跨域数据管理的内涵与挑战

柴云鹏 李 彤 范 举 等
中国人民大学

关键词：跨域数据管理 跨空间域 跨管辖域 跨信任域

数据跨域共享与协同

近年来，以数据为核心的数字经济蓬勃发展，但“数据孤岛”问题仍普遍存在于政务、教育、医疗、商业等行业中，如跨省市医保及健康码互认问题，严重制约数字化社会的进一步发展。数据的价值遵循著名的梅特卡夫法则（Metcalfe's law）：网络节点越多，每个节点价值越大，“增值”呈指数级变大。要使数据价值最大化，就需要打破“数据孤岛”，促进数据要素的共享与协同。数据要素共享与协同要求多个数据市场主体为了实现共同的目标，充分释放数据要素的价值，依靠多个市场主体共同的力量实现数据价值的最大化。然而，当前数据要素共享与协同过程中存在“不会、不愿、不敢”的关键问题，阻碍了数据价值有序释放。对数据要素跨域、高效、安全地共享与协同的需求，催生了跨域数据管理。

自20年前起，国家陆续启动南水北调、西气东输、西电东送等重大工程，进行物理世界资源的跨域管理；2022年初，国家完成全国一体化大数据中心体系总体布局设计，正式启动“东数西算”工程，在京津冀、长三角、粤港澳等八大区域部署国家算力枢纽节点，建设全国一体化算力网络。这一系列重大举措，为数字世界的数据跨域共享与协同提供了重要的基础设施，为跨域数据管理提供了基本条件。

跨域数据管理的内涵

数据管理是指基于计算机技术，高效、安全、经济地对数据进行摄取、处理、存储和使用的过程。简言之，数据管理是使数据跨越时间维度仍然保持高效与可用的技术。然而传统数据管理局限在单一企业、业务、数据中心等内部，即使少数分布式数据管理系统采用异地多中心方案，也主要用于容灾等极端情况。

伴随着数字经济时代数据要素流通的大趋势，以及算力网络等全国范围内广域基础设施的完善，数据管理正在从面向和限定于单域的孤立服务发展到跨域的共享与协同服务的阶段，即跨域数据管理。跨域为数据管理带来了全新的挑战：在通信层面，面临跨空间域的挑战，体现为不确定性网络的问题；在数据建模层面，面临跨管辖域的挑战，体现为异构模型融合的问题；在安全隐私保护层面，面临跨信任域的挑战，体现为隐私计算的问题。如图1所示，跨域数据管理的内涵可以分解为以下三个层次。

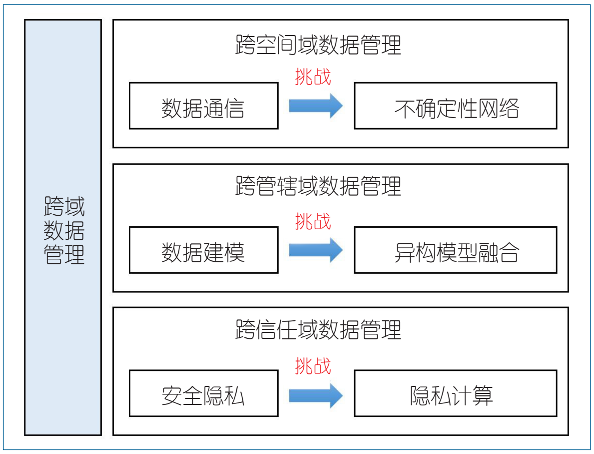

图1 跨域数据管理的内涵

1. 跨空间域数据管理。地域间的远距离决定了跨地域网络传输具有较高的基础时延（通常为几十到几百毫秒）；而广域网数据的端到端传输和介质共享特性，也为数据传输时延带来不确定性。跨地域将显著影响分布式数据管理系统中的事务处理效率，造成性能显著下降，甚至不可用。因此，跨空间域数据管理需要减少不确定性网络对事务处理等关键操作性能的影响。

2. 跨管辖域数据管理。数据管理跨越多个数据管辖域，这些数据管辖域的数据类型、模式和标准不统一，呈现异构特征。为了更好地支持全国范围内的数据要素共享与流通，跨域数据管理迫切需要广域范围内的统一查询体系，但统一查询非常复杂，且存在模型异构性、查询语言多样性、语义异构性等严峻挑战。因此，跨管辖域数据管理需要实现异构数据和模型的高效融合。

3. 跨信任域数据管理。跨域数据有不同的归属，数据所有者有版权保护、隐私安全的考虑，无法简单地将所有数据集中摄取、处理和存储；且全国大范围的数据集中、数据存储和访问规模等压力也非单一数据管理系统可以承受。因此，跨信任域数据管理需要实现隐私计算以支持不同信任域之间的安全数据流通。

跨空间域数据管理的挑战

跨地域网络节点间传输的时延和不确定性显著高于局域网。以“东数西算”场景为例，时延不确定性的主要影响因素有三个[1]：(1）西部超大型数据中心远离国家骨干网节点，与东部缺少直连链路，专线部署成本高，而公网时延具有极大不确定性；(2）由于行业竞争和统一维护等原因，移动、联通、电信三网互通的数据中心很难合规地建立（例如电信用户无法直接连接移动的服务器），这会导致数据传输时必须跨网跳转，直接增加了网络时延和不确定性；(3）出于成本考虑，西部超大型数据中心建立时，仅选择1~2家运营商网络接入，全运营商接入的数据中心比较少，使得跨网跳转不可避免。

由于域内和跨域节点间网络延时和不确定性之间的不平衡（通常情况下域内时延比域间时延低3个数量级，域内带宽比域间带宽高出1~2个数量级），跨域节点间的网络传输成为制约整个系统性能提升的瓶颈。以分布式事务为例，事务参与节点执行本地数据的读写操作，协调节点负责协调所有参与节点的提交或回滚，保证事务的原子性、一致性、隔离性和持久性。其中协调节点与参与节点之间的通信，包括读写操作、两阶段提交（Two-Phase Commit, 2PC），可能会涉及跨地域节点的网络传输。特别是2PC中两个阶段是串行进行的，只有协调节点完成与所有参与节点之间的通信之后，才能进入到提交阶段。跨地域环境下，由于传输基础时延大、时延抖动具有不确定性，一旦某个参与节点没有向协调节点返回确认信息，则事务处理无法进入到提交阶段。类似地，在提交阶段，只要存在参与节点因为网络无法完成与协调节点之间的通信，该事务就无法完成提交。在这两种情况下，事务处理的时延会增加，系统的吞吐会降低。

缓解跨地域通信时延不确定性对数据管理性能的影响，有两种基本的思路。(1）第一种思路是优化分布式数据管理过程，尽可能减少网络中传送信息的次数和数据量。例如尽可能将同步操作转换成异步操作，避免高延时低带宽网络造成的性能开销，也可以选择优化分布式事务和分布式共识协议等，特别是减少网络传输轮次，并与新兴网络技术紧密结合，实现协同优化。(2）第二种思路是优化网络传输通道，降低广域网传输的不确定性。首先需要以用户为中心构建骨干网、城市群、城市内等多级时延圈，满足不同业务诉求，并可以根据业务属性进行调度。这种通过网络架构构建的低时延圈，只能保证静态时延的确定性；而业务动态运行时的时延确定性，则需要确定性低时延传输技术来保障。当前业界提出了IEEE 802.1时间敏感网络和确定性网络[2]，分别在数据链路层和网络层通过资源预留实现确定性数据传输。然而，广域网中的链路和中间节点是不可控的，链路层和网络层资源预留和时间同步的方案开销大，在广域网中难以部署。针对

广域网，华为提出基础网络架构New IP[3]，包含确定性IP和新传输层技术，是确定性广域网研究的重要尝试，但这是一个全新的设计，其规模化部署仍然任重道远。因此，基于现有的IP架构进行面向确定性低时延传输技术创新和平滑演进，仍然是实现高性能跨地域数据管理的当务之急。

跨管辖域数据管理的挑战

不同的数据管辖域，存在模型各异、松散耦合、彼此自治的特点，给跨域数据管理带来了严峻的异构性挑战。首先，作为数据管理最核心的功能之一，跨域数据查询面临着两个难题：(1) 不同数据管辖域的数据库蕴含的数据模型和查询语言不同，无法直接进行统一的查询；(2) 缺乏统一的优化机制，不同数据模型的数据库适用的场景各异，结合上层应用特点，研究统一的查询优化方法对提升整体查询性能非常关键，也颇具挑战。其次，多个数据管辖域可能在数据类型、数据模式等方面存在显著的语义差异，为了支撑数据的跨域访问，需要在语义层面对跨域数据进行高质量的融合。数据融合（data integration，也称数据集成）的目标是整合多源异构数据，形成统一的数据视图，包括多源模式匹配、实体表示对齐、属性冲突消解等多个挑战性任务。因此，应对跨管辖域数据管理的挑战，可以从跨域数据的统一查询与优化，以及跨域异构大数据的语义融合两个方面展开探索。

针对跨域数据查询的挑战，解决方案有三类：物理汇聚、联邦数据库（federate database）、多存储数据库（polystore database）。由于前两类方法并不适用于跨域数据管理的场景，近些年多存储数据库系统受到了广泛关注，其目标是提供对多个异构数据库（如关系数据库、NoSQL数据库或文件系统）的统一访问，现有的技术方案可以分为松耦合、紧耦合以及混合系统。松耦合多存储系统[4]的基本思想是采用“中介器－包装器”架构：用户使用统一的查询语言进行查询；中介器将查询转换为若干子查询，每个子查询对应一个异构数据库，由相应的包装器负责处理；包装器将子查询翻译为自身数据库的查询语言，并反馈查询结果。最终，系统将来自包装器的结果进行集成，并返回给用户。紧耦合多存储系统[5]与松耦合系统的区别在于其查询处理器可以在查询执行期间直接访问底层数据库，使数据能够在不同数据库之间高效移动，从而优化整体的查询性能。混合系统[6]试图结合松耦合系统与紧耦合系统的优点，例如前者可以更好地支撑异构的数据库，而后者能够直接通过其接口高效访问某些数据。

针对异构数据融合的挑战，ACM图灵奖获得者迈克尔·斯通布雷克（Michael Stonebraker）教授将数据融合技术划分为三个发展阶段[7]，其中第三个阶段是大规模高质量数据融合，主要的特点是通过“人在回路”（human-in-the-loop）机制充分利用人的认知推理能力和机器的计算能力，解决大规模数据源的语义融合难题。人在回路的数据融合技术近年来也备受工业界和学术界的重视，包括沃尔玛、阿里巴巴在内的多家企业尝试利用人在回路的方法解决大规模的数据融合难题。学术界主要研究如何通过有效的人机交互机制，高效高质地完成数据转换、清洗、集成等相关操作[8]，实现大规模数据融合任务的“提质增效”。

总的来说，现有工作没有针对跨管辖域数据管理中存在的模型各异、松散耦合、彼此自治的挑战设计通用的方法和高效的系统，更多的是解决一些具体的技术与算法问题。因此，亟待研发通用的统一查询优化方法和高质量语义融合技术，更好地支持全国范围内的数据要素高效共享与流通，支撑跨管辖域数据增值。

跨信任域数据管理的挑战

信任域是指各企业、组织或机构设置本地认证服务形成相对独立的域，每个信任域包含不同的用户及不同的网络资产和数据对象，是控制访问的主要手段。然而，信任域也造成了分散的“数据孤岛”，阻碍机构间可能的数据共享与协同。

现实情况中，受到隐私保护的严格限制，构建有

效的跨信任域数据管理存在以下挑战：(1) 数据组织方面，共享资源的分散性和异质性限制了数据的汇聚和共享。跨域数据访问需要保护数据拥有者和访问者双方的隐私安全，而且各数据拥有方存在数据库系统异构、存储模式异构等问题。(2) 数据使用方面，保障数据在计算实体之间自由安全地流动是一项巨大挑战。专用芯片等集成技术的出现，使隐私保护计算成为可行解决方案。然而，由于性能瓶颈、技术缺乏可解释性等问题，该技术仅适用于小规模计算。(3) 数据存储方面，域内及域间的隐私保护需要加强。例如，杜绝可信用户非法访问域内敏感数据，避免跨域计算时存储系统遭受恶意攻击等。

许多新兴技术为实现信任域间数据安全高效共享开辟了新思路：(1) 数据联邦一定程度上解决了"数据孤岛"问题，使各数据拥有方能够在保护隐私的前提下完成联合查询。联邦计算凭借"数据不动计算动"的核心思想，将任务拆分至各方自行完成计算，最后汇总结果，避免敏感数据的跨域访问和传输。(2) 全同态加密（FHE）、安全多方计算（MPC）、差分隐私（DP）等新型隐私计算软件技术的通信开销巨大，需权衡实用性和保障性。(3) 基于硬件的可信执行环境（TEE）在硬件和操作系统层面为数据和程序提供隔离的运行环境，避免不可信的窃取或篡改[9]。(4) 以TEE为代表的新型硬件技术为密态数据处理系统注入了新的活力，可确保在所有状态（使用、传输、静态存储）下的数据保护。(5) 跨域数据共享对安全性和隐私保护的严格要求，与区块链去中心化和留痕不可篡改的特性高度符合，因此一些研究以区块链作为底层数据库实现多信任域分布式系统[10]。

结语

数字经济是继农业经济、工业经济之后的主要经济形态。数据成为继农副产品、工业品之后的全新流通要素，而促进数据要素在跨空间域、跨管辖域和跨信任域的流通以使数据价值最大化，正是跨域数据管理的意义所在。

柴云鹏

CCF专业会员，CCF教育工委委员，CCF数据库专委会执行委员、信息存储技术专委会执行委员。中国人民大学信息学院教授、计算机系主任。主要研究方向为云计算、数据库系统、分布式系统等。ypchai@ruc.edu.cn

李　彤

CCF专业会员。中国人民大学数据工程与知识工程教育部重点实验室副教授。主要研究方向为新型互联网体系结构、分布式系统和广域网数据管理等。tong.li@ruc.edu.cn

范　举

CCF专业会员，CCF数据库专委会执行委员。中国人民大学数据工程与知识工程教育部重点实验室教授。主要研究方向为AI4DB、人在回路的数据准备、大数据管理与分析等。fanj@ruc.edu.cn

其他作者：卢　卫　张　峰　杜小勇

参考文献

[1] 王建冬, 于施洋, 窦悦. 东数西算：我国数据跨域流通的总体框架和实施路径研究[J]. 电子政务, 2020 (3): 13-21.

[2] Farkas J, Varga B, Thubert P, et al. Deterministic Networking Architecture. RFC 8655. 2017.

[3] 郑秀丽, 蒋胜, 王闯. NewIP:开拓未来数据网络的新连接和新能力. 电信科学. 2019 (9): 1-10.

[4] Simitsis A, Wilkinson K, Castellanos M,et al. Optimizing Analytic Data Flows for Multiple Execution Engines[C]// *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data(SIGMOD'12)*, 2012:829-840.

[5] Alotaibi R, Bursztyn D, Deutsch A, Manolescu I,et al. Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue[C]// *Proceedings of the 2019 International Conference on Management of Data (SIGMOD'19)*, 2019:1660-1677.

[6] 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 (SIGMOD'15)*, 2015: 1383-1394.

[7] https://www.oreilly.com/library/view/getting-data-right/9781491935361/ch01.html.

[8] 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.
[9] 童咏昕, 李书缘, 毛睿. 浅谈数据库的隐私计算[J]. 中国计算机学会通讯, 2022, 18(6):39-45.
[10]Lu Y, Huang X, Dai Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT[J]. *IEEE Transactions on Industrial Informatics*, 2019, 16(6): 4177-4186.