社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－DPS－138 (3)
2009－CSEC－ 44 (3)
2009／3／5

Securing Provenance by Distributing the Provenance Storage

Amril Syalim
Graduate School of Information Science and Electrical Engineering,
Kyushu University, Fukuoka, Japan
amr@itslab.csce.kyushu-u.ac.jp

Yoshiaki Hori and Kouichi Sakurai
Faculty of Information Science and Electrical Engineering,
Kyushu University, Fukuoka, Japan
{hori, sakurai}@csce.kyushu-u.ac.jp

ABSTRACT

Provenance is defined in some literature as a complete documentation of processes that led to an object. Provenance can be applied in some systems: database systems, file systems and grid systems. Provenance is very important to verify the processes that led to an object. There are many research to develop a provenance system. Main provenance implementations use a centralized model or a centralized management of provenance system. Although many provenance stores may be used and the users may choose a trusted provenance store as the place to store their provenance, when recording a provenance information, the users should store a complete provenance information in a particular provenance store. We believe that this model has a weakness, that is the user can not restrict access of the administrator that maintains the provenance store where the user store their provenance. We propose an alternative to this model, that is a user may store some parts of the provenance information in a provenance store and some other parts in other provenance stores. We also discuss security advantages of this approach.

1 INTRODUCTION

Provenance of an object is defined in some literature as a complete documentation of process that led to the object. Provenance can be used in many contexts, e.g. database systems, file systems and grid systems. In a database, provenance of a data item is a complete documentation of process that led to the the data item. In a file system, provenance of a file is a complete documentation process that led to the file. In a grid system, provenance is a complete documentation of processes that led to the output of computation in the grid system.

Provenance is important because if we know the track record of data, we will understand the value of data. The data with a complete provenance will have higher value than those of without any documentation of process that produce them. By knowing provenance we may reproduce the data when we need them even without knowing/having the data. This is particularly important in the context of e-Science, when we use computation resources for solving scientific problem. In the e-Science infrastructure, the provenance is important because the data with a complete documentation will be easily reproduced with the same or different parameters. By knowing provenance the scientists can also easily verify the result of an experiment.

There has been considerable interests on the method to record the provenance information. The main provenance implementation use the concept of provenance store [4, 5, 8], that is a system that has interface to store and query provenance record (Figure 1). This architecture is much similar to the database system where users can do query to the provenance store that has interface for provenance management.

Provenance can be represented at some granularities. Although there are some efforts in developing the provenance representation [21, 22], in this paper we do not stick to any available provenance representation. We assume that provenance is stored in a database regardless of how to represent the provenance in database (relational or xml).

The problem in building a provenance system are [4] (a) provenance modeling: that is how to represent the provenance information in a storage, (b) scalability: how to manage a huge amount of provenance information and (c) security: how to secure the provenance information. In this paper we focus on the security issues of provenance especially in a distributed grid based system where the users of provenance store may come from many different organizations.

Main provenance implementations [4, 5, 8] use a centralized provenance system, where the provenance information is stored in a centralized provenance store (or with a distributed storage but a centralized management). The centralized provenance system has advantages in simplicity and easy to manage. Although the users may choose a trusted provenance store to store his/her provenance information, when recording the provenance information, the users should store a complete provenance information in a particular provenance store. We believe that this model has a weakness, that is the user can not restrict access of the administrators that maintain the provenance store where the user store their provenance.

A solution for this problem is that the user encrypt the provenance information to protect from unauthorized users/administrator. However, this method needs much cost in computation and will degrade performance. In this paper we propose an alternative to this model, that is a user may store some parts of the provenance information in a provenance store and some other parts in other provenance stores. Using this approach, an administrator in a provenance store can only access some part of the provenance information. To access all part of the provenance store the administrator should cooperate with all administrator of the provenance stores where the user store his/her provenance information.

Organization of this paper is as follow: first we discuss the provenance store, after that we discuss provenance representation including P-assertion and provenance graph. After then we discuss problem definition and proposed model. Before closing the paper with conclusion we discuss some requirements, security and feasibility of the proposed model and some related works.

2 Provenance Store

Main provenance systems use the concept of provenance store [4, 5, 8], that is a system that has interface to store and query provenance record (Figure 1). This architecture is much similar to the database system where user can do query to the provenance store that has interface for provenance management. The provenance store can be accessed by users to store and share the provenance record.

A provenance store, although has a centralized management, may have distributed storage. The centralized provenance system has advantages in simplicity and easy to manage. When recording provenance for data, a user should record complete provenance in a provenance store.

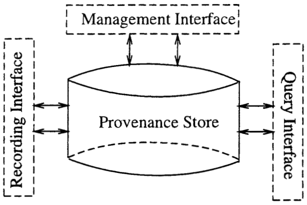

Figure 1: Provenance Store

3 P-Assertion and Provenance Graph

There is no standard model of provenance although there is an attempt to develop the standard [9]. The EU provenance project develop model of provenance record using the contept of p-assertion [4]. They defined p-assertion as an assertion that is made by an actor and pertains to a process. The provenance (documentation of a process) consists of a set of p-assertions made by the actors involved in the process.

There are two types of p-assertion to represent relationships between entities in the system. An interaction p-assertion is an assertion of the contents of a message by an actor that has sent or received that message, a relationship p-assertion is an assertion, made by an actor, that describes how the actor obtained output data or the whole message sent in that interaction by applying some function to input data or messages from other interactions [4]. With these types of p-assertion, we can make a model of interaction between entities in a service oriented architecture (SOA) provenance system.

Provenance can be represented by a directed cyclic graph (DAG) [6]. Each node in the graph represents an entity and each edge in the graph represents a causal relationship between two entities. Examples of entities include processes, people or data, i.e files.

An example of provenance in the figure 2 below. In the example below to produce data D3A in the figure 2 we need to execute process PB with input data D0B and D0C and process PC with input data D0D. The output of process PB is data D1B and the output of process PC is data D1C. After then we execute process PE with input data D1B and data D1C. The output of process PE is data D2B. We get data D3A from the output of process PF with input data D2B.

To produce data D2A in the figure we execute pro-

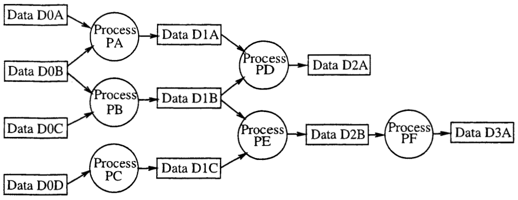

Figure 2: Provenance Graph

cess PA with input data D0A and D0B and process PB with input data D0B and data D0C. We send the result of process PA (data D1A and the result of process PB (data D1B) to process PD. Data D2A is result of computation at the process PD.

4 Problem Definition

In this section we discuss security problems in a provenance system. Before stating the problem definition, we identify four active entities in the provenance system (Figure 3). Active entities are the entity that has the possibility to access the provenance store (store or fetch). These active entities are:

1. User
   A user is a human being that has capability to record provenance to the provenance store and to query the provenance store
2. Administrator
   The administrator is a human being that has capability to manage the provenance information: change/add/delete/move any the provenance information in the provenance store.
3. Process
   A process is a computer program that has capability to record the provenance automatically.
4. Outsider
   An outsider is a human being or a process that is not authorized to access the provenance store but may have access to the network in the provenance store

The process to record provenance are as follow:

1. A user execute a workflow that will produce data
2. The processes that execute the workflow record the provenance information to a provenance store
3. The user may add additional provenance information to the provenance store

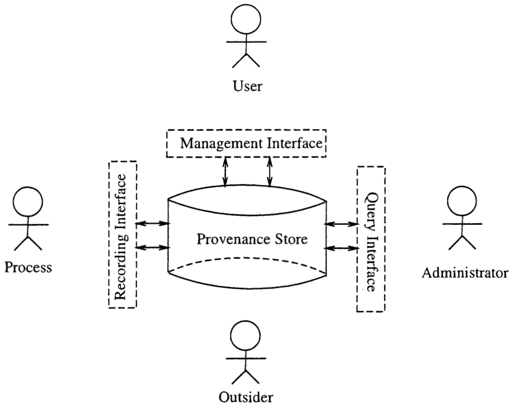

Figure 3: Active entities in the provenance store

An administrator have access to all of the provenance information that is stored in the provenance store maintained by the administrator. A user should store a complete provenance information in a provenance store whenever he/she need to store the provenance information. This model has a problem because a user can not restrict access by administrator of the provenance store to the provenance information. Although the user may restrict access by other users, the user should trust his/her provenance information to the administrator of the provenance store.

In the next section we discuss an alternative, that is by distributing provenance information in more than one provenance store. By using this approach an administrator can not access a complete provenance information. To access a complete provenance information the administrator should cooperate with all of administrators where the user store the provenance information.

5 Securing Provenance by Distributing Provenance Storage

In this section we discuss a method to secure provenance by distributing the provenance storage. First we discuss the basic model. After then we discuss the requirements for implementing the proposed model.

5.1 Basic Model

Our basic model is shown in the Figure 4. As shown in the Figure, whenever a process records the provenance information, the process should divide the provenance information into some parts and record those parts in some provenance stores. To get a complete provenance information a user should query all of the provenance stores that store the provenance information.

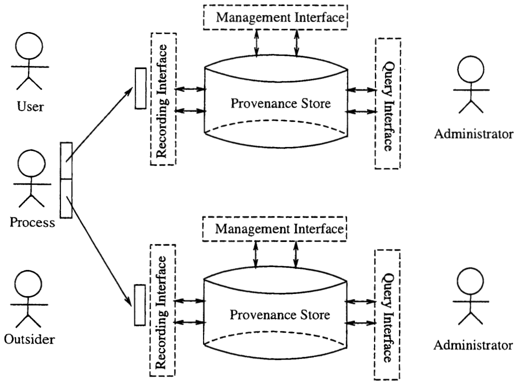

Figure 4: Distributed Storage Model

5.2 Requirements

We identify some requirements and challenges to implement this model:

1. The method to divide provenance information
   To implement this model, first we need a method to divide the provenance information into some parts. The method should divide the provenance information so that it can be reconstruct later easily.

2. The method to reconstruct the provenance information
   We also need a method to reconstruct the provenance so we can get a complete provenance after querying all of provenance stores where we store the provenance information.

6 Discussions

In this section we discuss security advantages of this approach. The security advantages include confidentiality, integrity and availability.

6.1 Confidentiality

By using this approach, we may improve confidentiality especially from attacks of administrators of the provenance stores. An administrator can not get a complete provenance information because the the provenance information is distributed at some provenance stores. To get all parts of the provenance information, the administrator should cooperate with all of other administrators where the user store the provenance information.

6.2 Integrity

The integrity may also be improved. By using a provenance store to store a complete provenance, the administrator may change all parts of the provenance information. However, by distributing the storage, an administrator can only change one part of the provenance information. To change all parts of the provenance information, the administrator should cooperate with all of other administrators where the user store the provenance information.

6.3 Availability

Availability can be increased and can also be decreased. Availability is increased because by separating the storage the queries for the storage (fetch or store) of provenance system are divided to some systems/storages. The bottleneck using one storage can be minimized. The availability can also be decreased because if one provenance store fail to work, the user cannot get a complete provenance store when he/she need the complete provenance information.

7 Related Works

Braun et al. and Tan et al. have discussed security issues on provenance [6, 1] although they did not propose any security system related to the issues. Braun et al. identified some of the security characteristics of provenance. The first is that provenance differs from data in that it forms a directed acyclic graph (DAG) so we need to have a security model for a directed acyclic graph (DAG). Second issue is that sensitivity level of data and its associated provenance may be different. It is possible that the provenance be more sensitive than data or vice versa. Tan et al. lists six security issues in a SOA-Based provenance system [1]. These security issues are (1) enforcing access control over process documentation, (2) trust framework for actors and provenance stores, (3) accountability and liability for p-assertions, (4) sensitivity of information in p-assertions, (5) long term storage of p-assertions, and (6) creating authorizations for new p-assertions. They emphasis that the first issue is unique to the provenance purposes because the requirements are different from regular data.

Groth et al. have proposed an architecture of provenance system including the security architecture in an EU sponsored project [4]. They have implemented the architecture in a SOA-based provenance store. They suggested that access control should be specifiable at the level of individual p-assertions and at individual elements within p-assertion if needed. They also suggested to use role-based access control and content-based access control although no detail explanation and implementation of their proposal.

Chebotko et al. [10] proposed a secure scientific workflow provenance querying with security view. Security view is a subset of data and processes. The main different of their work with ours is that they emphasis the use of view to enforce access control policy.

Another related work is the work Nagappan et al. nagappan1. They proposed a model of sharing confidential provenance information where an the actor who are willing to share the provenance information can share the query for that provenance information.

8 Conclusion

In this paper we have discussed a method to secure provenance by distributing the provenance storage. The main idea is that by dividing the provenance information into some parts and store those parts at some different provenance store. This method improve security (confidentiality, integrity and availability) because the administrator of a provenance store cannot access/change all parts of the provenance information and bottleneck of single provenance store can be minimized.

9 Acknowledgment

The first author is supported by the Japanese Government (Monbukagakusho) scholarship.

References

[1] Victor Tan, Paul Groth, Simon Miles, Sheng Jiang, Steve Munroe, Sofia Tsasakou and Luc Moreau, Security Issues in a SOA-Based Provenance System, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006.

[2] Liming Chen, Victor Tan, Fenglian Xu, Alexis Biller, Paul Groth, Simon Miles, John Ibbotson, Michael Luck, and Luc Moreau. A Proof of Concept: Provenance in a Service Oriented Architecture. In Proceedings of the Fourth All Hands Meeting (AHM), September 2005.

[3] Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, and Luc Moreau. An Architecture for Provenance Systems — Executive Summary. Technical report, University of Southampton, February 2006.

[4] Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, and Luc Moreau. An Architecture for Provenance Systems. Technical report, University of Southampton, November 2006.

[5] Groth, P., Miles, S. and Moreau, L. PReServ: Provenance Recording for Services. UK e-Science All Hands Meeting 2005, September 2005, Nottingham, UK.

[6] Uri Braun, Avraham Shinnar, and Margo Seltzer. Securing Provenance., In Proceedings of the 3rd USENIX Workshop on Hot Topics in Security (HotSec), San Jose, CA, July 2008.

[7] Shawn Bowers, Timothy M. McPhillips, Bertram Ludascher, Shirley Cohen, Susan B. Davidson: A Model for User-Oriented Data Provenance in Pipelined Scientific Workflows. IPAW 2006: 133-147

[8] Yogesh L. Simmhan , Beth Plale , Dennis Gannon, A Framework for Collecting Provenance in Data-Centric Scientific Workflows, Proceedings of the IEEE International Conference on Web Services, p.427-436, September 18-22, 2006

[9] Luc Moreau, Juliana Freire, Joe Futrelle, Robert E. McGrath, Jim Myers and Patrick Paulson, The Open Provenance Model: An Overview, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008.

[10] A. Chebotko, S. Chang, S. Lu, F Fotouhi and P. Yang,Secure Scientific Workflow Provenance Querying with Security Views, by 9th International Conference on Web-Age Information Management (WAIM), pages 349-356, IEEE press, 2008.

[11] Meiyappan Nagappan and Mladen A. Vouk, A Model for Sharing of Confidential Provenance Information in a Query Based System, Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17-18, 2008.

[12] Simon Miles, Paul Groth, Miguel Branco, Luc Moreau, "The requirements of recording and using provenance in e-science experiments," Journal of Grid Computing, 2005.

[13] Jonathan Ledlie, Chaki Ng, David A. Holland, Kiran-Kumar Muniswamy-Reddy, Uri Braun, and Margo Seltzer, "Provenance-Aware Sensor Data Storage," In Proceedings of NetDB 2005, Tokyo, Japan, April 2005.

[14] Todd J. Green, Zachary G. Ives, Grigoris Karvounarkis, Val Tannen, "Update Exchange with Mappings and Provenance," VLDB 2007.

[15] Uri Braun and Avi Shinnar, "A Security Model for Provenance," Technical Report TR-04-06, Harvard University, 2006.

[16] PASS Project. http://www.eecs.harvard.edu/~syrah/ pass/.

[17] Provenance Project. http://www. gridprovenance.org/.

[18] PReServ. http://twiki.gridprovenance.org/bin/ view/PASOA/SoftWare

[19] Provenance Store Service. http://www. gridprovenance.org/software/PService.html.

[20] Karma Provenance Framework. http://www. extreme.indiana.edu/karma/.

[21] http://openprovenance.org/

[22] Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J. and Paulson, P. (2007) The Open Provenance Model. Technical Report, ECS, University of Southampton.

[23] Oxford Advanced Learnerfs Dictionary. http://www. oup.com/elt/catalogue/teachersites/oald7/?cc=global

[24] Secure Provenance. http://www. ragibhasan.com/research/provenance.html

[25] R Hasan, R Sion, M Winslett, "Introducing secure provenance: problems and challenges," Proceedings of the 2007 ACM workshop on Storage security and survivability, October 2007.