面向数据要素高效共享流通的跨域数据管理关键技术研究

Research on Key Technologies of Cross Domain Data Management for Efficient Sharing and Circulation of Data Elements

范　举，杜小勇（中国人民大学，北京 100872）
Fan Ju，Du Xiaoyong（Renmin University of China，Beijing 100872，China）

摘　要：
现有数据管理技术难以解决跨域数据共享流通面临的跨域异质数据语义难统一、跨域共享流通隐私难保护、跨域数据查询性能难优化等问题，严重制约了数据共享流通的高效性。因此，跨域数据管理近年来受到了学术界和工业界的关注。介绍了跨域数据管理的基本概念与关键技术挑战，回顾了现有与跨域数据管理相关的研究，最后讨论了跨域数据管理的一些重要研究问题。

关键词：
数据管理；跨域共享；数据流通
doi：10.12045/j.issn.1007-3043.2025.05.007
文章编号：1007-3043（2025）05-0037-07
中图分类号：TN919.2
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Existing data management techniques struggle to address the challenges of cross-domain data sharing and circulation，including difficulties in unifying heterogeneous data semantics across domains，protecting privacy in cross-domain sharing and circulation，and optimizing query performance across domains. These issues severely limit the efficiency of data sharing and circulation. As a result，cross-domain data management has attracted increasing attention from both academia and industry in recent years. It introduces the basic concepts and key technical challenges of cross-domain data management，reviews existing research related to this field，and discusses potential research questions.

Keywords：
Data management；Cross-doman sharing；Data circulation

引用格式：范举，杜小勇. 面向数据要素高效共享流通的跨域数据管理关键技术研究[J]. 邮电设计技术，2025（5）：37-43.

0 引言

数据是数字时代的关键生产要素，具有倍增效应、叠加效应，能够赋值、赋能社会经济发展全过程，驱动社会经济数字化转型、网络化重构和智能化提升[1]。围绕数据价值的发挥，加快前沿数据技术融合和技术突破，有效支撑数据要素共享流通与价值释放，既是《关于构建数据基础制度更好发挥数据要素作用的意见》、《数字中国建设整体布局规划》等一系列国家层面战略规划共同关注的重点，也是培育数据要素市场和产业生态构建的迫切需求。

为了有效支撑这一系列国家层面的制度实践，培育数据要素高效共享流通的技术体系和产业生态十分迫切。以北京为例，围绕“数字经济国际标杆城市”和“数据基础制度先行区”建设，形成京津冀、长三角、粤港澳之间超大城市群的联动，亟需解决数据在城市治理复杂场景的共享流通与有效利用，从而以高质量的数据价值释放支撑高质量的经济社会发展。然而，随着数据要素共享流通规模的不断扩大和应用范围的不断扩展，越来越多的场景面临着因“跨域”而带来的数据管理难题。具体而言，在城市治理等复杂场景中，数据共享流通呈现出跨部门、跨层级、跨主体等显

收稿日期：2025-03-07

著的“跨域”特征，使数据管理的复杂度大大提升，同时面临严峻的高效性挑战，这对数据管理技术提出了新的要求。

为了有效应对数据跨域共享流通中的高效性挑战，跨域数据管理近年来受到了学术界和工业界的关注[2-3]。传统的以数据库管理系统为代表的数据管理技术主要关注单一企业、部门等单域场景，侧重于对域内数据进行存储、查询和分析。而跨域数据管理是指对分散于不同域（部门、层级等）的数据进行统一的管理，实现数据在不同域间进行高效且安全的共享流通，并为不同的应用场景提供统一的查询方式。

为了更好地对跨域数据管理进行说明，这里以某市金融数据专区为例（见图 1），这是某市大数据中心建设的全国首个公共数据专区，为多家金融机构提供数据服务，亟需解决个人/企业的大规模数据跨越公安、税收、民政等30多个部门高效共享流通问题。这里面存在一系列制约高效性的跨域数据管理问题。

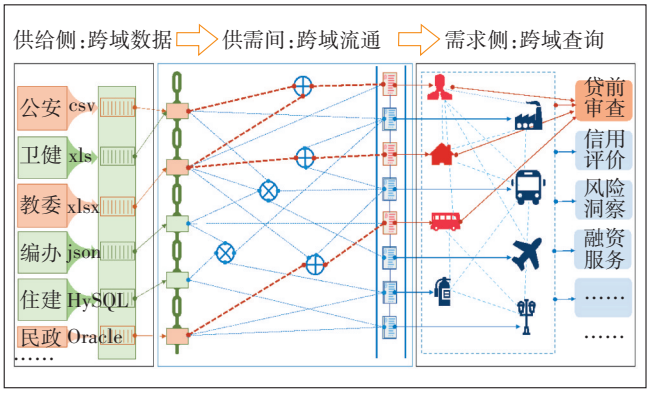

图1　某市金融数据专区跨域数据管理示意

首先在供给侧，也就是跨域数据层面，需要融合社保、税收等多个数据项，这些数据源分散在多个部门的多个数据源中，存在大量异质数据模式，存在大量同名不同义、同义不同名问题，因此语义理解难，数据找不准。其次在供需间，也就是跨域流通层面，金融专区大量的数据是隐私敏感数据，跨越多个部门流通。数据流经不同域时的访问权限和隐私保护要求千差万别：有些可以公开访问，有些需要经过差分隐私，有些则需要通过多方安全计算。因此跨域协同保护慢，数据流通不畅。最后在需求侧，专区通过查询的方式提供服务，需要支持日均十万级调用、毫秒级响应需求。但数据来自政务云、教育云等领域的私有云等资源异构的底层数据管理系统，它们的软硬件环境不一，给数据的实时查询带来挑战。

1 跨域数据管理的技术挑战

与传统数据管理技术相比，跨域数据管理亟需破解以下3个方面的主要矛盾（见图 2）。

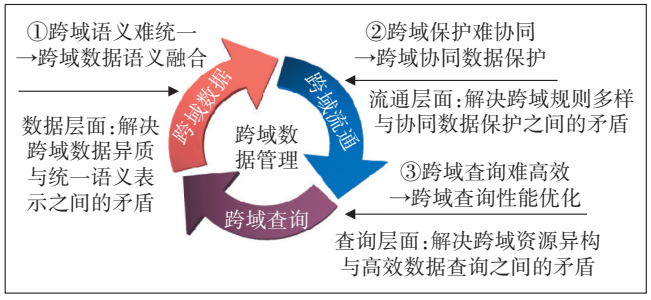

图2　跨域数据管理亟需解决的主要矛盾

a）在数据层面，亟需解决跨域数据异质与统一语义表示之间的矛盾。

b）在流通层面，亟需解决跨域规则多样与协同数据保护之间的矛盾。

c）在查询层面，亟需解决跨域资源异构与高效数据查询之间的矛盾。

1.1 跨域语义融合挑战

与传统数据管理相比，“跨域”带来了开放环境下的语义异质性挑战。具体而言，数据的语义既是异质的又是动态多变的，取决于其所在的域或是所处的应用需求。例如：学历信息不仅来自教育、公安等十余个部门、数十个数据源，而且被社会信用、城市治理等多种场景所使用，存在数据结构不统一、数据语义不一致等一系列问题，这些问题在开放环境下变得极具挑战性。

针对开放环境下的语义异质性挑战，需要研究跨域数据语义融合问题，形成物理跨域但语义统一的数据表示，并支撑高效的跨域数据发现。然而，现有的研究工作侧重解决面向特定领域或应用的语义融合问题，在跨域场景下难以支持有效的跨域语义表示与高效语义匹配。一方面，在开放环境下，跨域数据结构多样（结构化、半结构化、非结构化）、数据语义难对齐给跨域语义表示带来了严峻的挑战，现有工作采用基于规则或知识图谱的方法进行语义表示，存在规则不适用、知识图谱不完备等问题。另一方面，在开放环境下，跨域数据的需求开放、跨域数据源动态多变，如何为需求精准地匹配到语义相关的数据极具挑战。现有研究主要使用关键词匹配或借助知识图谱来度量用户需求与数据之间的相关性，存在过度依赖领域

知识、语义相关性度量效果不佳等难题。因此，十分迫切需要提出高效的语义匹配方法，以支持语义相关的跨域数据发现方法。

1.2 跨域协同数据保护挑战

与传统数据管理相比，“跨域”带来了数据在不同部门、层级间流通的隐私保护问题。例如：在群租房治理应用中，需要关联房屋数据（住建部门）、用水数据（自来水公司）、手机信令数据（电信运营商）等多源数据进行对比分析，而这些数据都有严格的隐私保护规则，不能像传统数据管理那样简单地以明文的方式出域并进行跨域流通。此外，即便是对同样的数据，不同部门也可能会有不同的隐私保护规则。

针对跨域数据在隐私保护层面上存在的规则多样挑战，需要形成能够避免跨域冲突，支持跨域协作的数据保护方案，以支持跨域数据在满足隐私保护的前提下进行高效共享流通。然而，现有隐私保护方法在跨域数据共享流通的场景下面临挑战。一方面，跨域数据的隐私保护规则多样，甚至可能会出现跨域相互冲突的情况，给现有的隐私保护方法带来了新的挑战，即需要判定跨域数据之间是否存在可行的协同隐私保护方案，若有规则冲突应如何消解冲突等。另一方面，隐私规则的多样性使得联合查询的隐私算子需要采用不同的隐私保护策略，而不同的隐私保护策略、算子执行顺序、数据查询粒度会产生不同的计算和通信时延，使得整个联合查询变成一个复杂的组合优化问题，因此有必要对这一问题进行建模和求解，提出跨域隐私保护算子的协同计算策略，以便有效提升数据跨域共享流通的高效性。

1.3 跨域查询性能优化挑战

为不同的应用场景提供统一的查询方式（如SQL查询语言）是跨域数据管理的核心问题。然而，与传统数据管理相比，跨域数据查询跨越多个域，呈现出跨域数源自治的特点，给高效联合查询带来了挑战。例如：为了应对紧急事件，管理部门为寻找风险目标常常需要联合查询交通、医疗等多个部门的数据，并要求在极短时间内返回结果，而各部门的数据存储于域内的数据管理系统，这给在线联合查询带来了极大的困难与挑战。

针对跨域查询层面的数据源自治性挑战，需要解决跨域查询的性能优化问题，实现跨域查询的高效执行与动态调度，支持跨域数据的高效联合查询。然而，现有数据管理中的查询优化研究在跨域场景下面临着新的研究挑战。一方面，跨域场景下存在查询负载多样、查询计划空间指数级增大的特点，给跨域查询优化带来挑战。现有研究难以针对跨域场景生成高效的查询计划与动态查询调度，因此难以支持跨域场景自适应的查询优化。另一方面，跨域底层数据库的类型多样性、语法差异性给在线查询的高效执行带来挑战，即各域的数据引擎按照自身语法规则进行算子执行，缺乏跨域自动查询翻译。因此，亟需提出场景适应的跨域查询优化方法，以提升查询性能。

2 跨域数据管理的研究现状

2.1 跨域数据语义融合的研究现状

现有跨域数据语义融合研究主要包括结构化数据抽取、表格数据的表示学习、最近邻检索3个方面。下面详细阐述这3个方面的相关工作。

首先，不同结构数据的结构化是一项复杂而长期存在的数据管理挑战，受到了广泛关注[4-5]，将异构数据转换为结构化数据库不仅需要处理不同格式和类型的数据，还需要解决各种语义和语法差异所带来的问题。目前主流的做法是利用深度学习技术，通过大规模数据训练模型来自动学习数据的语义和语法规律，具有更高的灵活性和智能性，能够适应不同的数据源和文档类型，而不依赖于预定义模式。对于从Web数据生成结构化数据抽取问题[6-8]，现有研究使用远程监督训练特定的抽取模型[9]，或依赖特定属性和值所在的位置[10-11]。这些方法仍然存在一些不足：它们需要大量领域和文档格式的特定训练；另外它们侧重于对句子进行推理，而不是长文档，并且依赖高质量的语言工具（例如依赖解析、词性标注、实体识别）来帮助在非结构化文本上引入结构[12-13]。

其次，表示学习是语义融合任务中的重要步骤，通过不同的编码器将数据元组、数据列、字符串等映射到向量空间。随着表示学习模型的不断迭代升级，语义融合任务中数据映射的方式也趋于多样化。DeepER中使用LSTM来获取表示向量[14]，这些表示向量在生成过程中考虑到了上下文信息，能够提取出文本数据中序列依赖关系和局部特征。GCN[15]设计了一种用于图结构数据处理的神经网络架构，通过图的结构信息来学习节点的有效表示，使得学习到的表示能够反映节点的拓扑信息。Ditto利用预训练语言模型（如BERT、RoBERTa等）学习文本的深层次表示，并在特定领域上利用标注数据进行模型微调[16]，使模型能

够捕捉实体的细微差异和复杂的上下文信息,提高了实体匹配任务的精度。

最后,最近邻检索是一种在机器学习、数据挖掘领域中的重要技术,在语义融合场景下,用于在给定的数据模式中查找与目标样本最相似的邻居数据模式样本,现有的工作着重于研究如何进行高效搜索。现有的最近邻搜索方法包括基于图的近似最近邻检索算法[17-18]和乘积量化方法[19]。其中,基于图的近似最近邻搜索算法以HNSW为代表[17],其基本思想是构建一种具有层级结构的图,其中每个节点都连接到一组“导航节点”,这些节点被选择为相似度较高的节点。这种层级结构的图使得在搜索过程中能够快速定位到潜在的最近邻节点,从而大大加速了近似最近邻检索的速度。乘积量化方法[19]的基本思想是将原始高维向量划分为若干个子向量,并对每个子向量进行量化。这样,一个高维向量就可以表示为一组离散的子向量索引,而不是原始的连续数值。然后,可以利用这些离散的子向量索引来进行近似最近邻检索。

总之,现有方法主要适合单域、提供大量训练数据的特定场景,面对海量跨域数据场景,存在跨域语义知识匮乏、训练数据不足和融合效率不高等问题。

2.2 跨域协同数据保护的研究现状

针对数据流通场景中的协同数据保护,目前SecretFlow[20]实现了支持SQL查询的统一框架下的多种隐私计算算子,HuFu[21]则实现了数据联邦下的空间查询。下面详细阐述这2个方面的相关工作。

首先,在跨域隐私规则建模方面,现有工作主要集中在对数据源的安全规则建模和对数据流的权限延伸控制。在对数据源的安全规则建模方面,目前的技术主要包括以单机数据共享为目的的主机访问控制和以单域内部数据共享为目的的访问控制等。还有一些模型通过基于属性的访问控制将时间[22]、空间位置[23]、访问历史[24]、运行上下文[25]等要素作为访问主体、访问客体和环境的属性来控制数据访问行为,通过定义属性间的关系来描述复杂的授权和访问控制约束。在延伸控制方面,目前的工作集中于基于起源的访问控制PBAC[26-27]和黏性策略SP技术[28-29],其中PBAC技术给出了形式化策略规约语言,而SP技术则使用加密机制将策略与数据相关联,但是目前这些技术未限定对哪些要素进行考量,且不能适应复杂的查询执行安全环境和复杂的数据关联关系。

其次,目前从多算子协同计算角度来提高隐私计算效率的技术尚未被充分探索,一些工作通过用于数据库连接算子的隐私集合求交技术[30]等密码学方法或近似查询处理[31]等方法提升对联邦数据库的查询效率,还有一些工作通过借助可信第三方聚合计算、TEE等安全工具[32]实现更高效的联合计算。但是这些工作仅聚焦于单一算子上隐私计算算法的优化,对多种隐私算子之间的联合优化探索不够充分,因此算子性能不佳。目前已有一些工作借助传统数据库查询优化技术对隐私算子树进行查询优化,以减少查询处理中加密计算部分的大小[33],如SMCQL通过简单地将明文算子上推或下推实现多方安全计算算子的减少使用。Secrecy基于秘密共享技术实现了多方联合分析[34],通过联合设计多层系统栈和预估多方安全计算算子执行时间进行基于成本的优化,另外还设计了一些逻辑优化规则对多种可能的查询树进行选择,以期优化查询执行中的计算代价和传输成本。但是目前这些工作只适用于单种隐私计算算子情况,并不能适用于复杂的跨域异质隐私规则环境。

总的来说,目前数据流通的隐私保护问题的相关研究无法应对跨域数据流通的新挑战,主要体现在安全规则较为简单、隐私计算技术架构较为单一且联合优化不足等方面。

2.3 跨域查询性能优化的研究现状

现有的研究工作对跨域数据查询优化涉及的关键技术进行了研究,主要包括基于代价估计的查询计划生成、基于规则转换的SQL翻译与转换和动态查询调度3个部分。首先,研究如何生成统计信息,然后以此进行代价估计与查询计划生成。其次,现有方法主要基于规则定义不同SQL方言的转换函数,然后进行不同方言之间的在线转换。最后,现有工作主要基于静态的场景进行查询的代价估计,然后进行在线并发查询的调度。下面详细阐述这3个方面的相关工作。

a) 查询计划优化主要包括代价估计与查询计划生成。首先,传统的统计信息生成方式在多维统计量上存在准确性不足的问题。对此,现在的研究工作通过对数据多维联合分布建模以估计统计信息,并未考虑跨域场景下如何进行隐私保护。例如,Naru[35]通过自回归模型[36],每一轮回归累计计算查询在该属性上的条件概率,最终得到数据所有属性上的联合概率分布。DeepDB[37]基于和积网络Sum-Product Network[38]对数据从行和列的角度递归切分成弱相关且分布简单的子集,从而建模数据的联合概率分布。除此之

外，还有UAE[39]、FLAT[40]等相似的方法。在跨域查询计划生成方面，文献[41]使用机器学习模型替代了传统优化器的代价模型，然而在计划枚举过程中需要将查询计划转换为向量，既产生非常高的开销又需要大量的训练数据来有效地训练机器学习模型。文献[42]则基于先进的数据处理系统LingoDB，在不牺牲性能的情况下保证灵活性和可扩展性。然而，由于引入了外部系统，这类方法也会带来较大的额外开销。

b）在查询自动化翻译方面，针对跨域场景下的查询方言兼容性问题，现有技术主要基于规则的方法进行识别并替换特定的关键字、调整语法结构或者转换特定的函数和操作来实现不同SQL方言之间的转换。例如，Luoma等人[43]通过设计若干个步骤逐步解析用户通过交互式接口生成的查询，将非规范化的SQL语句转换成规范化且可执行的SQL语句。JOOQ[44]则首先将查询转化为DSL，再基于规则将DSL转化为目标SQL方言。尽管这些方法在处理简单和标准化的查询时相对有效，但面对更复杂或特定方言的查询时，其准确性和灵活性受到限制。文献[45]和文献[46]基于大语言模型完成SQL方言的转化[47]。例如，智能SQL转换领航助手[47]是基于阿里千问大模型[46]的SQL方言转化工具。该工具通过将输入的SQL进行语法分析，生成合适的提示输入给千问模型，让模型输出转换后的SQL语句和对该语句的描述，初步证明了大型语言模型在SQL方言自动化翻译任务上的可行性。

c）在动态查询调度方面，Redshift提出自动化负载管理工具AutoWLM[48]，通过特征化查询，基于历史数据训练XGBoost模型[49]，在线预测查询的内存消耗量与执行时间，为用户的负载提前调整并发度，然而其机器学习模型未考虑查询计划的结构中算子之间的关系，面向异质负载的估计准确率不高。WiseDB[50]主要针对云场景进行负载管控，然而其关注指标主要面向云端的费用优化，无法针对跨域异质负载进行性能优化。在并发负载代价估计方面，BAL[51]基于线性函数进行负载的时延估计，然而其只能针对PostgreSQL的算子模型进行建模。文献[52]提出基于图神经网络的负载预测方法，针对并行场景下的OLTP类型负载进行性能预测，然而该方法主要面向事务处理类负载，无法解决分析类负载算子的可扩展问题，并且无法解决跨域场景下的动态资源调度。

总的来看，目前缺乏有效的面向跨域场景的查询优化技术，尤其是仍缺乏高准确度、低耗时，且具有良好隐私保护效用的统计信息生成技术。此外，目前缺乏高效的自适应查询优化技术，已有的计划生成方法无法面向跨域场景提供高效的查询优化。

3 跨域数据管理的研究问题

针对跨域数据语义融合、跨域协同数据保护、跨域查询性能优化这3个方面的挑战，围绕现有研究工作的局限性，本文提出应重点突破以下3个方面的研究问题（见图3）。

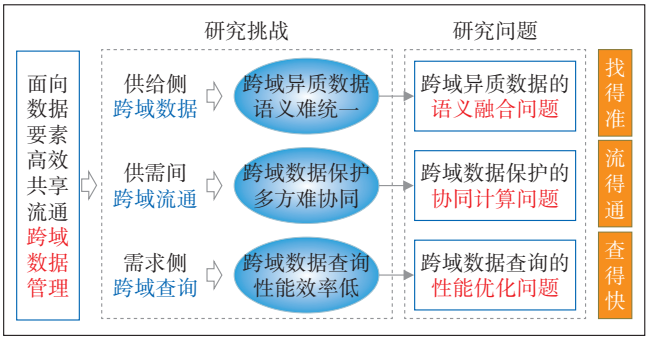

图3　跨域数据管理的研究问题

首先，跨域异质数据难统一，因此需要研究如何“找得准”，也就是跨域异质数据的语义融合问题。具体而言，需要研究如何围绕开放环境给跨域数据带来的数据结构不统一、数据语义难对齐、数据来源动态多变等挑战，针对开放环境下的跨域数据异质与统一语义表示之间的矛盾，为跨域数据生成统一的数据模式，实现跨域数据模式的统一语义表示，支持跨域数据模式的高效语义匹配，从而满足跨域数据管理对于统一语义表示与高效数据发现的核心需求。跨域异质数据的语义融合是支撑跨域数据管理的基础。

其次，跨域数据保护多方难协同，因此需要研究如何“流得通”，也就是跨域数据保护的协同计算问题。具体而言，需要研究如何围绕跨域数据安全规则多样异质、跨域多方隐私保护协同困难等挑战，针对跨域情况下的隐私规则多样与跨域隐私保护之间的矛盾，提出跨域隐私保护规则冲突避免的协同隐私保护方案，设计跨域隐私算子的协同计算方法，支持跨域隐私数据的可控生成，从而形成一套避免冲突、协作优化的跨域协同隐私保护方法。跨域共享流通的隐私保护为跨域数据高效共享流通提供了安全支撑。

最后，跨域数据查询性能效率低，因此需要研究如何“查得快”，也就是跨域数据查询的性能优化问题。具体而言，需要研究如何围绕数据源自治给跨域

查询处理带来的底层数据库类型多样、数据分布差异大、查询动态变化等挑战，针对关联分析下的跨域数源自治与高效联合查询之间的矛盾，综合考虑跨域数据分布情况与历史查询，生成代价感知的跨域查询计划，支持场景适应的跨域查询翻译与动态查询调度，从而满足跨域查询的高效性要求。跨域数据查询的性能优化是支撑跨域数据高效共享流通的关键。

4 结语

数据要素高效共享流通意义重大，既是数字基础制度建设等一系列国家层面战略规划的共同关注点，也有着十分重要的需求。本文介绍了跨域数据管理这一数据要素高效共享流通的关键支撑技术，深入探究了跨域数据语义融合、跨域协同数据保护、跨域查询性能优化3个方面的研究挑战，回顾了现有研究的相关工作。在此基础上，文章提出了3个方面的研究问题，包括针对跨域异质数据异质难统一，研究跨域异质数据的语义融合问题；针对跨域数据保护多方难协同，研究跨域数据保护的协同计算问题；针对跨域数据查询性能低效问题，研究跨域数据查询的性能优化问题。这些研究方向对推动跨域数据管理技术的不断进步，从而支撑数据要素高效共享流通，都有着十分重要的意义。

参考文献：

[1] 梅宏. 数据资源体系构建[C]//第六届数字中国建设峰会，2023.
[2] 杜小勇，李彤，卢卫，等. 跨域数据管理[J]. 计算机科学，2024，51(1)：4-12.
[3] 贾晓丰，高嵩，周琰，等. 一种面向超大城市治理的数据高效跨域流通技术框架[J]. 数据与计算发展前沿，2023，5(5)：35-45.
[4] CAFARELLA M J，SUCIU D，ETZIONI O. Navigating extracted data with schema discovery [C]//Proceedings of the 10th International Workshop on Web and Databases (WebDB 2007). Beijing：WebDB，2007：1-6.
[5] AGICHTEIN E，GRAVANO L. Snowball：extracting relations from large plain-text collections[C]//Proceedings of the 5th ACM International Conference on Digital Libraries. San Antonio：ACM，2000：1-10.
[6] CAFARELLA M J，RE C，SUCIU D，et al. Structured querying of Web text[C]//3rd Biennial Conference on Innovative Data Systems Research (CIDR). Asilomar：CIDR，2007：225-234.
[7] ETZIONI O，BANKO M，SODERLAND S，et al. Open information extraction from the Web[J]. Communications of the ACM，2008，51(12)：68-74.
[8] ETZIONI O，CAFARELLA M，DOWNEY D，et al. Unsupervised named-entity extraction from the Web：an experimental study[J]. Artificial Intelligence，2005，165(1)：91-134.
[9] LOCKARD C，SHIRALKAR P，DONG X L. Openceres：when open information extraction meets the semi-structured Web[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Volume 1 (Long and Short Papers). Minneapolis：Association for Computational Linguistics，2019：3047-3056.
[10] DENG X，SHIRALKAR P，LOCKARD C，et al. DOM-LM：learning generalizable representations for html documents[EB/OL]. [2024-08-26]. https://arxiv.org/abs/2201.10608.
[11] LIMAYE G，SARAWAGI S，CHAKRABARTI S. Annotating and searching Web tables using entities，types and relationships[J]. Proceedings of the VLDB Endowment，2010，3(1/2)：1338-1347.
[12] KOLLURU K，ADLAKHA V，AGGARWAL S，et al. Openie6：iterative grid labeling and coordination analysis for open information extraction[EB/OL]. [2024-08-26]. https://arxiv.org/abs/2010.03147.
[13] ZHOU S W，YU B W，SUN A X，et al. A survey on neural open information extraction：current status and future directions [EB/OL]. [2024-08-26]. https://arxiv.org/abs/2205.11725.
[14] EBRAHEEM M，THIRUMURUGANATHAN S，JOTY S. Distributed representations of tuples for entity resolution[J]. Proceedings of the VLDB Endowmen，2018，11(11)：2150-8097.
[15] KIPF T N，WELLING M. Semi-supervised classification with graph convolutional networks [EB/OL]. [2024-08-26]. https://arxiv. org/abs/1609.02907.
[16] LI Y L，LI J F，SUHARA Y，et al. Deep entity matching with pre-trained language models[EB/OL]. [2024-08-26]. https://arxiv.org/abs/2004.00584.
[17] MALKOV Y A，YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(4)：824-836.
[18] BAYER R，MCCREIGHT E. Organization and maintenance of large ordered indices[C]//SIGFIDET '70：Proceedings of the 1970 ACM SIGFIDET (now SIGMOD) Workshop on Data Description. Houston：Association for Computing Machinery，1970：107-141.
[19] JÉGOU H，DOUZE M，SCHMID C. Product quantization for nearest neighbor search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2011，33(1)：117-128.
[20] MA J M，ZHENG Y C，FENG J，et al. SecretFlow-SPU：a performant and user-friendly framework for privacy-preserving machine learning [C]//Proceedings of the 2023 USENIX Annual Technical Conference. Boston：USENIX，2023：17-33.
[21] TONG Y X，ZENG Y X，SONG Y，et al. Hu-Fu：efficient and secure spatial queries over data federation[J]. The VLDB Journal，2025，34(2)：19.
[22] HONG J N，XUE K P，XUE Y J，et al. TAFC：time and attribute factors combined access control for time-sensitive data in public cloud [J]. IEEE Transactions on Services Computing，2020，13(1)：158-171.
[23] XUE Y J，HONG J N，LI W，et al. LABAC：a location-aware attribute-based access control scheme for cloud storage[C]//2016 IEEE Global Communications Conference (GLOBECOM). Washing-

ton，DC：IEEE，2016：1-6.
[24] DECAT M，LAGAISSE B，JOOSEN W. Scalable and secure concurrent evaluation of history-based access control policies[C]//ACSAC '15：Proceedings of the 31st Annual Computer Security Applications Conference. Los Angeles：Association for Computing Machinery，2015：281-290.
[25] VERGINADIS Y，PATINIOTAKIS I，GOUVAS P，et al. Context-aware policy enforcement for PaaS-enabled access control[J]. IEEE Transactions on Cloud Computing，2022，10(1)：276-291.
[26] NGUYEN D，PARK J，SANDHU R. A provenance-based access control model for dynamic separation of duties[C]//2013 Eleventh Annual Conference on Privacy，Security and Trust. Tarragona：IEEE，2013：247-256.
[27] SUN L S，PARK J，NGUYEN D，et al. A provenance-aware access control framework with typed provenance[J]. IEEE Transactions on Dependable and Secure Computing，2016，13(4)：411-423.
[28] PEARSON S，CASASSA-MONT M. Sticky policies：an approach for managing privacy across multiple parties[J]. Computer，2011，44(9)：60-68.
[29] SPYRA G，BUCHANAN W J，EKONOMOU E. Sticky policies approach within cloud computing[J]. Computers & Security，2017(70)：366-375.
[30] NARAYAN A，HAEBERLEN A. Djoin：differentially private join queries over distributed databases[C]//10th USENIX Symposium on Operating Systems Design and Implementation (OSDI '12). Hollywood：USENIX Association，2012：149-162.
[31] BATER J，PARK Y，HE X，et al. Saqe：practical privacy-preserving approximate query processing for data federations[J]. Proceedings of the VLDB Endowment，2020，13(12)：2691-2705.
[32] VOLGUSHEV N，SCHWARZKOPF M，GETCHELL B，et al. Conclave：secure multi-party computation on big data[C]//EuroSys '19：Proceedings of the Fourteenth EuroSys Conference 2019. Dresden：Association for Computing Machinery，2019：1-18.
[33] BATER J，ELLIOTT G，EGGEN C，et al. SMCQL：secure querying for federated databases[EB/OL]. [2024-08-26]. https://arxiv.org/abs/1606.06808.
[34] LIAGOURIS J，KALAVRI V，FAISAL M，et al. SECRECY：secure collaborative analytics in untrusted clouds[C]//Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation. Boston：USENIX，2023：1031-1056.
[35] YANG Z H，LIANG E，KAMSETTY A，et al. Deep unsupervised cardinality estimation[EB/OL]. [2024-08-26]. https://arxiv.org/abs/1905.04278.
[36] GERMAIN M，GREGOR K，MURRAY I，et al. MADE：masked autoencoder for distribution estimation[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille：PMLR，2015：881-889.
[37] HILPRECHT B，SCHMIDT A，KULESSA M，et al. Deepdb：learn from data，not from queries![EB/OL]. [2024-08-26]. https://arxiv.org/abs/1909.00607.
[38] POON H，DOMINGOS P. Sum-product networks：a new deep architecture[C]//2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona：IEEE，2011：689-690.
[39] WU P Z，CONG G. A unified deep model of learning from both data and queries for cardinality estimation[C]//SIGMOD '21：Proceedings of the 2021 International Conference on Management of Data. Virtual Event：Association for Computing Machinery，2021：2009-2022.
[40] ZHU R，WU Z N，HAN Y X，et al. FLAT：fast，lightweight and accurate method for cardinality estimation[J]. Proceedings of the VLDB Endowment，2021，14(9)：1489-1502.
[41] JUNGMAIR M，GICEVA J. Declarative sub-operators for universal data processing[J]. Proceedings of the VLDB Endowment，2023，16(11)：3461-3474.
[42] Anon. lingo-db/lingo-db[EB/OL]. [2024-08-26]. https://github.com/lingo-db/lingo-db.
[43] LUOMA K，KUMAR A. Tech report：design and evaluation of an SQL-based dialect for spoken querying[J]. Proceedings of the VLDB Endowment，2020，14(1)：1-31.
[44] Anon. Great reasons for using jOOQ[EB/OL]. [2024-08-26]. https://www.jooq.org/.
[45] ACHIAM J，ADLER S，AGARWAL S，et al. Gpt-4 technical report[EB/OL]. [2024-08-26]. https://arxiv.org/abs/2303.08774.
[46] BAI J Z，BAI S，CHU Y F，et al. Qwen technical report[EB/OL]. [2024-08-26]. https://arxiv.org/abs/2309.16609.
[47] 阿里云．体验智能SQL转换领航助手(migration on pilot)[EB/OL]. [2024-08-26]. https://help. aliyun. com/zh/polardb/polardb-for-oracle/migration-on-pilot.
[48] SAXENA G，RAHMAN M，CHAINANI N，et al. Auto-WLM：machine learning enhanced workload management in Amazon redshift[C]//SIGMOD '23：Companion of the 2023 International Conference on Management of Data. Seattle：Association for Computing Machinery，2023：225-237.
[49] CHEN T Q，GUESTRIN C. XGBoost：a scalable tree boosting system[C]//KDD '16：Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco：Association for Computing Machinery，2016：785-794.
[50] MARCUS R，PAPAEMMANOUIL O. Wisedb：a learning-based workload management advisor for cloud databases[EB/OL]. [2024-08-26]. https://arxiv.org/abs/1601.08221.
[51] DUGGAN J，CETINTEMEL U，PAPAEMMANOUIL O，et al. Performance prediction for concurrent database workloads[C]//SIGMOD '11：Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. Athens：Association for Computing Machinery，2011：337-348.
[52] ZHOU X H，SUN J，LI G L，et al. Query performance prediction for concurrent queries using graph embedding[J]. Proceedings of the VLDB Endowment，2020，13(9)：1416-1428.

作者简介：
范举，教授，博士生导师，主要研究方向为跨域数据管理、智能数据库系统、大数据分析等；杜小勇，教授，博士生导师，主要研究方向为数据库系统、大数据管理、数据治理等。