DOI: http://dx.doi.org/10.26483/ijarcs.v10i3.6418

ISSN No. 0976-5697

Volume 10, No. 3, May-June 2019

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

DATA PROVENANCE TECHNIQUES AND SEMANTICS USING W7 MODEL-A REVIEW

Sonia Arora, Dr.Ritika Balhara & Dr.Pooja Sapra
Computer Science & Engineering Department
World College of Technology & Management
Gurugram, Haryana, India

Abstract: Data Provenance means tracing the lineage of the source of data, who performed manipulation on data, how it came to existence, Tracking data provenance helps in ensuring that the data which is being provided by many different providers and sources can be fully trusted and used significantly The area of data provenance is becoming very useful for scientific research work. Although data provenance can be extremely important for different applications, research on data provenance has started and there is still a lot of work that needs to be done. One of the main objective of this paper is to investigate the semantics or meaning of data provenance. In this paper we discuss with some data provenance categories like provenance granularity, data status, provenance computing, semantics of provenance, provenance storage and applications which are proposed to analyze present provenance techniques. We also review a W7 model for managing data for data provenance system.

Keywords: Data Lineage, Data Provenance, Data Warehouse,W7 Model

I. INTRODUCTION

Provenance means something's origin or the beginning of something's existence. Data provenance means the origin of data, the source of data. In today's scenario the actual data which is prepared when circulated does not reach the destination in the same format or quality. This is called data lineage. Data provenance is studying from where the data originated to find the lineage during the data flow. For Example Consider a Truck that contains Drugs loaded in it, The drugs are transported from one place to another. During its way the truck gets hijacked. And some counterfeit drug gets injected into the system. Data provenance is studying from where the counterfeit drug got injected, why it got injected and actions to recover from it.

Data Provenance also has a valuable role in scientific data. With the amount of data increasing in leaps and bounds and sharing of data, questions such as "where did this data originate from?","who other is using this data?" and "why this piece of data is here?" are becoming increasingly common. To ensure that data provided from other sources is trustworthy and used appropriately, it is important that the provenance of the data be recorded and made available to its users. Being a new field research is being going on to study how provenance strategies work and also to reduce the storage cost that is very high in data provenance.

The paper focuses on the meaning of data provenance with some examples on it, categories of provenance techniques and a brief review of W7 model for data management.

II. MEANING AND APPLICATION OF DATA PROVENANCE

[1] Defines provenance as "the history of ownership of a valued object or work of art or literature" [1]. According to [2], knowing the provenance of a work of art is of great importance in that it can "help to determine the authenticity and to establish the historical importance of a work by suggesting other artists who might have seen and been influenced by it, and to determine the legitimacy of current ownership. Some researchers define provenance as the origin of data and the process by which it arrived at the database [3], while others view it as metadata recording the process of experiment workflows, annotations and notes [4]. As suggested by [5], data provenance needs to be captured with the hope that it is comprehensive enough to be useful in the future. The meaning can be best explained with an example. Consider an organization manufacturing steering part of a missile system. In case of a failure the provenance system helps to back track the design of the system that specifies what material was used while manufacturing, why was that material used during manufacturing, what other alternatives were available and hence they are analyzed to check what was the properties of other material in comparison to the material used and hence at the end the best alternative is used for manufacturing process again.

The provenance system is a valuable process that also helps in the prediction of the sales of an organization by backward tracing the workflow cycle to predict the liking probability of a customer towards a product. It helps to improve the sales of an organization.

Provenance system can prove to be a valuable asset for debugging and verification. The error propagation path can be traced to find the actual source of error. Any buggy o stale source data can be removed during the processing. Provenance can also be useful during the auditing phase. The security loop holes can also be tracked very easily using the provenance system.

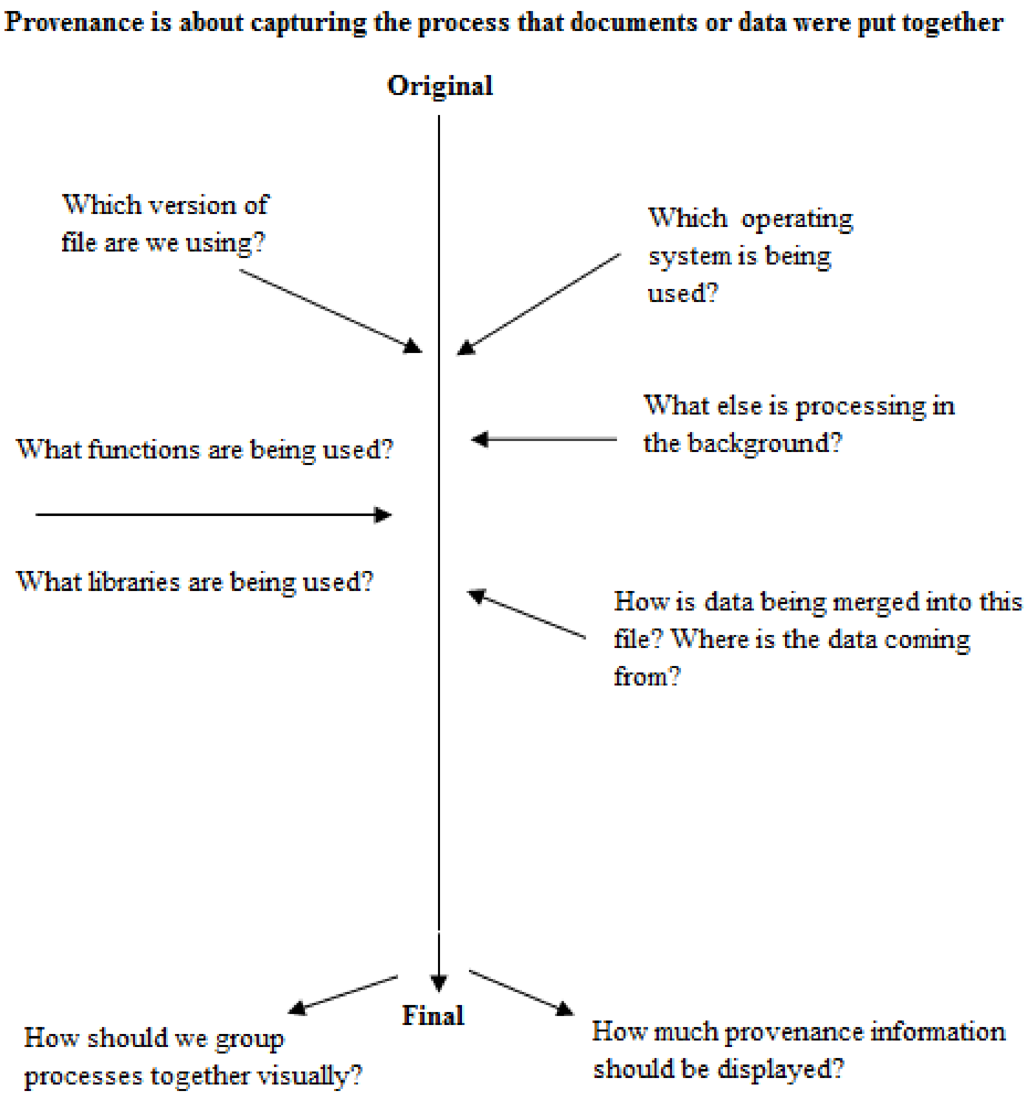
Provenance is about capturing the process that documents or data were put together
Original
Which version of file are we using?
Which operating system is being used?
What else is processing in the background?
What functions are being used?
What libraries are being used?
How is data being merged into this file? Where is the data coming from?
Final
How should we group processes together visually?
How much provenance information should be displayed?

Figure1: Semantics of Provenance

The different application areas of provenance can be categorized as:

1. Data Quality: The reliability and data quality can be estimated by data lineage that is based on the source of the data and how it transforms[6].Proof statements on data derivation can also be provided[9].

2. Audit Trail: Audit trail is another application of the provenance system[10].It also determines the resource used[11] and the errors that propagated during data generation[12].

3. Replication Recipes: Detailing a provenance system can allow repetition of data derived. The currency of the data is maintained [10] and could serve as a recipe for replication [13].

4 Attribution: Ancestry of data can establish the copyright and ownership of data, enable its citation [14], and determine liability in case of error prone data.

5. Informational: The inclusive use of pedigree is to query which is based on lineage metadata used for data discovery.

III. PROVENANCE TECHNIQUES

Provenance techniques can be categorized into seven categories: provenance granularity, data granularity, data status, provenance computing, and semantics of provenance, provenance storage and applications. A brief overview of these is given below:

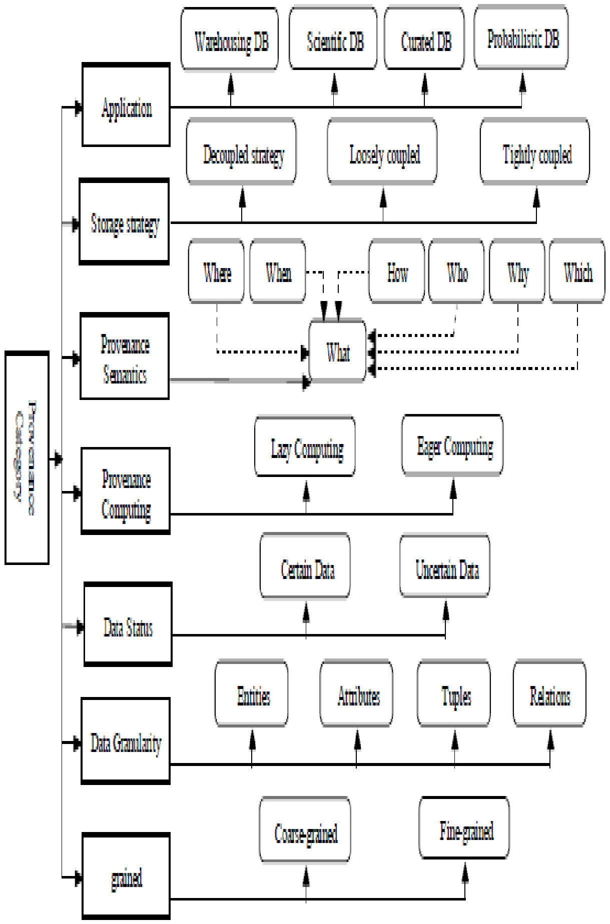

Figure 2: The Provenance Category

1. Provenance Granularity: Granularity a term which is associated with storage of provenance. It is divided into coarse grained or fine-grained provenance.

Coarse grained provenance basically records the complete history (or workflow) of the derivation of some dataset. It combines the pieces of information into one single piece. It records the input/output and transformations of derived data set into one single piece. It cannot however address any change in the database state. The storage cost however remains constant.

Fine Grained provenance stores the origin of each dataset. It provides with a detailed information on how the data is derived and the whole process involved in its transformation. However the storage cost is higher at this level since detailed storage of the dataset transformation is stored.

2. Data Granularity: Data granularity means the data exists in small pieces so that it becomes easier to abstract information from it. In DBMS data is stored in the form of attributes and when all attributes of similar interest are taken together they form a tuple or a record of data or information. The advantage of data granularity is that it can be molded in a way the scientists or researchers want.

3. Status of Data: A significant amount of data which is found in practice is imprecise, incomplete, and not reliable. Hence the data remains uncertain. The level of uncertainty in the data needs to be measured. It also needs to be recorded in order to estimate the confidence level of the results and find the potential sources of error. For some cases the dataset that is transformed needs to be examined in order to trace the errors involved and needs to be labeled accordingly. For this we need to back track to the input so that any further changes can be done to the input which could help in future processing. This is possible if the data is certain. But if the data is incomplete and the output is not labeled properly that is it is uncertain in that case the error sources could also be uncertain. Data provenance is useful in such cases for debugging scientific experiments which involves some sort of sensitive information.

4. Computing Data Provenance: There are two approaches for computing provenance. They are lazy and eager approaches. The eager approach is also known as the bookkeeping or annotation approach. In eager approach the query that is put to the dataset is re-engineered so that anything updated can be added to the output database during the process of transformation so that getting an answer to the provenance becomes simple. The result of it is that the provenance of output data can easily be derived from output database and the updates that are added.

The lazy approach on the other hand is also known as the non-annotation approach. It is a kind of approach that works only when a need arises. In this approach the provenance is calculated only when it is needed. When the need arises the source data, the output data and transformation is examined to compute the provenance. The advantage of this approach is that no space overhead is required and no further updates are necessary that incur cost.

In contrast in eager approach, extra space is required to store output database, transformations and updations and more cost is incurred to store the updates done time to time to the output database.

5. Semantics of Data Provenance: The actual semantics of the data provenance is still unclear and research is being done on what does it mean?, What are data provenance elements and many more questions to reveal its actual meaning. As pointed out by [43], a wide range of possible definitions exists in existing literature. Provenance can include literature references where data were first reported; its history in terms of how it had been stored, curated, and transferred; the series of experimental procedures, computations, and/or database queries through which it had been derived from other data; the sequence of ideas that lead to an experiment; its association with calibration and control data etc. Still at this point research is going on to conclude what does provenance means, what are its components, till what level of detail the provenance information is desired.

6. Provenance Storage: Provenance storage refers to the metadata that is recorded to compute provenance. Storage can be divided into three categories: Decoupled, Loosely

Coupled and Tightly Coupled.

Coupling refers to interdependence among data in the data provenance cycle. It means the degree to which data provenance relies on metadata.

In decoupled strategy of storage the provenance information is stored in separate stockroom which is independent of each other. Any change made to one stockroom would not affect the other stockroom.

In loosely coupled strategy of provenance storage, the data are stored independently in separate depositories, but there exist some mapping schemes which specifies how to track provenance using present data stored in different depositories.

In tightly coupled strategy of provenance storage there is dependence among the data stores and mapping exists to reveal data provenance from resent data. It is the most efficient storage scheme to track data provenance.

7. Database Applications: There are different database areas where how the data is stored and extracted helps to track data provenance. Some of these database applications are:

I. Data Warehouse: A data warehouse is a storehouse of an electronically stored data [17].A data warehouse contains data which is clean and is in standardized form which is been integrated from various sources of information systems built by an organization in a way which fulfills standardized analytical requirements of a system. Figure 3 shows a simple architecture for a data warehouse. Tracing provenance in a data warehouse has several uses, which includes in-depth analysis and mining of patterns relevant to an information system.

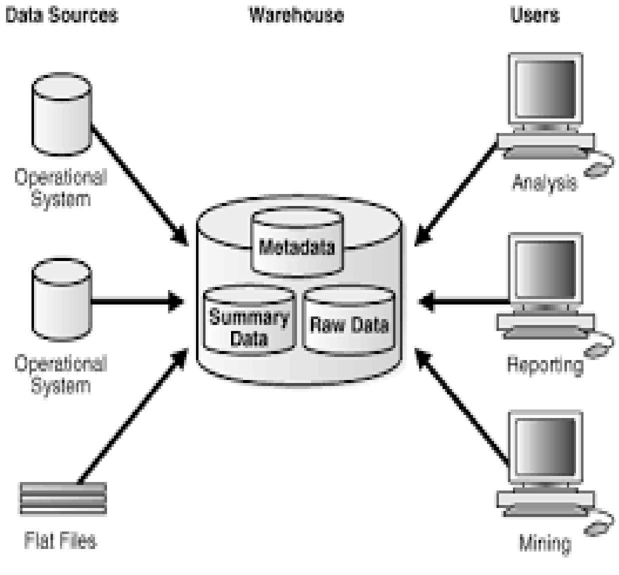

Figure 3. Architecture of Data Warehouse

II. Curated Database: A kind of structured depository such as a traditional database which has been created and updated with human efforts. The best example is the reference works like an encyclopedia, dictionaries newspapers etc. These types of databases contain updated data which can be used extensively at any time .It promotes the provenance process because the updated piece of data being traced by these referential processes.

III. Scientific Database: Scientific databases are the databases that can be represented in various fields such as astronomy, Biology, Biochemistry, Computer Science etc. These type of databases allows for integration of data sets that are dissimilar to each other and analyze them using cross-discipline approach that reveals a new data set. Data Provenance benefits from such databases by analyzing such tightly coupled databases that helps in improving data quality.

IV. Probabilistic Databases. Probabilistic databases are also known as databases with uncertainty. Any relational database can be represented without loss of information by a probabilistic database (Cavallo and Pittarelli, 1987).These databases are used to differentiate between physical data models and how data is represented in them.

IV. OVERVIEW OF W7 MODEL

It investigates the semantics of data provenance based on reference (Ram et al., 2006) ,which represents provenance as a combination of seven elements including ―What, Where, When, How, Who, Which and Why. The W7 model unifies data provenance and provides a clear picture of provenance.

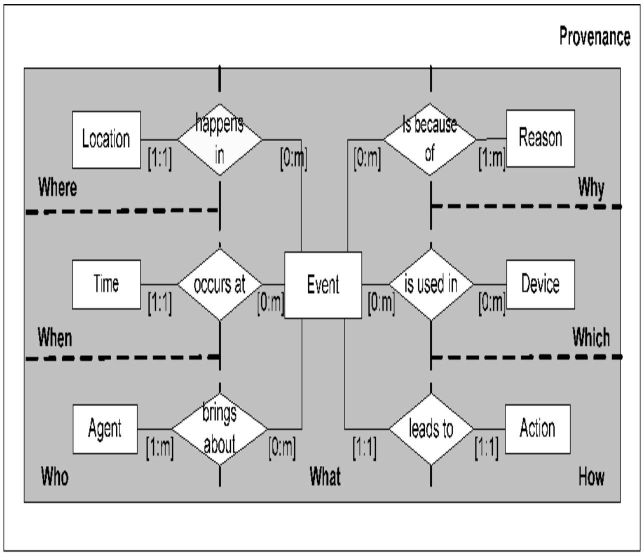

Figure 4. Overview of W7 Model

What is the main character of the W7 provenance model. It defines the actual sequence which the events follow that is affecting any data object during its life span. It also describes any detail regarding the events also.

Where describes the set of locations at which these events took place. The most common form of location representation is physical, geographical and transaction locations.

Physical locations specify where you are. In other words it tells the site or an area within a site where the event took place.

Geographical location refers to the specified location or the boundary of the occurrence of an event.

Transaction locations are the locations where an event takes place within a server or database.

The timestamps that represent various provenance events is the When part of W7 model. The model represents the actual time when the event occurred which affect the data items to be used during the lifecycle. The W7 provenance model also keeps a record of the total duration when the event took place from start point to end point. These details helps to keep a record of the history of data items.

How keeps a track of the relevant processing steps that lead to the occurrence of the event? These processing steps which are called as actions are performed so that a desired output could be achieved. Thus actions are the causes of events and vice-versa. The processing steps include the input, output, preconditions, sources and ways how they are performed.

Who designates the one who causes the occurrence of events; it can be any person associated, particular software or an organization.

Why defines any type of reasons for causing of various events in provenance. Beliefs and goals are its two sub-parts. Belief is what you think about an event that means the knowledge related to the event and goals are the ultimate result that needs to be achieved.

V. CONCLUSION

This paper gives a brief introduction to what data Provenance is. It describes the main application areas of Data provenance and the techniques used for data Provenance. At the end a survey of W7 model used for data management in data provenance is discussed. Future works are being done on this model to deeply explore the different characteristics of W7 model to improve the data quality used in the provenance system.

VI. REFERENCES

[1] Merriam-Webster, "Merriam-Webster Online. http://www.m-w.com/home.htm.

[2] H. U. A. Museums., "Provenance Research Project. http://www.artmuseums.harvard.edu/research/provenance/index.html

[3] Welsh, M. Jirotka, and D. Gavaghan, "Post-genomic science: cross-disciplinary and large-scale collaborative research and its organizational and technological challenges for the scientific research process," Philosophical Transactions: Mathematical, Physical and Engineering Sciences, vol. 364.

[4] P. Buneman, S. Khanna, and W. C. Tan, "Data Provenance: Some Basic Issues,"presented at FSTTCS, New Delhi, India, 2000.

[5] J. Frew and R. Bose, "Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products," presented at the 13th International Conference on Scientific and Statistical Database Management, Fairfax, VA,2001.

[6] H. V. Jagadish and F. Olken, "Database Management for Life Sciences Research," in SIGMOD Record,vol. 33, 2004, pp. 15-20.

[7]]Buneman P., Chapman v and Cheney J., ―Provenance management in curated databases. In SIGMOD, pp.539--550, 2006.

[8] CHENEY, J., CHITICARIU, L., AND TAN, W.-C. Provenance in databases: Why, how, and Where. Foundations and Trends in Databases 1, 4 (2009), 379–474..

[9] P. P. da~Silva, D. L. McGuinness, and R. McCool, "Knowledge Provenance Infrastructure," in IEEE Data Engineering Bulletin, vol. 26, 2003, pp. 26-32.

[10] S. Miles, P. Groth, M. Branco, and L. Moreau, "The requirements of recording and using provenance in e-Science experiments," in Technical Report, Electronics and Computer Science, University of Southampton,2005.

[11] M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn, "Provenance of e-Science Experiments - experience from Bioinformatics," in Proceedings of the UK OST e- Science second All Hands Meeting, 2003.

[12] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita, "Improving Data Cleaning Quality Using a Data Lineage Facility," in DMDW, 2001, pp. 3.

[13] I. T. Foster, J.-S. Vöckler, M. Wilde, and Y. Zhao, "The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration," in CIDR, 2003.

[14] H. V. Jagadish and F. Olken, "Database Management for Life Sciences Research," in SIGMOD Record,vol. 33, 2004, pp. 15-20.

[15] https://docs.oracle.com/database/121/DWHSG/concept.htm #DWHSG8070

[16] Yogesh L. Simmhan, Beth Plale, Dennis Gannon," A Survey of Data Provenance Techniques", Technical Report IUB-CS-TR618.

[17] Tan W. C., ―Provenance in Databases: Past, Current, and Future‖, In IEEE Data Eng. Bull. vol.30, no. 4, pp. 3-12, 2007.

[18] Ikeda R. and Widom J., ―Panda: A System for Provenance and Data, Technical Report. Stanford InfoLab,2009.

[19] Ram S., Liu J., and George R. T., ―PROMS: A system for harvesting and managing data provenance‖, In WITS, 2006. url: http://kartik.eller.arizona.edu/WITS_DEMO_final.pdf

[20] Y. Cui, J. Widom, and J. L. Wiener, "Tracing the Lineage of View Data in a Data Warehousing Environment," in Technical Report:University of California, 1997.