

面向元宇宙的云边端协同 大数据管理

朱锐, 王宏志, 崔双双, 张恺欣, 燕钰

哈尔滨工业大学计算机科学与技术学院, 黑龙江 哈尔滨 150000

摘要

随着元宇宙的用户不断增多, 数据量也相应地增长, 这给元宇宙的数据管理带来了挑战。大数据管理技术对元宇宙的实现至关重要。对元宇宙中的数据管理技术进行了讨论, 将元宇宙分解为云、边、端3个层次, 分析了元宇宙中的大数据的特点和数据管理特点, 讨论了面向元宇宙的云边端数据管理的4个挑战, 并从数据模型、数据同步、数据存取、查询优化4个方面提出相应的研究路线。

关键词

元宇宙; 数据管理; 数据库

中图分类号: TP391.9

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2023011

Cloud-edge-end collaborative big data management for metaverse

ZHU Rui, WANG Hongzhi, CUI Shuangshuang, ZHANG Kaixin, YAN Yu

Faculty of Computing, Harbin Institute of Technology, Harbin 150000, China

Abstract

With the increasing number of users in the metaverse, the data also increases accordingly, which brings challenges to the data management of the metaverse. Big data management techniques are essential to realizing the metaverse. Therefore, data management technology in the metaverse was discussed. The metaverse was decomposed into three levels cloud, edge, and end. The massive data in the metaverse was analyzed. The four challenges of data management in the metaverse were discussed, and the corresponding research routes were put forward from four aspects of data synchronization, data access, data model, and query optimization.

Key words

metaverse, data management, database

0 引言

元宇宙 (metaverse) 是由 meta (意为超越) 和 verse (宇宙的缩写) 组合而成的术语, 在尼尔·斯蒂芬森 (Neal Stephenson) 于1992年撰写的《雪崩》中第一次被提到。《雪崩》描述了一个叫作矩阵的虚拟现实世界, 小说中的人们通过他们的数字化身探索一个数字世界, 以逃避现实世界的痛苦。到目前为止, 有一些“精简版”的元宇宙, 它们主要是从大型多人在线游戏发展而来的, 如 Pokemon Go、Second Life 和 World of Warcraft^[1]等。

元宇宙是计算机生成的、网络化的扩展现实 (extended reality, XR)^[2-3] 世界。扩展现实包含增强现实 (augmented reality, AR)、混合现实 (mixed reality, MR) 和虚拟现实 (virtual reality, VR) 的所有方面。参考文献[4]提出元宇宙可以是完全虚拟的, 像VR一样; 也可以是部分虚拟的, 如通过AR设备与虚拟世界交互; 同时提出了一个框架, 该框架具有3个特征 (共享、持久、去中心化), 以区分元宇宙与AR和VR。元宇宙的优势是允许用户在物理空间之外创建自己的内容, 并在网上体验只有在线下才能体验到的东西。但是, 访问元宇宙需要使用特定类型的设备 (如AR设备) 或安装程序, 这降低了元宇宙的可及性。目前, 元宇宙被定义为一个共享的虚拟3D世界, 可以通过交互和协作为用户提供全面的沉浸式体验。在元宇宙中, 现实世界中的人类与虚拟世界的自动化实体 (如虚拟人物) 进行交互。随着新兴技术 (如区块链^[5]、人工智能等) 的发展, 元宇宙被定义为一个与现实世界并行、相互映射的复杂虚拟世界和数字世界的融合。人们可以使用移动设备的AR应用与元

宇宙中的虚拟世界进行交互, 或者在游戏中体验更沉浸式的互动。

近年来, 现实世界的许多行为发生改变, 如远程会议、线上购物等。随着人们越来越习惯于在“虚拟世界”中进行这些日常活动, 元宇宙的概念重新流行起来。此外, 各种新兴技术的发展使元宇宙实现的可能性越来越大。例如, 增强移动宽带 (enhanced mobile broadband, eMBB) 和超可靠低时延通信 (ultra reliable low latency communications, URLLC)^[6-7] 可以支持AR、VR等技术, 使用户在感官上更沉浸于元宇宙的虚拟世界中。这些新型技术的快速发展使元宇宙在教育^[8]、医疗^[9]、电子商务^[10]、智能产业^[11]以及许多活动中占据优势。通过将更多的数据注入虚拟世界和物理世界中, 并且使用人工智能辅助系统, 为用户创造全新的空间和体验。

在技术层面上, 元宇宙可以被看作大数据与信息技术的融合载体。用户在元宇宙中的信息等以数据的形式被记录在各种形式的文件中。随着用户的增加, 元宇宙中将产生大量的数据, 从而形成一个大数据网络。同时该网络会继续扩大, 人们需要使用智能的数据分析工具来获取有用的信息, 使决策更加可预测和准确, 更有效地指导生产和生活, 因此给数据处理技术^[12]带来了巨大挑战。元宇宙中虚拟世界和物理世界的无缝连接需要众多物联网设备的支持, 这些设备要实时地收集和物理世界中的数据。虚拟世界与现实世界的融合、互动是大数据技术的关键任务。处理大数据的能力对于元宇宙来说非常重要。

值得注意的是, 元宇宙不是简单地将物理世界与虚拟世界结合起来。相反, 它是物理世界在虚拟世界中的延续, 是一个融合了物理世界与虚拟世界的生态系统。

虚拟世界依赖于分布式技术来协调、沟通和完成各项活动。为了解决元宇宙中地理、时间等密切相关但不连续的问题,需要合理协调虚拟世界与物理世界的各项活动,使人们可以在两个世界间有效沟通。尽管有成熟的协作工具,但是仍然无法满足元宇宙因地理分散等原因带来的相关挑战^[13-15]。

目前,元宇宙中的虚拟世界依赖于集中式的架构,每个虚拟世界在被划分为更小的区域之后使用专用服务器进行管理,但是一些计算(如仿真)需要中央服务器执行。边缘计算可以减轻中央服务器的计算压力。边缘计算将云计算扩展到网络边缘,支持资源有限的移动设备将其任务上传到边缘服务器进行处理^[16]。然而,随着元宇宙的出现,连接到通信网络的移动设备数量将急剧增加,这可能会使边缘服务器由于资源有限^[17-18]而拥塞,因此可以采用分布式协同计算解决元宇宙^[9]中的计算密集型任务。云边端协同技术综合考虑了云边端不同层次的特征,可以进一步优化元宇宙中的分布式问题。

元宇宙中虚拟世界和物理世界需要云的大型计算能力和边、端共同协作来完成工作。然而,现有工作并没有从云边端协同的角度关注元宇宙中数据存取等相关挑战。本文从元宇宙的云边端协同的角度出发,提出大数据解决元宇宙数据管理问题的方案。

本文主要考虑以下几个问题。

- 如何跨越通信和网络障碍,在时延和数据速率的要求下,实现虚拟世界与物理世界的数据统一。物理世界与虚拟世界之间存在大量的数据和信息流动,以保证两个世界的同步。因此,综合考虑带宽的限制和要传输的大量数据,笔者希望尽可能让物理世界与虚拟世界具有高度一致性。

- 如何设计数据存取结构,使用户在虚拟世界与物理世界中高效管理数据。元宇宙平台拥有大量的网络用户和扩展现实设备的物理用户,需要用好的数据结构管理产生的大量数据,以便元宇宙平台及所有用户对数据进行操作。

- 同一数据在虚拟世界与物理世界中的语义可能不同,如何定义数据模型以便管理数据。例如,元宇宙中不同世界数据的优先级是不同的,针对不同世界的同一数据需要不同的处理方案,因此需要设计统一的数据模型对数据进行约束,作为查询计划等的理论基础。

- 如何执行数据查询以使用户在两个世界无缝切换,使元宇宙更具有沉浸性。虚拟世界与物理世界的查询任务可能是不相同的,结果的精度也存在差异。因此,需要根据不同查询指定相应的查询计划。同时,元宇宙中的查询更加复杂,如元宇宙游戏中可能涉及移动查询,查询任务会根据用户的地理位置不同而变化。因此执行良好的查询计划可以更好地融合虚拟世界与物理世界。

1 现有工作

Thomason J^[19]研究了元宇宙对协作工作、教育、临床护理、健康和货币化的影响。参考文献[20]提出了一种管理游戏中用户生成数据的解决方案。Yang Q L等人^[21]讨论了AI和区块链的融合如何促进元宇宙中的服务交付。还有一些工作讨论了人工智能如何在开发元宇宙中发挥作用,如参考文献[22]中根据自然语言处理来创建智能聊天机器人或机器视觉,以允许AR/VR设备有效地分析和理解用户环境。Ning H S等人^[23]重点讨论了元宇宙目前的应用和工业发展,并从工业的角度对未来的前景进

行了预测。

现有的“精简版”元宇宙，如在游戏 Second Life 中，用户可以通过客户端与游戏中的虚拟化身互动。在游戏中，虚拟化身的活动与区域有关，而区域与服务器相关联，任何客户端与任何服务器都可以交互。因此，这种模式与云边端协同的管理模式相似。

为了支撑不同场景下的数据管理，研究人员提出了多种数据模型，主要包括关系型数据模型、非关系型数据模型、多模态数据模型3种。对应每类数据模型有不同的研究，包括基础数据结构、基本数据操作及其代数理论、数据完整性约束研究。最经典的数据模型是关系型数据模型^[24]，相关知识包括基础关系数据结构、基础关系数据操作、关系代数和关系数据完整性约束理论。随着不同类型数据的出现，多种非关系型数据模型被提出。其中比较有代表性的是 XML 数据模型^[25]、JSON 数据模型、RDF 数据模型、Property 图模型^[26]等。

当前云边端数据的存储策略主要有集中和分布两种。参考文献[27]提出了一种数据定期从传感器节点收集并发送到集中式数据库的方法，但由于该方法会生成大量数据，可能在服务器中因资源争夺形成瓶颈^[28]。Kanzaki A 等人^[29]提出了基于集中方法的无线传感器网络测试平台 X-Sensor，该平台将从所有节点获取的数据存储在集中数据库中。Elias A G F 等人^[30]提出了基于集中存储的无线传感器网络监测方法，该方法收集传感数据，并将其存储在集中式关系数据库中。在支持云边端的数据库产品中，OpenTSDB^[31]/HBase^[32]仅支持云侧数据存储；InfluxDB^[33]支持云侧和边侧的数据存储；TDengine、Machbase 和 Apache IoTDB 均支持云边端3类数据存储，但均不支持协同处理。

随着不同云存储系统的出现，如 GFS^[34]、Cassandra^[35]、Dynamo^[36]等，研究人员已经提出了多种索引方案来有效地支持大规模分析作业和高并发联机事务处理过程 (on-line transaction processing, OLTP) 查询^[37-39]。为了有效支持复杂查询，多维索引策略被提出。RTCAN 在 CAN 中使用了基于 R-tree 的索引^[40]，Zhang X Y 等人^[41]通过使用 R-tree 和 KD-tree 的组合来支持多维索引，Cheng C L 等人^[42]提出了一种基于多维索引的 CAN，Chen G 等人^[43]在云基础设施中提出了一个类似数据库管理系统 (database management system, DBMS) 的索引。

近年来，随着云计算、边缘计算的兴起，也有一些新兴数据库可以更好地支持云边端协同数据管理。例如时序数据库 InfluxDB^[33]被用来支持云边数据查询；Apache IoTDB 虽然支持云边端上的数据管理，但其仅支持有限的几种基于规则的查询优化技术^[44]，且无法实现元宇宙云边端协同查询处理和优化。

2 元宇宙中云边端协同的挑战

首先，与目前在数据流和传感器上进行的相对简单的聚合不同，元宇宙中的数据管理需要对这些数据源进行更复杂的逻辑推理。其次，与旨在为一组异构数据库派生通用模式的数据集成不同，元宇宙中数据管理基于这些数据发生的事件，并在元宇宙中准确而有效地描述这些事件。显然，需要开发数据管理机制来有效地处理这两个问题。元宇宙中的活动可以被看作云边端设备共同协作的结果。以元宇宙购物为例，用户的终端设备（如手机等）可以被看作端，实体商场可以被看作边，而虚拟的线上商店可以

被看作云。因此，元宇宙中的数据管理可以被看作云边端协同的数据管理。下面从云边端的角度讨论数据管理的优势和挑战。

云边端协同可以充分整合云计算与边缘计算的优势。以在线购物为例，位于用户端侧的设备（如手机）需要处理大量的用户数据及需求，使数据量激增；而位于边侧的商场和位于云侧的线上商店需要管理商品信息、处理用户需求等，这需要准确、快速的计算能力支撑。若全部在端侧或边侧处理，端、边侧设备的计算能力无法支撑，且无法实现全局数据分析；若将端侧的全量数据传至云端处理，则无法满足数据处理的实时性要求，且会严重消耗网络带宽。此时云边端协同的优势显著体现。在端侧对用户数据进行过滤、平滑、降采样等处理，在边侧进行特征提取等操作，在云侧进行关联规则挖掘、深度学习等高代价的分析任务。此时，终端算力上移，云端算力下沉，可有效解决云上集中处理带宽消耗大、响应速度慢的问题，也能有效解决边端处理能力不足的问题，进而为用户提供良好的购物体验。

云边端协同的新计算模式对数据管理提出了新的要求，需要有效融合云边端的计算能力，构筑云边端数据处理、通信、存储等能力全面协同的数据管理平台，这可以带来以下3个方面的优势。

- 强时效。由于云的计算能力更强，而边缘侧和端侧距离数据源更近，云边端协同能够为数据存储和查询、处理任务分配最优计算节点，减少中间数据传输，从而提高数据存取性能，确保实时处理，使虚拟世界与物理世界更好地融合。例如，为了保证元宇宙购物的高时效性，需要在用户端侧监控数据实时处理端侧数据，而大量交易信息交换可以在云侧完成。

- 高安全性。边缘和端上的数据处理

基于本地，涉及用户隐私的敏感数据无须上传到云端，这避免了网络传输带来的泄漏风险。数据受到攻击，只会影响本地数据，不会影响云上数据，从而提高数据安全性，为元宇宙提供安全保障。例如，在军事演习中，如果将所有的战略信息上传到云端，可能会造成军事信息泄露，因此可以采用联邦学习框架，在边端完成面向敏感数据的学习任务并加密上传到云，在云上基于密态数据完成分析，这样可以在不泄露军事机密的前提下完成元宇宙军事演习。

- 低代价。由于数据无须全部上传到云，面向云边端协同的数据管理无须使用太多的网络带宽，充分利用云的协同能力和边端的计算能力，降低本地设备管理数据的成本和能耗，从而提高计算效率。例如，在元宇宙游戏中，用户的扩展现实设备采集的数据不需要全部上传到云端，在用户端进行信息筛选，在云端运行复杂模型进行处理。这既节省了带宽，又避免在边端上执行代价高昂的复杂运算。

尽管面向云边端协同的数据管理具有诸多优势，但是云边端的计算、存储资源状况和对数据管理的需求带来了一些挑战，如图1所示。

- 异构性。元宇宙中云边端协同的异构性主要包括设备异构和数据异构两个方面。一方面，异构设备（如不同的用户XR设备、承载虚拟世界的服务器）之间的计算能力和通信资源存在巨大的差异，导致将同样的模型和方法应用于不同设备会产生不同代价。同时元宇宙各个设备的存储能力和计算能力有巨大差异。云节点存储计算能力强，但查询时延大，且I/O带宽容易成为瓶颈；边缘节点靠近终端，查询时延低，但存储计算能力较弱，通常只保存局部数据，难以执行大规模复杂查询；终端节点存储计算能力差，但其兼顾数据采

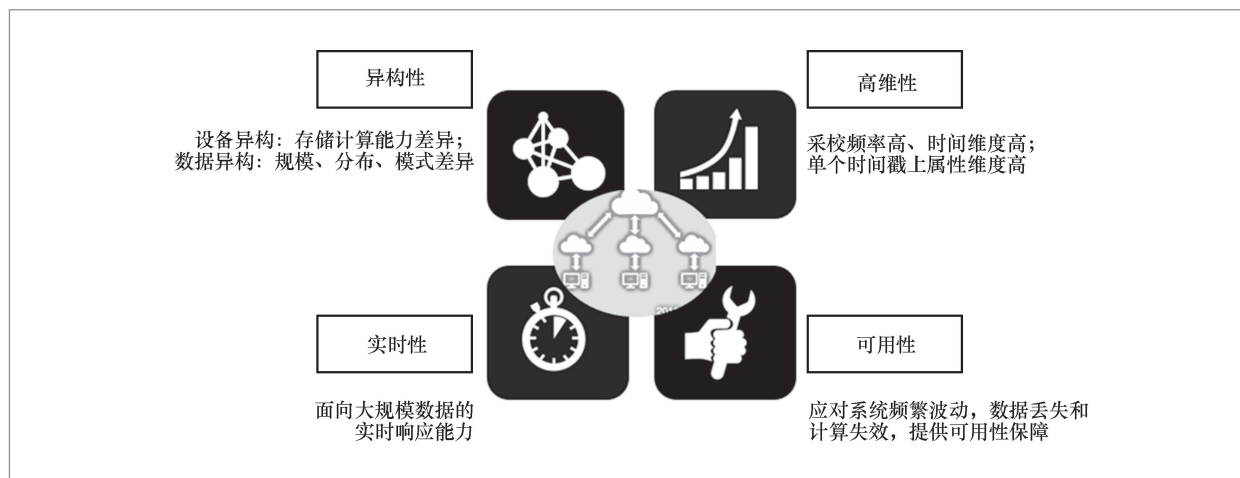


图1 面向元宇宙的云边端协同大数据管理的挑战

集功能, 适用于数据过滤等轻量级操作。因此, 实现异构设备的兼容协调及资源的合理分配与管理为数据管理带来挑战。另一方面, 元宇宙的云边端协同中, 终端设备上产生海量数据, 包括数值型数据、图像数据和音频数据等; 同时, 边缘设备和云有大量结构化数据和非结构化数据。现有数据管理方法难以有效处理设备异构和应用场景差异, 以及其带来的数据规模、分布和模式的差异, 因此这是面向元宇宙的云边端协同数据管理的第一个挑战。

- **高维性。**包括数据在时间维度的高维性和单个时间戳上各个属性维度的高维性两个方面。一方面, 在元宇宙的云边端协同场景下, 物理世界终端设备上产生的传感器监测数据随采样时间有序变化, 是典型的时序数据。由于端设备采样频率高、时间维度高, 时序数据的长度很大, 需要更加快速的数据写入能力和更高效的压缩技术。例如, 在元宇宙游戏中, 用户终端设备上的摄像头需要实时监控用户的行为, 并将相应数据上传到虚拟世界中, 如使用高速摄像头进行高频率的图像

采集, 可达到单点每秒采集100 MB以上的数据。另一方面, 在元宇宙的云边端协同场景下, 端侧设备(如各类传感器以及检测设备)数量巨大, 这使得同一实体在各个采样点上的属性具有超高维度。例如, 元宇宙游戏中同一时间玩家视野下需要采样数百个点才能在虚拟世界中建模。现有的数据索引和查询处理技术难以高效处理这类超高维度的数据, 这成为面向元宇宙的云边端协同数据管理的第二个挑战。

- **实时性。**在云边端协同场景中, 物理世界中海量终端设备以较高的频率采集数据, 使数据规模进一步扩大。而云边端协同的业务需求, 如商场实时购物、游戏中的奖励、基于元宇宙的远程医疗等, 常需要实时响应。因此, 面向元宇宙大规模数据的实时响应能力成为面向元宇宙的云边端协同数据管理的第三个挑战。

- **可用性。**受网络信号、能量损耗和设备故障等因素的影响, 物理世界的设备会频繁地上线或下线, 边云侧也会出现不稳定的情况。例如, 元宇宙购物可能会因网络故障无法交易、云端用户可能发生堵塞

等。因此,元宇宙的云边端协同的显著特点是系统存在频繁波动。现有数据管理理论和方法均未考虑到这一问题,因而,如何应对元宇宙的故障、保证系统可用性成为面向元宇宙的云边端协同数据管理的第四个挑战。

当前针对云边端的数据管理理论和技术研究主要集中在云侧,面向边和端的研究较少,面向云边端协同的数据管理理论和技术的研究更少,难以应对上述挑战,因此很难将其应用在云边端协同场景下,更无法应用在元宇宙场景中。因而,本文讨论面向元宇宙的云边端协同大数据管理技术,重点讨论面向元宇宙的数据统一技术、数据存取技术、数据模型技术以及查询优化技术。

3 面向元宇宙的大数据管理技术

针对上述元宇宙中数据管理的挑战,本文提出面向元宇宙数据管理的4个研究内容,如图2所示。首先,本文提出了面向元宇宙的云边端协同数据模型,将其作为元宇宙数据管理的基础;其次,本文分别提出了面向元宇宙的云边端数据同步管理以及高效存储索引,解决元宇宙的数据同步、存储、索引等问题;最后,本文提出面向元宇宙的云边端协同查询优化,在数据管理的基础上解决元宇宙中的各种复杂查询。

3.1 数据模型

数据模型是数据管理的基础。在面向元宇宙的数据管理中,传统的数据模型存在3个问题。第一,数据和设备异构带来了元宇宙中不同的数据模型需求及元宇宙的数据模型新需求,现有数据模型或者数据

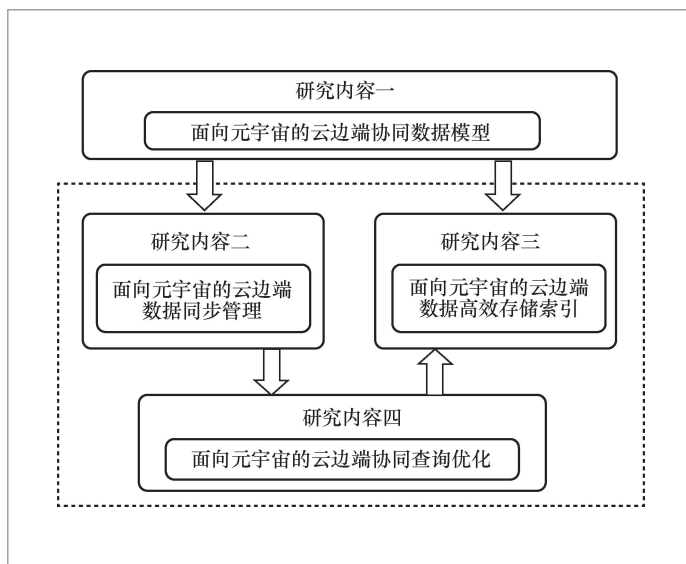


图2 面向元宇宙数据管理的研究内容

模型的简单叠加难以同时满足元宇宙异构数据管理和协同关系管理需求。第二,面向元宇宙的数据库中数据操作更加复杂,难以用传统代数运算来描述,很多查询需要虚拟世界和物理世界协同处理,传统数据模型难以直接支撑复杂的协同数据处理,而多种数据操作的简单组合又会降低数据操作执行的效率。第三,元宇宙不同世界的异构数据结构带来了新的数据完整性约束需求,目前尚未有面向元宇宙的数据完整性约束理论,而现有的完整性约束理论难以支撑元宇宙数据协同中不同层次异构数据转换和数据协同查询所需的完整性判定。

综上,本文针对上述3个问题,提出面向元宇宙的云边端协同数据模型,为数据管理奠定理论基础。重点研究面向元宇宙的异构数据映射模型、异构模型协同数据操作与代数运算,以及多元弹性数据完整性约束。

为了完成虚拟世界与现实世界的交互,元宇宙通常利用多源数据与多台设备进行交互,从而带来了设备和模型异构的

数据管理需求。本文基于云边端平台,针对虚拟数据和现实数据设计异构数据映射模型,定义不同元数据的具体数据模式和数据模式在云边端层内、层间的关联、依赖关系,以有效支撑对虚拟世界与现实世界的高效管理,还能够有效支撑元宇宙中云边端协同的数据管理,为元宇宙管理的数据操作和数据完整性约束理论奠定基础。

其次,元宇宙除了要支撑虚拟数据和现实数据的存取,还要支撑虚拟世界与现实世界的自动链接,传统数据平台的范围过滤、时序过滤、近似过滤以及聚集过滤等操作已经难以满足元宇宙的数据操作需求。因而,要研究元宇宙中异构模型协同数据操作,包括多源虚拟数据模型、多源现实数据模型和异构映射模型的数据操作;定义面向元宇宙中新型数据结构的基本数据操作,并研究底层代数运算,为查询处理与优化奠定基础。

最后面向元宇宙的异构映射数据模型对数据完整性约束理论提出了新的泛化需求,元宇宙数据具有不同的数据约束语义,并且现实数据和虚拟数据之间存在复杂映射约束语义,传统面向关系模型或高维数据模型设计的数据完整性约束理论不能支撑面向元宇宙数据管理的复杂多样语义需求。因此要研究面向异构映射数据模型和数据操作的多元弹性数据完整性约束理论,以支持元宇宙数据的多样化语义约束、一对多依赖语义约束以及约束理论的推理规则,为元宇宙数据存储、索引、查询优化奠定理论基础。

3.2 数据同步管理

元宇宙数据的来源多样、类型不同,且在不同应用的数据处理中,虚拟世界与物理世界对数据存储的要求不同。元宇

宙数据处理要求低时延、高度自治,而传统数据存储技术无法兼顾两个世界。针对元宇宙对数据存储的复杂需求,要设计基于深度学习的虚实结合分级分层数据存储策略,分别为物理世界和虚拟世界数据设计相应的存储架构、存储结构、存储布局等智能存储策略。将物理世界设备的敏捷性与虚拟世界数据的弹性融为一体,需要以最小化存储空间、最大化负载性能为目标设计元宇宙存储弹性自适应算法,实现元宇宙不同世界数据的协同管理。

元宇宙中数据来源复杂,物理世界需要支撑有一定复杂度的数据处理任务,实时管理和低时延响应要求高。而传统的数据存储结构较单一、存储方式较固定,仅能在数据采集阶段确定存储结构,不能在协同处理过程中提供灵活多变的存储结构方案,难以满足面向元宇宙的存传算一体、计算力下沉需求,实现元宇宙中海量数据的高效并发存取。为了支撑元宇宙多源异构数据的存储,满足数据处理能力下沉至边端侧的需求,需设计智能化存储方案,设计基于深度学习的智能存储决策算法,最小化存储代价,确保物理世界和虚拟世界均可存储与数据处理任务最匹配的数据,消除不必要的数据传输开销,实现元宇宙低时延特性。同时,为了实现数据管理过程中存算有机耦合,可以在存储结构中设计智能数据处理方法,实现存储与数据处理相融合的需求,完成数智化存储;针对视频、语音、图像等多模态数据,开发虚实结合的多模态存储引擎;根据负载和数据变化,综合现有数据存储布局的优缺点,提出智能切换数据布局的方法,面向元宇宙传输需求实现存储结构自适应转换,提高元宇宙数据实时并发读写能力。

元宇宙中设备间、虚拟世界和物理世

界间数据传输频繁,需要高效选择需传输的数据,提高数据传输效率。现有方法虽然可以结合索引实现数据过滤等操作,但是索引结构单一,针对传输任务的变化及多样化处理不灵活,限制了任务传输效率。例如,在元宇宙军事演习中,重要数据必须实时同步以保证演习的正常进行。时延可能导致策略失效,进而导致演习失败。针对这个问题,需研究面向元宇宙中通信任务的索引结构,以降低元宇宙中的通信代价,满足实时性要求;同时结合工作负载等特征,对传输数据进行适当聚合和过滤,对不同世界不同设备的传输任务自适应地选择恰当的索引结构,针对传输任务的变化自动对索引结构进行调整,达到数据在元宇宙设备上高效传输、同步的目的。

3.3 数据高效存储索引

元宇宙中的数据量巨大,原有的数据管理方法不再有效,因此提供高效数据管理机制对元宇宙虚拟世界和物理世界互通是非常必要的。

(1) 面向元宇宙的数据存储技术

面向元宇宙的数据管理的基础是元宇宙两个世界一体的数据协同存储。虚拟世界和物理世界数据的来源多样且规模巨大,包含多种传感器采集的实时数据。随着物理世界的计算力逐渐下沉到边端侧,边端侧存储面临存算一体的设计挑战。为了提升元宇宙中数据的交互能力,元宇宙中的数据与计算力同样需要下沉,传统数据存储方法难以满足要求。

面向元宇宙的数据管理需要在两个世界存储热度不同的数据,传统的存储方法没有考虑面向元宇宙存储中不同特征的数据区分方法。物理世界需要对采集的数据或历史数据进行长期的存储,同时,数据

会在协同处理时,由物理世界上传至虚拟世界,物理世界也会在处理过程中提供原始数据支持。例如,元宇宙游戏中,一些仅与玩家有关的数据可以在本地处理,当虚拟世界需要该数据时再进行上传操作。基于此,为了节省存储空间,应对不同节点可用性不同的挑战,需要研究智能数据热度分级管理方案,从而自动区分处理过程中数据的热度,基于热度选择不同位置和模式进行存储,并能够预测出不同方案的存储代价,以加快面向元宇宙的数据处理速度。

物理世界的数据规模整体上非常巨大且数据存在大量冗余,给数据传输和存储带来负担,因此需要面向元宇宙数据的特点对数据进行压缩。为了应对元宇宙海量高维时序数据存储的挑战,最大化存储空间利用率,需要设计适用于元宇宙的数据弹性压缩方法,按照数据分布、负载等特征进行自适应压缩,实现虚拟世界和物理世界数据的智能化协同压缩。针对多模态数据,设计基于卷积神经网络的多模态数据压缩算法,最大限度降低不同世界间的数据传输代价。

(2) 面向元宇宙的索引技术

有效的索引对提升数据处理效率起着重要作用,快速获取数据可以在很大程度上淡化虚拟世界与物理世界的边界,使用户在两个世界中切换自如。

现有索引缺少元宇宙中云边端整体协同及元宇宙中数据不断更新的综合考虑。由于这些索引未能考虑到元宇宙两个世界以及云边端3个层次之间的协同,简单的叠加会导致查询处理过程中数据或索引的冗余访问,进而影响效率,因此需要综合考虑元宇宙协同数据索引技术,设计一体化新型索引。例如,在元宇宙图书馆中,图书库存不断变化,一旦虚拟世界与物理世界的数据不匹配,就会出现购买数

量超过实际库存的情况。因此针对高维时序数据分布复杂、节点数量多、数据量大而目前尚无系统性设计索引的现状,需综合考虑元宇宙中云边端协同的问题,研究面向元宇宙的高维索引,充分考虑高维时序数据特征和元宇宙中云边端协同的综合特征,以处理效率最大化为目标设计元宇宙索引分布策略。为虚拟世界和物理世界数据的映射设计索引,支持元宇宙中云边端各个节点的高效查找,实现元宇宙高效协同。考虑到元宇宙中数据的不稳定更新,需要研究支持快速更新的索引结构,解决由数据更新过快引发的索引不同步问题。

由于元宇宙设备的异构性,不同世界之间的数据难以统一索引且同一个世界的索引很难扩展到其他世界,尚无有效方法根据不同设备处理任务的能力自适应地生成或选择相应的索引。针对元宇宙不同设备处理能力和承担的計算任务不同、现有索引结构单一且难以扩展的挑战,需要研究以提升元宇宙数据访问整体任务的效率为目的的自适应索引,降低空间复杂度、加速查询处理,保证面向元宇宙不同设备、数据和负载特征的索引智能构建和更新。

3.4 查询优化

面向元宇宙的查询处理和优化带来了数据海量异构、协作约束复杂、优化目标多样、设备网络异构和节点稳定性低5个方面新的挑战,当前的技术难以应对这些挑战,因此元宇宙中的查询处理和优化需要新的机制。针对上述挑战,本文提出4个查询优化技术路线。

首先元宇宙中存在着海量异构数据,现有数据库查询技术无法对这些复杂异构数据做出有效管理。例如,在元宇宙购物、

图书馆和游戏等场景下,元宇宙环境中存在着海量的三维模型数据、文本数据和键值数据等。面对元宇宙中用户的复杂交互行为,需要针对其海量异构的数据特点,研究面向元宇宙的逻辑查询计划表示模型,并设计支持异构数据交叉查询的查询执行引擎。

元宇宙异构节点间普遍存在着多种协作形式,现有数据库的查询处理和优化方法难以适应这种分层架构下复杂的协作和异构节点中多样的约束条件。例如,在元宇宙购物场景中,商品需要显示高精度的三维模型和详细的文字描述信息。为了减小客户端侧的交互时延,提升元宇宙的沉浸性,需要利用边侧设备对用户常用或元宇宙空间中用户周围的商品进行缓存以加速查询。不同用户使用的边端设备的计算能力、存储能力千差万别。为了有效利用元宇宙不同层级设备的处理能力,打通元宇宙异构设备的屏障,实现异构数据操作的协同处理,需要研究面向元宇宙数据管理中一系列基本数据操作的高效协同算法,在基本数据操作层面实现对元宇宙分层异构节点的自适应,作为面向元宇宙查询处理的基础。

在元宇宙中,存在不同优化目标的多类查询,如连续、聚集和复杂查询等,当前尚未有面向多样优化目标的协同查询处理和优化方法。例如,远程手术协助等对查询时延要求极高的场景要求数据库能够快速查询、传递场景中的信息,其中的生理指标监控、器械信息和病历查询等涉及连续查询和复杂聚集查询。同时,医院内的边端设备还受到能源效率和计算效率的约束,其各自有不同的优化目标,这进一步为元宇宙场景下云边的协同查询优化带来了困难。为了适应元宇宙中复杂多样的查询类型、极高的时延要求和异构的优化目标,需要研究面向元宇宙的物理查

询计划模型,为协同查询优化奠定基础;并针对不同查询类型、协作形式的要求与约束,面向元宇宙异构设备与协议研究协同查询优化算法。以总体效率最大化、云边负载平衡、端侧能效最优为目标,设计协同查询重写、云边操作下推、端侧设备参数动态自适应等方法,可以实现各类查询在元宇宙上的协同优化,在查询效率最大化的同时,满足各级异构设备的约束需求。

最后,元宇宙中节点的安全性、稳定性按虚拟世界、物理世界、云边端顺序递减。在实际生产环境中,用户终端节点可能随时因网络不稳、设备故障等原因断线,目前尚无针对此特性的鲁棒查询处理方法。面向元宇宙中不同层次的节点异构分层且安全性和稳定性存在较大差异的挑战,为了协调元宇宙各节点上的动态查询执行,并有效应对终端设备出现断线、重连以及系统节点出现安全性问题等异常情况,需要设计面向元宇宙的高效查询执行方法,最小化因节点或传输异常而导致的异常终止、结果错误等问题,达到提高查询处理效率、鲁棒性和稳定性的目的。

4 结束语

随着新兴技术的快速发展,在过去的几年中研究人员越来越关注元宇宙,并且基于元宇宙进行了大量的研究。本文针对元宇宙中的数据管理技术进行了分析。已有的成果缺乏对元宇宙中大数据管理技术的阐述。因此,本文首先回顾了关于元宇宙及其应用的文献,总结了元宇宙的应用现状;然后,本文针对云边端协同的元宇宙数据管理提出了未来的研究方向。希望本文的讨论,能够对元宇

宙与数据管理之间的关系进行详细的解释,并为今后的研究提供一些有益的研究方向。

参考文献:

- [1] WIEDERHOLD B K. Ready (or not) player one: initial musings on the metaverse[J]. *Cyberpsychology, Behavior and Social Networking*, 2022, 25(1): 1-2.
- [2] RATCLIFFE J, SOAVE F, BRYAN-KINNS N, et al. Extended reality (XR) remote research: a survey of drawbacks and opportunities[C]//*Proceedings of 2021 CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 2021: 1-13.
- [3] XI N N, CHEN J, GAMA F, et al. The challenges of entering the metaverse: an experiment on the effect of extended reality on workload[J]. *Information Systems Frontiers: A Journal of Research and Innovation*, 2022: Accepted.
- [4] HWANG G J, CHIEN S Y. Definition, roles, and potential research issues of the metaverse in education: an artificial intelligence perspective[J]. *Computers and Education: Artificial Intelligence*, 2022, 3: 100082.
- [5] MOHANTA B K, JENA D, PANDA S S, et al. Blockchain technology: a survey on applications and security privacy Challenges[J]. *Internet of Things*, 2019, 8: 100107.
- [6] YOU X H, WANG C X, HUANG J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts[J]. *Science China Information Sciences*, 2020, 64(1): 1-74.
- [7] TANG F X, CHEN X H, ZHAO M, et al.

- The roadmap of communication and networking in 6G for the metaverse[J]. *IEEE Wireless Communications*, 2022(99): 1–15.
- [8] NG W C, LIM W Y B, NG J S, et al. Optimal stochastic coded computation offloading in unmanned aerial vehicles network[C]//*Proceedings of 2021 IEEE Global Communications Conference*. Piscataway: IEEE Press, 2021: 1–6.
- [9] YU X D, OWENS D, KHAZANCHI D. Building socioemotional environments in metaverses for virtual teams in healthcare: a conceptual exploration[C]//*Health Information Science*. [S.l.:s.n.], 2012: 4–12.
- [10] JEONG H, YI Y, KIM D. An innovative e-commerce platform incorporating metaverse to live commerce[J]. *International Journal of Innovative Computing, Information and Control*, 2022, 18(1): 221–229.
- [11] KWON C. Smart city-based metaverse a study on the solution of urban problems[J]. *Journal of the Chosun Natural Science*, 2021, 14(1): 21–26.
- [12] OOI B C, TAN K L, TUNG A, et al. Sense the physical, walkthrough the virtual, manage the metaverse: a data-centric perspective[J]. *arXiv preprint*, 2022, arXiv:2206.10326.
- [13] COUSINS K C, ROBEY D. Human agency in a wireless world: patterns of technology use in nomadic computing environments[J]. *Information and Organization*, 2005, 15(2): 151–180.
- [14] JIN L, ROBEY D. Bridging social and technical interfaces in organizations: an interpretive analysis of time-space distancing[J]. *Information and Organization*, 2008, 18(3): 177–204.
- [15] PINSONNEAULT A, CAYA O. Virtual teams[J]. *International Journal of e-Collaboration*, 2005, 1(3): 1–16.
- [16] QIAN L P, WU Y, YU N N, et al. Learning driven NOMA assisted vehicular edge computing via underlay spectrum sharing[J]. *IEEE Transactions on Vehicular Technology*, 2021, 70(1): 977–992.
- [17] GOUDARZI M, WU H M, PALANISWAMI M, et al. An application placement technique for concurrent IoT applications in edge and fog computing environments[J]. *IEEE Transactions on Mobile Computing*, 2021, 20(4): 1298–1311.
- [18] DAI M J, ZHENG Z Y, ZHANG S L, et al. SAZD: a low computational load coded distributed computing framework for IoT systems[J]. *IEEE Internet of Things Journal*, 2020, 7(4): 3640–3649.
- [19] THOMASON J. Metahealth—how will the metaverse change health care? [J]. *Journal of Metaverse*, 2021, 1(1): 13–16.
- [20] KASAPAKIS V, GAVALAS D. User-generated content in pervasive games[J]. *Computers in Entertainment*, 2017, 16(1): 1–23.
- [21] YANG Q L, ZHAO Y T, HUANG H W, et al. Fusing blockchain and AI with metaverse: a survey[J]. *IEEE Open Journal of the Computer Society*, 2022, 3: 122–136.
- [22] HUYNH-THE T, PHAM Q V, PHAM X Q, et al. Artificial intelligence for the metaverse: a survey[J]. *Engineering Applications of Artificial Intelligence*, 2023, 117: 105581.
- [23] NING H S, WANG H, LIN Y J, et al. A survey on metaverse: the state-of-the-art, technologies, applications, and challenges[J]. *arXiv preprint*, 2021, arXiv: 2111.09673.
- [24] DATE C J. *A guide to the SQL*

- standard(Second Edition)[M]. [S.l.]: Addison-Wesley, 1989.
- [25] DEUTSCH A, SUI L, VIANU V. XML path language (xpath) version 1.0. w3c recommendation, the world wide web consortium[J]. Journal of Computer and System Sciences, 1999, 73 (3): 442-474.
- [26] REST O V, HONG S, KIM J, et al. PGQL: a property graph query language[C]// Proceedings of the 4th International Workshop on Graph Data Management Experiences and Systems. New York: ACM Press, 2016.
- [27] BONNET P, GEHRKE J, SESHADRI P. Towards sensor database systems[C]// Proceedings of the 2nd International Conference on Mobile Data Management. New York: ACM Press, 2001: 3-14.
- [28] DIALLO O, RODRIGUES J J P C, SENE M. Real-time data management on wireless sensor networks: a survey[J]. Journal of Network and Computer Applications, 2012, 35(3): 1013-1021.
- [29] KANZAKI A, HARA T, ISHI Y, et al. X-Sensor: wireless sensor network testbed integrating multiple networks[M]//Wireless sensor network technologies for the Information explosion era. Heidelberg: Springer, 2010: 249-271.
- [30] ELIAS A G F, RODRIGUES J J P C, OLIVEIRA L M L, et al. A ubiquitous model for wireless sensor networks monitoring[C]//Proceedings of 2012 6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. Piscataway: IEEE Press, 2012: 835-839.
- [31] PRASAD S, AVINASH S B. Smart meter data analytics using OpenTSDB and Hadoop[C]//Proceedings of 2013 IEEE Innovative Smart Grid Technologies-Asia. Piscataway: IEEE Press, 2013: 1-6.
- [32] QIU Y J, ZHU X H, LU J. Fitness monitoring system based on Internet of Things and big data analysis[J]. IEEE Access, 2021, 9: 8054-8068.
- [33] BEERMANN T, ALEKSEEV A, BABERIS D, et al. Implementation of ATLAS Distributed Computing monitoring dashboards using InfluxDB and Grafana[J]. EPJ Web of Conferences, 2020, 245: 03031.
- [34] GHEMAWAT S, GOBIOFF H, LEUNG S T. The Google file system[J]. ACM SIGOPS Operating Systems Review, 2003, 37(5): 29-43.
- [35] LAKSHMAN A, MALIK P. Cassandra: a decentralized structured storage system[J]. ACM SIGOPS Operating Systems Review, 2010, 44(2): 35-40.
- [36] DECANDIA G, HASTORUN D, JAMPANI M, et al. Dynamo: amazon's highly available key-value store[J]. ACM SIGOPS Operating Systems Review, 2007, 41(6): 205-220.
- [37] WU S, WU K L. An indexing framework for efficient retrieval on the cloud[J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2009, 32(1): 75-82.
- [38] WU S, JIANG D W, OOI B C, et al. Efficient B-tree based indexing for cloud data processing[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1207-1218.
- [39] WANG J B, WU S, GAO H, et al. Indexing multi-dimensional data in a cloud system[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2010: 591-602.
- [40] NGUYEN T D T, YONG C H, PHAM X Q, et al. An index scheme for similarity search on cloud computing using MapReduce over Docker container[C]//

- Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. New York: ACM Press, 2016: 1-6.
- [41] ZHANG X Y, AI J, WANG Z Y, et al. An efficient multi-dimensional index for cloud data management[C]//Proceedings of the 1st International Workshop on Cloud Data Management. New York: ACM Press, 2009: 17-24.
- [42] CHENG C L, SUN C J, XU X L, et al. A multi-dimensional index structure based on improved VA-file and CAN in the cloud[J]. International Journal of Automation and Computing, 2014, 11(1): 109-117.
- [43] CHEN G, VO H T, WU S, et al. A framework for supporting DBMS-like indexes in the cloud[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 702-713.
- [44] TAN K L. Query optimization in sensor networks[M]//Encyclopedia of database systems (2nd ed.). [S.l.:s.n.], 2018.

作者简介



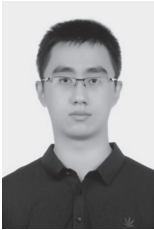
朱锐 (1998-), 女, 哈尔滨工业大学计算机科学与技术学院博士生, 发表EI检索论文1篇, 主要研究方向为AI4DB、大数据、数据库索引。



王宏志 (1978-), 男, 哈尔滨工业大学计算学部长聘教授、博士生导师, 计算机科学与工程系主任, 海量数据计算研究中心主任, 数据科学与大数据技术专业负责人, 美国加州大学欧文分校博士后, 微软亚洲研究院铸星计划访问学者, 青年龙江学者。主要研究方向为大数据治理、大数据管理与分析、数据库系统、工业大数据等。在国内外重要会议和期刊发表学术论文300余篇, 出版学术专著4本。论文被SCI收录百余次, 他引4 000余次, 授权发明专利32项。获微软学者、中国优秀数据库工程师、IBM博士英才等称号, 获黑龙江省自然科学奖一等奖和教育部高等学校科技进步奖一等奖, 获黑龙江省青年科技奖、宝钢优秀教师奖、CSC-IBM奖教金。主持国家自然科学基金重点项目、国家支撑计划课题、国家博士后特别资助项目等10余个项目, 主讲的课程获批国家一流线上课程。中国计算机学会杰出会员, ACM SIGMOD中国秘书长、教育部高等学校计算机类专业教学指导委员会计算机系统专家委员会委员、中国计算机学会数据库专业委员会常务委员、中国计算机学会大数据专家委员会委员、哈尔滨工业大学计算学部校友会秘书长、ACM数据科学学科标准编写组专家、黑龙江省“头雁团队”成员。



崔双双 (1997-), 女, 哈尔滨工业大学计算机科学与技术学院博士生, 发表EI检索论文两篇, 主要研究方向为数据库存储、AI4DB。



张恺欣(1999-),男,哈尔滨工业大学计算机科学与技术学院博士生,在TKDE期刊发表过论文1篇,公示专利两篇,主要研究方向为数据库系统、AI4DB。



燕钰(1996-),女,哈尔滨工业大学计算机科学与技术学院博士生,发表SCI检索论文1篇、EI检索论文3篇,主要研究方向为AI4DB、大数据。

收稿日期: 2022-12-08

通信作者: 王宏志, wangzh@hit.du.cn

基金项目: 国家自然科学基金资助项目(No. 62232005)

Foundation Item: The National Natural Science Foundation of China(No. 62232005)