● 魏银珍，邓仲华（武汉大学 信息管理学院，湖北 武汉 430072）

云环境下科学工作流的溯源数据收集和查询框架研究*

摘 要：在大数据时代，科学研究第四范式已经成为一种根本研究范式，云计算可以解决数据密集型科学研究中数据的存储、管理、注解和共享等，但仍然存在一些全新的挑战。文章提出云计算环境下科学工作流的数据溯源基本框架，详细阐述了该框架模型中溯源数据的收集、存储、查询的设计。这个溯源框架对科学工作流本身的性能无显著影响，具有最小入侵性；同时，允许用户指定从3个不同层级收集和查询溯源信息，来保证溯源的保真度，提高数据溯源的灵活性。

关键词：云计算；溯源；科学工作流；数据收集

Abstract: In the era of big data, the fourth paradigm of scientific research becomes a fundamental paradigm, in the mean while, cloud computing can solve the storage, management, annotation and sharing of data in data-intensive scientific research. However, there are still many new open challenges. This paper proposes a provenance framework for scientific workflow in cloud environment, and expounds the design of collection, storage and query for provenance data in the framework mode. The provenance framework, with minimal invasive, proposed in the paper has no significant effect on the performance of scientific workflow, and allow the users to collect and query provenance information from 3 different levels in order to ensure the fidelity of provenance and improve the flexibility of provenance.

Keywords: cloud computing; provenance; scientific workflow; data collection

1 问题的提出

科学计算加速了科学数据的产生，导致了在很多领域的信息爆炸，分析和理解这些科学数据需要借助于复杂的计算处理过程，通常要整合一些松散耦合的资源，尤其是专业图书馆、云资源和Web服务，这些过程会产生很多数据产品以及中间数据，增加了科学家处理数据的难度。同时，科学家需要付出极大的努力管理、记录溯源信息来回答一些基础性的问题，如谁在什么时候创建了该数据？什么时候又是被谁修改过？两个不同的数据产品是否可以追溯到同一个原始数据？回答这些问题不仅费时而且容易出错[1]。

因此，科学工作流广泛应用于科学研究。科学工作流不仅支持自动的任务执行，而且可以在不同级别捕获复杂的分析处理过程，系统地捕获衍生数据产品的溯源信息。溯源信息也指数据产品的审计线索、世系和谱系，包含处理过程信息和追溯数据产品的信息[2]。它是保存了确定数据的质量和作者、并重现和验证结果的关键文档。这是所有科学过程的重要要求。从而，科学工作流溯源至关重要，而且越来越重要。

科学工作流系统的主要优势之一是它们可以很容易地配置自动捕获溯源——可以通过系统的API直接访问这些溯源信息。早期的工作流系统如Taverna和Kepler扩充了溯源捕获功能，新开发的系统如VisTrails在设计之初就支持溯源。

在过去的几年中，已有多个溯源推荐模型[3-4]。所有这些模型都支持某种形式的溯源，其中有些支持溯源捕获和标注。虽然这些模型的捕获和存储方式不同，但它们都具有基本的信息类型：过程和数据产生。最近的研究表明，整合不同的溯源模型是非常必要的。

虽然已有很多种溯源和建模的方法，溯源的存储、访问、查询在近期才受到关注。溯源可以使用户发现和更好地理解工作流执行的结果，能够查询工作流的溯源信息，使知识重用[1]。

在大数据时代，数据密集型科学如今已经与理论科学、实验科学和计算科学比肩，共同成为一种根本的研究范式——科学研究第四范式。这个科学范式的目标是拥有一个所有科学文献和科学数据都在线且能够彼此交互的世界（如图1所示）[5]。科学家可以利用云计算的资源和业界已有的数据库来提高他们的发明能力，在阅读论文的同时查看相关的原始数据，重作分析；或者在查看数据时阅

* 本文为国家自然科学基金项目“大数据环境下面向科学研究第四范式的信息资源云研究”的成果，项目编号：71373191。

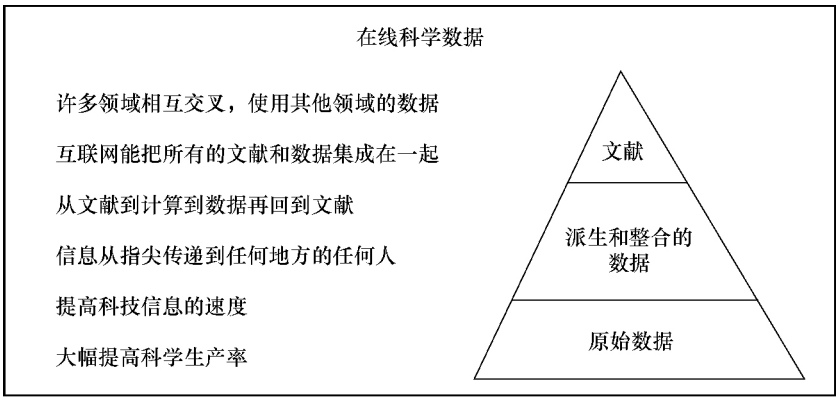

图 1　科学数据层次

读相关的文献，藉此来提高科学的“信息速率”，促进研究人员的科学生产力。

本文首先分析云环境下科学工作流执行的优势和劣势，支持云环境下溯源收集存储的挑战，其次提出云计算环境下科学工作流的数据溯源基本框架，最后详细阐述了该框架模型中溯源数据的收集、存储、查询的设计目标和实现过程。本文所提出的溯源框架对科学工作流本身的性能无显著影响，具有最小入侵性，同时，允许用户指定 3 个级别的溯源信息收集，来保证溯源的保真度，提高数据溯源的灵活性。

2　云计算应用于科学工作流的优势与劣势

由于研究方法的转变以及多种快速发展的技术相互融合，科学数据体量十分庞大，对科学数据生成并经过分析后，需要存储、管理、注解、归档和共享。研究人员发现，云计算以“现用现付”的方式访问云数据中心，其资源能够高效地满足科学数据分析的某些新兴计算需求。

云计算通常有 3 种模式：一是“基础架构即服务”模式，程序员对虚拟机操作系统的配置有全部的访问权限；二是“平台即服务”模式，提供更高级的编程模型和数据库服务；三是“软件即服务”模式，用户能够访问全部软件服务。这 3 种模式都能很好地应用于数据密集型科学。

对科学工作流和云计算进行深入分析后，笔者总结出云环境下执行工作流的几个优势：①不仅利用了云计算服务的资源优势，而且也克服了地域限制，这对科学合作至关重要。②云计算以数据为中心的资源调度和科学工作流的数据驱动模式相匹配。③云计算的动态资源调度与科学工作流的动态执行相匹配。④云计算平台可以以低廉的费用、较高的性能为科学工作流提供所需的高性能计算资源和海量存储资源。

然而，在云计算环境下，一项科学实验的数据和程序将分布在成百上千台的计算节点，因此给科学工作流中数据的管理带来了全新的挑战，尤其是在数据溯源方面。首先，科学工作流的数据密集型特征需要进行大量的数据传输，单云计算中的各个数据中心地理上可能不在一个地方。数据中心之间的网络带宽有限，所以，数据传输就是一大挑战。其次，科学工作流中任务执行可能会产生被后续任务使用的中间数据，这些中间数据会占据一定的存储空间，但是这些中间数据也可以通过重新计算得到。这就要求捕获数据溯源来提供数据的上下文，进行重复计算，数据溯源就是数据世系，是从历史的角度审视数据产品[6]。

3　溯源框架

云计算环境中，通常有成千上万个节点协同工作完成一个大型模拟实验，通过用高性能计算系统完成分析和处理大量数据的工作。跟踪云环境下数据的运行，追踪实验数据世系成为难点 。目前，在科学工作流中，科学界用户通常通过在文件名中嵌入时间戳和一些简单描述，或者通过手工方法，间接维护溯源信息。首先，这种做法不具有普适性且容易出错，所以不宜扩展到大量的数据系统中。其次，工作流级的溯源集合在灵活性方面不能满足用户需求，也使其难以捕捉资源特性。最后，这些溯源集合对多数依靠脚本来管理计算和数据用户是不能访问的。

本文提出一个溯源收集和存储系统框架，并描述了实现该框架的基本思路。溯源信息的收集、存储和访问受限于系统的可用存储，当扩展到具有海量数据的大型系统时，支持半结构化溯源信息收集和更复杂的溯源分析变得十分必要。本研究构建了分层体系结构模型，将收集存储和分析存储分离，该模型灵活支持多种溯源数据模型，具有良好的可扩展性、半透明性。

3.1　溯源系统分层体系结构

图 2 表示出了本文所设想的面向科学应用溯源采集和查询框架。这是一个分层的溯源收集和分析架构，其中溯源收集的存储与查询和分析的存储是分离的。第一层负责收集溯源数据。大数据环境下，溯源数据是结构化和半结构化的格式；第二层主要解决溯源数据存储问题，为不同类型的溯源数据选择合适的存储方式；第三层负责溯源数据分析，为用户提供查询功能。

分层结构中的第一层是该体系结构中负责溯源收集的部件，以具有最小入侵性的机制从云计算系统中收集溯源

数据，其所收集的信息是数据溯源产品、资源以及环境的混合信息，用户可以选择在不同级别收集溯源数据（将在3.3节中详细阐述）。

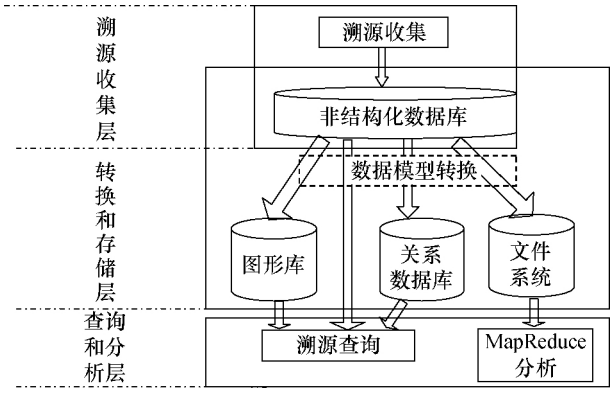

图2　溯源收集和分析框架

溯源数据量的大小会因科学应用和用户所使用的实验设备的不同而有差别，由于系统中所捕获到的是形式和大小各异的溯源信息，因此，本研究建议不以良好的结构存储这些溯源信息，而是优先考虑溯源数据能够快速有效地写入存储。该框架可以收集系统所有的原始溯源数据，并将其存储在溯源缓冲区中，提供快速有效的查询，也可以作进一步处理，转化成不同格式的结构化的溯源信息。与结构相关联的数据留给下游的溯源系统（如KARMA[4]）或其他数据分析工具（如图形分析或MapReduce）。溯源收集存储和查询存储的分离，使得在这两个层上的优化都成为可能。

3.2　应用情景

本文考虑了两种情况，第一种情况是用户提交作业。跟踪提交给高性能计算系统的批处理队列的任务脚本中的数据处理溯源信息，包括任务提交、执行和完成，这样，管理员和普通用户均可跟踪执行每一步。第二种情况是在命令提示下，用户通常完成一个数字金字塔的任务。该任务可能包括运行时的数据准备和将数据移动至文档系统。

3.3　设计目标

本文所设计的溯源框架拟达到以下目标：①支持不同溯源数据模型。溯源系统在很大程度上依赖于后台的结构化的关系数据库，溯源收集系统在执行周期的已知点（如工作流启动时）捕获溯源记录集合，然而，高性能系统中的作业脚本有很大差异，因此，开发一个基于强架构的数据模型是非常困难的。所以，针对不同的系统、不同的用户和不同的应用支持半结构化的溯源数据模型尤为重要。②低开销。海量的数据处理中，高性能是关键。因此降低溯源收集对应用性能的影响也很重要。③半透明化。溯源收集需要对用户的工作流具有最小侵入。使用户透明地初始化溯源收集工作。在工作流中，用户的脚本自动化是我们拟解决的问题。④支持用户标注。除了自动化装置，捕获用户笔记和元数据也是必需的，因此，需要一个用户界面，允许用户在实验前、实验中，新增有关实验和数据的科学记录。⑤分阶段的溯源层级。云计算环境的应用往往在需求和使用方面有很大不同，例如，例行模拟运行只需要基础级的溯源来表明任务开始和结束时间或者其他数据的基本特征。然而，有些应用或使用情景在执行过程中可能需要更细粒度的数据溯源的收集。⑥可扩展性。高性能计算系统中急剧增长的模拟数据量使得有必要扩展溯源收集和存储机制。这在即将到来的百亿亿次时代尤其如此。

4　溯源框架设计

图3显示的是溯源框架中的存储和查询部件设计。溯源收集机制的启动将触发用户会话溯源，同时捕获作业脚本执行时的溯源数据，并将其保存于NoSQL数据存储中，继而可以通过命令行或者Web查询界面访问这些溯源数据。用户将作业提交给批处理队列系统，等待在资源可用时执行。框架中的溯源模块首先捕获原始作业脚本，并将其存储于NoSQL数据库中；其次，在作业脚本中添加溯源信息，并将带有溯源信息的工作脚本提交给批处理队列。在作业执行期间，加载了的溯源信息的脚本将溯源数据存入溯源数据存储器。这样，溯源数据就能够通过命令行界面或者Web界面被用户访问。

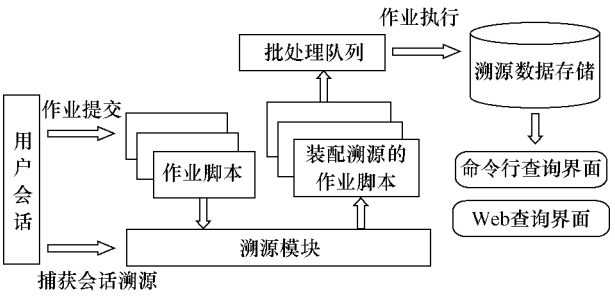

图3　溯源框架体系结构

用户通过高性能计算系统内置的溯源能力，依靠捕获的溯源信息跟踪作业和数据。同时，科学家将溯源系统作为“笔记本”来记录科学应用。用户可以执行检索，识别实验的异同。另外，高性能计算系统的管理员也可以通过使用溯源数据调整溯源收集来帮助科学应用的调试。

4.1　溯源收集

本文所设计的溯源收集部件负责脚本执行和用户会话的溯源收集，在上述框架中，用户会话溯源被完全捕获。对溯源收集，本文将分为3个层次：第一层由基本溯源信

息构成，包括脚本信息，任务运行的输出数据，任务提交和用户标注相关的基本环境信息，这是一个科学用户感兴趣的最基本的溯源信息；第二层包含来自第一层的所有信息和参与计算的资源信息，本层的溯源信息对系统管理员更有价值，同时也有助于对特殊类型问题的资源使用分析；第三层捕获第二层的所有溯源信息任务脚本中的命令所使用的输入输出详细追踪信息。

框架中的收集分层的设计基于目前用户（科学家和系统管理员）需求。然而，分级实现将随资源类型和开发方式的不同而存在差异，在实现时，可通过提高系统的可扩展性来解决此问题。

4.2 存储

系统中捕获的溯源信息数量庞大，类型多样，因此，数据存储必须能够支持半结构化的溯源文档。在实现时，可以使用 NoSQL 数据库，如 MongoDB 即可满足该要求。MongoDB 具有可扩展性、开源的、性能优等特点，还具有支持半结构化数据和分片特征[7]，能够比较理想地解决半结构化的问题。另外，利用 MapReduce、MongoDB 可以执行复杂聚合分析。既可以通过本地 MapReduce 框架，也可以通过 MongoDB-Hadoop 连接器支持多样化的溯源分析工具。

捕获溯源按作业编号 ID（又批处理系统指定的唯一标识符）或文件标识符（基于本地的）分组。然而，数据存储的每个入口无须固定其输入格式，不同的文件和任务之间的差别很大，这使得我们能够在多个用户会话和多个系统之间使用同一个存储。

4.3 查询

工作流产品数据的溯源信息可以通过支持简单查询语言（仅关注任务编号或文件名）的查询界面获得，设计查询语言使其能够抽象数据存储层的底层语言，允许用户以直观的方式查询溯源信息而无须懂得太多的数据存储细节。无论是命令行还是图形化的 Web 界面都为不同的用户设计。命令行界面主要针对熟悉命令行的用户设计，而 Web 界面主要为那些类似于 Google 的溯源收集，支持开放的溯源集合的查询。

5 相关工作

近年来，已有很多面向科学工作流的多重溯源收集系统被开发，如 VisTrails[8]，Kepler[9]，及 Pegasus[10]。最近出现的 RAMP[11] 主要针对 MapReduce 工作流的溯源收集问题。Karma[12] 是一个在半结构化的 eScience 环境下支持溯源捕获的工作流系统，Provenance Aware Storage System（PASS）[13] 主要是操作系统和文件系统级的溯源捕获系统。

Gadelha Jr 等[14] 描述了溯源收集和查询是如何帮助管理在不同环境中大规模科学计算的；Jones 等[15] 详细阐述了一种能够帮助科学家更好地管理他们的数据的文件级别的溯源系统。

本文从应用的不同级别捕获和表示溯源信息，包括应用在高性能计算环境下所执行的资源和环境。同时，提出的溯源框架能够捕获溯源轨迹，用户可以决定其想要的溯源的保真级别。本文的创新之处在于溯源信息的捕获具有轻量级和最小侵入性，同时具有用户的保真度。此外，本文的设计亦允许特定使用情况下简单化。

6 结束语

本文介绍了云环境下科学工作流的溯源数据收集和查询框架的设计和实现建议，主要创新之处在于该框架模型具有最小入侵、轻量级、多级别的溯源收集，将溯源数据的存储和查询分离，支持来自高性能系统中任务和用户命令的半结构化溯源数据。下一步的工作中，笔者将搭建实证研究环境，评估该框架模型的性能并进行优化。□

参考文献

[1] DAVIDSON S B, FREIRE J. Provenance and scientific workflows: challenges and opportunities [C] //Proceedings of the 2008 ACM SIGMOD International Conference on Management of data. ACM, 2008: 1345-1350.

[2] DEELMAN E, GIL Y, ZEMANKOVA M. Report on the 2006 NSF workshop on the challenges of scientific workflows [EB/OL]. [2006-05-02]. http://www.isi.edu/nsf-workflows06.

[3] MILES S, GROTH P, MUNROE S, et al. Extracting causal graphs from an open provenance data model [J]. Concurrency and Computation: Practice and Experience, 2008, 20 (5): 577-586.

[4] SIMMHAN Y L, PLALE B, GANNON D, et al. Performance evaluation of the karma provenance framework for scientific workflows [M] //MOREAU L, et al. Provenance and annotation of data. Springer Berlin Heidelberg, 2006: 222-236.

[5] HEY T, TANSLEY S, TOLLE K. The fourth paradigm [M]. Microsoft Press, 2009: 17-18.

[6] CHEAH Y W, CANON R, PLALE B, et al. Milieu: lightweight and configurable big data provenance for science [C] //Big Data (BigData Congress), 2013 IEEE International Congress on. IEEE, 2013: 46-53.

[7] CATTELL R. Scalable SQL and NoSQL data stores [J]. ACM SIGMOD Record, 2011, 39 (4): 12-27.

（下转第 114 页）

[14] 宋新平，梅强，田红云，等．复杂系统理论视角下的中小企业竞争情报系统建设研究［J］．情报杂志，2010（3）：83-88.

[15] MANYIKA J，CHUI M，BROWN B，et al. Big data：the next frontier for innovation，competition，and productivity［R］. McKinsey Global Institute，2011.

[16] GANTZ J，REINSEL D. IDC：the digital universe in 2020：big data，bigger digital shadows，and biggest growth in the Far East［R］. IDC iView：IDC Analyze the Future，2012.

[17] MCAFEE A，BRYNJOLFSSON E. Big data：the management revolution［J］. Harvard Business Review，2012，90（10）：60-68.

[18] 孟小峰，慈祥．大数据管理：概念、技术与挑战［J］．计算机研究与发展，2013（1）：146-169.

[19] RUSSOM P. Big data：analytics［J］. Armonk：IBM & TD-WI，2012.

[20] FAN Wei，BIFET A. Mining big data：current status，and forecast to the future［J］. SIGKDD Explorations，2012，14（2）：1-5.

[21] 吴华珠，赵斐．江苏省中小企业竞争情报需求与供给现状研究［J］．情报杂志，2012（4）：80-84.

[22] 王磊．我国中小型企业竞争情报系统的构建框架研究［D］．哈尔滨：黑龙江大学，2006.

[23] 王文清，陈凌．CALIS 数字图书馆云服务平台模型［J］．大学图书馆学报，2009（4）：13-17.

[24] 刘冰．动态环境中企业竞争情报发展趋势［J］．图书情报知识，2007（6）：21-24.

[25] 李萱格，司有和．企业竞争情报系统（CIS）模型新探［J］．图书与情报，2008（6）：75-78，102.

[26] 官思发，李进华，刘齐平．基于 Web 2.0 的企业竞争情报系统模型构建［J］．情报科学，2014（1）：47-53.

[27] HERRING J P. Key intelligence topics：a process to identify and define intelligence needs［J］. Competitive Intelligence Review，1999，10（2）：4-14.

[28] 谢新洲．企业信息化与竞争情报［M］．北京：北京大学出版社，2006：87-90.

[29] 司有和．竞争情报理论与方法［M］．北京：清华大学出版社，2009：266.

[30] 宋海沂．复合企业竞争情报系统模型研究［J］．情报科学，2011（12）：1785-1790.

[31] 迈克尔·波特．竞争优势［M］．陈小悦，译．北京：华夏出版社，2003：44-51.

[32] 查先进．情报学研究进展［M］．武汉：武汉大学出版社，2007：190-195.

[33] 周海炜，王洪亮，郝云剑．云计算环境下中小企业竞争情报安全模型构建［J］．图书馆理论与实践，2013（11）：38-41.

[34] 王洪亮，郝云剑．云时代我国中小企业竞争情报安全子系统构建［J］．江苏商论，2013（3）：39-41.

作者简介：王洪亮，男，1989 年生，硕士生。研究方向：企业竞争情报。

张琪，女，1993 年生。研究方向：信息管理与信息系统。

朱延涛，男，1993 年生。研究方向：水利水电工程。

收稿日期：2014－12－10

（上接第 118 页）

[8] CALLAHAN S P，FREIRE J，SANTOS E，et al. VisTrails：visualization meets data management［C］//Proceedings of the 2006 ACM SIGMOD International Conference on Management of data. ACM，2006：745-747.

[9] LUDÄSCHER B，ALTINTAS I，BERKLEY C，et al. Scientific workflow management and the Kepler system［J］. Concurrency and Computation：Practice and Experience，2006，18（10）：1039-1065.

[10] KIM J，DEELMAN E，GIL Y，et al. Provenance trails in the wings/pegasus system［J］. Concurrency and Computation：Practice and Experience，2008，20（5）：587-597.

[11] PARK H，IKEDA R，WIDOM J. Ramp：a system for capturing and tracing provenance in mapreduce workflows［C］//Proceedings of the VLDB Endowment，2011.

[12] SIMMHAN Y L，PLALE B，GANNON D. Karma2：provenance management for data-driven workflows［J］. International Journal of Web Services Research（IJWSR），2008，5（2）：1-22.

[13] MUNISWAMY-REDDY K K，HOLLAND D A，BRAUN U，et al. Provenance-aware storage systems［C］. USENIX Annual Technical Conference，General Track，2006：43-56.

[14] GADELHA JR L M R，WILDE M，MATTOSO M，et al. Exploring provenance in high performance scientific computing［C］//Proceedings of the First Annual Workshop on High Performance Computing Meets Databases（HPCDB 2011）. 2011：17-20.

[15] JONES S N，STRONG C R，PARKER-WOOD A，et al. Easing the burdens of HPC file management［C］//Proceedings of the Sixth Workshop on Parallel Data Storage. ACM，2011：25-30.

作者介绍：魏银珍，女，1976 年生，博士生，高级工程师。研究方向：知识组织与知识服务。通讯作者。

邓仲华，男，1957 年生，教授，博士生导师。研究方向：知识组织，知识构建与处理。

收稿日期：2014－12－11