Big Data Provenance: State-Of-The-Art Analysis and Emerging Research Challenges

Alfredo Cuzzocrea
DIA Department
University of Trieste and ICAR-CNR
Italy
alfredo.cuzzocrea@dia.units.it

ABSTRACT

This paper focuses the attention on *big data provenance issues*, and provides a comprehensive survey on state-of-the-art analysis and emerging research challenges in this scientific field. Big data provenance is actually one of the most relevant problem in big data research, as confirmed by the great deal of attention devoted to this topic by larger and larger database and data mining research communities. This contribution aims at representing a milestone in the exciting big data provenance research road.

CCS Concepts

•**Theory of computation** → **Data provenance;**

Keywords

Big Data Provenance, Privacy of Big Data, Big Data Lineage, Big Data Derivation

1. INTRODUCTION

In big data research, *privacy and security of big data* (e.g., [13, 14, 12]) play a major role. Along with these topics, *provenance of big data* (e.g., [16, 17, 10, 18, 4]) is relevant as well. Data provenance concerns with the problem of detecting the origin, the creation and the propagation process of data within a data-intensive system. In other words, data provenance consists in the *lineage* (e.g., [27]) and *derivation* (e.g., [22]) of data and data objects, and it puts its conceptual roots in extensively studies performed in the past in the contexts of arts, literary works, manuscripts, sculptures, and so forth. Another concept that is close to the "*data provenance*" one is represented by the so-called *ownership of data* (e.g., [21]), which refers to the issue of defining and providing information about the rightful owner of data assets, and to the acquisition, use and distribution policy implemented by the data owner. This way, data ownership primarily shapes itself like a *data governance process* that details an organization's legal ownership of enterprise-wide data.

When applied to big data, provenance problems become prohibitive (e.g., [10]), mostly due to the enormous size of big data. For instance, one of the most successful data provenance techniques consists in the so-called *annotation-based approaches* (e.g., [22]) that propose modifying the input database queries in order to support data provenance tasks, while being able to access *all* the target data set. Obviously, the latter requirement becomes very hard when applied to big data repositories. Many others research challenges and open issues still arise in big data provenance research. For instance, advanced concepts like *confidentiality* of the data provenance process, *secure and privacy-preserving* big data provenance, *flexible big data provenance query tools*, and so forth, still need to be deeply investigated.

Inspired by these considerations, in this paper we provide an overview of relevant research issues and challenges of the above-introduced big data provenance problems, by also highlighting possible future efforts within these research directions.

The remaining part of this paper is organized as follows. Section 2 contains a comprehensive analysis of state-of-the-art proposals that focus on big data provenance issues. In Section 3, we recognize and report on emerging challenges in big data provenance research, by highlighting possible promising directions. Finally, Section 4 draws the conclusions of our research.

2. STATE-OF-THE-ART ANALYSIS

Data provenance is relevant for a wide spectrum of typical enterprise data tasks, such as: (*i*) data validation (e.g., [7]); (*ii*) data debugging (e.g., [20]); (*iii*) data auditing (e.g., [26]); (*iv*) data quality (e.g., [24]); (*v*) data reliability (e.g., [3]). Application-wise, the provenance problem has been typically addressed in database management systems (e.g., [9]), but several efforts even arise in the contexts of *workflow management systems* (e.g., [15]) and *distributed systems* (e.g., [25]).

As regards the proper research side, there are several research initiatives that composes the state-of-the-art. Here, we review some of them.

[11] describes a framework for modeling and capturing provenance in *MapReduce jobs* and deriving *MapReduce tasks*, called *Kepler*. The approach is distributed in nature, and it exploits the *MySQL Cluster* distributed database system [2].

[19, 23] proposed an extension of *Hadoop* [1] called *Reduce and Map Provenance* (RAMP). It introduces a wrapper-based method that can be easily deployed on top of Hadoop yet resulting transparent to it. Tracing of data-intensive processes is supported as well.

[5] describes an extension of Hadoop for implementing provenance detection in MapReduce jobs, called *Hadoop-Prov.* The goal of HadoopProv is to minimize overheads introduced by computing provenance, which is usually a resource-consuming task. The proposed system provides flexible tools for querying the so-built big data provenance graph.

Pig Lipstick [6] is a kind of *hybrid big data provenance system* that combines the management of *fine-grained dependencies*, which are typical of database-oriented provenance systems, with the management of *workflow-grained dependencies*, which are typical of workflow-oriented provenance systems. The internal model for reasoning on big data provenance is graph-like in nature.

[4] proposes anatomy and functionalities of a *layer-based architecture* for supporting big data provenance. In particular, the architecture is composite in nature and it focuses on the provenance collection, querying and visualization of provenance in the specialized context of scientific applications.

[17] considers the problem of managing fine-grained provenance in *Data Stream Management Systems* (DSMS). Indeed, this problem is recognized as particularly hard due to the fact of the need of supporting *flexible analysis tools over the so-computed provenance*, such as revision processing or query debugging. With this goal in mind, the paper proposes a novel big data provenance framework based on the concept of *operator instrumentation.* It consists in modifying the behavior of operators in order to generate and propagate fine-grained provenance through several operators of a query.

CloudProv, a framework for integrating, modeling and monitoring data provenance in *Cloud environments*, is presented in [18]. The proposed framework is based on a method that allows us to model collected provenance information as to continuously acquire and monitor such information for *real-time applications*, according to a *service-oriented paradigm.*

Finally, *Oruta*, an innovative *privacy-preserving public auditing mechanism for supporting data sharing in untrusted Cloud environments* is proposed in [26]. The proposed mechanism makes use of *homomorphism authenticators* [8] that allows the third party auditor to check the integrity of shared data from a given user group, yet not superimposing the need for accessing *all* data.

3. EMERGING RESEARCH CHALLENGES

A relevant number of issues and challenges in big data provenance research arise. In the following, we will introduce and discuss some noticeable ones.

Accessing Big Data Big data are prominently enormous-in-size, hence accessing the entire big data set become problematic. Accessing data is a strict requirement for data provenance techniques, hence this makes applying classical methods not suitable to the particular context of dealing with big data provenance.

Analyzing Big Data In order to apply data provenance methods, state-of-the-art techniques require to analyze the target (big) data set. Here, a major problem is represented by the *scalability* of big data, which can be really explosive.

Scalability Issues When dealing with big data, one of the most problematic drawbacks is represented by scalability, as highlighted before. This again occurs with provenance of big data, as provenance techniques are *multi-step in nature* and they need to access and process target data repetitively. This poses relevant issues, as big data are typically growing-in-size and large-scale.

Information Sharing Data provenance methods very often require the need for *sharing information* among the actors that perform the same data provenance task. The latter is not easy when dealing with big data, as such data are typically distributed over *large-scale network environments*, hence information sharing introduces relevant research challenges as well as technological drawbacks.

Minimum Computational Overhead Requirement Data provenance techniques may be data-intensive and resource-consuming. This imposes the need for devising and implementing techniques that introduce a *minimum computational overhead*, in order to avoid impacting on the performance of the target system, e.g. workflow management systems.

Query Optimization Issues Data provenance techniques need to access and query data in order to determine their provenance, even in an *interactive manner.* This applicative requirement introduces severe drawbacks when these techniques run over big data, as querying big data is a crucial open problem at now.

Transformation Issues During data provenance tasks, data sources need to be *transformed* among different data formats. Tracing provenance must be introduced accordingly, in order to track all the different transformations occurred. This topic is a first-class one in the family of big data provenance research issues, which also has several points in common with the *data exchange research area.*

When Computing Provenance? There exist two alternatives for computing provenance. One predicates to compute provenance only when the same provenance is required (this is called *lazy provenance model*). The other one argues to compute provenance every time data are transformed (this is called *eagerly provenance model*). Both models have pros and cons. They also imply different computational overheads. This one is still an open problem to be considered in future efforts.

Data Modeling Support for Provenance When data sources are processed to detect their provenance, several transformations must be applied, as mentioned above. This also implies the need of devising ad-hoc *data models for supporting provenance*, as data sources may be significantly different. In this case, *semantic techniques* seem promising to this direction.

Heterogeneity of Data Source Models Data provenance techniques usually run over *heterogeneous data sources* hence they need to cope with *heterogeneous data models* as well. Therefore, heterogeneity of data sources is a big challenge for such techniques, as data sources expose different formats, (data) types, and schema.

User Annotation Support for Provenance The data provenance process is usually enriched by *user annotation*, according to which domain experts are devoted to annotate data in order to enhance the effectiveness of this process. As a consequence, data provenance tools need to introduce *ad-hoc software modules* capable of supporting user annotation over big data.

Secure and Privacy-Preserving Provenance Provenance can represent a *security and privacy breach* for target data sources. Therefore, a relevant issue for future efforts is represented by the need for secure and privacy-preserving big data provenance techniques. Possible solutions are those based on accepting a sort of *compromise* among security and privacy of data sources from a side, and provenance of data sources from the other side.

Flexible Provenance Query Tools Provenance needs to be used not only to detect the lineage and the derivation of data and data objects, but also in the vest of enabling methodology for flexible query tools focused to support next-generation *cybersecurity systems* where users may be interested in tracking records generated by a particular person in a specific research lab, or detecting the confidentiality of tracked records, i.e. understanding who may have looked these tracked records beyond to authorized people.

Provenance Visualization Tools *Visualization tools* are extremely important for big data provenance techniques, as the provenance one is an *interactive* process that typically requires intelligent tools for visualizing actual results and supporting next-step decisions. This will be a relevant research challenge in future years.

4. CONCLUSIONS

This paper has provided a comprehensive survey on state-of-the-art analysis and emerging research challenges in the context of big data provenance research. We have highlighted benefits and limitations of most relevant proposals, and we have described possible research directions in the exciting big data provenance research road.

5. REFERENCES

[1] Apache Hadoop. http://wiki.apache.org/hadoop. Accessed: 2015-01-15.

[2] MySQL Cluster CGE. https://www.mysql.com/products/cluster/. Accessed: 2015-01-15.

[3] N. Agmon and N. Ahituv. Assessing data reliability in an information system. *J. of Management Information Systems*, 4(2):34–44, 1987.

[4] R. Agrawal, A. Imran, C. Seay, and J. Walker. A layer based architecture for provenance in big data. In *2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27-30, 2014*, pages 1–7, 2014.

[5] S. Akoush, R. Sohan, and A. Hopper. Hadoopprov: Towards provenance as a first class citizen in mapreduce. In *5th Workshop on the Theory and Practice of Provenance, TaPP'13, Lombard, IL, USA, April 2-3, 2013*, 2013.

[6] Y. Amsterdamer, S. B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen. Putting lipstick on pig: Enabling database-style workflow provenance. *PVLDB*, 5(4):346–357, 2011.

[7] A. Assaf, A. Senart, and R. Troncy. Roomba: Automatic validation, correction and generation of dataset metadata. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 159–162, 2015.

[8] G. Ateniese, R. C. Burns, R. Curtmola, J. Herring, L. Kissner, Z. N. J. Peterson, and D. X. Song. Provable data possession at untrusted stores. In *Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007, Alexandria, Virginia, USA, October 28-31, 2007*, pages 598–609, 2007.

[9] P. Buneman, A. Chapman, and J. Cheney. Provenance management in curated databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006*, pages 539–550, 2006.

[10] Y. Cheah, S. R. Canon, B. Plale, and L. Ramakrishnan. Milieu: Lightweight and configurable big data provenance for science. In *IEEE International Congress on Big Data, BigData Congress 2013, June 27 2013-July 2, 2013*, pages 46–53, 2013.

[11] D. Crawl, J. Wang, and I. Altintas. Provenance for mapreduce-based data-intensive workflows. In *WORKS'11, Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science, co-located with , SC11, Seattle, WA, USA, November 14, 2011*, pages 21–30, 2011.

[12] A. Cuzzocrea. Privacy and security of big data: Current challenges and future research perspectives. In *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data, PSBD@CIKM 2014, Shanghai, China, November 7, 2014*, pages 45–47, 2014.

[13] A. Cuzzocrea, V. Russo, and D. Saccà. A robust sampling-based framework for privacy preserving OLAP. In *Data Warehousing and Knowledge Discovery, 10th International Conference, DaWaK 2008, Turin, Italy, September 2-5, 2008, Proceedings*, pages 97–114, 2008.

[14] A. Cuzzocrea and D. Saccà. Balancing accuracy and privacy of OLAP aggregations on data cubes. In *DOLAP 2010, ACM 13th International Workshop on Data Warehousing and OLAP, Toronto, Ontario, Canada, October 30, 2010, Proceedings*, pages 93–98, 2010.

[15] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008,*

Vancouver, BC, Canada, June 10-12, 2008, pages 1345–1350, 2008.

[16] B. Glavic, K. S. Esmaili, P. M. Fischer, and N. Tatbul. Ariadne: managing fine-grained provenance on data streams. In *The 7th ACM International Conference on Distributed Event-Based Systems, DEBS '13, Arlington, TX, USA - June 29 - July 03, 2013*, pages 39–50, 2013.

[17] B. Glavic, K. S. Esmaili, P. M. Fischer, and N. Tatbul. Efficient stream provenance via operator instrumentation. *ACM Trans. Internet Techn.*, 14(1):7:1–7:26, 2014.

[18] R. Hammad and C. Wu. Provenance as a service: A data-centric approach for real-time monitoring. In *2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014*, pages 258–265, 2014.

[19] R. Ikeda, H. Park, and J. Widom. Provenance for generalized map and reduce workflows. In *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 9-12, 2011, Online Proceedings*, pages 273–283, 2011.

[20] D. Kontokostas, M. Brümmer, S. Hellmann, J. Lehmann, and L. Ioannidis. NLP data cleansing based on linguistic ontology constraints. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 224–239, 2014.

[21] M. Mizan, M. L. Rahman, R. Khan, M. M. Haque, and R. Hasan. Accountable proof of ownership for data using timing element in cloud services. In *International Conference on High Performance Computing & Simulation, HPCS 2013, Helsinki, Finland, July 1-5, 2013*, pages 57–64, 2013.

[22] I. Nunes, Y. Chen, S. Miles, M. Luck, and C. J. P. de Lucena. Transparent provenance derivation for user decisions. In *Provenance and Annotation of Data and Processes - 4th International Provenance and Annotation Workshop, IPAW 2012, Santa Barbara, CA, USA, June 19-21, 2012, Revised Selected Papers*, pages 111–125, 2012.

[23] H. Park, R. Ikeda, and J. Widom. RAMP: A system for capturing and tracing provenance in mapreduce workflows. *PVLDB*, 4(12):1351–1354, 2011.

[24] L. Pipino. Information quality assessment. In *Encyclopedia of Database Systems*, pages 1512–1515. 2009.

[25] Y. S. Tan, R. K. L. Ko, and G. Holmes. Security and data accountability in distributed systems: A provenance survey. In *10th IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, HPCC/EUC 2013, Zhangjiajie, China, November 13-15, 2013*, pages 1571–1578, 2013.

[26] B. Wang, B. Li, and H. Li. Oruta: Privacy-preserving public auditingfor shared data in the cloud. *IEEE T. Cloud Computing*, 2(1):43–56, 2014.

[27] E. Wu, S. Madden, and M. Stonebraker. Subzero: A fine-grained lineage system for scientific databases. In *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, pages 865–876, 2013.