arXiv:2302.01713v4 [cs.AI] 6 Jun 2024

迈向避免数据混乱：来自数据网格实施的行业洞察

JAN BODE*，IBM，德国NIKLAS KÜHL，KIT，德国DOMINIK KREUZBERGER*，IBM，德国
SEBASTIAN HIRSCHL*，IBM，德国。

随着数据和人工智能的重要性日益提升，组织致力于成为数据驱动型组织。然而，当前的数据架构未必能够跟上数据规模和数据与分析用例范围的增长。事实上，现有架构往往无法兑现与之相关的预期价值。数据网格是一种用于企业数据管理的社会技术型、去中心化、分布式概念。由于数据网格概念仍属新兴事物，业界尚缺乏实证洞察。具体而言，人们尚不了解引入数据网格的动机因素、相关挑战、实施策略、业务影响以及潜在原型。为弥补这一空白，我们与行业专家进行了15次半结构化访谈。我们的结果显示，除其他发现外，组织在向与数据网格概念相关的联邦式数据治理转型、转移数据产品开发、供给和维护的责任，以及理解整体概念方面面临困难。在本研究中，我们归纳出多种实施策略，并建议组织设立跨领域指导单元、关注数据产品的使用情况、在早期阶段取得快速见效成果，并优先考虑规模较小、专注于数据产品的专职团队。我们承认，组织需要根据自身需求采用相应的实施策略；此外，我们还归纳出两种原型，可提供更为具体的建议。我们的发现汇总了行业专家的洞察，为研究人员和专业人员提供了成功采用数据网格的初步指南。

1 引言

随着数据量持续增长，组织正努力变得更加数据驱动，以在竞争中脱颖而出 [14]。国际数据公司（IDC）预测，2022—2026年间，数据量将增长一倍以上，其中私营组织将引领这一增长 [32]。

然而，在快速演变的数据管理领域，基于数据仓库和数据湖的传统集中式数据架构的局限性正日益显现。这些系统难以跟上数据量和种类不断增加的步伐，给中央 IT 部门带来了重大挑战 [10, 11, 24]。

以日益去中心化的方式产生、且数量不断增加的数据，给这些部门的能力带来压力，并导致数据请求的响应时间延长 [39]。这种延迟是一个关键瓶颈，影响数据消费者访问相关数据的可访问性，并降低以数据为中心的环境中组织的整体敏捷性和响应能力。

更复杂的是，数据种类日益增多，又带来了新的复杂性。中央 IT 部门通常缺乏有效管理这些数据所需的特定领域知识。这种领域专属知识的缺失，使数据无法得到准确、高效的处理，导致数据提供与不同组织单元的实际需求不匹配 [7]。

*本文不代表 IBM 的官方声明。[0]版本 1.0，截至 31/01/2024

作者地址 例如：Jan Bode，jan.bode1@ibm.com，IBM，德国；Niklas Kühl，kuehl@kit.edu，KIT，德国；Dominik Kreuzberger，dominik kreuzberger@ibm .com，IBM，德国；Sebastian Hirschl，sebastian.hirschl@de.ibm.com，IBM，德国。

此外，在更广泛的数据治理领域中，集中式的数据管理方法引发了人们对数据所有权的重大担忧 [11]。如果缺乏明确的数据所有权，数据质量和维护工作的责任就会变得模糊，进而可能导致数据完整性和质量方面的问题。

这些挑战共同阻碍了对高质量数据的无缝访问，削弱了数据完整性，并延长了数据驱动型倡议的上市时间和价值实现周期 [7]。因此，数据和人工智能（AI）应用的潜在规模与成效受到限制，阻碍了组织发展成为完全数据驱动型实体。

为了解决集中式数据管理方法的不足，Dehghani [10] 提出了数据网格这一全新的社会技术数据概念。该概念通过将数据所有权分散至特定领域，直接应对中央 IT 部门过载造成的低效问题，并借助领域专属洞见提升数据管理质量。它提出了一种将数据视为产品的范式，强调数据的质量、可用性和生命周期，从而解决有关数据完整性和所有权清晰度的顾虑。赋予用户自助式平台，可大幅缩短数据访问的延迟，旨在提升组织敏捷性。此外，联邦式数据治理确保维护数据标准，同时不扼杀创新，并与各领域的独特需求保持一致。这种原则上的战略协同旨在缓解当前的数据管理挑战，并推动数据民主化[1]。

尽管 Dehghani 的 [11] 工作为去中心化数据架构奠定了基础，但关于这一主题的实证研究仍然稀少 [3, 24, 39, 参见]。近期出版物探讨了技术型数据网格架构[3, 7],隐私挑战 [29],整体概念 [24, 30], ，或介绍了单个案例研究 [20, 39]。

然而，尚无科学研究通过实证方式考察行业专家如何在多个行业中应用数据网格这一高层次概念。这是个问题，因为企业数据管理的重大变革需要付出巨大的组织努力。了解其他组织在实施过程中的经验——识别他们遇到的挑战以及采取的策略——对于避免代价高昂的陷阱、防止数据混乱至关重要。此外，探究推动采用数据网格的动机因素也十分关键，因为这有助于组织判断数据网格是否可行、是否适合自身的具体需求。考察采用数据网格所产生的整体影响同样不可或缺，因为这有助于了解此类转变带来的更广泛影响。此外，运用原型可以提供更具针对性的建议，同时认识到不同类型的组织有着各自不同的需求和挑战。因此，我们提出以下研究问题：

研究问题：沿着动机因素、典型挑战、实施策略、现实影响和由此形成的原型这些类别，我们可以为数据网格采用得出哪些初步见解？

为回答这一问题，我们首先开展探索性文献综述，随后在多个行业进行一系列半结构化专家访谈。本研究为正在采用或计划采用数据网格这一架构性数据治理范式的专业人员提供了有益的初步见解。我们强调，本研究的贡献属于理论和概念层面，主要探讨与数据网格的组织与社会维度相关的方面。关于技术实施，请参阅公开可用的指南 [34, 参见]以及近期出版物 [3, 参见]。

本文其余部分的结构如下。第2节介绍理论基础并讨论相关研究。随后，第3节介绍方法论，而第4节则呈现对

[1]数据民主化是指组织内所有员工都能普遍访问数据 [5]。

专家访谈的分析结果。我们将在第5节重点阐述本文的贡献，并指出未来工作方向和局限性。最后，我们将在第6节总结本文。

2 基础与相关工作

为了在这一领域保持竞争力，组织必须不断重新思考并采用其数据战略、架构和管理系统，以便从日益增长的数据量中创造价值 [33]。过去，围绕相关概念出现了各种术语，包括但不限于“数据仓库”“数据湖”，以及近来出现的“数据湖仓”“数据网格”和“数据织物”。在本基础章节中，我们将阐明这些术语、其核心概念及相互联系。

一般而言，数据仓库和数据湖侧重于数据管理，而数据湖仓、数据网格和数据织物则广泛指代数据架构 [16]。数据管理系统与架构的抽象层级各不相同。例如，数据架构可能包含并协调多个数据管理系统 [26]。

数据仓库是特定类型的数据库，包含来自多个来源的结构化数据，主要用作已处理数据的集中式存储 [38]。传统上，它们存储数据以用于商业智能和报告，因此不会存储数据供未来探索 [19]。相比之下，数据湖能够以更快的速度摄取数据，存储更大的数据量以及不同类型的数据 [9]。与数据仓库不同，数据湖还会存储原始数据，供未来探索和可能的业务活动使用。因此，它们对机器学习（ML）应用至关重要。近期出现的数据湖仓架构将数据湖的灵活存储与数据仓库的分析结构相结合，为管理和分析多种数据类型提供了可扩展解决方案。这种混合模式增强了数据可访问性和分析能力，以满足大数据管理不断变化的需求 [17]。根据相关文献，我们总结并定义以下三个术语：

定义 1：数据仓库。数据仓库是支持管理层决策的、面向主题的、集成的、非易失的、随时间变化的数据集合 [19]。

定义 2：数据湖。数据湖是一种中央存储库系统，用于以原始格式存储、处理和分析非结构化、半结构化或结构化的原始数据 [9, 基于]。

定义 3：数据湖仓。数据湖仓是一种基于低成本且可直接访问的存储（用于存储结构化、半结构化和非结构化数据）的数据管理系统，进一步提供传统分析型数据库管理系统的管理功能和性能特性 [4, 基于]。

一个相关但不同的术语是数据织物。数据织物是一种技术架构，将跨越多个数据源的异构数据汇集在一起；它允许组织无论数据位于何处，都能在考虑适当的数据治理和数据编目的情况下监控和管理数据 [22]。因此，其主要重点在于集成多个数据管理系统，包括数据仓库、数据湖或数据湖仓。为了向整个组织中的用户提供访问，数据织物使用丰富的元数据 [36] 和数据虚拟化层 [22]。元数据的使用不仅对于访问、发现和理解数据至关重要，对于自动化数据集成、工程和数据治理活动也同样重要。这包括对数据访问、隐私和合规相关主题进行集中管理。

定义 4：数据织物。数据织物是一种新兴的数据管理设计，旨在实现灵活、可复用且增强型的数据集成管道、服务和语义 [15]。

相比之下，数据网格是一种社会技术概念，包含架构方面。它还纳入了去中心化和所有权等社会和组织方面。与数据织物类似，数据网格通常由多个数据管理系统组成，并辅以集成与治理层，同时配合去中心化的

组织结构。根据参考文献 [11]，数据网格由四项主要原则构成，使组织能够大规模管理数据。第一，面向域的去中心化数据所有权：各个领域拥有自己产生的数据，并利用其领域知识提升数据质量。我们将领域定义为组织特有的划分，用以界定组织相关的竞争边界 [35]。因此，领域知识是指某人通过工作经验、教育或培训，在某个特定领域或范围内获得的专业知识。例如，生产部门作为一个领域，拥有所有与生产相关的数据，因为他们具备最丰富的专业知识，能够理解数据所反映的复杂技术关系。第二，数据即产品：数据被视为产品，并由相关方承担端到端责任。数据产品[2] 包含元数据、可访问性选项（如 API）以及实际数据。它相当于软件产品，也需要额外的服务，例如安全更新或手册。此外，数据产品应具备以下易用性特征：可发现、可寻址、易于理解、可信、可访问、可互操作、有价值且安全 [11]。第三项原则，自助式数据平台，指的是一个专用的数据平台，为各领域提供高层抽象基础设施，使其能够高度自主地开展工作。这对于各领域至关重要，可避免重复投入技术工作，转而专注于创建高质量数据产品。第四项也是最后一项原则：联邦式数据治理 定义了数据产品的治理结构。领域数据产品负责人和相关利益相关者共同决定通用标准和政策，并在各领域内自动执行，以确保数据产品的互操作性。这一点至关重要，因为数据产品在组合使用时才能创造最大的价值。综合运用这四项原则，组织能够克服集中式数据架构的局限，并推动自身成为数据驱动型组织 [11]。图 1展示了由此形成的架构概念。

采用数据网格涉及三个关键阶段：探索与启动、扩展与规模化和提取与持续 [11]。在初期，选定的领域同时充当数据提供方和使用方，建立基础实践，并将数据整合到相互匹配的产品中。随着数据网格在扩展与规模化阶段不断发展，越来越多的领域加入其中，通过标准化技术和组织模式，实现快速扩展并整合遗留系统。最后，在提取与持续阶段，各领域实现自主的数据所有权，专注于优化和完善数据产品的交付与使用情况，最终形成成熟、协调的数据生态系统。每个阶段都以前一阶段为基础，提升组织内部的可扩展性和集成能力。

我们想强调，数据网格具有前文所述的架构方面，但其核心是一种社会技术概念。因此，我们根据参考文献 [11], 对数据网格作如下定义。

定义 5：数据网格。数据网格是一种用于企业数据管理的社会技术、去中心化、分布式概念。

鉴于目前人们普遍混淆数据网格与数据织物的区别， [30,参见], ，我们将在下文重点介绍二者的相似之处与差异。

数据网格和数据织物在数据可访问性和互操作性方面有着共同的目标，最终旨在赋能数据驱动型组织。二者都致力于打破数据孤岛，让数据在整个组织内顺畅流动，并借助自助访问，确保有需要的人能够获取数据。此外，数据网格和数据织物都非常重视可扩展性，以适应不断增长的数据量。

[2]供给 数据产品包括初始开发、使数据可供访问以及维护数据产品。下文中，我们仅 y
指供给。

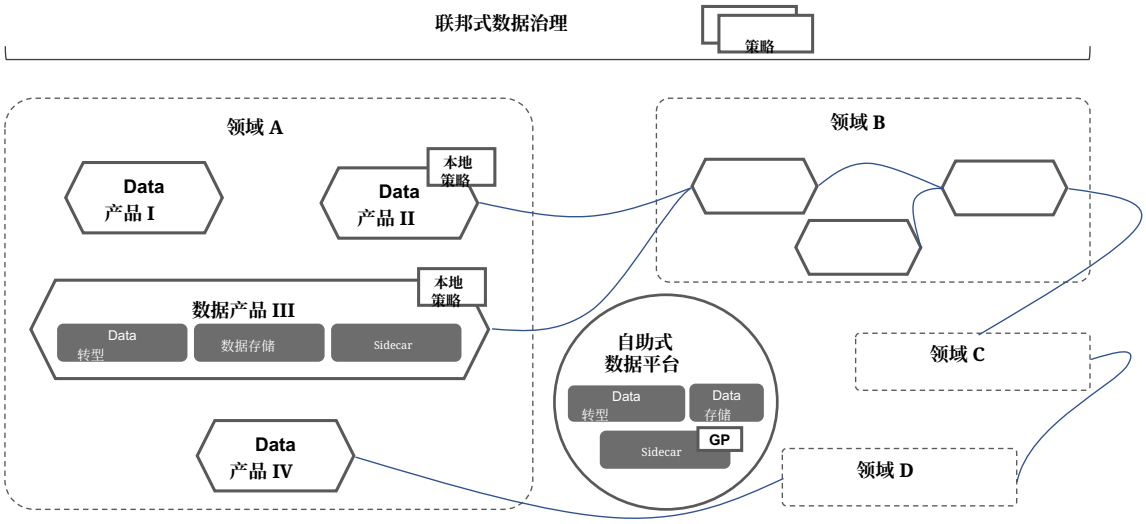

图1. 基于四项关键原则的数据网格概念概览：1）面向域的去中心化数据所有权，2）数据即产品，3）自助式数据平台，4）联邦式数据治理。图中展示了不同的粒度层级（左侧较高，右侧较低）。

然而，二者在组织方法上存在很大差异。数据网格倡导在去中心化组织中实现数据所有权的去中心化，而数据织物并未明确界定所有权这一概念。因此，数据网格涵盖了更多社会和组织方面的内容，而数据织物则着重关注技术解决方案。

总之，数据仓库为存储数据和支持商业智能应用提供了解决方案，而数据湖则充当数据存储库，用于存储数据并支持机器学习和数据科学项目。数据湖仓融合了这些功能，既提供存储，也提供先进的分析能力，涵盖商业智能、机器学习和数据科学。

数据织物和数据网格通常都涵盖并编排多个数据管理系统，重点关注集成、治理和可访问性。此外，数据网格还涵盖组织和社会层面，而数据织物则更侧重技术。

基于这些基础，我们回顾了近期有关数据网格及其相关概念的文献。参考文献 [10] 首次于2019年5月提出了数据网格一词。此后，许多组织分享了关于数据网格领域的知识、观点和工作经验 [37, 40, 参见]。然而，关于“数据网格”这一新兴主题、能够为知识库作出贡献的同行评审出版物仍然寥寥无几。近期的同行评审出版物阐释了数据网格一词 [24, 30] ，并介绍了多种数据网格采用案例。这些描述缺乏细节，仅旨在解释或阐释这一概念。近期的两项案例研究 [20, 39] 深入介绍了挪威劳动与福利管理局 和盛宝银行 采用数据网格的情况。两项研究都贡献了宝贵知识，但均局限于单一案例研究。

总而言之，关于数据网格这一主题已有相关研究，但就当前聚焦于动机因素、挑战和实施策略的研究问题而言，仍缺乏来自现实行业的洞察。因此，我们

选择对该领域的专家开展定性访谈研究，以此作为实证研究的方法，探索这一新兴领域。

3 方法论

为全面了解数据网格采用的动机因素、挑战和实施策略，我们于2022年11月至2023年1月开展了15次半结构化专家访谈。选择半结构化访谈，是因为这种方法能够在封闭式问题的结构性与开放式问题的灵活性之间取得平衡。在探讨数据网格这类复杂且新颖的主题时，这种平衡至关重要，同时也能让访谈过程中涌现出新的想法和主题 [1]。遵循参考文献 [18], 中的方法，我们使用访谈指南来组织访谈内容，围绕概述的主题展开。指南最初根据研究问题的核心主题制定。通过试点测试，我们评估了这一方法的有效性，并据此完善指南。在整个访谈过程中，我们进一步调整指南，将受访者提出的主题纳入其中，以确保全面探讨研究对象。访谈指南的主要版本在第四次访谈后确定。我们采用目的性抽样方法 [12] ，从不同行业中遴选访谈对象，以纳入多种观点和应用场景，从而全面探讨数据网格的特性。具体而言，我们通过专家抽样 [12] ，根据专家在领英上的职位和活动（尤其是领英帖子和评论）来识别专家。此外，我们还通过领英直接联系在公开的数据网格成功案例中发挥关键作用的关键利益相关者。

要符合访谈过程的纳入条件，候选人必须在数据网格领域至少拥有一年工作经验，并在数据与人工智能领域拥有五年经验。有一位受访者在数据网格领域仅有六个月工作经验；不过，由于该参与者在密切相关的分布式数据架构领域拥有丰富背景，因此仍被纳入。虽然一年的工作经验要求看似不高，但对于新兴的数据网格领域而言，这代表着相当丰富的专业知识，而这一领域直到2019年才被提出。

我们的研究涵盖了不同规模、对该主题的经验水平各异的公司。表1概述了受访者及其特征。此外，图2概述了访谈中讨论的主要概念。

由于新冠疫情，所有访谈均通过视频通话进行，并在征得同意后录制。此外，转为数字形式使我们能够接触到分布在不同地点的更多专家，简化录制流程，并在安排时间方面提供了更大的灵活性，以配合专家的日程。

访谈结束后，对录音进行了转录，并使用开放式编码、主轴编码和选择性编码进行定性内容分析 [13, 21, 23, 41] ，并借助 MAXQDA2020 [31]。

每次访谈都先通过开放式编码进行初步的独立分析，随后再结合主轴编码，以归纳内容并从不同访谈中推导结论。最后，使用选择性编码将主轴编码归入各个主题。此过程重复了 15 次，每次迭代对应一次访谈。这种方法有助于提取关键见解，并将其纳入更广泛的框架 [41]。

更具体地说，在每次迭代中，我们首先对自动生成的转录文本进行全面审查，以确保内容的准确性。随后，我们对文本片段进行释义和压缩，以提供

[3]一位参与者 表示自己有五年经验；其给出的理由是，在“数据网格”一词引入之前，曾有该领域的工作经验 .



图2. 访谈中讨论的主要概念词云图。图中排除了停用词，并对词语进行了词形还原。字体越大的概念，表示其在访谈中的相关性越高，相关性依据线性缩放计算。作为参考，*mesh* 出现了570次，而*ml* 出现了51次。由于术语*data* 出现频率高达2447次，会使其他概念的可视化效果产生偏差，因此将其排除。

表1. 受访者概况，包括其首字母缩略词、职位、工作经验、行业、公司规模以及访谈时长。

Name	职位	工作经验 数据与人工智能（年）	工作经验 数据网格（年）	行业	公司 规模（员工人数）	时长 （分:秒）
A	解决方案架构师	9	2	运动服饰	57.000	58:52
B	数据与分析总监	15	2	电子商务	17.000	23:11
Γ	高级顾问	5	1	咨询	150	41:38
Δ	数据与分析总监	10	3	咨询	300.000	36:52
E	数据分析师	6	3	电子商务	17.000	50:29
Z	数据与人工智能技术专家	30	3	IT	290.000	51:21
H	首席工程师	7	3	汽车行业	120.000	39:07
Θ	企业转型架构师	20	3	软件	110.000	42:11
I	工程总监	10	2	外卖	5.000	45:23
K	解决方案顾问首席专家	40	3	软件	110.000	46:22
Λ	诊断数据办公室负责人	20	3	医疗保健	100.000	45:59
M	高级经理、解决方案架构师	8	5	软件	4.000	23:42
N	数据与人工智能技术主管（DACH）	6	0.5	IT	290.000	25:45
Ξ	数据与人工智能战略高级经理	12	2	咨询	415.000	49:39
O	首席数据架构师	25	3	汽车行业	37.000	52:55

更清晰的概览 [13]。之后，我们基于释义片段，在访谈层级建立了第一轮编码迭代。此后，我们对编码进行审查，以确保其准确反映访谈内容。在第二阶段，我们将访谈层级的编码集成到总体框架中。为此，我们首先将编码归入以下主要主题：理论理解、案例描述、动机因素、挑战、实施策略、影响、准备度、展望和原型，以便围绕既定研究问题进行推论。图3通过饼图展示了各主要主题的相对分布和绝对分布。

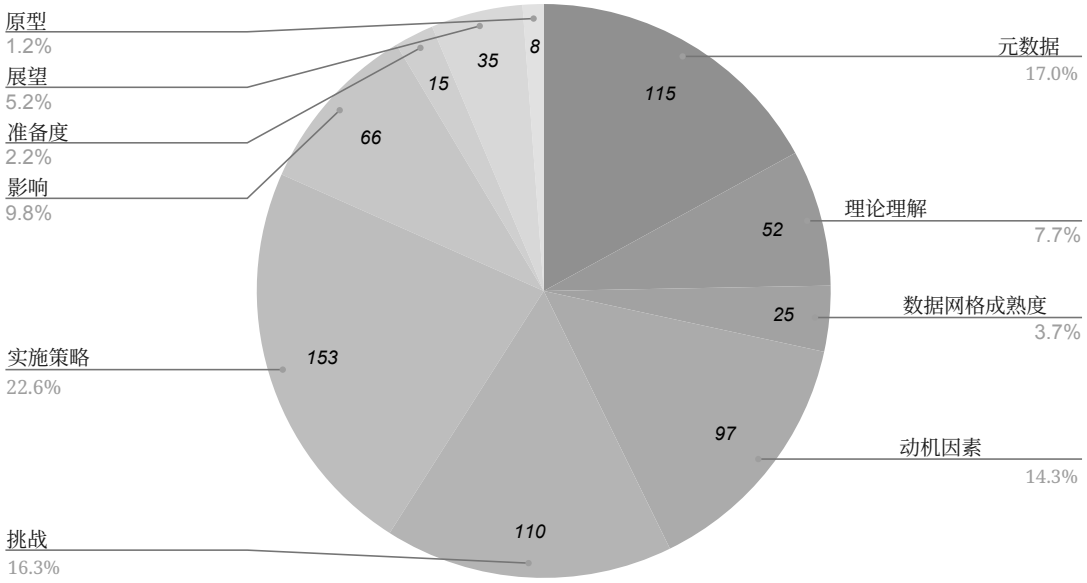

图 3. 访谈主题饼图。主题按照访谈指南顺时针排列。档案主题中的 39 个编码未列出。

在每次迭代中，我们都会创建并调整各主题下的子代码，将多位受访者的相似陈述归类为主轴编码 [23]。完成主轴编码后，我们进一步细化各子组，得出选择性主题，分别对应以下章节中呈现的动机因素、挑战、实施策略、影响和原型。总计，我们在 15 位受访者的访谈材料中得出 717 个（子）代码，编码工作耗时 48 小时。图 4 展示了各受访者编码片段的绝对数量。

4 结果

在本节中，我们综合了访谈结果。我们将介绍受访者对理论理解、促使其采用数据网格的动机因素、他们面临的挑战以及由此形成的实施策略。此外，我们还关注受访者观察到的影响。最后，我们介绍采用数据网格概念的组织的两种原型。我们仅关注与数据网格应用高度相关的方面。不过，一些挑战和实施策略并非数据网格所独有，而是与变革管理和技术引入这一更广泛的主题相关 [8, 25, 参见]。

4.1 理论理解

受访者普遍支持 Dehghani 提出的数据网格核心原则（$\Theta - M,\ O$），并强调要成功采用数据网格，四项核心原则缺一不可。然而，当被要求在这些原则中择其一时，受访者认为面向域的去中心化数据所有权和数据即产品这两项原则最为重要。此外，受访者表示，需要进一步制定指南，以帮助采用和运营数据网格（$A,\ \Delta,\ E,\ \Theta,\ K$）。$A$ 表示：“这一概念被描述得过于理论化，我

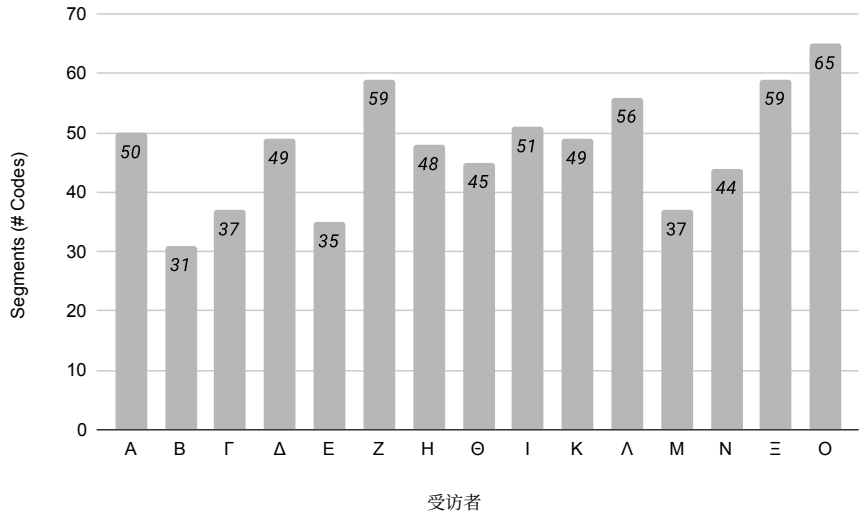

图4. 各受访者编码片段的条形图。纵轴为编码片段数量；横轴为表1中定义的受访者。

认为，许多组织都面临着将这一理论付诸实践的挑战。”而 *E* 则认为，若要直接采用数据网格，其核心原则“过于宽泛”——这凸显了为专业人员提供面向实践的指南、以支持他们今后开展工作的重要性。

4.2 动机因素

在采用数据网格概念之前，受访者面临各种问题，因此考虑了多种动机因素。

动机因素1：减少瓶颈 ($A-E, H, \Theta, K-O$)。第一个因素结合了瓶颈、可扩展性和速度等相关概念，与 Dehghani [11] 最初提出的方案一致。瓶颈是指中央数据管理系统或负责数据提供的团队缺乏处理能力，导致来自多个领域的数据请求积压。因此，各领域无法以期望的速度扩展数据驱动用例。例如， *M* 表示：“在外卖领域，他们再也无法扩展了，因为中间这个大数据仓库成了瓶颈”。速度这一概念有两个维度：一是中央团队提供所请求数据所需的时间，二是新产品和服务的上市时间； *B* 强调了其重要性：“快速交付客户功能、改善客户旅程并拓展更多市场。我们是一家超高速增长的公司。在超高速增长阶段，速度至关重要。”

动机因素2：利用领域知识 ($A-Z, I, K, \Xi, O$)。在集中治理的组织中，IT部门提供数据，但他们缺乏领域知识，无法确保数据达到任务特定质量。受访者强调领域知识对于提供更高质量的数据至关重要。 Γ 强调说：“那个人并不具备领域知识 [...] 所以结果很糟糕”，而 *E* 则进一步指出：“我们需要想办法把处理数据的能力交还给真正了解这些用例到底是什么的人。”

MF3：打破信息孤岛 ($A, \Gamma, E-I, \Lambda, M, \Xi, O$**)**。对于跨领域数据驱动用例而言，可访问性至关重要。然而，组织内部的信息孤岛妨碍其他领域访问数据。例如，营销部门（领域 A）可能需要向销售部门（领域 B）申请数据，以追踪区域营销活动的成效。但是，如果销售部门未能提供数据，营销部门就会错失宝贵的洞察。 O 举了一个真实案例：“我们发现，公司里没有人完全清楚有哪些数据，以及这些数据可以在哪里访问。”

MF4：确立数据所有权 (I, Λ, N**)**。第四个动机因素强调了数据所有权的重要性。它指的是对数据产品的质量和可访问性承担端到端责任。这一概念与传统的产品所有权角色类似。 I 以如下方式描述了潜在问题：“我们遇到的问题是，当我们将数据推送到中央数据平台时，数据所有权就丧失了。”此外， I 将更强的所有权与更高的数据质量联系起来：“我认为这非常简单。我们这样做的主要原因是提高数据质量。我们的设想是，更强的数据所有权会带来更高的数据质量。”

MF5：采用现代架构 (Z, K, Λ**)**。组织内的管理者和员工都意识到数据架构领域即将兴起的趋势——他们看到整个行业的组织都在采用数据网格，“不想错失良机”（Λ）；“有些公司只是听说了这股热潮，所以才开始关注它”（K）。然而，数据网格并非放之四海而皆准的方法。因此，这一发现令人十分担忧，因为它可能会让组织或业务单元出于错误的原因采用数据网格。

MF6：减少冗余（B， I， O）。信息孤岛、沟通不畅和缺乏透明度，可能导致大型组织中出现重复工作，包括数据准备等数据相关领域（B， I）或整个用例（O）。不同数据消费者会分别为各自的用例准备数据，却没有与组织共享结果——“我们在研究中发现，团队显著改善了数据，却没有把数据分享回公司”（B）。同样，I 指出：“我们一直把问题追溯到数据生产者，希望他们能够解决[数据质量] 问题，而不是让下游反复处理这个问题。”因此，跨领域的重复工作会挤占用于价值创造的资源（B）。此外，来自大型去中心化组织的受访者抱怨称，他们耗费大量资源为其他部门提供数据，而数据网格旨在简化这项工作（I， O）。

讨论完动机因素后，至关重要的是认识到，人们在使用数据网格概念时可能遇到的挑战。现在，我们来分析可能阻碍数据网格采用进程的障碍。

4.3 挑战

组织在采用数据网格的过程中面临诸多挑战。我们主要关注数据网格相关主题中独特且高度相关的挑战。

C1：联合数据治理 (A， $\Delta - O$**)**。我们发现，从集中式数据治理转向联合数据治理，是专业人员面临的主要挑战。受访者表示，联合式方法给以往由中央集中管理的活动和职责带来了困难。他们强调，联合数据治理对于根据领域需求制定规则至关重要；与此同时，受访者指出，自动化执行存在局限，尤其是在安全、监管和隐私相关主题方面（A， $E-\Theta$， N， O）。 O 指出，领域内的员工并不了解哪些数据受到保护和监管。 N 警告称，管理者可能会因未遵守数据保护法规而“最终一只脚踏进监狱”。此外，受访者表示，缺少中央部门导致各领域之间的用例和项目优先级排序不足（Δ， Z， I）。 I 认为，去中心化所有权和联合治理导致可观测性不足，进而使合理的优先级排序更加困难。

C2：责任转移 (B，Γ，Z，Θ，$I-\Lambda$，N**)**。参考文献 [11] 提出的首要原则是，所有权从中央权威机构转移至各领域。然而，领域内的去中心化所有权意味着要对数据产品的可访问性和质量承担端到端责任。这项新的责任带来了一系列次级挑战。首先，数据产品负责人认为，为其他领域提供数据是一项额外工作（Γ，Θ，K）——“他们不得不和待办事项列表中的其他所有事项争抢资源，而且看不到这样做有什么价值”（Γ）。此外，各领域通常无法因有益于其他领域的数据提供工作而获得直接补偿（Θ，K，N）。而且，各领域的业务活动通常并非围绕数据产品供给展开，因此会降低这项任务的优先级（B，Γ，Z，I，Λ）。

C3：元数据质量(E, Θ, I**)**。受访者指出的第三项挑战与数据产品模型有关。数据产品存储各种数据类型，并提供元数据，帮助缺乏领域知识的用户解读数据 [11, 16]。然而，受访者指出，提供方领域所提供的描述与使用方领域正确解读数据所需的信息之间存在差距。E 表示：“数据建模在大多数公司中是最关键的部分，但目前还没有相关标准。”随后又补充说：“如果你解决了数据建模 [...]，如今就很可能掌握了数据领域的金矿。”元数据质量方面的挑战并非数据网格独有——“以前这个问题就已经存在，只是没人真正着手处理”（Θ）——但由于对数据产品及其用例的预期不明确，这一问题变得更加复杂（I）。

C4：理解($\Delta, Z, \Theta, \Lambda, X$**)**。第四项挑战与对数据网格概念缺乏理解有关。受访者表示，组织内的员工只是使用新术语，却没有带来组织或技术层面的变革，例如把所有数据集都称为数据产品（Δ），或将高度集中的数据存储归类为数据网格（Λ）。Λ 进一步解释说，他/她觉得“昨天的仓库如今被称为网格，就像今天的会议被称为同步会一样”。这表明从业者对数据网格术语存在严重误解，使数据网格倡议难以推进。

受访者指出了其他值得关注的挑战，尽管这些挑战并非数据网格所独有。这些挑战值得重视，因为它们有助于制定针对数据网格的具体实施策略；忽视这些挑战可能会危及数据网格的成功采用。

C5：资源限制 ($A-\Delta$，$Z-\Theta$，$K-M$，Ξ**)**。C5描述了资金、技术和人力资源的匮乏——“构建数据网格以及相应的数据产品非常耗费资源。你需要很多掌握ETL知识的数据工程师。”（A）。Θ 进一步阐述道：“每个人都希望实现数据驱动，但谈到所需投入时，企业通常不愿意投入更多人力或资金。”

C6：接受问题与抵制 (Γ，Z，Θ，Λ，M，Ξ，O**)**。C6涉及组织内部的接受度与抵制，例如，O 表示：“大多数时候，我就像个.....怎么说呢，像个幼儿园老师。是的，只能告诉大家，请改变想法，请转换视角，请开始从数据的角度看待数据。”

C5 和 C6 在变革管理研究中都已得到充分确立，因此我们不再进一步讨论，而是参照这一主题的现有文献 [2, 8, 25]。

4.4 实施策略

在访谈中，我们确定了能够帮助组织克服上一节提出的挑战的实施策略。我们重点说明挑战与实施策略之间的关系，使组织能够根据自身需求应对挑战。这些实施策略是本研究的主要贡献，因为它们为专业人员和研究人员提供了初步指南。接下来，

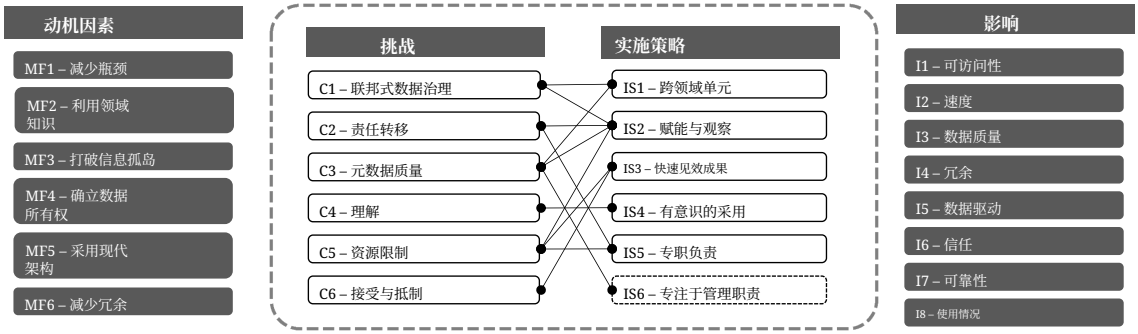

图 5. 数据网格：由动机因素和影响构成的挑战与实施策略之间的关系。

我们将介绍实施策略并提供详细说明。第 9 页的图 5 总结了我们的核心发现。

IS1：跨领域单元 ($A, B, \Delta, Z-I, N$)。受访者认为，向联邦式数据治理转型（C1）是主要挑战。为应对这一挑战，我们认为组织应设立一个跨领域指导单元，负责战略规划、用例优先级排序（A, Δ, Z, H），并落实具体的治理规则，尤其是涉及安全、监管和隐私相关主题的规则。在尚无法实现全面自动化时，这一点尤为必要。Δ 进一步指出，在早期阶段，从中央团队或支持团队调配人力资源至关重要，因为这有助于那些用例价值高但技术专长不足的领域。不过，跨领域指导单元也有助于推动制定高标准的元数据标准（C3）（Δ, Z, H, Θ, I）。此外，受访者表示，集中供给关键数据产品可能会带来益处（B, Δ, N）。B 认为，早期实行集中所有权，能够为各领域提供榜样，使其效仿并供给领域所有的数据产品。此外，这也能确保战略价值较高的数据顺畅运行（B, N）。B 补充道："我们所做的一件事，就是选出 20 个关键程度很高的数据集。这些数据集由我们集中所有，并转化为数据产品。这样一来，我们就能为其他领域树立榜样。"不过，我们强调，跨领域指导单元只能补充和支持联邦式治理结构；随着数据网格日趋成熟，它可能会逐渐变得不再必要，例如平台集中提供的策略执行实现全面自动化之后。

IS2：赋能与观察 ($B,\ \Delta,\ Z-K,\ M-O$)。第二种实施策略旨在解决责任转移（C2）、元数据质量相关问题（C3），以及向联邦式数据治理过渡的问题（C1）。数据网格的核心理念之一是利用领域知识。不过，中央部门需要提供技术基础设施、支持、模板和指南，以促进高质量数据产品的供给（$B,\ \Delta,\ Z-I,\ N$）。这一发现与数据网格的第三项主要原则——自助式数据平台——相一致，该原则强调赋能各领域的重要性。在访谈过程中，受访者多次强调观察数据产品质量和使用情况的重要性（$Z-K,\ M,\ \Xi,\ O$），并指出了两大主要益处。首先，数据产品的使用数据可以激励各领域的数据产品负责人，因为这些数据直观呈现了数据产品所产生的影响。这些数据有助于各领域凸显自身在组织中的重要性，并为其争取额外资源提供筹码（C5）。此外，我们建议组织观察并自动评估数据产品质量，最好由独立单元负责。评分应包括元数据评分、数据质量评分和可用性指标。根据数据的关键程度和数据产品质量标准，组织还应考虑安排一名人在回路的审核人员负责初步审批（Δ）。O 认为，数据产品

质量评分能够激励数据产品负责人确保元数据具有较高质量（C3）。 Δ 进一步强调：“不需要层层升级处理 [.....] 只要知道这件事会被报告 [.....]他们现在处理问题的方式就大不相同了”。因此，我们建议中央指导小组跟踪并排列关键数据产品，以推动数据产品负责人提供高质量数据产品。

IS3：快速见效成果（$A-O$）。尽管组织内部的接受问题和抵触（C6）已得到充分证实，但数据网格专属实施策略有助于应对组织挑战。我们综合受访者的发现，制定了一个细化的多步骤流程，围绕组织挑战指导专业人员完成数据网格探索与启动阶段。首先，数据网格倡议应从“预算所在之处”开始，因为财务资源对于数据网格采用的成功至关重要（Z，C5）。如果中央 IT 部门拥有启动向数据网格转型所需的资源，则应由其负责。如果组织已经高度去中心化，则应由具有影响力的领域率先行动。无论采取哪种方式，在迭代转型阶段，中央 IT 部门与多个领域之间的沟通都至关重要。下一步，各领域与中央 IT 部门应制定一个兼顾所有利益相关者利益的概念，以确保未来的采用，例如使用调查问卷或访谈。随后，数据网格倡议的主要推动者应选择多个试点数据产品。试点数据产品应具备以下特征：覆盖多个用例（A、Z、Θ、Λ），成本低廉（A、Δ、$K-M$），规模小（A、Λ、M）但影响显著（A、Z、$I-\Lambda$、N、O），并且能够轻松快速地取得成果（A、H、Λ、Ξ、O）。因此，数据网格倡议的推动者能够展示成功的用例，以获得高层管理层的批准，并在整个组织内推广数据网格（H、Ξ、O）。与此同时，成本低廉且规模较小的方法能够确保初始试点计划获得接受度。此外，教育举措和社区建设工作应支持数据产品的初始开发。 Δ 强调了通过项目推介之夜、社区聚会或黑客马拉松等活动庆祝早期成果的重要性，以便在组织内部形成发展势头。

IS4：有意识的采用 (B，Δ，H，$I-\Lambda$，Ξ，O)。在组织决定采用数据网格之前，应全面评估现有的数据管理系统、架构及其不足之处，以及数据网格可能带来的益处。此外，还应考虑组织结构和规模。如果组织决定继续推进，至关重要的是有意识、谨慎地引入数据网格相关术语。 Λ & O 甚至建议完全不要使用数据网格相关术语。 Λ & O 认为，这样做有助于深入理解其底层概念，并避免炒作引发的误解（C4）。为加深理解，I 表示：“我通常会把它联系到人们熟悉的事物上，以此说服他们，而这通常是应用程序接口。基本上，可以把数据产品看作应用程序接口。你不会在没有团队支持、也没有明确定义其内容的情况下，就向全世界发布一个应用程序接口。”

IS5：专职负责(A，Λ，O)。责任转移（C2）凸显了一个挑战：领域员工认为提供数据产品属于额外工作，因此不会优先处理这项任务。由于组织必须在有限资源（C5）的条件下开展工作，管理者需要仔细考虑员工人数，以及员工能够投入数据产品提供工作的精力。我们的访谈表明，出于三个主要原因，管理者应优先选择由专人负责数据产品提供工作的较小团队，而不是承诺度较低的较大团队。首先，不会出现优先级冲突，因为首要任务就是提供数据产品。此外，作为首要任务的工作岗位不易受到组织内部可能发生的重组和重构工作的影响——“重组来临时会发生什么？ [...] 所有不是你的主要工作的事情都会被取消。你要么做好自己的主要工作，要么就得换一份工作”（Λ）。最后，专职数据产品负责人能够积累更深厚的专业知识，尤其是在处理敏感数据和满足监管要求方面（O）。

IS6：专注于管理职责(*O*)。数据管家是组织或数据产品团队中的一个角色，负责整体数据管理，例如制定数据交互规则 [27]。根据 *O*，数据产品管家拥有数据产品，并确保其质量、适用性和可用性，包括元数据质量（C3）。此外，数据产品管家全面了解数据产品现有的各类用例；这是对数据产品负责人角色的关键扩展，后者专注于提供数据产品以供自助服务。 *O*概述了一套协议：数据消费者提交数据产品访问请求，并简要说明预期用途。管家有多种决策路径可供选择：第一，批准请求；第二，批准请求，并协助申请者联系相关的现有用例，以探索潜在协同效益；第三，因监管或安全约束而拒绝请求；第四，因存在重叠的现有用例而拒绝请求。例如， *O* 详细描述了一个实例：某业务单元打算销售已由组织另一部门推向市场的数据。若按此计划行事，会减少现有销售收入，并导致组织结构重复。因此，管家可以运用自身知识，有效减少重复工作（MF6）。然而，专业人员需要权衡这些优势与可能造成瓶颈（MF1）和信息孤岛（MF3）的风险，因为事先审批可能会牺牲快速开发和自助服务。处于数据网格环境中时，数据产品负责人可以承担这些额外职责，也可以与数据产品团队的另一名成员分担这些职责。

4.5 影响

在 15 位受访者中，有十位分享了他们在较晚的探索与启动阶段向数据网格转型的工作经验（Γ， Δ， *Z*， Θ， *K* – *O*）。此外，有八位受访者分享了第二阶段扩展与规模化的见解（*A*， *B*， Δ， *E*， *H*， *I*， Ξ， *O*）。因此，所有受访者都能够提供有关其数据网格采用所产生影响的（初步）细节。

I1：可访问性 （*A* – Γ， Δ， *Z* – Θ， Λ， *N*， *O*）。我们观察到，数据网格采用最显著的影响是可访问性的提升，这与瓶颈和信息孤岛的减少直接相关（MF1、MF3）；“你可以自行进行数据分析，不用询问任何人。如果发现有需要改进的地方，你只需自己动手。同样，你也不必问任何人。”（Γ）。

I2：速度(Γ – *E*, *H*, *I* – Λ, Ξ**)**。此外，受访者观察到速度有所提升。正如我们在 MF1 中所述，瓶颈的减少从两个方面带来了速度提升。首先是获取访问权限时间，其次是新产品和服务的上市时间。受访者观察到这两方面都有所改善：“过去需要五个月才能完成的事情，现在三个为期六周的冲刺周期就能完成”（Ξ）。

I3：数据质量(*A*, Γ, Δ, *M* – Ξ**)**。此外，受访者观察到，由于职责界定明确，数据质量有所提升；“由一位数据所有者专注于特定数据域，数据质量会更好”（*A*）。 Γ 称数据“更新及时得多，出错的可能性也更小”，而 Δ 认为，由于问责力度加强，数据所有者更有动力交付高质量数据产品：“如果数据被跟踪，就会被报告；当然，没人想排在名单的末尾。”

I4：冗余减少(*B*， Δ， *H* – *I*， *M*， *O***)**。“一次开发，多次使用”(Δ)；这与 MF6 直接相关，受访者观察到冗余成功减少。具体而言， *B* 指出，以经过整理、干净的格式提供最受欢迎的数据集，能带来“巨大的价值，因为有许多工程团队会重复使用这些数据集”。此外， Θ 观察到，由于标准化，提取、转换、加载（ETL）流程更加高效：“他们为数百个系统实施了数百条数据管道，投入了数百倍的工作量。有了数据网格，

只需向数据市场提供标准化 API。” *O* 还提到，领域知识可以更轻松地转移。

I5：数据驱动 (*M* – Ξ**)**。除此之外，我们还观察到，组织变得更加数据驱动。这与组织的整体战略一致；据受访者称，这也与员工心态的转变以及一系列广泛措施相关。

I6：信任 (*I*，*K*，*M***)**。此外，受访者指出，由于领域知识（MF2）和数据质量有所提高，数据信任也随之增强；“突然之间，你对数据更加信任了，因为数据背后有负责这些数据的业务人员；他们负责质量、数据和产品经理。”（*M*）。

I7：可靠性 (*A***)**。此外，*A* 报告称，与数据仓库概念相比，可靠性有所提高，这归功于可访问性和数据供给的改善；此前，需求高峰期并行查询会导致系统过载和故障，中央数据库也会因此崩溃，“由于采用分布式访问，我们观察到数据访问的停机时间有所减少”（*A*）。

I8：使用情况(*B*，*H*，Ξ，*O***)**。最后，受访者报告称，数据可访问性的提升使得用例数量出现了“爆炸式增长”（*H*）。更确切地说，由于以往的访问障碍不复存在，员工能够在短得多的时间内开发新用例并迭代现有用例。

我们的观察表明，采用数据网格总体上会带来积极影响，这说明组织即使处于转型初期也能从中受益。

4.6 组织原型

我们承认，组织在数据管理系统和架构方面采取的方法具有很强的个体差异。因此，不存在放之四海而皆准的方法，组织需要权衡利弊，并根据自身需求采用相应的实施策略。尽管如此，我们在研究结果中识别出两种原型，它们各自代表着具有相似动机、挑战和实施策略的组织。在本小节中，我们将介绍这两种原型，并着重阐述最重要的研究发现。

A1：初创与规模化企业(*B*，Γ，Δ，*E*，*I*，*M*，Ξ**)**。初创与规模化企业是处于动态超高速增长环境中的年轻组织。它们是较早采用数据网格的组织，平均而言，准备度和成熟度也更高。此外，初创与规模化企业通常由较年轻的员工队伍构成，拥有进步的组织文化，并以成熟的数据文化为支撑；数据与人工智能是其产品或服务不可或缺的一部分。由于组织成立时间较短，初创与规模化企业通常没有遗留系统，能够从绿地式方法入手开展架构变更。此外，由于成立时间较近且产品或服务范围较窄，这些组织往往具有更集中的结构。因此，数据网格倡议通常由上而下推动，并得到高层管理层的大力支持。

由于处于超高速增长环境中，他们的重点是速度和可扩展性（MF1）。然而，利用领域知识来提升数据质量（MF2）和减少冗余（MF6），对于增强机器学习能力并避免数据准备中的重复工作同样至关重要。MF1-2 和 MF6 导致了以下主要挑战。

首先，动态环境以及数据驱动的产品和服务供给通常要求具备较高的元（数据）质量（C3）。此外，转向联邦式组织结构可能颇具挑战（C1）。相比之下，由于组织层级较扁平且采用自上而下的方法，对新举措的接受度与抵制（C6）以及理解（C4）通常不那么棘手。资金资源（C5）通常也不是主要挑战，因为组织更看重速度和增长，而非盈利能力。

鉴于这些情况，若干实施策略至关重要。建议设立跨领域指导单元（IS1），以应对标准化方面的挑战，并为各领域提供明确指导。尽管在这些自上而下推动的倡议中，快速见效成果（IS3）不那么关键，但赋能团队并观察成效（IS2）对于建立适当的激励机制、确保产出高质量数据产品至关重要。此外，在快节奏的环境中，专职负责（IS5）对于明确职责和问责十分重要，也是数据网格倡议持续取得成功并实现可扩展性的关键。应谨慎采用侧重于数据管护的IS6，因为它可能会阻碍新产品或服务的快速开发。对于层级较少的小型组织而言，重复开发完整用例的可能性较低，因此这一点尤为重要。

这种方法有助于初创企业和规模化企业有效应对各自独特的挑战，并充分发挥其固有的敏捷性，以及采用创新数据管理实践的准备度。

A2：成熟组织(A，Γ，Δ，Z，H，Θ，K，Λ，N，Ξ，O)。成熟组织历史悠久，员工人数和营收规模都很大。它们拥有一支资深员工队伍，并采用层级结构，由覆盖多个业务领域的半自治单元组成。由于公司的历史原因，成熟组织通常拥有（多个）遗留系统，组织文化也更为保守。它们通常依赖数据仓库或数据湖等集中式数据存储概念。

成熟组织采用数据网格的主要动机因素包括：减少瓶颈和数据孤岛，以提高可访问性（MF1、MF3）；提升（元）数据质量（MF2）；明确数据所有权（MF4）；以及减少冗余（MF6）。

他们面临的主要障碍是组织挑战。组织框架往往缺乏灵活性，导致组织内部在联邦式数据治理（C1）和抵制（C6）方面出现问题。分配财务资源（C5）的僵化方法，阻碍了拥抱数据网格所需的转型。这一点尤其值得注意，因为量化数据质量和可访问性提升带来的财务优势可能并非易事。因此，这些优势无法为内部预算谈判提供有力依据。此外，昂贵的遗留系统也让数据网格的商业论证更难成立。

鉴于上述挑战，在采用的早期阶段取得快速、有效且成本低廉的成果至关重要（IS3）。借助这种方法，数据网格倡议可以在试点阶段悄然推进，若取得成功，便可利用试点数据产品或领域来推动广泛采用并争取预算。

此外，我们建议设立跨领域指导单元（IS1），整合多个数据网格倡议，提升整个组织的数据可访问性。还应考虑实行专职负责（IS5）以及赋能与监督策略（IS2），因为这些做法与成熟组织的层级结构十分契合，能够明确职责，并在数据治理框架内促进问责。

5 讨论

在本　节中，我们讨论结果，突出贡献，承认局限性，并指出未来的工作方向　　.

5.1 总结与文献嵌入

数据网格是一种包含架构方面的社会技术概念，能够帮助组织变得更加数据驱动。然而，数据网格的架构概念是新颖的，仍需进一步开展科学研究，尤其是在动机因素、挑战、实施策略、影响和原型方面。为获得行业洞察，我们开展了15次半结构化专家访谈。

我们发现，受访者对数据网格的看法及相关动机因素总体上与 Dehghani 提出的概念一致。受访者指出，需要更快速、可扩展的解决方案来减少瓶颈、利用领域知识并打破信息孤岛（MF1-3）。此外，确立数据所有权和减少数据冗余也有充分的记录（MF4、MF6）。我们还发现，当前围绕数据网格的热潮促使组织内的利益相关者为了采用而采用数据网格（MF5）。然而，数据网格并非适用于所有数据架构的通用概念，数据网格采用不应成为目的本身。这一洞察凸显了开展清晰、以事实为依据的讨论和沟通的必要性。参考文献 [30] 通过对选定大数据架构的初步评述回应了这一关切。

除了动机因素外，我们还发现了若干挑战，其中部分挑战已在现有文献中有所体现 [28, 参见]。首先，我们将向联邦式数据治理转型确定为一项主要挑战（C1）。参考文献 [39] 为这一挑战提供了进一步证据：作者认为，在某些特定情况下，联邦式治理可能难以实施，并强调中央指导单元对于提供跨领域洞察的重要性。此外，他们还对数据产品的合法访问提出了担忧 [39]。参考文献 [29]进一步拓展了这一论点，并针对数据网格概念中的隐私挑战进行了广泛分析。此外，研究人员将领域间共识和合规问题描述为“不断变化的目标” [20]。与参考文献 [39], 的观点一致，我们将责任转移确定为一项主要挑战（C2）。作者认为，各领域缺乏提供数据产品所需的能力、动力和资源 [39]。此外，有限资源（C5）以及接受度和抵制问题（C6）是组织与文化变革管理领域中公认的挑战；专业人员和研究人员应参考丰富的相关文献，以补充本研究的发现 [8, 20, 25, 39, 参见]。访谈显示，采用现代架构（MF5）可能与对数据网格概念理解不足（C4）有关，因为采用者不想“错失良机”，而不是基于扎实、切实的动机。这种关联在近期出版物中鲜有涉及；令人意外的是，受访者在多个用例中都提到了这一挑战。数据网格这一特定语境下的元数据质量（C3）也同样鲜少受到关注。

针对上述挑战，我们制定了实施策略（IS1-6）。首先，为应对C1和C3，我们认为专业人员应设立跨领域指导单元（IS1）。一项近期研究支持这一观点。参考文献 [40] 指出，组织应为下一代数据平台兼顾去中心化与中心化能力。此外，我们认为，为应对C2、C3和C5，组织需要赋能各领域，并密切观察和跟踪其进展（IS2）。与参考文献 [11], 一致，我们阐述了赋能各领域的必要性。然而，将观察数据产品作为激励提供方和进行优先级排序的依据，具有重要意义；这一观点颇具新意，仍需进一步研究。此外，参考文献 [11] 建议采用迭代方法来采用数据网格概念。通过分析访谈记录，我们证实了这一方法，并在IS3中进一步探讨了如何应对组织挑战（C5、C6）。

有意识的采用（IS4）是实施数据网格时的一个重要方面。组织在采用数据网格之前，应先全面评估当前的数据架构，而不是出于纯粹追逐炒作的动机 [6, c.f.]。我们还建议，专业人员在供给数据产品时，优先选择小型专职团队，而不是大量员工（IS5）。这一实施策略旨在以有限资源优先开发数据产品（C2, C5） [39]。最后，我们认为，组织应考虑扩大数据产品负责人的职责，或设立数据产品管家这一角色（IS6），以确保数据质量、合规性、适用性和可访问性。对于需要严格控制数据产品访问权限并全面了解用例的组织而言，这或许值得考虑。不过，我们承认，IS6 在某种程度上与

Dehghani [11] 强调数据产品应能在整个组织内轻松访问的观点相冲突，并且可能造成新的瓶颈或信息孤岛（MF1,MF3）——因此，必须谨慎实施 IS6。

在所有访谈中，受访者都指出了其数据网格采用工作的积极影响。我们的研究结果包括可访问性提高（I1）、速度提升（I2）、数据质量改善（I3）、冗余减少（I4）、组织更加数据驱动（I5）、信任度提升（I6）、可靠性增强（I7）以及数据与分析用例增加（I8）。总体而言，这些积极影响表明，组织即使处于采用早期阶段，也能够获得预期收益，从而证实了 Dehghani 关于数据网格影响的 [11]观点。

我们提出了两种采用数据网格概念的组织原型：初创企业与规模化企业（A1），以及成熟组织（A2）。专业人员应根据自身独特情况评估实施策略；为此，A1 和 A2 为制定更具体的指南提供了初步步骤。

5.2 贡献与影响

本研究加深了对数据网格的实证理解，弥补了现有科学文献中的一个重要空白。由于该研究领域较为新颖，现有文献目前数量有限，且大多停留在理论层面。鉴于探索数据网格实践方面的独立学术研究数量较少，本文提供了全面且科学严谨的重要见解。

我们识别出组织采用数据网格的动机因素（MF1-6），例如对敏捷性、可扩展性和提升数据可访问性的需求。了解这些因素有助于论证向数据网格架构转型所需的投资和投入。

此外，本研究识别出若干挑战（C1-C6），帮助考虑实施数据网格的组织做好准备，并制定策略，主动应对这些挑战，从而降低可能危及数据网格实施的风险。

接下来，我们概述了针对特定挑战量身定制的实施策略（IS1-6），这些初步指南有助于成功实施并降低相关风险。

此外，研究采用数据网格所产生的影响（I1-8）至关重要，因为这能提供有关其益处的实证证据，帮助组织评估该架构在现实环境中的有效性和可扩展性，从而更好地进行战略规划。

最后，基于组织特征划分的原型可指导企业针对数据网格采用情况定制方法，指出更可能出现的挑战及相应的实施策略。

本研究的意义深远，尤其是考虑到针对这一主题的独立研究十分匮乏。对于从业者而言，本文可作为一份初步指南，帮助他们理解并实施数据网格，使其既具备战略性，又能应对各自特定的组织挑战。本文提出的见解有助于应对常见陷阱并运用有效策略，从而显著降低数据网格倡议的风险，提高其成功率。

本研究也填补了学术界的一项空白，为数据网格在不同行业中的实施情况提供了有据可查的分析。它为未来研究奠定了基础，尤其有助于进一步探讨随着数据网格日趋成熟并得到更广泛的采用，这些策略与挑战将如何演变。

最后，这项工作为（大）数据管理领域的学术界和从业者提供了宝贵的参考。它帮助读者更清晰地理解动机、策略和影响，并以透明的方式传达这些方面，同时也突出了潜在的挑战。通过这种方式，它有助于将数据网格置于更广泛的（大）数据管理领域中。

5.3 局限性与未来研究

为回答我们的研究问题，我们开展了15次半结构化专家访谈。我们工作的定性性质导致其定量效度有限。然而，鉴于研究主题的新颖性，我们认为采用定性方法是合理的。尽管如此，未来研究仍应在定量层面进一步检验这些发现，例如使用调查问卷。

关于抽样方法，我们承认可能存在偏差：受访者可能会将其数据网格实施描述得比实际情况更成功，尤其是当他们希望公开树立自己作为数据网格领域领导者的形象时。为应对这一挑战，我们已明确说明研究重点仅限于研究活动。此外，我们会对受访者数据进行匿名处理，鼓励他们畅所欲言，谈论自己在组织中的工作经验，无论是积极还是消极的方面。此方法旨在减少偏差，从而更准确、细致地了解数据网格实施情况。

我们的主要贡献是为采用数据网格的专业人员提供行业洞察。不过，我们承认，这些实践的适用程度取决于各组织的具体情况。因此，专业人员需要相应地调整策略，仅纳入相关方面。

本文提出的组织原型，是根据具体特征为组织制定更细致指南的第一步。不过，我们承认，行业洞察整体而言，尤其是这些原型，缺乏定量证据。这为未来研究创造了巨大的机会，因为可以通过定量研究补充这一探索性定性方法。在此背景下，研究人员可以考察中小型组织中的数据网格，以拓展原型框架。

此外，未来的研究应更深入地探讨数据网格概念的技术实现。这可能包括详细设计技术型数据产品，并整合数据仓库、数据湖或 Blob存储，以实现数据网格概念。此外，还可以进一步探索可能的数据网格拓扑，以及理想数据网格节点规模的相关考量。

总而言之，前方有着充满希望的研究领域。

6 结论

本文基于15次半结构化专家访谈获得的见解，全面探讨了数据网格概念在各个行业中的采用情况。本研究弥补了现有文献中的空白，针对数据网格实施的动机、挑战、实施策略和影响，提供了实践性、实证性的见解。由于这一概念尚属新颖，迄今为止相关研究大多停留在理论层面。

研究结果表明，采用数据网格的动机包括减少瓶颈、利用领域知识、改进数据所有权以及打破数据孤岛，所有这些举措都旨在提升数据可访问性和质量。这些动机与 Dehghani 关于数据网格的基础框架所提出的理论优势高度一致 [11]。

然而，向数据网格架构转型并非没有挑战。这些挑战包括转向联合治理的复杂性、管理去中心化数据所有权所带来的责任、确保元数据质量，以及应对数据管理实践发生重大变化时可能出现的组织阻力。为克服这些障碍，研究提出了若干实施策略，例如建立跨领域单元、赋能并密切关注各领域的工作、争取快速见效成果、促进有意识的采用、落实专职负责，以及考虑数据管理员的作用。

数据网格早期实施所呈现的影响令人鼓舞，包括提升可访问性和数据访问速度、改善数据质量、减少冗余，以及推动组织整体朝着更加数据驱动的方向发展。这些影响证实，数据网格有望显著改善组织的数据管理实践。

本研究还初步识别出两种能够从定制化数据网格实施方法中获益的组织原型——初创企业与规模化企业，以及成熟组织。这种区分有助于理解如何调整数据网格，以满足不同类型组织的特定需求和特点。

本文通过详细分析数据网格在现实中的应用经验，为学术研究和数据管理实践应用均作出了贡献。本文为进一步研究奠定了基础，并帮助组织更好地应对采用数据网格架构所带来的挑战与机遇。随着数据网格日趋成熟、采用范围不断扩大，未来研究应继续以互补的定量方式探讨这些主题。

参考文献

[1] Omolola A. Adeoye-Olatunde 和 Nicole L. Olenik。2021。研究与学术方法：半结构化访谈。《美国临床药学学院杂志》4（2021年10月），1358–1367。第10期。https://doi.org/10.1002/JAC5.1441[2] Monica Anastassiu 和 Gleison Santos。2020。软件过程改进中的变革抵制——对原因、影响和应对措施的调查。载于第19届巴西软件质量研讨会。1–11。[3] Inês Araújo Machado、Carlos Costa 和 Maribel Yasmina Santos。2022。借助数据网格实施推进数据架构。商业信息处理讲义 452（2022），10–18。https://doi.org/10.1007/978-3-031-07481-3_2[4] Michael Armbrust、Ali Ghodsi、Reynold Xin 和 Matei Zaharia。2021。湖仓：新一代开放平台，统一数据仓库与高级分析。载于 CIDR 论文集，第8卷。[5] Pranjal Awasthi 和 Jordana George。2020。数据民主化的必要性。AMCIS 2020 论文集（2020年8月）。https://aisel.aisnet.org/amcis2020/ data_科学_分析_用于_决策_支持/data_科学_分析_用于_决策_支持/23[6] Andrew Burgess。2018。别轻信炒作。载于《人工智能执行指南：如何识别并在组织中实施人工智能应用》。Springer International Publishing，Cham，1–9。https://doi.org/10.1007/978-3-319-63820-1_1[7] Vijay Kumar Butte 和 Sujata Butte。2022。企业数据战略：去中心化的数据网格方法。2022年商业与工业数据分析国际会议（ICDABI 2022）（2022），62–66。https://doi.org/10.1109/ICDABI56818.2022.10041672[8] Rune Todnem By。2007。组织变革管理：批判性综述。《变革管理杂志》5，4（2007），369–380。https: //doi.org/10.1080/14697010500359250[9] Julia Couto、Olimar Teixeira Borges、Duncan D Ruiz、Sabrina Marczak 和 Rafael Prikladnicki。2019。关于数据湖的映射研究：改进的定义与可能的架构。载于 SEKE。453–578。[10] Zhamak Dehghani。2019。如何摆脱单体数据湖，迈向分布式数据网格。检索于2023年9月28日，网址：https://martinfowler.com/articles/data-monolith-to-mesh.html[11] Zhamak Dehghani。2022。《数据网格：大规模交付数据驱动价值》。O'Reilly Media。[12] Ilker Etikan、Sulaiman Abubakar Musa、Rukayya Sunusi Alkassim 等。2016。便利抽样与目的抽样的比较。《美国理论与应用统计杂志》5，1（2016），1–4。[13] Uwe Flick。2022。《定性研究导论》。《定性研究导论》（2022），1–100。[14] Susan Galer。2022。成为数据驱动型企业有何价值？检索于2023年9月28日，网址：https://www.forbes.com/sites/forbesbusinessdevelopmentcouncil/2022/01/07/whats-the-value-of-becoming-a-data-driven-business/?sh=21dfb9463776[15] Gartner。2023。数据织物的定义——Gartner 信息技术术语表。检索于2023年9月28日，网址：https://www.gartner.com/en/information-technology/glossary/data-fabric[16] Christoph Gröger。2021。没有数据，就没有人工智能。《Commun. ACM》64，11（2021），98–108。[17] Ahmed A. Harby 和 Farhana Zulkernine。2022。从数据仓库到湖仓：比较综述。《2022年IEEE大数据国际会议论文集》（Big Data 2022）（2022），389–395。https://doi.org/10.1109/BIGDATA55660.2022.10020719[18] Cornelia Helfferich。2011。《定性数据的质量》。第4卷。Springer。[19] William H Inmon。2005。《构建数据仓库》。John wiley & sons。[20] Divya Joshi、Sheetal Pratik 和 Madhu Podila Rao。2021。数据网格基础设施中的数据治理：盛宝银行案例研究。国际电子商务会议（ICEB）论文集 21（2021），599–604。[21] Klaus Krippendorff。2018。《内容分析：方法论导论》。SAGE 出版社。

[22] Xinchi Li、Mingchuan Yang、Xiaoqing Xia、Kaicheng Zhang 和 Kang Liu。2022。基于元数据知识图谱的分布式数据织物架构。2022年第五届数据科学与信息技术国际会议（DSIT 2022）论文集（2022）。https://doi.org/10. 1109/DSIT55514.2022.9943831[23] Yvonna S Lincoln 和 Egon G Guba。1985。自然主义探究。SAGE。[24] Inês Araújo Machado、Carlos Costa 和 Maribel Yasmina Santos。2021。数据网格：数据架构范式转变的概念与原则。《Procedia Computer Science》196（2021），263–271。https://doi.org/10.1016/J.PROCS.2021.12.013[25] J. S. Oakland 和 Stephen Tanner。2007。成功的变革管理。《全面质量管理与卓越经营》18，1-2（2007年1月），1–19。https://doi.org/10.1080/14783360601042890[26] Sergio Orenga-Roglá 和 Ricardo Chalmeta。2018。在组织中实施大数据生态系统的框架。《Commun. ACM》62，1（2018），58–65。[27] Boris Otto。2011。数据治理组织形态。《欧洲信息系统会议》（2011）。[28] Boris Otto。2022。欧洲数据空间的联邦式基础设施。《Commun. ACM》65，4（2022），44–45。[29] Nikolai J. Podlesny、Anne V.D.M. Kayem 和 Christoph Meinel。2022。知识状态综述：与数据网格相关的隐私挑战综述。《计算机科学讲义》（含人工智能讲义和生物信息学讲义子系列）13426 LNCS（2022），85–102。https://doi.org/10.1007/978-3-031-12423-5_7[30] Torsten Priebe、Sebastian Neumaier 和 Stefan Markus。2021。探索大数据架构丛林。《2021年IEEE大数据国际会议论文集》（Big Data 2021）（2021），5994–5996。https://doi.org/10.1109/BIGDATA52589.2021.9671862[31] Stefan Rädiker 和 Udo Kuckartz。2020。MAXQDA PRESS：使用 MAXQDA 对定性访谈进行聚焦分析的分步指南。（2020）。[32] David Reinsel、John Gantz 和 John Rydning。2018。从边缘到核心的世界数字化。检索于2023年9月28日，网址：https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf[33] Seref Sagiroglu 和 Duygu Sinanc。2013。大数据：综述。载于2013年协作技术与系统国际会议（CTS 2013）论文集（2013），42–47。https://doi.org/10.1109/CTS.2013.6567202[34] Nivas Shankar、Ian Meyers、Zach Mitchell 和 Roy Hasson。2021。使用 AWS Lake Formation 和 AWS Glue 设计数据网格架构 |AWS大数据博客。检索于2023年9月28日，网址：https://aws.amazon.com/blogs/big-data/design-a-data-mesh-architecture-using-aws-lake- formation-and-aws-glue/[35] Jatinder S. Sidhu、Edwin J. Nijssen 和 Harry R. Commandeur。2000。业务领域定义实践：它会影响组织绩效吗？《长期规划》33，3（2000年6月），376–401。https://doi.org/10.1016/S0024-6301(00)00037-6[36] Julia Stoyanovich、Serge Abiteboul、Bill Howe、HV Jagadish 和 Sebastian Schelter。2022。负责任的数据管理。《Commun. ACM》65，6（2022），64–74。[37] Thoughtworks。2023。数据网格 | Thoughtworks。检索于2023年9月28日，网址：https://www.thoughtworks.com/en-de/what-we-do/data- and-ai/data-mesh[38] Alejandro Vaisman 和 Esteban Zimányi。2014。数据仓库概念。《数据仓库系统》（2014），53–87。https://doi.org/10.1007/978-3-642-54655-6 3_[39] Kathrine Vestues、Geir Kjetil Hanssen、Marius Mikalsen、Thor Aleksander Buan 和 Kieran Conboy。2022。NAV 中的敏捷数据管理：案例研究。《商业信息处理讲义》445 LNBIP（2022），220–235。https://doi.org/10.1007/978-3-031-08169-9_14[40] Martin Whyte、Andreas Odenkirchen、Stephan Bautz、Agnes Heringer 和 Oliver Krukow。2022。数据网格——又一个流行词，还是下一代数据平台？PwC研究 2022：数据平台变革（2022）。https://www.pwc.de/en/digitale-transformation/data-mesh-the-next- generation-enterprise-data-platform.html[41] Michael Williams 和 Tami Moser。2019。定性研究中的编码艺术与主题探索。《国际管理评论》15，1（2019），45–55。