RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Toward Effective Big Data Analysis in Continuous Auditing

By Juan Zhang, Xiongsheng Yang, and Deniz Appelbaum

Introduction

New sources: emails, phone calls, click stream traffic, social media, news media, sensor recordings and videos, and RFID tags

4 V's of Big Data (Laney, 2001; IBM, 2012):

- Volume
- Variety
- Velocity
- Veracity

Purpose of paper: To identify these gaps and challenges and to point out the need for updating the CA system to accommodate Big Data analysis

Big Data and Transforming the Continuous Audit

Continuous Auditing (CA) is *"to provide assurance simultaneously with, or a short period of time after, the occurrence of events underlying the subject matter"* (CICA/AICPA, 1999)

- CA is needed to access and process Big Data; BD creates chance for wider scale and range of auditing
- How to deliver value from the Big Data repositories?

  - Connections and relations are paramount

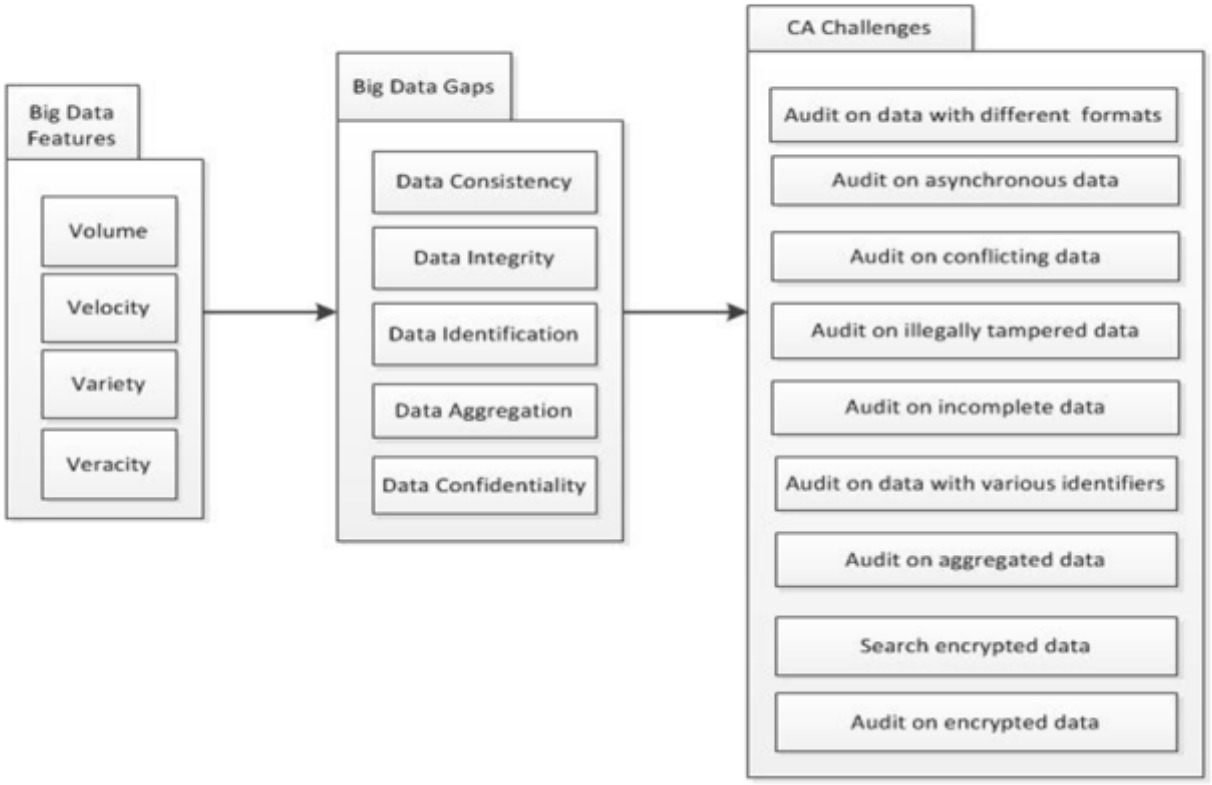
4 V's creates gaps of data consistency, data identification, data aggregation, and data integrity

- The main elements of CA architecture, such as data provisioning, data filtering, and data diagnosis will need to be altered for the Big Data environment

Big Data and Transforming the Continuous Audit

FIGURE 1

Effective big data analytics in continuous auditing



Big Data Gaps and Challenges to Continuous Auditing

Data Consistency: summarizes the validity, accuracy, usability and integrity of related data between applications and across an organization.

- Big Data = more gaps in data consistency
- Big Data = Map Reduce/Hadoop

3 types of data inconsistency:

- data formats
- data synchronization
- data contradiction

Integrating methodologies that can address these issues in a CA system of Big Data without losing efficiency is an area of future research

Big Data Gaps and Challenges to Continuous Auditing

Data Integrity: maintaining and assuring the accuracy and consistency of data over its entire life cycle.

- lack of data integrity = data tampering or incomplete data? (physical or logical causation)
- Traditional methods of integrity checks such as reasonability, edit checks, and comparison with other sources may not be currently practical in a Big Data environment

2 types of data integrity issues:

- Modification: unintentional or intentional (could be detected with message digest keys…Menezes et al, 1996)
- Incomplete Data: various causes from human error to system glitches, various classic but infeasible methods

Big Data Gaps and Challenges to Continuous Auditing

Data Identification: regarding records that link two or more separately recorded bits of information about the same individual or entity (Newcombe et al, 1959)

- Challenging to connect disparate types of data from varied sources concerning the same event/entity

Two main methods to address this issue:

- Semantic rules: based on the experience and knowledge required to set the identification rules
- Similarity Measures: various algorithms based on fuzzy matches of distance and frequency.

Challenge: most provide just probabilistic measures on data identification, and this audit risk to the CA system is magnified with Big Data

Big Data Gaps and Challenges to Continuous Auditing

Data Aggregation: Data aggregation is any process in which information is gathered and expressed in a summary form

- Kogan et al (2014): aggregation of exception data derived from Big Data can aid in the identification of patterns
- Alles et al (2006): alarms should be aggregated to pinpoint weaknesses in certain control areas
- Perols and Murthy (2011): layered framework that aggregates the alarms instead of the raw data

Challenge: the more aggregated the data, the more normal it will appear to be and detections may be missed at the detail level. This issue is compounded with BD!

"Here's a list of 100,000 warehouses full of data. I'd like you to condense them down to one meaningful warehouse."

Big Data Gaps and Challenges to Continuous Auditing

Data Confidentiality: certain data, or more often the associations among data points, are sensitive and cannot be released to others (Ciriani et al, 2009)

- If sensitive data is associated with other data sources in Big Data, the damage can be significant to individuals and businesses

Two issues about data confidentiality in a CA system:

- Data searching: search encrypted data?
- Audit the encrypted data: techniques with encryption algorithms are suggested, but time consuming

Challenge: how to audit private Big Data while protecting its confidentiality and preserving some utility?

Concluding Remarks

Focused on Big Data challenges to data analytics in CA

The Four Big Data qualities:

- Volume
- Variety
- Velocity
- Veracity

Which Create these Big Data Gaps:

- Data Consistency
- Data Integrity
- Data Identification
- Data Aggregation
- Data Confidentiality

Research is needed in these areas to increase the applicability of CA in Big Data!!