2011 IEEE International Symposium on Policies for Distributed Systems and Networks

Policy-Based Integration of Provenance Metadata

Ashish Gehani Dawood Tariq Basim Baig
SRI International, Menlo Park, CA 94025, USA
Email: {ashish.gehani,dawood.tariq,basim.baig}@sri.com

Tanu Malik
University of Chicago, Chicago, IL 60637, USA
Email: tanum@ci.uchicago.edu

***Abstract*—Reproducibility has been a cornerstone of the scientific method for hundreds of years. The range of sources from which data now originates, the diversity of the individual manipulations performed, and the complexity of the orchestrations of these operations all limit the reproducibility that a scientist can ensure solely by manually recording their actions. We use an architecture where aggregation, fusion, and composition policies define how provenance records can be automatically merged to facilitate the analysis and reproducibility of experiments. We show that the overhead of collecting and storing provenance metadata can vary dramatically depending on the policy used to integrate it.**

***Keywords*-lineage; aggregation; fusion; flexible**

I. Introduction

Reproducibility has been a cornerstone of the scientific method for hundreds of years, dating back at least to Galileo [11]. Experimental scientists have traditionally maintained laboratory notebooks, recording the steps they followed in measuring phenomena and obtaining data. Theoretical scientists have provided formal proofs showing how their results can be derived. The American Physical Society observes that the "success and credibility of science are anchored in the willingness of scientists to expose their ideas and results to independent testing and replication by others. This requires the open exchange of data, procedures and materials." [3]

Access to data has always played a critical role for verifying scientific theories. Without the astronomical observations of the Danish nobleman Tycho Brahe, the German mathematician Johannes Kepler would not have been able to address the critiques of his early efforts to characterize planetary motion [12]. The X-ray diffraction images of British chemist Rosalind Franklin were instrumental for the American biologist James Watson and British physicist Francis Crick when they developed their model of the structure of DNA [30]. History abounds with similar examples. However, the advent of the Web has revolutionized this aspect of science over the past two decades. Today, a biologist can freely download data from the Protein Data Bank, which contains structural information that cost an average of $70,000 to produce for each protein or nucleic acid [6]. Physicists around the world have access to the data from experiments at the Large Hadron Collider, which is estimated to have cost €5-10 billion [21] to build. Scientific data in every active field of research is increasingly available to anyone with an Internet connection.

At the same time that data has become ubiquitously accessible, the cost of computing has dropped dramatically, with a 12-core 1.7 GHz chip costing less than $800 [18] and TeraGrid [27] providing over 2 petaflops today. The theoretical and experimental branches of science are now complemented by a branch based on computational exploration of data [8], [29]. In this paradigm the range of sources from which the data originates, the diversity of the individual manipulations performed, and the complexity of the orchestrations that compose these operations all limit the reproducibility that a scientist can ensure solely by manually recording their actions [9]. Systems to track scientific workflows began to develop in response – for example, CMCS helps chemists document combustion research [19], $^{my}Grid$ [31] with Taverna [1] aids biologists, and ESSW is used by earth scientists [10].

Since most infrastructure being developed to record the provenance of scientific data targeted specific fields, the projects could not easily be repurposed for different domains. The systems differed with respect to what data was captured, the types of operations performed, how the data was stored, and the kinds of queries supported. Over the past five years, a community of two dozen research groups interested in data annotation, derivation, and provenance has met "to understand the capabilities of different provenance systems and the expressiveness of their provenance representations," and then iteratively created an Open Provenance Model (OPM) aimed at increasing the interoperability of systems [16].

"The Open Provenance Model aims to capture the causal dependencies between the artifacts, processes, and agents" as "a directed acyclic graph, enriched with annotations capturing further information pertaining to execution." It does not "specify the internal representations that systems have to adopt to store and manipulate provenance internally", nor does it "specify protocols to store such provenance information in provenance repositories" or "protocols to query provenance repositories" [16]. Indeed, a recent effort to use MITRE's PLUS system to import, query, and visualize provenance exported in OPM format from Harvard's Provenance-Aware Storage System [22] demonstrated that OPM needed to be augmented to facilitate query interoperability [5].

978-0-7695-4330-7/11 $26.00 © 2011 IEEE
DOI 10.1109/POLICY.2011.12

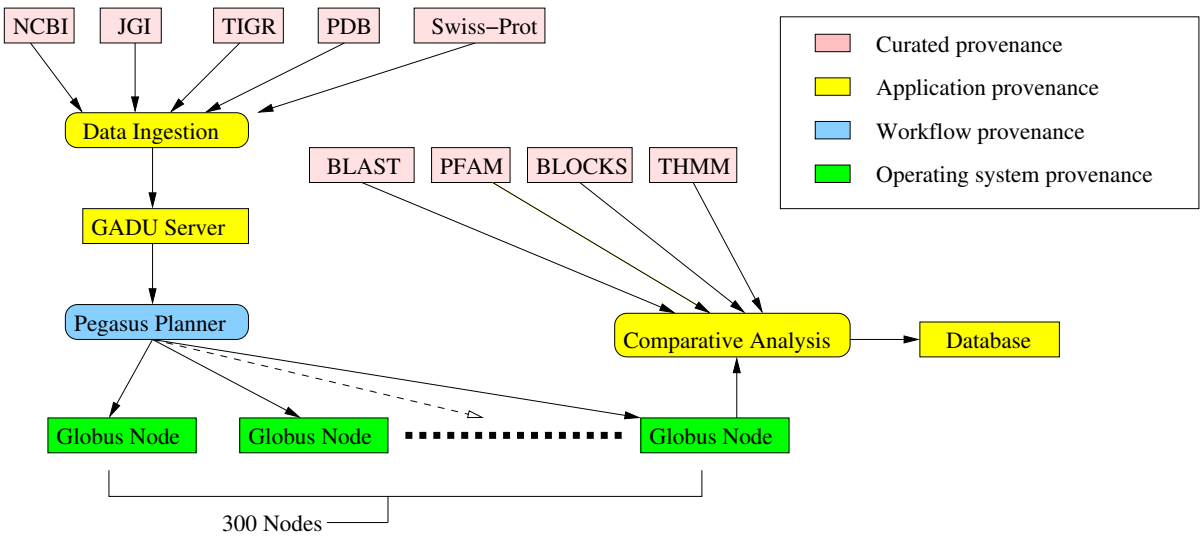

Figure 1. The provenance of a GADU record requires the integration of provenance from manual curation in data banks, the GADU components, the workflow system, and the operating system on numerous machines.

II. Challenge

As scientists begin to get access to data sets that are accompanied by provenance records, they are faced with the challenge of integrating and analyzing this metadata. Independent sources are likely to have captured provenance at distinct levels of abstraction, have different levels of completeness, used separate sets of identifiers to refer to the same artifacts, processes, and agents, and introduced dissimilar semantics in the annotations. The issue is illustrated by considering a representative example (depicted in Figure 1) – the provenance of records in the Genome Analysis and Database Update (GADU) system, which is designed to automate the assignment of functions to genes [24].

GADU works by periodically querying the National Center for Biotechnology Information (NCBI) [17], Joint Genome Institute (JGI) [14], The Institute for Genomic Research (TIGR) [28], Protein Data Bank (PDB) [20], and Swiss-Prot [26] data banks. If any new data is found, it is downloaded to the GADU server. The Pegasus planner [7] dispatches sequence data to hundreds of remote nodes. At each node, reference data is drawn from BLAST [2], PFAM [4], BLOCKS [13], and THMM [15] data banks for different types of comparative analyses. The resulting output for each then goes into a database.

As each of the constituent systems starts maintaining provenance records, the output of a genome analysis workflow will have associated metadata that includes curated provenance from NCBI, JGI, TIGR, PDB, Swiss-Prot, BLAST, PFAM, BLOCKS, and THMM, application-level provenance from the GADU software infrastructure, workflow provenance from Pegasus, and operating system-level provenance collected from the Grid nodes where parts of the analysis were executed. A scientist who wants to study the notes associated with the sources of a specific genome analysis, determine which database entries are dependent on particular biological data, or conduct a broad study of the relationship between certain molecules and properties known about them would need to assemble the pieces from the provenance records, reconciling variations in the syntax and semantics, and then construct suitable queries.

III. Policy-Based Integration

Combining provenance metadata that has a range of sampling granularities, abstraction levels, and attribute schema creates new problems. In particular, the resulting information can require large amounts of storage, degrade analytic performance, and substantially complicate query construction. We have developed an architecture for integrating and analyzing provenance metadata that arises from diverse sources. It provides sufficient flexibility to handle the needs of a wide range of applications. Rather than imposing arbitrary choices about how the information is combined, the system allows aggregation, fusion, and composition policies to define tradeoffs that are appropriate for the target domains.

The current version of SPADE [25] is the second generation of our data provenance collection and management software infrastructure. It includes a provenance kernel that exposes a non-blocking interface to the modules that report provenance. This minimizes the possibility of events being dropped while waiting for the kernel to return control. Internally, the kernel maintains a buffer for each producer from which it ingests events, utilizing the aggregation, fusion, and composition filters to reconcile the provenance elements where possible. By specifying policies in these filters, data provenance can be integrated semi-automatically.

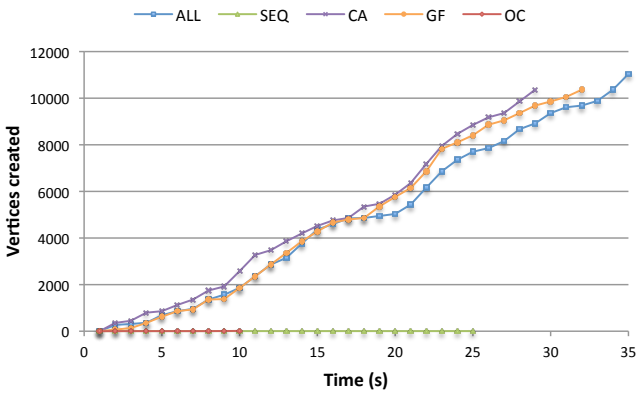

Figure 2. The number of Open Provenance Model *Artifact* vertices generated by different *aggregation policies* over time.

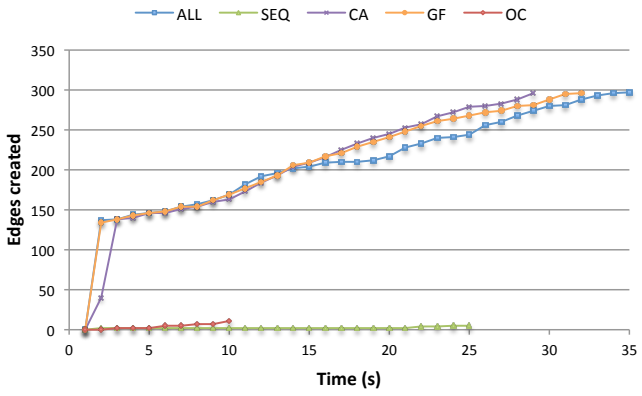

Figure 3. The number of Open Provenance Model *used* edges generated by different *aggregation policies* over time.

IV. Case Study

An important application of a provenance record is the ability it provides for understanding the relationship between data objects. However, the fidelity with which such analysis can be performed depends on the granularity at which changes are tracked by the storage system. At one extreme, a new version of a file can be created each time it is written to, an algorithm denoted by *ALL*. While this provides sufficient detail for any analysis, it results in significant storage overhead. At the other extreme, new file versions can be created only when a file is closed after it has been modified. This approach is labeled *OC* since it treats as equivalent all versions between an *open()* and a *close()* call. It incurs the least storage overhead but is particularly prone to creating cycles in the provenance graph, preventing inferences about the direction of data flow in the system.

Harvard's PASS [22] group implemented two algorithms that lie between the above extremes [23]. *Cycle Avoidance (CA)* tracks the ancestors of a file and creates a new version each time a new ancestor is encountered. *Graph Finesse (GF)* tracks the entire lineage graph of a file and

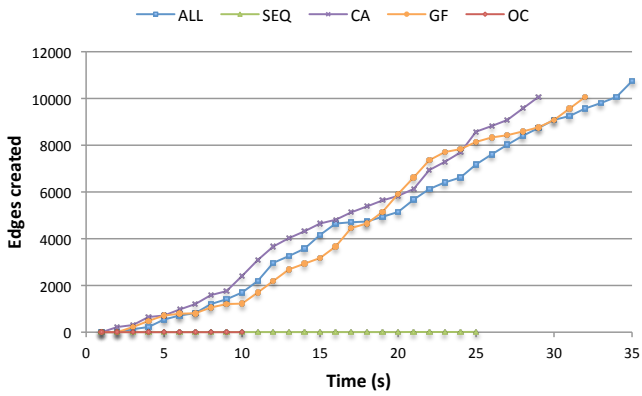

Figure 4. The number of Open Provenance Model *wasGeneratedBy* edges generated by different *aggregation policies* over time.

Policy	Lines of code
ALL	38
SEQ	73
CA	63
GF	80
OC	6

Table I
Size of different *aggregation policy* filters.

creates a new version if a new edge would have created a cycle. Finally, *SEQ* refers to the algorithm used in the first generation of SPADE, where a new version is created after a sequence of consecutive writes is succeeded by a read or a write by a different process.

We implemented policies for each of the five algorithms as separate SPADE *aggregation policy* filters. The number of lines of Java code it took to create each is shown in Table I. We ran a BLAST [2] workload with each of the five policies. The numbers of Open Provenance Model *Artifact* vertices, *used* edges, and *wasGeneratedBy* edges emitted after aggregation are plotted in Figures 2, 3, and 4, respectively. The fidelity of the provenance record differs substantially depending on the aggregation policy used.

The vertex and edge counts are plotted as a function of the amount of time that has elapsed. Each figure has plots for the five different policies. Though the same workload is utilized for each of the policies, the length of time needed to complete the execution varies substantially. The run with the *OC* policy finished within a few seconds since it generates few Open Provenance Model elements. In contrast, running the workload with the *GF* policy takes close to half a minute since it must incrementally compute the transitive closure of the provenance graphs of the files being modified. This illustrates one dimension of the tradeoff from using different aggregation policies. Our experience implementing the aggregation filters shows that it is possible to change the provenance integration behavior of SPADE with very little effort needed to develop the policies.

V. Conclusion

As datasets begin to be accompanied by provenance records, scientists are faced with the challenge of integrating the metadata. We described how SPADE uses policies to integrate the metadata with minimal development effort. Depending on the policy used, the overhead of storing, querying, and integrating new provenance metadata can vary dramatically. It is therefore important to be able to customize the integration policy to the application for which the provenance records will be utilized. We illustrated how different policies could be utilized through a case study of five different aggregation policies for operating system activity. Implementing each policy required half to seven dozen lines of Java code, several orders of magnitude less development than would be required to construct a system customized to the specific provenance sources.

VI. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant OCI-0722068. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Nedim Alpdemir, Arijit Mukherjee, Norman W. Paton, Alvaro A. A. Fernandes, Paul Watson, Kevin Glover, Chris Greenhalgh, Tom Oinn, and Hannah Tipney, Contextualised workflow execution in myGrid, *European Grid Conference, Springer-Verlag Lecture Notes in Computer Science*, Vol. 3470, 2005.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Research*, Vol. 25, 1997.

[3] American Physical Society, http://www.aps.org/policy/statements/99_6.cfm

[4] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer, The Pfam protein families database, *Nucleic Acids Research*, Vol. 30, 2002.

[5] U. Braun, M. Seltzer, A. Chapman, B. Blaustein, M. D. Allen, and L. Seligman, Towards query interoperability: PASSing PLUS, *2nd Workshop on the Theory and Practice of Provenance*, 2010.

[6] S.K. Burley, A. Joachimiak, G.T. Montelione, and I.A. Wilson, Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers, *Structure*, Vol. 16(1), 2008.

[7] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, K. Blackburn, A. Lazzarini, A. Arbree, R. Cavanaugh, and S. Koranda, Mapping abstract complex workflows onto Grid environments, *Journal of Grid Computing*, Vol. 1(1), 2003.

[8] Sergey Fomel and Jon F. Claerbout, Reproducible research, *Computing in Science and Engineering*, Vol. 1(1), 2009.

[9] J. Freire, D. Koop, E. Santos, and C.T. Silva, Provenance for computational tasks: A survey, *Computing in Science and Engineering*, Vol. 10(3), 2008.

[10] J. Frew and R. Bose, Earth System Science Workbench: A data management infrastructure for earth science products, *Scientific and Statistical Database Management Conference*, 2001.

[11] Galileo Galilei, *Discourses and Mathematical Demonstrations Relating to Two New Sciences*, 1638.

[12] S. Hawking (Editor), *On the Shoulders of Giants*, Running Press, 2004.

[13] S. Henikoff, J. G. Henikoff, and S. Pietrokovski, Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations, *Bioinformatics*, Vol. 15, 1999.

[14] Joint Genome Institute, http://www.jgi.doe.gov/

[15] A. Krogh, Prediction of transmembrane helices in proteins, http://www.cbs.dtu.dk/services/TMHMM/

[16] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche, The Open Provenance Model core specification (v1.1), *Future Generation Computer Systems*, 2010.

[17] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/

[18] AMD Opteron 6100, http://www.amd.com/us/press-releases/Pages/amd-sets-the-new-standard-29mar2010.aspx

[19] C. Pancerella, J. Hewson, W. Koegler, D. Leahy, M. Lee, L. Rahn, C. Yang, J. D. Myers, B. Didier, R. McCoy, K. Schuchardt, E. Stephan, T. Windus, K. Amin, S. Bittner, C. Lansing, M. Minkoff, S. Nijsure, G. v. Laszewski, R. Pinzon, B. Ruscic, Al Wagner, B. Wang, W. Pitz, Y. L. Ho, D. Montoya, L. Xu, T. C. Allison, W. H. Green, Jr., and M. Frenklach, Metadata in the collaboratory for multi-scale chemical science, *Dublin Core Conference*, 2003.

[20] Protein Data Bank, http://www.rcsb.org/pdb/

[21] M. L. Perl, The Tau Lepton and thirty years of changes in elementary particle physics research, *Physical Review Letters*, American Physical Society, Vol. 100(7), 2008.

[22] K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer, Provenance-aware storage systems, *USENIX Annual Technical Conference*, 2006.

[23] K. Muniswamy-Reddy and D. A. Holland, Causality-based versioning, ACM Transactions on Storage, Vol. 5(4), 2009.

[24] A. Rodriguez, D. Sulakhe, E. Marland, V. Nefedova, M. Wilde, and N. Maltsev, Grid enabled server for high-throughput analysis of genomes, *Workshop on Case Studies on Grid Applications*, 2004.

[25] Support for Provenance Auditing in Distributed Environments, http://spade.csl.sri.com/

[26] Swiss-Prot Protein Knowledgebase, http://us.expasy.org/sprot/

[27] TeraGrid, http://teragrid.org/

[28] The Institute for Genomic Research, http://www.tigr.org/

[29] D. van Dyk, Data-Driven Science: The view of a statistician, *Workshop on Interdisciplinary Strategic Issues in e-Science and Cyber-Infrastructure*, 2007.

[30] J. D. Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*, Atheneum, 1968.

[31] J. Zhao, C. A. Goble, R. Stevens, and S. Bechhofer, Semantically linking and browsing provenance logs for E-science, *1st IFIP International Conference on Semantics of a Networked World*, 2004.