DATA PROVENANCE AND FINANCIAL SYSTEMIC RISK

(Case Study)

Len Seligman, Shaun Brady, Barbara Blaustein, Paula Mutchler,
Adriane Chapman, Charles Worrell
The MITRE Corporation, USA
{seligman, sbrady, bblaustein, paulam, achapman, cworrell}@mitre.org

Abstract: We describe the needs for data provenance in a large-scale analytic environment to support financial systemic risk analysis. Government financial regulators need to make sense of the outputs of thousands to tens of thousands of simulation runs invoked by a large analytic staff; automatic capture of data provenance (dataset sources and processing steps) supports analysts without adding to their workloads. We present an architecture for automated provenance capture from both simulations and data transformation tools. Finally, we describe a prototype implementation and next steps.

Key Words: Data Provenance, Information Product Theory and Practice

1. INTRODUCTION

A common mantra of CIOs is that information should be managed as a strategic, enterprise asset. This requires active management of data quality, so that users understand the quality of the information on which they base their decisions. A theory of information products has emerged, based on parallels between the production of physical and information products [2, 13, 15]. Instead of processing physical raw materials, information systems transform raw data inputs into information products.

This is rarely a one-step process. The information products produced by one process are often refined further by other processes, in turn producing additional value-added information products. In such an environment, users need *data provenance*—informally, a data "family tree"—to help them interpret the information properly and determine how much they should trust it. Data provenance is *"information that helps determine the derivation history of a data product...[It includes] the ancestral data product(s) from which this data product evolved, and the process of transformation of these ancestral data product(s).*" [9]

This case study describes the need for data provenance among government regulators of the financial system. Our work is motivated by consideration of the needs of regulators like the Office of Financial Research (OFR), an organization within the U.S. Department of the Treasury with the mission of collecting, integrating, and analyzing diverse data in order to better track and analyze financial systemic risk. We then present an architecture for a Financial Modeling and Analysis Environment which addresses the provenance capture challenge by integrating a provenance manager, an open source data transformation tool, and a novel simulation execution environment. We briefly describe a prototype implementation and our future directions.

2. THE NEED FOR DATA PROVENANCE

Systemic risk analysis requires collection of data from hundreds (and ultimately, thousands) of financial firms as well as data on interest rates, equity and commodity indices, labor productivity, employment rates, and inflation. This data must be integrated (e.g., by matching corresponding entities), and transformed to meet the needs of diverse systemic risk analysis models. Some of the mismatches that need reconciliation include differences in semantics (e.g., profit being calculated before vs. after taxes),

time frame (e.g., a model that assesses risk for a three month period that is four quarters in the future vs. one that does so for a 12 month period one fiscal year in the future), data resolution (e.g., annual, quarterly, monthly, daily, vs. tick-level data), and statistical precision [12].

Once transformed and integrated, this highly heterogeneous data feeds diverse information production processes, consisting mostly of computer-based simulation models. The OFR has developed an analytical framework around six basic functions of the financial system – credit allocation and leverage, maturity transformation, risk transfer, price discovery, liquidity provision, and facilitation of payments—and an assessment of how threats may disrupt their functioning [4].

Currently, care and feeding of these simulation models is mostly ad hoc, with analysts (or their programmers) creating custom scripts and running their own data transformation and movement tools. With hundreds of analysts creating tens to hundreds of thousands of simulation runs on thousands of diverse and sometimes overlapping data sets, it becomes extremely difficult for analysts to:

- Understand the data assumptions behind different simulation runs, and
- Locate model runs and simulation results by a variety of criteria.

Analysts need tools to help them answer questions like the following:

- I ran a flow of funds model from MIT back in May. Which version did I use? What transformations did I perform on the input data sets?
- Which model runs used the 1Q 2011 version of the FDIC's Uniform Bank Performance Reports?
- Who is running Prof. Jones' model? What input data are they using it with and with what parameters?

The considerable body of research in data provenance, primarily motivated by similar challenges in large-scale scientific applications, partially addresses these needs [14]. There is widespread agreement among researchers that provenance is best represented as a directed, acyclic graph that includes two kinds of nodes: data and processes [8].

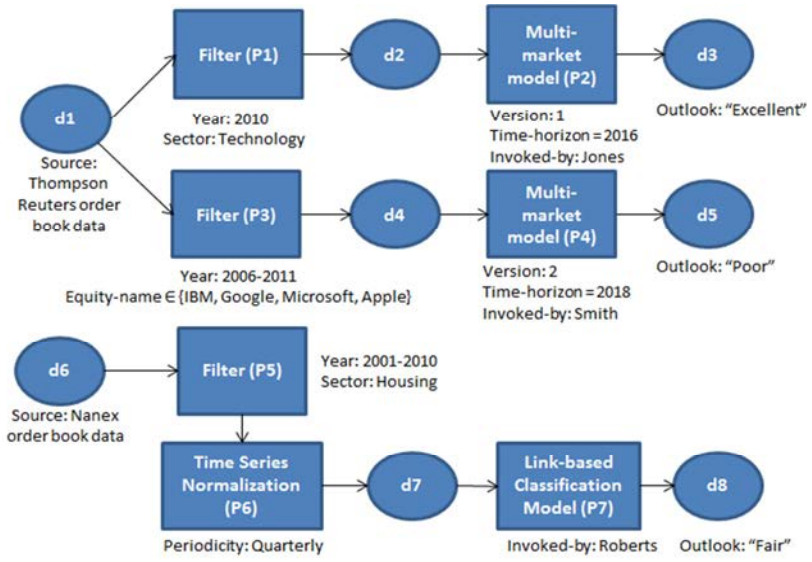

Figure 1: Sample provenance graph. Ovals represent data, while rectangles represent processes. Arcs represent data flows.

Figure 1 shows a simple provenance graph with data transformations and simulation runs of financial models. Ovals represent information about datasets, while rectangles represent processes, such as running a data transformation or a simulation model. Arcs represent data flow among the nodes. In the example graph, two processing chains emanate from Thompson Reuters order book data (node d1), while another emanates from Nanex order book data (d6). The top row shows one processing chain: a filter (node P1) is run on d1 producing dataset d2 which in turn feeds a run of a multi-market model (P2) and produces dataset d3. Metadata can be associated with any node; for example, the run of the multi-market model (P2) was invoked by analyst Jones using version 1 of the model and a time horizon of 2016.

Provenance addresses the challenges described above. First, it helps analysts understand the data assumptions behind different simulation runs. For example, in Figure 1, nodes P2 and P4 represent runs of the multi-market model that produce outputs d3 and d5 respectively. In this simple example, d3 and d5 differ in their assessments of the economic outlook ("excellent" vs. "poor"). What might account for the differences? Data provenance allows the user to see exactly what input data and model version were used, what transformations were performed on the data, and the parameter settings used. Without automatic capture of this information as simulations and data transformation tools run, it is very difficult to recreate the provenance retrospectively by examining scripts and individual analysts' notes.

Provenance also helps analysts find collaborators. For example, a new analyst Wilson could query to see who is using the multi-market model (Smith and Jones), how they are using it, and with what data. Alternatively, Wilson could query to see who is using order book data (Smith, Jones, and Roberts), what sources they are using for it (Thompson Reuters and Nanex), and what models they are using the data for. These queries would be very difficult to answer in a large-scale analytic environment without provenance.

3. AUTOMATIC PROVENANCE CAPTURE

The previous section illustrated how provenance information enables analysts to better understand the data and assumptions used for potentially vast numbers of simulation runs. However, it is not enough to provide data structures, query mechanisms, and graph renderings for provenance; one also needs a scalable strategy for collecting provenance.

The current state of the practice is manual population (also called manual tagging); that is, some human must manually enter one or more data fields that describe the provenance. Typical field names include "source", "pedigree", "lineage", or "provenance." There are many limitations of the manual population approach. First, because it relies on a manual process, the provenance is often left blank. Second, even when it is filled in, there is no mechanism to ensure that it is done consistently—i.e., is it the original source or the immediate ancestor? Third, when it is filled in, it is "one-hop" at best; indicating an immediate ancestor, with no linkage further back; no information is available about what intervening processes (e.g., algorithms, human review) acted upon the data. Because the field is often a free-form text field, there is no way to capture data's "family tree." In sum, for provenance to be widely and consistently captured (and therefore useful to financial analysts and decision makers), there must be mechanisms to automatically capture it and to populate the full family tree.

A number of research efforts have explored automatic capture of provenance for scientific workflows [6, 10], derivation of data in relational databases [3], and provenance-aware storage systems [9]. The PLUS system [5] extended that work to support automatic provenance capture in distributed, heterogeneous environments in a way that minimizes the need to modify legacy applications. Similar to [7], PLUS supplies an API that any application can call to log provenance information. However, in addition to this basic service, PLUS provides provenance capture at "coordination" points that are often used in distributed systems. For example, an Enterprise Service Bus (ESB) is often used to coordinate applications comprised of data and components from many different organizations. PLUS includes a provenance collector for Mule, a popular open source ESB, to automatically capture and report provenance for all messages passed [1]. The PLUS provenance manager tracks this provenance information and integrates it with provenance captured in other systems.

4. PROVENANCE CAPTURE FOR SYSTEMIC RISK ANALYSIS

While PLUS previously demonstrated automatic provenance capture at selected system coordination points, we still needed to identify high-value coordination points for financial systemic risk analysis. [16] identifies critical needs for this domain, two of which offer excellent provenance capture points:

- Data transformation, to slice and dice data to meet the needs of diverse models
- A facility that supports easy specification and efficient batch execution of large numbers of simulation runs

To address these needs, we have implemented a PLUS provenance capture module for Pentaho's Kettle data transformation tool and are currently implementing one for the MITRE Elastic Goal-directed (MEG) simulation middleware [11].

Pentaho Kettle (kettle.pentaho.com) is an open source Extract-Transform-Load (ETL) tool. It includes a large library of built-in, atomic data-transform steps (e.g., input-from-CSV, output-to-tab-delimited, filtering, joins, string manipulation, arithmetic conversions, etc.) and provides a GUI for graphically tying together these steps into a "transformation." The steps of the transformation are captured in an XML file, which is then executed by Kettle's run-time engine. We added hooks in the run-time engine to parse the XML file and record the steps of the transformation as they are being run using the PLUS provenance capture API. As a result, any transformation that can be defined by Kettle can be captured in our provenance store. To our knowledge, this is the first example of a general purpose provenance tool automatically capturing provenance from an ETL tool. This is much more powerful than provenance that is internal to the ETL tool, since it supports queries over provenance that spans multiple system boundaries.

The MEG simulation middleware [11] provides a number of powerful features for analysts running financial simulations, including:

- A Design of Experiments (DoE) tool that supports automatic spawning of large numbers of simulation runs to explore a space of solutions, either via parameter sweeps or optimization of some function over the simulation's output values (i.e., optimization via simulation)
- Automatic, efficient scheduling and execution of these runs across various high-performance computing clusters.

A sister project of ours has wrapped several financial simulations to run within MEG. We are currently designing a provenance capture module for MEG, which will enable automatic provenance capture for any simulation run within MEG.

Prototype Implementation

Figure 2 illustrates the architecture of a prototype Financial Modeling and Analysis Environment (FMAE) under development at the MITRE Corporation. FMAE is built on top of the PLUS provenance manager, which uses the MySQL open source database management system as a back-end store. We have extended the PLUS schema to capture metadata important to financial datasets (e.g., periodicity, start and end dates) and executable models (e.g., version, language). An analyst populates this metadata using the dataset/model registration module. As noted above, we have completed provenance capture for the Pentaho Kettle data transformation tool and are in the process of building one for MEG. The PLUS provenance manager will seamlessly integrate provenance captured for both Pentaho Kettle and MEG. Analysts will

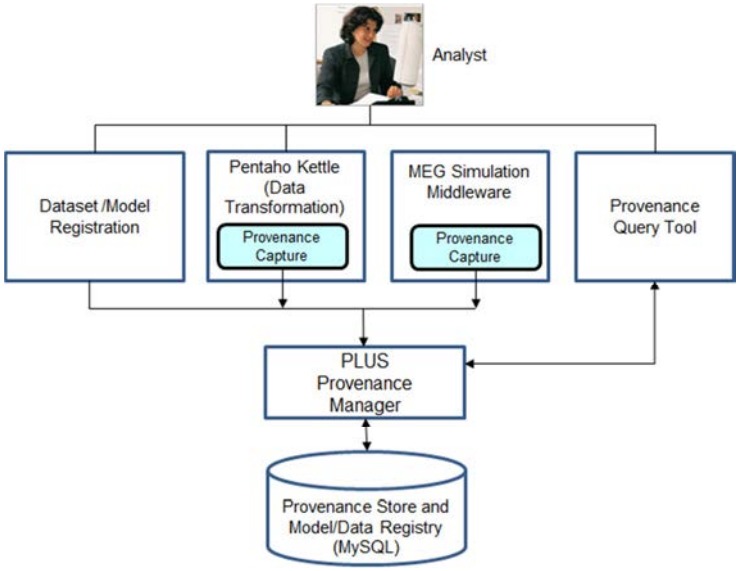

Figure 2: Financial Modeling and Analysis Environment Architecture

then be able to browse and navigate provenance graphs using PLUS' generic GUI, which provides an early instantiation of FMAE's provenance query tool.

Future Work

PLUS's current graphical displays will probably not be adequate for financial analysts, given the large expected size of provenance graphs in this domain. To address this need, we plan to develop a provenance query tool designed specifically to support financial analysts' most important queries. Next, we will conduct pilots to assess the usability and scalability of these tools. Our expectation is that the end product will greatly improve financial analysts' understanding of the provenance of data in a large-scale analytic environment.

Acknowledgements

We thank Raj Shenoy, Matt McMahon, and David Allen for help with the MEG and PLUS systems, Jeff Hoyt for development wizardry, and Brian Tivnan for helpful comments.

REFERENCES

[1] Allen, M. D., Chapman, A. Blaustein, B. and Seligman, L. Provenance capture in the wild, *International Provenance and Annotation Workshop (IPAW)*, 2010

[2] Ballou, D. P., R. Y. Wang, H. Pazer and G. K. Tayi, Modeling information manufacturing systems to determine information product quality. *Management Science*, *44*(4) 1998, pp. 462-484.

[3] Benjelloun, O., A.D. Sarma, C. Hayworth, and J. Widom. An introduction to ULDBs and the Trio system, *IEEE Data Engineering Bulletin*, 29(1), 2006

[4] Bisias, D., Flood, M., Lo, A.W., Valavanis, S. A survey of systemic risk analytics, Working Paper #0001, Office of Financial Research, January 5, 2012

[5] Chapman, A., M.D. Allen, B. Blaustein, L. Seligman. PLUS: a provenance manager for integrated information, *IEEE International Conference on Information Reuse and Integration*, Las Vegas, NV, August 2011

[6] Davidson, S., S. Cohen-Boulakia, A. Eyal, B. Ludascher, T. McPhillips, S. Bowers, and J. Freire, Provenance in scientific workflow systems, *IEEE Data Engineering Bulletin*, 32(4), 2007

[7] Groth, P. Miles, S., and Moreau, L. PReServ: Provenance recording for services, *UK e-Science All Hands Meeting*, 2005.

[8] Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J., and Paulson, P. The open provenance model: an overview, *International Provenance and Annotation Workshop (IPAW)*, 2008

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

[10] Oinn, T., M. Greenwood, M. Addis, M.N. Alpdemir, J. Ferris, et al. Taverna: lessons in creating a workflow environment for the life sciences, *Concurrency and Computtion : Practice & Experience*, 18(10), 2006.

[11] Page, E.H., Litwin, L., McMahon, M.T., Wickham, B., Shadid, M., and Chang, E.. Goal-directed grid-enabled computing for legacy simulations, *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, May 2012

[12] Rosenthal, A. and L. Seligman. Data integration for systemic risk in the financial system, in Fouque, J-P, Langsam, J. (eds.), *Handbook of Systemic Risk*, Cambridge Univ. Press, to appear, 2012

[13] Shankaranarayan, G., Ziad, M., Wang, R.Y., Managing data quality in dynamic decision environments: an information product approach, *Journal of Database Management*, 14(4), 2003

[14] Simmhan, Y., Plale, B., Gannon, D. A survey of data provenance in e-science, *SIGMOD Record*, 34(4), September 2005

[15] Wang, R.Y., A product perspective on total data quality management, *Communications of the ACM*, 41(2), February 1998.

[16] Worrell, C., Guharay, S., McMahon, M., Seligman, L., Shenoy, R. Operational considerations in an analytic environment for systemic risk, in Fouque, J-P, Langsam, J. (eds.), *Handbook of Systemic Risk*, Cambridge Univ. Press, to appear, 2012