Study of Constructing Data Supply Chain Based on PROV

Jiewei Lan[✉], Xiyun Liu, Hong Luo, and Peng Li

Department of Computer Science, Beijing University of Posts
and Telecommunication, Beijing 100876, China
{lanvivian,spooons,luoh,lipeng1106}@bupt.edu.cn

Abstract. In the era of big data, the value of data can be better explored during data flowing and processing. If a data supply chain from the source to the destination is constructed across data platforms where data flows through, then it will help users analyze and use these data more safely and effectively. Due to the complexity and diversity of data platforms, there is no uniform data supply chain model specification. To solve the problem, we construct a distributed data supply chain model based on PROV, a data provenance specification presented by W3C to standardize information records of data activities in corresponding data platforms. On this basis, we design Data Supply Chain Service Module (DSCSM), so as to provide effective accessing methods for data traceability information on distributed platforms. Finally, we deploy the proposed model to real data platforms we built to verify the effectiveness and feasibility of solution.

Keywords: PROV · Data supply chain · Distributed

1 Introduction

In the era of big data, decision-making in commercial, military or other fields is increasingly dependent on big data analysis rather than experience and intuition, which puts forward higher requirements on the data management. To meet needs from various applications, data stored in dispersed data platforms flows frequently and is complicatedly processed. In order to ensure the security and authentication of data, researches on data provenance are more extensive [1]. Cause records of data activities are stored in distributed data platforms, it is difficult to get it effectively and safely. If we can sort out data relationships among platforms using provenance information, it will not only bring benefits to data analysis but also provide a safe and efficient method to access the data.

There is no uniform standard for data provenance record until now, which is not conducive for data integration and analysis in distributed environment. PROV [2] presented by W3C is a data provenance specfication. Because of its strong analyticity and semantic feature, PROV is suitable to be used as data provenance standard. Data provenance researches mostly focus on how to record

Y. Wang et al. (Eds.): BigCom 2015, LNCS 9196, pp. 69–78, 2015.
DOI: 10.1007/978-3-319-22047-5_6

provenance information and how to analyze it. The information are usually stored in a central server or follow the data flows. However, provenance information increases with the multiple data transfer. When circulation gets more frequent, provenance data may be more massive than the data itself. And security of centralized storage is poor. So this model does not apply to information environment in the era of big data. Thus, considering about the reliability and security, we store data provenance information in distributed platforms and construct a dynamic and efficient data supply chain to obtain information quickly and safely. It will also do good to dynamic data management and access control.

Since massive provenance information are stored in distributed data platforms, how to get these information becomes a big challenge. As provenance is analogous to object tracking information of supply chain and discovery service (DS) in EPCGlobal network is used to access tracking information, we learn from SCOR (Supply Chain Operations Reference model) and analyze discovery service. It turns out that distributed index model in DS works better for data acquisition. Therefore, we design data supply chain service model combined with distributed index model.

Based on the PROV specification, we propose a data provenance model to formally describe how the data is received, processed and provided, which standardizes records of data activities in corresponding data platforms. Further more, we construct data supply chain service model to obtain provenance information combined with distributed index model in EPCGlobal framework and new characteristics of data supply chain. It can organize scattered data stored in the heterogeneous platforms to one or more data supply chains, which will provide convenience to form data ecosystem and further explore data value.

The remainder of this paper is organized as follows. We discuss the related work in Section 2. Section 3 describes the prov data model. The proposed Data Supply Chain Service Module is detailed in Section 4. In Section 5, we give our experimental study and simulations. We conclude the paper in Section 6.

2 Related Work

Currently, researches about big data mainly aim at data mining [3], data credibility [4], privacy protection [5], access control [6] and etc. However, there are relatively fewer studies about data traceability information among distributed platforms. We intend to construct data supply chain based on data provenance. And here we present the research states of data provenance.

2.1 Data Provenance

In the era of information data mostly includes raw data and derived data. Due to the complexity and variety of derivation, data users usually have doubt on the reliability of data. In fact, many mistakes of derived data have nothing to do with the raw data. Therefore, it is necessary to record provenance information of data so that people can build data supply chain for scattered data.

The data supply chain construct model mainly includes the record and acquisition of data provenance. When it comes to the record of data provenance, how to get provenance information and which information should be recorded are taken into consideration. Early studies stored data and intermediate data together, which would cause mismanagement easily. With the expansion of data volume and penetrating of study, Buneman et al presented a provenance model [7] of Why and Where. Based on the model, Vansummeren proposed a model called "how provenance" [8]. Then Sudha et al brought a 7w model [9] which means who, when, where, how, which, what and why. 7w model records comprehensive details of data provenance but it will increase storage space.

Although researches about data provenance model are quite aplenty, it is still difficult to consolidate variety information from different platforms and build data supply chain cause lacking of uniform data provenance standard. The PROV specification presented by W3C gives a standard description of data provenance. It has strong analyticity and semantic feature cause it can be recorded in forms like XML, JSON, OWL2 and etc. What's more, PROV defines inference rule to consolidate provenance information, which makes it possible to construct data supply chain and achieve data provenance among distributed platforms. As a conclusion, we choose PROV as the standard of data supply chain model.

However, cause massive provenance information are stored on distributed data platforms, how to get these information becomes a big challenge. Since provenance is analogous to object tracking information of supply chain in EPCGlobal network [10], it brings a good reference to this issue.

2.2 Supply Chain

Provenance information of data supply chain has some similarities with the object tracking information of supply chain. Discovery service technology in EPCGlobal framework is designed to achieve object tracking information dynamically. However, data relationships and data processes in data supply chain are more complex and diverse. So we should build flexible data supply chain to acquire provenance information based on supply chain research combined with characteristics of data supply chain. There are three kinds of discovery service technology in EPCGlobal framework.

Centralized Server Model. There is a central server in this model [11]. Information of supply chain will not only be put in local-storage but also sent to central server, which is suitable for logistic query. It is easy to accomplish this method but there is no guarantee on safety since central server can be overload easily.

Centralized Index Model. In Centralized index model, supply chain data is put in the local-storage and at the same time the index related will be sent to central server [12]. By using this model, storage pressure will be reduced and privacy of supply chain data can also be protected. But once the central server is broken down, information traceability will be influenced heavily.

Distributed Index Model. In this model, central server is replaced by distributed structure. Each local server is bound with a query engine. When users want to query for data of supply chain, only need to send the query to the related query engine. Then a process of distributed index model called "Process and Forward" [13] will send queries to related participants spontaneously and recursively. Distributed index model reduces storage and safety demand to central server and intercurrent pattern can also improve query efficiency. In conclusion, this model offers valuable references to data supply chain construction.

Combined with characteristics of data supply chain and distributed index model, we propose data supply chain service model to get provenance information stored in distributed system. It will bring benefits to build flexible and dynamic data supply chain so as to manage and use big data in better ways.

3 PROV Data Model

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. PROV is a standard data model defined by W3C. It allows users to standardly describe provenance that exist in different information systems by using widely available formats like XML, JSON, RDF, etc, which offers great convenience in operations. Meanwhile, the PROV specification also presents the standardized and readable description format-PROV Nation (PROV-N). The conceptions of the PROV data model are introduced in the following contents.

Entity. In PROV, an entity is a physical, digital or other kind of thing whose provenance information is to be recorded. The provenance of one entity might be related to other entities. For instance, an document entity includes a data table and the data is derived from a database. Therefore, the document is in connection with the table and the database.

Activity. Activities are how entities come into existence and how their attributes change to become new entities, often making use of previously existing entities to achieve this. For instance, the process of translating a document into another language, which creates a new document, is called activity.

Agent. An agent is a person, an organization, maybe a software that have responsibility for an activity that take place. The relationship between agent and its corresponding activity is called "associate". Several agents can be associated with same activity. Entities, which involved in the activity, are attributed to the related agent. For instance, an activity tabulates the data into a table. The person who makes data table and the software used by this person are agents that associated to this activity. And the table is the entity that attributed to these two agents.

Generation and Usage. Generation is the production of a new entity by an activity. For instance, the activity of writing a document generates an entity called document. Usage is the utilization of an entity by an activity. For instance, correcting the spelling mistakes will use the former edition.

Derivation and Revision. When part of an entity, such as existence, content, property and so on, can be dated from the other entity, we defined that the former is derived from the latter. Like, a table is derived from its used data. Revision is a special form of derivation. Entities, like documents, might be revised for many times and every time the revision will create new entities.

The following Fig.1. provides a high level overview of the structure of PROV records, limited to some key PROV concepts discussed in this paper.

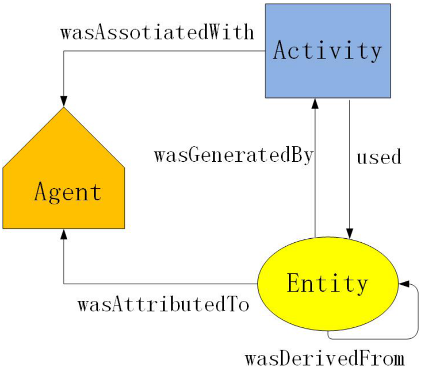

Fig. 1. Structure of PROV data model

4 Data Supply Chain

In the era of big data, due to the numerous and jumbled information, the various provenance and disparate data structures, it is difficult to manage data uniformly among different platforms. PROV provides the standard data model for data provenance. Based on this specification, it is convenient to construct the data supply chain in heterogeneous environment.

4.1 Data Supply Chain Model

Suppose there are information platforms of three different companies: A, B and C. Data d flows from platform A to B. After being disposed by B, it turned into d1 and then flows to platform C. Also disposed by C, d1 turned into d2. As shown in Fig.2.

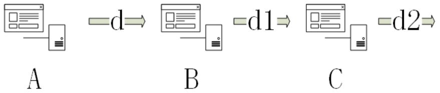

Fig. 2. Data flows among information platforms

Entity. Data d, d1 and d2 are entities based on the definition of PROV data model. According to the description of PROV-N, cA, cB and cC are called prefix, serving as organizations that entities belong to. PROV-N describes entity d1 as follows.

```
entity (cB:d1)
```

Activity. The process to dispose data, executed by the information platform A,B and C, is called activity, based on the definition of PROV data model. We call these two activities process A, process B and process C. PROV-N describes activity processB as follows. "tB" is occurrence time.

```
activity (cB:processB, tB)
```

Agent. The programs used by platform A, B and C to dispose data are called a, b and c. Therefore, program a, b, c and orgnization A, B, C that programs belong to are "agent". Agent can record the information of agent and its relationship with entity, activity. PROV-N describes agent B as follows.

```
agent (cB:b, [prov:type="SoftwareAgent", foaf:givenName="b"])
agent (cB:B, [prov:type="Organization", foaf:givenName="B"])
actedOnBehalfOf (cB:b, cB:B)
wasAssociatedWith (cB:processB, cB:b, -)
wasAttributedTo (cB:d1, cB:b, -)
```

Generation and Usage. The relationship between activity and entity is described by generation and usage. They are described by PROV-N as follows (tB is occurrence time).

```
used (cB:processB, cA:d, -)
wasGeneratedBy (cB:d1, cB:processB, tB)
```

Derivation and Revision. When data need to be revised because of mistakes or information update, the processes to complete the action are called derivation and revision. For instance, data d of information platform A is revised into newd and then the data of information platform B and C turn into new1 and new2. PROV-N describes derivation and revision as followed.

```
wasDerivedFrom (cB:newd1, cA:newd)
```

4.2 Dynamic Data Supply Chain Service

Using PROV specification to build data provenance model can standardize records of data activities in corresponding data platforms. On this basis, we can construct dynamic data supply chain. Considering about the variety of information platform, distributed data chain construction has greater advantages over the centralized ones.

We design Data Supply Chain Service Module (DSCSM) combined with discovery service mechanism of supply chain and the characteristics of data supply chain. Each DSCSM communicates with related upstream and downstream data platforms, so as to provide effective accessing methods for data traceability information and further protect the data privacy. The structure of the system and process of data query are shown in Fig.3. Construction of distributed data chain can be divided into four stages.

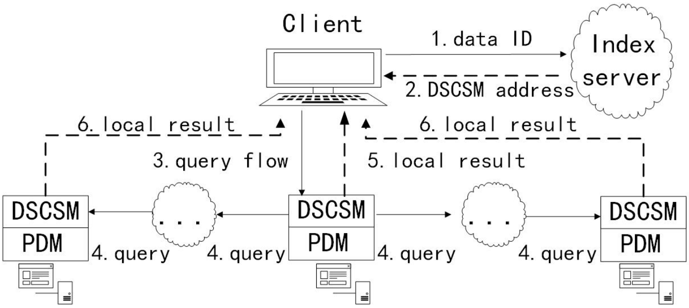

Fig. 3. Structure of data supply chain construct system

Send Query. Firstly, client generates a data chain query with a unique query ID. Query parameter "data ID" refers to the data and its current attribution-prefix and entity, like cB:d1 (step1). Then client will submit the query parameter to index server. After analyzing, server sends back a related DSCCM address (step2). Next, client sends a query to the DSCCM address (step3).

Process and Forward. After DSCCM receives the query, firstly it will visit local PDM in order to get local PORV records. Then DSCCM will generate some new queries to its related upstream and downstream according to the PROV records (step4). Query parameter which is submitted to upstream, is the source data (includes organization it belongs to) of local record "used". Query parameter which is submitted to downstream is in accordance with the query parameter received by now. For instance, information platform B send query parameter "cA:d" to upstream A and send "cB:b" to downstream C. Also, each DSCCM will send the local query result back to client (step5 and 6). When receiving the query from downstream, DSCCM will send back PROV records whose "waGeneratedBy" records match the query parameter. While receiving the query from upstream, DSCCM will send back PROV records whose "used" records match the query parameter. For instance, after receiving a query parameter “cB:d1” sent by B, C should send back PROV records corresponding with “used (cC:processC, cB:d1, -)”. If A receives “cA:d” from B, PROV records match “wasGeneratedBy (cA:d, cA:processA, tA)” will be sent back. The query will be forwarded to the related DSCCM of data chain recursively until each query stream reaches the boundary of data chain or runs into another query stream. Cloud in Fig.3. represent the recursive process.

Avoid Query Crash. DSCCM will ignore the queries that have already been processed. For instance, if query submitted by upstream comes earlier than queries from downstream, the queries from downstream will be ignored. This mechanism can avoid redundancy query and deadlock between partial nodes.

Merge Query Results. Client will merge and sort results from different nodes and get the final query result in predetermined time. Since results reported from DSCSM nodes concurrently, client needs to distinguish which results belong to the same initial query. Therefore, we have to guarantee that all forwarded queries correspond with the initial query. After merging all query results, we will have an integrated PROV record about data provenance. Then we can construct data supply chain based on the complete provenance information.

5 Experiment

5.1 Experiment Environment

We set up four data platforms, and deploy the data supply chain service module to the real data platforms to verify the effectiveness and feasibility of the solution. Computer processor is Intel (R) Core (TM) 2 Duo CPU 3.00GHZ, and memory is 4.00GB. We use java programming language, MySQL database and J2EE framework includes struts, spring and hibernate. Data platform communicates with each other by web service.

5.2 Case Study

Airline company purchases data from travel company in order to analyze market situation and develop route plan for next year. Travel company collects data from its own travel website, processes and sells to airline company. After the data analysis, some results turn out to be unreasonable. To essure the quality of business decisions, airline company hope to figure out which part in data chain mishandles the data. At this point, it is necessary to construct supply chain and acquire entire data activities through the chain.

We use three data platforms to simulate travel website, travel company, airline company and one data platform to simulate client that send queries. Through the process of distributed DSCSM, we get data supply chain shown in Fig.4.

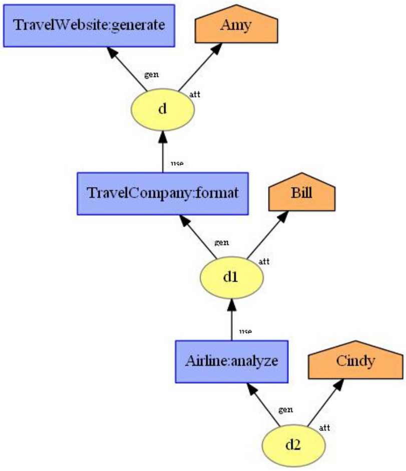

Fig. 4. Graphical data supply chain

As can be seen, data flows through three data platforms, from travel website to travel company and arrives at airline company finally. Data d of travel website is produced by process named generate and the operator is Amy. Bill from travel company uses process format to integrate data d collected from travel company and sell the output data d1 to airline company. Finally Cindy from airline company gets data d2 using process analyze to handle data d1.

Results from experiment above show that DSCSM can construct data supply chain among distributed data platforms automatically and display it in a easily understood way. Thus, the proposed model is feasible and effective for data supply chain construction in the era of big data.

6 Conclusion

We construct a distributed data supply chain service model based on PROV, a data provenance specification presented by W3C. Firstly we introduce PROV data model and describe it using formal language PROV-N. On this basis, we design a data supply chain construct model combined with discovery service of supply chain in EPCGlobal framework and new characteristics of data supply chain. Our model standardizes information records of data activities in corresponding data platforms. And DSCSM built on it can achieve data provenance dynamically from distributed platforms, which provides great convenience for further data minning. In future work, we will focus on the design of more flexible and efficient data supply chain service model to improve the quality of data supply chain construction and adapt to the increasingly complex and diverse information systems.

References

1. Groth, P.: Transparency and reliability in the data supply chain. IEEE Internet Computing **17**(2), 69–71 (2013)
2. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
3. Altiparmak, F., Ferhatosmanoglu, H., Erdal, S., et al.: Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases. IEEE Transactions on Information Technology in Biomedicine **10**(2), 254–263 (2006)
4. Liu, B., Terlecky, P., Bar-Noy, A., et al.: Optimizing information credibility in social swarming applications. IEEE Transactions on Parallel and Distributed Systems **23**(6), 1147–1158 (2012)
5. Goryczka, S., Xiong, L., Fung, B.C.M.: m-Privacy for collaborative data publishing. In: 2011 7th International Conference onCollaborative Computing: Networking, Applications and Worksharing (CollaborateCom), pp. 1–10. IEEE (2011)
6. Shu, Y., Gu, Y.J., Chen, J.: Dynamic authentication with sensory information for the access control systems. IEEE Transactions on Parallel and Distributed Systems **25**(2), 427–436 (2014)
7. Buneman, P., Khanna, S., Tan, W.-C.: Why and where: a characterization of data provenance. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 316–330. Springer, Heidelberg (2000)
8. Green, T.J., Karvounarakis, G., Tannen, V.: Provenance semirings. In: Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 31–40. ACM (2007)
9. Ram, S., Liu, J.: A New Perspective on Semantics of Data Provenance. SWPM, 526 (2009)
10. Uckelmann, D.: Quantifying the value of RFID and the EPCglobal Architecture Framework in Logistics. Springer Science Business Media (2012)
11. Yu, G., Du, X.: Unstructured discovery service method based on extended ONS. In: 2011 International Conference on Internet Technology and Applications (iTAP), pp. 1–4. IEEE (2011)
12. Schuster, E.W, Allen, S.J., Brock, D.L.: Global RFID: the value of the EPC global network for supply chain management. Springer Science Business Media (2007)
13. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Forward-and-backward diffusion processes for adaptive image enhancement and denoising. IEEE Transactions on Image Processing **11**(7), 689–703 (2002)