Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 12, December 2013, pg.272 – 277

RESEARCH ARTICLE

A STUDY ON DATA REPERTORY ACUMEN SCHEMA TO MANAGE DATA PROVENANCE IN GEOSCIENCE APPLICATION

M.HEMALATHA[1], K.S.KANNAN[2]
PG STUDENT, ASSISTANT PROFESSOR
NPR COLLEGE OF ENGINEERING AND TECHNOLOGY, TAMIL NADU, INDIA
EMAIL: *mshema90@gmail.com, saikannan2012@gmail.com*

Abstract: **Data provenance accepts and approves the scientists to model as to investigate the beginning of an unexpected value. It can be used as a duplicate recipe for output data products. The capturing provenance requires enormous effort by scientists in terms of time, training and need to design the workflow of the scientific model i.e., workflow source, which requires both time and training. Scientists may not document any workflow source before the model execution due to lack of time and training. And it is needed to capture provenance while the model is running, i.e., fine-grained data provenance. Explicit documentation of fine-grained provenance is not feasible because of the massive storage consumption by source data in the applications, including those from the geosciences domain where data are continuously arriving and are processed. This work proposed an inference-based framework, which provides both workflow and fine-grained data provenance at a minimal cost in terms of time, training, and disk consumption. The proposed framework is applicable to any given scientific model, which is capable of handling different model dynamics .The variation in the processing time as well as input data products arrival design. The framework used to especially shows the data that are proposed. Proposed framework is relevant and suitable for scientists using geosciences domains for their research work.**

Index terms: Data Provenance; Geo Processing workflow; Geographic information system

1. INTRODUCTION

The word provenance is used synonymously with the word lineage in the database community. It is also sometimes referred to as source attribution or source tagging. Provenance means origin or source. It also means the history of ownership of a valued object or work of art or literature. The knowledge of provenance is especially important for works of art, as it directly determines the value of the artwork. The same applies to digital artifacts or results that are generated by scientific applications. Information about provenance constitutes the proof of correctness of scientific results and in turn, determines the quality and amount of trust one places on the results. For these reasons, the provenance of a scientific result is typically regarded to be as important as the result itself. There are two granularities of provenance considered in literature: workflow provenance and data (or fine-grained) provenance. In what follows, we provide an overview of workflow and data provenance.

Many of these data-intensive e-Science applications are focusing on geoscientific research. In a geoscientific research, scientists collect geospatial data, i.e., measurements or sensor readings with time and space, from different sources. Later, this data are processed to produce the output, i.e., a data product. A framework for managing both workflow and fine-grained data provenance for data-intensive, geoscientific applications is to be developed. Provenance is defined in many different contexts. One of the earlier definitions was given in the context of geographic information system (GIS). In all contexts, provenance can be defined at different levels of granularity. Fine-grained data provenance is defined at the value-level of a data product, which refers to the determination of how that data product has been created and processed starting from its input values. There are three design factors such as the framework would be generic; the framework should be storage-efficient. The framework should be self-adaptable to cope with any given scientific model and the model dynamics, such as processing delay, data arrival pattern etc. An inference-based framework to manage both workflow and fine-grained data provenance for geoscientific applications is proposed. Scientists may not document any workflow provenance before the model execution due to the lack of time and training. Scientists from many domains, such as physical, geological, environmental, biological etc. facilitate data-intensive e-Science applications to study and better understand these complex systems. In these applications, the data collection contains both in-situ data collected from the field and streaming data sent by sensors. Scientists use this data fitting into their model describing processes in the physical world and get the output, which is used to facilitate either a process control application or a decision support system.

The provenance of data products generated by complex transformations such as workflows is considerable value to scientists. From it, one can ensure the quality of the data, based on its ancestral data and derivations track back the sources of errors, allow automated re-enactment of derivations to update a data, and provide attribution of data sources. Provenance is also essential to the business domain, where it can be used to drill down the source of data in a data warehouse, track the creation of intellectual property, and provide an audit trail for regulatory purposes. The growing number and size of computational data resources coupled with uniform access mechanisms provided by a common Grid middleware stack is allowing scientists to perform advanced scientific tasks in collaboratory environments.

Data provenance describes how a particular piece of data has been produced. It is generated once the data is processed. An auditor can obtain it by querying the store where it is recorded. Data provenance plays a vital role in forensic analysis, enabling the collection of digital evidence by a post-incident investigation. It is widely used not only for forensics analysis but also for scientific collaborations and in legal proceedings. Generally, data provenance may include, but not limited to, what action was taken, who took it, where it was taken, why it was taken, how it was taken, when it was taken, in which environment it was taken and what the sequence of those actions.

Provenance is defined in many different contexts. One of the earlier definitions was given in the context of geographic information system (GIS). In GIS, data provenance is known as lineage, which explicates the relationship among events and source data in constructing the data product. In the context of database systems, data provenance provides the description of how a data product is achieved through the transformation activities from its input data. In a scientific workflow, data provenance refers to the derivation history of a data product starting from its origin.

In the context of the geoscientific domain, geospatial data provenance is defined as the processing history of a geospatial data product. Decision makers are now trying to integrate meteorological data with societal data. They usually have to find ways to convert the meteorological data into a compatible Geographic Information System (GIS) format so they can be used in their GIS tools. However, the current GIS tools do not handle the time varying, multidimensional datasets that are often used in meteorological analysis and prediction very well and provide little support for time or spatial animation.

The hydrologic cycle is not a continuous mechanism through which water moves steadily at a constant rate. The movement of water through the cycle is erratic both in time and over area. On occasion nature provides excessive rains floods and other times a little. In adjacent areas the variation in the cycle may be quite different. It is precisely these extremes of flood and drought that are of most interest to the engineering hydrologist, for hydraulic engineering projects are designed to protect against the ill effect of extremes. It must be able to deal quantitatively with the interrelation between factors so they can predict the influence of human activities on these relationships. It must concern themselves with the frequencies with which extremes of the cycle may occur. Although the concept of

the hydrologic cycle is simple, the phenomenon is enormously complex and intricate. It is not just one large cycle but rather is composed of many interrelated cycles of continental, regional and local extend. Although total volume of water in the global hydrologic cycle remains essentially constant, the distribution of this water is continuously changing on continents, in regions and within local drainage basin.

Data provenance is referred to as the lineage of GIS data products. In the Geographic Information - Metadata standard, the lineage of geospatial data is defined as the information about the events or source data used in constructing the data. The events could be a single step of geo-processing or a large aggregated geo-processing. In the context of database system, data provenance is defined as the description of the origins of a piece of data and process by which it arrived in the database. Such provenance for relational views in the database system is referred to as a view data lineage problem, where the origin of the data is associated to the base data items or tables and process is associated to the relational algebra operations that yield the database view data. In the context of scientific workflow, data provenance refers to the derivation history of a data product. Such derivation history includes both the source data and processes or transformations used to derive that data product.

2. LITERATURE SURVEY

2.1. GEOSPATIAL DATA PROVENANCE IN CYBER INFRASTRUCTURE

A new generation of information infrastructure, Cyber infrastructure, is being developed to support the next generation of geoscientific research. An important part of this effort is to serve distributed geospatial resources such as data and analysis functions over the Web. The access to these distributed geospatial a resource is enabled by advanced information technologies. Service technology is a typical choice, since it defines standard interfaces to allow service-to service interactions and integration of multiple services. A set of technologies called geospatial Web services has been widely used in the geospatial domain, including the Open Geospatial Consortium (OGC) standards-compliant services such as Web Feature Service (WFS), Web Map Service (WMS), Web Coverage Service (WCS), Sensor Observation Service (SOS), Catalogue Service for Web (CSW), and Web Processing Service (WPS). As a result, we can see a network of distributed geospatial services, each exposed by different vendors, and each providing original geospatial resources or value-added geospatial data products through geo-processing workflows. Constructing distributed and complex geoprocessing workflows is also a special attention in Cyber infrastructure, since it can help to enhance a data-rich geoscientific research environment to an analysis-rich environment. As more and more scientists use geo-processing workflow or service chains to integrate various distributed data and services to perform geoscientific analyses, it becomes important to capture the provenance, also known as lineage, of a particular derived data product, so that researchers can determine the reliability of the data products, validate and reproduce scientific results in Cyber infrastructure when needed.

The development of Cyber infrastructure in the geospatial context, or called Geo-Cyber infrastructure, opens the challenge for provenance-aware applications. Provenance aware applications provide informed ways to understand the reliability of derived geospatial data products. This paper introduces techniques for provenance-aware applications and current progress related to the geospatial domain. The key considerations discussed in this paper offer a guide to the further exploration of this subject.

2.2 A PROVENANCE FRAMEWORK FOR WEB GEOPROCESSING WORKFLOWS

With the advancement of e-Science or Cyber infrastructure, geoscientific workflows are important activities now to conduct scientific discoveries. To promote the automation of scientific discoveries using a large number of heterogeneous and distributed geospatial data and computational resources, it is crucial for programs to discover and access specific resources using uniform interfaces when developing workflows. Service technologies, in particular Web services technologies, can provide such interfaces and are widely employed in geospatial domain. In a service-oriented geoscientific research environment, where geospatial data and various processing functions are available as interconnected services, individual geospatial services must be chained together as Web geoprocessing workflows to solve a complex geoscientific problem. The development of Web geoprocessing workflows can be divided into three phases: process modeling, which generates an abstract composite process model consisting of the control flow and data flow among process nodes; process model instantiation, where the abstract process is instantiated into a concrete workflow or an executable service chain; and workflow execution, where the chaining result or workflow is executed in the workflow engine to generate value-added data products. The data products derived from the

geoprocessing workflows can take advantages of automation and dynamism from such a procedure. In the first phase, the process models can be generated either manually by domain experts or automatically using Artificial Intelligence (AI) planning methods. The process models can be archived as existing knowledge from modelers and allow knowledge based construction more complex geospatial process models. In the second phase, distributed geoprocessing services are discovered and bound to process models dynamically.

Provenance is crucial in making more effective use of scientific workflows. This paper presents a framework for collecting, organizing, storing, and serving the provenance information generated during the application of Web geoprocessing workflows. A prototype implementation is provided by extending an existing Web geoprocessing system. The work demonstrates how geospatial provenance can be organized in the CSW and incorporated into applications of geoprocessing workflow. The three levels of geospatial provenance support re-orchestration of geoprocessing workflows. Future work will provide a user friendly interface that allows users to browse, navigate, and use provenance easily in Web geoprocessing systems.

2.3. TRACKING AND SKETCHING DISTRIBUTED DATA PROVENANCE

Current provenance collection systems typically gather metadata on remote hosts and submit it to a central server. In contrast, several data-intensive scientific applications require a decentralized architecture in which each host maintains an authoritative local repository of the provenance metadata gathered on that host. The latter approach allows the system to handle the large amounts of metadata generated when auditing occurs at fine granularity, and allows users to retain control over their provenance records. The decentralized architecture, however, increases the complexity of auditing, tracking, and querying distributed provenance. We describe a system for capturing data provenance in distributed applications, and the use of provenance sketches to optimize subsequent data provenance queries. Experiments with data gathered from distributed workflow applications demonstrate the feasibility of a decentralized provenance management system and improvements in the efficiency of provenance queries.

A distributed provenance model introduces challenges as well. The first question that arises is how to track the movement of files so that there is no loss of coupling between the file content and the associated metadata. Even if a file moves and its metadata does not (due to its large size), a user should be able to subsequently retrieve the file's associated metadata. Further, gathering application-agnostic provenance requires fine-grained auditing of processes and network connections. Ideally, the auditing should not require modification of extant user programs. Another challenge is how to efficiently trace distributed provenance. The audited metadata can be viewed as a directed graph data structure. Tracing a path in a directed graph by recursively querying antecedents is known to be a computationally expensive operation .In the case of distributed provenance, it becomes expensive in terms of network operations as well, since part of the provenance metadata is likely to be located remotely.

2.4 PROVENANCE IN DATABASES: PAST, CURRENT, AND FUTURE

The need to understand and manage provenance arises in almost every scientific application. In many cases, information about provenance constitutes the proof of correctness of results that are generated by scientific applications. It also determines the quality and amount of trust one places on the results. For these reasons, the knowledge of provenance of a scientific result is typically regarded to be as important as the result itself. In this paper, we provide an overview of research in provenance in databases and discuss some future research directions.

In the scientific domain, a workflow is typically used to perform complex data processing tasks. A workflow can be thought of as a program which is an interconnection of computation steps and human-machine interaction steps. Workflow provenance refers to the record of the entire history of the derivation of the final output of the workflow. The amount of information recorded for workflow provenance varies. It may include a complete record of the sequence of steps taken in a workflow to arrive at some dataset. In some cases, this entails a detailed record of the versions of soft wares used, as well as the models and makes of hardware equipments used in the workflow.

Most research efforts on data provenance have focused on reasoning about the behavior of provenance and keeping track of annotations or metadata through SQL queries. While SQL queries are fundamental building blocks of many database applications, knowing how to reason about the provenance and flow of data through SQL queries alone is still insufficient for a complete end-to-end tracking of the provenance and flow of data in many database

applications. For example, a Web application that is powered by a database backend may only use SQL queries to retrieve data from (or store data into) the underlying database system. Data that is retrieved may still undergo various transformations (e.g., cleansing or formatting transformations) before they are displayed on a Web page. To make matters worse, many Web applications today (e.g., mashups) are based on other Web applications where information is extracted and integrated though public application programming interfaces and appropriate programming languages. In particular, the process by which information is extracted and integrated is typically not described by SQL queries. Therefore, a major unsolved challenge for data provenance research is to provide a uniform and seamless framework for reasoning about the provenance (and flow) of data through different data transformation paradigms. We list three aspects of research on data provenance next that would make progress towards resolving this challenge.

2.5 SECURE PROVENANCE TRANSMISSION FOR STREAMING DATA

Many application domains, such as real-time financial analysis, e-healthcare systems, sensor networks, are characterized by continuous data streaming from multiple sources and through intermediate processing by multiple aggregators. Keeping track of data provenance in such highly dynamic context is an important requirement, since data provenance is a key factor in assessing data trustworthiness which is crucial for many applications. Provenance management for streaming data requires addressing several challenges, including the assurance of high processing throughput, low bandwidth consumption, storage efficiency and secure transmission. In this paper, we propose a novel approach to securely transmit provenance for streaming data (focusing on sensor network) by embedding provenance into the interpacket timing domain while addressing the above mentioned issues. As provenance is hidden in another host-medium, our solution can be conceptualized as watermarking technique. However, unlike traditional watermarking approaches, we embed provenance over the interpacket delays (IPDs) rather than in the sensor data themselves, hence avoiding the problem of data degradation due to watermarking. Provenance is extracted by the data receiver utilizing an optimal threshold-based mechanism which minimizes the probability of provenance decoding errors. The resiliency of the scheme against outside and inside attackers is established through an extensive security analysis. Experiments show that our technique can recover provenance up to a certain level against perturbations to inter-packet timing characteristics.

The novel problem of securely transmitting provenance for data streams. We propose a spread-spectrum watermarking-based solution that embeds provenance over the interpacket delays. The security features of the scheme make it able to survive against various sensor network or flow watermarking attacks. The experimental results show that our scheme is scalable and extremely resilient in provenance retrieval against various attacks. In future, we will investigate the feasibility of this technique for large sized provenance.

3. CONCLUSION

From this paper we perform literature survey on data repertory acumen schema to efficiently manage data provenance in geosciences application. Scientists understand the importance of provenance data. However, provenance data were rarely maintained due to the lack of time and proper training to use the workflow engines and other tools. Furthermore, an integrated framework to provide both workflow and fine-grained data provenance was much needed due to the increasing popularity of data intensive applications.

REFERENCE

[1] Mohammad Rezwanul Huq, Peter M. G. Apers, and Andreas Wombacher 2013 ”An Inference-Based Framework to Manage Data Provenance in Geoscience Applications” IEEE Transaction on issue:99

[2] Salmin Sultana, Mohamed Shehab, Elisa Bertino 2013”Secure Provenance Transmission for Streaming Data” IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, August 2013

[3] Tanu Malik, Ligia Nistor, Ashish Gehani 2012”Tracking and Sketching Distributed Data Provenance” IEEE sixth international conference

[4] Wang-Chiew Tan 2009 “Provenance in Databases: Past, Current, and Future” Foundations and Trends in Databases Volume 1 Issue 4, April 2009

[5] P. Yue, Z. Sun, J. Gong, L. Di, and X. Lu in Jul. 2011 “A Provenance Framework For Web Geoprocessing Workflows” Proc. IEEE Int. Geosci. Remote Sens. Symp., Jul. 2011, pp. 3811–3814

[6] P.Yue and L.He 2009 ”Geospatial Data Provenance in Cyber infrastructure” Proc. IEEE 17th Int. Conf. Geoinformat., Aug. 2009, pp. 1–4.

