3-1-2016

Securing Big Data Provenance for Auditors The Big Data Provenance Black Box As Reliable Evidence

Deniz Appelbaum
Montclair State University, appelbaumd@mail.montclair.edu

JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING
Vol. 13, No. 1
Spring 2016
pp. 17–36

American Accounting Association
DOI: 10.2308/jeta-51473

Securing Big Data Provenance for Auditors: The Big Data Provenance Black Box as Reliable Evidence

Deniz Appelbaum
Rutgers, The State University of New Jersey, Newark

ABSTRACT: The purpose of this article is to highlight a main issue regarding reliable audit evidence derived from Big Data—that of secure data provenance. Traditionally, audit evidence external to the client has been regarded as superior to other forms of evidence. However, regarding external "messy" Big Data sources that may be material to aspects of the audit, these sources may lack provenance and verifiability. That is, the origins of the data may be unclear and its log files incomplete. According to the standards, such evidence should be considered as less reliable for audit evidence. External auditors, as outsiders of the client, should be able to reproduce the data lifecycle or transaction path, which may not be possible in an electronic environment with incomplete provenance. Furthermore, this mapping or provenance of the data origins and history should be securely maintained so that it cannot be thwarted. This need for secure data provenance has been largely ignored by the business community in its haste to utilize Big Data, but has been acknowledged by extant systems research as being an area that requires attention. This paper contributes to the discussion of Big Data provenance through the lens of public company auditing, where the provenance and reliability of data sources and audit evidence are of paramount importance. This paper also proposes a system of secure provenance collection, the Big Data Provenance Black Box, which is derived from several streams of extant research.

Keywords: auditing; Big Data; data provenance; audit evidence; Hadoop/MapReduce.

INTRODUCTION

Big Data has become the new business currency (CompTIA 2015). To this end, businesses are now collecting more data than they have in the past 2,000 years (Warren, Moffitt, and Byrnes 2015). These businesses regard Big Data as a potential firm asset (Warren et al. 2015; Brown, Chui, and Manyika 2011) and have been reported to have attained 5 to 6 percent gains in productivity from analysis of this data (Brynjolfsson, Hammerbacher, and Stevens 2011). There is an enormous quantity of data now available in many forms from many different sources that is being generated very quickly—2.5 quintillion bytes of data are being generated daily (IBM 2015; Jagadish et al. 2014)—a Big Data deluge (Hey and Trefethen 2003). Most of these datasets are unstructured, derived from social media, sensors, and the Internet of Things (IoT) (Bauer and Schreckling 2013). As such, Big Data is dynamic data with volume, variety, and velocity (Laney 2001) and, more recently, veracity (IBM 2012). Big Data may be defined as the large flows of widely differing data and the aggregation of datasets that cannot be processed using traditional database management tools (Polato, Ré, Goldman, and Kon 2014; Mittal 2013; Zikopoulos and Eaton 2011). Furthermore, the origins and treatments of these datasets are largely unknown, as they often originate outside of the business that is absorbing and analyzing them (Taylor, Haggerty, Gresty, and Hegarty 2010; Tan 2007; Cui and Widom 2003).

For decision makers, researchers, auditors, and regulators, the ability to verify the accuracy of information is of paramount importance (Liao and Squicciarini 2015; Ikeda, Park, and Widom 2011; Li, Roge, Rydl, and Hughes 2007; Nearon 2005; Alles, Kogan, and Vasarhelyi 2002; Elliott 2002, 1996). External auditors may be interested in Big Data for two reasons: first, their clients may be utilizing Big Data for decision making and accounting judgments that could materially affect the financial statements if the data are flawed; and second, auditors themselves may want to access Big Data sources for industry and client assessment, risk analysis, confirmations, and reasonableness tests—if the data are reliable.

The author thanks Dr. Miklos A. Vasarhelyi, the editorial staff, and the anonymous reviewers for their helpful guidance and suggestions. The author also thanks Dr. Nabil Adam for his assistance in the Big Data Seminar and Dr. Graham Gal for his discussion at the 2015 AAA Annual Meeting in Chicago.

Editor's note: Accepted by Miklos A. Vasarhelyi.

Submitted: November 2014
Accepted: March 2016
Published Online: April 2016

The audit standards (Public Company Accounting Oversight Board [PCAOB] 2010, AS 15; American Institute of Certified Public Accountants [AICPA] 2012, SAS 122; International Auditing and Assurance Standards Board [IAASB] 2009, ISA 500) specify that external sources of evidence and information are generally more reliable for verification. However, Big Data potentially poses an opposite situation: due to its possible lack of provenance and veracity, it could be a less reliable source of evidence for auditors. Big Data may not be trustworthy if the organization utilizing it has not employed certain procedures to address its risks (Zhang, Yang, and Appelbaum 2015; Mittal 2013). Basically, with Big Data, much of the innovation has been directed toward processing and analyzing this data of such volume, variety, and velocity and not tracing its veracity or origins and transformations (Liao and Squicciarini 2015). Until very recently, little attention has been paid to the provenance[1] of this Big Data, its pedigree, or lineage (Liao and Squicciarini 2015; Ikeda et al. 2011).

Big Data, due to its volume and velocity, has compelled business organizations to utilize the Cloud for data storage and enterprise applications (Polato et al. 2014). Big Data, due to its immense volume, great variety of format, and streaming velocity of occurrence, has forced numerous firms to utilize applications such as Hadoop MapReduce to process and prepare the data in a form that is manageable for analysis and understanding (Akoush, Sohan, and Hopper 2013; Lin and Ryaboy 2013; Dean and Ghemawat 2008). However, both the Cloud and MapReduce processing create additional challenges to the auditor for evidence verification (Cohen and Acharya 2014; Polato et al. 2014; Lin and Ryaboy 2013). The Cloud is a data repository that resides outside of the business enterprise or Cloud client, the result of which is that the enterprise has partially lost control of the data in an environment where provenance tracking is challenging. Hadoop and MapReduce process the streams of data and may alter and transform it without complete tracking of these alterations. For an enterprise processing Big Data with a Hadoop platform in the Cloud, these provenance issues could be magnified. Audit techniques should take into account the impact of this reliance on messy Big Data by the client. This Big Data may not be providing verifiable evidence for auditors and regulators, particularly if these data materially impact the financial statements.

The auditor, whether internal or external, should be able to access the desired level of provenance of the electronic information under examination, and this provenance tracking should be secure and trustworthy (Bates, Mood, Valafar, and Butler 2013; McDaniel et al. 2010; Hasan, Sion, and Winslett 2009; Braun, Shinnar, and Seltzer 2008). The internal auditor could be utilizing Big Data from sensor streams and social media texts to perform efficiency and fraud auditing more efficiently and effectively (Warren et al. 2015). As such, the origins and paths of the lifecycle of these data should be verifiable by the auditor, and this recording of the lifecycle, the data provenance, should be secure and unalterable. Similarly, external auditors could access Big Data in many forms, primarily from social media and the web, for example, to augment the initial client evaluation decision, to verify the client's fair value assessment of intangible assets, or to evaluate the determination of going concern (Warren et al. 2015).

To summarize, it is envisioned that the external auditor would directly access Big Data to enhance the following typical audit phases:

1. To supplement the auditor's industry and client knowledge acquisition during the Engagement Phase.
2. To supplement the auditor in the risk assessment process of the Audit Planning Phase, similar to the Engagement Phase.
3. As part of Substantive Testing, particularly if re-performing client calculations and analyses that utilized information derived from Big Data. For example, verifying the client's Fair Value assessment of intangible assets that has been partially based on social media information is one task that would require the auditor to access Big Data.
4. During the review stage, the auditor may want to view all the audit results in a greater context and in a comparative sense against the client's own industry and associated Internet media. Critical to this analysis would be any direct social media or macro-economic/demographic Big Data that would indicate a probable Going Concern issue.
5. Big Data may also enhance the auditor's knowledge regarding the client in the Continuous Activities phase, similar to the Engagement and Planning phases.

Big Data could expand the auditor's client and industry knowledge beyond that provided from the client's own data. Evidence collection in this Big Data scenario could not only assist in traditional financial statement verification, but also enhance auditor knowledge for client assessment.

Essentially, the traditional view of audit evidence collection may no longer be sufficient in this more advanced technical business environment (Brown-Liburd and Vasarhelyi 2015). The customary characteristics that define traditional audit

[1] Traditionally, provenance has meant the chronology of the ownership, custody, or location of a historical object. The term was originally mostly used in art, but is now used in a number of domains such as archaeology, paleontology, archives, manuscripts, printed books, medical sciences, and computing. "The primary purpose of tracing the provenance of an object or entity is normally to provide contextual and circumstantial evidence for its original production or discovery, by establishing, as far as practicable, its later history, especially the sequences of its formal ownership, custody, and places of storage. The practice has a particular value in helping authenticate objects. Comparative techniques, expert opinions, and the results of scientific tests may also be used to these ends, but establishing provenance is essentially a matter of documentation" (see: https://en.wikipedia.org/wiki/Provenance).

evidence may not be adequate, and have been proposed as a future research issue (Brown-Liburd and Vasarhelyi 2015). Previously, when the bulk of electronic data were internally generated and quantitative, provenance information was readily available to auditors via system log files (Caster and Verardo 2007; M. V. Cerullo and M. J. Cerullo 2003). In contrast, Big Data may not be internally generated and most likely has been processed outside the client. The provenance tracking that is missing for many Big Data and Cloud systems would appear to challenge the long-held view in the audit profession that external data equal reliable data.

To expand upon this concern, the purpose of this innovation paper is to discuss the challenge of provenance evidence verification facing the auditor in the current electronic Big Data business environment, to identify the current gaps in the audit and systems research regarding secure Big Data provenance, and to propose a model and direction for future research—the Big Data Provenance Black Box. The "Introduction" section is followed by a review of the Auditing Standards on Evidence Collection, where the evidence attributes are discussed and issues of digital evidence collection, with an emphasis toward external auditors, are highlighted. These attributes and evidence collection issues will shape the remainder of this discussion. The third section offers an overview of provenance collection, emphasizing security. The fourth section discusses Hadoop/MapReduce and Hadoop in the Cloud and their impact on reliable evidence collection. The Big Data Provenance Black Box is proposed next, and the final section offers a conclusion and commentary on areas for future research regarding evidence collection in the current Big Data business environment and the external auditor.

THE AUDITING STANDARDS ON EVIDENCE COLLECTION

The main purpose of the work conducted by an auditor in an external engagement is to obtain reasonable assurance that the client's financial statements are basically free from material misstatements and to subsequently express an opinion regarding these financial statements in the auditor's report. To accomplish this task, the auditor must design and perform audit procedures to obtain sufficient appropriate evidence; furthermore, the Audit Standards require auditors to examine physical evidence as part of the risk assessment process (PCAOB 2010, AS 15; AICPA 2012, SAS 122; IAASB 2009, ISA 500).

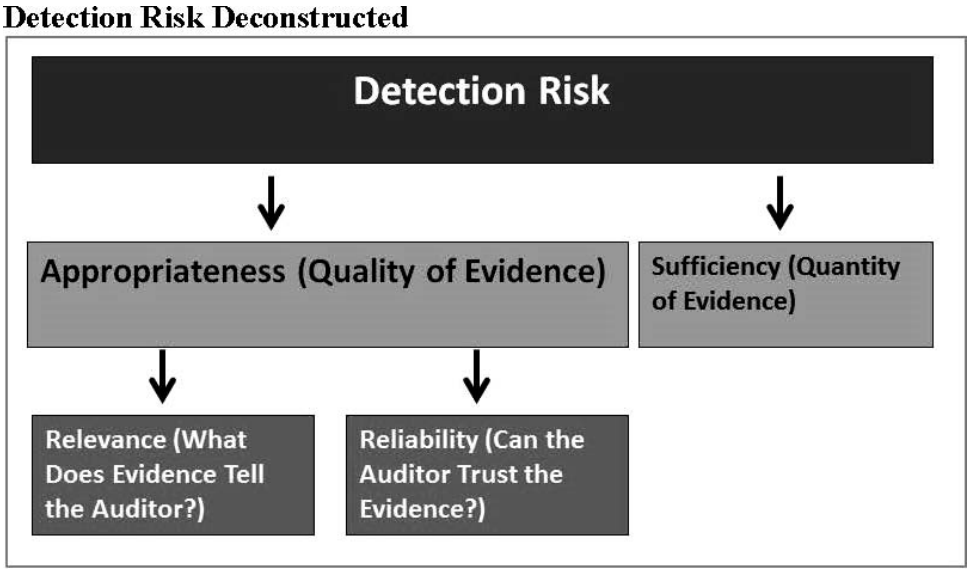
Additionally, the Sarbanes-Oxley Act (SOX) demands that public auditing firms maintain the provenance of an audit report (and all of its supporting information) for at least seven years after its issuance (U.S. House of Representatives 2002; Tsai et al. 2007). The Sarbanes-Oxley Act also mandates that auditors verify the accuracy of the information or evidence that forms the basis of their audit opinion. Management also needs to be able to audit and verify each step of every transaction, with all its data inflows and outflows. The client's document management, access to data, and storage of information must provide auditing (vouching, verifying, and tracing) capabilities (Li et al. 2007). As such, many public companies have sought to reduce compliance costs by collecting data in a real-time fashion to provide continuous monitoring of 100 percent of the transactions.

Audit evidence is all the information used by the auditors to form the audit opinion (PCAOB 2010, AS 15). This audit evidence must be both sufficient and appropriate; the degree of each is determined by the other (see Figure 1). Sufficiency is the measure of the quantity, the amount of which is determined by Detection Risk, determined by the auditor, and the level of quality of the evidence or its Appropriateness (PCAOB 2010, AS 15). Appropriateness is the measure of Relevance (what does the evidence tell the auditor) and Reliability (can the auditor trust the evidence). Basically, if the underlying information is not reliable and its provenance or lineage is not verifiable, then more evidence will need to be collected and reviewed—to a certain degree. Poor-quality evidence cannot always be compensated by collecting a larger amount (PCAOB 2010, AS 15). If the evidence is relevant and reliable, possessing trustworthy provenance, then the auditor can proceed confidently with substantive testing and other analytical procedures (PCAOB 2010, AS 15). Traditionally, much of this evidence has been paper, observations, inquiries, and other physical formats. As shown in Figure 1, the aspect of Appropriateness is quite significant to the determination of Detection Risk.

However, in today's complex IT environment and Big Data, the nature and competence of this audit evidence has changed (Brown and Vasarhelyi 2015; Caster and Verardo 2007; Nearon 2005). Every phase of a transaction is computer-generated and recorded and can only be verified electronically. For example, every phase of a purchase or a sale may occur within the electronic system. Or, with additional information available from external Big Data, intangible assets might be partially valued by the client from information derived from text analysis of aggregated tweets and web scraping of social media. With more than 90 percent of these records in easily alterable digital formats that possess many iterations and possibilities, provenance of data sources and provenance of log files become of paramount importance (Nearon 2005). To summarize and expand, Table 1 displays the following differences that exist between paper evidence and electronic evidence (Brown-Liburd and Vasarhelyi 2015; Colbert and Smalling 2011; Ratcliffe and Munter 2002).

The implications for electronic accounting data and evidence collection are substantially different from those of manual, paper-based examination. Many of the characteristics that are strengths with paper-based evidence pose issues for electronic evidence. It could be said that technology has weakened a number of traditional forms of audit evidence (Caster and Verardo 2007). Whereas paper documentation is considered not to be easily altered, electronic data may be easily changed and these

FIGURE 1
Depiction of the Role of Appropriateness and Reliability of Evidence in Detection Risk



alterations might not be detected, absent the appropriate controls. In paper-based evidence collection, sources that are verified external to the client are considered to be highly reliable, whereas external electronic evidence is difficult to verify for veracity, origin, and reliability. External data also frequently lack evidence of approvals and signatures. Paper-based evidence is easy to evaluate and understand, whereas electronic data and evidence may require a high level of technical expertise of the auditor. Also, whereas manual, paper-based information is competent for re-performance and recalculation, electronic evidence may require additional complex procedures due to its random and dynamic condition. These five characteristics will assist in the evaluation of Big Data and the suggested provenance collection system in the remainder of the paper.

Statement of Auditing Standards No. 80 (SAS 80), *Amendment to Statement on Auditing Standards No. 31, Evidential Matter,* was released to provide guidance regarding audit evidence collection in electronic environments (Auditing Standards Board [ASB] 1996). SAS 80 clarifies that tests of IT controls, together with substantive testing, may provide sufficient evidence to form an audit opinion if the client's reliance on IT is so great that detection risk cannot be limited to substantive testing alone (ASB 1996). IT controls may be examined by inspection of log file activity for compliance verification. Log files record the dynamics, activity flows, and events in a system. A log file will record the data or transaction origin if this information was provided, and any subsequent changes with time/location/authorization/actor stamps and identifiers (Accorsi 2006). Logging, in fact, has typically been recognized as the recording of significant events that may need to be identified in a future audit. These log entries should be considered as evidence of origins, authorizations, permutations, alterations, IP addresses, and time strings (Vaughan, Jia, Mazurak, and Zdancewic 2008). Log files are also considered to be the starting point for process mining (Jans, Alles, and Vasarhelyi 2010; van der Aalst, van Hee, van Werf, and Verdonk 2010), where the systemic, reliable, and trustworthy recording of events and data (business provenance) is required. Additional future research and discussion could focus on how provenance log files may provide sufficient evidence for internal control compliance evaluations in an electronic or Big Data environment.

Nearon (2005) proposed that an appropriately skeptical auditor should inquire as follows regarding electronic evidence, log files, and IT controls:

- Is the electronic evidence subject to alteration without an audit trail or evidence of this change?
- Is there an audit trail that clearly ties the digital evidence back to the initiating entry or data origin? Or can this trail lead forward to the point of inclusion on the face of the financial statements?
- Does the electronic evidence include metadata that identifies who made the entry and when?
- What are the controls designed to prevent unauthorized changes to the digital evidence after it was created?
- Who has or had access rights to change the digital evidence?
- How does the auditor know that the digital evidence has not been intentionally altered?
- Has the audit logging process been configured to record all access attempts, regardless of whether successful or not?

TABLE 1

Review of Evidence Characteristics
Adapted from Brown and Vasarhelyi (2015), Colbert and Smalling (2011), and Ratcliffe and Munter (2002)

Review of Evidence Characteristics

Evidence Characteristics	Paper Evidence	Electronic Evidence
Alterability Easily altered evidence lacks credibility; evidence should be difficult to alter	Difficult to alter without detection	Alterations may be difficult to detect without performing specifically designed tests
Prima Facie Credibility SAS 80 establishes a hierarchy of credibility—outside sources e\nhance credibility when independent of the client and confirmable	Outside sources of paper and documentary evidence and submitted directly to the auditor enhance credibility; inside sources of paper evidence that have been reviewed and processed by outsiders is also reliable	An electronic document derives its credibility primarily from the controls within the system. Outside electronic documentation/data is missing the assurance of system controls that the document or data is not fraudulent or altered
Completeness of Documents All essential terms of a transaction are verifiable	Typically, all essential terms are included on its surface in a text/human readable form	An electronic system may substitute codes or cross-references to other data files that may not be accessible
Evidence of Approvals This essential aspect of internal controls should be easily verifiable and transparent	Approvals integrated into paper documentation add to completeness	Electronic approvals may be similarly integrated, but need additional verification
Ease of Use Simplicity of application and access encourages compliance	Paper evidence can usually be evaluated without the use of additional tools and/or skills	Electronic evidence may require extraction of data by an expert
Clarity Competent evidence should allow for the same re-performance and conclusions by other auditors	The nature of paper documentation is readily clear	The nature of electronic evidence is not always so clear, particularly in the absence of appropriate controls

- Have the audit logs been reviewed independently?
- Has the continuity of logs been maintained and any gaps justified?
- Have the logs been frequently copied to offline, read-only media and stored in a separate secure location, inaccessible to those who might be motivated to change it?
- Has the access to the logs and their security settings been recorded, and limited to only authorized persons?

All of these questions could potentially be satisfied with an appropriate secure provenance system, which will be discussed in the sections that follow. Basically, business provenance provides assurance of traceability, verifying the lineage of the event or transaction. With the assurance provided by the reliable and trustworthy recording of event logs and audit trails, known as provenance of process flows, the auditor can embark on a risk assessment analysis based on a secure foundation of accurate accounting data, event log files, and process flows.

DATA PROVENANCE

Data Provenance

Provenance, by definition, means origin and lineage, and is used quite extensively in the arts, antiques, and scientific domains to describe lineage or ownership of different items (Moreau et al. 2008a; Moreau et al. 2008b; Moreau et al. 2008c). When applied to data, provenance may be metadata or log files/audit trails pertaining to the lineage of a data event, capturing and recording its origins, derivations, and transformations, and has been used extensively in the sciences (Bose and Frew 2005;

TABLE 2

Review of Generally Suggested Provenance Types per Audit Task

Suggested Provenance Types for Audit Tasks

Purpose/Audit Task	Provenance Type	Qualities
Internal Controls Verification/Re-performance	Coarse-grained	Work flows- or process flows-based; data at schema level; DAG models
Evidence Collection/Verification; Recalculations	Fine-grained	Data elements/metadata; DAG models

Moreau et al. 2008c; Simmhan, Plale, and Gannon 2005a, 2005b). As businesses increasingly depend on data from sources outside the firm, such as Big Data, the need for provenance of this data grows exponentially (Cheah and Plale 2012).

As the available data have become larger, i.e., Big Data, the analysis required to achieve knowledge discovery requires more complex and distributed processing (Crawl, Wang, and Altintas 2011; Davidson and Freire 2008; Frew, Metzger, and Slaughter 2008). Therefore, it is quite possible that the originating data could have been entirely different from the data that the organization now possesses, due to pre-processing applications (Cheney, Chiticariu, and Tan 2009; Glavic and Dittrich 2007; Scheidegger et al. 2008; Simmhan et al. 2005a). Hence, provenance is essential to the business domain as it may be used to provide an audit trail for regulatory and audit engagement purposes (Simmhan et al. 2005a, 2005b). For the purposes of this paper, data provenance is considered to be all the information that assists in determining the origin, derivations, and transformations of a data product or dataset (files, tables, process flows, log files, virtual collections) (Cheah and Plale 2012). Two main features of data provenance are the originating data product itself and the process flows that record the activity and locate points of transformations of the originating data product to its current form (Ikeda and Widom 2010; Tsai et al. 2007).

Data provenance can be available explicitly or deduced indirectly. The explicit model, or data-based model, collects lineage metadata about the data and transformations directly. A provenance Directed Acyclic Graph (DAG)[2] is directly associated with the data product whose lineage it describes. The indirect model, or process-oriented model, describes the deriving processes that contribute to a dataset's existence.

Provenance may also be fine-grained (explicit and detailed) or coarse-grained (deduced and processed through a workflow) (Tsai et al. 2007). The size of provenance information may exceed that of the dataset's and the storage costs may be substantial. The storage location and format of the provenance should also be determined by the frequency and application of use. The granularity (and hence the cost) of the provenance to be recorded will depend on the inherent risk of the business cycle, the origins of the data (internal/external), the type of dataset (structured/unstructured), and the impact or potential materiality of the dataset on the financial statements. Table 2 summarizes the provenance types generally applicable to the audit examination tasks.

The business domain has typically worked with organized, quantitative, and mostly internally generated data, where the structure and semantics of the data are organization-wide. However, many businesses are now collecting and analyzing data that are messy and unstructured, whose issues are further compounded by their aggregation to a data warehouse (CompTIA 2015). Basically, the data are required to be extracted, cleansed, and transformed from many different operational databases and external sources before they are placed in a data warehouse or a Cloud. Provenance is also essential in a warehouse environment, as warehouse data are built upon layers of data views, with one layer derived from layers below it, and where lineage information is essential for vouching and tracing. This warehouse provenance data product and its transformations may be conceptualized graphically as a DAG with nodes representing the different iterations of the data product and with edges representing the transformation processes.

Goble (2002) summarized the feasible applications for provenance information and that research has been adopted and modified in this paper to the external audit domain as follows:

- *Data Quality:* Lineage can estimate and verify data *quality* and data *reliability* based on the source information and transformations (Simmhan et al. 2005a). The level of data included in the provenance determines the extent to which

[2] A Directed Acyclic Graph (DAG) is a design from computer science that models a wide variety of activities or process flows. The DAG consists of the following elements: Nodes, which represent objects or points of data; Directed Edges, which are directional arrows or edges from one node to another; a Root Node, which has no parents and only children; and Leaf Nodes, which have no children. Arrows in a DAG may not form a cycle, where these arrows illustrate the basis. A DAG may be considered to be a tree-like data structure, similar to decision trees—see: http://ericsink.com/vcbe/html/directed_acyclic_graphs.html

the quality can be estimated—the more fine-grained (detailed) the provenance, the more precise the estimation of data quality. The more coarse the provenance (summary level), the less detailed the estimation. The granularity of provenance to be recorded may vary based on the inherent risk of the business cycle, the origins of the data (internal/external), the type of dataset (structured/unstructured), and the impact or potential materiality of the dataset on the financial statements.

- *Audit Trail:* Provenance can provide a means by which to audit the *veracity* of the data and the process by which they evolved. This information is important for accounting and auditing purposes, particularly for data that are ambiguous. The standards stipulate that uncertain evidence or data must be thoroughly examined with substantive procedures such as re-performance, recalculation, trend analysis, analytical procedures, and vouching/tracing (ASB 1996, AS 80). Lineage can help identify any exceptions that took place in data creation. Provenance can also be used to back-track and identify the source of errors and violations of controls (Galhardas, Florescu, Shasha, Simon, and Saita 2001).
- *Replication Recipes:* Detailed or fine-grained provenance can allow repetition of data derivation and be a recipe for its re-performance or recalculation. Re-performance and recalculation are integral procedures for most audits of financial statements. With provenance, the auditor can vouch and trace from the dataset origin to the face of the financial statement and *vice versa*. Many current applications of provenance have adopted XML for representing lineage information (Bose and Frew 2004). As a suggestion for future research, XBRL, as an XML derivative, may present possibilities to the business domain as a provenance metadata standard, particularly since public companies are currently required to prepare their financial statements in XBRL.
- *Attribution:* Pedigree or lineage can help determine or verify ownership of the source data used to generate certain estimates or calculations. An auditor can verify the creators of intellectual property and copyrights or look at the lineage chain to see who has had access. Lineage is also the means by which citations are tracked in the academic publications domain (Cameron 2003). Provenance can also be used to assign liability in case of errors in the dataset (Cameron 2003).
- *Informational:* A more generic use of provenance is as a metadata categorization that may be utilized for queries, with the trail of any particular query available for re-performance, avoiding duplication of effort. Annotations that accompany the provenance may help interpret the data in the context required, particularly for archived data that are accessed long after they were generated (Simmhan et al. 2005a, 2005b).

Actually, without assurance that this data provenance has been collected and maintained securely, the audit records of the origins and transformations of these data are suspect (Cheah and Plale 2012; van der Aalst et al. 2010; Buneman and Tan 2007; Buneman, Khanna, and Tan 2001, 2000). The use of any provenance as a basis for decision making, whether by the client or the auditor, depends on the trustworthiness of that provenance information itself (Bier 2013; Aldeco-Pérez and Moreau 2010; Simmhan et al. 2005a, 2005b). There should be assurances that the provenance information was not tampered with, and securing provenance with digital signatures has been a common solution (Aldeco-Pérez and Moreau 2010; Simmhan et al. 2005a, 2005b). Securing provenance information will significantly enhance its usefulness and value for auditors as a reliable source of examination evidence and accounting data.

Secure Data Provenance

Provenance has been recognized, due to its ability to track causal dependencies between data and events that explain the data's current state, as a means to achieve information accountability (Aldeco-Pérez and Moreau 2010; Moreau et al. 2008b; Weitzner et al. 2008). Provenance provides transparency of the datasets it reflects and is auditable, allowing auditors to decide whether information is credible or has been used in the proper way. However, the integrity of this provenance information and its graphs are critical to guaranteeing the quality of a data provenance-based audit. Basically, the auditor should be able to verify that the information tracking the subject datasets has not itself been altered. Most research to date has suggested digital signatures to be the most feasible means of securing the provenance documentation (Bier 2013; Aldeco-Pérez and Moreau 2010; Accorsi 2009, 2006; Simmhan et al. 2005a, 2005b). The provenance information flows should be recorded securely in these four stages in order to guarantee a correct audit report (Aldeco-Pérez and Moreau 2010):

- Recording of any process documentations in which influential components make assertions about the actions they perform on the dataset, in addition to the alterations.
- Storage of the provenance information in which it is continually stored in a Secure Provenance Repository separately located, with highly enforced access controls, and is read-only.
- Querying of the provenance information should also be recorded.
- Analysis of provenance information should be recorded, which provides the basis for the audit report.

If the provenance data and DAGs are secured via digital signatures at the formation, recording, storage, querying, and analysis stages, then the provenance data may be regarded as reliable for auditors (Accorsi 2009; Alles, Kogan, and Vasarhelyi

TABLE 3

Summary of Satisfaction of Audit Evidence Characteristics by Evidence Type

Audit Evidence Characteristics by Evidence Type			
Evidence Characteristics	**Paper Evidence**	**Electronic Evidence**	**Secure Data Provenance**
Difficult to alter	√		√
Credible	√	√ for internal data	√
Complete	√		√
Evidence of approvals	√		√
Easy to use	√		
Clear	√		√

2004). With the use of digital signatures, security is assured in the transmission and storage phases. In the transmission phase, origin authentication, message confidentiality, message integrity, message uniqueness, and reliable delivery are assured with digital signatures. Similarly, in the storage phase, entry accountability, entry integrity, entry confidentiality, and tamper prevention are assured.

With digital signatures, a small change to the original data results in a huge difference to the hashed message (digital signature). It is computationally impossible to create two different documents that have the same digest; so if one document is altered, then it would be impossible to create another document with the exact same digital signature. A digital signature does not reveal any information about the content of the provenance data itself, only if the content has been altered (Alles et al. 2004). With digital signatures, not only is the transmission and storage of provenance records secure, but also this security itself is assured. With digital signatures, the provenance information cannot be thwarted.

The ability of secure provenance to satisfy the requirements of audit evidence that were discussed in the "Auditing Standards on Evidence Collection" section are shown in Table 3.

Table 3 summarizes the information from Table 1, extended with the attributes of a secure data provenance storage system using digital signatures. Although secure data provenance comes close to meeting the attributes of audit evidence as required by the audit standards, it is not considered to rank highly for ease of use generally. For auditors to navigate a secure provenance data warehouse, applications would need to have been scripted that would be interactive and provide a simple interface. Such applications have been proposed by academics using Python, Perl, or Matlab (Simmhan et al. 2005a, 2005b).

However, to date, there has not been research published specifically about secure provenance of Big Data in Hadoop. This may be due to the rapidly expanding exposure and availability of Big Data, in which common applications such as MapReduce and high-capacity storage locations such as the Cloud have neglected provenance issues until recently (Polato et al. 2014). There are many studies of Hadoop or MapReduce in the area of Big Data, but only a few that discuss data provenance in Big Data or Hadoop (Chen and Plale 2015; Imran, Agrawal, Walker, and Gomes 2014; Akoush et al. 2013; Che, Safran, and Peng 2013; Ghoshal and Plale 2013; Crawl et al. 2011; Park, Ikeda, and Widom 2011; Simmhan et al. 2005a, 2005b). Furthermore, none of the studies provide for a secure form of data provenance in Big Data applications (Ikeda et al. 2011; Margo and Smogor 2010; Aggarwal 2009; Bao, Cohen-Boulakia, Davidson, Eyal, and Khanna 2009; Muniswamy-Reddy et al. 2009; Souiah, Francalanza, and Sassone 2009; Cohen-Boulakia, Biton, Cohen, and Davidson 2008; Freire, Koop, Santos and Silva 2008; Buneman and Tan 2007; Davidson et al. 2007; Glavic and Dittrich 2007; Muniswamy-Reddy, Holland, Braun, and Seltzer 2006; Simmhan et al. 2005a, 2005b; Tan 2004; Buneman et al. 2001). Basically, if the provenance information about the Big Data cannot be stored securely, then there is no point in collecting it for auditing purposes. Without security measures, the data provenance recording is not reliable (Buneman and Davidson 2010). For auditors, unreliable information equals poor-quality evidence.

HADOOP/MAPREDUCE AND THE CLOUD

Hadoop/MapReduce

In the realm of Big Data, MapReduce applications such as open source Hadoop have been widely adopted (Akoush et al. 2013; Dean and Ghemawat 2008). Hadoop as a MapReduce agent has become synonymous with Big Data processing and analysis (Crawl et al. 2011), particularly in larger public companies (CompTIA 2015). Hadoop was designed as an open source software framework that would provide a scalable distributed storage and parallel processing system for structured and unstructured Big Data sets (Cohen and Acharya 2014). If an internal or external auditor is working with Big Data, then, most

likely, he or she will be referring to datasets that have been processed with Hadoop. Many social media sources and aggregators of Big Data, such as Facebook, Twitter, Yahoo, and Google, employ various forms of Hadoop or MapReduce (Lin and Ryaboy 2013; Patil 2012; Hammerbacher 2009).

Not only do these social media data generators utilize Hadoop and MapReduce, much of their qualitative, textual, video, and audio feeds must be transformed and integrated before analysis (Lin and Ryaboy 2013). These processes may have altered the data and may not have been completely recorded or logged unless provenance collecting applications were added. Furthermore, Twitter, which has become a predominant social media source for business promotion, customer service, political campaigning, medical services, health care, marketing, and stock market prediction (Chu, Gianvecchio, Wang, and Jajodia 2012; Bollen, Mao, and Zeng 2011; Hughes and Palen 2009), is plagued with issues of fraudulent accounts and spam campaigns whose origins are not clear/traceable (Cresci, Di Pietro, Petrocchi, Spognardi, and Tesconi 2015; Duncan 2015; Chu et al. 2012; Castillo, Mendoza, and Poblete 2011; Thomas, Grier, Song, and Paxson 2011). Why is this important for auditors? Depending on the client industry and business cycle, Twitter data and other social media sources may have been used by the client in its analytics to gain additional insights beyond mere quantitative analysis (Lin and Ryaboy 2013; Bollen et al. 2011). If the results of these analytics contribute to information that is material to the financial statements, then auditors should be concerned about the provenance of the contributing social media Big Data, as the risk of material misstatement has increased.

Hadoop consists of two functions: Map and Reduce. The user-provided Map function reads, filters, and transforms data from an input file, creating a set of intermediate records. These intermediate records are then usually split via a certain hash function into different buckets. Then the user-provided Reduce function processes and combines all of the intermediate records associated with that hash value into new records that are written into parallel output files. Essentially, the system splits large datasets into smaller pieces, distributes them to as many output files as possible, and then processes the data in each parallel folder so that it is tightly aggregated (Cohen and Acharya 2014). Processing speed and data replication were the core goals behind Hadoop's evolution, with provenance and security a secondary concern. Programs developed with the Hadoop model are parallel because there are no inter-key data dependencies. As such, MapReduce is tolerant of system failures as problematic functions can be restarted independently of the other parallel operations. MapReduce functions are usually expressed as a series of jobs creating a computational workflow. Provenance metadata are captured only at two main points within the core Hadoop platform, unless there have been additional specific provenance process applications added to the Hadoop software (Cohen and Acharya 2014).

Provenance metadata in the basic Hadoop are captured at the storage level and at the resource management level (Alabi, Beckman, Dark, and Springer 2015). The storage-level metadata captures such information as file location, ownership settings, file type, permissions settings, and transaction history—all useful information for provenance. The resource management collects and tracks the data provenance related to the application of Hadoop, but at two points only (Alabi et al. 2015). Therefore, much of the current research in Hadoop provenance is related to enhancing another aspect of provenance, the tracking and the lineage of the Hadoop application workflows (Alabi et al. 2015; Akoush et al. 2013). This additional coarse-grained provenance serves the purpose for tracing and vouching the data outputs back to their associated input activities and origins, and *vice versa,* for the detection of data alterations or any type of suspicious activity.

As can be imagined, the complex MapReduce processes could result in an even more extensive provenance, larger than the workflow that it records and resulting in significant overhead; therefore, current research has been focused on establishing feasible provenance collection in Hadoop (Alabi et al. 2015). For example, one extension of Hadoop that was developed to support provenance capture and tracing for workflows of MapReduce jobs is Reduce and Map Provenance or RAMP (Park et al. 2011; Ikeda et al. 2011). However, there was a fairly large runtime overhead of 76 percent on unstructured Twitter data. Another study presented an application of MapReduce in Kepler,[3] a Kepler+Hadoop framework, to record provenance of workflows (Crawl et al. 2011). However, word count tests took 2.5 times longer to execute when the provenance capture was enabled (Crawl et al. 2011).

A more recent application of provenance in Hadoop is HadoopProv (Akoush et al. 2013). HadoopProv was designed as a modification of Hadoop that takes advantage of the metadata that Hadoop captures while also tracking lineage of data at the process log level. The authors claim that provenance capture overheads are reduced by treating the Map and Reduce phases separately and deferring construction of the provenance Directed Acyclic Graph (DAG) to the query stage. HadoopProv was also designed to capture provenance at the record level, and this level of fine-grained tracking allows for incremental process and log analysis. The temporal overhead of HadoopProv was 10 percent on a typical MapReduce workload (Akoush et al. 2013). In all three approaches, security measures of the provenance files were suggested by the authors as an area for future research.

[3] Kepler is an open source software application for the modeling and processes of scientific data; see: https://code.kepler-project.org/code/kepler-docs/trunk/outreach/documentation/shipping/2.5/getting-started-guide.pdf

Hadoop/MapReduce in the Cloud

Further compounding the issue of feasible provenance collection of Big Data is the recent migration of Hadoop platforms to the Cloud[4] (Olavsrud 2016). The Cloud has become a popular pay-as-you-go location for data storage, due to its flexibility and scalability (Assunção, Calheiros, Bianchi, Netto, and Buyya 2014). Clouds are known for their ability to scale dynamically upward or downward depending on demand and workload. Hadoop and other MapReduce systems have also been established with Cloud providers as Platform as a Service (PaaS). However, the Cloud is perceived as being insecure (O'Driscoll, Daugelaite, and Sleator 2013; Armbrust et al. 2010), providing scanty locational provenance as a result of this scalability and flexibility. Clouds are generally untrusted since the guarantees provided regarding data transformations and locations are minimal (Sakka, Defude, and Tellez 2010). Furthermore, most Cloud providers offer clients little capability on data, application, and service interoperability. Most Cloud storage services are not designed to effectively and efficiently store provenance data, due to the cyclic nature of provenance—its need to be stored separately, yet linked to the data objects (Muniswamy-Reddy and Seltzer 2006). Currently, provenance of the Cloud persists as an open research problem (Assunção et al. 2014). For auditors, the use of the Cloud for either processing or storage of Big Data by a client may likely increase the risk that the relevant data are not reliable as audit evidence due to the minimal provenance of transactions.

As discussed in the second section, an appropriately skeptical auditor should inquire as follows regarding Big Data electronic evidence, log files, and IT controls in the core Hadoop platform or Hadoop in the Cloud Big Data context:

- Is the Big Data electronic evidence subject to alteration without an audit trail or evidence of this change?—Quite possibly, the data have been altered in core Hadoop with minimal provenance. Ideally, the provenance flows should be continually linked to the subject data and should be recording any permutations.
- Is there an audit trail that clearly ties the digital evidence back to the initiating entry or data origin? Or can this trail lead forward to the point of inclusion on the face of the financial statements?—Not offered in the core Hadoop platform, but this aspect of provenance may be added.
- Does the Big Data electronic evidence include metadata that identifies who made the entry and when?—Core Hadoop does not record metadata outside of the storage and resource management points, but could be built into the Hadoop platform modifications.
- What are the controls designed to prevent unauthorized changes to the Big Data digital evidence after it was created?—The evidence of the enforcement of these controls is available through access activity log files, which are minimally recorded in base Hadoop.
- Who has or had access rights to change the Big Data digital evidence?—The evidence of enforcement of access rights is available only at two points in Hadoop.
- How does the auditor know that the Big Data digital evidence has not been intentionally altered?—Core Hadoop can only provide metadata at two points.
- Has the audit logging process been configured to record all access attempts, whether successful or not?—Core Hadoop is not configured for that degree of logging.
- Have the audit logs been reviewed independently?—This control is independent of the Hadoop platform.
- Has the continuity of logs been maintained and any gaps justified?—Core Hadoop does not provide enough metadata to determine this.
- Have the logs been frequently copied to offline, read-only media and stored in a separate secure location, inaccessible to those who might be motivated to change them?—Hadoop, as originally configured, does not copy this information.
- Has the access to the logs and their security settings been recorded, and limited to only authorized persons?—This information is not provided by core Hadoop and additional applications are required.

Clearly, Hadoop requires applications that may contribute additional aspects of provenance to the basic platform (Lin and Ryaboy 2013). The next section proposes secure provenance recording, extended to the HadoopProv framework discussed earlier in this section.

[4] According to the National Institute of Standards (NIST), "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This Cloud model is composed of five essential characteristics, three service models, and four deployment models." The five essential characteristics are: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The three service models are Software as Service (SaaS), Platform as Service (PaaS), and Infrastructure as a Service (IaaS). The four deployment models are as Private Cloud, Community Cloud, Public Cloud, and Hybrid Cloud; see: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

THE BIG DATA PROVENANCE BLACK BOX AND EVIDENCE COLLECTION

The Big Data Provenance Black Box

All of the proposed systems to date make use of separate Big Data provenance storage files (Akoush et al. 2013; Park and Lee 2013; Crawl et al. 2011; Ikeda et al. 2011; Park et al. 2011). However, there is scant detail provided about a critical aspect of provenance for auditors: secure provenance record storage. Furthermore, these files are likely to be much larger than the Big Data files that they describe, as a many-to-one scenario (Ghoshal and Plale 2013; Buneman et al. 2011). As such, the storage of provenance ought to be kept separate from the main files, so as to not encumber any processing overhead (Hasan et al. 2009). However, if the provenance is being frequently queried, then there could be partial or full connections to the main workflow (Braun et al. 2008; Glavic 2014; Bao et al. 2009).

Storage of Big Data provenance files is as critical an aspect as the recording of the Big Data origins and transformations, since the storage should be secure (Hasan et al. 2009). Maintaining the integrity and security of data provenance is further complicated by the fact that it is linked to the data itself. These linkages are also expressed as provenance and audit workflows. Basically, assurance needs to be provided that the provenance records of the data and the audit workflows themselves have not been altered or thwarted (Aldeco-Pérez and Moreau 2010; Braun et al. 2008) while being simultaneously connected to the Big Data itself.

This paper proposes a conceptual framework by which to achieve this secure storage of Big Data Provenance—that of a Big Data Provenance Black Box (BDPBB). The concept of a Black Box for provenance or log file storage is not a new concept and has been proposed previously (Stamatogiannakis, Groth, and Bos 2015; Accorsi 2009; Alles et al. 2004; Oppliger and Rytz 2003). In fact, Oppliger and Rytz (2003) explain at length how digital signatures, although feasible for securing provenance information, should be deployed in digital black boxes to truly provide reliable and trustworthy evidence. This paper extends the concept of this digital black box to the issue of secure provenance tracking of Big Data in Hadoop, in support of reliable evidence collection for auditors.

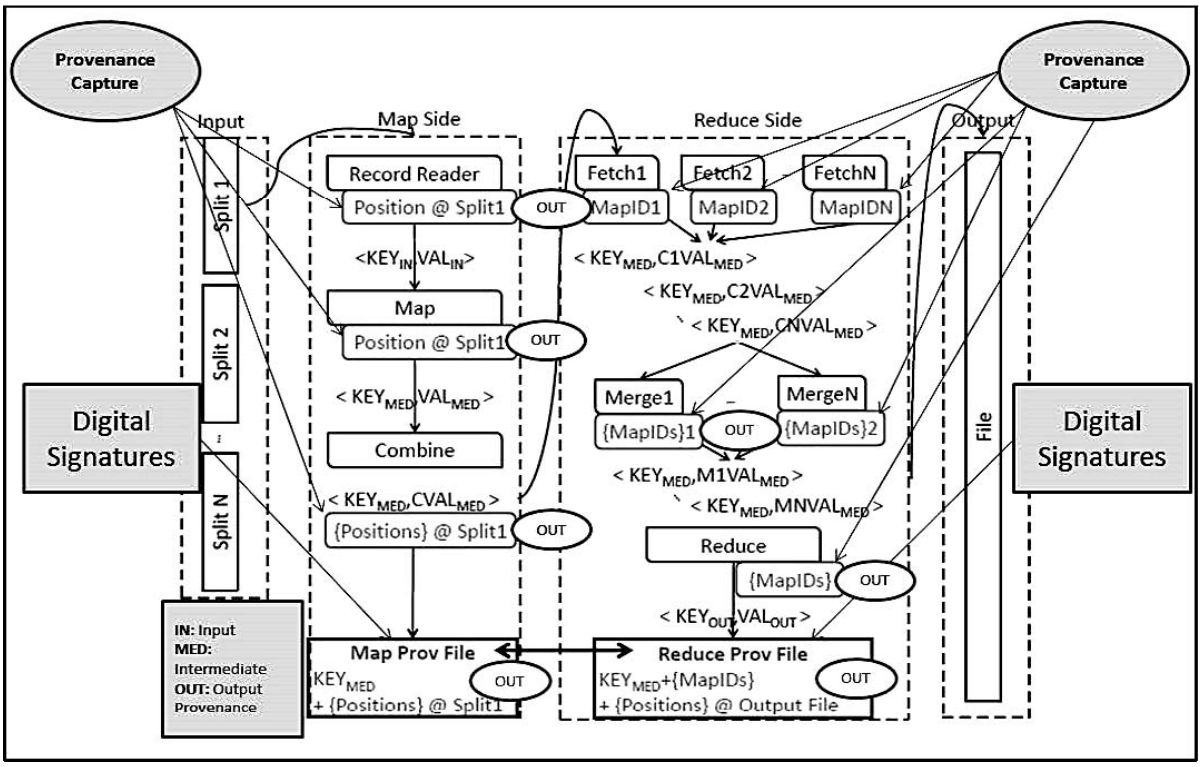
Black Boxes on airplanes record cockpit conversations and sounds, as well as numerous digital measurements sent from many different sensors. The concept here is that everything is being recorded and stored in an orderly fashion, as separate logs of activities in case these actions need to be analyzed or audited in the future. Black Boxes may be regarded as a type of log recorder. Recording data provenance is basically creating logs of data about the activities of data point(s) or document (Glavic 2014; Ghoshal and Plale 2013; Muniswamy-Reddy, Macko, and Seltzer 2010; Souiah et al. 2009). Expanding on an earlier work (Alles et al. 2004) where Black Boxes were conceptualized as an internal audit tool and Black Box (BB) Log file, this paper proposes that such a BB concept would serve well in the capacity of a Big Data provenance collection system. The main difference with a Big Data Provenance Black Box (BDPBB) and the BB log file is that the former is primarily concerned with all provenance data connected with a particular firm, whereas the latter is primarily interested in data pertaining to the audit of that firm (Alles et al. 2004). The BDPBB would generate a much larger Big Data than it records, so it would be magnitudes larger than the data collected in the BB log file of Alles et al. (2004). However, given the rapidly decreasing cost of data storage, it is possible that cost might be less of a prohibiting factor for the collection and storage of huge provenance files.

The BDPBB could record every transaction and alteration of the Big Data into the provenance files. It could also record less granular provenance or work flows, the level of which to be suggested by the auditor and undertaken by management. The provenance data could be recorded in a standardized format, determined by and particular to each host, and that would enable search algorithms to find certain data points at certain time recordings. This standard is necessary to avoid the BDPBB becoming a data dump, where finding anything would be prohibitive in effort and cost. No entry to the log could be altered after it is recorded; it would be read-only. This read-only quality would make the BDPBB feasible for an audit trail (Bishop 2006). The provenance production would be write-once and the provenance query would be read-only.

The most important assurances for the BDPBB to provide are those of integrity, security, and confidentiality, as these qualities provide security (Braun et al. 2008; Cheah and Plale 2012). The BDPBB has to maintain privacy and security with its contents as read-only. Furthermore, stringent access controls should be applied utilizing a role-based approach (Ferraiolo and Kuhn 2009; Bishop 2006). Protecting the BDPBB against tampering and alteration could be achieved with write-once mediums. However, these mediums can be destroyed. Another possibility is to hand the BDPBB over to a trusted third party for protection (McDaniel et al. 2010). However, this transfer would create its own set of security issues.

Or perhaps the firm could compute and transfer a digital signature of the BDPBB to this third party. After all, it is possible to detect if the BDPBB has been altered by using digital signatures (Stamatogiannakis et al. 2015; Accorsi 2009; Hasan et al. 2009; Tsai et al. 2007; Bishop 2006; Alles et al. 2004; Oppliger and Rytz 2003). With digital signatures, a small change to the original data results in a huge difference to the hashed message (digital signature). It is computationally impossible to create two different documents that have the same digest, so if one document is altered, then it would be impossible to create another document with the exact same digital signature. A digital signature would not reveal any information about the content of the BDPBB, only if the content has been altered (Accorsi 2009; Alles et al. 2004). With digital signatures, not only is the storage of Big Data provenance

FIGURE 2
Big Data Provenance Black Box Illustration
Modification of HadoopProv from Akoush et al. (2013)



records secure, but also this security is assured. Furthermore, in a related study of secure Hadoop (Park and Lee 2013), the authors established that encryption and decryption measures only added about 5 percent overhead to MapReduce jobs.

The BDPBB would be made available to appropriate regulators and auditors; however, even access and read, which are not active changes, will be recorded as part of the Big Data or document provenance. This BDPBB takes advantage of the digitization of the firm and the capacities of its enterprise resource planning (ERP) system at little additional cost (Park and Lee 2013). The provenance of the Big Data is maintained securely with the Black Box concept in a provenance-enabled Hadoop platform, such as HadoopProv, as mentioned earlier in the "Hadoop/MapReduce and the Cloud" section and shown in Figure 2. Provenance is captured at multiple points, as indicated in both Map and Reduce, and is recorded at Map Prov File and Reduce Prov File.

The creators of HadoopProv suggested that the security of their provenance information is an area for future research (Akoush et al. 2013). HadoopProv was conceived as an open source template that could be modified by others as needed. Provenance is captured and securely stored at the two separate phases of Map and Reduce, with the secure provenance graph construction occurring later. This paper amends HadoopProv by suggesting that the provenance information be recorded as digital signatures and stored in a digital Black Box.

Evidence Collection with BDPBB and the Audit Standards Revisited

Businesses and their IT systems are becoming increasingly more complex and are constantly evolving, forcing the audit profession to constantly adjust examination processes. One such complexity is the use of external Big Data by clients to improve effectiveness and efficiency of business analytics. The auditor should regard external Big Data with increased professional skepticism. The BDPBB may be regarded as one additional component in an integrated audit (ASB 2001, SAS 94), where the client is utilizing external Big Data and where the risk of insufficient competent evidence is greater. Thus, in the risk model of AR = IR × CR × DR, where audit risk (AR) is set low and inherent risk (IR) and control risk (CR) are assessed to calculate detection risk (DR), Big Data may significantly increase IR and CR. Detection Risk is the level of risk that the auditors could allow—high

FIGURE 3
Proposed DR Assessment for Each Data Type

Evidence Evaluation and Detection Risk Assessment			
Data Type	Secure Provenance Recorded or Available?	Missing Origins or Steps?	Preliminary Detection Risk Assessment of Data Type
Paper external:	Yes	Yes	low/medium
		No	high
	No	Yes	low
		No	low/medium
Paper internal:	Yes	Yes	medium
		No	high
	No	Yes	low
		No	low/medium
Electronic external:	Yes	Yes	low/medium
		No	high
	No	Yes	low
		No	low/medium
Electronic internal:	Yes	Yes	medium
		No	high
	No	Yes	low
		No	low/medium
Big Data external:	Yes	Yes	low/medium
		No	high
	No	Yes	low
		No	low
Big Data internal:	Yes	Yes	medium/low
		No	high
	No	Yes	low/medium
		No	low/medium

means that the auditor can afford less effective testing, and low means the auditor will need more effective testing. Inherent Risk could be assessed high if the Big Data is external and the business process required substantial client judgement. CR could be high if the Big Data originated outside the client and was stored in the Cloud. For high-risk IR and CR assertions and disclosures, the Big Data should be verified with fine-grained provenance, with coarse provenance reserved for less risky areas. If the provenance does not exist or is not in BDPBB format, then DR would be at a low level; see Figure 3.

In Figure 3, levels of DR for each data type are proposed, irrespective of whether the data are qualitative or quantitative. The lowest DR assessments for all data types exist when secure provenance does not exist for those data and there are gaps in either their origins or other intermediate steps. Transactions or data types that are slightly less risky are indicated as low/medium, and those that pose medium risk are indicated in the lightest shade. Data types that are high DR pose less risk of material misstatement to the auditor—the auditor, based on secure provenance of the data and their completeness, should be able to afford less effective testing.

The high-DR scenarios all assume that the client is recording fine and coarse provenance in a BDPBB format wherever and whenever external Big Data is acquired and that the client has agreed to secure and store this BDPBB outside its control for the benefit of auditors and regulators. Currently, this provenance recording may depend on the client's own assessment of its exposure to the risk of false information from external Big Data. However, businesses that have greater reliance on external Big Data may have a greater probability of being negatively impacted by faulty analyses derived from incompetent external Big Data.

Businesses such as Amazon, Twitter, Facebook, and several large banks and insurance companies have all experienced incidents due to faulty external Big Data social media and have responded with increased provenance collection efforts (Lin and Ryaboy 2013; Castillo et al. 2011). Twitter can currently collect provenance on reads and writes, but not the source control, due to the immensity of its data with its thousands of Hadoop nodes that process over 340 million tweets or 100 terabytes daily

(Lin and Ryaboy 2013)—an improved provenance, but not complete. This lack of provenance origin is troublesome, as Twitter has disclosed that fraudulent accounts and tweet spam could diminish its platform (Twitter 2014). Furthermore, 10 percent of Twitter's revenue originates from data licensing, where data "partners" are allowed to access, search, and analyze public tweets and their content (Twitter 2014).

However, as businesses rely more and more on external Big Data, it is hoped that the long-term issues presented by the four Vs (one of which is veracity, or provenance) will be successfully addressed by vendors, systems experts, and academics. Although businesses may be realizing short-term benefits from acquiring and analyzing external Big Data, eventually, the complexities presented by its four Vs should be addressed. Secure provenance collection and storage of external Big Data will, hopefully, become standard processes.

The BDPBB would appear to be somewhat computationally expensive at this time, based on the studies of HadoopProv, secure Hadoop, and digital signatures. It would seem that the more provenance tracking to be collected as BDPBB and the more fine this provenance, the more expensive the process. The actual application of the BDPBB (or a similar platform) is an area for future case study research regarding computational and monetary costs.

The second and fourth sections reviewed how a skeptical auditor should regard electronic evidence, log files, and IT controls. These conditions can now be addressed again with the perspective of the proposed BDPBB:

- Is the Big Data electronic evidence subject to alteration without an audit trail or evidence of this change?—The audit trail is securely recorded in BDPBB, where any alteration that occurs with the subject data is recorded and this recording is write-once, read-only.
- Is there an audit trail that clearly ties the Big Data digital evidence back to the initiating entry or data origin? Or can this trail lead forward to the point of inclusion on the face of the financial statements?—With recording of provenance flows, this trail is available.
- Does the Big Data electronic evidence include metadata that identifies who made the entry and when?—This metadata is now available from more points in the Hadoop process.
- What are the controls designed to prevent unauthorized changes to the Big Data digital evidence after it was created?—Evidence of IC compliance is available through process logs that have been securely recorded in BDPBB.
- Who has or had access rights to change the Big Data digital evidence?—Evidence of access rights compliance is available through additional metadata that is available in BDPBB.
- How does the auditor know that the Big Data digital evidence has not been intentionally altered?—This information is securely recorded in the BDPBB.
- Has the audit logging process been configured to record all access attempts, whether successful or not?—This information is securely recorded in the BDPBB.
- Have the audit logs been reviewed independently?—This control is maintained by limiting access to external auditors, internal auditors, and appropriate regulators.
- Has the continuity of logs been maintained and any gaps justified?—Any changes to the provenance logs are securely maintained in the BDPBB.
- Have the logs been frequently copied to offline, read-only media and stored in a separate secure location, inaccessible to those who might be motivated to change them?—The provenance logs are continually updated as read-only and stored separately as digital signatures in a secure location with limited access.
- Has the access to the logs and their security settings been recorded, and limited to only authorized persons?—The BDPBB records read-only information of all access attempts.

Additionally, the audit standards specify attributes for reliable evidence, which may now be revisited in the context of the BDPBB (see Table 4).

Many of the concerns about audit evidence in electronic environments may be satisfied with secure provenance of the datasets. Metadata, log files, and provenance graphs can be recorded and stored securely for reference by the auditor regarding the evidence characteristics. Secure provenance enables the auditor to ascertain whether the data has been altered, or whether the origins of the data have been accounted for. Using provenance information, the auditor may more confidently and accurately assess the level of risk that the data pose to certain business accounting judgements, processes, and assumptions.

CONCLUDING REMARKS

Big Data is now an important component of many businesses due to the rapid development of social media, sensors, and IoT, concurrent with increased data collection capabilities and storage capacity. Businesses or audit clients may be generating this Big Data internally or accessing it from external sources. Furthermore, these data have attributes of massive volume, high

TABLE 4

Evidence Characteristics of Paper, Electronic, and BDPBB Formats

Summary of Evidence Characteristics with BDPBB

Evidence Characteristics	Paper Evidence	Electronic Evidence	BDPBB Evidence
Alterability easily altered evidence lacks credibility; evidence should be difficult to alter	Difficult to alter without detection	Alterations may be difficult to detect without performing specifically designed tests	Alterations of the data are easy to detect and verify with BDPBB files
Prima Facie Credibility SAS 80 establishes a hierarchy of credibility—outside sources enhance credibility when independent of the client and confirmable	Outside sources of paper and documentary evidence and submitted directly to the auditor enhance credibility; inside sources of paper evidence that have been reviewed and processed by outsiders is also reliable	An electronic document derives its credibility primarily from the controls within the system. Outside electronic documentation/data is missing the assurance of system controls that the document or data is not fraudulent or altered	Outside sources are credible to the extent that their provenance has been securely recorded with the BDPBB. Auditors can readily determine the degree of veracity of the dataset based on its secure provenance
Completeness of Documents All essential terms of a transaction are verifiable	Typically, all essential terms are included on its surface in a text/human readable form	An electronic system may substitute codes or cross-references to other data files that may not be accessible	The BDPBB file is complete in that it will show what has been altered and where the transaction evidence is incomplete
Evidence of Approvals This essential aspect of internal controls should be easily verifiable and transparent	Approvals integrated into paper documentation add to completeness	Electronic approvals may be similarly integrated, but need additional verification	BDPBB data can record the approvals as metadata/coarse grained provenance
Ease of Use Simplicity of application and access encourages compliance	Paper evidence can usually be evaluated without the use of additional tools and/or skills	Electronic evidence may require extraction of data by an expert	BDPBB could be designed with a simple interface for auditor interaction/query
Clarity competent evidence should allow for the same re-performance and conclusions by other auditors	The nature of paper documentation is readily clear	The nature of electronic evidence is not always so clear, particularly in the absence of appropriate controls	BDPBB offers a straightforward recording of whether the provenance information has been altered or not and the entire lineage of the dataset that is possible to record

velocity, wide variety, and uncertain veracity (Zhang et al. 2015). These four Vs of Big Data persist as issues for entities attempting to unlock additional value from Big Data (CompTIA 2015). Basically, the Big Data trend may exhibit evidence of Amara's Law[5]—the tendency to overestimate the effects of a technology in the short run and underestimate the effects in the long run. The Big Data attribute of uncertain veracity is particularly troubling, as this challenges the requirement of reliable

[5] Amara's Law is the statement for which Dr. Roy Charles Amara, researcher and scientist, is well known: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run" (see: https://en.wikipedia.org/wiki/Roy_Amara).

competent audit evidence in the audit standards. Uncertain veracity in data means that the data lineage and transformations are not verifiable and not readily available. Lack of provenance in this instance equals unreliable data. Therefore, in a Big Data client environment, auditors may need to be more cognizant of secure data provenance.

The standards require that auditors ensure that the information generated through the client's system is reliable before the audit opinion is generated (Li et al. 2007; Alles et al. 2002; Elliott 1996). This requirement exists regardless of whether the auditor examines few (sampling) or all transactions (continuous monitoring). Furthermore, Sarbanes-Oxley requires that auditors verify that the management report regarding Internal Controls is accurate, and such auditor attestation requires re-performance of transactions and controls. In an electronic environment, the only "map" of a transaction or dataset may very well be the provenance record, also known as an audit trail. As Big Data increases in ubiquity of usage across businesses and industries, external auditors will be increasingly pressed to validate the reliability of this Big Data, particularly external Big Data and its attributes. This external Big Data may be messy, which clients may tolerate in the short term since the benefits of using the Big Data appear to outweigh the costs (Cukier and Mayer-Schoenberger 2013). However, SOX still requires management to provide auditable data, and auditors are not given the license, according to the current standards, to overlook the quality, reliability, and veracity of material audit evidence.

In this research, the BDPBB has been suggested as a possible means to provide secure data provenance of external Big Data that may serve as reliable audit evidence. Other solutions may exist that can address this issue—hopefully, this paper will stimulate more discussion about the secure provenance of Big Data for auditing. Basically, how should the truthfulness of the results of data analysis be validated by auditors when the data origins and/or permutations are unknown, as is often the case with MapReduce/Hadoop Big Data platforms? As such, without this provenance, Big Data that has been processed in MapReduce/Hadoop poses a huge risk as unreliable audit evidence when conducting audit examinations.

This paper poses a conceptual model of a BDPBB based on HadoopProv, which has been demonstrated to be the most cost- and workload-efficient of any Hadoop provenance collection application to date (Alabi et al. 2015; Akoush et al. 2013). However, HadoopProv was not proposed as a secure system, and has been modified as the BDPBB here. Subsequent application and demonstration of the BDPBB in a Hadoop Big Data environment is an area for future research and exploration. Efficiency performance of a secure provenance system in Hadoop should be evaluated, as should computational costs.

External Big Data that has been processed with Hadoop presents unique challenges of complexity and, possibly, high computational costs to the client and, subsequently, the auditing profession. In this context, to what extent should the auditing profession regard external Big Data as competent evidence and under what circumstances?

The Audit Standards should address the unique situation posed by Big Data: external evidence in the form of external Big Data may not be reliable unless secure data provenance of that data has been recorded.

Finally, internal auditors may have more exposure than public auditors to examinations of business decisions and observations that were generated from "messy" external Big Data that was processed with Hadoop. In a survey by the Institute of Internal Auditors (IIA), nearly half of the auditors had little or no involvement with data quality evaluation, despite the fact that 23 percent of them had only slight or no confidence in that quality (Tysiac 2016). Perhaps the genesis of a solution that addresses the challenges of external Big Data audit evidence could occur initially within the internal auditing profession.

This paper has contributed to the discussion regarding secure data provenance in the Big Data environment from a public auditing context. As businesses proceed to embrace Big Data and its potential for impactful and insightful analytics, this complex challenge of scant Big Data provenance and the subsequent erosion of evidence reliability should not be ignored by the audit profession, regulators, and academics.

REFERENCES

Accorsi, R. 2006. On the relationship of privacy and secure remote logging in dynamic systems. In *Security and Privacy in Dynamic Environments,* 329–339. New York, NY: Springer U.S.

Accorsi, R. 2009. Safe-keeping digital evidence with secure logging protocols: State of the art and challenges. In *Proceedings of the Fifth International Conference on IT Security Incident Management and IT Forensics,* 94–110. IEEE. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5277863&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F5277833%2F5277834%2F05277863.pdf%3Farnumber%3D5277863

Aggarwal, C. C. 2009. Trio: A system for data uncertainty and lineage. In *Managing and Mining Uncertain Data*, 1–35. New York, NY: Springer.

Akoush, S., R. Sohan, and A. Hopper. 2013. HadoopProv: Towards provenance as a first class citizen in MapReduce. In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance*, 11–14. Berkeley, CA: USENIX Association.

Alabi, O., J. Beckman, M. Dark, and J. Springer. 2015. Toward a data spillage prevention process in Hadoop using data provenance. In *Proceedings of the 2015 Workshop on Changing Landscapes in HPC Security,* 9–13. ACM. Available at: http://dl.acm.org/citation.cfm?id=2752502

Aldeco-Pérez, R., and L. Moreau. 2010. Securing provenance-based audits. In *Provenance and Annotation of Data and Processes,* 148–164. Berlin/Heidelberg, Germany: Springer.

Alles, M. G., A. Kogan, and M. A. Vasarhelyi. 2002. Feasibility and economics of continuous assurance. *Auditing: A Journal of Practice & Theory* 21(1): 125–138.

Alles, M. G., A. Kogan, and M. A. Vasarhelyi. 2004. Restoring auditor credibility: Tertiary monitoring and logging of continuous assurance systems. *International Journal of Accounting Information Systems* 5 (2):183–202.

American Institute of Certified Public Accountants (AICPA). 2012. *Audit Evidence*. Statement on Auditing Standards No. 122, AU-C Section 500. New York, NY: AICPA.

Armbrust, M., A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. 2010. A view of cloud computing. *Communications of the ACM* 53 (4): 50–58.

Assunção, M. D., R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya. 2015. Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing* 79: 3–15.

Auditing Standards Board (ASB). 1996. *Amendment to SAS 31, Evidential Matter*. Statement on Auditing Standards No. 80. New York, NY: ASB.

Auditing Standards Board (ASB). 2001. *The Effect of Information Technology on the Auditor's Consideration of Internal Control in a Financial Statement Audit*. Statement on Auditing Standards No. 94 (Amends Statement on Auditing Standards No. 55, *Consideration of Internal Control in a Financial Statement Audit*). New York, NY: ASB.

Bao, Z., S. Cohen-Boulakia, S. B. Davidson, A. Eyal, and S. Khanna. 2009. Differencing provenance in scientific workflows. In *25th International Conference on Data Engineering,* 808–819. IEEE. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4812456&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4812456

Bates, A., B. Mood, M. Valafar, and K. Butler. 2013. Towards secure provenance-based access control in cloud environments. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy,* 277–284. ACM. Available at: http://dl.acm.org/citation.cfm?id=2435389

Bauer, S., and D. Schreckling. 2013. Data provenance in the Internet of things. Presented at EU Project COMPOSE, Conference Seminar. Available at: https://web.sec.uni-passau.de/projects/compose/papers/Bauer_Data_Provenance_in_the_Internet_of_Things.pdf

Bier, C. 2013. How usage control and provenance tracking get together—A data protection perspective. In *Proceedings of Security and Privacy Workshops* (*SPW*), 13–17. IEEE. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6565222

Bishop, M. 2006. *Introduction to Computer Security*. Upper Saddle River, NJ: Pearson Education.

Bollen, J., H. Mao, and X. Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2 (1): 1–8.

Bose, R., and J. Frew. 2005. Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys (CSUR)* 37 (1): 1–28.

Braun, U., A. Shinnar, and M. I. Seltzer. 2008. Securing provenance. In *Proceedings of the 3rd Conference on Hot Topics in Security* (HOTSEC '08). Available at: http://dl.acm.org/citation.cfm?id=1496675

Brown, B., M. Chui, and J. Manyika. 2011. Are you ready for the era of "Big Data"? *McKinsey Quarterly* 4: 24–35.

Brown-Liburd, H., and M. A. Vasarhelyi. 2015. Big Data and audit evidence. *Journal of Emerging Technologies in Accounting* 12 (1): 1–16.

Brynjolfsson, E., J. Hammerbacher, and B. Stevens. 2011. Competing through data: Three experts offer their game plans. *McKinsey Quarterly* 4: 36–47.

Buneman, P., S. Khanna, and W. C. Tan. 2000. Data provenance: Some basic issues. In *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science,* 87–93. Berlin/Heidelberg, Germany: Springer.

Buneman, P., S. Khanna, and W. C. Tan. 2001. Why and where: A characterization of data provenance. In *Database Theory—ICDT 2001,* 316–330. Berlin/Heidelberg, Germany: Springer.

Buneman, P., and W. C. Tan. 2007. Provenance in databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of data,* 1171–1173.

Buneman, P., and S. B. Davidson. 2010. *Data Provenance—The Foundation of Data Quality*. Available at: http://www.sei.cmu.edu/measurement/research/upload/Davidson.pdf

Cameron, G. 2003. Provenance and pragmatics. In *Workshop on Data Provenance and Annotation, Edinburgh*. Available at: http://citeseerx.ist.psu.edu/showciting;jsessionid=6A9559F7C488770E91D02D9244EA34D5?cid=3621642

Caster, P., and D. Verardo. 2007. Technology changes the form and competence of audit evidence. *CPA Journal* 77 (1): 68–70.

Castillo, C., M. Mendoza, and B. Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web,* 675–684. ACM. Available at: http://dl.acm.org/citation.cfm?id=1963500

Cerullo, M. V., and M. J. Cerullo. 2003. Impact of SAS No. 94 on computer audit techniques. *Information Systems Control Journal* 1: 53–58.

Che, D., M. Safran, and Z. Peng. 2013. From Big Data to Big Data mining: Challenges, issues, and opportunities. In *Database Systems for Advanced Applications,* 1–15. Berlin/Heidelberg, Germany: Springer.

Cheah, Y. W., and B. Plale. 2012. Provenance analysis: Towards quality provenance. In *IEEE 8th International Conference on E-Science (e-Science),* 1–8. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6404480&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6404480

Chen, P., and B. A. Plale. 2015. Big Data provenance analysis and visualization. In *15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid),* 797–800. IEEE. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7152560&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7152560

Cheney, J., L. Chiticariu, and W. C. Tan. 2009. Provenance in databases: Why, how, and where. *Foundations and Trends® in Databases* 1 (4): 379–474.

Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia. 2012. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing* 9 (6): 811–824.

Cohen, J. C., and S. Acharya. 2014. Towards a trusted HDFS storage platform: Mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection. *Journal of Information Security and Applications* 19 (3): 224–244.

Cohen-Boulakia, S., O. Biton, S. Cohen, and S. Davidson. 2008. Addressing the provenance challenge using ZOOM. *Concurrency and Computation: Practice and Experience* 20 (5): 497–506.

Colbert, J. L., and J. Smalling. 2011. EDI and the financial auditor. *Review of Business Information Systems* 2 (4): 9–16.

CompTIA. 2015. *Big Data Insights and Opportunities* Research Report (November). Available at: https://www.comptia.org/resources/big-data-insights-and-opportunities-2015

Crawl, D., J. Wang, and I. Altintas. 2011. Provenance for MapReduce-based data-intensive workflows. In *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science,* 21–30. Available at: http://dl.acm.org/citation.cfm?id=2110501

Cresci, S., R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems* 80: 56–71.

Cui, Y., and J. Widom. 2003. Lineage tracing for general data warehouse transformations. *VLDB Journal* 12 (1): 41–58.

Cukier, K., and V. Mayer-Schoenberger. 2013. The rise of Big Data: How it's changing the way we think about the world. *Foreign Affairs* 92 (3): 28–40.

Davidson, S. B., S. Cohen-Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. Kumar Anand, and J. Freire. 2007. Provenance in scientific workflow systems. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 30 (4): 44–50.

Davidson, S. B., and J. Freire. 2008. Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* 1345–1350. Available at: http://dl.acm.org/citation.cfm?id=1376772

Dean, J., and S. Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Communications of the ACM* 51 (1): 107–113.

Duncan, M. 2015. *Future Casting Influence Capability in Online Social Networks*. Available at: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc195/p802328_A1b.pdf

Elliott, R. K. 1996. Assurance service opportunities: Implications for academia. *Accounting Horizons* 11 (4): 61–74.

Elliott, R. K. 2002. Twenty-first century assurance. *Auditing: A Journal of Practice & Theory* 21 (1): 139–146.

Ferraiolo, D. F., and D. R. Kuhn. 2009. *Role-Based Access Controls*. Available at: http://arxiv.org/abs/0903.2171

Freire, J., D. Koop, E. Santos, and C. T. Silva. 2008. Provenance for computational tasks: A survey. *Computing in Science and Engineering* 10 (3):11–21.

Frew, J., D. Metzger, and P. Slaughter. 2008. Automatic capture and reconstruction of computational provenance. *Concurrency and Computation: Practice and Experience* 20 (5): 485–496.

Galhardas, H., D. Florescu, D. Shasha, E. Simon, and C. A. Saita. 2001. Improving data cleaning quality using a data lineage facility. In *Proceedings of Design and Management of Data Warehouses,* 3. Available at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.8584

Ghoshal, D., and B. Plale. 2013. Provenance from log files: A BigData problem. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops,* 290–297. Available at: http://dl.acm.org/citation.cfm?id=2457366

Glavic, B., and K. R. Dittrich. 2007. Data provenance: A categorization of existing approaches. *BTW* 103: 227–241.

Glavic, B. 2014. Big Data provenance: Challenges and implications for benchmarking. In *Specifying Big Data Benchmarks,* 72–80. Available at: http://link.springer.com/chapter/10.1007%2F978-3-642-53974-9_7

Goble, C. 2002. Position statement: Musings on provenance, workflow and (semantic web) annotations for bioinformatics. In *Workshop on Data Derivation and Provenance, Chicago*, Vol. 3. Available at: http://www.ipaw.info/chicago02/papers/provenance_workshop_3.doc

Hammerbacher, J. 2009. Information platforms and the rise of the data scientist. In *Beautiful Data: The Stories Behind Elegant Data Solutions,* 73–84. Sebastopol, CA: O'Reilly Media.

Hasan, R., R. Sion, and M. Winslett. 2009. The case of the fake Picasso: Preventing history forgery with secure provenance. In *Proceedings of the 7th Conference on File and Storage Technologies* 9: 1–14. Available at: http://dl.acm.org/citation.cfm?id=1525909

Hey, T., and A. E. Trefethen. 2003. The data deluge: An e-science perspective. In *Grid Computing: Making the Global Infrastructure a Reality,* edited by Berman, F., Fox, G. and T. Hey, 809–824. Chichester, U.K.: John Wiley & Sons, Ltd.

Hughes, A. L., and L. Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management* 6 (3/4): 248–260.

IBM. 2012. *The Four V's of Big Data*. Available at: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

IBM. 2015. *Clearing Out Digital Debris with Informational Governance*. Available at: http://www.bankinfosecurity.com/whitepapers/clearing-out-digital-debris-information-governance-w-1795

Ikeda, R., and J. Widom. 2010. Panda: A system for provenance and data. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 33 (3): 42–49.

Ikeda, R., H. Park, and J. Widom. 2011. *Provenance for Generalized Map and Reduce Workflows*. Available at: http://ilpubs.stanford.edu:8090/985/2/cidr_prov_camera2.pdf

Imran, A., R. Agrawal, J. Walker, and A. Gomes. 2014. A layer based architecture for provenance in Big Data. In *Proceedings of the 2014 IEEE International Conference on Big Data (Big Data),* 29–31. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7004468&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7004468.

International Auditing and Assurance Standards Board (IAASB). 2009. *Audit Evidence*. International Standard of Auditing No. 500. New York, NY: IAASB.

Jagadish, H. V., J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. 2014. Big Data and its technical challenges. *Communications of the ACM* 57 (7): 86–94.

Jans, M. J., M. Alles, and M. A. Vasarhelyi. 2010. *Process Mining of Event Logs in Auditing: Opportunities and Challenges*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1578912

Laney, D. 2001. *3D Data Management: Controlling Data Volume, Velocity and Variety*. META Group, Research Note 6. Available at: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Li, Y., J. N. Roge, L. Rydl, and J. Hughes. 2007. Achieving Sarbanes-Oxley compliance with XBRL-based ERP and continuous auditing. *Issues in Information Systems* 8 (2): 430–436.

Liao, C., and A. Squicciarini. 2015. Towards provenance-based anomaly detection in MapReduce. In *Proceedings of the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid),* 647–656. IEEE. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7152530&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7152530

Lin, J., and D. Ryaboy. 2013. Scaling Big Data mining infrastructure: The Twitter experience. *ACM SIGKDD Explorations Newsletter* 14 (2): 6–19.

Margo, D., and R. Smogor. 2010. Using provenance to extract semantic file attributes. In *Proceedings of the 2nd Conference on Theory and Practice of Provenance,* Vol. 7. Available at: https://www.usenix.org/legacy/event/tapp10/tech/full_papers/margo.pdf

McDaniel, P., K. Butler, S. McLaughlin, R. Sion, E. Zadok, and M. Winslett. 2010. Towards a secure and efficient system for end-to-end provenance. In *Proceedings of the 2nd Conference on Theory and Practice of Provenance*. Available at: http://dl.acm.org/citation.cfm?id=1855797

Mittal, A. 2013. Trustworthiness of Big Data. *International Journal of Computer Applications* 80 (9).

Moreau, L., P. Groth, S. Miles, J. Vazquez-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan, and L. Varga. 2008a. The provenance of electronic data. *Communications of the ACM* 51 (4): 52–58.

Moreau, L., B. Ludäscher, I. Altintas, R. S. Barga, S. Bowers, S. Callahan, et al. 2008b. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience* 20 (5): 409–418.

Moreau, L., J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson. 2008c. The open provenance model: An overview. In *Provenance and Annotation of Data and Processes,* 323–326. Berlin/Heidelberg, Germany: Springer.

Muniswamy-Reddy, K. K., D. A. Holland, U. Braun, and M. I. Seltzer. 2006. Provenance-aware storage systems. In *Proceedings of the USENIX Annual Technical Conference, General Track,* 43–56. Available at: https://www.usenix.org/legacy/event/usenix06/tech/full_papers/muniswamy-reddy/muniswamy-reddy_html/

Muniswamy-Reddy, K. K., U. Braun, D. A. Holland, P. Macko, D. Maclean, D. Margo, M. Seltzer, and R. Smogor. 2009. Layering in provenance systems. In *Proceedings of the 2009 USENIX Annual Technical Conference*. Available at: https://www.usenix.org/conference/usenix-09/layering-provenance-systems

Muniswamy-Reddy, K. K., P. Macko, and M. I. Seltzer. 2010. Provenance for the cloud. In *Proceedings of the 8th USENIX Conference on File and Storage* 10: 15–14. Available at: http://dl.acm.org/citation.cfm?id=1855526

Nearon, B. H. 2005. Foundations in auditing and digital evidence. *CPA Journal* 75 (1): 32.

O'Driscoll, A., J. Daugelaite, and R. D. Sleator. 2013. "Big Data," Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics* 46 (5): 774–781.

Olavsrud, T. 2016. *21 Data and Analytic Trends that Will Dominate 2016*. Available at: http://www.cio.com/article/3023838/analytics/21-data-and-analytics-trends-that-will-dominate-2016.html

Oppliger, R., and R. Rytz. 2003. Digital evidence: Dream and reality. *IEEE Security and Privacy* 1 (5): 44–48.

Park, H., R. Ikeda, and J. Widom. 2011. *RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows*. Available at: http://ilpubs.stanford.edu:8090/995/1/ramp-demo.pdf

Park, S., and Y. Lee. 2013. Secure Hadoop with encrypted HDFS. In *Grid and Pervasive Computing,* 134–141. Berlin/Heidelberg, Germany: Springer.

Patil, D. J. 2012. *Data Jujitsu: The Art of Turning Data into Product*. Sebastopol, CA: O'Reilly Media, Inc.

Polato, I., R. Ré, A. Goldman, and F. Kon. 2014. A comprehensive view of Hadoop research—A systematic literature review. *Journal of Network and Computer Applications* 46: 1–25.

Public Company Accounting Oversight Board (PCAOB). 2010. *Audit Evidence*. Auditing Standard No. 15. Washington, DC: PCAOB.

Ratcliffe, T. A., and P. Munter. 2002. Information technology, internal control, and financial statement audits. *CPA Journal* 72 (4): 40.

Sakka, M. A., B. Defude, and J. Tellez. 2010. Document provenance in the cloud: Constraints and challenges. In *Networked Services and Applications-Engineering, Control and Management,* 107–117. Berlin/Heidelberg, Germany: Springer.

Scheidegger, C., D. Koop, E. Santos, H. Vo, S. Callahan, J. Freire, and C. Silva. 2008. Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience* 20 (5): 473–483.

Simmhan, Y. L., B. Plale, and D. Gannon. 2005a. *A Survey of Data Provenance Techniques*. Computer Science Department, Indiana University. Available at: http://www.cs.indiana.edu/cgi-bin/techreports/TRNNN.cgi?trnum=TR618

Simmhan, Y. L., B. Plale, and D. Gannon. 2005b. A survey of data provenance in e-science. *ACM SIGMOD Record* 34 (3): 31–36.

Souiah, I., A. Francalanza, and V. Sassone. 2009. A formal model of provenance in distributed systems. In *Workshop on the Theory and Practice of Provenance,* 1–11. Available at: http://eprints.soton.ac.uk/268600/

Stamatogiannakis, M., P. Groth, and H. Bos. 2014. Looking inside the black-box: Capturing data provenance using dynamic instrumentation. In *Provenance and Annotation of Data and Processes,* 155–167. Cham, Switzerland: Springer International Publishing.

Tan, W. C. 2004. Research problems in data provenance. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 27 (4): 45–52.

Tan, W. C. 2007. Provenance in databases: Past, current, and future. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 30 (4): 3–12.

Taylor, M., J. Haggerty, D. Gresty, and R. Hegarty. 2010. Digital evidence in cloud computing systems. *Computer Law and Security Review* 26 (3): 304–308.

Thomas, K., C. Grier, D. Song, and V. Paxson. 2011. Suspended accounts in retrospect: An analysis of Twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference,* 243–258. ACM. Available at: http://dl.acm.org/citation.cfm?id=2068840

Tsai, W. T., X. Wei, Y. Chen, R. Paul, J. Y. Chung, and D. Zhang. 2007. Data provenance in SOA: Security, reliability, and integrity. *Service Oriented Computing and Applications* 1 (4): 223–247.

Twitter. 2014. *Annual Report 2014*. Available at: http://www.viewproxy.com/twitter/2015/1/annualreport2014.pdf

Tysiac, K. 2016. Internal auditors challenged by cybersecurity, data quality. *Journal of Accountancy* (February 16). Available at: http://www.journalofaccountancy.com/news/2016/feb/internal-audit-challenges-201613894.html

U.S. House of Representatives. 2002. The Sarbanes-Oxley Act of 2002. Public Law 107-204 [H.R. 3763]. Washington, DC: Government Printing Office.

van der Aalst, W. M., K. M. van Hee, J. M. van Werf, and M. Verdonk. 2010. Auditing 2.0: Using process mining to support tomorrow's auditor. *Computer* 43 (3): 90–93.

Vaughan, J. A., L. Jia, K. Mazurak, and S. Zdancewic. 2008. Evidence-based audit. In *2008 21st IEEE Computer Security Foundations Symposium,* 177–191. Available at: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4556686&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4556686

Warren, J. D., Jr, K. C. Moffitt, and P. Byrnes. 2015. How Big Data will change accounting. *Accounting Horizons* 29 (2): 397–407.

Weitzner, D. J., H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman. 2008. Information accountability. *Communications of the ACM* 51 (6): 82–87.

Zhang, J., X. Yang, and D. Appelbaum. 2015. Toward effective Big Data analysis in continuous auditing. *Accounting Horizons* 29 (2): 469–476.

Zikopoulos, P., and C. Eaton. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY: McGraw-Hill Osborne Media.