

数据血缘关系构建方法综述

吕琳¹, 田庆业², 焦冬冬¹, 郭金磊¹, 房志奇¹, 陈瑞¹

(1. 华北计算机系统工程研究所, 北京 100083;

2. 中电智能科技有限公司, 北京 102200)

摘要: 随着数据量的急剧增长, 如何管理和利用数据面临严峻挑战。而数据血缘作为数据治理的核心组成部分, 在数据治理中有重要作用, 如提升数据质量、保障数据安全等。研究了构建数据血缘关系的方法, 包括系统跟踪法、基于 SQL 解析的方法、逆置函数法、标注法和机器学习法, 分析了不同数据血缘构建方法的优缺点和应用场景, 并探讨未来研究方向, 为数据血缘关系的应用及后续研究提供参考。

关键词: 数据血缘; 元数据; 数据治理; 大数据

中图分类号: TP301

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2025.12.001

引用格式: 吕琳, 田庆业, 焦冬冬, 等. 数据血缘关系构建方法综述 [J]. 网络安全与数据治理, 2025, 44(12): 1-5.

A review of data lineage relationship construction methods

Lv Lin¹, Tian Qingye², Jiao Dongdong¹, Guo Jinlei¹, Fang Zhiqi¹, Chen Rui¹

(1. National Computer System Engineering Research Institute of China, Beijing 100083, China;

2. Intelligence Technology of CEC Co., Ltd., Beijing 102200, China)

Abstract: With the rapid growth of data volume, how to manage and utilize data is facing severe challenges. As a core component of data governance, data lineage plays an important role in data governance, such as improving data quality and ensuring data security. Therefore, this paper studies the construction methods of data lineage, including system tracking method, SQL parsing-based method, inverse function method, annotation method and machine learning method. It also explores the advantages and disadvantages of different data lineage construction methods and their application scenarios, and discusses future research directions, providing a reference for the application of data lineage relationships and subsequent research.

Key words: data lineage; metadata; data governance; big data

0 引言

在当今数据驱动的时代, 数据已成为企业核心资产之一^[1]。随着数据规模的指数级增长, 如何管理和利用数据成为挑战^[2]。

数据血缘 (Data Lineage) 作为数据治理的核心组成部分, 旨在通过追踪数据从采集、加工、存储到消费的全生命周期路径, 揭示数据的演化关系与依赖链条, 优化资源配置, 从而提高公司的决策水平。

近年来, 学术界形成了多种数据血缘构建方法。然而, 不同方法在自动化程度、粒度等方面仍存在显著差异。本文梳理了目前构建数据血缘关系的方法, 对比分析其优缺点和应用场景, 并探讨未来研究方向, 为数据血缘关系的应用及后续研究提供参考。

1 数据血缘相关概念

1.1 数据血缘概念

数据血缘, 又称为数据血统、数据起源、数据谱系, 用于描述数据与数据之间的关系^[3], 具体是指数据在进行传输、加工、处理时, 表、字段、数据之间存在的血缘关系^[4], 呈现数据之间的上下游来源去向——数据从哪里来, 到哪里去。

数据血缘不仅涉及数据的物理流动, 还包括数据的逻辑关系和转换过程。数据血缘是以某数据对象为起点, 追踪数据处理过程, 定位全生命周期内与之关联的数据节点, 并构建这些节点间的关系。数据血缘记录了数据起源、处理的整个过程^[5], 对于理解数据的来源、加工方式、映射关系以及数据出口至关重要。

1.2 数据血缘层次

数据血缘具有层次性,结构化数据血缘通常可以分为四个层次,如图1所示。数据库、表和字段,是数据的存储结构。

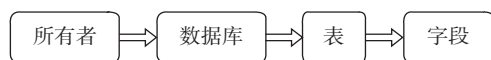


图1 数据血缘关系层次结构图

对于非结构化数据,以文件服务器存储的非结构化数据为例,数据血缘层次为所有者、文件服务器、文件目录、文件。

1.3 数据血缘关系在数据治理中的作用

数据血缘关系在数据治理中发挥着重要作用:

(1) 提升数据质量。数据血缘分析能够揭示数据元素间的深层关联,展现数据的来源、处理过程和变化情况,动态监控数据质量,快速发现数据异常或错误,确保数据准确性和完整性。

(2) 保障数据安全。数据血缘展现了数据的流转路径,通过监控数据流向,敏感数据得到适当保护,从而保障数据安全。例如,Porkodi等^[6]在物联网设备数据采集集中引入数据血缘,结合混合加密与区块链技术,有效防范未授权访问和数据篡改。

(3) 优化数据管理流程。数据血缘关系通过标识数据处理的各个环节,促进了数据管理流程的优化,确保了数据的可追溯性和可审计性。

(4) 提升数据问题排查效率。当应用服务发生异常时,快速定位问题是关键。依赖人工逐行排查代码效率低下,显著延长故障恢复时间。数据血缘关系可以辅助进行数据问题排查与运维响应,从而快速定位并解决问题。

2 发展历史及研究进展

国内外对数据血缘的研究主要经历两个阶段。

2.1 概念研究

早期是概念研究阶段。1990年Wang等^[7]首次提出数据血缘,数据血缘被认为是数据及其在数据库间运动的起源。1991年Lanter^[8]首次用“Data Lineage”(数据血缘)描述数据流转过程。在此基础上,1997年Woodruff等^[9]将数据血缘的概念拓展为数据信息处理过程的集合。

2000年,Cui等^[10]认为数据血缘分析是给定具体视图中的数据项,确定生成该数据的源数据以及生成该数据的过程。同年,Buneman等^[11]将其定义为追踪和记录数据及其在数据库之间移动的过程。2001年,他们又从不同角度进行描述,将数据血缘分为为什么来源和在哪里来源两层含义^[12]。自此,数据血缘开始作为正式术语使用。

2003年,Greenwood等^[13]认为数据血缘是关于实验

过程的元数据,包括数据的派生路径、数据来源和质量等信息。Goble^[14]将数据血缘定义为对数据在其全生命周期中所经历的转换与处理过程的系统性描述。2005年,Simmhan等^[15]认为数据血缘体现数据的演化历史。2007年,Glavic等^[16]定义数据血缘来展现数据项产生、处理过程和源数据信息。2010年,Malik等^[17]认为数据血缘是记录数据起源和转化的元数据。

2.2 构建方法研究

在概念的基础上,学者们开始研究如何构建数据血缘关系。2000年,Cui等^[10]构建数据血缘知识体系,设计针对不同类型数据库查询的追踪算法。在此基础上,2002年,戴超凡^[18]整体性总结数据血缘追踪方法。之后刘喜平等^[19]归纳其常见方法和应用,高明等^[20]概括数据世系管理方法。

同时,也有不少学者研究数据血缘的具体构建方法。2004年,Buneman等^[12]使用辅助数据库作为数据血缘管理的工具,依据使用者对辅助数据库的操作来追踪使用者操作。Li等^[21]设计针对物化视图的数据血缘追踪算法。RAMP基于源码的改造实现了对MapReduce作业的血缘解析。此外还有元数据血缘关系映射技术^[22],以及其他关键技术研究^[23-24]等。

3 数据血缘构建方法

构建数据血缘关系的方法主要有系统跟踪法、基于SQL解析的方法、逆置函数法、标注法及机器学习法。

3.1 系统跟踪法

系统跟踪法是一种通过监控数据系统在运行时产生的行为信息捕获数据流动路径的方法。例如,Spline系统和SAC(SAP Analytics Cloud)系统便是基于此方法的典型系统。该方法的优点是血缘关系精准、细粒度、支持动态追踪;缺点是该方法具有一定的侵入性,且依赖数据治理平台。

3.2 基于SQL解析的方法

基于SQL解析的数据血缘分析,是目前业界的主流方法。该方法通过自动解析SQL脚本,追溯数据在表与字段层级间的依赖与转换关系。2000年Cui提出该方法^[10],但目前还没有很好的方法来验证解析结果是否正确。

其他研究还包括:杨彬^[25]设计了基于SQL抽象语法树的数据血缘分析方法;潘晓华等^[26]提出基于SQL表达式词法分析和抽象语法树的构建方法,将粒度提升至单条数据;滕召嘉^[27]实现了从Hive SQL和Spark SQL中获取数据血缘关系,实现任务级数据血缘分析。

3.3 逆置函数法

逆置函数法是指在已知目标数据生成规则(如计算公式、转换函数)的前提下,通过数学或逻辑逆运算,

推导出生成该数据所需的原始输入集及其关联关系。该方法适用于复杂计算场景的血缘追溯。

逆置函数法的优点是只需存储少量元数据，不需要存储标注信息。缺点是该方法具有局限性，并不是所有函数都可逆，例如哈希函数、随机数生成等不可逆，无法推导。

1997 年，Woodruff 等^[9]对原始查询求逆运算得到源元组。2002 年，戴超凡^[18]依据可逆和弱可逆映射，构造出数据的弱起源集。2010 年 Ikeda 等^[28]以及 2017 年林悦邦^[29]探讨了支持全特性查询语言的逆置函数法。

2017 年，张苒^[30]基于 ETL 中转换操作可逆性的研究，设计逆操作，利用 PROV 模型对数据操作过程及其逆过程进行全面详细分析描述，在保证溯源结果正确的前提下，追求溯源过程高效率。

3.4 标注法^[31]

标注法是通过元数据标注来构建数据血缘关系的方法，数据的出处、流转信息等都可以作为标注，通过查询数据标注，构建数据血缘关系。该方法的优点是实现比较简单，易于理解。缺点是维护成本高，存储开销大，频繁数据标注可能带来额外性能开销，影响处理效率。

1993 年，Hachem^[32]等提出通过中间数据集存储标注信息来实现数据血缘追踪，但人工维护数据转换日志的过程增加了时间与人力成本。

2005 年，Chiticariu^[33]等开发数据血缘分析系统 DB-Notes，通过标注符号的传递完成数据血缘关系的分析。2009 年，Glavic 等^[16]提出基于标注的关系模型原型系统 Mondrian，该系统在 DBNotes 的基础上进行优化，能更灵活地展现血缘关系。在此基础上，2015 年沈晓谦^[34]提出

主键注释法，获取查询语句的层次结构数据，从而构建相应的中间视图和中间表。

2018 年，Senellart 等^[35]设计了半自动化完成血缘解析的 ProvSQL^[36]系统，该系统只需要一次数据标注，就能自动实现标注转换。Green 等^[31]基于代数结构半环构造了能够记录元组的演化过程的起源半环模型，能捕捉元组级的数据血缘关系。

在实际应用中，标注法因存储开销大、处理效率低而受到限制。这一挑战推动了学术界对数据血缘构建方法的优化研究^[37]，进而催生了基于机器学习的新方法。

3.5 机器学习法

机器学习法是利用机器学习技术，自动化识别、追踪数据的来源、流转过程。这种方法通过训练模型来识别数据之间的关系和依赖，从而构建数据血缘。该方法的优点是可应用海量数据场景；缺点是结果可解释性较弱。国外从 2020 年开始利用机器学习技术进行数据血缘关系的研究，目前研究成果较少，准确度不够理想。该领域具有广阔的研究前景。

3.6 五种方法总结

以上讨论的几种方法各有利弊，各方法的优缺点如表 1 所示。系统跟踪法精准实时但侵入性强，依赖数据治理平台；基于 SQL 解析的方法通用性强，但对复杂语句解析和结果验证存在挑战；逆置函数法存储高效但受限于函数可逆性；标注法灵活易实现，却面临存储开销大和查询效率低的瓶颈；机器学习法可用于海量数据且潜力大，但精度不足且研究尚处早期。在实际应用中，企业通常会根据自身需求选择合适的方法，或者结合多种方法，以提高血缘分析的准确性和效率。

表 1 数据血缘关系构建方法对比

方法	过程	粒度	优点	缺点
系统跟踪法	数据加工过程，程序自动发出血缘信息	表、字段、元组	血缘关系精准，及时，细粒度	具有一定侵入性
基于 SQL 解析法	自动解析 SQL 文件、存储过程、ETL 过程	表、字段	应用最广泛的血缘分析方法	需要人工检验解析结果是否正确
逆置函数法	通过构造逆置函数对查询求逆	元组	占用少部分存储空间	并非所有函数都可逆
标注法（手工）	手动标注数据，添加数据间的血缘关系	元组	最直接的方式	消耗大量人力和存储空间
标注法（半自动）	手动添加一次标注后自动生成半环多项式	元组	人力消耗少，实现半自动化	存在存储开销大和查询效率低的问题
机器学习法（有脚本依赖）	依赖数据集之间的关系计算相似度	元组	避免标注法的缺点	需要人工手动确认其精确率
机器学习法（无脚本依赖）	通过模型预测目标数据的源数据	表、字段、元组	对业务和工具均没有依赖	以丢失一些信息为代价

4 数据血缘的应用领域

数据血缘与它的应用一直被广泛关注。在具体应用上,很多学者设计了数据血缘分析平台,如冀欣欣^[38]、叶天琦^[39]等。很多企业也开始研究数据治理,可视化展现数据血缘关系,帮助用户更好地了解数据的流转情况。

在金融行业,数据血缘被用于风险控制、反欺诈等领域。通过血缘关系,用户能快速定位数据质量问题的根源,发现潜在风险^[40]。例如,Gao等^[41]提出金融数据监管的大数据溯源模型,金旭君等^[42]探讨了数据血缘在银行 EAST 治理中的应用。

在医疗领域,数据血缘被用于临床试验溯源、隐私保护等。Woodman等^[43]将数据血缘和性能数据相结合预测未来的工作负载,以快速通知决策者。

其他领域也有不少学者进行数据血缘研究。例如,沈晓谦^[34]提出面向审计领域的数据血缘追踪技术,杨彬^[25]研究了油田数据血缘构建方法,张海涛等^[44]构建面向烟草行业的大数据分析平台,张宁^[45]将数据血缘运用到图书馆数据管理。

5 结论

随着大数据时代的到来,数据生成规模激增,数据血缘受到越来越多的关注,本文梳理了当前数据血缘关系构建方法,并对比了几种方法的优缺点。不同的数据血缘分析方法各有利弊。

当前数据血缘构建的研究仍存在若干难点:首先,实时数据血缘与动态数据流的支持仍不成熟;其次,血缘信息的可视化探索及其与数据质量管理、数据安全管控的深度集成尚未形成体系化解决方案。

未来,数据血缘构建方法将呈现以下发展趋势:一是向自动化与智能化方向演进,通过融合知识图谱、自然语言处理等技术,提升血缘发现的准确性与覆盖范围;二是朝着细粒度、实时化血缘采集方向深化,以满足故障快速定位等场景的需求;三是血缘驱动的主动式数据治理将成为关键方向,即基于血缘关系实现数据质量预警、权限管控等活动的自动响应。

参考文献

- [1] DEV D, PATGIRI R. Dr. Hadoop: an infinite scalable metadata management for Hadoop—how the baby elephant becomes immortal [J]. *Frontiers of Information Technology & Electronic Engineering*, 2016, 17 (1): 15–31.
- [2] 马方舟,陈政,吴骏. 基于大数据技术的数据库管理策略分析 [J]. *电子技术*, 2024, 53 (10): 414–415.
- [3] 李春梅,张星,耿慧拯,等. 基于数据血缘构建数据分析方法 [J]. *中国新通信*, 2020, 22 (20): 50–51.
- [4] 朱青. 基于数据血缘分析的电网企业数据资产价值评估 [J]. *信息与电脑 (理论版)*, 2023, 35 (13): 91–93.
- [5] 金泳,高扬华,潘晓华,等. 采用数据血缘的数据热度预测方法 [J]. *计算机应用*, 2023, 43 (S1): 119–125.
- [6] PORKODI S, KESAVARAJA D. Secure data provenance in Internet of Things using hybrid attribute based crypt technique [J]. *Wireless Personal Communications*, 2021, 118 (4): 2821–2842.
- [7] WANG Y R, MADNICK S E. A polygen model for heterogeneous database systems: the source tagging perspective [C]//*Proceedings of the 16th International Conference on Very Large Data Bases (VLDB)*. Calif.: Morgan Kaufmann Publishers, 1990: 519–538.
- [8] LANTER D P. Design of a lineage-based meta-data base for GIS [J]. *Cartography and Geographic Information Systems*, 1991, 18 (4): 255–261.
- [9] WOODRUFF A, STONEBRAKER M. Supporting fine-grained data lineage in a database visualization environment [C]//*Proceedings of the Thirteenth International Conference on Data Engineering*. Los Alamitos, Calif.: IEEE Computer Society, 1997: 91–102.
- [10] CUI Y W, WIENER J L, WIDOM J. Tracing the lineage of view data in a warehousing environment [J]. *ACM Transactions on Database Systems*, 2000, 25 (2): 179–227.
- [11] BUNEMAN P, KHANNA S, TAN W C. Data provenance: some basic issues [C]//*FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*. Springer Berlin Heidelberg, 2000: 87–93.
- [12] BUNEMAN P, KHANNA S, TAN W C. Why and where: a characterization of data provenance [C]//*International Conference on Database Theory (ICDT)*. Berlin, Heidelberg: Springer, 2001: 316–330.
- [13] GREENWOOD M, GOBLE C, STEVENS R, et al. Provenance of e-science experiments—experience from bioinformatics [J]. *ResearchGate*, 2003, 32 (4): 58–49.
- [14] GOBLE C. Position statement: musings on provenance, workflow and (semantic web) annotations for bioinformatics [J]. *Workshop on Data Derivation and Provenance*, 2002, 31 (4): 1–5.
- [15] SIMMHAN Y L, PLAILE B, GANNON D. A survey of data provenance in e-science [J]. *ACM Sigmod Record*, 2005, 34 (3): 31–36.
- [16] GLAVIC B, DITTRICH K R. Data provenance: a categorization of existing approaches [J]. *BTW*, 2007, 7 (12): 227–241.
- [17] MALIK T, NISTOR L, GEHANI A. Tracking and sketching distributed data provenance [C]//*2010 IEEE Sixth International Conference on e-Science*. Brisbane, QLD, Australia:

- IEEE, 2010: 190 – 197.
- [18] 戴超凡. 数据仓库中数据志追踪的理论和方法研究 [D]. 长沙: 国防科技大学, 2002.
- [19] 刘喜平. 数据起源研究综述 [J]. 科技广场, 2005 (1): 47 – 52.
- [20] 高明, 金澈清, 王晓玲, 等. 数据世系管理技术研究综述 [J]. 计算机学报, 2010, 33 (3): 373 – 389.
- [21] LI J, LI X, LV J. Selecting materialized views based on top-k query algorithm for lineage tracing [C] //2012 Third Global Congress on Intelligent Systems. IEEE, 2012: 146 – 150.
- [22] 程世豪. 面向元数据血缘关系的映射技术及实现 [D]. 成都: 西南财经大学, 2020.
- [23] 于俊锋. 面向电子邮件的近似世系关系抽取技术 [D]. 沈阳: 东北大学, 2011.
- [24] 高元照. 面向监管的大数据世系关键技术研究 [D]. 郑州: 战略支援部队信息工程大学, 2021.
- [25] 杨彬. 油田结构化数据血缘分析方法研究 [D]. 大庆: 东北石油大学, 2023.
- [26] 潘晓华, 金泳, 高扬华, 等. 面向复杂数据审计需求的数据血缘构建方法 [J]. 计算机应用研究, 2024, 41 (1): 76 – 82.
- [27] 滕召嘉. 基于血缘关系的元数据管理系统的设计与实现 [D]. 北京: 北京交通大学, 2022.
- [28] IKEDA R, WIDOM J. Panda: A system for provenance and data [J]. Bulletin of the Technical Committee on Data Engineering, 2010, 33 (3): 42 – 49.
- [29] 林悦邦. 面向关系数据库的数据起源研究与设计 [D]. 广州: 华南理工大学, 2017.
- [30] 张苒. 基于 PROV 的 ETL 溯源方法研究 [D]. 长沙: 国防科学技术大学, 2017.
- [31] GREEN T J, KARVOUNAKIS G, TANNEN V. Provenance semirings [C] //Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2007: 31 – 40.
- [32] HACHEM N I, QIU K, GENNERT M A, et al. Managing derived data in the Gaea scientific DBMS [C] //Proceedings of the 19th International Conference on Very Large Data Bases, 1993: 1 – 12.
- [33] CHITICARIU L, TAN W C, VIJAYVARGIYA G. DBNotes: a post-it system for relational databases based on provenance [C] //Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, 2005: 942 – 944.
- [34] 沈晓谦. 基于数据血缘追踪的审计报告生成技术研究 [D]. 哈尔滨: 哈尔滨工程大学.
- [35] SENELLART P, JACHET L, MANIU S, et al. Provenance and probability management in PostgreSQL [J]. Proceedings of the VLDB Endowment (PVLDB), 2018, 11 (12): 2034 – 2037.
- [36] SENELLART P. Provenance and probabilities in relational databases [J]. ACM SIGMOD Record, 2018, 46 (4): 5 – 15.
- [37] 王晓庆, 孙战伟, 吴军红, 等. 基于数据要素流通视角的数据溯源研究进展 [J]. 数据分析与知识发现, 2022, 6 (1): 43 – 54.
- [38] 冀欣欣. 基于大数据环境的数据血缘分析系统的设计与实现 [D]. 南京: 南京大学, 2019.
- [39] 叶天琦, 沈春锋. 数据血缘可视化分析平台研究与应用 [J]. 信息技术与标准化, 2020 (11): 17 – 20.
- [40] 王哲, 赵梓荣, 岳丰, 等. 面向监管报送数据治理的元数据血缘图谱研究 [J]. 中国金融电脑, 2023 (1): 58 – 62.
- [41] GAO Y, CHEN X, DU X. A big data provenance model for data security supervision based on PROV-DM model [J]. IEEE Access, 2020, 8: 38742 – 38752.
- [42] 金旭君, 沈叶红, 汪梦婷, 等. 数据血缘管理在银行 EAST 治理中的应用 [J]. 金融会计, 2024 (9): 54 – 63.
- [43] WOODMAN S, HIDDEN H, WATSON P. Applications of provenance in performance prediction and data storage optimisation [J]. Future Generation Computer Systems, 2017, 75 (8): 299 – 309.
- [44] 张海涛, 薛翔. 面向烟草行业研发应用的大数据分析平台 [J]. 计算机应用与软件, 2021, 38 (6): 18 – 22, 93.
- [45] 张宁. 赋能图书馆效能评估: 数据血缘关系驱动的指标管理与应用探析 [J]. 图书馆, 2024 (4): 53 – 61, 68.

(收稿日期: 2025 – 04 – 01)

作者简介:

吕琳 (1999 –), 女, 硕士, 主要研究方向: 大数据。

田庆业 (1976 –), 男, 硕士, 工程师, 主要研究方向: 工业控制、数据通信、信息系统。

焦冬冬 (1987 –), 男, 硕士, 高级工程师, 主要研究方向: 分布式软件、物联网、大数据。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com