A P, LARSON J A. Federated database systems for managing distributed, heterogeneous, and autonomous databases[J]. ACM Computing Surveys(CSUR), 1990, 22(3):183-236.
- [5] SUWANMANEE S, BENSLIMANE D, CHAMPIN P A, et al. Wrapping and Integrating Heterogeneous Relational Data with OWL[C]//ICEIS, 2005:11-18.
- [6] GADEPALLY V, CHEN P, DUGGAN J, et al. The BigDAWG polystore system and architecture[C]//2016 IEEE High Performance Extreme Computing Conference(HPEC). IEEE, 2016: 1-6.
- [7] GU Z, CORCOGLIONITI F, LANTI D, et al. A Systematic Overview of Data Federation Systems[J/OL]. <https://content.iospress.com/articles/semantic-web/sw223201>.
- [8] KHINE P P, WANG Z S. Data lake: a new ideology in big data era[C]//ITM Web of Conferences, 2018.
- [9] FRANKLIN M, HALEVY A, MAIER D. From databases to dataspace: a new abstraction for information management[J]. ACM Sigmod Record, 2005, 34(4):27-33.
- [10] FRANKLIN M, HALEVY A, MAIER D. A first tutorial on dataspace[C]//Proceedings of the VLDB Endowment, 2008: 1516-1517.
- [11] LI Y K, MENG X F, ZHANG X Y. Research on data space Technology[J]. Journal of Software, 2008, 19(8):2018-2031.
- [12] SINGH M, JAIN S K. A survey on dataspace[C]//International Conference on Network Security and Applications, Berlin: Springer, 2011:608-621.
- [13] XIAO G, CALVANESE D, KONTCHAKOV R, et al. Ontology-based data access: a survey[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018:5511-5519.
- [14] LI P, CHENG K, JIANG P, et al. Investigation on industrial dataspace for advanced machining workshops: enabling machining operations control with domain knowledge and application case studies[J]. Journal of Intelligent Manufacturing, 2022, 33: 103-119.
- [15] JIANG P, LIU C, LI P, et al. Industrial dataspace: A broker to run cyber-physical-social production system in level of machining workshops[C]//2019 IEEE 15th International Conference on Automation Science and Engineering(CASE). IEEE, 2019: 1402-1407.
- [16] NADAL FRANCESCH S, RABBANI K, ROMERO MORAL Ó, et al. ODIN: A dataspace management system[C]//Proceedings of the ISWC 2019 Satellite Tracks(Posters & Demonstrations, Industry, and Outrageous Ideas): co-located with 18th International Semantic Web Conference(ISWC 2019). Auckland, New Zealand, 2019:185-188.
- [17] HERNANDEZ J, MCKENNA L, BRENNAN R. TIKD: A trusted integrated knowledge dataspace for sensitive healthcare data sharing[C]//2021 IEEE 45th Annual Computers, Software, and Applications Conference(COMPSAC). IEEE, 2021:1855-1860.
- [18] BEN HAMADOU H, GALLINUCCI E, GOLFARELLI M. Answering GPSJ queries in a polystore: A dataspace-based approach[C]//International Conference on Conceptual Modeling, Cham: Springer, 2019:189-203.
- [19] MATSUBARA M, MIYAMAE T, ITO A, et al. Improving Reliability of Data Distribution Across Categories of Business and Industries with Chain Data Lineage[J]. Fujitsu Scientific & Technical Journal, 2020, 56(1):52-59.
- [20] MÖLLER J, HAHN A. Searching on Heterogeneous and Decentralized Data: A Short Review[C]//3rd International Open Search Symposium, 2021.
- [21] ZHENG Y. Methodologies for cross-domain data fusion: An overview[J]. IEEE Transactions on Big Data, 2015, 1(1):16-34.
- [22] BLUNSCHI L, DITTRICH J P, GIRARD O R, et al. A dataspace odyssey: The iMeMex personal dataspace management system[C]//CIDR, 2007:114-119.
- [23] DONG X. Providing best-effort services in dataspace systems[D]. Seattle: University of Washington, 2007.
- [24] ELSAYED I, MUSLIMOVIC A, BREZANY P. Intelligent dataspace for e-science[C]//7th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, 2008:94-100.
- [25] YANG D, SHEN D, NIE T, et al. Layered graph data model for data management of dataspace support platform[C]//International Conference on Web-Age Information Management, Berlin: Springer, 2011:353-365.
- [26] KOUTRAS C, SIACHAMIS G, IONESCU A, et al. Valentine: Evaluating matching techniques for dataset discovery[C]//2021 IEEE 37th International Conference on Data Engineering(ICDE). IEEE, 2021:468-479.
- [27] KOUTRAS C, FRAGKOULIS M, KATSIFODIMOS A, et al. REMA: Graph Embeddings-based Relational Schema Matching[C]//EDBT/ICDT Workshops, 2020.
- [28] HÄTTASCH B, TRUONG-NGOC M, SCHMIDT A, et al. It's AI Match: A Two-Step Approach for Schema Matching Using Embeddings[J]. arXiv:2203.04366, 2022.
- [29] RAHM E, BERNSTEIN P A. A survey of approaches to automatic schema matching[J]. the VLDB Journal, 2001, 10(4): 334-350.
- [30] COHEN W W, RAVIKUMAR P, FIENBERG S E. A Comparison of String Distance Metrics for Name-Matching Tasks[C]//IWeb, 2003:73-78.
- [31] JEFFERY S R, FRANKLIN M J, HALEVY A Y. Pay-as-you-go user feedback for dataspace systems[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008:847-860.
- [32] DOAN A H, NAUGHTON J, BAID A, et al. The case for a structured approach to managing unstructured data[J]. arXiv: 0909.1783, 2009.
- [33] GAL A, ROITMAN H, SHRAGA R. Heterogeneous data integration by learning to rerank schema matches[C]//2018 IEEE International Conference on Data Mining(ICDM). IEEE, 2018: 959-964.
- [34] KHINE P P, WANG Z. A review of polyglot persistence in the big data world[J]. Information, 2019, 10(4):141.
- [35] KOLONKO M, MÜLLENBACH S. Polyglot persistence in conceptual modeling for information analysis[C]//2020 10th International Conference on Advanced Computer Information Tech-

- nologies(ACIT). IEEE,2020;590-594.
- [36] SANTANA L H Z,DOS SANTOS MELLO R. A middleware for polyglot persistence of RDF data into NoSQL databases [C]//2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science(IRI). IEEE,2019;237-244.
- [37] DAS SARMA A,DONG X,HALEVY A. Bootstrapping pay-as-you-go data integration systems[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008;861-874.
- [38] SALLES M A V,DITTRICH J P,KARAKASHIAN S K,et al. iTrails: Pay-as-you-go Information Integration in Dataspaces [C]//VLDB. 2007;663-674.
- [39] HOWE B,MAIER D,RAYNER N,et al. Quarrying dataspace: Schemaless profiling of unfamiliar information sources [C] // 2008 IEEE 24th International Conference on Data Engineering Workshop. IEEE,2008;270-277.
- [40] ATZORI M,DESSÌ N. Dataspace:where structure and schema meet[M]// Learning Structure and Schemas from Documents. Berlin:Springer,2011;97-119.
- [41] NÄPILÄ T,NIEMI T. An approach for developing a schemaless XML dataspace profiling system[J]. Journal of Information Science,2012,38(3):234-257.
- [42] HILLENBRAND A,SCHERZINGER S,STÖRL U. Remaining in Control of the Impact of Schema Evolution in NoSQL Databases[C]// International Conference on Conceptual Modeling. Cham:Springer,2021;149-159.
- [43] STIEMER A,VOGT M,SCHULD T H,et al. PolyMigrate:Dynamic schema evolution and data migration in a distributed polystore[M]// Heterogeneous Data Management, Polstores, and Analytics for Healthcare. Cham:Springer,2020;42-53.
- [44] VEINHARDT L I. Schema Inference for NoSQL Databases [D]. Czech Republic:Charles University in Prague,2021.
- [45] ČONTOŠ P,SVOBODA M. JSON schema inference approaches [C]//International Conference on Conceptual Modeling. Cham: Springer,2020;173-183.
- [46] BAAZIZI M A,LAHMAR H B,COLAZZO D,et al. Schema inference for massive JSON datasets[C]// Extending Database Technology(EDBT). 2017.
- [47] BAAZIZI M A,COLAZZO D,GHELLI G,et al. Parametric schema inference for massive JSON datasets[J]. The VLDB Journal,2019,28(4):497-521.
- [48] BAAZIZI M A,BERTI C,COLAZZO D,et al. Human-in-the-loop schema inference for massive JSON datasets[C]//23rd International Conference on Extending Database Technology(EDBT 2020). OpenProceedings. org,2020;635-638.
- [49] BAAZIZI M A,COLAZZO D,GHELLI G,et al. A typesystem for interactive JSON schema inference[C]//46th International Colloquium on Automata, Languages, and Programming(ICALP 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.
- [50] LBATH H,BONIFATI A,HARMER R. Schema inference for property graphs[C]//24th International Conference on Extending Database Technology(EDBT 2021). 2021;499-504.
- [51] CURINO C A,MOON H J,ZANIOLO C. Graceful database schema evolution:the prism workbench[C]//Proceedings of the VLDB Endowment. 2008;761-772.
- [52] CURINO C,MOON H J,DEUTSCH A,et al. Automating the database schema evolution process [J]. The VLDB Journal, 2013,22(1):73-98.
- [53] HERRMANN K,VOIGT H,PEDERSEN T B,et al. Multi-schema-version data management: data independence in the twenty-first century[J]. The VLDB Journal,2018,27(4):547-571.
- [54] HILLENBRAND A,STÖRL U,LEVCHENKO M,et al. Towards self-adapting data migration in the context of schema evolution in NoSQL databases[C]//2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW). IEEE, 2020;133-138.
- [55] DONG X L,HALEVY A,YU C. Data integration with uncertainty[J]. The VLDB Journal,2009,18(2):469-500.
- [56] MAGNANI M,MONTESI D. Uncertainty in data integration: current approaches and open problems[C]//MUD. 2007;18-32.
- [57] CHRISTODOULOU K,SERRANO F R S,FERNANDES A A A,et al. Quantifying and propagating uncertainty in automated linked data integration[M]// Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXVII. Berlin: Springer, 2018;81-112.
- [58] VAN KEULEN M. Managing uncertainty:The road towards better data interoperability [J]. IT-Information Technology, 2012,54(3):138-146.
- [59] CAMACHO-RODRÍGUEZ J,CHAUHAN A,GATES A,et al. Apache hive:From mapreduce to enterprise-grade big data warehousing[C]//Proceedings of the 2019 International Conference on Management of Data. 2019;1773-1786.
- [60] ARMBRUST M,XIN R S,LIAN C,et al. Spark sql:Relational data processing in spark[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of data. 2015;1383-1394.
- [61] SETHI R,TRAVERSO M,SUNDSTROM D,et al. Presto:SQL on everything[C]//2019 IEEE 35th International Conference on Data Engineering(ICDE). IEEE,2019;1802-1813.
- [62] FERNANDES S,BERNARDINO J. What is bigquery? [C]// Proceedings of the 19th International Database Engineering & Applications Symposium. 2015;202-203.
- [63] DAGEVILLE B,CRUANES T,ZUKOWSKI M,et al. The snowflake elastic data warehouse[C]//Proceedings of the 2016 International Conference on Management of DaTA. 2016;215-226.
- [64] AGUILAR-SABORIT J,RAMAKRISHNAN R,SRINIVASAN K,et al. POLARIS:the distributed SQL engine in azure synapse [C]//Proceedings of the VLDB Endowment. 2020;3204-3216.
- [65] WANG Y,SONG S,CHEN L. A survey on accessing dataspace [J]. ACM SIGMOD Record,2016,45(2):33-44.
- [66] LI Y,MENG X. Research on personal dataspace management [C]//Proceedings of the 2nd SIGMOD PhD Workshop on Innovative Database Research. 2008;7-12.
- [67] SONG S,CHEN L,YUAN M. Materialization and decomposition of dataspace for efficient search[J]. IEEE Transactions on Knowledge and Data Engineering,2010,23(12):1872-1887.

- [68] DONG X, HALEVY A. Indexing dataspaces[C]// Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 2007:43-54.
- [69] LAL N, SINGH M, PANDEY S, et al. A Proposed Ranked Clustering Approach for Unstructured Data from Dataspace using VSM[C]// 2020 20th International Conference on Computational Science and Its Applications (ICCSA). IEEE, 2020:80-86.
- [70] SALLES M A V, DITTRICH J, BLUNSCHI L. Intensional associations in dataspaces[C]// 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010). IEEE, 2010:984-987.
- [71] YANG D, LI L, ZHANG M. Semantic Keyword Query Mechanism Based on Entity Association Graph in Dataspaces[C]// 2012 Fourth International Conference on Computational Intelligence and Communication Networks. IEEE, 2012:912-914.
- [72] DURNER D, LEIS V, NEUMANN T. JSON tiles: Fast analytics on semi-structured data[C]// Proceedings of the 2021 International Conference on Management of Data, 2021:445-458.
- [73] KASSELLA E, KONSTANTINOU I, KOZIRIS N. Towards a Multi-engine Query Optimizer for Complex SQL Queries on Big Data[C]// 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019:6095-6097.
- [74] LAL N, QAMAR S. Comparison of ranking algorithms with dataspace[C]// 2015 International Conference on Advances in Computer Engineering and Applications. IEEE, 2015:565-572.
- [75] LAL N, QAMAR S, SHIWANI S. Search ranking for heterogeneous data over dataspace[J]. Indian Journal of Science and Technology, 2016, 9(36):1-9.
- [76] CAO Y, FAN W, WANG Y, et al. Querying shared data with security heterogeneity[C]// Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020:575-585.
- [77] DITTRICH J P. iMeMex: A platform for personal dataspace management[C]// Proceedings of SIGIR PIM Workshop, ACM, 2006:40-43.
- [78] DONG X L, HALEVY A. A platform for personal information management and integration[C]// Proceedings of VLDB 2005 PhD Workshop, 2005:26-30.
- [79] CHUNG P W H, LIAO Z. Cross-organisation dataspace (COD)-architecture and implementation[C]// 2008 International Conference on Computer Science and Software Engineering. IEEE, 2008:448-451.
- [80] KUICHEU N C, WANG N, NARCISSE F T G, et al. Building semantic relationships incrementally in DataSpace[C]// 2009 First International Conference on Information Science and Engineering. IEEE, 2009:2288-2291.
- [81] JIANG X, SUN X, ZHUGE H. A Resource space model for dataspace[C]// 2010 Sixth International Conference on Semantics, Knowledge and Grids. IEEE, 2010:33-41.
- [82] ELSAYED I, BREZANY P. Dataspace support platform for e-science[J]. Computer Science, 2012, 13(1):49-61.
- [83] ANGRISH A, STARLY B, LEE Y S, et al. A flexible data schema and system architecture for the virtualization of manufacturing machines (VMM) [J]. Journal of Manufacturing Systems, 2017, 45:236-247.
- [84] MCHUGH J, CUDDIHY P E, WILLIAMS J W, et al. Integrated access to big data polystores through a knowledge-driven framework[C]// 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017:1494-1503.
- [85] ELSAYED I, BREZANY P. Towards large-scale scientific dataspaces for e-science applications[C]// International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2010:69-80.
- [86] BORJIGIN C, ZHANG Y, XING C, et al. Dataspace and its application in digital libraries[J]. The Electronic Library, 2013, 31(6):688-702.
- [87] MACHADO J C, AMORA P R P. The Impact of Privacy Regulations on DB Systems[J]. Journal of Information and Data Management, 2021, 12(5):428-439.
- [88] LIU Y F, WANG N, WANG Z G, et al. Collection and analysis of multi-dimensional category data under the mix-up differential privacy model [J]. Journal of Software, 2021, 33(3):1093-1110.
- [89] ZHOU C X, SUN Y, WANG D G, et al. A review of federal learning research[J]. Journal of Network and Information Security, 2021, 7(5):77-92.
- [90] GENTRY C. Fully homomorphic encryption using ideal lattices [C]// Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, 2009:169-178.
- [91] CUI J J, LONG J, MIN E X, et al. A review of homomorphic encryption applications in cryptographic machine learning [J]. Computer Science, 2018, 45(4):46-52.
- [92] WEI S, ZHIHUA H, ZIKAI W, et al. A method and application for constructing a authentic data space[C]// 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). IEEE, 2019:218-224.
- [93] DOLGUI A, IVANOV D, POTRYASAEV S, et al. Blockchain-oriented dynamic modelling of smart contract design and execution in the supply chain[J]. International Journal of Production Research, 2020, 58(7):2184-2199.
- [94] NATHAN S, GOVINDARAJAN C, SARAF A, et al. Blockchain meets database: Design and implementation of a blockchain relational database[J]. arXiv:1903.01919, 2019.
- [95] BELOTTI M, BOŽIĆ N, PUJOLLE G, et al. A vademecum on blockchain technologies: When, which, and how[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4):3796-3838.



FAN Shuhuan, born in 1988, Ph.D, is a member of China Computer Federation. Her main research interests include data management and big data analysis.



HOU Mengshu, born in 1971, professor, Ph.D supervisor, is a senior member of China Computer Federation. His main research interests include data management and natural language processing.