

## **Antecedents of big data quality**

### **An empirical examination in financial service organizations**

Haryadi, Adiska Fardani; Hulstijn, Joris; Wahyudi, Agung; van der Voort, Haiko; Janssen, Marijn

**DOI**

[10.1109/BigData.2016.7840595](https://doi.org/10.1109/BigData.2016.7840595)

**Publication date**

2017

**Document Version**

Peer reviewed version

**Published in**

Proceedings of 2016 IEEE International Conference on Big Data, Big Data 2016

**Citation (APA)**

Haryadi, A. F., Hulstijn, J., Wahyudi, A., Van Der Voort, H., & Janssen, M. (2017). Antecedents of big data quality: An empirical examination in financial service organizations. In Proceedings of 2016 IEEE International Conference on Big Data, Big Data 2016 (pp. 116-121). [7840595] Institute of Electrical and Electronics Engineers Inc.. DOI: 10.1109/BigData.2016.7840595

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

# Antecedents of Big Data Quality<sup>1</sup>

## An Empirical Examination in Financial Service Organizations

Adiska Fardani Haryadi<sup>1</sup>, Joris Hulstijn<sup>2</sup>, Agung Wahyudi<sup>1</sup>, Haiko van der Voort<sup>1</sup>, and Marijn Janssen<sup>1</sup>

<sup>1</sup>Faculty of Technology, Policy, and Management, <sup>2</sup>Department of Management

<sup>1</sup>Delft University of Technology, <sup>2</sup>Tilburg University

<sup>1</sup>Delft, <sup>2</sup>Tilburg, The Netherlands

a.f.haryadi@student.tudelft.nl, m.f.w.h.a.janssen@tudelft.nl, j.hulstijn@uvt.nl, h.g.vandervoort@tudelft.nl, a.wahyudi@tudelft.nl

**Abstract**—Big data has been acknowledged for its enormous potential. In contrast to the potential, in a recent survey more than half of financial service organizations reported that big data has not delivered the expected value. One of the main reasons for this is related to data quality. The objective of this research is to identify the antecedents of big data quality in financial institutions. This will help to understand how data quality from big data analysis can be improved. For this, a literature review was performed and data was collected using three case studies, followed by content analysis. The overall findings indicate that there are no fundamentally new data quality issues in big data projects. Nevertheless, the complexity of the issues is higher, which makes it harder to assess and attain data quality in big data projects compared to the traditional projects. Ten antecedents of big data quality were identified encompassing data, technology, people, process and procedure, organization, and external aspects.

**Keywords**—big data; data quality; big data quality; antecedents; finance

### I. INTRODUCTION

Nine out of ten business leaders now perceive data as the fourth fundamental resource for business after land, labor, and capital. Likewise, big data has been apprehended as the promising type of data to create value in the organizations. Among all sectors, finance is the one with the highest potential to capture the value of big data. Some promises of big data in this domain are better risk management, more advanced marketing, and deeper knowledge about customer's behavior. To achieve these aims, numerous projects and applications based on big data are initiated by banks and insurance companies.

However, exploiting big data proves to be challenging. It is reported that only 40% of financial service organizations considered their expectations from big data utilization have been “delivered” and “over delivered”, while the rest determined as either “undelivered” or “too early to tell” [1]. There are many issues that hinder financial service organizations to create value from big data. This research focuses on one of the main issues, e.g. data quality. Data quality issues affect the results from big data analysis and decision-making [2]. The objective of this research is to identify antecedents of information quality from big data utilization (big data quality) in financial institutions. This will help to understand how big data quality can be improved.

This paper is structured as follows. This section explains the project background and the problem solving it is contributing to. Section II describes the research approach to meet the determined research objective. Section III elaborates the latest development of big data applications in financial industry. Section IV explains the data quality in big data context. Section V and VI present the main findings of this research. Thereafter, section VII exhibits the relevant discussions. Lastly, this paper is closed with the conclusions presented in section VIII.

### II. RESEARCH APPROACH

Adopting from the Data Quality Management (DQM) approach, the quest to identify the antecedents of data quality consists of three activities: 1) defining information output in focus, 2) assessing significant quality traits of the information output (can be the important and/or the problematic ones), 3) and analyzing root causes or factors that affect data quality traits [3, 4] To assure a comprehensive result, triangulation between three main sources are employed, namely systematic review to academic literature, multiple case studies, and content analysis to non-academic literature to collect as many stories of big data implementation in financial domain as possible. Section V describes in detail the conducted method, the collected information from each source, and the analysis to perform the first and second activities, followed by similar finding of the third activity in section VI. Beforehand, a brief overview about big data in financial domain and data quality in big data context will be explained in the next two sections to set the common ground of this research.

### III. BIG DATA IN FINANCIAL INDUSTRY

Despite its ubiquitous use, there is still no single uniform definition of big data. Nevertheless, some academic works have collated the list of recurring terms and definitions related to big data [5, 6]. The frequently recognized elements are 1) Volume

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: [10.1109/BigData.2016.7840595](https://doi.org/10.1109/BigData.2016.7840595)

(indicated by the use of serious computing power to process sizeable dataset; 2) Velocity (indicated by data that are being formed at an unprecedented speed and must be dealt with in a timely manner); 3) Variety (indicated by diversity of data formats employed such as structured data, semi-structured data, and unstructured data); 4) Veracity (indicated by uncertainty of data trustworthiness from external open source); and 5) Value (indicated by the impact it possibly yields for the organizations).

Big data has been widely utilized by financial sectors in the recent years. From a review to 32 white papers and online articles that are selected from a thorough search, six clusters of big data applications in financial industry are identified. They are 1) customer acquisition and retention, 2) risk management, 3) fraud and cybercrime detection, 4) investment and asset management, 5) online monitoring, and 6) new service proposition.

#### IV. DATA QUALITY IN BIG DATA CONTEXT

Data quality is a key factor for decision making in financial industry [7]. This section explains about data quality definition, data quality issues, and data quality issues in financial industry.

##### A. Data Quality

Many information system researchers see data quality as equivalent with information quality [8, 9]. In this paper, both refer to the same, but a distinction is made on whether the focus is on the data level (before analysis) or information level (after analysis). Hence, unless explicitly differentiated, ‘data quality’ term encompasses data and information quality.

Quality is rather a subjective term. The notion of data quality depends not only on its intrinsic quality (conformance to specification), but also the actual use of data (conformance with customer’s expectation) [10, 11]. Data quality is a multidimensional concept [12]. Some frequently mentioned dimensions are accuracy, timeliness, currency, completeness, consistency, accessibility, and relevance. However, there is neither a consensus on what constitute the dimensions of data quality, nor the exact meaning of each dimension [12, 13]. The quest for dimensions that may constitute big data quality, particularly in financial industry, is one of the aims of this study and will be elaborated further in the next section.

##### B. Data Quality Issues

Data quality issues mean the unfulfillment of fitness for use condition by the customers, and may lead to poor information quality (Figure 1). Low data quality can lead to increasing administrative costs due to inefficient business operation. Eventually, it may also affect customer’s satisfaction, which later influence dropping in sales. Dealing with data quality issues is not an easy challenge which may incur considerable amount of resources [14].

##### C. Data Quality Issues in Financial Industry

Data quality is an essential property in driving business performance and compliance, with no exception for financial service organizations. Typical data quality issues in these organizations can be inconsistency, incompleteness, incorrectness and redundancy. Nevertheless, the existing works on data quality are mainly focusing on structured data rather than (big) unstructured data [15].

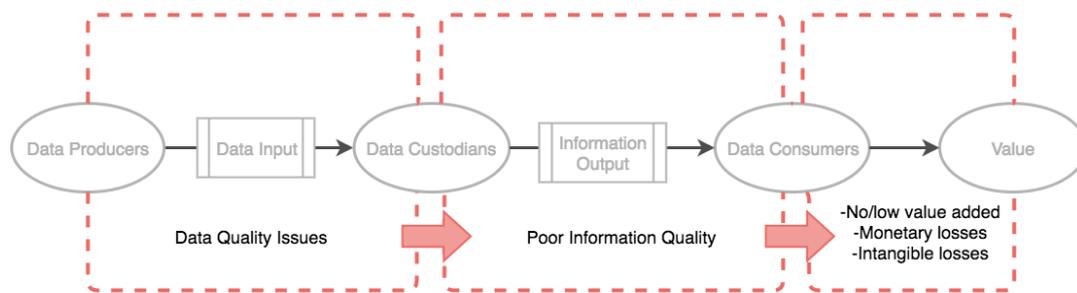


Figure 1. Implications of Data Quality Issues

#### V. BIG DATA QUALITY

This section explains findings of the first two activities mentioned earlier in section II, which are defining information output and assessing important big data quality traits in financial industry. These findings are collected from content analysis and case study.

##### A. Content Analysis

A desk research to white papers and online articles is performed through Google Search as the largest search engine by using keyword: “big data” <institution name>. The search is narrowed to 10 biggest banks and 10 biggest insurance companies in Europe. By focusing on articles within five-years timeframe (2011-2016) and after eliminating irrelevant search results,

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: 10.1109/BigData.2016.7840595

from 2,000 search results (10 Google Search pages of 10 search result per page for each institution), seven articles are marked as containing sufficient level of details pertaining big data projects in financial institutions. Thereafter, content analysis is performed using Nvivo software to support the coding and analysis processes. As the result, information output from each big data project and the important quality traits can be identified and summarized in Table 1.

Table 1. Overview of Content Analysis

| No | Institution        | Information Output   | Information Quality Goal                              |
|----|--------------------|--|---|
| 1  | ING Bank           | Tailored advertisement   | Relevancy, Validity, Timeliness                       |
| 2  | Barclays           | Vendor recommendation with better rates                        | Relevancy, Validity, Timeliness                       |
| 3  | UBS Bank           | Financial and reputational risks of a company                  | Comprehensiveness, Believability                      |
| 4  | Allianz Insurance  | Suspicious claim and network of organized insurance fraudsters | Accuracy, Timeliness                                  |
| 5  | ING Bank           | Notification of fraud  | Accuracy, Timeliness                                  |
| 6  | Barclays, RBS Bank | Complaint trends and report                                    | Accuracy, Timeliness, Traceability, Comprehensiveness |
| 7  | BBVA               | Card transactions in Spain during July-August 2014             | Relevancy   |

*Accuracy* means representing real-world state. *Relevancy* means addressing customer’s needs. *Believability* means the extent to which information is regarded as credible. *Comprehensiveness* means covering of larger content or scope. *Timeliness* means availability of data on time. *Traceability* means the extent to which information can be traced back to the original source. *Validity* means information is generated with respect to applicable regulations and procedures.

### B. Case Study

A set of questions to identify big data quality issues is being asked to three case study organizations. Each of these organizations must fulfill three criteria: 1) a bank or an insurance company, 2) has provided financial services for minimum five years, 3) has undergone any big data initiatives, and 4) is willing to cooperate in this study. Summary of their cases is listed in Table 2. The organization’s names are not made explicit due to confidentiality reason.

Table 2. Three Cases of Case Study [16]

| No | Institution | Information Output                                 | Information Quality Goal  |
|----|-------------|--|---|
| 1  | Bank X      | Package of mortgage files                          | Accuracy, Completeness, Currency, Consistency, Timeliness, Uniqueness, Validity, Traceability |
| 2  | Bank Y      | Credit risk level, most suitable loan for customer | Believability, Comprehensiveness, Relevancy, Validity   |
| 3  | Insurance Z | A single customer 360 <sup>0</sup> profile         | Uniqueness, Accuracy, Completeness, Currency, Timeliness, Validity, Comprehensiveness         |

Four additional traits of data quality are identified from case study. *Completeness* means the degree to which all possible relevant states are represented in the information. *Currency* means how up to date the information is. *Consistency* means no contradiction between the information and the data source. *Uniqueness* means no redundant information. The elaborated version of case study finding and analysis can be observed in [16].

## VI. ANTECEDENTS OF BIG DATA QUALITY

After determining the information output in focus, as well as data quality traits that need to be taken into account, analysis to reveal the antecedents of data quality traits is performed in this section. Three methods are employed to perform this analysis. First, content analysis using Nvivo software is conducted towards the seven articles in V.A, by coding every phrase indicated as antecedents of big data quality. Second, the search continues to academic literature. However, in this stage the scope is not limited only to finance domain to open more possibilities. Hence, the keyword that is being used to find expected publications in Scopus (only from the title) is: “big data” quality. As the result, 76 papers are found. After briefly reading each paper, only five of them are found to have a clue about antecedents of big data quality. Third, data are collected from the same three case study organizations as mentioned in V.B. Findings from each method are described in turn.

### A. Content Analysis

To start with, four traits of big data (Volume, Velocity, Variety, and Veracity; Value is not considered as it is rather the ‘impact’ than the intrinsic characteristics of the big data) that may affect big data quality are identified in each case based on the criteria below.

- Volume is indicated by phrases that mention the use of big data technology to process large amount of data which until recently would not be feasible or require too much time for organization.
- Velocity is indicated by phrases that mention the use of data that are being formed at an unprecedented speed or data that must be dealt in (almost) real time.

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: 10.1109/BigData.2016.7840595

- Variety is indicated by phrases that mention the use of semi or unstructured data in addition to structured data.
- Veracity is indicated by phrases that mention the use of data in which organization has no or unclear authority over its use (e.g. customer's individual transaction data), and control over its quality and trustworthiness (e.g. news, blogs, social media).

Next, the relationships between the big data traits and their influence to information quality are drawn as depicted in **Figure 2**. **Error! Reference source not found.** Apparently, not all four big data traits (strongly) appear in every project. Most of them employ high volume data, but not always high variety and high veracity data. Hence, it is interesting to see from this small sample that European financial service organizations currently rely on their internal data, and the use of unstructured data and external open data are still limited. Moreover, only one of the seven cases explicitly confirms the presence of high velocity of data either due to lack of explicit mention of this trait in the data source, or lack of clarity regarding threshold for 'high velocity' (the rest is unknown, hence the dashed line).

The four big data traits are found to affect big data quality in the following ways. First, high **volume of data** could enable discovery of hidden pattern, such as finding the network of suspicious fraud. In addition, bigger data volume most likely leads to higher representativeness. However, it should also be kept in mind that in the same time big data are prone to bias if the sample only reflecting views of certain groups. Second, high **velocity of data** could heavily affect currency of data. Third, high **variety of data** directly translates into higher level of traceability. For example, complaints trends, which incorporate data from various sources, will require ability to trace the details of each trend up to the origination from each source. In addition, high variety of data means more data format involved during data collection and analysis. Consequently, organization must be able to precisely interpret data from each source and correctly integrate data that refer to the same meaning to yield accurate and comprehensive information. Failure to interpret sentiment from a tweet, for example, could result into incorrect figure of negative feedbacks from customers, which could mislead decision to be taken. Another implication from high variety of data is likelihood of inconsistent data content. This inconsistency eventually triggers organization to make subjective judgment to choose which news is the correct one, which then could lower the believability of the reputational risk of a company. Fourth, high **veracity of data** threatens reliability and credibility of data content, because the source may not be credible (e.g. personal blogs, social media) and/or the process of generating the data may involve errors (e.g. incomplete feedbacks, biased news). As a result, a pile of incomplete complaint data could lead to fully incorrect trend and thus misleading decision to be taken. Veracity also means uncertainty of the right to use data, such as customer's transaction data in individual level, as well as data from social media and personal blogs.

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: [10.1109/BigData.2016.7840595](https://doi.org/10.1109/BigData.2016.7840595)

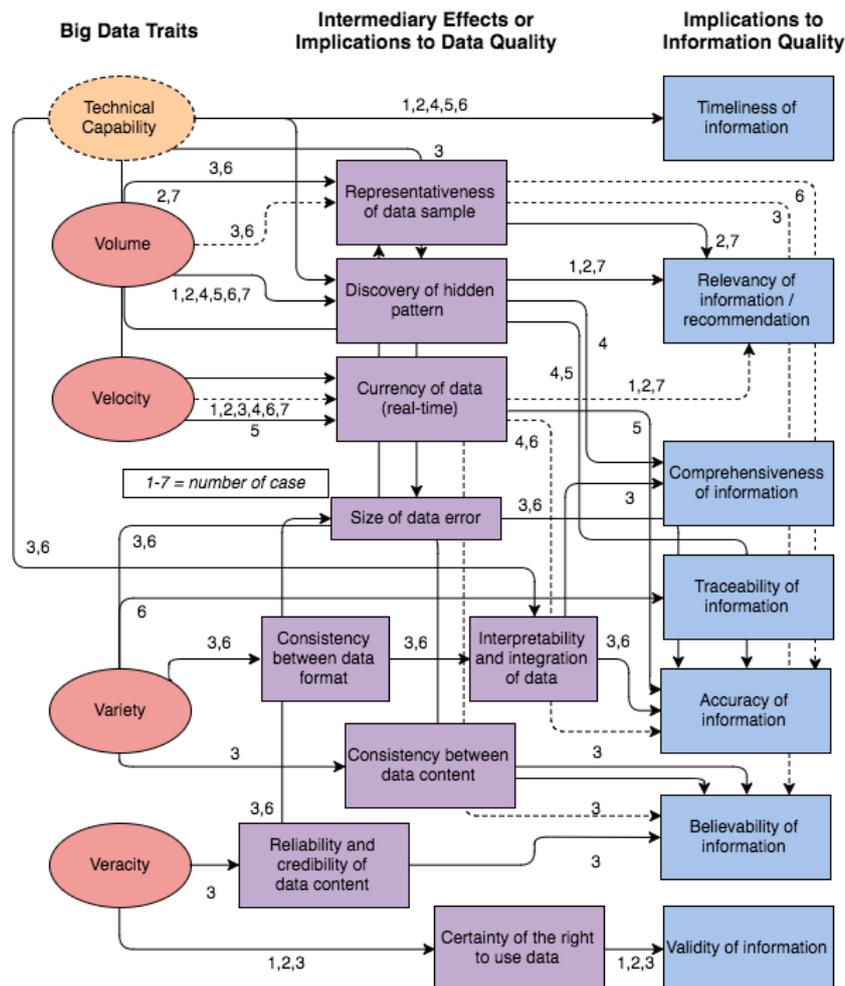


Figure 2. Influence of Big Data Traits from Content Analysis

Moreover, findings from content analysis indicate that beyond the four traits of big data, *big data technology* and *data analytics technique* have also been proven to substantially affect information quality in all of the seven cases. These factors can be considered as the main enabler to extract, process, and analyze large volume of data with high velocity and wide variety. Moreover, *analytics skills* to draw insight from a bunch of data are also implied from several cases. Next in the line is the use of *drill-down visualization tool* to ease traceability of particular information. Moreover, according to one of the sources, it is also important that all required data are stored in a *centralized data repository* to improve accessibility. If data is inaccessible, all other data qualities are irrelevant. Lastly is *certainty of permission to use data*, which tends to touch the ethical side of using data that are made available in public. This (un)certainty eventually leads to (un)certainty about the right to use data for analysis.

Other than that, there are also two indirect factors revealed from analysis in this stage, which are *internal capability to be agile* (e.g. working in short cycle, adapting with changes, accepting mistakes) and *compelling business case*. Although not being explicitly mentioned how these factors affect data quality, the interim guess is that they support organization's capability to experiment with big data. Success story and trial-error are very relevant to discover hidden pattern from big data, in which the value is mostly unknown in the beginning.

### B. Literature Review

As from the literature review, three antecedents are identified. First is the *authenticity of (big) data collecting method*, which is rather opportunistic than scientific [17-19]. This factor may significantly affect representativeness of data sample being used for analysis. For example, drawing a statistical conclusion from Twitter may seem representing a large population,

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: 10.1109/BigData.2016.7840595

while in fact it only represents people who live in big cities with convenient access to Internet. In addition, it is oftentimes difficult to find all the possibly matching keywords to mine data about particular issues from different data sources, while avoiding redundant data. Second and third are the *credibility of data source and data content* [18, 20, 21], which can be threat for reliability and credibility of data being used for analysis.

### C. Case Studies

The impacts of four V traits are also analyzed from case study to three organizations as displayed earlier in V.B. As the result, some implications are found to be similar with the findings from content analysis. Two main differences are described as follows. *Volume issue* brings the issue of data redundancy due to legacy system consolidations in financial institutions that store multiple data. These data refer to the same objects but with different formats, and thus threaten the uniqueness of information. *Veracity issue* does not only appear in when using external data, but also internal data. In Bank X for example, accuracy, completeness, and currency of manual data input heavily affect the accuracy, completeness, and currency of information from analysis. As the size of data increases ('Volume' trait), the size of error carried by these data tends to be snowballed. All implications of big data traits in data level contribute to the information level. For example, accuracy, believability, and relevancy are determined by discovery of the correct pattern, representativeness of data sample, currency of data, ability to interpret and combine various unstructured data, consistency between data content, and reliability of data content times size of data error.

Analysis from case studies confirms some findings from content analysis and literature review. However, three new antecedents are also identified. First, *clarity of metadata* is very important for data custodian in understanding what is actually inside a data, especially for unstructured data. Second, *unclear policy about the use of data* leads analysts to the confusion: 'which data can be used for which purpose'. Absence of this policy could result into information that is formed by invalid or unauthorized data. Third, *formation of multidisciplinary team* is very essential to discover the hidden value of big data. Business should be able to seek compelling business case, while IT is expected to realize it from technical side.

In addition, it is found that antecedents of big data quality do not only encompass specific factors related to big data, but also factors related to general data management, such as multiple legacy systems, compliance to data access policy, and top management support. Nevertheless, to keep the focus of the study, factors that correspond to traditional data management are not considered for further analysis.

## VII. DISCUSSIONS

Section V and VI have reported the outcomes of data collection in this study. Now, several intriguing findings that are discovered throughout this research will be succinctly discussed.

### A. Big Data Perception in Financial Industry

Big data definition could vary in diverse organizations and industries. Based on what is perceived as 'big data' by the studied organizations in this research, all four V's of big data are apparent in different projects. The most salient commonality in all case is the (undiscovered) Value of big data, while Velocity is hardly noticeable. Apart from that, 80% of the cases still rely on internal big data sources rather than utilizing external sources such as social media, online news, and open data.

### B. Dimensions of Big Data Quality

According to the findings from this study, there are 11 important dimensions of big data quality in financial domain that should be taken into consideration for its improvement. They could be traits that serve as the aim of the big data project or traits that are likely problematic. They are accuracy, believability, relevancy, currency, completeness, comprehensiveness, consistency, uniqueness, timeliness, validity, and traceability. As observed from case study and content analysis, not all the information quality dimensions are perceived as important in different projects. This means the dimensions can be context independent or context dependent. Context independent dimensions are likely present in both case study and content analysis, for example validity and traceability. Context dependent means that the data quality rules could not be specified beforehand, because they depend on several contextual factors, such as the purpose of big data projects, the type of information output, time-criticality of information, and the existence of rules to cross-reference. This accounts for the remaining nine data quality dimensions.

### C. Comparing Traditional and Big Data Quality

Is big data quality different from traditional data quality? Findings from this research indicate that there is no fundamentally new data quality issue in big data context, which confirms discovery from a previous research [21]. However, after tracing thoroughly the implications of the four V traits of big data to information quality, three distinguishing points may be worth to consider, as follows.

- 1) *Big data quality contains wider array of quality indicators that constitute information quality dimensions.*

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: [10.1109/BigData.2016.7840595](https://doi.org/10.1109/BigData.2016.7840595)

Variety and veracity traits enhance the number of indicators that affect information quality dimensions. For example, accuracy, believability, and relevancy are now also determined by interpretability and consistency of content. As the implication, it becomes more complex to assess quality in big data context due to higher number of quality indicators to consider.

2) *Big data quality has longer range of quality level.*

Larger number of indicators that correspond to information quality implicates higher variation on its overall level of quality, and thereby longer range on its level of quality than in traditional data quality. This means that information resulting from big data analysis could generate highly valuable insight or decision, but on the other side, could also generate extremely misleading decision.

3) *Big data quality contains higher proportion of indicators that convey negative effects than positive effects to the total quality.*

As can be observed from **Error! Reference source not found.**, there are some adverse impacts from the four V traits of big data to data quality in the data level, such as bias representativeness of data sample, higher level of data redundancy, shorter currency of data, snowballed data errors, inexplicable data, complexity of data integration, inconsistent data, untrustworthy data, and uncertainty of the right to use data. These issues might not be novel, but have higher level of complexity in big data projects. Hypothetically, this means it could be more difficult to attain good quality level in big data context rather than in traditional data context.

#### D. Antecedents of Big Data Quality

There are at least 10 big data quality antecedents revealed throughout the section V. They can be grouped into five clusters, namely 1) data, 2) technology, 3) people, 4) organization, and 5) external environment. However, it can be observed that not all of them directly influence big data quality, although the absence of these factors may lead to the non-existence of result from big data analysis. First, big data processing technology is highly required to manage and process huge and heterogeneous data for analysis. If it is not present, organization will not be able to yield any analysis result. Second, in big data lies undiscovered value for the organizations. To discover the value, organization should have agile capability to encourage experiment with big data and compelling business case (as the trigger to encourage experiment with big data). This discovery process is essential because the definition of big data quality for a particular project will be based on the value that organization would like to realize. Third, it is important to ensure the accessibility of data for analysis in a centralized repository, since big data tend to consist of multiple types of data from multiple sources. The centralization of data is intended to ensure that all information possibly associated with subject being analyzed can be obtained easily. Hence, big data quality is supported with capabilities to discover hidden value of big data, to ensure accessibility of data, and to enable the data operation. The mapping of three aspects that support big data quality is depicted in Figure 3.

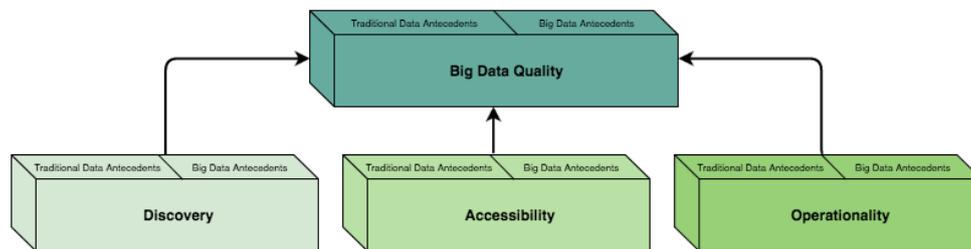


Figure 3 Three Supporting Aspects of Big Data Quality

## VIII. CONCLUSIONS

This study is among the first to collate the list of the antecedents of big data quality from empirical perspective. It discovers 11 important big data quality dimensions for financial institutions, as well as 10 antecedents of big data quality that they should take into consideration when aiming to improve data quality from big data analysis. This study has shown that there are no new data quality issues in big data projects. However, the complexity of these issues is higher, which makes it more difficult to assess and attain data quality in big data projects compared to the traditional ones. Moreover, most of data quality rules in big data projects cannot be specified independently from its context and thus cannot be determined a priori.

As the focus of this study, antecedents of big data quality have been unearthed from case study, literature review, and content analysis. The antecedents consist of 1) four big data V traits, 2) clarity of metadata, 3) credibility of data source and data content, 4) big data tools, technique, and technology, 5) authenticity of data collecting method, 6) clarity of policy about use of data, 7) analytics skills and multidisciplinary team, 8) centralized data repository 9) agile capability, and 10) compelling business case. Nevertheless, not all them directly influence big data quality, but affect big data projects and influence the

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: [10.1109/BigData.2016.7840595](https://doi.org/10.1109/BigData.2016.7840595)

difficulty of creating value from big data. These supporting aspects are grouped into three clusters of capabilities: discovery of hidden value from big data, assurance of data accessibility, and enablement of data operation. Low data quality can result in the need for additional processing tasks and the consuming resources.

Our findings suggest that data and information quality are central issues in the value creation process from big data.

#### REFERENCES

- [1] C. Beattie and B. Meara, "How Big is Big Data? Big Data Usage and Attitudes among North American Financial Service Firms," Oliver Wyman 2013.
- [2] M. Janssen, H. Van der Voort, and A. Wahyudi, "Factors influencing big data decision-making quality," *Journal of Business Research*, 2016.
- [3] L. English, "Total Quality data Management (TQdM), Methodology for Information Quality Improvement, Information and Database Quality," ed: Kluwer Academic Publishers, 2002.
- [4] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, pp. 5-33, 1996.
- [5] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," in *AIP Conference Proceedings*, 2015, pp. 97-104.
- [6] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, p. 2, 2015.
- [7] D. Zahay, J. Peltier, and A. S. Krishen, "Building the foundation for customer data quality in CRM systems for financial services firms," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, pp. 5-16, 2012.
- [8] J. E. Boritz, "IS practitioners' views on core concepts of information integrity," *International Journal of Accounting Information Systems*, vol. 6, pp. 260-279, 2005.
- [9] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, pp. 211-218, 2002.
- [10] B. K. Kahn and D. M. Strong, "Product and Service Performance Model for Information Quality: An Update," in *IQ*, 1998, pp. 102-115.
- [11] M. Mouzhi Ge, "A review of information quality research," in *Proceedings: 12th International Conference on Information Quality*, 2007, pp. 76-91.
- [12] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, pp. 86-95, 1996.
- [13] R. R. Nelson, P. A. Todd, and B. H. Wixom, "Antecedents of information and system quality: an empirical examination within the context of data warehousing," *Journal of management information systems*, vol. 21, pp. 199-235, 2005.
- [14] I. Caballero, Ó. Gómez, and M. Piattini, "Getting Better Information Quality By Assessing And Improving Information Quality Management," in *IQ*, 2004, pp. 44-58.
- [15] C. Batini, F. Cabitza, C. Cappiello, and C. Francalanci, "A comprehensive data quality methodology for web and structured data," *International Journal of Innovative Computing and Applications*, vol. 1, pp. 205-218, 2008.
- [16] A. F. Haryadi, "Requirements on and Antecedents of Big Data Quality: An Empirical Examination to Improve Big Data Quality in Financial Service Organizations," Master, Faculty of Technology, Policy, and Management, Delft University of Technology, Delft, 2016.
- [17] S. Kelling, D. Fink, F. A. La Sorte, A. Johnston, N. E. Bruns, and W. M. Hochachka, "Taking a 'Big Data' approach to data quality in a citizen science project," *Ambio*, vol. 44, pp. 601-611, 2015.
- [18] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 134-142, 2016.
- [19] J. Maślankowski, "Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology," in *International Conference: Beyond Databases, Architectures and Structures*, 2014, pp. 92-101.
- [20] I. Caballero, M. Serrano, and M. Piattini, "A data quality in use model for big data," in *International Conference on Conceptual Modeling*, 2014, pp. 65-74.
- [21] D. Becker, T. D. King, and B. McMullen, "Big data, big data quality problem," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015, pp. 2644-2653.

<sup>1</sup> Fardani Haryadi, A., Hulstijn, J., Wahyudi, A., van der Voort, H., & Janssen, M (2016). Antecedents of big data quality: An empirical examination in financial service organizations. In: *Proceedings of the IEEE International Conference on Big Data*. pp. 116-121. DOI: [10.1109/BigData.2016.7840595](https://doi.org/10.1109/BigData.2016.7840595)