# 数据网格：系统性灰色文献综述

ABEL GOEDEGEBUURE、INDIKA KUMARA、STEFAN DRIESSEN 和 WILLEM-JAN VAN DEN HEUVEL，荷兰蒂尔堡大学GEERT MONSIEUR 和 DAMIAN ANDREW TAMBURRI，荷兰埃因霍温理工大学DARIO DI NUCCI，意大利萨勒诺大学

arXiv:2304.01062v3 [cs.SE] 7 Aug 2024

数据网格是一种新兴的领域驱动型去中心化数据架构，旨在最大限度地减少或避免企业采用集中式单体数据架构所带来的运维瓶颈。该主题引起了从业者的兴趣，并且已有大量相关灰色文献。与此同时，我们发现，学术界尚缺乏对这一概念进行定义并在此基础上开展研究的尝试。因此，本文旨在从基础出发，从设计原则、架构组件、业务能力和组织角色等方面刻画数据网格架构。我们系统地收集、分析并综合了114篇工业灰色文献。该综述揭示了从业者对于数据网格四项关键原则的观点：数据即产品、数据的领域所有权、自助式数据平台以及联邦式计算治理。此外，鉴于数据网格与SOA（面向服务的架构）具有可比性，我们将灰色文献中的发现映射到SOA学术文献中的参考架构，从而构建用于描述数据网格三个关键维度的参考架构：业务能力和角色的组织、开发以及运行时。最后，我们讨论了数据网格中尚待解决的研究问题，其中部分讨论基于灰色文献中的发现。

**ACM 引用格式：**
Abel Goedegebuure、Indika Kumara、Stefan Driessen、Willem-Jan van den Heuvel、Geert Monsieur、Damian Andrew Tamburri 和 DarioDi Nucci。2024。DataMesh：系统性灰色文献综述。1, 1（2024 年 8 月），35 页。https://doi.org/10.1145/nn nnnnn.nnnnnnn

## 1 引言

我们正生活在数据黄金时代。国际数据公司（IDC）预测，到 2025 年，企业和个人生成的数字数据量将增长 61%，达到 175 泽字节 [137]。分析数据生态的发展使组织能够将不断增长的原始数据转化为价值，但大规模整合、管理和治理数据的困难，阻碍了组织数据分析项目的成功 [69,129]。在当前的集中式数据管理实践中，领域团队依赖中央数据管理团队来收集、处理和管理领域数据。这个中央团队正成为释放领域数据价值的瓶颈。因此，需要将领域数据的责任从中央团队转移到领域团队 [40]。

数据网格作为一种新兴的去中心化方法，通过采用面向领域的自助式设计和产品思维，实现大规模数据管理 [40]。Zhamak Dehghani 首次于 2019 [37] 定义了“数据网格”这一术语。图 1 展示了 Google 趋势中2019年1月至2023年12月间“Data Mesh”的搜索指数（月平均值）。

作者地址：Abel Goedegebuure，a.a.goedegebuure@tilburguniversity.edu；Indika Kumara，i.p.k.weerasinghadewage@tilburguniversity.edu；Stefan Driessen，s.w.driessen@tilburguniversity.edu；Willem-Jan van den Heuvel，w.j.a.m.vdnHeuvel@tilburguniversity.edu，蒂尔堡大学，Warandelaan 2，蒂尔堡，北布拉班特省，荷兰，5037 AB；Geert Monsieur，g.monsieur@tue.nl；Damian Andrew Tamburri，d.a.tamburri@tue.nl，埃因霍温理工大学，Groene Loper 3，埃因霍温，北布拉班特省，荷兰，5612 AZ；DarioDi Nucci，ddinucci@unisa.it，萨莱诺大学，Via Giovanni Paolo II, 132，菲夏诺 SA，萨莱诺，意大利，84084。

2024. 提交给 ACM 的稿件



© 所有者/作者 2024。本内容为作者版本，仅供您个人使用。不得再分发。

人们对数据网格的关注度逐渐上升，直到2022年初。此后关注度急剧增长，并总体保持在较高水平。尽管数据网格广受关注，关于这一主题的学术文献却很少。然而，随着组织对数据网格的调查日益增多，我们看到了系统梳理灰色文献的机会，以便更好地界定这一概念并推动研究议程的发展。

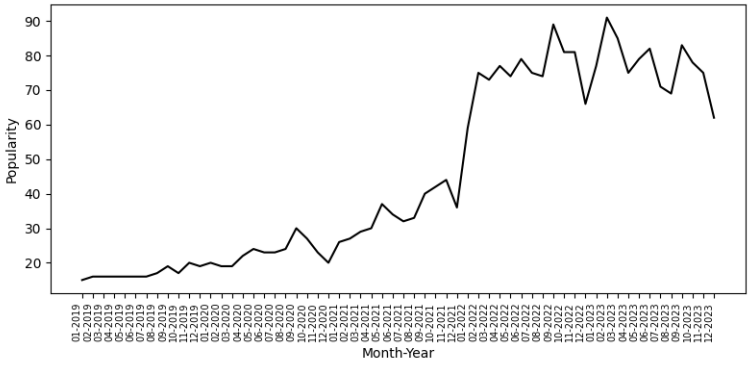



图 1. “数据网格”一词的 Google 趋势。

Zhamak [40] 根据四项原则定义了数据网格架构：数据即产品、数据的领域所有权、自助式数据平台和联邦式计算治理。组织中的数据资产被作为产品提供，使消费者能够高效、自主地发现和使用领域数据。最接近数据源头的领域团队负责创建和管理数据产品。当各领域相互使用对方的数据产品时，便形成了数据产品网格。为确保数据产品具有互操作性，并符合内部和外部法规及策略，联邦治理团队负责制定并执行适用于整个数据网格的产品互操作标准和合规政策。最后，自助式数据平台通过提供构建、维护和管理数据产品所需的基础设施及平台服务，为各领域提供支持。

在一定程度上，我们观察到数据网格与面向服务的架构（SOA）之间存在很强的相似性，尤其是与领域驱动微服务的相似之处 [37, 40, 71]。SOA 将软件应用程序视为一组松散耦合的协作服务 [118]。它倡导通过发现并调用网络可用服务，将多方现有应用程序组合起来，以此构建应用程序。微服务架构是 SOA 的一种变体。面向领域的设计和微服务架构的成功启发了数据网格架构的提出 [37, 40, 41]。数据产品通常以服务的形式对外提供，并通过 API 使用。数据网格将数据产品关联并连接起来，这与服务网格类似。产品负责人可以发布其数据产品，消费者则可以发现并连接所需的产品，以开展跨领域数据分析并构建新的增值数据产品，这与服务组合类似。

在本文中，我们报告了一项系统性灰色文献综述（SGLR），旨在识别并评审有关数据网格的灰色文献，从而综合从业者对数据网格的理解，并为学术界识别研究挑战。我们选取并分析了114篇灰色文献文章。研究结果包括：从业者对数据网格四项原则的描述、数据网格中的常见架构组件，以及数据网格的潜在优势与不足。此外，面向服务的架构（SOA）与数据网格之间的相似性促使我们采用SOA中经过充分验证的参考架构，将系统性灰色文献综述的研究结果整理为三种参考架构。我们对数据 Manuscript submitted to ACM

网格与SOA的映射还有助于研究界理解数据网格中的研究挑战，以及利用SOA研究成果应对这些挑战的潜在方式。

本文其余部分安排如下。第2节介绍我们的系统性文献综述方法论。基于灰色文献的研究发现，第3节探讨数据网格及其四项原则，第4节则介绍其潜在优势和缺点。第5节介绍我们为数据网格提出的三种参考架构，第6节列出潜在的研究挑战。第7节总结了有关数据网格的学术研究以及与本综述相关的文献综述研究。最后，我们在第8节讨论研究有效性面临的威胁，并在第9节总结全文。

## 2 方法论

由于数据网格范式较新，关于这一主题的学术文献很少。与此同时，数据网格概念源于行业，而非学术界 [37, 41],，因此有关这一主题的灰色文献（如博客文章和白皮书）十分丰富。结构化文献综述（SLR）的质量很大程度上取决于是否有足够的源材料，因此我们开展了一项灰色文献综述（GLR），以探究数据网格的最新研究进展。在本研究中，我们遵循 Garousi 等人提供的指南 [56], ，他们对 Kitchenham 和 Charters 的 SLR 指南进行了专门调整，使其适用于 GLR [75] 。按照这些方法论，我们的 GLR 分为三个阶段：规划与设计阶段、执行阶段和报告阶段。图2展示了这些阶段及其步骤的高层次概览。

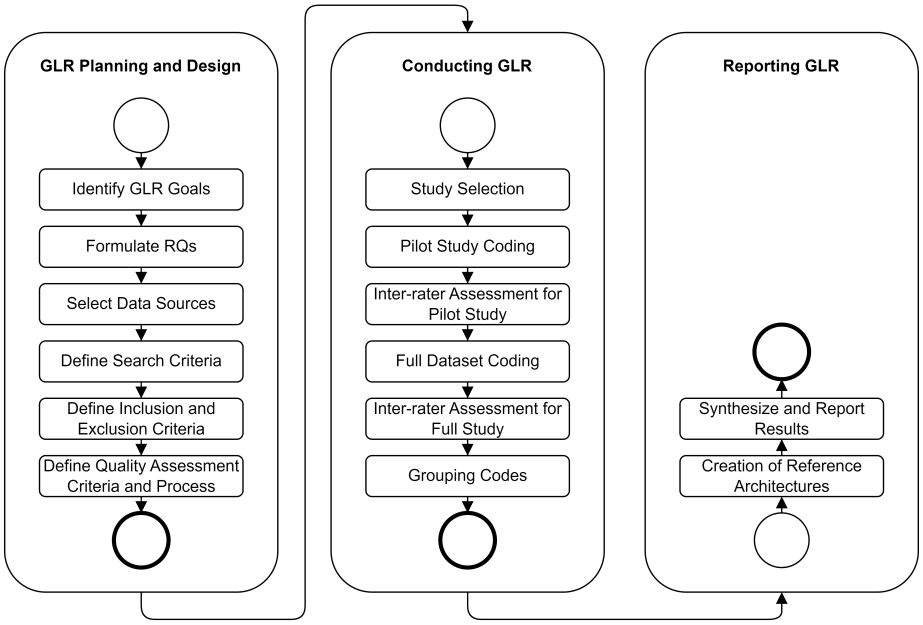



图2. 系统性灰色文献综述流程。

### 2.1 灰色文献综述的规划与设计

我们的主要目标是为数据网格概念提供统一定义，创建数据网格参考架构，并确定研究挑战。为实现这些目标，我们提出了四个研究问题：

**研究问题1：什么是数据网格？**数据网格体现了四项设计原则 [37, 40, 41]：数据即产品、以领域为驱动的数据所有权、自助式数据平台，以及联邦式计算治理。从业者对这些原则以及应用这些原则的经验可能有不同看法。因此，研究问题1旨在从从业者的角度理解并阐述数据网格的四项原则。

**研究问题2：为什么需要数据网格？**具体而言，采用数据网格有哪些益处和顾虑，以及应在何时使用？数据网格有其优缺点，并非适用于所有数据管理场景的通用方案。因此，研究问题2旨在了解数据网格试图解决的数据管理问题，以及数据网格可能带来的益处、用例、实施挑战和缺点。

**研究问题3：组织应如何构建数据网格？**一些组织已经开始探索数据网格，并分享了自身经验，包括开发和运维指南、方案及示例。因此，研究问题3旨在分析相关灰色文献文章，构建一组参考架构，指导组织在开启数据网格之旅时做出更好的设计选择。此外，鉴于数据网格架构尚属新兴事物，我们旨在将灰色文献中的发现映射到相关领域所采用的架构中，以此构建参考架构。

**研究问题4：关于数据网格，有哪些研究挑战？**从业者在灰色文献文章中报告了关键挑战；因此，研究问题4旨在根据从业者面临的挑战，确定与数据网格相关的研究挑战。

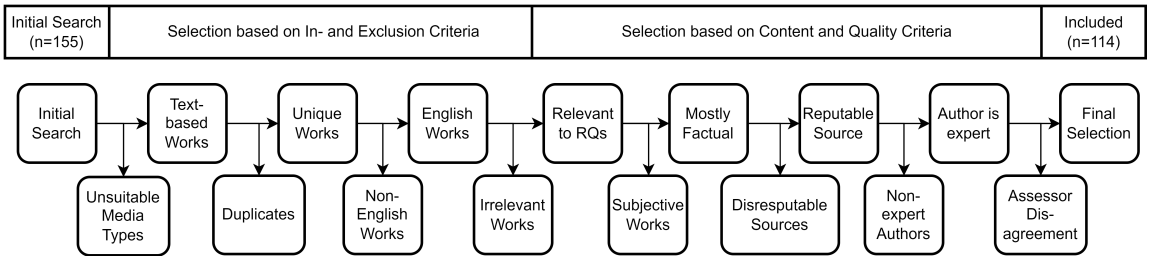



图 3. 搜索与筛选流程。各步骤依次进行，每一步都会减少候选来源的数量，直至最终选定来源。完成最后一步排除后，我们与 $\kappa = 0.79$ 进行了一致性检验，采用成熟的 Cohen Kappa 系数计算方法，以确保评审者偏差没有高得不恰当。

遵循 Kitchenham 等人的研究方案 [75], ，我们制定了若干搜索字符串，以查找相关灰色文献来源。随后，将每个字符串应用于选定的搜索引擎 Google，再根据预先制定的纳入和排除标准筛选结果，如图 3 所示。

按照灰色文献综述中的惯例，我们通过使用不同搜索字符串进行试探性查询来选择搜索字符串，这有助于判断该搜索字符串是否有效 [153]。我们从研究问题中使用的关键词入手，不断改进关键词组合，最终确定了以下查询式；该查询式为回答研究问题带来了令人鼓舞的结果。

**查询：** 数据网格（需求 | 优势 | 缺点 |关注点 | 优点 |劣势 |挑战 | 组件 | 架构 | 设计）提交给 ACM 的稿件

随后于 $1^{st}$ 2021在Google搜索引擎中执行了查询，并根据我们的纳入和排除标准检查了所有搜索结果。由于数据网格相对较新，当时的信息和数据源大多仅限于通用的数据网格原则。因此，灰色文献综述的范围扩展至2022年5月 $1^{st}$ 2022，纳入有关数据网格架构设计的数据源。此步骤共筛选出155个来自2019年至2022年的数据源。

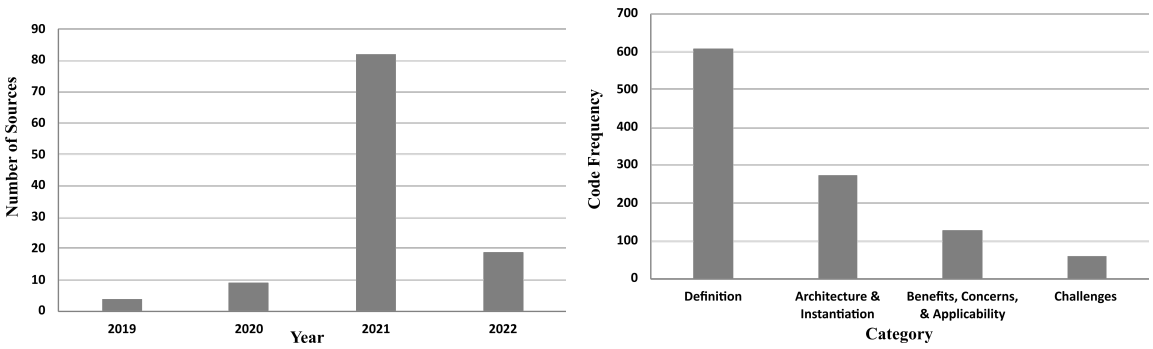
为按照扎根理论有效处理我们的来源，我们将灰色文献来源限定为文本类资料，例如报告、博客文章、白皮书、供应商的官方文档、演示文稿，以及主题演讲和网络研讨会的文字记录。此外，我们剔除了重复来源和非英语来源，以消除偏差并确保作者能够阅读这些文本。

为确保本次灰色文献综述所用来源的相关性和质量，我们参考以往的灰色文献综述，采用了若干质量标准 [21, 56, 79]。首先，我们剔除了未明确包含理论讨论或数据网格实际实施内容的文献，因为这些内容有助于回答我们提出的研究问题。前者的例子包括数据网格的形式化定义、示例架构，或对数据网格实施的优势与挑战的概述；后者的例子包括对数据网格工具或其（组成部分）实例化的讨论。

接下来，我们排除了主要由主观陈述构成、旨在表达作者观点而非提供事实信息的研究。此外，我们还核查了出版组织在信息技术领域是否享有声誉，以及作者是否是声誉良好的个人，即作者是否隶属于声誉良好的组织、具备良好的专业资历，或曾在数据管理领域发表过其他文章。

最后，两位作者依据内容和质量标准审核了每篇文章，以消除潜在的选择偏差。只有双方达成一致，文章才会被纳入最终数据集；如有分歧，作者会进行讨论。由此得到的评估者间信度评估 [139] 显示，科恩卡帕系数为 0.79，表明两位作者之间具有较高的一致性。

最终筛选出114篇文献。图4a展示了所选来源的分销情况。每个来源的完整概览可在我们的在线附录中找到（见第2.3节）。




(a) 所选来源的发表日期。 (b) 所分析研究中讨论的主题。

图 4. 来源统计。

### 2.2 开展并报告灰色文献综述

第二组步骤与依照定性分析方法开展综述有关。我们使用结构化编码来初步了解主题，随后使用描述性编码总结主题的部分内容。结构化编码用于识别围绕广泛主题的大段文本。随后，通过将基本主题概括为一个词或短语，使用描述性编码对这些主题进行更深入的分析 [140]。我们在编码过程中使用了 Atlas.ti[1]。

例如，随机选取并分析了 [79]，20 个初始来源，以建立一套初始编码。论文的前三位作者对此进行了讨论。就初始编码达成共识后，对整个数据集进行了一次编码。通过论文前三位作者的讨论，确定了编码的主题分组和类别识别。

我们首先使用从灰色文献中提取的数据，依据既定编码回答研究问题1和研究问题2。接着，为回答研究问题3，我们将研究发现映射到面向服务的计算领域中使用的参考架构 [80, 85, 116, 117]，据此创建了三个参考架构。论文前三位作者通过一系列会议协作创建这些参考架构。其他作者对这些架构进行了验证。我们通过讨论解决了任何分歧。我们采用类似的方法回答研究问题4。我们将从业者在灰色文献文章中提到的困难映射到面向服务的计算 [117, 118, 174] 和数据管理 [5, 45, 69, 152]领域的相关研究问题，从而识别潜在的研究挑战。

本文其余部分将呈现系统性灰色文献的研究结果。如图 4b 所示，我们定义了四个主要类别。在线附录中，我们详细介绍了各类别，包括主题、底层原子编码，以及涉及这些编码的来源。

### 2.3 复现包

我们已将相关数据在线公开[2]，以便进一步验证和复现本研究。其中包含完整的来源清单、使用*Atlas.ti* 工具完成的定性分析（即编码、分组和分析），以及编码表。

## 3 研究问题1：什么是数据网格

简而言之，数据网格是一种面向领域的去中心化架构，用于大规模管理（分析）数据。它将组织中单体式的分析数据空间拆分为与业务领域相一致的数据领域。通过这种拆分，管理数据并提供高质量数据和有价值洞察的责任，便从传统的中央数据团队转移到深入了解数据的领域团队。

Zhamak [37, 40, 41] 以四项原则定义数据网格概念：(i) 数据产品思维，(ii) 数据的领域所有权，(iii) 联邦式计算治理，(iv) 自助式数据平台。根据灰色文献的研究发现，第 3.1 至 3.4 节旨在探讨从业者对这些原则的看法，以及他们如何拓展或实施这些原则。

### 3.1 原则1：数据即产品

这一原则将产品思维应用于分析数据，将其作为宝贵资产提供给潜在的数据消费者。

[1]https://atlasti.com/ [2]
https://github.com/IndikaKuma/DataMesh 提交给 ACM 的稿件

数据产品的构成要素。Zhamak [37, 40, 41] 将数据产品定义为一种架构量子（一个可独立部署、高内聚的组件，包含实现其功能所需的所有结构要素 [53]）。数据产品由其所有关键要素组合而成：1）数据和元数据；2）用于摄取、处理、提供和治理数据的代码；3）用于构建、部署和执行代码，以及存储数据和元数据的基础设施。数据产品的元数据可能包括产品负责人、访问端点、数据模型、访问策略、数据质量指标和数据血缘等信息。数据产品是数据网格中的一个节点，也是一个自主入口，它从一个或多个来源接收或获取输入数据，对接收到的数据进行某种计算/转换，并生成和提供消费者所需的结果。

尽管大多数灰色文献普遍认同 Zhamak 将数据产品定义为架构量子，但有些来源 [67, 108]将精选数据集（不含代码和基础设施）视为数据产品。还有多个来源提出了数据契约 [19, 71, 126, 155, 156, 177] 这一概念，用来表示数据产品的接口。数据契约可以包含数据产品所提供数据的数据访问端点，以及其语法、语义、质量和使用条款。从业者提出，可通过存储相关元数据或使用开放 API [3]等标准将数据契约定义为 API 来实现。

Zhamak 建议采用基础设施即代码（IaC）方法，自动为数据产品配置基础设施 [40]。灰色文献也强调了 IaC 的重要性 [4, 38, 87, 130, 177] ，并提供了多个使用 IaC 部署数据产品的示例 [95, 108, 115, 126]。IaC 使用源代码来管理和配置基础设施资源，并部署和配置应用程序，从而支持应用软件工程工具和最佳实践 [101]。

数据产品的类型。 Zhamak [37, 40, 41] 提到了三种数据产品：源对齐、聚合型和消费者对齐。灰色文献也使用相同的术语对数据产品进行分类。源对齐数据产品 [71, 97, 131, 143, 157, 160, 162, 166] 从运营数据库、外部应用程序接口和传感器等源系统中摄取并处理数据。处理后的数据随后作为产品提供，由组织的最终用户或下游数据产品使用。聚合型数据产品 [36, 71, 78] 整合上游数据产品中的数据，为下游使用者生成增值数据。与聚合型数据产品相比，消费者对齐数据产品 [71, 96, 122, 130, 143, 157, 159, 162, 166] 通常不会作为数据产品出售给其他领域；相反，它们为组织的最终用户用例提供数据。

图5展示了分布在各个领域中的数据产品。一些数据产品，例如产品搜索、订单和发票，属于源对齐产品。它们主要从底层运营数据源（例如产品和订单数据库）摄取原始数据，对其进行清理和转换，并在本地存储。它们还提供接口，将本地域数据提供给其他数据领域。聚合型数据产品，例如360度客户视图和产品推荐，会聚合一个或多个领域的数据，并将其作为增值产品提供给一个或多个领域。一些数据产品，例如漏斗分析和CRM仪表板，主要服务于最终用户（即消费者对齐数据产品）。

数据产品特征。Zhamak最初为数据产品提出了六项属性 [37] ，后来又增加了两项属性 [40]。我们尝试从灰色文献中识别从业者对这些属性的看法。

**可发现性**。 数据消费者应能够轻松发现组织中的数据产品。灰色文献中提出的一种常见方法是使用中央数据/产品目录来存储数据的元数据

[3]https://www.openapis.org/

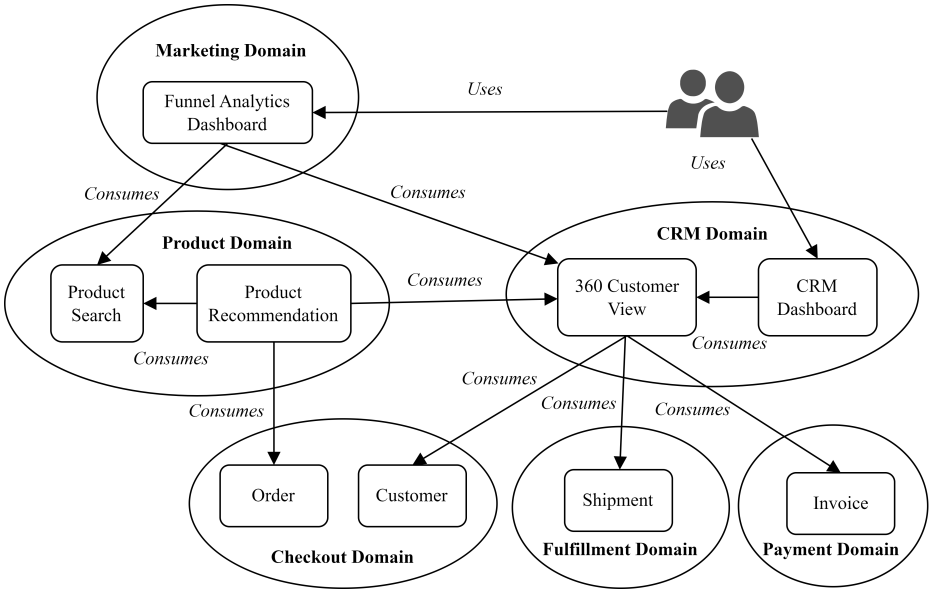



图5. 不同领域中的数据产品示例，改编自 [71]。

产品 [4, 9, 14–18、27、31、33、58、64、71、72、74、91、92、100、102、103、107、108、112、113、115、120、124、126、127、130、134、143–146, 158, 166, 170, 171, 175, 178]。数据目录可以让消费者使用元数据搜索数据产品。然而，Zhamak [40] 不建议使用由独立团队管理的中央数据目录，因为数据产品负责决定并提供其可发现性信息。相反，数据产品可以提供发现接口，用于在数据网格层面构建产品搜索和探索能力。

**可互操作**。应当能够（相对）轻松地将来自不同数据产品的数据组合或集成为复合数据产品。在各数据产品中一致地采用开放标准和约定，可以提高数据产品之间的互操作性。灰色文献提到了数据网格架构应涵盖的若干标准和实践 [14, 15, 18, 27, 31, 55, 71, 130, 135, 144, 178, 180]。例如，用于描述数据产品及其组合的开放标准，包括接口、元数据以及（访问和管理）策略，可以确保组织内部及跨组织的数据产品具有互操作性。此外，标准元数据字段和业务术语表模型可以确保所有数据产品都得到明确无歧义的描述。最后，通用的数据质量建模方法可以确保所有领域都依据相同的关键绩效指标（KPI）来定义其产品的数据质量。

**可寻址**。 数据产品的接口应提供唯一、可通过编程访问的端点。接口包括用于数据摄取与提供、数据产品发现与管理的接口 [40]。灰色文献主要关注数据访问端点，并提及了 Web 服务 API [11, 65, 67, 112, 113, 130, 156, 177, 178], 、存储桶 [9, 28, 100, 108, 110–112, 114, 130, 143, 170, 171], 、发布-订阅消息系统中的主题 [11, 14, 36, 67, 71, 110–113, 115, 125, 126, 130, 142, 143, 151, 178], 以及 SQL 端点 [7, 61, 67]。

**原生可访问**。 Zhamak [40] 强调，数据产品应提供多个端点来消费其数据，以便消费者可以使用自己偏好的方式访问数据产品。同样，灰色文献提到，数据产品需要支持多语言数据输出接口 [28, 38, 91, 112, 113, 145, 160]。**易于理解**。 数据产品的数据模型和自然语言文档应帮助潜在消费者理解产品的语义、语法和行为/用法 [40]。例如，数据产品的文档可以包括该产品消费和生成的数据中的实体、这些实体之间的关系、与数据网格中其他数据产品之间的关系，以及有关访问模式和如何请求访问数据的信息 [3, 38, 55, 86, 131, 151, 178]。有关产品负责人和主题专家的信息可以确保消费者在需要了解更多与数据相关的信息时，知道应联系谁。所有数据视图的模式都可用于描述数据结构。服务级别目标（SLOs）可以描述产品的质量，并充当数据契约。数据血缘和来源信息的元数据可以展示数据的来源，并让消费者信任数据产品，因为数据是如何创建和处理的变得清晰可见。文献表明，使用通用的数据建模语言或标准有助于增进对数据产品之间交换的数据/消息的理解 [151, 178]。

**安全**。 数据产品必须实施必要的安全策略，确保消费者能够以安全且符合监管要求的方式使用该产品 [40]。这些策略包括但不限于数据访问控制、静态数据加密、传输中的数据加密、数据机密性和数据保留。在数据网格架构中，安全策略主要由中央统一管理，但以分散的方式执行 [40]。Zhamak [40] 及其他从业者 [71, 112, 113, 135, 155, 156] 建议采用策略即代码 方法，即使用高级语言编写代码来定义和执行策略，以管理数据产品的安全策略。

**可信**。 数据消费者应当能够信任数据产品提供的数据。灰色文献提供了若干提升数据产品可信度的指南。产品负责人可以使用联邦治理团队提供的方法论，对数据质量进行建模，并从多个维度（如完整性、准确性和及时性）定义其产品的质量 [70, 103, 108, 177]。此外，自助式平台可以支持产品开发人员在产品生命周期内确保质量（例如，自动化监控数据质量问题并执行纠正措施） [10, 151]。产品负责人和消费者还可以使用正式契约，定义一组服务级别目标（SLOs），以反映每个数据产品的目标质量 [91, 131, 134, 156]。最后，提升数据可信度的另一种方法是向产品使用者提供数据来源信息和数据血缘信息，使他们能够检查所用数据的来源及其历史 [28, 166, 178]。

**本身具有价值**。 产品提供的数据应对组织中的其他部分具有长期价值。即使不与其他数据产品聚合，数据产品也应支持有意义的用例 [40]。不过，灰色文献大多提供的是这样的用例：某个领域的数据与其他领域的数据结合或关联后，便对组织产生价值，例如，用户行为数据与销售信息 [143]，以及智能设备数据与企业数据 [177]。应持续监控和维护数据产品的价值。灰色文献提到采用数据运维方法 [105]，以自动化向消费者持续交付高质量数据产品 [87, 113, 130, 135]。

### 3.2 原则 2：数据的领域所有权 a

在传统数据架构中，中央数据团队通常负责管理和共享分析数据。相比之下，数据网格将数据的所有权分配给最接近稿件提交至 ACM 的团队或个人（属于组织中的领域），这些团队或个人

最了解数据并熟悉数据的细节 [40]。自主领域团队负责基于领域数据构建和管理数据产品。Zhamak [37, 40, 41] 倡导应用领域驱动设计（DDD）方法 [51] ，将组织中的分析数据分解为与组织单元相一致的数据领域。

领域划分。关于使用领域驱动设计识别数据领域及其数据产品，灰色文献与 Zhamak 的观点一致 [38, 48, 71, 78, 86, 92, 95, 96, 112, 113, 115, 122, 127, 130, 151, 157, 162]。领域驱动设计将组织的业务职能和业务能力分解为多个领域，从而能够基于底层领域的模型来设计软件系统 [51]。领域模型使用领域专家、用户和开发人员共享的词汇（即抽象）来描述领域的某些方面。组织已经通过将自身划分为不同的业务领域（如产品、销售和市场营销）来实现职责分离。可以围绕这些业务领域产生的运维数据和分析数据构建数据产品。因此，业务领域构成了分配数据所有权的逻辑单元，并且可以进一步分解为子领域，以便更好地管理其复杂性。第5.2节讨论了如何应用领域驱动设计来构建数据产品。

职责。Zhamak [37, 40, 41]指出了领域团队的两项关键职责：构建和维护数据产品，以及在领域层面开展治理活动。灰色文献也提到了类似的职责 [11, 22, 28, 31, 55, 61, 67, 74, 91, 97, 99, 102, 103, 108, 111, 112, 114, 125–127, 135, 141, 144, 146, 150, 162, 164–166]。领域团队对领域数据有深入了解，包括数据的潜在用例和消费者，以及维护数据质量的流程和技术。因此，他们能够确定哪些领域数据资产应作为数据产品提供。他们可以使用自助式数据平台提供的工具和服务来创建、管理和提供数据产品（参见第3.4节）。领域团队还会开展治理活动，以确保数据产品符合相关法规、组织政策和标准的合规性与符合性（参见第3.3节）。

### 3.3 原则3：联邦式计算治理

数据网格采用去中心化治理模型，由自治领域管理其数据产品 [37, 40, 41]。数据产品网格层面的全局治理主要旨在确保数据产品之间的互操作性，并落实标准化规范和全组织政策 [14, 19, 22, 28, 36, 47, 55, 70, 74, 81, 87, 96, 102, 103, 109, 135, 146, 158, 164, 165, 178]。在联邦式治理模型中，平衡集中化与去中心化的程度可能颇具挑战。一方面，集中化能加强控制，以确保互操作性和合规性；另一方面，各领域需要足够的自主权，才能构建与其业务相契合的数据产品。

全局治理。 我们将联邦治理团队在数据网格层面开展的治理活动称为全局治理。联邦治理团队由不同类型的代表组成，例如领域专家、主题专家、治理专家、法务部门成员、平台团队成员以及管理层成员 [40]。常见的全局治理活动包括：

**定义和执行组织级标准**。 标准的一个关键目标是提高数据产品的互操作性 [37, 40, 41]。标准还可用于确保数据产品符合联邦治理团队制定的安全与隐私要求。Zhamak [40] 强调了数据产品接口、数据与查询建模以及数据血缘的标准化。灰色文献提出了几项标准：用于产品元数据建模的 W3C 数据目录词汇表 [144], 用于事件数据的 CloudEvents [97],提交至 ACM 的稿件

用于产品接口的开放 API [65, 71], 用于数据建模的 JSON 模式 [71, 135], 以及用于机器学习模型的 ONNX（开放神经网络交换） [71]。

**定义并执行数据治理政策**。 数据治理政策规定规则或指南，以确保在组织内一致且正确地收集、存储、访问、使用和管理数据资产。Zhamak [40] 建议将法规合规、用户身份管理和产品所有权管理等跨领域关注事项纳入网格级治理政策。多篇灰色文献介绍了使用集中式身份与访问管理（IAM）提供方来管理用户及其权限 [92, 108] ，以及使用数据目录自动发现并分类敏感数据的示例 [92, 108, 126]。

**制定用于定义和评估数据质量的方法论**。 联合治理团队还负责制定一种方法论，以定义数据产品负责人应执行的数据质量要求 [40]。根据文献来源 [70, 103, 124, 135, 151], ，标准化的方法论和工具可以确保所有数据产品依据一组共同的质量维度定义质量，并采用共享的定义。缺乏一致的数据质量执行机制会导致数据沼泽 [115, 127]。从业者建议采用统一格式来表示数据质量约束和结果 [164] ，并使用元数据规则，确保在各领域之间一致地建模共享概念 [91]。

**维护通用业务术语表**。 组织会使用业务术语及其定义构成的业务术语表，确保在分析数据时，整个组织采用相同的定义。数据网格原则不鼓励为整个组织创建单一的规范数据模型，因为这与去中心化背道而驰 [40]。不过，一些灰色文献来源认识到，在整个组织内共享所用概念的定义（例如客户和货币代码等主数据和参考数据）十分重要，因为这有助于提升数据产品的互操作性 [36, 154, 158, 161]。

**监控数据网格**。 联邦治理团队应持续观察和评估数据网格的整体健康状况，并采取必要的干预措施。Zhamak [37, 40, 41] 讨论了收集各项指标，例如更改数据产品的交付周期、数据产品消费者的数量以及数据质量维度。这些指标可用于确定数据产品的价值，并采取诸如提升或降级产品等干预措施。灰色文献来源还强调了定义和衡量关键绩效指标（KPI）的重要性，例如数据产品的互操作性、合规性和使用情况 [11, 49, 71, 102, 166]。

**创建激励模型**。 资源配置数据产品会增加领域团队的日常工作量，而且可能不符合他们传统的工作职责 [135]。例如，供应链部门的主要任务是优化配送路线，可能不太愿意向其他领域提供高质量的数据。若没有适当的激励，领域可能看不到构建数据产品的理由。因此，实施数据网格架构或向其转型的组织应制定激励模型，鼓励各领域将其数据作为产品进行资源配置，并持续治理所提供的产品 [37, 40, 41]。多篇灰色文献强调，需要协调数据生产者与数据消费者的激励，并鼓励数据生产者生成评估数据产品质量和可信度所需的运营元数据 [37, 41, 61, 67, 87,108, 151]。

本地治理。 我们将产品/领域团队在产品/领域层级开展的治理活动称为本地治理。产品负责人可能与其他领域成员协作，负责开展本地治理活动。Zhamak [37, 40, 41] 指出了领域层级的若干治理活动

：领域数据建模、数据质量保障，以及数据访问控制策略的执行。灰色文献提到了本地治理团队类似的职责：

**管理产品数据模型**。数据产品团队掌握领域知识，因此能够准确地对产品数据进行建模 [61, 67, 87, 91, 108, 151]。此类模型的模式需要相应演进，以反映领域数据源的变化 [10, 108]。**管理数据访问**。领域专家充分了解数据，因此最能判断谁应该以及不应该访问哪些数据（或数据的哪些部分） [78, 91, 112, 130]。此外，收集数据血缘、实施数据保留策略以及加密数据等任务，也可以在产品级实施 [108, 120, 130]。**管理合规性与符合性**。与管理访问控制类似，要在产品的整个生命周期内确保合规性与符合性，需要深入了解领域数据 [40, 78, 120]。灰色文献中提到的示例包括遵循数据建模标准 [22, 146] 和互操作性标准 [124, 175], , 以及确保隐私合规 [120, 130]。**管理数据质量**。产品团队使用联邦治理团队制定的质量评估方法，根据一组质量维度确定产品的目标质量 [40]。灰色文献中提到的数据质量测试包括衡量数据质量维度、验证模式，以及检查是否违反领域特定的业务规则 [91, 107, 146, 151, 166]。**监控数据产品健康状况**在产品的整个生命周期内，产品负责人会监控其运行状况，并确保其具备满足消费者需求的能力 [49, 71, 102]。灰色文献提到了若干用于评估数据产品运行状况的指标：数据质量指标、正常运行时间、性能和成本 [11, 49, 102]。

治理活动自动化。 在联邦式计算治理模型中，自动化发挥着重要作用，可确保该模型在整个组织中有效扩展 [37, 40, 41]。根据灰色文献来源 [71,107, 130], , 自动化是应对数据治理挑战的关键；这些挑战源于法规频繁变更，以及产品使用者和数据源数量的增加。文献还列举了可自动化的数据治理操作示例 [71, 107, 130, 135, 166]: 数据分类（例如，敏感度标签）、数据匿名化和加密、数据质量检查、为确保合规性而进行的数据保留和删除，以及自动化监控，包括告警和自动响应（例如，扩展资源以满足需求）。

### 3.4 原则4：自助式数据平台

数据网格倡导构建领域无关基础设施和平台服务，并以自助服务的方式提供这些服务，以赋能参与创建、部署和维护数据产品的不同角色 [37, 40, 41]。基础设施服务用于配置和管理计算资源，例如虚拟机、网络资源和存储磁盘。平台服务则支持数据产品、治理应用及其组件（例如数据管道、机器学习管道和微服务）的设计、开发、部署和管理。

目标。 根据 Zhamak [37, 40, 41], , 自助式平台旨在优化数据网格利益相关者（例如产品开发人员、消费者和治理团队）执行各自任务的体验，即开发、使用和治理数据产品。灰色文献与 Zhamak 的观点一致。首先，自助式平台可以降低产品开发人员所需的专业化程度，因为他们无需深入了解如何管理基础设施和平台服务 。提交给 ACM 的稿件

其次，由于各领域无需重复投入精力来管理平台服务 [11, 28, 55, 119,132, 145, 150, 151]，因此可以提高开发和管理产品的效率。最后，共享平台服务可以提升数据产品的一致性，减少领域之间的摩擦，并促进数据产品之间的互操作性 [74, 81, 121, 166]。

构建自助式数据平台。一支由高度专业化的基础设施和平台服务开发人员组成的团队负责构建和维护该平台 [40]。中央平台团队必须与所有领域保持密切联系，以识别并确定平台服务、其功能及接口的需求优先级 [78, 104]。灰色文献建议使用基础设施即代码（IaC）蓝图，简化通过自助式平台配置数据产品的流程 [28, 38, 93, 108, 121, 130, 145, 155, 158, 177]。IaC 模板包含基础设施资源和平台服务的预定义配置，并将联邦治理团队制定的安全措施、策略和标准集成其中。例如，数据平台团队可以针对不同的数据产品类型（如数据管道、机器学习管道和机器学习模型服务）以及不同的产品组件类型（如使用 Kafka 消息代理的数据摄取组件、使用 Google 对象存储的数据存储组件，以及使用 AWS SageMaker 的训练管道执行器组件）开发 IaC 模板。产品开发人员可以找到合适的 IaC 模板，根据需要进行定制，并使用更新后的模板来配置和管理其数据产品。

自助式数据平台的组件。Zhamak 提议将平台从逻辑上分解为三个层：数据基础设施实用层、数据产品体验层和网格体验层 [40, 41]。这些层旨在封装管理基础设施和平台资源、构建数据产品以及治理数据网格所需的业务能力。一些灰色文献来源讨论了这一自助式平台的逻辑架构 [131]。不过，许多从业者都提供了平台组件的示例。

**基础计算资源**。平台应提供计算资源、网络资源和存储资源，包括虚拟机、网络设备和对象存储。文献来源大多提到使用公共 IaaS（基础设施即服务）提供商提供的资源 [7, 71, 92, 108, 113, 143, 150]。

**多语言数据存储**。关于数据存储，从业者 [28, 38, 70, 72,87, 103, 130, 145, 169, 175] 和 Zhamak [37, 40]的普遍建议是采用多语言数据存储，使系统能够透明地使用多种数据存储技术。数据产品可能包含需要不同数据存储的数据类型。例如，某个数据产品可能由高度结构化的表格数据组成，并存储在关系数据库中。相比之下，其他数据产品可能由高度非结构化的数据组成，最适合存储在对象存储中（例如视频或图像）。因此，平台必须支持多语言数据的存储和提供。

**数据产品组件服务**。 数据产品通常由数据摄取、数据（ETL）管道、机器学习训练管道、机器学习模型存储、机器学习特征存储、消息代理和 API 等组件构成 [10,42, 87, 91, 103, 111, 126, 130, 145, 169, 177]。产品开发人员应能够使用平台服务来实现和管理此类产品组件。例如，开发人员应能够以自助服务的方式使用平台提供的管道编排工具，开发、部署、调度、执行和监控数据管道及机器学习管道。

**元数据存储库/数据目录**。 产品、其组件及其制品（例如，机器学习模型和数据库表）都包含各种元数据，因此，平台需要提供一项服务来生成、存储和管理元数据。此外，还应提供公司级元数据，例如业务术语和参考数据的标准化定义，以便产品开发人员将其纳入数据产品中 [154, 156, 161]。如前所述，

第3.1节中，数据目录可用于开发产品目录服务，该服务存储产品元数据，以实现数据产品的发现。

**联邦查询引擎**。 自助式平台应提供联邦查询引擎，以查询来自不同领域和数据源的数据，使产品聚合者能够轻松创建复合数据产品 [61, 71, 87, 130, 169]。**监控**。 平台应提供工具，使领域团队能够监控其数据产品，并使联邦式全局治理团队能够监控整个数据网格 [40]。这些工具应包括但不限于能够监控运行状况、数据血缘、成本以及与数据质量指标相关的性能的功能 [11, 49, 102]。它们还应支持日志记录、告警和自动化干预（例如，扩展资源） [74, 102]。最后，监控工具必须与其他相关平台服务集成，例如，为数据管道和机器学习管道添加监控功能 [10]。

**产品生命周期管理**。 平台应提供相应业务能力（例如，产品版本控制、源代码管理和测试），以便产品开发人员管理数据产品的生命周期 [28, 71, 91, 102]。开发人员在部署和运维数据产品时应能够采用 DevOps [46] 实践 [169, 180]。例如，他们应能够使用持续集成与持续交付/部署平台服务，以原子方式构建和测试产品所用的数据管道和 API，并通过自助式平台的管道编排和容器托管服务部署这些管道和 API。

**安全与隐私**。 平台服务必须支持实现数据产品的安全与隐私要求 [37, 40]。例如，平台可以提供数据加密服务，以保护静态数据和传输中的数据 [78, 108, 120]。此外，还可以提供身份与访问管理（身份与访问管理）服务，用于管理用户和产品身份，并为数据产品及其他平台服务实施访问控制策略 [71, 92, 112, 120, 134, 145]。

**策略执行**。 平台应提供工具，使本地和全局联邦治理团队能够定义、存储、关联、监控和执行治理策略 [37, 40]。如第3.1节所述，灰色文献建议采用策略即代码 方法来自动执行策略。因此，平台需要支持必要的工具，例如开放策略代理（开放策略代理）引擎 [112, 113]。此外，平台还必须提供支持策略实施的服务，例如数据匿名化服务和数据质量指标计算服务 [38, 42, 71, 115, 126]。

**商业智能（BI）工具**。 商业智能仪表板和报告应用程序是常见的消费者对齐数据产品类型 [71, 72, 157]。因此，该平台应能让业务用户以自助方式探索数据、创建可视化、生成洞察并创建报告 [3, 91, 130, 169, 177]。

## 4 研究问题2：动机、优势、顾虑与适用性

本节介绍采用数据网格方法的动机、其被认为的优缺点，以及决定组织是否采用该方法的因素。

动机。 数据网格是针对集中式单体数据管理方法而开发的直接应对方案。这类方法长期以来一直是行业标准，例如数据仓库和数据湖。根据本综述所考察的文献，我们发现有三个因素促使组织在自身组织中采用数据网格范式。

**随数据源扩展**。随着组织内数据源数量不断增加，要全面掌握各类数据源并确保其提供高质量数据，变得愈发困难。大数据兴起期间，这一问题尤为突出，也成为开发数据湖以取代数据仓库的主要动因之一。

尽管数据湖已被广泛采用，灰色文献表明，组织仍难以发现并提供组织内的所有数据 [25, 28, 49, 64, 103, 163],，使各类数据源实现可互操作 [25, 49, 113, 127]，并在数据源头确保数据质量和一致性 [25, 70, 127, 165, 166]。数据网格提出，应在数据源头解决这些问题，而非集中处理。

**集中式团队成为瓶颈**。 数据源数量不断增加的另一面，是组织中数据驱动用例也日益增多。通常，这些用例所需的数据由一支高度专业化的数据工程师组成的集中式团队提供，该团队还负责管理用于聚合数据的单体数据平台。然而，依赖集中式团队会阻碍数据交付的可扩展性和敏捷性，而这正是满足各类数据消费者多样化需求所必需的 [12, 19, 36, 70, 71, 102, 111, 127, 133, 136, 142, 150, 165, 166]。

**提供方与消费者专业知识的分离**。 另一个问题（前述问题使这一问题更加严重）是，在大多数组织中，数据提供方专业知识与数据消费者专业知识彼此分离。有效利用数据需要深入理解数据本身及拟议用例。依赖集中式团队同时理解提供方领域和消费方领域的背景，会进一步加剧现有瓶颈；而数据提供方和数据消费者则可以依靠各自对所属领域的了解 [55, 61, 108, 150]。

优势。 我们从灰色文献中归纳出数据网格相较于传统数据架构的以下潜在优势。我们简要讨论每项潜在优势与上述推动数据网格转型的问题之间的关系。

**可扩展性**。 数据网格架构比集中式数据架构更具可扩展性，因为它减少了对中央管理平台团队的依赖，使组织能够横向扩展（即增加数据源）和纵向扩展（即增加消费者），而不会形成瓶颈 [19, 42, 47, 49, 70, 102,108, 109, 120, 143, 164]。

**敏捷性提升**。 上述优势使组织更加敏捷，因为它们减少了对集中式团队的依赖，并让数据提供方与数据生产者直接接触。消除瓶颈后，组织可以更快地交付新的数据应用，从而迅速响应不断变化的需求和目标 [58, 70, 76, 120, 164]。总体而言，组织可以创建更多数据驱动应用，释放数据的价值 [119]。

**更高的数据质量**。 通过让最接近数据源头的人员负责将数据作为数据产品提供，便确立了对数据质量的问责机制 [19, 64, 72, 87, 108, 159, 162, 177]。领域团队可以运用其领域专业知识，构建对不同客户有价值的数据产品 [25, 110, 111, 145]。联邦治理团队会制定一套全局标准和策略，要求各领域在创建数据产品时遵守 [55, 70, 164]。

**更好的数据发现** 在数据网格中，领域团队受到激励，会尽可能提高领域数据产品的使用率，因此也会让产品更容易被发现 [19, 58, 63]。因此，他们会创建自描述数据产品，并将产品元数据发布到中央产品注册表／目录中，以便轻松发现 [72, 162]。提升数据（产品）的可发现性，有助于组织有效扩展其生态系统中的数据规模。

**更好的治理**。 数据网格还能够实现更好的大数据治理。通过将治理责任分配给接近数据来源的领域，去中心化方法更适合治理大量数据。稿件提交至 ACM

由于数据提供方与数据消费者直接接触，因此他们能够更好地判断谁应该或不应该在特定条件下使用某些数据，并帮助组织遵守外部法规 [19, 96, 120]。

**缩短数据交付周期**。最后，各领域可以独立配置并维护数据产品，消费者也可以通过自助服务模式独立使用这些产品，无需依赖任何集中式团队。此外，自助服务平台赋能了各领域和联邦治理团队，大幅缩短了整个组织的数据交付周期 [14, 28, 70, 110, 165]。

顾虑与缺点。数据网格也面临着一系列挑战。下面概述灰色文献中指出的数据网格主要顾虑。

**变革管理**。采用数据网格架构会给组织带来技术和组织层面的挑战 [61, 70, 164]。组织已经建立了数据管理流程和层级组织结构，而这些很可能需要进行调整 [61, 120, 127, 162]。员工可能会对变化感到不安并产生抵触。为克服这一挑战，组织可以围绕文化转变（工作方式）对团队进行培训 [127, 146], , 向利益相关者说明整体愿景以及转向数据网格的总体益处 [103, 146, 162],, 并为现有业务领域团队配备数据工程师 [146]。组织还可以组建赋能团队，与领域团队密切合作，帮助其接入和使用数据，并培训和指导团队践行数据网格最佳实践 [71, 159]。

**人才短缺**。 数据网格要求领域团队开始独立运作，因此团队需要具备必要技术技能（例如数据工程和机器学习）的成员。自助式平台的一个重要目标是赋能技术专长较少的开发人员开发数据产品，但构建自助式平台服务可能颇具挑战，需要投入时间和精力进行研究与开发。不幸的是，经验丰富的数据工程师十分短缺 [48, 49, 61, 141, 165], , 这也是数据仓库和数据湖架构中常见的瓶颈 [165]。

**数据重复**。 随着各领域将数据重新用于复合数据产品中的新用例，源领域的数据可能会被复制和重复。尤其是在大规模实施时，数据网格可能会增加数据管理成本 [47, 67, 70, 120, 159, 164]。此外，由于数据存储在组织内不同位置，数据治理将变得更具挑战性 [49, 141, 159]。从业者提出的解决方案包括创建数据虚拟化技术（例如零拷贝数据访问），以便在不复制数据的情况下摄取数据，以及开发开放协议，以安全、高效地共享数据 [62, 127, 141]。

**工作重复**。 尽管自助式平台试图减少与数据平台管理相关的工作重复，但与由一个团队负责集中式数据相比，在不同领域构建和维护数据管道仍会存在一定程度的工作重复和技能重复 [47, 55,102, 104, 127]。

**技术债务**。 由独立团队开发的多个数据产品和数据管道可能不遵循一套通用的软件开发最佳实践和原则 [70, 135, 164]。在开发数据产品时采用不当实践可能导致巨大的技术债务。因此，成熟的联邦式治理模型可以通过准确识别并强制执行标准、最佳实践等，减少技术债务 [40, 70, 135]。

**复杂性**。 数据网格具有分布式和去中心化的特性，因此需要管理和协调许多彼此独立且各不相同的单元，从而增加了整体复杂性 [38, 112, 159]。数据网格还可能给稿件引入 ACM

采用多云方法的组织面临更多互操作性挑战 [127]。随着自助服务平台和计算治理模型日趋成熟，复杂性可能会降低 [40]。

适用性。 灰色文献指出，在若干情况下，数据网格比集中式数据架构更适用。

**大型且多样的数据格局**。 数据格局庞大，且数据提供方和使用方众多的组织，可能会受益于去中心化架构 [4, 70, 73, 83, 99, 102, 103, 109, 136, 145, 164]。此外，当组织拥有多样化且经常随时间变化的数据集时，数据网格可能比具有众多点对点连接的集中式方法更适用 [38, 55, 73, 99]。

**对敏捷性的需求**。 如果组织的数据平台和中央数据供给团队削弱了其创新能力，而组织又需要缩短数据应用的上市时间，那么采用数据网格架构可能会带来益处 [42,70, 99, 102, 127, 136]。此外，如果组织需要频繁进行数据实验，那么数据网格比集中式架构更有优势，因为它消除了对规模过大、负荷过重的数据供给团队的依赖 [12, 102]。

**更好地进行数据治理的需求**。如果缺乏数据所有权和治理模型，集中式数据湖很快就会沦为数据沼泽。对于需要明确数据所有权并基于领域知识对数据实行严格治理的组织，数据网格可以解决这些问题 [55, 103, 145]。此外，面临许多外部法规的组织可以利用数据网格确保合规性，因为数据网格有助于改善数据治理 [52, 108,120]。

数据网格可能并非适用于所有组织的最实用数据架构。灰色文献提到了若干情形，在这些情形下，数据网格方法可能并非最合适的选择。

**缺乏领域导向**。数据网格预设组织采用由相互独立的跨职能团队组成的组织结构，并使这些团队与业务领域的结构相对应。如果组织（例如，通常是中小型企业）不遵循领域驱动设计原则，也没有多个能够独立构建数据产品的团队，那么数据网格方法可能不适合该组织 [64, 71, 73, 127, 136, 159]。**数据治理成熟度低**。数据网格需要成熟的数据治理支持，以防止数据孤岛，并确保自治领域的数据处理符合监管要求。因此，对于数据治理成熟度较低的组织（例如，没有自动化的数据质量评估或法规遵从性检查），数据网格可能并不适用 [42, 71, 73, 146, 159]。**数据平台成熟度低**。自助式数据平台是降低数据产品所有权成本、并在各领域实施数据治理的关键。对于缺乏经验丰富的数据平台工程师的组织而言，构建自助式数据平台可能并不切实际 [71, 73]。**缺乏数据用例**。如果一个组织只有少量有用的分析型数据源和少数几个数据用例，那么转向数据网格所带来的成本和风险将远远超过其收益 [42, 71, 73]。

## 5 研究问题3：组织应如何构建数据网格?

本节提出一组参考架构，详细描述数据网格架构。我们借鉴了面向服务的计算领域中使用的相关架构，以及灰色文献综述所得的洞察。

### 5.1 数据网格中能力与角色的分层架构

本节介绍一种分层模型，用于从逻辑上组织数据网格架构中的不同构成要素、角色和职责。我们将灰色文献研究中的相关发现映射到 xSOA 模型 [117, 118], 中，该模型旨在对复杂的面向服务的架构应用程序的业务能力进行分组和逻辑结构化。图6展示了数据网格的分层模型。分层突出体现了以下不同需求：i）提供实现和使用数据产品所需的基础设施和平台服务；ii）创建、使用和组合数据产品；以及 iii）在数据产品的整个生命周期中对其进行管理。需要强调的是，尽管较高层建立在较低层之上，但这种划分仅代表关注点的分离，并不意味着层级关系。

如第3.1节所述，Zhamak 和灰色文献将数据产品分为源对齐、聚合型和消费者对齐。借鉴面向服务的架构（SOA）中的服务类型，我们将其分为原子型和复合数据产品。与组合式 Web 服务类似，复合数据产品使用其他数据产品（原子型或复合型）作为其源数据，并将其整合以创造新的增值产品。因此，聚合型和消费者对齐型产品属于复合产品。源对齐的数据产品则是原子型的，因为它们只从运维源系统中摄取数据。

图6所示的分层架构并未呈现数据网格架构中的全部职责。相反，它展示了创建由个人管理的数据产品所需承担的职责。例如，联邦治理团队创建业务术语表的职责未包含在此架构中，因为这与在运行时构建特定的受管数据产品无关。

角色。 我们确定了数据网格参与者的六种不同角色：

**数据平台团队** 负责构建和维护自助式数据平台 [37, 78, 99, 104, 111,126]。他们还会创建基础设施即代码蓝图，供数据产品开发人员用于配置和设置基础设施资源及平台服务，以托管和管理数据产品。他们需要具备软件和数据工程技能 [99]。**数据产品负责人** 负责在各自领域提供和治理数据产品 [14, 16, 37, 41, 47, 61, 70, 74, 78, 125–127, 144, 150, 164, 166]。他们必须确保数据产品可互操作，并满足消费者的需求。他们通常负责领域级治理活动（参见第3节 3.3）。**数据产品开发人员** 使用自助式平台提供的工具/服务，在数据网格上构建和发布数据产品 [33, 37, 41, 47, 61, 127, 135]。**产品聚合者** 构建并提供复合数据产品。灰色文献并未明确提及聚合者这一角色。我们引入这一角色，以清晰区分构建源对齐的原子数据产品与复合数据产品所涉及的职责和业务能力。**产品使用者** 使用数据产品中的数据 [1, 72, 81, 108, 155]。产品聚合者也是产品使用者，因为他们会使用多个其他数据产品，并将其聚合为复合数据产品。**联邦治理团队** 由来自组织内不同领域的代表组成，并在全局范围内治理数据网格 [41, 61, 71, 120, 135, 164]。第3节 3.3 讨论了联邦治理团队的治理活动。

自助服务层。 自助式平台以自助服务的方式提供一系列工具/服务，供数据网格架构中的参与者用来履行各自的职责（见第3.4节）。例如，产品开发人员可以使用平台服务来开发、测试、部署和迭代其数据产品。相比之下，产品负责人和提交至 ACM 的稿件

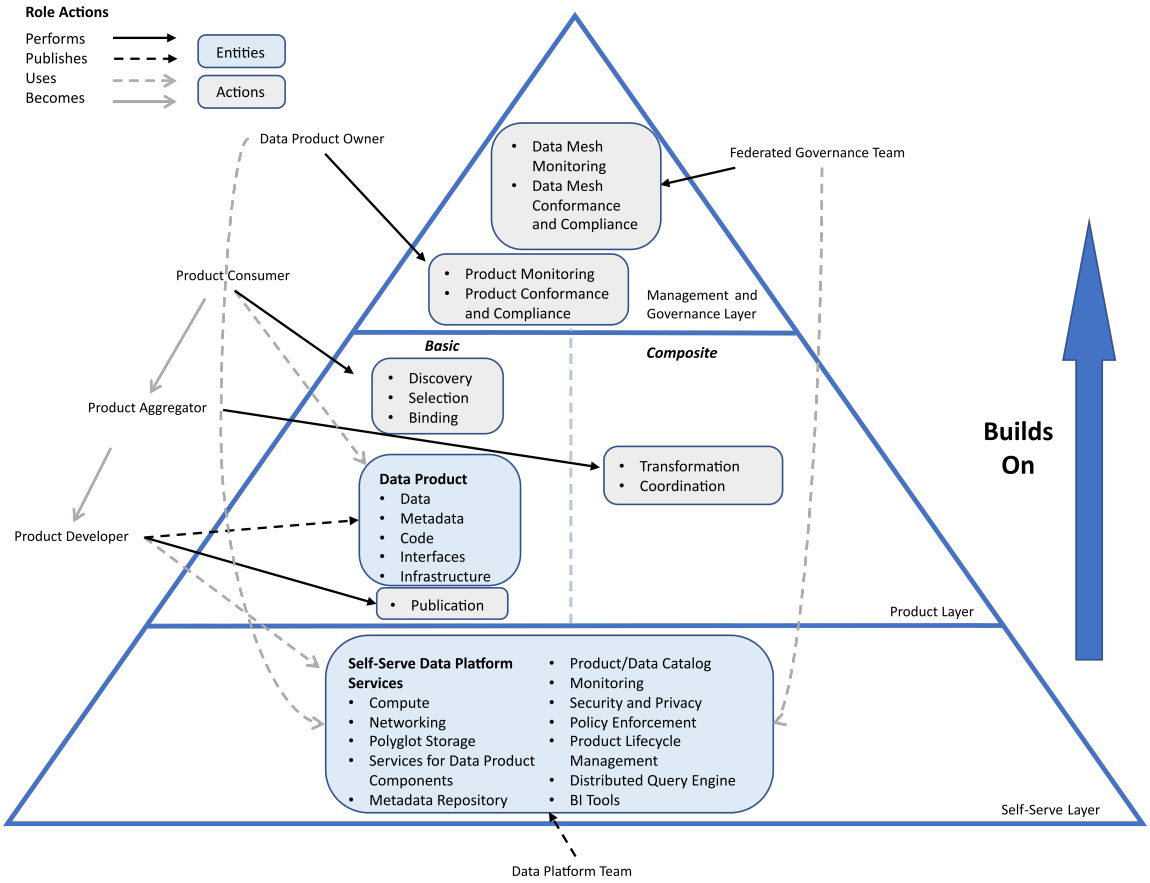



图6. 数据网格中业务能力与角色的分层架构。

联邦治理团队可以使用平台服务来监控和治理各个数据产品以及整个数据网格。

数据产品层。 数据产品层位于自助服务层之上。如第3.1节所述，数据产品由数据、元数据、代码、接口以及用于部署产品的基础设施构成。产品开发人员定义并实现所有组件后，即可将数据产品发布到产品目录中。产品使用者可以浏览目录，查找适用于其用例的候选数据产品，并查看相应的元数据，以确认这些产品是否满足自身需求。产品元数据还可以帮助消费者自动连接（绑定）到数据产品，以获取其中的数据。例如，如果某个数据产品使用存储桶作为服务端点，就可以将元数据（例如存储桶的 URL，以及结果对象的名称和模式）发布到目录中 [115]。

复合数据产品使用其他（原子或复合）数据产品中的数据，可能会应用复杂的转换，并将结果作为新的数据产品发布（参见第3.1节）。复合数据产品包含与原子数据产品相同的组件。不过，产品聚合者还会执行其他操作来构建复合数据产品，这与 xSOA 中的额外职责类似 [118]。为区分这些职责，

图6中的分层架构区分了产品开发人员和产品聚合者：前者构建原子数据产品，后者创建复合数据产品。

复合数据产品需要为组织创造新的价值。简单地集成两个数据源很容易被复制，而且可能无法提供足够的额外价值。因此，产品聚合者可能需要对来自上游数据产品的数据执行复杂的转换，以创造增值数据。此外，组合产品可能涉及协调一系列步骤，例如，从多个数据产品摄取数据、合并其中一部分产品的数据、清洗和标准化数据，以及验证并存储中间结果和最终结果。

复合数据产品在其整个生命周期内都必须与上游数据产品保持兼容。例如，上游产品开发人员可能会更改其数据产品（例如模式和产品接口），这可能会影响下游消费者 [10, 28, 71, 71, 151]。产品聚合者必须通过采用适当的策略来应对此类问题，例如实施产品版本控制并使用防腐层（适配器层） [10, 71]。

管理与治理层。 数据产品必须在运行时进行管理和治理，以确保它们为组织创造价值。如第3.3节所述，治理分为两个层级：本地治理和全局治理。本地治理（针对每个产品）负责确保个人数据产品的质量和符合性。例如，本地治理团队必须确保数据产品在整个生命周期内始终按照指定的服务级别目标运行，并遵守本地和全局政策以及外部法规。复合产品的所有者还需要根据服务级别目标（SLO）监控上游数据产品的性能。复合数据产品的性能可能会受到上游数据产品性能的影响。因此，产品聚合方需要监控上游产品的性能，以确定是否存在服务级别目标违规，并确保复合产品在其服务级别目标范围内运行。运行时全局治理的主要目标是确保网格架构能够在整个组织中有效扩展。联邦治理团队会监控网格，评估数据网格的整体健康状况，并执行全局政策、互操作性标准和指南。

### 5.2 数据网格开发的分层架构

数据网格文献并未明确提供一种用于在组织内部或跨组织实施数据网格的端到端方法。然而，相关文献提倡应用领域驱动设计（DDD）、产品思维和基于平台的开发原则。特别是，应将领域驱动设计应用于组织业务领域所产生的数据，以识别数据产品 [1, 11, 28, 33, 37, 39, 41, 48, 71, 87, 92, 95, 103, 113, 115, 122, 124, 130, 134, 151, 154, 157, 158, 162]。受面向服务的架构开发分层模型的启发 [116],，我们提出了一种用于数据网格开发的分层模型，将前述原则加以整合。图7展示了所提出的数据网格开发分层方法。该模型由七个不同层次组成：业务领域、数据领域、数据限界上下文、数据产品、数据产品组件、自助式平台服务和运行基础设施。

遵循DDD方法，第一层将组织的业务划分为互不重叠的领域。业务领域由一组共享共同业务能力的业务流程组成，这些流程协作以实现该领域的业务目标 [116]。这些流程为该领域及组织的其他部分产生有价值的数据，因此可以将其作为数据产品提供。例如，制造业组织可能包括分销、制造、财务和人力资源等领域 [116]。分销领域可能包含消费者采购、订单管理和库存等关键业务流程。采购流程可能会产生包含产品或服务信息、购买时间和支出金额的数据。这份稿件已提交至ACM

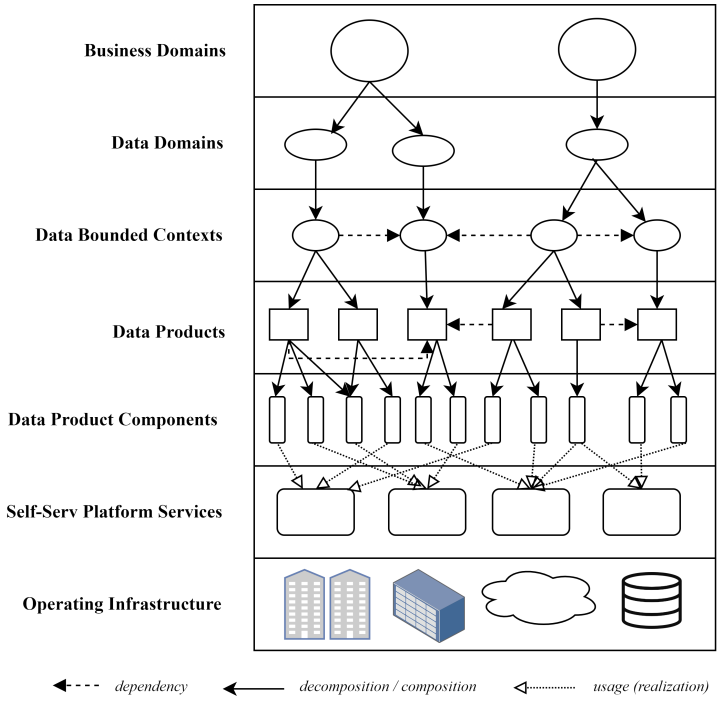



图 7. 数据网格开发的分层模型。

这些数据可用于构建数据产品，以回答有关购买历史、客户购买模式以及库存可用性和产品外观等其他相关细节的查询。

第二层将业务领域的分析数据划分为一个或多个数据领域。数据领域为分析数据划定边界。从数据网格和数据治理的角度来看，数据领域对数据进行分类和组织，以明确数据的问责与责任。如上所述，一个业务领域中的不同业务流程可能会产生多种数据。由于流程的多样性和复杂性，每个流程的责任可能会分配给不同的团队 [32]。同样，各流程产生的分析数据也可能由不同团队所有，从而使同一业务领域形成多个数据领域。灰色文献还建议，让业务领域与数据领域保持一致，并酌情为共享数据创建新的领域 [11, 28, 33, 39, 48, 87, 95,113, 133, 151, 157, 158, 162]。

第三层确定每个数据领域中的限界上下文。灰色文献建议使用限界上下文来界定数据产品的范围 [41, 115, 127, 130, 134, 154, 157, 162]。领域驱动设计基于领域模型来设计和实现软件系统。领域模型是各领域的对象模型，包含领域中的行为和数据 [51, 54]。对于规模更大、复杂度更高的领域，如果多个团队分别负责不同的子系统，创建并维护统一模型可能并不可行，或不具备成本效益 [51, 54]。为解决这一问题，领域驱动设计将某个领域的解决方案空间划分为多个限界上下文，每个限界上下文都可以拥有独立的领域模型。数据领域也可能很复杂，并包含多个数据逻辑分组。每个分组都可以有独立的数据所有者和模型（即用于分解数据领域的限界上下文）。例如，客户可能分为不同类别：仅线上、仅实体渠道、第三方、潜在客户、个人、团体、公司、政府和慈善机构。

管理这些类别中的每一类或其中一部分的数据，方式可能各不相同，也可能由多个团队或个人负责。因此，客户数据领域中会形成多个限界上下文。

数据限界上下文之间可能存在各种依赖关系（设计时和运行时）。灰色文献建议使用 DDD 的战略设计模式（例如客户-供应商和防腐层）来集成限界上下文（即上下文映射） [158]。例如，复合产品的限界上下文可以使用防腐层，尽量减少上游限界上下文中的数据产品数据模型变更所带来的影响 [71]。例如，用于销售漏斗分析的数据产品会使用客户搜索数据产品来获取客户的浏览历史。销售漏斗分析产品可以实现防腐层，以防止浏览历史数据格式变更带来的不利影响。

第四层用于识别每个限界上下文中的数据产品。为了构建限界上下文的应用程序，DDD 建议应用所谓的战术设计模式：实体、值对象、聚合、服务、仓储、工厂、事件和模块。这些模式可用于丰富限界上下文的领域模型，并可根据丰富后的模型推导出应用架构组件。例如，在微服务架构中，可以使用战术设计模式从领域模型推导出服务、API 和事件 [149]。数据网格灰色文献 [11, 71, 92, 122] 也建议应用这些模式，例如，使用领域主事件来摄取和提供数据，并实现数据变更日志；使用实体和聚合型来识别产品，以及创建数据库模式和产品 API。

第五层由实现数据产品功能性和非功能性需求的组件构成。数据产品有多种类型，例如数据集（以数据库视图或 API 的形式提供）、数据管道、机器学习训练管道和机器学习预测服务。不同的数据产品可能采用不同的架构。例如，数据管道产品可能采用 Lambda 和 Kappa 等流行的数据架构风格 [35], ，而机器学习预测服务可能采用模型即服务或模型即依赖等架构风格 [173]。最终的数据产品可以包含流式数据摄取、ETL工作流、原始数据、清洗后数据和增强数据的存储、模型和特征存储、API、数据质量评估器以及报告生成器等组件。数据产品也可以共用某些组件的实现，例如通用的数据验证服务或模型验证服务。

第六层由支持数据产品及其组件端到端生命周期的自助式平台服务构成，涵盖设计、实现、测试、部署、执行、监控和演进。例如，数据产品开发人员应能够通过自助方式使用 ETL平台服务，为其数据产品创建 ETL管道，按需或按计划部署和执行所创建的管道，监控已执行管道的性能、资源使用情况和错误，并根据需要对管道进行修改。第3.4节列出了我们从灰色文献中发现的常见平台服务。平台团队负责提供自助式平台服务。他们可以从头开发平台服务、重复使用并定制第三方软件、使用云服务提供商提供的平台服务，或通过融入自助服务能力来改造现有平台服务。

第七层主要包括平台服务和数据产品所使用的基础设施资源，以及组织中充当数据源系统的现有系统。

在数据网格中，除了数据产品之外，联邦治理团队还会使用管理和治理应用来监控和控制数据产品。图7并未明确包含这些应用。不过，由于这些应用可能会使用数据网格的监控数据，开发人员可以开发并将其作为数据产品提供。此外，管理应用还应能够使用自助式平台服务。

### 5.3 数据网格的逻辑运行时结构

本节介绍数据网格的逻辑运行时视图。灰色文献来源中介绍的数据网格架构主要关注由跨数据领域相互连接的受管理（受治理）数据产品构成的网络的运行时模型 [4, 11, 28, 31, 41, 50, 55, 71, 72, 78, 103, 111–113, 127, 131, 170, 171, 175]。我们从这些架构设计中提取了核心组件，并参考 Web 服务管理架构 [80, 85, 116, 117]，创建了一个整合模型。图8展示了运行时的参考架构。

产品容器是数据产品及其管理代理的部署和运行时环境。我们借鉴了 Web 服务容器和管理代理的概念 [116]。与 Web 服务管理架构 [85, 116, 117], 一样，我们将数据产品、源系统和最终用户之间的通信（数据通道和输入／输出接口），与管理代理和管理／治理平面之间的通信（管理通道和管理接口）分开。输入接口 [3, 10,38, 71, 78, 92, 114, 177] 使数据产品能够从源系统和其他数据产品接收数据。输出接口 [3, 10, 38, 78, 92, 114, 124, 177] 使数据产品能够向消费者或复合数据产品提供数据。最后，管理接口 [38, 71, 78, 92] 可监控和控制数据产品的行为（例如，监控产品性能、发出告警、控制数据路由，以及强制执行安全策略和数据契约）。通信模式可以是拉取式（例如，通过 REST API [11, 14, 15, 41, 91, 95, 113, 154, 158, 177] 检索数据、从共享数据库／存储桶 [38, 67, 78, 100, 108, 130, 142, 157, 162, 177]读取数据，或从消息代理 [14, 15, 17, 18, 67, 71, 78, 91, 112, 113, 154, 158, 177]轮询数据），也可以是推送式（例如，数据一旦可用便将其推送给下游使用者 [10, 18, 40, 71, 108]）。

产品的管理代理 [37, 38, 112, 113] 在产品级执行管理任务，例如标记并执行策略决策、收集性能指标，以及管理产品生命周期（即创建、更新和销毁产品）。灰色文献建议采用边车设计模式和策略即代码方法，实现管理代理 [15, 39, 112, 113, 156]的监控与策略执行业务能力。边车模式 [20] 将策略决策制定与数据产品解耦，从而可在各数据产品间复用监控与管理逻辑，并确保其实现一致。策略即代码方法使用编程代码编写策略，以便应用软件开发最佳实践，例如版本控制、自动化测试和自动化部署。策略以一组规则的形式呈现，用于规定管理代理的行为，例如如何对数据产品访问执行身份验证和授权策略、何时触发告警，以及将消息路由到何处。

管理与治理平面中的组件/应用程序可以使用产品的管理接口，从数据产品接收信息（例如性能数据、访问日志和告警），并向数据产品发送配置和策略更新。我们从灰色文献中确定了该平面的几个关键组件：数据产品注册与发现服务 [16, 102, 130, 166],策略编写服务和策略存储库 [28, 71],全局元数据存储库 [28, 31, 71, 161], 以及数据网格仪表板 [28, 71]。产品注册表和发现服务连接到管理通道。产品开发人员通过管理通道发布其产品的元数据。消费者可以利用已发布的元数据发现数据产品。全局元数据存储库可以存储跨领域共享的元数据，例如主数据和参考数据。联邦治理团队可以使用策略编写服务创建全局策略（例如互操作性、安全和合规性），将其存储在策略存储库中以便重复使用和进行版本控制，并通过各个产品提供的管理接口使这些策略生效。

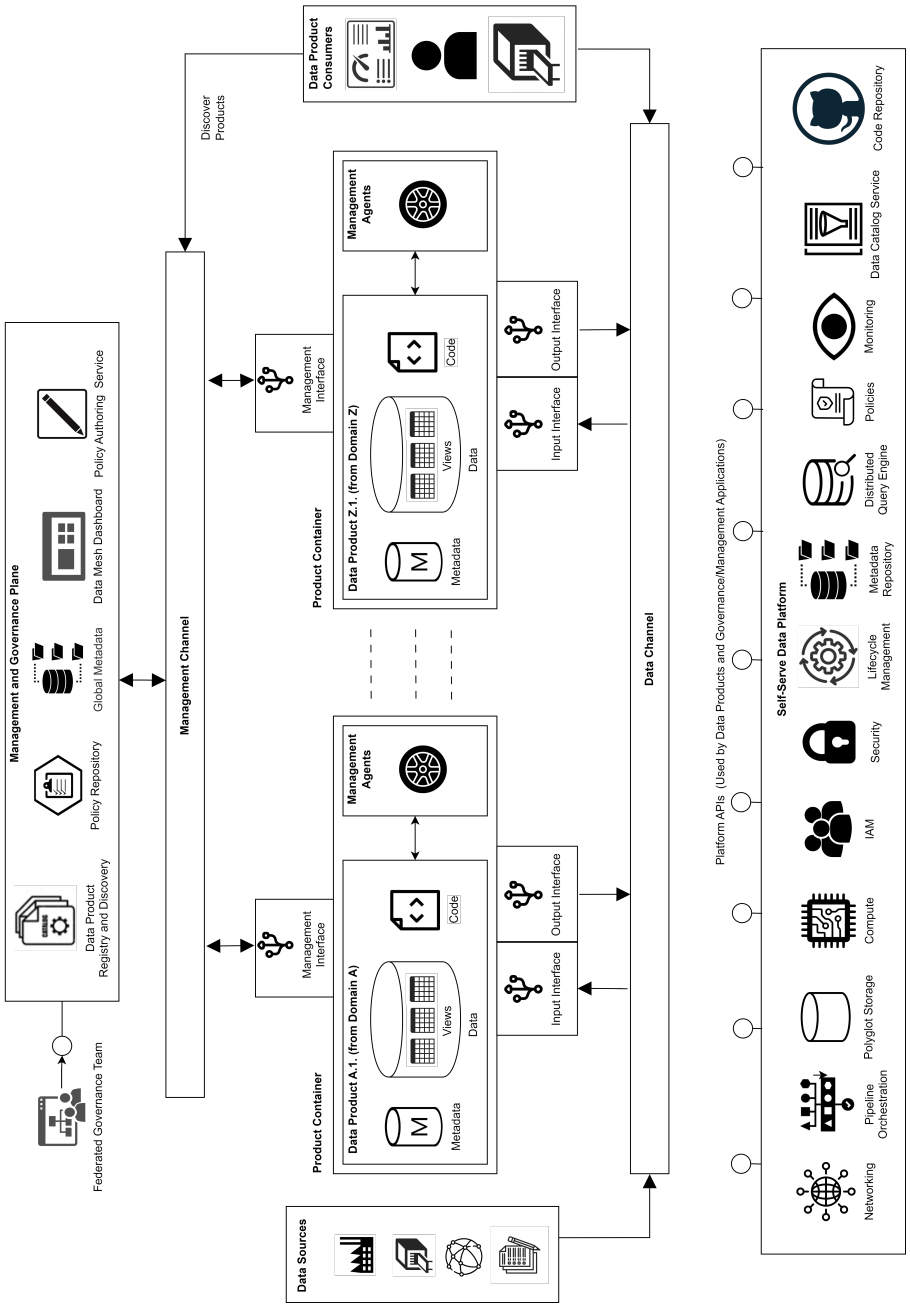



图 8. 数据网格的逻辑运行时结构。

团队可以使用数据网格仪表板监控数据网格的运行状况，包括跨域访问和数据产品的使用情况，以及政策合规程度。稿件已提交至 ACM

自助式平台平面提供基础设施资源和平台服务，用于开发、测试、构建、部署、执行、监控和管理数据产品及治理组件。例如，数据产品可以使用机器学习工作流平台服务来创建和运行训练流水线。产品注册表和发现服务可以使用数据/产品目录或元数据存储库平台服务。第3.4节讨论了自助式平台的关键平台服务。

## 6 研究问题4：关于数据网格的研究挑战有哪些?

本节提出了一系列亟待解决的数据网格研究挑战。我们将灰色文献中从业者提及的挑战与面向服务的计算 [117, 118, 174] 和数据管理 [5, 45, 69, 152]中的相关研究问题进行对应，从而识别出这些挑战。

**数据网格标准化**。 数据网格中的数据产品应当具有互操作性，以实现产品的无缝组合并促进跨领域协作。数据产品还需要具备可移植性，以便在几乎不修改或完全不修改数据产品的情况下，修改或替换自助式平台服务和治理应用。数据网格的标准化能够提升其互操作性和可移植性 [4, 12, 36, 71, 127, 158]。以下方面需要标准：i）数据产品及其接口和组合的定义；ii）数据产品发布与发现、质量评估、部署和治理；iii）自助式平台服务及其接口；以及 iv）数据产品使用或生成的制品（例如数据模型、元数据和机器学习模型）。面向服务的架构和云领域的标准化工作有望帮助应对这些挑战。例如，Web 服务描述语言（WSDL）支持对具有不同端点、消息类型和访问策略的服务进行建模（例如 HTTP 端点、发布-订阅主题和 FTP 端点）[117]。因此，WSDL 及其他相关 Web 服务标准或许有助于确定用于对不同类型数据产品建模的概念。云应用拓扑与编排规范（TOSCA）是 OASIS 的一项标准语言，用于对复杂应用程序的部署拓扑进行建模 [13]。它可作为一种合适的、与技术无关的基础设施即代码语言，用于定义数据产品和自助式平台组件可互操作的部署蓝图。

**数据网格开发与运营的方法论和工具**。 数据网格文献 [41, 102, 111,112, 120, 127, 130, 158] 提供了开发数据网格架构的一般原则，例如领域驱动设计（DDD）、产品思维、平台思维和计算治理。然而，目前还不存在用于将这些原则应用于数据网格开发与运营的工具和详细的（经实证验证的）方法论。组织还需要相关指导和工具，以便在评估成本与收益的同时，系统地将其遗留数据架构迁移到数据网格。这些研究问题可能受益于以下方面的丰富知识：应用领域驱动设计来设计和实现微服务 [183]，以及将遗留单体架构迁移到基于云的微服务架构 [66]。

**数据产品生命周期**。 数据产品是数据网格的核心。数据产品生命周期的每个阶段（例如识别、设计、开发、测试、部署、发现、组合、运营、监控和演进）都可能存在尚未解决的挑战 [41, 43, 115, 127, 130, 134, 154, 157, 162]。例如，产品识别阶段会将组织的分析数据格局分解为一组数据产品。分区过程需要支持正确应用领域驱动设计，确保数据产品的模块化和效用，并考虑复用现有数据架构中的制品，例如数据源系统、数据管道、机器学习管道、文档和执行日志。还需要开展进一步的探索性研究

需要这些内容，才能充分理解产品生命周期各阶段的研究挑战。鉴于数据产品与服务具有可比性，我们认为，SOA 和微服务领域的相关现有研究有助于理解并应对上述产品挑战。例如，已有关于 Web 服务发现 [106] 和组合 [147], 、微服务开发中的领域驱动设计（DDD） [183], ，以及数据存储 [148] 和 Web 服务 [167]的元数据管理技术的文献。

**自助式平台服务**。 领域团队需要能够以自助方式使用与领域无关的平台服务，从而轻松且独立地构建和管理其数据产品。尽管已有针对自助式系统的研究，例如自助式商业智能（商业智能） [82]和面向云应用的自助服务门户 [98], ，但这些研究并未涵盖每种平台服务类型，例如数据管道、机器学习管道，以及基于基础设施即代码的数据产品配置。此外，推荐系统也可以集成到平台服务中，协助用户构建数据产品，例如在创建数据管道或基础设施即代码蓝图时交互式地提供建议。最后，由于组织可能已经在使用数据平台，因此可能需要有关如何为其遗留数据平台添加自助服务能力的指导和技术 [28, 37, 61, 74, 78, 150, 177]。关于软件工程推荐系统的现有文献探讨了如何简化软件开发生命周期各阶段的工作并为用户提供协助 [6, 57]。许多研究都探索了如何根据自然语言描述生成代码，包括 SQL 和数据处理代码 [179]。大语言模型也在这一研究领域展现出巨大潜力 [128, 182]。

**数据网格治理**。 尽管数据治理数十年来一直是活跃的研究主题，正如近期多篇综述所指出的 [2, 34, 69], ，仍有许多尚待解决的研究问题。数据网格，尤其是其数据去中心化所有权和联邦式计算治理原则，可能为数据治理研究问题带来新的维度。例如，跨越领域边界（甚至可能跨越组织边界）高效组合数据产品，取决于联邦治理机制能否确保数据产品可互操作、符合服务质量（QoS）和策略要求，并与公司级数据战略保持一致 [41, 78, 120, 127, 158]。此外，采用策略即代码方法（即计算治理）实现端到端数据治理自动化，尚未得到充分探索。策略即代码方法已应用于多种用例，例如在微服务架构中（借助服务网格）实施策略 [26] 、在基于基础设施即代码的部署管道中实施策略 [181], ，以及在机器学习环境中执行契约 [168]。

**组织变革管理**。 如第5节所述，数据网格为组织中的团队和个人引入了一系列新的角色与职责。具体而言，领域数据的责任被转移至更接近去中心化自治领域团队的位置。因此，对于采用集中式运营模式和组织设计的组织而言，踏上数据网格之旅可能需要进行重大组织变革（此外还需要技术变革）。数据网格还可能对组织现有的数据管理和治理流程与实践产生重大影响。因此，有必要开展更多研究，以了解数据网格的组织影响以及采用数据网格的障碍，并制定指南和框架，以实施和管理所需的组织变革 [12, 37, 61, 74, 120, 127, 162]。有关将本地部署系统迁移至云端、将遗留应用迁移至微服务的研究，也探讨了相关的组织挑战及潜在解决方案 [59, 172]。进一步的研究也有助于了解并应对数据网格转型中的组织挑战。

## 7 相关工作

本节报告了有关数据网格及相关综述的研究。

### 7.1 数据网格研究

Loukiala 等人介绍了某个北欧大型制造业国家的一家公司，如何从集中式单体数据平台架构迁移到类似数据网格的分布式架构 [88]。他们着重说明了迁移过程中可以采取的具体步骤，并对传统数据仓库架构与数据网格进行了高层次的对比。Machado 等人发表了多篇关于数据网格剖析与架构的研究。第一篇概述并解释了数据网格的概念，并以行业中的两种实现为例加以说明（一家在线零售商 Zalando，以及一个流媒体平台 Netflix） [8]。随后，他们在另一篇论文中提出了一个“领域模型”，其中包含数据网格的抽象类以及高层次架构 [89]。最后，他们基于这些研究，考察了可用于实现数据网格的现有（商业）工具，并通过概念验证实现演示验证其架构模型 [90]。类似地，Butte 和 Butte [23] 提出了一个更完整的参考架构。然而，由于论文没有明确引用来源，因此无法确定其架构的依据。目前，这些是仅有的明确讨论数据网格架构的已发表学术研究。不过，它们对数据网格概念的描述缺乏对数据网格最新研究进展的深入探讨，几乎完全基于 Zhamak Dehghani 的博客文章 [37, 41], 以及两份企业演示材料：Zalando [94] 和 Netflix [29]。

Kumar 和 Sujata [24] 介绍了数据网格的关键概念，提出了领域内数据处理管道的模板，并提供了一个使用 AWS 云进行数据网格部署的示例。Li 等人 [84] 介绍了将数据网格概念应用于电力系统数据管理的案例研究。他们在原始数据源之上使用联邦式数据处理引擎和元数据层，以构建增值数据服务。Podlesny 等人 [123] 讨论了分布式数据网格可能存在的数据隐私问题，重点关注通过准标识符暴露个人身份信息（PII）数据。Kraska 等人 [77] 提出了一个具有前瞻性的系统，旨在为联邦式数据处理引擎集合提供统一数据处理接口，使非专业终端用户能够利用数据访问不同的专用引擎。Dahdal 等人 [30] 提出了一种边缘中间件系统架构，可将面向战术战场的数据密集型应用作为数据产品提供。该中间件提供数据摄取、存储、安全和 API 等常见的数据管理功能。Hooshmand 等人 [68] 提出将数据网格方法应用于工程产品生命周期管理（PLM）。他们还建议应用领域驱动设计，并使用本体进行数据建模。

### 7.2 相关综述

Oliveira 等人在 2018 [138] 中调查了有关“数据生态系统”的最新研究进展。数据网格符合他们对数据生态系统的定义，其研究也讨论了许多与数据网格相关的要素。然而，他们提及的研究中，没有一项涉及数据网格或任何密切相关的术语。Davoudian 和 Liu [35] 还开展了一项元研究，探讨大数据（生态）系统中的活动与挑战及其对架构的影响。同样，本文指出的活动与挑战与数据网格中的活动与挑战直接相关。然而，本文讨论的文献完全聚焦于作为最新研究进展的传统数据湖和数据仓库。直到最近，数据网格才开始出现在综述文献中。例如，Driessen 等人 [44] 明确讨论了数据网格，并将其视为一种特殊的（内部）数据市场。不过，在这项研究中，数据网格仅在数据交换生态系统这一更大背景下被抽象地考察，并未对数据网格架构进行深入讨论

已有相关内容。因此，基于现有文献，我们得出结论：数据网格架构领域存在知识空白：一方面，明确讨论数据网格架构的论文（[88, 90]）并未考虑广泛的来源来作为其架构定义的依据。另一方面，探讨数据生态系统最新研究进展的现有元研究 [35, 44, 138]几乎没有讨论数据网格相关来源，也未提供参考架构。

鉴于上述工作，目前尚无针对数据网格文献的综述。此外，该主题的学术文献寥寥无几，主要基于少数灰色文献文章。因此，本文对这类文献来源进行了系统分析、评估和总结。

## 8 效度威胁

本节讨论可能影响本研究的外部效度、构念效度和内部效度威胁 [176] 。首先，我们指出，在系统性灰色文献综述中存在遗漏相关研究的风险。由于该领域发展迅速，过一段时间后重新开展灰色文献综述，观察随着行业进一步采用数据网格，这一主题如何发展成熟，可能会很有意义。因此，我们将研究范围扩展至涵盖2021年9月至2022年5月的来源，并由此新增了35个来源。

此外，在方法论的文章筛选过程中，我们依赖作者的（主观）专业判断来确定来源的质量和相关性。这一过程可能会给最终筛选结果引入偏差。因此，两位作者分别评估了每个来源，并计算了评定者间信度卡帕系数，以确保偏差没有过高。

最后，我们通过系统地考察各类来源构建了参考架构。然而，如上所述，可供比较的完整、整体性架构实例寥寥无几，而且许多来源只考察了部分架构。这些部分架构实例可能导致整体参考架构未臻优化。同时，我们指出，这个问题在某种程度上形成了循环：参考架构可以显著增加完整架构实例化的数量，而这反过来又能改进未来的参考架构。

## 9 结论与未来工作

尽管企业日益走向数据驱动，其集中式单体数据架构仍使其无法充分发挥分析数据的价值。数据网格是一种新兴的、领域驱动的去中心化架构与社会技术范式，有望解决集中式企业数据架构的局限。尽管数据网格在行业内受到广泛关注，但针对这一主题的学术研究显然十分匮乏。本文系统地筛选并评述了114篇关于数据网格的工业灰色文献，以识别关键设计原则、架构组件、组织角色、效益与顾虑。我们进一步结合相近领域的学术文献，丰富了评述结果，并据此构建了数据网格关键维度的参考架构。最后，我们确定了实现数据网格愿景所需应对的研究挑战。我们认为，所识别的开放问题和参考架构能够为未来的数据网格研究提供一个框架。

关于未来工作，我们将提出领域特定语言，以定义数据网格三大支柱——数据即产品、自助式平台和联邦式治理——所需的数据产品和决策平台。我们将创建数据产品和契约的元模型，并利用本文提供的参考架构，开发一种用于设计和实例化数据网格架构的方法。

## 参考文献

[1] Tareq Abedrabbo。2021。现实世界中的数据网格：从金融市场汲取的经验。
ttps://www.infoq.com/presentations/cmc-markets- challenges/[访问日期：3月 2023]。提交至 ACM 的稿件

[2] Rene Abraham、Johannes Schneider 和 Jan vom Brocke。2019。数据治理：概念框架、结构化综述与研究议程。International Journal of Information Management 49（2019），424–438。[3] Adevinta。2021。在 Adevinta 构建数据网格，以支持数据产品生态系统。https://medium.com/adevinta-tech-blog/building-a-data-mesh-to-support-an-ecosystem-of-data-products-at-adevinta-4c057d06824d[访问日期：3月 2023]。[4] AgileLab。2021。数据网格解析：成功的数据驱动型公司如何以及为何采用数据网格。https://medium.com/agile-lab-engineering/data-mesh-explanation-a207fac61341/ [访问日期：3月 2023]。[5] Majid Al-Ruithe、Elhadj Benkhelifa 和 Khawar Hameed。2019。数据治理与云数据治理的系统性文献综述。Personal and Ubiquitous Computing 23（2019），839–859。[6] Lissette Almonte、Esther Guerra、Iván Cantador 和 Juan de Lara。2022。模型驱动工程中的推荐系统。Software and Systems Modeling 21，1（2022年2月1日），249–280。[7] Paul Andrew。2022。Azure 中构建数据网格架构系列。https://mrpaulandrew.com/2022/01/07/building-a-data-mesh-architecture- in-azure-part-3/[访问日期：3月 2023]。[8] Inês Araújo Machado、Carlos Costa 和 Maribel Yasmina Santos。2022。通过数据网格实现推进数据架构。载于《Intelligent Information Systems》，Jochen De Weerdt 和 Artem Polyvyanyy（编）。Springer International Publishing，Cham，10–18。[9] AWS。2021。参考架构。https://docs.aws.amazon.com/wellarchitected/latest/analytics-lens/data-mesh-reference-architecture.htm l[访问日期：3月 2023]。[10] Tristan Baker。2021。通过数据网格在 Netflix Studio 中实现数据移动。https://netflixtechblog.com/data-movement-in-netflix-studio-via-data-mesh- 3fddcceb1059[访问日期：3月 2023]。[11] Tristan Baker。2021。Intuit 的数据网格战略。https://medium.com/intuit-engineering/intuits-data-mesh-strategy-778e3edaa017[访问日期：3月 2023]。[12] Oliver Bauer。2021。开启网格之旅。https://vistaprint.io/blog/making-a-mesh-start[访问日期：3月 2023]。[13] Julian Bellendorf 和 Zoltán Ádám Mann。2020。使用 TOSCA 规范云拓扑并进行编排：一项调查。Computing 102，8（2020），1793–1815。[14] Eric Broda。2022。数据网格架构。https://towardsdatascience.com/an-architecture-for-the-data-mesh-32ff4a15f16f[访问日期：3月 2023]。[15] Eric Broda。2022。数据网格架构以及 API 和 JSON Schema 的作用。https://towardsdatascience.com/data-mesh-architecture-and-the-role-of-apis-json-schemas-3dc616650960[访问日期：3月 2023]。[16] Eric Broda。2022。数据网格架构模式。https://towardsdatascience.com/data-mesh-architecture-patterns-98cc1014f251[访问日期：3月 2023]。[17] Eric Broda。2022。数据网格模式：变更数据捕获。https://towardsdatascience.com/data-mesh-pattern-deep-dive-change-data-capture- eb3090178c34[访问日期：3月 2023]。[18] Eric Broda。2022。数据网格模式：事件流骨干。https://towardsdatascience.com/data-mesh-pattern-deep-dive-event-streaming- backbone-99a5bb2a7cbf[访问日期：3月 2023]。[19] Seb Bulpin。2021。为何选择数据网格 101：数据发现的艺术与科学。https://www.linkedin.com/pulse/why-data-mesh-101-art-science-killer-discovery-seb-bulpin/[访问日期：3月 2023]。[20] Brendan Burns 和 David Oppenheimer。2016。基于容器的分布式系统设计模式。载于《第八届 USENIX 云计算热点主题会议论文集》（美国科罗拉多州丹佛）（HotCloud'16）。USENIX Association，美国，108–113。[21] Bert-Jan Butijn、Damian A. Tamburri 和 Willem-Jan Van Den Heuvel。2020。区块链：系统性多源文献综述。ACM Computing Surveys（CSUR）53（2020），1–37。第 3 期。[22] Chris Butler。2021。人员分析需要支持企业数据网格。https://www.onemodel.co/blog/people-analytics-needs-to-support- the-enterprise-data-mesh[访问日期：3月 2023]。[23] Vijay Kumar Butte 和 Sujata Butte。2022。企业数据战略：去中心化数据网格方法。载于《2022年商业与工业数据分析国际会议》（ICDABI）。62–66。https://doi.org/10.1109/ICDABI56818.2022.10041672[24] Vijay Kumar Butte 和 Sujata Butte。2022。企业数据战略：去中心化数据网格方法。载于《2022年商业与工业数据分析国际会议》（ICDABI）。62–66。[25] Andrew Carr。2021。数据网格到底是什么？它真的是一种东西吗？https://blog.scottlogic.com/2021/05/28/what-actually-is-a-data-mesh- and-is-it-really-a-thing.html [访问日期：3月 2023]。[26] Ramaswamy Chandramouli、Zack Butcher 等。2020。使用服务网格架构构建安全的基于微服务的应用程序。NIST 特别出版物 800（2020），204A。[27] Muthu Chinnasamy 和 Vinod Menon。2021。借助 Infosys 数据网格和 MongoDB，改造孤岛式数据仓库之间的数据。https: // www.mongodb.com/blog/post/modernize-data-between-siloed-data-warehouses-infosys-data-mesh-monogdb [访问日期：3月 2023]。[28] Jon Cooke。2021。以业务速度部署数据产品。http://dataception.com/Data-Mesh-Deploying-Data-Products-at-the-speed- of-the-business.html[访问日期：3月 2023]。[29] Justin Cunningham。2020。Netflix 数据网格：可组合数据处理。https://www.youtube.com/watch?v=TO_IiN06jJ4[30] Simon Dahdal、Filippo Poltronieri、Mauro Tortonesi、Cesare Stefanelli 和 Niranjan Suri。2023。采用数据网格方法，在战术边缘支持以数据为中心的应用程序。载于《2023年军事通信与信息系统国际会议》（ICMCIS）。1–9。M

https://doi.org/10.1109/ICMCIS59922.2023.10253568[31] Kenneth Dalvik. 2020. SAP 数据仓库云数据网格：事实。https://blogs.sap.com/2020/10/22/sap-data-warehouse-cloud-data-mesh- true/[访问日期：3月 2023].[32] Kjersti Berg Danilova. 2019. 业务流程管理中的流程负责人：系统性文献综述。《业务流程管理期刊》25, 6 (2019), 1377–1412.[33] Devayon Das. 2020. 数据网格：如何克服数据湖面临的挑战。https://www.zaloni.com/resources/data-mesh-data-lake- challenges/[访问日期：3月 2023].[34] Elizabeth Davidson, Lauri Wessel, Jenifer Sunrise Winter, and Susan Winter. 2023. 数据治理、数字创新与重大挑战领域学术研究的未来方向。《信息与组织》33, 1 (2023), 100454. https://www.sciencedirect.com/science/article/pii/S1471772 723000088[35] Ali Davoudian and Mengchi Liu. 2020. 大数据系统：软件工程视角。《ACM计算调查》53, 5, Article 110 (2020年9月), 39页。https://doi.org/10.1145/3408314[36] Zhamak Dehghani. 2019. 与 Zhamak Dehghani 探讨数据网格。https://softwareengineeringdaily.com/2019/07/29/data-mesh-with-zhamak- deghani/[访问日期：3月 2023].[37] Zhamak Dehghani. 2019. 如何从单体式数据湖迈向分布式数据网格。https://martinfowler.com/articles/data- monolith-to-mesh.html [访问日期：3月 2023].[38] Zhamak Dehghani. 2021. 数据网格：架构深入解析。https://www.infoq.com/presentations/data-mesh-concepts/[访问日期：3月2023].[39] Zhamak Dehghani. 2021. 数据网格：与 Zhamak Dehghani 探讨如何定义新的数据架构范式。https://www.torocloud.com/podcast/defin ing-data-mesh-zhamak-dehghani/[访问日期：3月 2023].[40] Z Dehghani. 2022. 数据网格：大规模实现数据驱动价值。O'Reilly Media, Inc.[41] Zhamak Dehghani. 2022. 数据网格原则与逻辑架构。https://martinfowler.com/articles/data-mesh-principles.html [访问日期：3月 2023].[42] Heather Devane. 2021. 什么是数据网格? https://www.immuta.com/articles/what-is-a-data-mesh/ [访问日期：3月 2023].[43] Chris Dowsett. 2021. 使用微服务构建和扩展数据功能。https://towardsdatascience.com/using-microservices-to-build-and- scale-data-functions-28d47f400419[访问日期：3月 2023].[44] S.W. Driessen, G. Monsieur, and W. Van Den Heuvel. 2022. 数据市场设计：系统性文献综述。《IEEE开放获取》10 (2022), 1–1.[45] Schahram Dustdar, Reinhard Pichler, Vadim Savenkov, and Hong-Linh Truong. 2012. 质量感知的面向服务数据集成：需求、技术现状与开放挑战。《ACM SIGMOD记录》41, 1 (2012), 11–19.[46] Christof Ebert, Gorka Gallardo, Josune Hernantes, and Nicolas Serrano. 2016. DevOps。《IEEE软件》33, 3 (2016), 94–100.[47] ELEKS. 2021. 数据网格：分布式架构的四项原则。https://medium.datadriveninvestor.com/data-mesh-the-four-principles- of-a-distributed-architecture-59514eba1e52 [访问日期：3月 2023].[48] Dodo Engineering. 2021. 数据网格：如何在没有单体的情况下处理数据。https://prog.world/data-mesh-how-to-work-with-data-without-a- monolith/[访问日期：3月 2023].[49] Spark Equation. 2021. 关于数据网格，你需要了解的内容。https://sparkequation.com/2021/02/24/data-mesh/[访问日期：3月 2023].[50] Vanessa Ericsson. 2021. 数据网格战略与架构。https://drive.google.com/file/d/1Xi-Ri_6HrjBtCHssmi2E4XvXqEK25anV/vi ew[访问日期：3月 2023].[51] Eric Evans and Eric J Evans. 2004. 领域驱动设计：应对软件核心领域的复杂性。Addison-Wesley Professional.[52] Chloe Feingold. 2021. 为什么治理是数据网格的关键纽带（以及如何避免把"网格化"搞砸）。https://blog.privacera.com/why-governance-is-the-critical-stitch-in-data-mesh-and-how-to-avoid-meshing-it-up-ade91a896ae5[访问日期：3月 2023].[53] Neal Ford, Rebecca Parsons, Patrick Kua, and Pramod Sadalage. 2022. 构建演进式架构。" O'Reilly Media, Inc.".[54] Martin Fowler. 2012. 企业应用架构模式：企业应用架构模式。Addison-Wesley.[55] Bhavesh Furia. 2021. 数据网格：重新思考企业数据架构。https://www.cuelogic.com/blog/data-mesh [访问日期：3月 2023].[56] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. 在软件工程中纳入灰色文献并开展多声部文献综述的指南。《信息与软件技术》106 (2019), 101–121。期号：2018年9月。https://doi.org/10.1016/j.infsof.2018.09.006[57] Marko Gasparic and Andrea Janes. 2016. 软件工程推荐系统会推荐什么：系统性文献综述。《系统与软件杂志》113 (2016), 101–113.[58] Chandan Gaur. 2021. 数据网格及其优势全面概览。https://www.xenonstack.com/blog/data-mesh[访问日期：3月 2023].[59] Mahdi Fahmideh Gholami, Farhad Daneshgar, Graham Low, and Ghassan Beydoun. 2016. 云迁移流程——调查、评估框架与开放挑战。《系统与软件杂志》120 (2016), 31–69.[60] Corinna Giebler, Christoph Gröger, Eva Hoos, Holger Schwarz, and Bernhard Mitschang. 2019. 发挥数据湖的价值：现状与挑战。载于《大数据分析与知识发现》，Carlos Ordonez, Il-Yeol Song, Gabriele Anderst-Kotsis, A Min Tjoa, and Ismail Khalil（编）。Springer International Publishing, Cham, 179–188.[61] Paul Gillin. 2021. 数据仓储存在问题，数据网格或许能解决。https://siliconangle.com/2021/08/06/data-warehousing- problems-data-mesh-solution/ [访问日期：3月 2023]。

[62] Jason Gilmore. 2019. 借助 DreamFactory 数据网格简化数据集市。
https://blog.dreamfactory.com/easier-data-marts-with-dreamfactory-data- mesh/[访问日期：3月 2023].[63] Joe Gleinser. 2021. 你的应用程序是否应考虑采用数据网格连接？ https://www.forbes.com/sites/forbestechcouncil/2020/05/07/should-your-application-consider-data-mesh-connectivity/?sh=5ab427756d7f[访问日期：3月 2023].[64] Mathias golombek. 2021. 实践中的数据网格：从客户旅程中获得的经验。https://www.dataversity.net/data-mesh-in-practice-learnings- from-a-customer-journey/ [访问日期：3月 2023].[65] Adam Gugliciello. 2021. 迈向企业数据网格之旅。
https://blog.dataiku.com/the-journey-to-an-enterprise-data-mesh [访问日期：3月 2023].[66] Muhammad Hafiz Hasan, Mohd Hafeez Osman, Novia Indriaty Admodisastro, and Muhamad Sufri Muhammad. 2023. 从遗留系统迁移到云：架构视角综述。《系统与软件杂志》202 (2023), 111702. https://doi.org/10.1016/j.jss.2023.111 702[67] Trey Hicks. 2021. 在数据网格中捕获数据：原则（第一部分）。https://medium.com/nerd-for-tech/catching-data-in-a-data-mesh-principles- part-i-2b2e11e9e33a [访问日期：3月 2023].[68] Y. Hooshmand, J. Resch, P. Wischnewski, and P. Patil. 2022. 从单体式PLM格局迈向联邦式领域和数据网格。《设计学会论文集》2 (2022), 713–722.[69] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. 数据治理：为可信人工智能组织数据。《政府信息季刊》37, 3 (2020), 101493.[70] Can Yurtseven Jarvin Mutatiina and Ernst Blaauw. 2020. 从数据混乱走向数据网格。https://www2.deloitte.com/nl/nl/pages/strategy-analytics- and-ma/articles/from-data-mess-to-a-data-mesh.html [访问日期：3月 2023].[71] Larysa Vsiengeriyeva Jochen Christ and Simon Harrer. 2022. 数据网格架构的工程视角。
https://www.datamesh- architecture.com/[访问日期：3月 2023].[72] 摩根大通. 2021. 数据网格架构的演进可为现代企业创造显著价值。https://www.jpmorgan.com/technology/te chnology-blog/evolution-of-data-mesh-architecture [访问日期：3月 2023].[73] Kainos. 2021. 实施数据网格前应提出的6个问题。https://www.kainos.com/insights/blogs/thinking-about-data-mesh [访问日期：3月 2023].[74] Keboola. 2021. 数据网格：集中式数据架构失效问题的答案。https://www.keboola.com/blog/data-mesh [访问日期：3月2023].[75] Staffs Keele 等. 2007. 软件工程系统性文献综述实施指南。(2007).[76] 毕马威. 2021. 混合云数据网格。
https://www.kpmg.us/alliances/kpmg-ibm/data-mesh-hybrid-cloud.html[访问日期：3月 2023].[77] Tim Kraska, Tianyu Li, Samuel Madden, Markos Markakis, Amadou Ngom, Ziniu Wu, and Geoffrey X. Yu. 2023. 看看 BRAD 这个大脑：借助学习型自动化数据网格简化云数据处理。《VLDB基金会会刊》16, 11 (2023年7月), 3293–3301. https://doi.org/10.14778/3611479.3611526[78] Saurabh Kumar. 2022. 解构数据网格原则。https://medium.com/slalom-data-ai/data-mesh-232e50f42e66[访问日期：3月 2023].[79] Indika Kumara, Martín Garriga, Angel Urbano Romeu, Dario Di Nucci, Fabio Palomba, Damian Andrew Tamburri, and Willem-Jan van den Heuvel. 2021. 基础设施代码的注意事项：系统性灰色文献综述。《信息与软件技术》137 (2021), 106593.[80] Indika Kumara, Jun Han, Alan Colman, and Malinda Kapuruge. 2017. 软件定义服务网络：多租户共享云应用中的性能差异化。《IEEE服务计算汇刊》10, 1 (2017), 9–22.[81] Martin Lam. 2021. 为什么 Snowflake 适合用于实施数据网格。
https://www.capgemini.com/no-no/2021/05/why-snowflake-is-a- good-match-for-implementing-data-mesh/ [访问日期：3月 2023].[82] Christian Lennerholt, Joeri Van Laere, and Eva Söderström. 2018. 自助式商业智能的实施挑战：文献综述。(2018), 5055–5063.[83] Kevin Lewis. 2021. 数据网格及其维系纽带。https://www.teradata.com/Blogs/Data-Mesh-and-the-Threads-that-Hold-it-Together[访问日期：3月 2023].[84] Jin Li, Shixia Cai, Lei Wang, Mingyang Li, Jiamao Li, and Hui Tu. 2022. 基于数据网格理念的园区级电力系统数据处理框架创新设计。载于《2022年IEEE能源互联网国际会议（ICEI）》。153–158. https://doi.org/10.1109/ICEI57064.2022.00032[85] Wubin Li, Yves Lemieux, Jing Gao, Zhuofeng Zhao, and Yanbo Han. 2019. 服务网格：挑战、技术现状与未来研究机会。载于《2019年IEEE面向服务系统工程国际会议（SOSE）》。IEEE, 122–1225.[86] Rufus Lidman. 2021. 数据网格的权变模型。https://towardsdatascience.com/the-contingency-model-of-data-mesh- c4bbe57577d6[访问日期：3月 2023].[87] Advancing Analytics Limited. 2021. 数据网格深入解析。
https://www.advancinganalytics.co.uk/blog/2021/8/5/data-mesh-deep-dive[访问日期：3月 2023].[88] Antti Loukiala, Juha-Pekka Joutsenlahti, Mikko Raatikainen, Tommi Mikkonen, and Timo Lehtonen. 2021. 从集中式数据仓库迁移到去中心化数据平台架构。载于《以产品为中心的软件过程改进》， Luca Ardito, Andreas Jedlitschka, Maurizio Morisio, and Marco Torchiano（编）。Springer International Publishing, Cham, 36–48.[89] Inês Machado, Carlos Costa, and Maribel Yasmina Santos. 2021. 数据驱动型信息系统：数据网格范式转变。(2021).

[90] Inês Araújo Machado、Carlos Costa 和 Maribel Yasmina Santos。2022年。《数据网格：数据架构范式转变的概念与原则》。Procedia Computer Science 196（2022年），263–271。[91] Anil Madan。2021年。《现代数据平台——如何构建？》https://www.linkedin.com/pulse/modern-data-platform-how-build-one-anil- madan/[访问日期：3月 2023]。[92] Mansi Maharana。2021年。《在 GCP 上使用分布式数据架构和智能数据平台实现数据即产品（DaaP）》。https: //
m
edi
um.com/google-cloud/implementing-data-as-a-product-daap-using-distributed-data-architecture-and-smart-data-platform-on-c2fcd64c67d5[访问日期：3月 2023]。[93] Valdas Maksimavicius。2021年。《在 Lf 启动 Databricks》。https://medium.com/if-tech/launching-databricks-at-if-819be388aa8a[访问日期：3月2023]。[94] Schultze Max 和 Wider Arif。2020年。《实践中的数据网格：欧洲领先的时尚在线平台如何超越数据湖》。https://www.youtube.com/watch?v=eiUhV56uVUc[95] Alexis McKenzie。2021年。《数据仓储的数据网格方法》。https://towardsdatascience.com/a-data-mesh-approach-to-data-warehousing-cd71e55490ba[访问日期：3月 2023]。[96] Matt McLarty。2021年。《API 管理如何与数据网格相结合，嗯？》https://blogs.mulesoft.com/api-integration/api-management-and-data- mesh/[访问日期：3月 2023]。[97] Jesse Menning。2021年。《采用事件驱动数据网格，避免淹没在（数据）湖中》。https://solace.com/blog/event-driven-data-mesh/[访问日期：3月 2023]。[98] Ralph Mietzner 和 Frank Leymann。2010年。《面向基于服务的应用程序的自助服务门户》。载于《2010年 IEEE 面向服务的计算与应用国际会议（SOCA）》。1–8。[99] Pawel Mitrus。2021年。《数据网格详解：最新数据架构趋势的端到端指南》。https://lingarogroup.com/blog/data-mesh- explained-an-end-to-end-guide-to-the-latest-data-architecture-trend [访问日期：3月 2023]。[100] Adit Modi。2021年。《数据网格简介》。https://dev.to/aws-builders/introduction-to-data-mesh-3f1b[访问日期：3月 2023]。[101] Kief Morris。2020年。《基础设施即代码》，第2版。O'Reilly Media, Inc。[102] Barr Moses。2020年。《什么是数据网格——以及如何避免把它搞砸》。https://towardsdatascience.com/what-is-a-data-mesh-and-how-not-to-mesh- it-up-210710bb41e0 [访问日期：3月 2023]。[103] Barr Moses。2021年。《数据网格入门：如何开始》。https://www.montecarlodata.com/blog-data-mesh-101-everything-you-need-to-know-to-get- started/[访问日期：3月 2023]。[104] Barr Moses。2022年。《如何将数据视为产品》。https://towardsdatascience.com/how-to-treat-your-data-like-a-product-73731ec5f131[访问日期：3月 2023]。[105] Aiswarya Raj Munappy、David Issa Mattos、Jan Bosch、Helena Holmström Olsson 和 Anas Dakkak。2020年。《从临时数据分析到数据运维》。载于《软件与系统过程国际会议论文集》（韩国首尔）（ICSSP '20）。计算机协会，纽约，美国，165–174。[106] Le Duy Ngan 和 Rajaraman Kanagasabai。2013年。《语义网服务发现：技术现状与研究挑战》。Personal and Ubiquitous Computing 17, 8（2013年12月1日），1741–1752。[107] Francois Nguyen。2021年。《迈向数据网格（第2部分）：架构与技术》。https://francois-nguyen.blog/2021/03/22/toward-a-data- mesh-part-2-architecture-technologies/[访问日期：3月 2023]。[108] Zach Mitchell Nivas Shankar、Ian Meyers 和 Roy Hasson。2021年。《使用 AWS Lake Formation 和 AWS Glue 设计数据网格架构》。https://aws.amazon.com/blogs/big-data/design-a-data-mesh-architecture-using-aws-lake-formation-and-aws-glue/ [访问日期：3月 2023]。[109] Abraham Enyo one Musa。2020年。《数据管理架构——单体数据架构与分布式数据网格》。https: //
t
o
wardsdatascience.co
m/data-management-architectures-monolithic-data-architectures-and-distributed-data-mesh-63743794966c[访问日期：3月 2023]。[110] Gerben Oostra。2021年。《数据网格说明》。https://medium.com/bigdatarepublic/two-steps-towards-a-modern-data-platform- 37c74e7c104b[访问日期：3月 2023]。[111] Gerben Oostra。2021年。《迈向现代数据平台的两步》。https://medium.com/bigdatarepublic/two-steps-towards-a-modern-data-platform- 37c74e7c104b[访问日期：3月 2023]。[112] Oracle。2021年。《企业数据网格》。https://www.oracle.com/a/ocom/docs/datamesh-ebook.pdf [访问日期：3月 2023]。[113] Oracle。2021年。《技术简报：借助 Oracle GoldenGate 平台实现动态数据织物与可信数据网格》。https://www.oracle.com/a/oco
m/docs/techbrief-enterprisedatameshandgoldengate.pdf[访问日期：3月 2023]。[114] Einat Orr。2020年。《数据网格应用：如何借助 lakeFS 超越数据湖》。https://lakefs.io/data-mesh-applied-how-to-move-beyond- the-data-lake-with-lakefs/[访问日期：3月 2023]。[115] Patryk Orwat。2021年。《AWS 上的数据网格》。https://dev.to/aws-builders/data-mesh-on-aws-57ah[访问日期：3月 2023]。[116] Michael Papazoglou。2008年。《Web 服务：原则与技术》。Pearson Education。[117] Michael P. Papazoglou、Paolo Traverso、Schahram Dustdar 和 Frank Leymann。2007年。《面向服务的计算：技术现状与研究挑战》。Computer 40, 11（2007年），38–45。[118] Mike P Papazoglou 和 Willem-Jan Van Den Heuvel。2007年。《面向服务的架构：方法、技术与研究问题》。The VLDB journal 16, 3（2007年），389–415。稿件已提交至 ACM

[119] Courtney Perio。2022年。《充分利用数据分析微服务》。https://towardsdatascience.com/get-more-out-of-your-data-with-analytics-microservices-9a5a34a3ad2f[访问日期：3月 2023]。[120] Yval Perlov。2021年。《数据网格：通过数据编织实现的架构、用例与落地》。https://www.k2view.com/blog/data-mesh/ [访问日期：3月 2023]。[121] Paolo Platter。2021年。《数据网格如何以及为何推动数据管理演进》。https://medium.com/agile-lab-engineering/yet-another- datamesh-article-83378b62f334[访问日期：3月 2023]。[122] Paolo Platter。2021年。《如何识别数据产品？欢迎了解“数据产品流”》。https://medium.com/agile-lab-engineering/how-to-identify- data-products-welcome-data-product-flow-76d7d85d23af[访问日期：3月 2023]。[123] Nikolai J Podlesny、Anne VDM Kayem 和 Christoph Meinel。2022年。《综述：数据网格相关的隐私挑战调查》。载于《数据库与专家系统应用国际会议》。施普林格，85–102。[124] Jeffrey Pollock。2021年。《数据网格不是数据湖！》https://www.linkedin.com/pulse/data-mesh-lake-jeffrey-t-pollock/ [访问日期：3月 2023]。[125] Taavi Pungas。2021年。《避开数据瓶颈——Starship 的数据网格》。https://medium.com/starshiptechnologies/dodging-the-data-bottleneck-data-mesh-at-starship-5925a2de45e6[访问日期：3月 2023]。[126] Ilan Raab 和 Marco Chiapusso。2021年。《如何借助数据网格构建现代 CPG 数据架构》。https://aws.amazon.com/blogs/industries/how- to-create-a-modern-cpg-data-architecture-with-data-mesh/[访问日期：3月 2023]。[127] Rajesh Rajagopalan。2021年。《揭开数据网格的神秘面纱》。https://www.peerislands.io/demystifying-data-mesh/[访问日期：3月 2023]。[128] Asha Rajbhoj、Akanksha Somase、Piyush Kulkarni 和 Vinay Kulkarni。2024年。《利用生成式人工智能加速软件开发：ChatGPT 案例研究》。载于《第17届软件工程创新会议论文集》（印度）（ISEC ’24）。计算机协会，纽约，美国，文章5，共11页。[129] Karthik Ramachandran。2020年。《数据管理对人工智能成功的阻碍》。https://www2.deloitte.com/xe/en/insights/industry/technology/ai-and-data- management.html[130] Javier Ramos。2021年。《构建数据网格：初学者指南》。https://itnext.io/introduction-to-data-mesh-59e6f3a4c15e[访问日期：3月 2023]。[131] Javier Ramos。2022年。《什么是数据产品？》https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/architectures/what-is-data-product[访问日期：3月 2023]。[132] Fernando Raposo。2021年。《FindHotel 数据架构的演进——第一部分》。https://blog.findhotel.net/2021/07/the-evolution-of-findhotels- data-architecture-part-i/[访问日期：3月 2023]。[133] Azaz Rasool。2021年。《数据仓库/数据湖正在消亡，还是需要全新视角？》https://www.linkedin.com/pulse/datamesh-paradigm-shift- data-warehouses-lakes-dying-azaz-rasool/[访问日期：3月 2023]。[134] James Reid。2021年。《在摩根大通实施数据网格架构》。https://www.dremio.com/subsurface/implementing-a-data-mesh-architecture- at-jpmc/[访问日期：3月 2023]。[135] Chris Riccomini。2021年。《数据网格究竟是什么？！》https://cnr.sh/essays/what-the-heck-data-mesh [访问日期：3月 2023]。[136] Aleksi Roima。2021年。《组织何时应考虑采用数据网格？》https://futurice.com/blog/when-should-organizations-consider-data-mesh[访问日期：3月 2023]。[137] David Reinsel-John Gantz-John Rydning、J Reinsel 和 J Gantz。2018年。《从边缘到核心的世界数字化》。马萨诸塞州弗雷明汉：国际数据公司 16（2018年）。[138] Marcelo Iury S. Oliveira、Glória de Fátima Barros Lima 和 Bernadette Farias Lóscio。2019年。《数据生态系统研究：一项系统映射研究》。知识与信息系统 61（2019年），589–630。[139] Frank E Saal、Ronald G Downey 和 Mary A Lahey。1980年。《评估评分：评定评分数据的心理测量质量》。心理学通报 88, 2（1980年），413。[140] Johnny Saldaña。2021年。《定性研究者编码手册》。赛奇出版社。[141] Saket Saurabh。2021年。《数据网格：设计、优势、炒作与现实》。https://www.nexla.com/data-mesh-design-benefits-hype-reality/[访问日期：3月 2023]。[142] Max Schultze 和 Arif Wider。2020年。《实践中的数据网格》。https://www.iteblog.com/ppt/sparkaisummit-north-america-2020-iteblog/data-mesh-in-practice-how-europes-leading-online-platform-for-fashion-goes-beyond-the-data-lake-iteblog.com.pdf[访问日期：3月 2023]。[143] Max Schultze 和 Arif Wider。2020年。《实践中的数据网格：欧洲领先的时尚在线平台如何超越数据湖》。https://databricks.com/session_na20/data-mesh-in-practice-how-europes-leading-online-platform-for-fashion-goes-beyond-the-data-lake[访问日期：3月 2023]。[144] Juan Sequeda。2021年。《数据网格究竟是怎么回事？》http://www.juansequeda.com/blog/2021/02/22/what-is-the-deal-with-the-data- mesh/[访问日期：3月 2023]。[145] James Serra。2021年。《数据网格定义》。{https://www.jamesserra.com/archive/2021/02/data-mesh/}[访问日期：3月2023]。[146] Nazia Shahrin。2021年。《构建成功的数据网格——不仅仅是一项技术举措》。https://www.linkedin.com/pulse/building-successful-data-mesh-more-than-just-nazia-shahrin/[访问日期：3月 2023]。[147] Quan Z. Sheng、Xiaoqiang Qiao、Athanasios V. Vasilakos、Claudia Szabo、Scott Bourne 和 Xiaofei Xu。2014年。《Web 服务组合：十年综述》。信息科学 280（2014年），218–238。[148] Harcharan Jit Singh 和 Seema Bawa。2018年。《面向超大规模分布式存储系统的可扩展元数据管理技术——系统综述》。ACM计算调查 51, 4，文章 82 （ 2018 年 7 月），共 37 页。稿件已提交至 ACM

[149] Apitchaka Singjai、Uwe Zdun 和 Olaf Zimmermann。2021。微服务 API 与领域驱动设计之间相互关系的实践者观点：一项基于扎根理论的灰色文献研究。载于《2021 年 IEEE 第 18 届国际软件架构大会（ICSA）》。25–35。[150] Snowflake。2021。借助基于 Snowflake 构建的数据网格赋能数据团队。https://www.snowflake.com/blog/empower-data-teams-with-a-data-mesh-built-on-snowflake/[访问日期：3 月 2023]。[151] Graham Stirling。2021。盛宝银行基于数据网格构建分布式领域驱动架构的最佳实践。https://www.confluent. io/blog/distributed-domain-driven-architecture-data-mesh-best-practices/[访问日期：3 月 2023]。[152] Michael Stonebraker、Ihab F Ilyas 等。2018。数据集成：现状与未来方向。《IEEE 数据工程通报》41，2（2018），3–9。[153] Anselm Straus 和 Juliet Corbin。2014。《定性研究基础：扎根理论程序与技术》。[154] Piethein Strengholt。2020。荷兰银行的数据与集成网格。https://www.linkedin.com/pulse/abn-amros-data-integration-mesh-piethein- strengholt/[访问日期：3 月 2023]。[155] Piethein Strengholt。2021。数据网格拓扑。https://towardsdatascience.com/data-mesh-topologies-85f4cad14bf2[访问日期：3 月 2023]。[156] Piethein Strengholt。2022。数据契约——确保数据网格架构的稳健性。https://towardsdatascience.com/data-contracts-ensure- robustness-in-your-data-mesh-architecture-69a3c38f07db[访问日期：3 月 2023]。[157] Piethein Strengholt。2022。数据领域——从何处着手? https://towardsdatascience.com/data-domains-where-do-i-start-a6d52fef95d1[访问日期：3 月 2023]。[158] Piethein Strengholt。2022。数据领域与数据产品。https://towardsdatascience.com/data-domains-and-data-products-64cc9d2828 3e[访问日期：3 月 2023]。[159] Piethein Strengholt。2022。数据网格：集中化与去中心化之间的平衡。https://towardsdatascience.com/data-mesh-the-balancing-act-of-centralization-and-decentralization-f5dc0bb54bcf[访问日期：3 月 2023]。[160] Piethein Strengholt。2022。在 Azure 上实施数据网格。https://towardsdatascience.com/implementing-data-mesh-on-azure- c01ee94306cd[访问日期：3 月 2023]。[161] Piethein Strengholt。2022。数据网格中的主数据管理。https://towardsdatascience.com/master-data-management-in-data-mesh- 594d21f3ee10[访问日期：3 月 2023]。[162] Piethein Strenholt 和 Andrea Courtright。2022。数据网格的金融机构场景。https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/data-management/architectures/reference-architecture-data-mesh[访问日期：3 月 2023]。[163] Firat Tekiner。2021。在 Google Cloud 上构建统一的分析数据平台。https://cloud.google.com/blog/products/data-analytics/building- unified-analytics-data-platform-google-cloud[访问日期：3 月 2023]。[164] Meenakshisundaram Thandavarayan。2021。什么是数据网格（以及谁应该使用它）？https://dataintegration.info/what-is-data-mesh-and- who-should-be-using-it [访问日期：3 月 2023]。[165] Thoughtworks。2020。数据网格白皮书：数据网格转型。https://www.thoughtworks.com/insights/whitepapers/the-data-mesh- shift[访问日期：3 月 2023]。[166] Thoughtworks。2021。利用数据网格实现新加坡公共部门可信数据共享。https://www.thoughtworks.com/content/dam/thoughtworks/documents/whitepaper/whitepaper-tw-102021-ebook-data-mesh-singapore.pdf[访问日期：3 月 2023]。[167] Davide Tosi 和 Sandro Morasca。2015。支持 Web 服务半自动语义标注：一项系统性文献综述。《信息与软件技术》61（2015），16–32。[168] Hong-Linh Truong 和 Tri-Minh Nguyen。2021。QoA4ML——支持机器学习服务契约的框架。载于《2021 年 IEEE 国际 Web 服务大会（ICWS）》。IEEE，465–475。[169] Sandeep Uttamachandani。2022。自助式数据平台的 Google 技术格局。https://www.linkedin.com/pulse/landscape- technologies-self-service-data-platform-sandeep [访问日期：3 月 2023]。[170] Dave Velante。2021。数据新时代：深入了解摩根大通如何在 AWS 云上运行数据网格。https://siliconangle.com/2021/07/10/new-era-data-deep-look-jpmorgan-chase-runs-data-mesh-aws-cloud/[访问日期：3 月 2023]。[171] David Vellante。2021。分析深度报道：摩根大通如何在 AWS 云上实施数据网格。https://wikibon.com/breaking-analysis-how-jp-morgan-is-implementing-a-data-mesh-on-the-aws-cloud/ [访问日期：3 月 2023]。[172] Padmal Vitharana 和 Shahir A. Daya。2024。采用并持续推进基于微服务的软件开发：组织挑战可能比技术挑战更难应对。《ACM 通讯》（2024 年 3 月），9 页。[173] Stephen John Warnett 和 Uwe Zdun。2022。机器学习部署的架构设计决策。载于《2022 年 IEEE 第 19 届国际软件架构大会（ICSA）》。90–100。[174] Yi Wei 和 M. Brian Blake。2010。面向服务的计算与云计算：挑战与机遇。《IEEE 互联网计算》14，6（2010），72–75。[175] Dave Wells。2021。数据架构：复杂与繁杂。https://www.eckerson.com/articles/data-architecture-complex-vs- complicated[访问日期：3 月 2023]。[176] Claes Wohlin、Per Runeson、Martin Höst、Magnus C Ohlsson、Björn Regnell 和 Anders Wesslén。2012。《软件工程实验》。Springer 科学与商业媒体。[177] Tom De Wolf。2022。将数据网格原则应用于物联网数据架构。https://www.acagroup.be/en/blog/applying-data-mesh-principles- to-an-iot-data-architecture/[访问日期：3 月 2023]。

[178] Olivier Wulveryck。2021。POV：面向数据网格的流式/通信平台。https://blog.octo.com/en/pov-a-streaming-communication-platform-for-the-data-mesh/ [访问日期：3 月 2023]。[179] Frank F. Xu、Bogdan Vasilescu 和 Graham Neubig。2022。通过自然语言进行 IDE 内代码生成：前景与挑战。《*ACM* 软件工程方法论汇刊》31，2，文章 29（2022 年 3 月），47 页。[180] Doichin Yordanov。2021。数据网格：即将兴起的新数据范式。https://helecloud.com/blog/data-mesh-the-new-data-paradigm-set-to-rise/[访问日期：3 月 2023]。[181] Ammar Zeini、Ruth G. Lennon 和 Patrick Lennon。2023。基础设施即代码安全方法的初步调查。载于《第八届国际信息与通信技术大会论文集》。施普林格·自然新加坡，新加坡，763–783。[182] Kun Zhang、Jiali Zeng、Fandong Meng、Yuanzhuo Wang、Shiqi Sun、Long Bai、Huawei Shen 和 Jie Zhou。2024。用于大语言模型复杂问题回答的思维树问题分解。载于《*AAAI* 人工智能大会论文集》，第 38 卷。19560–19568。[183] Chenxing Zhong、Shanshan Li、Huang Huang、Xiaodong Liu、Zhikun Chen、Yi Zhang 和 He Zhang。2024。微服务的领域驱动设计：一项基于证据的调查。《*IEEE* 软件工程汇刊》（2024），1–25。