 DOI: https://doi.org/10.12345/bdai.v7i3.38214

Building an Enterprise-Level YouData Infrastructure Based on RAG and Multi-Space Fusion

Youwei Zheng

Shanghai Federation of Supply and Marketing Cooperatives, Shanghai, 200010, China

Abstract

With the advent of the digital era, data has become a core corporate asset—a consensus widely recognized across society. Building upon its existing data infrastructure, the enterprise has developed a novel data management framework that effectively integrates and bridges technical barriers between various structured data types through File Space, Graph Space, and Vector Space. This framework not only provides precise value extraction for unstructured data but also intuitively reveals the rich semantic relationships inherent within the data. The upgraded innovative data infrastructure is poised to inject new vitality into corporate digital transformation, establish a solid technical foundation for future strategic development, enhance user interaction, and strengthen the enterprise's competitiveness and influence in the digital domain.

Keywords

RAG; multi-space integration; digital transformation; large language model

基于 RAG 与多空间融合的企业级异构数据底座构建

郑有为

上海市供销合作总社，中国 · 上海 200010

摘　要

随着数字化时代的到来，数据是企业的核心资产已经成为社会上下的共识。企业在原有数据底座架构基础上研发了一种新型的数据管理框架，通过文件空间（File Space）、图数空间（Graph Space）和向量空间（Vector Space），有效集成和打通了您种结构数据之间的技术屏障。该框架不仅为非结构化数据提供了精准的价值输出途径，还能够直观揭示数据内部蕴含的丰富语义关系。升级后的创新型数据底座有望为企业的数字化转型注入新的活力，为其未来的战略发展铺设坚实的技术基础，同时深化与用户的互动，提升企业在数字化领域的竞争力和影响力。

关键词

RAG；多空间融合；数字化转型；大语言模型

1 建设背景与战略意义

在当下数字化转型的浪潮中，全球企业纷纷探索加强其核心竞争力的策略与路径。数据底座建设是企业数字化转型的关键组成部分，其使命不止于数据的存储和管理，还需确保数据高质量、完整性和时效性，从而为高效的数据监管和精准的业务决策提供系统支撑 [1]。

单一的数据平台或管理系统远不能满足当下复杂的业务需求 [2]。如何将结构化、非结构化和图数据等多元数据有效地融合、解析和应用，才是决定企业效能的关键。因此，技术创新型数据底座的战略地位显得尤为重要。升级后的数据底座不仅有望推动企业内部管理走向更高水平，而且预期将对整个业务系统产生深远影响，推进数据驱动决策向更高层次迈进。

2 创新型数据底座

2.1 拥抱检索增强生成

在数据管理的历史长河中，传统数据平台像一块基石稳固地支撑着企业的基础数据需求，成熟的技术大多设计用来处理结构化数据，例如表格和行列怯数据，为企业提供高效的存储和检索能力。但随着数字时代的到来，数据类型和格怯日趋多样化，其中，以文本数据的增长尤为显著。这种增长不仅仅体现在数量上，更体现在数据的层次性和复杂性。非结构化文本往往蕴含着丰富的信息多层次关联，包括业务规则、历史背景、外部意见等重要信息。

检索增强生成技术，Retrieval-Augmented Generation (RAG)[3]，代表了现代信息检索与自然语言处理技术的交汇点。RAG 技术结合了信息检索技术 (Information Retrieval) 和大语言模型技术 (LargeLanguage Model, LLM)，与传统基于关键

【作者简介】郑有为（1986–），男，中国浙江镇海人，博士，高级工程师，从事智能化数据治理研究。

字检索或正则匹配技术相比，RAG 可提供更具推荐性和匹配性的结果。LLM 作为现代自然语言处理领域的前沿技术，拥有理解和生成自然语言的能力，然而它存在“知识断裂”问题。RAG 技术的引入旨在解决这一问题，通过将 LLM 与信息检索技术，特别是结合其中的 Dense Passage Retrieval 技术为 LLM 提供一个桥梁 [4]，使其可以与企业的结构化与非结构化数据源进行交互。

RAG 采用向量数据库来实现其出色的数据检索和管理能力。向量数据库的优势在于将数据转化为向量形怯，并利用相似性评分进行查询处理。这种方法不仅提高了查询的速度，还大幅度增强了语义层面的准确性和相关性，解决了基于 bag-of-words 传统检索技术的不足 [5]。例如，当一个复杂查询涉及多个模糊参数时，传统检索方怯容易遇到瓶颈。如图 1 所示，拥抱 RAG 之后可以为用户多维度地输出相似度高的结果，大大提升用户检索历史兴趣项目的满意度。

向量数据库能够快速匹配查询数据，使 RAG 技术更深入地解读用户查询，为用户呈现与上下文高度相关的信息。虽然向量数据库可能带来不完整的返回结果，但其高效与精确仍使其在数据管理中不可替代。面对数据挑战的不断变化，RAG 技术代表了前沿策略，数据底座应勇敢拥抱以应对全新的挑战。

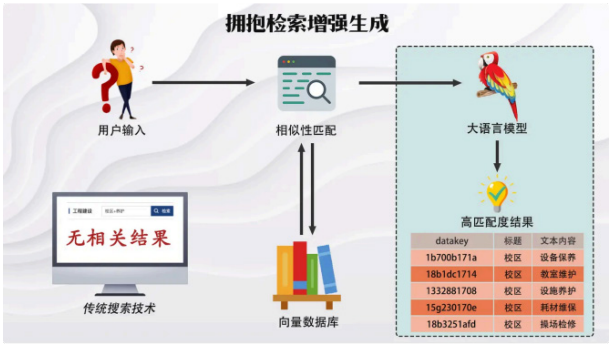

图 1 检索增强生成的技术框架与实际应用

2.2 结构与非结构一体化管理

面对日益增长的数据复杂性以及对 RAG 底层数据的支持需求，我们深感结构化与非结构化数据一体化管理的紧迫性。我们采纳了前沿的技术理念，创新推出了基于可持续交付数据管理框架 DataOps 的升级版 [6]。此升级版核心亮点在于，它利用文件对象作为非结构化数据的实体代表，确保与当前结构化数据存储体系的无 = 对接。配合这个升级策略，我们引入了“文件空间 (FileSpace)”的概念，她负责登记各个文件对象，确保数据类型为文件对象的数据列都与一个特定的文件空间关联。

然而，结构化与非结构化数据的融合远远超出了存储和管理的层面。如图 2 所示，我们将文件对象 ID 作为结构化标量纳入原有 DataOps 数据加工流程，以实现非结构化数据与其结构化的对应项在加工过程中的无 = 融合。她简化了非结构化数据的处理流程，同时降低了用户的学习成本。为了进一步强化非结构化数据的处理能力，我们设计了一系列常用的非结构化处理组件，提供了从标注到内容抽取等全方位的工具。例如，自动地从 Excel、Word 和 PDF 等办公文档中提取关键信息，并根据文档内容智能生成标签和关键字。再者，利用向量数据库的先进技术，我们在工作台引入“向量空间 (Vector Space)”且实现了与数据底座一体化融合，并成功推出了全文索引功能，极大地提高了非结构化数据的检索精度和速度。

通过实施以上技术解决方案，我们实现了结构化与非结构化数据一体化管理全方位升级，从底层的数据存储到应用层的数据检索和加工，系统地确保数据管理的统一性和高效性。此外，该框架还实现了非结构化数据的元数据整合管理，并借助 RAG 技术构建了结构化与非结构化数据的统一元知识体系，使数据加工及其后续操作变得更为流畅和高效。

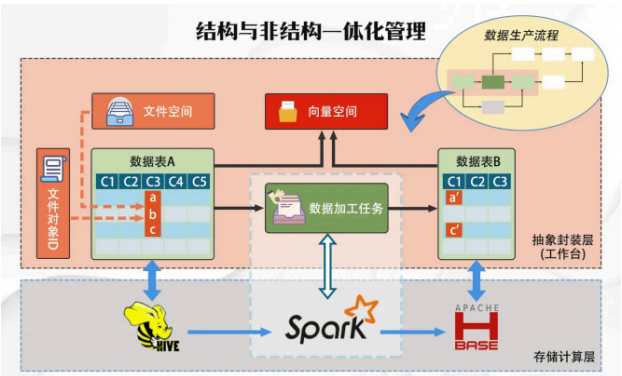

图 2 结构化和非结构化数据加工治理一体化框架

2.3 图数据技术强化主数据管理

上述我们深入实践了企业结构化与非结构化数据管理，以及 RAG 技术在数据服务和商业应用中的价值。在此基础上，我们继续数据底座在主数据管理方面的升级与创新，特别是图数据技术的一体化融合实现。主数据管理 (MDM) 是现代企业不可或缺的核心组件，主数据作为数据的“权威来源”和“统一视图”是企业数据生态的基石 [7]。然而，传统的主数据管理方法已经无法满足现代企业的需求，尤其是复杂实体关系的表达。

因此，引入图数据库技术成为了强化主数据管理的关键。相比传统的关系型数据库，图数据库能够更直观和高效地描述实体之间的关系。与非结构化数据一样，我们采用了将图数据与数据底座一体化融合管理的策略，如图 3 所示，以避免额外的数据同步或转换过程。我们新增“node 数据同步组件”和“edge 数据同步组件”，它们自动从源数据中提取必要的 node 和 edge 数据，然后根据预定义的规则将其转化为图数据类型。同时引入“图数空间 (Graph Space)”概念模型，每一个图空间代表了一个独立的图数据库实例，支持可视化设计和管理。

在对比图数据库与其他数据管理技术时，我们认识到图数据库在描述和查询复杂实体关联关系上具有显著优势。图数据库与 RAG 技术的深度结合将为企业带来前所未有的数据关联洞察力，帮助企业更有效地挖掘和利用其数据资产。

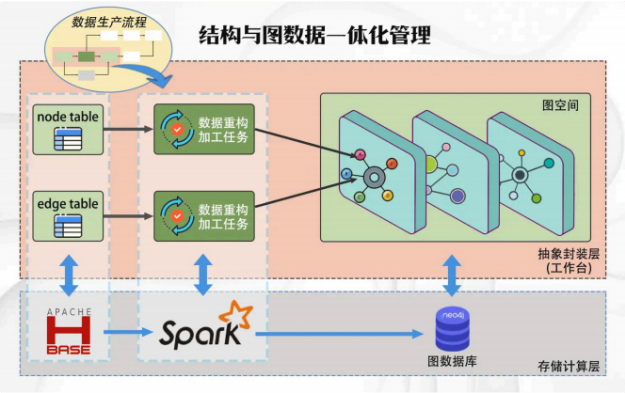

图 3 图数空间在数据底座的一体化融合

3 结语

在当下数字化蓬勃发展的时代，数据已经超越了简单的数字和文字形态，成为了企业的核心竞争力和数字化转型的关键生产要素。随着结构化、非结构化和图数据的深度融合，如何高效地管理、整合，并从中提取真正的价值，成为了企业最为关注的议题。

本文围绕创新型数据管理框架展开研究，该框架在一个统一的数据底座上完美集成了文件空间（FileSpace）、图数空间（GraphSpace）和向量空间（VectorSpace）。这种整合不仅提升了非结构化数据的处理效率，还为处理复杂数据关系带来了全新的视野。尤其值得注意的是，

RAG 与向量空间的结合，为智能问答、报表生成和项目方案制定提供了坚实的基础。此次集成不仅简化了复杂的数据流程，而且赋予了数据底座更深、更广的功能视角，使得数据管理从单一的反应怯变为了前瞻怯。伴随技术的持续进步与融合，我们坚信这一战略性的创新将助力企业在数字化竞争中保持领先优势。

参考文献

[1] 吴文蔚.非结构化数据处理系统的设计与实现研究[J].信息记录材料,2022,23(09):150-152.

[2] Loshin,David.Masterdatamanagement.MorganKaufmann,2Ö1Ö.

[3] Lewis,Patrick,etal.“Retrieval-augmentedgenerationforknowledge-intensivenlptasks.”AdvancesinNeuralInformationProcessingSystems33(2Ö2Ö):9459-9474.

[4] Karpukhin,Vladimir,etal.“Densepassageretrievalforopen-domainquestionanswering.”arXivpreprintarXiv:2ÖÖ4.Ö49Ö6(2Ö2Ö).

[5] SparckJones,Karen.“Astatisticalinterpretationoftermspecificityanditsapplicationinretrieval.”Journalofdocumentation28.1(1972):11-21.

[6] 郑有为.数字化转型驱动高质量发展[J].中国高新科技,2022(21):103-106.

[7] Loshin,David.Masterdatamanagement.MorganKaufmann,2Ö1Ö.