A Primer on Database Provenance

Boris Glavic

IIT DB Group Technical Report
IIT/CS-DB-2014-01

2014-09

http://www.cs.iit.edu/~dbgroup/

A Primer on Database Provenance

A Beginner's Introduction to Provenance and Its Use Cases

Boris Glavic

September 22, 2014

This paper provides an introduction to the concept of data provenance. Furthermore, we discuss which types of applications can benefit from provenance management and illustrate use cases for provenance using a set of easy to understand example. The target audience of this paper are professionals with database background that are new to provenance.

1 What is Data Provenance?

In computer science, the term **data provenance** denotes information about the creation process and origin of data. Some work uses the terms lineage or data pedigree to describe the same concept. In this paper we will stick to data provenance or just provenance, because it is the most broadly accepted term. Provenance has been studied in the context of databases, software engineering, scientific workflows, data integration, functional programming languages, and distributed systems. In this paper we focus on database provenance. Provenance can be used to answer questions such as

- "Who did create this piece of data?"
- "This analysis result looks suspicious. How can I find errors in the input data that lead to the suspicious result?"
- "Based on which input data did we compute this result of the analysis?"
- "I know which data in my database to trust. Can I trust this query result?"
- "My employees should have access to all documents that are based on documents they have written. How can I enforce this access control rule?"
- "User Bob's account has been compromised. Which data in my database depends on data created or modified by Bob's account?"

Before discussing the different types of information that are considered to be data provenance, we first present an illustrative example.

Example 1. *Consider a database table storing account information (account number, branch, account holder, and balance) for a hypothetical bank X. Clerk Bob at branch Y has helped a customer Alice to make a withdrawal of $10,000. However, Bob has accidentally subtracted $10,000,000 from Alice's account instead of the correct amount of $10,000. Later on, bank manager Peter computes the total balance of all accounts per branch. Peter sees a negative balance for branch Y and now wants to investigate what lead to this negative balance. Was it an error, the result of an unusually high withdrawal, or caused by a large number of small withdrawals? Provenance information can help Peter to find the answer to his question. Tracing the provenance of the result for branch Y, Peter will realize that one of the account table rows used to compute this result is Alice's account. Furthermore, provenance will reveal that this row was last updated by Bob to a large negative value.*

1.1 Types of Provenance Information

The term provenance has been used to denote different types of information. On an abstract level any type of analysis or modification of data can be viewed as applying a transformation to a collection of existing data items to creates a collection of new data items. For example, an SQL query is a transformation that takes as input one of more database tables and returns a result table. Similarly, an update is a transformation that takes the current version of a database table and returns an updated version of this table. We introduce this abstract view of data transformations, because it allows us to classify different types of provenance information in the abstract view and illustrate the relationship to provenance for more specific types of data transformations. We distinguish between three types of provenance information for a data item d.

- **Data**: This type of provenance models which data was used by some transformation to derive data item d. For instance, in the banking example shown above, the fact that Alice's account was used to compute the total balance for all accounts of branch Y is provenance of this type.

- **Transformation**: This type of provenance models which transformations were involved in deriving a data item d. For instance, in the banking example, the query run by Peter was responsible for computing the total balance for branch Y.

- **Agents, Auxiliary, and Environment**: This type of provenance models any type of information that does not fall into the first two categories such as the user that executed a particular transformation, the environment of a transformation (e.g., the machine it was executed on), and so on.

1.2 Isn't Provenance Just ...?

When introducing the concept of data provenance to new users, we often get questions like "Isn't this just X?" where typically X is *temporal databases* or *audit logging*. We now provide answers to these questions and in the meanwhile discuss how provenance is related to these concepts.

Temporal databases: Temporal databases enable access to past versions of tables in a database. Many database systems use unique identifiers for tuples that are not affected by update operations. Thus, such identifiers can be used to track the derivation history of a certain row t through different versions of the database. This can be considered as a type of data provenance information for this row t. However, provenance provides additional information that is not available in temporal databases: 1) Temporal databases can be used to track back different versions of the same row through update operations, but do not provide any help with tracking the provenance of queries. For instance, in the banking example, using a temporal database Peter can unearth that the current version of the row representing Alice's account was produced from a previous version of this row. However, this information is meaningless, unless Peter has first discovered that this row is in the provenance of the row representing the total balance for branch Y produced by his query. 2) Temporal databases provide no record of how new rows have been created (e.g., by combining the information from several existing rows). For example, assume that Peter has inserted the results of his query into a new table. A temporal database has no information on how the rows in the new table have been produced. 3) Temporal databases only provide access to committed versions of rows. Intermediate row versions created by some update in a transactions are discarded. However, for use cases such as auditing, access to these intermediate results can be crucial.

Audit logging: In audit logging, the database keeps track of which SQL statements were executed at which time (and potentially stores additional information such as the user which executed the statement). An audit log can be seen as a type of transformation provenance. The difference between transformation provenance and audit logging is, that an audit log does not relate the transformation to data whereas transformation provenance records which transformations were used to derive a certain data item. To some extend this disadvantage can be overcome by combining audit logging with temporal databases. For example, Peter could use an audit log to figure out that Bob has updated Alice's account and then use a temporal database to retrieve the old version of the row representing Alice's account. However, both the audit log and a temporal database do not help Peter to identify that Alice's account was used to compute the total balance of branch Y in his query.

2 Use Cases for Provenance

We now discuss how provenance information can be used in different application domains.

2.1 Auditing and Compliance

Many companies have to fulfill strict auditing and compliance requirements, e.g., because of government regulations. For instance, a company may have to be able to prove for each data item in the database how it was produced, from which data, and by whom. Provenance provides exactly this type of information. Auditing requires provenance for queries as well as for transactional updates.

2.2 Data Debugging

Provenance information can be used to trace erroneous or suspicious outputs of a transformation back to the relevant inputs from the transformation, because it relates the output to all inputs that were used to produce it. In our running example, Peter was applying this method to trace the balance for branch Y back to the inputs from the accounts table that were used to compute this result.

2.3 Access Control

Provenance can be used to implement new types of access control such as allowing users to access data if it was derived from data they have created. For example, under this rule if Peter has created a document and Bob has used part of Peter's document in a new document, then Peter would be granted access to the new document. Provenance based access control enables new types of access control rules that are not expressible using traditional approaches.

2.4 View Maintainance and Updates

Provenance has been used to update views based on changes to base data (this is usually called *view maintenance* in database research) and to translate updates to views into updates to base data (this is the well-known *view update* problem). Similarly, provenance is used to compute answers to "What-if"-queries (how would my query results change if I change the inputs in a certain way) and "How-to"-queries (how to achieve a certain change to the result of query by modifying its input data).

2.5 Data Quality and Trust

Assessing the quality of data and trust in data is a complex problem which cannot be solved though provenance alone. However, if quality or trust information is available for data in a database, then provenance can be used to compute quality and trust measure for the results of transformations. For example, consider a database storing traveling routes between cities. For each route we store the origin, destination, travel distance, and whether this information is trusted or not. From this database, we can compute indirect routes between cities by combining multiple direct routes. However, it is unclear how to determine whether we should trust a certain indirect route. Provenance provides a well-founded foundation for computing such trust measures.

3 A few Pointers to Research

How to model and compute the provenance of queries is relatively well understood [9, 13, 14, 26, 32, 22, 3, 36, 10, 38, 39, 44, 2, 24, 35, 31, 7, 33, 24, 52]. Several papers have also addressed this problem for update operations [33, 8, 51, 6]. However, Arab et al. [5, 4] are the first to provide a solution for computing the provenance of transactions using temporal databases, audit logging, and query rewrite techniques. Provenance-based access control has been discussed in [45, 41, 42, 43]. Using provenance for computing quality and trust measure has been studied intensively [12, 20, 15, 16, 37, 1, 53, 6, 48, 26, 32, 30, 27, 25, 29, 34, 33]. Examples for approaches that use provenance for What-if and How-to queries are Deutch et al. [19] and Meliou et al. [40]. Note that this list is by no means complete. The interested reader can find additional literature references in one of the many surveys on data provenance (e.g., [49, 46, 47, 23, 50, 17, 18, 21, 28, 13, 11]).

4 Conclusions

We have given an overview of the emerging concept of data provenance, have discussed several important use cases for provenance, and highlighted some recent research findings in this area that are likely to be the foundations of practical implementations of provenance management.

References

[1] M. D. Allen, A. Chapman, L. Seligman, and B. Blaustein. Provenance for collaboration: Detecting suspicious behaviors and assessing trust in information. In *CollaborateCom*, pages 342–351, 2011.

[2] Y. Amsterdamer, D. Deutch, T. Milo, and V. Tannen. On provenance minimization. In *PODS*, pages 141–152, 2011.

[3] Y. Amsterdamer, D. Deutch, and V. Tannen. Provenance for Aggregate Queries. In *PODS*, pages 153–164, 2011.

[4] B. Arab, D. Gawlick, V. Krishnaswamy, V. Radhakrishnan, and B. Glavic. Reenacting transactions to compute their provenance. Technical report, Illinois Institute of Technology, 2014.

[5] B. Arab, D. Gawlick, V. Radhakrishnan, H. Guo, and B. Glavic. A generic provenance middleware for database queries, updates, and transactions. In *TaPP*, 2014.

[6] D. W. Archer, L. M. Delcambre, and D. Maier. User trust and judgments in a curated database with explicit provenance. In *In Search of Elegance in the Theory and Practice of Computation*, pages 89–111. Springer, 2013.

[7] D. Bhagwat, L. Chiticariu, W.-C. Tan, and G. Vijayvargiya. An Annotation Management System for Relational Databases. *VLDB Journal*, 14(4):373–396, 2005.

[8] P. Buneman, J. Cheney, and S. Vansummeren. On the Expressiveness of Implicit Provenance in Query and Update Languages. *TODS*, 33(4):1–47, 2008.

[9] P. Buneman, S. Khanna, and W.-C. Tan. Why and Where: A Characterization of Data Provenance. In *ICDT*, pages 316–330, 2001.

[10] P. Buneman, E. V. Kostylev, and S. Vansummeren. Annotations are relative. In *ICDT*, pages 177–188, 2013.

[11] L. Carata, S. Akoush, N. Balakrishnan, T. Bytheway, R. Sohan, M. Selter, and A. Hopper. A primer on provenance. *Communications of the ACM*, 57(5):52–60, 2014.

[12] J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *WWW*, page 622, 2005.

[13] J. Cheney, L. Chiticariu, and W.-C. Tan. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4):379–474, 2009.

[14] Y. Cui, J. Widom, and J. L. Wiener. Tracing the Lineage of View Data in a Warehousing Environment. *TODS*, 25(2):179–227, 2000.

[15] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. An Approach to Evaluate Data Trustworthiness Based on Data Provenance. In *VLDB SDM workshop*, pages 82–98, 2008.

[16] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. Trust evaluation of data provenance. Technical report, CERIAS, 2008.

[17] S. B. Davidson, S. Cohen-Boulakia, A. Eyal, B. Ludäscher, T. McPhillips, S. Bowers, and J. Freire. Provenance in Scientific Workflow Systems. *IEEE Data Engineering Bulletin*, 32(4):44–50, 2007.

[18] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *SIGMOD*, pages 1345–1350. ACM, 2008.

[19] D. Deutch, Z. Ives, T. Milo, and V. Tannen. Caravan: Provisioning for what-if analysis. *CIDR*, 2013.

[20] L. Ding, P. Kolari, T. Finin, A. Joshi, Y. Peng, and Y. Yesha. On homeland security and the Semantic Web: A provenance and trust aware inference framework. In *Proceedings of the AAAI Spring Symposium on AI Technologies for Homeland Security*, 2005.

[21] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.

[22] F. Geerts and A. Poggi. On database query languages for K-relations. *Journal of Applied Logic*, 8(2):173–185, 2010.

[23] B. Glavic and K. R. Dittrich. Data Provenance: A Categorization of Existing Approaches. In *BTW*, pages 227–241, 2007.

[24] B. Glavic, R. J. Miller, and G. Alonso. Using sql for efficient generation and querying of provenance information. In *In search of elegance in the theory and practice of computation*, pages 291–320. Springer, 2013.

[25] T. J. Green, G. Karvounarakis, Z. G. Ives, and V. Tannen. Update Exchange with Mappings and Provenance. In *VLDB*, pages 675–686, 2007.

[26] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance Semirings. In *PODS*, pages 31–40, 2007.

[27] T. J. Green, G. Karvounarakis, N. E. Taylor, O. Biton, Z. G. Ives, and V. Tannen. ORCHESTRA: Facilitating Collaborative Data Sharing. In *SIGMOD*, 2007.

[28] R. Ikeda and J. Widom. Data Lineage: A Survey. Technical report, Stanford University, 2009.

[29] Z. G. Ives, T. J. Green, G. Karvounarakis, N. E. Taylor, V. Tannen, P. P. Talukdar, M. Jacob, and F. Pereira. The ORCHESTRA Collaborative Data Sharing System. *SIGMOD Record*, 37(2):26–32, 2008.

[30] Z. G. Ives, N. Khandelwal, A. Kapur, and M. Cakir. ORCHESTRA: Rapid, Collaborative Sharing of Dynamic Data. In *CIDR*, 2005.

[31] G. Karvounarakis. *Provenance in collaborative data sharing.* PhD thesis, University of Pennsylvania, 2009.

[32] G. Karvounarakis and T. Green. Semiring-annotated data: Queries and provenance. *SIGMOD Record*, 41(3):5–14, 2012.

[33] G. Karvounarakis, T. J. Green, Z. G. Ives, and V. Tannen. Collaborative data sharing via update exchange and provenance. *TODS*, 38(3):19, 2013.

[34] G. Karvounarakis, Z. Ives, and V. Tannen. Querying data provenance. In *SIGMOD*, pages 951–962, 2010.

[35] S. Köhler, B. Ludäscher, and D. Zinn. First-order provenance games. In *In Search of Elegance in the Theory and Practice of Computation*, pages 382–399. Springer, 2013.

[36] E. V. Kostylev and P. Buneman. Combining dependent annotations for relational algebra. In *ICDT*, pages 196–207, 2012.

[37] J. Lyle, A. Martin, et al. Trusted computing and provenance: Better together. In *TaPP*, 2010.

[38] A. Meliou, W. Gatterbauer, J. Halpern, C. Koch, K. Moore, and D. Suciu. Causality in databases. *IEEE Data Engineering Bulletin*, 2010.

[39] A. Meliou, W. Gatterbauer, K. Moore, and D. Suciu. The Complexity of Causality and Responsibility for Query Answers and non-Answers. *PVLDB*, 4(1):34–45, 2010.

[40] A. Meliou and D. Suciu. Tiresias: The database oracle for how-to queries. In *SIGMOD*, pages 337–348, 2012.

[41] D. Nguyen, J. Park, and R. Sandhu. Dependency path patterns as the foundation of access control in provenance-aware systems. In *TaPP*, 2012.

[42] D. Nguyen, J. Park, and R. Sandhu. A provenance-based access control model for dynamic separation of duties. In *PST*, pages 247–256, 2013.

[43] D. Nguyen, J. Park, and R. Sandhu. Adopting provenance-based access control in openstack cloud iaas. In *NSS*, 2014.

[44] D. Olteanu and J. Závodný. On factorisation of provenance polynomials. In *TaPP*, 2011.

[45] J. Park, D. Nguyen, and R. Sandhu. A provenance-based access control model. In *PST*, pages 137–144, 2012.

[46] Y. L. Simmhan, B. Plale, and D. Gannon. A Survey of Data Provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005.

[47] Y. L. Simmhan, B. Plale, and D. Gannon. A Survey of Data Provenance Techniques. Technical report, Indiana University, 2005.

[48] M. Stamatogiannakis, P. Groth, and H. Bos. Facilitating trust on data through provenance. In *Trust and Trustworthy Computing*, pages 220–221. Springer, 2014.

[49] W.-C. Tan. Research Problems in Data Provenance. *IEEE Data Engineering Bulletin*, 27(4):42–52, 2004.

[50] W.-C. Tan. Provenance in Databases: Past, Current, and Future. *IEEE Data Engineering Bulletin*, 30(4):3–12, 2007.

[51] S. Vansummeren and J. Cheney. Recording Provenance for SQL Queries and Updates. *IEEE Data Engineering Bulletin*, 30(4):29–37, 2007.

[52] J. Zhang and H. Jagadish. Lost source provenance. In *EDBT*, pages 311–322, 2010.

[53] O. Zhang, R. Ko, M. Kirchberg, C. Suen, P. Jagadpramana, and B. Lee. How to track your data: Rule-based data provenance tracing algorithms. In *TrustCom*, pages 1429–1437, 2012.