Accounting Horizons
Vol. 29, No. 2
2015
pp. 469–476

American Accounting Association
DOI: 10.2308/acch-51070

Toward Effective Big Data Analysis in Continuous Auditing

Juan Zhang, Xiongsheng Yang, and Deniz Appelbaum

SYNOPSIS: Big Data now pervades every sector and function of the global economy. This paper focuses on the gaps between Big Data and the current capabilities of data analysis in continuous auditing (CA). It identifies four dimensions of Big Data and five subsequent gaps: namely, data consistency, integrity, aggregation, identification, and confidentiality. For each gap, the paper outlines challenges and possible solutions derived from traditional data systems, which can be further applied to CA systems in an era of Big Data.

Keywords: Big Data; gap analysis; continuous auditing.

INTRODUCTION

Big Data originates from traditional transactions systems, as well as new sources such as emails, phone calls, Internet activities, social media, news media, sensor recordings and videos, and RFID tags. Since much of this Big Data informs and affects corporate decisions that are important to both internal and external corporate stakeholders, auditors will need to expand their current scope of data analysis (Cao, Chychyla, and Stewart 2015).

Certain qualities, known as the four Vs, define the term Big Data: namely, massive Volume or size of the database, high Velocity of data added on a continuous basis, large Variety of types of data, and uncertain Veracity (Laney 2001; IBM 2012). Due to volume and velocity, the application of continuous auditing (CA) has become increasingly relevant for the automation and real-time analysis of Big Data (Vasarhelyi, Alles, and Williams 2010). However, massive volume and high velocity also introduce gaps between the present state of audit analytics and the requirements of Big Data analytics in a continuous audit context. Moreover, variety and uncertain veracity present challenges beyond the capability of current CA methods. The purpose of this paper is to identify these gaps and challenges and to point out the need for updating the CA system to accommodate Big Data analysis.

Juan Zhang is an Associate Professor and Xiongsheng Yang is a Professor, both at Nanjing University, and Deniz Appelbaum is a Ph.D. student at Rutgers, The State University of New Jersey, Newark.

We thank editors, two anonymous referees, and participants at the 32nd World Continuous Auditing and Reporting Symposium and workshops at Rutgers, The State University of New Jersey for their helpful comments. This paper is supported by the National Natural Science Foundation of China (NFSC-71402070).

Submitted: February 2015
Accepted: February 2015
Published Online: February 2015
Corresponding author: Juan Zhang
Email: zjtider@nju.edu.cn

Big Data and Transforming the Continuous Audit

A survey by the Institute of Internal Auditors (IIA) states "the push for real-time financial reporting and the drive to automate resource draining manual audits are nudging organizations to adopt continuous auditing now" (Brannen 2006). Continuous auditing is "to provide assurance simultaneously with, or a short period of time after, the occurrence of events underlying the subject matter" (CICA/AICPA 1999) by automatically identifying exceptions or misstatements as defined by some prespecified criteria in the embedded audit modules (Groomer and Murthy 1989) or a monitoring control layer (Vasarhelyi, Alles, and Kogan 2004). With the push of Big Data, CA is needed to access and process much additional relevant information due to the vastly increased volumes of data and transactions. Auditors with competence in data analytics will have better opportunities to widen the range and increase the scale of auditing on a more frequent basis via continuous data monitoring (CDM), continuous control monitoring (CCM), and continuous risk monitoring and assessment (CRMA) (Vasarhelyi et al. 2010).

However, one of the most important questions looming in audit departments is how to effectively deliver value from increasingly expanding Big Data repositories. For example, a torrential flow of Big Data from multiple sources serves no potential benefit if there is no connection between it. Such a connection is heavily relied upon by CA tools and data analytics to identify anomalies and misstatements automatically and in a timely fashion. Considering the four Vs of Big Data, auditors should realize that there are gaps between Big Data and the capabilities of modern CA systems, as shown in Figure 1. Specifically, the original three Vs of huge Volume, high Velocity, and huge Variety introduce the gaps of data consistency, data identification, and data aggregation to link databases in a CA system. Veracity has been added to the original three Vs of Big Data, as the integrity of the information becomes an issue without secure log files or data provenance. Meanwhile, as Big Data becomes an important source for analytics, its confidentiality is another source of concern for both corporate and CA systems. Each of the gaps creates the corresponding challenges in Figure 1, which we will discuss further in the following parts of the paper.

Since the gaps and challenges will become common for audit process of Big Data, the feasible solutions beg for an updated application of CA. The main elements of the continuous auditing architecture, such as data provisioning, data filtering, and the data diagnostic layer, must be adapted to accommodate the challenges presented by Big Data (Vasarhelyi et al. 2004).

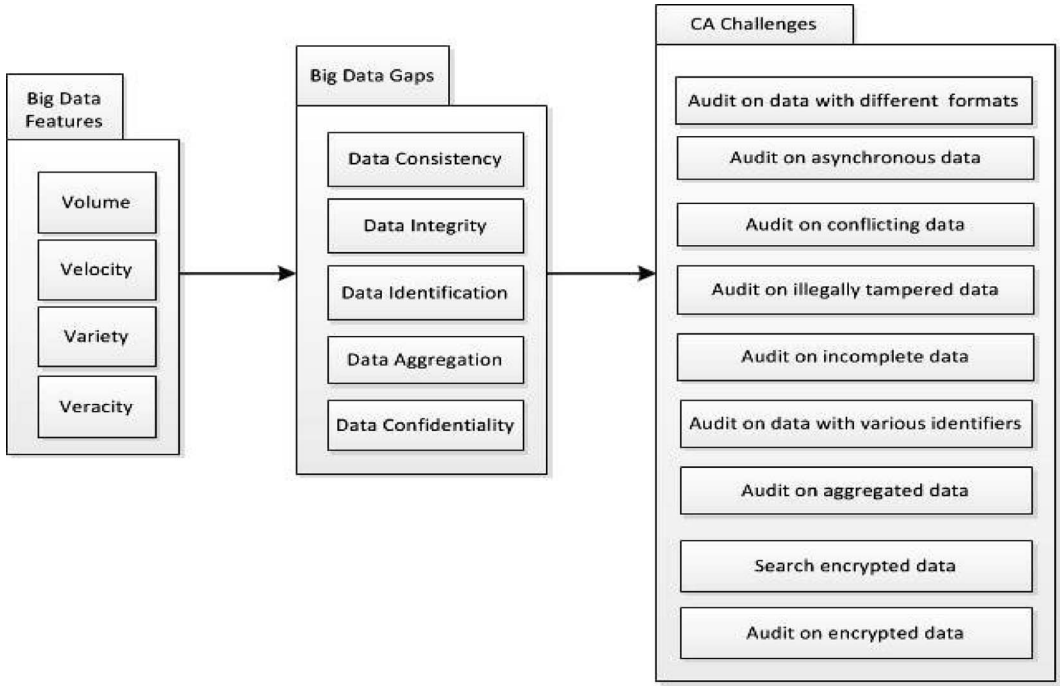
Big Data Gaps and Challenges to Continuous Auditing

Data Consistency

Data consistency is the most important issue for CA of Big Data and relates to interdependent data between applications and across an organization (Sheth and Rusinkiewicz 1990). Since there are an increasing number of different data sources in today's era of Big Data, the frequency of data conflicts has inevitably increased. The main cause here is that Big Data systems supporting key business processes usually consist of a patchwork of different systems, where data may be fully or partially replicated, the informational content may be overlapped, and more derived data may be stored. This situation gives rise to the serious gap in data consistency. For example, data may be mapped and reduced in a Hadoop® platform.[1] Therefore, the new CA approach needs to verify the relationship among data sources and to check data consistency.

[1] Hadoop has been the most efficient approach to date for dealing with the massive streams of complex and unstructured data (Crawl, Wang, and Altintas 2011). Almost all Big Data is collected and aggregated in a Hadoop (MapReduce) system where the incoming data are mapped according to their elements, and then these attributes are reduced equally across available nodes (Akoush, Sohan, and Hopper 2013; Dean and Ghemawat 2008). Basically, the Map/Shuffle function breaks down large unstructured multidimensional input data into smaller data via keys and values. These resulting key-value pairs are then exchanged and sorted via a Reduce function that combines similar data values in a parallel fashion across several output nodes. The Reduce/Sort function simplifies the data into more structured and meaningful parts for the end user. Unfortunately, this treatment, while efficient for organizing Big Data, may result in data inconsistency and questionable provenance that is not being logged in most Hadoop systems (Cheah and Plale 2012). Developers are just now beginning to address this lack of data provenance in MapReduce systems.

FIGURE 1
Effective Big Data Analytics in Continuous Auditing



There are at least three types of data inconsistencies that should be managed by a CA system. The first inconsistency occurs in *data formats,* where data from different sources may vary in terms of structure. Interestingly, Koudas, Saha, Srivastava, and Venkatasubramanian (2009) developed a model to resolve this type of inconsistency, where varying data formats can be related based on their underlying dependencies (i.e., relations between attributes in databases). For example, by including the exchange rate into the dependency model, the CA system can automatically relate the sales of 1 million U.S. dollars recorded by headquarters located in the U.S. with the sales of 6.2 million RMB in the database of subsidiaries in China. The second issue is *data synchronization,* which ensures that data are in sync across the entire organization. One example of such synchronization was illustrated by Golab, Karloff, Korn, Saha, and Srivastava (2009). In this study, underlying sequential dependencies (SD) derived from seemingly disparate data were used to construct the relationships and synchronicity with real-world data. The third and perhaps the most serious inconsistency is that of *data contradiction:* data from one source may contradict data from another source. Chiang and Miller (2008) provided an algorithm based on conditional functional dependencies to address this form of inconsistency. Plus, there are additional research studies on data consistency evaluations that can also be used to update the CA system. For example, Fan and Geerts (2010) and Fan, Geerts, and Wijsen (2012) designed a data consistency check system in a storage environment where the data were separated, and Fuxman and Miller (2007) used a data query method to repair simple inconsistent data.

Even if we have the experience of applying the above-mentioned techniques to the three issues of data inconsistency in data warehousing, how to integrate those techniques into a CA system of Big Data without it losing its efficiency is worthy of future research. At the same time, to

fundamentally avoid the low-efficiency issue, a universal audit data format standard for CA may be a potential solution to address the issue of data format inconsistency. Moreover, we need further discussion about how this universal data standard could include additional data attributes, such as data sequential dependency and data source, in an effort to minimize these asynchrony and contradiction issues.

Data Integrity

Data integrity is a term broad in scope. In the auditing realm, the lack of data integrity usually relates to data tampering (Menezes, Oorschot, and Vanstone 1996) and incomplete data (Motro 1989). In the Big Data environment, the volume and types of data are so expansive that it becomes more difficult to identify individual data as well as data sets that have been modified/deleted/hidden/destroyed because of operating error, procedural error, illegal access, and/or network transmission failures. This difficulty in identifying integrity issues can create a domino effect that causes other reliable data to lose their value for audit analysis purposes, thus increasing audit risk in a Big Data, continuous audit environment. Traditional methods of verifying data integrity, such as reasonability, edit checks, and comparison with other sources, may not currently be practical for Big Data audit applications. Approaches for repairing data integrity problems must be developed that not only enhance the real-time and automated audit, but also improve the effectiveness of management's continuous monitoring routines.

There are two types of Big Data integrity issues that should be addressed. The first is *data modification*. Data modification may refer to the unintentional modification that can occur with such Big Data aggregation systems, such as Hadoop MapReduce, as discussed earlier. Or, with the increased number of individuals involved with producing, processing, and transforming data, the probability of intentional data modification is increasing. Message digest techniques provide an efficient way to test modified data, especially key data, and Menezes et al. (1996) have summarized a variety of message digest algorithms that can be used in continuous auditing. The second integrity issue is *incomplete data,* which can occur when an employee enters partial records of unfinished transactions or records with missing fields into an enterprise database. The classic methods of dealing with incomplete data are listwise deletion and mean substitution, but neither of them is particularly effective (Schafer and Graham 2002). Therefore, constructive ways such as estimation and imputation are proposed. For example, Motro (1989) offered a scheme for searching complete query results from an incomplete database. The algorithm is based on the framework that Integrity = Validity + Completeness. In addition, Mayfield, Neville, and Prabhakar (2010) proposed a probability model named ERACER for estimating the value and effect of incomplete data in a system.

Although the above incomplete-data repairing techniques are effective, they are not efficient when processing a large amount of data. Therefore, new and highly efficient data repairing techniques are imperative. At the same time, new analytical paradigms for auditing Big Data containing incomplete information are required so that the inefficient data repairing process can be avoided.

Data Identification

Data identification refers to records that link two or more separately recorded pieces of information about the same individual or entity (Newcombe, Kennedy, Axford, and James 1959). When data are structured, identification is easy. However, in the Big Data audit where much of the data might be unstructured, identification becomes difficult. For example, the revenue amount for a given sale can be easily identified by the continuous auditing system, but it can be challenging to

automatically connect this information with the associated sales terms and conditions, which are in an unstructured textual format.

There are many innovative methods to address the data identification issue, and their suitability for the continuous auditing of Big Data should be examined. Those methods can be generally summarized in two ways: one is based on *semantic rules* and the other is based on *similarity measures*. As for the former, it is often based on the experience and knowledge needed to set the identification rules. Lim, Srivastava, Prabhakar, and Richardson (1996), and Hernandez and Stolfo (1998) proposed simple methods to determine the equivalency of data points using key identifying attributes stored as instances. Moreover, Arasu, Chaudhuri, and Kaushik (2008) presented a transformation-based framework to capture variations and general forms of the data. Fan, Jia, Li, and Ma (2009) deduced alternative attributes of the data to develop matching data identities. Getoor (2010) introduced the notion of graph identification to label the data that are ambiguous. As for the similarity measures, there are fuzzy matches based on the Levenshtein distance and other similar algorithms, which can be applied in a CA system. The greater the Levenshtein distance, the more different the two entities are. Ananthakrishna, Chaudhuri, and Ganti (2002) proposed another solution by measuring the simultaneous frequency of A and B with C. For example, the same IP address and login/logout time can be used as features to identify the activities of one staff member whose login name is Tom for one data source, while it is John for another data source.

Besides the above data identification techniques, developers have proposed numerous identification techniques for specified scenarios that make it necessary to review and categorize those techniques so as to offer a CA system with the applicable menu. Moreover, most of the current techniques just provide probabilistic measures on data identification, which will in turn affect the audit risk of CA system. And this influence should be further researched.

Data Aggregation

In the Big Data context, the large volume of data flowing to the CA system may incur excessive computing resource demands. Furthermore, data sets with large variances may lead to unstable audit analysis models that could either trigger too many alarms or experience deflated detection power (Alles, Brennan, Kogan, and Vasarhelyi 2006; Perols and Murthy 2012). Thus, data aggregation is necessary for the normal operation of continuous auditing using Big Data and to meaningfully summarize and simplify the Big Data that is most likely coming from different sources.

Existing continuous auditing architectures focus not only on the aggregation of raw data for collecting exception data, but also on the aggregation and analysis of exception data itself. Kogan, Alles, Vasarhelyi, and Wu (2014) proposed that aggregation of exception data based on raw Big Data could facilitate the identification of general patterns over a period of time. For example, the number of transactions can differ greatly between individual transactions, as well as the lag times between order and delivery, and delivery and payment. By aggregating the individual transactions, this variance can be significantly reduced, allowing more effective detection of material anomalies. Alles et al. (2006) suggested that although a CA system automatically generates alarms for critical exceptions, such as individual accounts without passwords, these alarms should also be aggregated to pinpoint weaknesses in certain control areas (e.g., segregation of duties). Perols and Murthy (2012) suggested a layered framework that aggregates alarms rather than the raw data. Basically, the exceptions detected in the monitoring layer are grouped in the aggregation layer according to their association to specific auditing objectives.

However, there is trade-off in aggregation. Specifically, the more aggregated the data, the more normal the analytical relationships tend to be, thus increasing statistical power in the analysis. However, normality comes at the price of missed detections at the detail level. So, the choice of

aggregation levels has to be made on a case-by-case basis taking into account the inherent characteristics and risk level of the underlying transactional data. These challenges introduced by the aggregation become more evident in a continuous audit of Big Data, and they present topics for much future research.

Data Confidentiality

Data confidentiality means certain data, or more often the associations among data points, are sensitive and cannot be released to others (Ciriani et al. 2009). In the era of Big Data, data can easily be associated with other data. Once some sensitive data are leaked, they can propagate with high velocity and connect to a large amount of related data. Thus, Big Data confidentiality becomes even more urgent and important to preserve brand image and secure competitive advantage.

CA is attractive for auditing confidential Big Data because the automation software protects the data from routine human observation and handling. However, any data stream seeking to identify a trend will need to be retained in some format, at least until the trend is established and documented in a CA system, and this retention of data may carry potential security and confidentiality risks (Alles et al. 2006). For example, accounts payable and receivables data may need to match invoices or orders with receipts and will be retained in a CA database because of time series length requirements for analysis. Under these circumstances, one of the extensively applied solutions for this Big Data confidentiality problem is the encryption of data by corporate personnel. However, this encryption solution then leads to the following two issues in a CA system: The first issue involves *data searching*. While certain data, including some sensitive financial data, are typically encrypted, a CA system may need to search and access these data. To address this issue, schemes based on Song, Wagner, and Perrig (2000) can be applied to CA system to search on encrypted data through keywords. But the schemes require the person who encrypts data to create simultaneously the proofs (e.g., signatures) of keywords, which are used by the CA system to check whether some keywords are in the encrypted data without knowing the plaintext. The second issue entails how to *audit the encrypted data*. There are techniques regarding some encryption algorithms that can be used by auditors, such as those suggested by Rivest, Adleman, and Dertouzos (1978) and Gentry (2009). For example, only certain senior auditors would have access to the sensitive plaintext and would encrypt it into ciphertext, leaving junior auditors to work on the ciphertext. In this manner, the sensitive data may be known only by few authorized senior auditors, and the probability of leaking information during an auditing process can be reduced.

However, there remain several issues when the above searching and encryption algorithms are applied to CA system: the searching algorithm can only be used to judge whether a keyword is in the ciphertext, and the encryption algorithm is quite time consuming. Therefore, continued research to optimize encryption algorithms is desired. The principal challenge of technical development is protecting data privacy while guaranteeing utility for the audit, so the establishment of theoretical trade-offs between privacy and utility is an important open area for research.

CONCLUDING REMARKS

In this paper, we focused on the Big Data challenges to data analytics in CA. Based on the features of Big Data, we have identified the gaps between the demands that Big Data analysis presents and the current state of data analysis in most CA systems, and have described the specific five gaps of Big Data (see Figure 1). To summarize, the Big Data qualities of Volume, Velocity, Variety, and Veracity contribute to the creation of the following Big Data Gaps: Data Consistency, Data Integrity, Data Identification, Data Aggregation, and Data Confidentiality. These Big Data Gaps create challenges for current CA systems, as shown in Figure 1 and discussed throughout this paper. The paper outlines possible solutions to these gaps along with needed research topics with

the aim of increasing the applicability of continuous auditing systems to Big Data. Big Data is a business phenomenon that is here to stay, and CA systems need to adapt to its challenges.

REFERENCES

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

Alles, M., G. Brennan, A. Kogan, and M. A. Vasarhelyi. 2006. Continuous monitoring of business process controls: A pilot implementation of a continuous auditing system at Siemens. *International Journal of Accounting Information* Systems 7 (2): 137–161.

Ananthakrishna, R., S. Chaudhuri, and V. Ganti. 2002. Eliminating fuzzy duplicates in data warehouses. *Proceedings of the 28th International Conference on Very Large Databases:* 586–597.

Arasu, A., S. Chaudhuri, and R. Kaushik. 2008. Transformation based framework for record matching. *Proceedings of IEEE 24th International Conference on Data Engineering:* 40–49.

Brannen, L. 2006. *Upfront: Continuous Auditing Is Ready for Prime Time*. Available at: http://businessfinancemag.com/risk-management/upfront-continuous-auditing-ready-prime-time

Canadian Institute of Chartered Accountants/American Institute of Certified Public Accountants (CICA/AICPA). 1999. *Research Report: Continuous Auditing*. Toronto, Canada: CICA, AICPA.

Cao, M., R. Chychyla, and T. Stewart. 2015. Big Data analytics in financial statement audits. *Accounting Horizons* 29 (2).

Cheah, Y. W., and B. Plale. 2012. Provenance analysis: Towards quality provenance. *Proceedings of IEEE 8th International Conference on E-Science:* 1–8.

Chiang, F., and J. Miller. 2008. Discovering data quality rules. *Proceedings of the VLDB Endowment:* 1166–1177.

Ciriani, V., S. Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati. 2009. Keep a few: Outsourcing data while maintaining confidentiality. *Proceedings of European Symposium on Research in Computer Security:* 440–455.

Crawl, D., J. Wang, and I. Altintas. 2011. Provenance for MapReduce-based data-intensive workflows. *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science:* 21–30.

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

Fan, W., and F. Geerts. 2010. Capturing missing tuples and missing values. *Proceedings of the 29th ACM SIGMOD-SIGACT:* 169–178.

Fan, W., F. Geerts, and J. Wijsen. 2012. Determining the currency of data. *Proceedings of the 30th ACM Transactions on Database Systems:* 71–82.

Fan, W., X. Jia, J. Li, and S. Ma. 2009. Reasoning about record matching rules. *Proceedings of the VLDB Endowment* 2 (1): 407–418.

Fuxman, D., and J. Miller. 2007. First-order query rewriting for inconsistent databases. *Journal of Computer and System Sciences* 73 (4): 610–635.

Gentry, C. 2009. Fully homomorphic encryption using ideal lattices. *Proceedings of the 41st ACM Symposium on Theory of Computing:* 169–178.

Getoor, L. 2010. *Graph Identification*. Berlin, Germany: Springer.

Golab, L., H. Karloff, F. Korn, A. Saha, and D. Srivastava. 2009. Sequential dependencies. *Proceedings of the VLDB Endowment* 2: 574–585.

Groomer, S. M., and U. S. Murthy. 1989. Continuous auditing of database applications: An embedded audit module approach. *Journal of Information Systems* 3 (2): 53–69.

Hernandez, A., and J. Stolfo. 1998. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2 (1): 9–37.

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

Kogan, A., M. Alles, M. A. Vasarhelyi, and J. Wu. 2014. *Analytical Procedures for Continuous Data Level Auditing: Continuity Equations*. Available at: http://raw.rutgers.edu/docs/Innovations/Continuity%20Equations.pdf

Koudas, N., A. Saha, D. Srivastava, and S. Venkatasubramanian. 2009. Metric functional dependencies. *Proceedings of IEEE 25th International Conference on Data Engineering:* 1275–1278.

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

Lim, E., J. Srivastava, S. Prabhakar, and J. Richardson. 1996. Entity identification in database integration. *Proceedings of 9th International Conference on Data Engineering:* 294–301.

Mayfield, C., J. Neville, and S. Prabhakar. 2010. ERACER: A database approach for statistical inference and data cleaning. *Proceedings of the 2010 International Conference on Management of Data:* 75–86.

Menezes, A., P. Oorschot, and S. Vanstone. 1996. *Handbook of Applied Cryptography*. Boca Raton, FL: CRC Press.

Motro, A. 1989. Integrity = validity + completeness. *ACM Transactions on Database Systems* 14 (4): 480–502.

Newcombe, H., J. Kennedy, S. Axford, and A. James. 1959. Automatic linkage of vital records. *Science* 130 (3381): 954–959.

Perols, J., and U. Murthy. 2012. Information fusion in continuous auditing. *Journal of Information Systems* 26 (2): 35–52.

Rivest, R., L. Adleman, and M. Dertouzos. 1978. *On Data Banks and Privacy Homomorphisms. Foundations of Secure Computation,* 169–179. New York, NY: Academic Press.

Schafer, L., and W. Graham. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7 (2): 147–177.

Sheth, A., and M. Rusinkiewicz. 1990. Management of interdependent data: Specifying dependency and consistency requirements. *Proceedings of the Workshop on the Management of Replicated Data:* 133–136.

Song, D., D. Wagner, and A. Perrig. 2000. Practical techniques for searches on encrypted data. *Proceedings of IEEE Symposium on Security and Privacy:* 44–55.

Vasarhelyi, M. A., M. Alles, and A. Kogan. 2004. Principles of analytic monitoring for continuous assurance. *Journal of Emerging Technologies in Accounting* 1: 1–21.

Vasarhelyi, M. A., M. Alles, and K. T. Williams. 2010. *Continuous Assurance for the Now Economy*. Available at: http://www.charteredaccountants.com.au/Industry-Topics/Audit-and-assurance/Publications-and-tools/Other-audit-resources/Resources/Continuous-assurance-for-the-now-economy