# Relating Big Data and Data Quality in Financial Service Organizations

Agung Wahyudi, Adiska Farhani, Marijn Janssen



# Relating Big Data and Data Quality in Financial Service Organizations

Agung Wahyudi, Adiska Farhani, and Marijn Janssen

Delft University of Technology, Jaffalaan 5, 2628 BX Delft,
The Netherlands

{a.wahyudi,M.F.W.H.A.Janssen}@tudelft.nl,
A.F.Haryadi@student.tudelft.nl

**Abstract.** Today's financial service organizations have a data deluge. A number of V's are often used to characterize big data, whereas traditional data quality is characterized by a number of dimensions. Our objective is to investigate the complex relationship between big data and data quality. We do this by comparing the big data characteristics with data quality dimensions. Data quality has been researched for decades and there are well-defined dimensions which were adopted, whereas big data characteristics represented by eleven V's were used to characterize big data. Literature review and ten cases in financial service organizations were invested to analyze the relationship between data quality and big data. Whereas the big data characteristics and data quality have been viewed as separated domain ours findings show that these domains are intertwined and closely related. Findings from this study suggest that variety is the most dominant big data characteristic relating with most data quality dimensions, such as accuracy, objectivity, believability, understandability, interpretability, consistent representation, accessibility, ease of operations, relevance, completeness, timeliness, and value-added. Not surprisingly, the most dominant data quality dimension is value-added which relates with variety, validity, visibility, and vast resources. The most mentioned pair of big data characteristic and data quality dimension is Velocity-Timeliness. Our findings suggest that term 'big data' is misleading as that mostly volume ('big') was not an issue and variety, validity and veracity were found to be more important.

**Keywords:** big data, 11V, data quality, variety, value, finance service organization

## 1 Introduction

Todays' organizations are harvesting more and more data using technologies such as mobile computing, social networks, cloud computing, and internet of things (IoT) (Akerkar, 2013). This data deluge can be used to create a competitive advantage over

competitors and create significant benefits (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2013) such as better understanding of customer's behavior, more effective and efficient marketing, more precise market forecasting, and more manageable asset risks (Beattie & Meara, 2013; PricewaterhouseCoopers, 2013). Manyika (2011) argues that finance and insurance organizations have one of the highest potential to take advantage from big data.

However, creating value from big data is a daunting task. Reid's (2015) study revealed that two thirds of businesses across Europe and North America failed to extract value from their data. A number of challenges impede the creation of value from data by the financial service organizations (The Economist Intelligence Unit, 2012). Data quality is one of the challenges that are frequently mentioned in the literature impeding value creation from big data (Chen & Zhang, 2014; Fan, Han, & Liu, 2014; Janssen, Van Der Voort, & Wahyudi, 2016; Leavitt, 2013; Marx, 2013; Zhou, Chawla, Jin, & Williams, 2014; Zicari, 2014).

Data quality is a multi-dimensional construct (Eppler, 2001; Fox, Levitin, & Redman, 1994; Miller, 1996; Tayi & Ballou, 1998; Wang & Strong, 1996). In data quality the role of the data custodian is a key elements in the relationship between colleting and creating value from data. Data custodians process data from data producers/providers and generate information for data consumer. Wang and Strong's (1996) definition of data quality embraces the data custodian's perspective, " data quality is data that is fit for use by data custodian" (p. 6). To be fit for data custodian' task, the data should not only be intrinsically good, but also have proper representation, properly accessed and retrieved from the source, as well as appropriate for contextual use.

Insufficient data quality hinders the value creation from the data (Verhoef, Kooge, & Walk, 2015). Redman (1998) found that lack of data quality results in disadvantages a the operational, tactical and strategic level, including:

- Operational level: lower customer satisfaction, an increase in costs, and lower employee satisfaction;
- Tactical level: poorer decision making, longer time to make decision, more difficulties to implement data warehouse, more difficulties to reengineer, and increased organization mistrust;
- Strategy level: more difficulties to set strategy, more difficulties to execute strategy, contribution to issues of data ownership, compromise ability to align organizations, and diverting management attention.

Moreover, poor data quality is also associated with great amount of quality cost. According to Eckerson (2002) poor data quality costs US businesses $600 billion annually (3.5% of GDP).

Our objective is to understand the relationship between big data and data quality in financial service organizations. This research is among the first that studied the relationship between big data and data quality. For this purpose, we formulated a research approach which is presented in Section 2. We then discussed key concepts

and theories on the basis of state-of-the-art literature in Section 3. Big data will be measured by looking at its defining characteristics (the V's) and data quality will be measured using the commonly found dimensions in the literature. Next case studies and the corresponding findings is presented in Section 4. This resulted in the relationship between the big data characteristics and data quality dimensions. Finally, conclusions will be drawn in Section 5.

## 2 Research Approach

To attain our objective, i.e. investigating correlation between big data and data quality, three main steps were taken:

1. Literature review to further detail the big data and the data quality. This resulted in big data construct which is represented by its characteristics (V's) and data quality construct which is represented by its dimensions. The constructs are employed as the basis for investigating the case studies.
2. Online case studies from financial service organizations by content analysis to extract data quality issues and the corresponding big data characteristics. The result is list of data quality issues as a consequence of big data characteristics. These cases did not enable us to understand the causal relation;
3. In-depth case studies at financial service organizations to cross-reference and further refine the findings from online case studies. The refined list of data quality issues is mapped to the corresponding data quality dimensions.

First literature about big data characteristics and data quality dimensions were investigated. To review big data characteristics, we surveyed the literatures during 20112016 for any statements of 'big data' or 'data-intensive' in Scopus. 22,362 documents were found. After carefully checked the contents, we focused on nine papers that are strongly relevant with big data characteristics. The same approach was utilized to study the data quality concepts. Using the statements 'data quality' or 'information quality', we found 7,468 documents in Scopus. However, we concentrated to 13 articles that discussed comprehensively about data quality and its dimensions.

The aim of the desk research was to find relevant cases. To explore the relationship between big data characteristics and data quality in financial industry, a desk research to online articles and corresponding white papers was conducted with systematic approach. The search started with narrowing down 10 biggest banks Europe based on Banks Daily's ranking [1] and 10 biggest insurance companies in Europe based on Relbanks's ranking [2] to keep the focus of this research. The search is conducted through Google Search with keyword "big data" <institution name> (e.g. "big data" Barclays). From the 2000 search results (10 Google Search pages of 10 search result

[1] See http://www.banksdaily.com/topbanks/Europe/market-cap-2015.html

[2] See http://www.relbanks.com/top-insurance-companies/europe

per page for each institution), 2 of the authors independently selected relevant articles which results in a list 32 articles that were relevant with big data quality and produced within
5-years' timeframe (2011-2016). After further analysis, seven online cases were selected providing sufficient details (e.g. mentioning data input, information output, and problematic big data quality issues) for being able to analyze them, as described in Table 1. The cases were analyzed for its big data characteristics and data quality dimension using content analysis of the case studies' documents and interview transcripts using NVivo software. Content analysis has been widely used in qualitative study to analyze and extract information from text, web pages, and various documents (Hsieh & Shannon, 2005).

**Table 1.** Online cases that are used in this study

| Case | Organization | Big Data Objective | Source |
|---|---|---|---|
| **1** | ING Bank | Customer retention | https://goo.gl/RTWLh9 |
| **2** | Barclays | Customer retention | https://goo.gl/BEWqOI |
| **3** | UBS Bank | Risk identification | https://goo.gl/ZNwO6H |
| **4** | Allianz Insurance | Fraud detection | https://goo.gl/XPLwLo |
| **5** | ING Bank | Fraud detection | https://goo.gl/KaomAQ |
| **6** | Barclays, RBS Bank | Complaint monitoring | https://goo.gl/hQHxCe<br>https://goo.gl/MS8c1Z |
| **7** | BBVA | New product proposition | https://goo.gl/KUtXn5 |

In addition, we conducted three in-depth case studies to confirm and refine our findings from the previous step. It is important to see how the findings implemented in real-life practices as well as to find out the possible missing challenges. The criteria of case study selection were defined as follows: 1) the organization must be an information-intensive financial service organization; 2) the organization should make use of big data; 3) The organization is willing to cooperate and share information that are required to conduct this study. Three case studies were created by conducting interviews and investigating documents. The summary of offline case studies are presented in Table 2.

**Table 2.** In-depth cases that are used in the study

| | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| **Organization** | Retail banking | Retail banking | Insurance |
| **Big Data Case** | Balance Sheet Reduction (Risk Management) | Credit Risk Assessment (Risk Management) | Single Customer View (Customer Acquisition and Retention) |

| | | | |
|---|---|---|---|
| **Project's Goal** | Ensuring mortgage data quality meets the buyer's expectation | -Assessing the most appropriate credit risk level of a company<br>-Providing the most suitable loan | Obtaining a single view of a customer from multiple databases to improve customer service experience |
| **Information output** | Mortgage files (supporting data about mortgages) | -Credit risk level<br>-Most suitable loan for the company | A single customer view/profile |

# 3 Literature background: Key Concepts

## 3.1 Big Data Concept

Big data is used in various ways and has no uniform definition (Chemitiganti, 2016; Ward & Barker, 2013). Big data is often described in through white papers, reports, and articles about emerging trends and technology. A lack of formal definition may lead to research into multiple and inconsistent paths. Nevertheless, there is consensus about what constitutes the characteristics of big data. The big data have changed over time. As the initial big data characteristic the three V's of Volume, Velocity, and Variety were introduced by Gartner (2001). Later, IBM added a new V called Veracity, which addresses the uncertainty and trustworthiness of data and data source (2012).

The V's continues to evolve to 5 V's (Leboeuf, 2016), 8 V's (m-Brain, n.d.), and 9 V's (Fernández et al., 2014). Our literature review that 11 different V's are mentioned in the literature and reports. As our objective is to take a comprehensive view we take all V's into account and define these V's to avoid any confusion about overlap between these characteristics. The characteristics and their definitions are presented in Table 3. These will be used to analyze the big data used in the case studies.

**Table 3.** Big data characteristics

| No | Big data Characteristics | Defined characteristic of the data |
|---|---|---|
| **1** | Volume | Huge size of the data (Douglas, 2001) |
| **2** | Velocity | Unprecedented speed of data creation and data must be must be processed in a timely manner (Douglas, 2001) |
| **3** | Variety | Various sources of the data and diverse format of the data (structured, semi-structured, unstructured data) (Douglas, 2001) |
| **4** | Variability | Changing meanings and interpretations for the data based on its context (Owais & Hussein, 2016) |
| **5** | Veracity | Questionable trustworthiness of the data (authenticity, origin/reputation, availability, accountability) (Tee, 2013) |
| **6** | Validity | Questionable data generation with respect to regulations and procedures (compliance) (Hulstijn, Jagt, & Heijboer, 2011) |
| **7** | Volatility | Huge and up-to-date data needed for temporary and quick action (Owais & Hussein, 2016) |
| **8** | Visibility | Many invisible relationship from the contents inside the data (Owais & Hussein, 2016) |

| | | |
|---|---|---|
| **9** | Viability | Too many contents inside the data, but only few are useful (Dini, 2016) |
| **10** | Vast resources | The data need very high network bandwidth, huge computing power, large memory/storage for retrieving and processing (Dini, 2016) |
| **11** | Value | Questionable benefit derived from the data (Owais & Hussein, 2016) |

### 3.2 Data Quality (DQ) Concept

Data is the lifeblood of financial industry and DQ is key to the success of any financial organization (Zahay, Peltier, & Krishen, 2012). Financial players such as analysts, risk managers, and traders rely on data in their value chain. Poor DQ such an inaccurate or biased data may lead to misleading insights and even wrong conclusions. Financial industry was reported to loss $10 billion annually from poor DQ (Klaus, 2011). In addition, as a highly regulated industry, finance service organizations must conform to several regulations which require high DQ (Glowalla & Sunyaev, 2012).

Quality is rather a subjective term, i.e. the interpretation of 'high quality' may differ from person to person. Moreover, the notion may change based on the circumstances. Various definitions of DQ are found in the literature (Eppler, 2001; Huang, Lee, & Wang, 1998; Kahn & Strong, 1998; Miller, 1996; Mouzhi & Helfert, 2007; Tayi & Ballou, 1998; Wang, 1998; Wang, Kon, & Madnick, 1993). Overall, the term DQ depends not only on its intrinsic quality (conformance to specification), but also the actual use of the data (conformance with customer's expectation) (Wang & Strong, 1996). Knowing the customers and their business needs is a precursor to understand how DQ will be perceived.

DQ is a multidimensional concept (Eppler, 2001; Fox et al., 1994; Miller, 1996; Tayi & Ballou, 1998; Wang & Strong, 1996). However, there is neither a consensus on what constitute the dimensions of DQ, nor the exact meaning of each dimension (Nelson, Todd, & Wixom, 2005). The dimensions of DQ vary among scholars (Bovee, M., Srivastava, R., and Mak, 2001; Fox et al., 1994; Miller, 1996; Naumann, 2002; Wang & Strong, 1996). However, the most cited DQ dimensions are the dimensions of Wang and Strong (1996), They list sixteen DQ dimensions categorized into four thematic, namely intrinsic, accessibility, contextual, and representational quality, as shown in Figure 2.

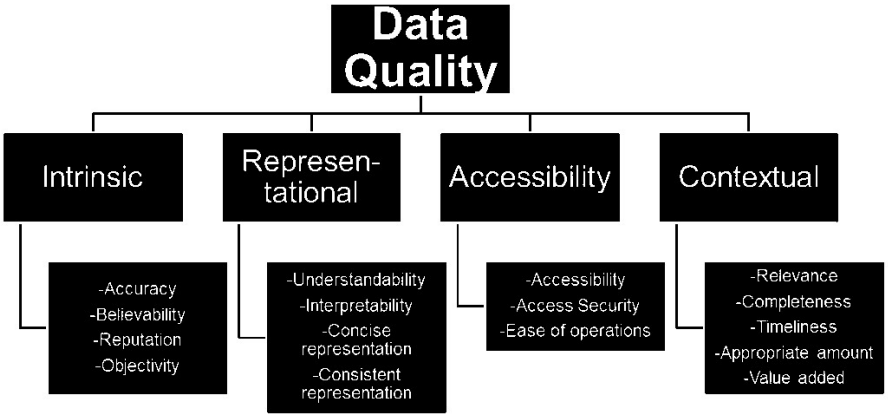


**Fig. 1.** DQ category and dimensions (adapted from Wang and Strong (1996))

*Intrinsic* quality is referring to internal properties of the data, e.g. accuracy, objectivity, believability, and reputation. *Accesibility* quality emphasizes the importance of computer systems that store and provide access to data. *Representational* quality consists of understandability, interpretability, concise representation, and consistent representation. *Contextual* quality, which highlights the requirement that DQ must be considered within the context of the task at hand, consists of value-added, relevance, timeliness, completeness, and appropriate amount.

## 4 Correlation between Big Data and Data Quality in Financial Service Organizations

Our aim was to investigate the relationship between big data characteristics and DQ dimensions as depicted in Figure 2. The big data characteristics and DQ dimensions are used to investigate the case studies. Using content analysis these are mapped and the relationship explored. There are eleven Vs that represent big data (their definition were given in Section 3) and four category of DQ that includes 16 dimensions (See Section 3 for their definition). We conducted seven cases that were carefully selected as explained in Section 2 to study the correlation. Three more in-depth case studies were performed to confirm and refine the findings and investigate the relationship in detail. DQ issues emerged from big data characteristics mentioned in case studies were explained as follow. Although big data characteristics and DQ dimensions are different, we found both ‘value’ refers to the same definition. Therefore we opted only one ‘value’ in the matrix, i.e. ‘value’ as a DQ dimension.

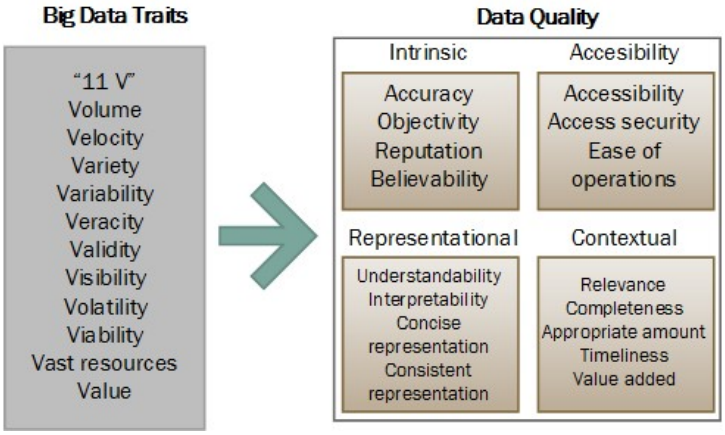


**Fig. 2.** Relating big data characteristics to DQ dimension

### 4.1 Volume

Volume was not frequently mentioned affecting DQ issue in the case. Huge size of data could increase chance to discover hidden patterns, such as finding a suspicious fraud. In addition, larger volume most likely leads to higher representativeness. However, bigger size could also bring troubles. In case 3 and 7, *information overload* was caused by volume of the data. It affected the level of amount of the data that is needed for the task in hand. For example, UBS Bank found in several situations that the transaction data for risk identification was too large for pre-processing.

### 4.2 Velocity

Many financial service organizations need real-time data for their activities such as fraud detection, complaint monitoring, and customer retention. Therefore, they were very concerned with the timeliness of the data. *Outdated data* is mentioned as an important issue by most cases (case 1, 2, 4, 5, 6, and 7). For example, data like credit card transactions is useful for the fraud detection and avoiding the fraud can have a huge impact, but becomes useless if it is not processed in real-time to predict and prevent the subsequent fraud.

### 4.3 Variety

Most cases mentioned the necessity to combine data from multiple sources in order to reveal more insightful value. However, incorporating many data sources results in a number of DQ issues, such as:

1) *Different value was reported by same field from multiple data* (case 3 and 6). An example is having a different zip code for the same person in different data sources;

2) *Inconsistent field's accuracy from multiple data* (case 3 and 6), e.g. which one is the accurate one from multiple zipcodes for the same person?;

3) *Varied population representativeness from multiple data* (case 3 and 6), e.g. some data have true objectivity but others like social media data tend to be biased and the data represents only certain group of population (e.g. youth, people with good internet connection);

4) *Inconsistent field's format from multiple data* (case 3 and 6, also confirmed in in-depth case 3). An simple example is that the content of field 'name' is varied in multiple data (e.g. John Clarke Doe, J. Doe, J. C. Doe);

5) *Inconsistent field's content from multiple data* (case 3 and 6). An example is having 'male' and 'man' in the 'sex' field;

6) *Different terminologies/semantics/definitions from multiple data* (case 2, 4, and 5). For example the term 'risk' in the data differs across data sources from various domains, especially data from non-specific finance domain;

7) *Various requirements for access from multiple data producers/providers* (case 1, 5, and 7). Some data providers provide a secure API, whereas others may prefer insecure API or even refer plain data transfer to ensure a high speed;

8) *Complex structure of the data* (case 1, 2, 4, 5, 6, and 7). An example is unstructured content from social media that contains lexical complexity;

9) *Duplicate and redundant data sources* (case 1 and 6, confirmed in in-depth case 1, 2, and 3). In offline case 1, there are two legacy systems for mortgages for the private banking and for the company which keep different record of information, but refer to the same mortgage;

10) *Incomplete content of the field in the data* (case 2 and 6, confirmed in in-depth case 1). In in-depth case 1, previously customers can use post bus as an address, but based on new regulation now they must use postal code. Because the postal code data was not required previously, the absence of this data would make the mortgage information considered as incomplete;

11) *Timeliness from multiple data* (case 3, 4, and 7) causes difficulties to combine those data in the same timeframe, e.g. statistics data from Eurostat or World Bank was collected at different points in time and cannot be combined to infer at useful insights;

12) *Complex relationship among data* (case 1, 2, 4, 5, 6, and 7). The more varied and numerous data fed into the system, the more complex the relationship resides in those data and the more complex it is to be combined In these cases we found that the data could not be combined as the data analysts were not able to unravel the complexity.

### 4.4 Variability

Variability of the data is rarely mentioned in the cases. The DQ issues originate from the use of social media data. In case 3, *different contextual meaning and sentiment for same content in the data* occurs, e.g. 'happy' and 'happy? ☹'. Real sentiments are hard to express. It brings difficulties to operate the data if the organization uses a traditional way (e.g. static algorithm) to process the content. Moreover, *the meaning of the words changes dependent on the context and the time* which brings in the need to dynamically interpret the sentiment. The word could

change from positive sentiment to neutral sentiment or even to negative sentiment after contextually use by communities along the time. For example, the word ‘advertisement’ which formerly gave a neutral sentiment currently shifts to a negative sentiment. It’s because nowadays people are annoyed by too many digital ads in web pages. On the contrary, some words may shift from neutral or negative sentiment to positive sentiment, such as ‘vegetarian’ that before was neutral now becoming more positive due to people’s conscience of nature reservation and personal health.

### 4.5 Veracity

Since many organizations involve many data sources into their data processing, they may face trustworthy issues on the authenticity, origin/reputation, availability, and accountability of the data, especially with the data is freely available in the Internet. The following DQ issues were found

1) *Inaccurate content often found from self-reported data* like social media (case
2). For example complaint came from black campaigner or fake account;

2) *Unclear reliability and credibility of data providers* (case 3, confirmed in indepth case 2), e.g. blogs or untrusted media;

3) *Unclear ownership of the data* (case 2, confirmed in in-depth case 2) may discourage organizations to use the data because they might not able to access the data if there is dispute in the future regarding commercial use of the data;

4) *Unclear responsibility to maintain content of the data* (case 2) might hinder use of the data for long term because the data could be complete and timely at the moment but useless in the future if the content and update of the data is not managed properly; the data from untrusted data source such as social media probably tends to have low objectivity, i.e. representing only portion of population (case 2, 3, 6, and 7).

### 4.6 Validity

Validity strongly represents the compliance of data generation with respect to procedures and regulations. Finance service organizations are among institutions that are mandated to strictly comply with external regulations such as privacy law and confidentiality agreement, as well as internal regulations and procedures, such as SOPs for data entry, service level agreements with partners and among internal units. Hence, the validity of the data should be carefully assessed beforehand because invalid data may bring trouble in the future.

Validity impacts the following DQ issues are the following.

1) *Inaccurate content of the field in the data due to manual entry* (raised from offline case 1 and 3) creates difficulties to understand the data, e.g. wrong address, wrong postal code, or wrong spelling in mortgage data because of disobedience to DQ control procedures;

2) Wrong coding or tagging in the data (case 3);

3) *Uncertainty about the right to use the data*. For example no knowledge about licenses or the impact of the privacy regulation (case 1, 2, and 3, confirmed in in-depth case 1) might limit or even remove access of the organizations to personal data;

4) *Difficult to extract value from anonymous data* (case 1, 2, and 3) as a consequence of privacy compliance because person-related field (e.g. name, phone number, email address) is the primary key of multiple data that are going to be combined; 5) Anonymous field makes the data become incomplete for the task in hand (case 1, 2, and 3).

### 4.7 Visibility

Almost all the cases mentioned that it is difficult to discover the relationship among variables within the data. For example, it's difficult to reveal which group of ages that have increasing internet banking usage over time in certain country by only viewing the data. Moreover, the more sources combined in the process, the more variables are added and the more complex relationship among the variables. *Unless the organizations build capability to visualize big data, that relationship is difficult to discover* (case 1, 2, 3, 4, 5, 6, and 7).

### 4.8 Vast resource

Some cases mentioned that *vast resources are essentially required in order to retrieve and process the data* (case 2 and 5). Retrieving huge size, very rapid generation, variety of the data needs, sufficient network bandwidth (especially if the organizations decided to put the data analytics platform in the cloud), computing power, and storage. Moreover, *data engineering skills* are required to retrieve and operate the data. Besides that, to discover the relationship among variables in the data and finally get the insight from the data organizations require *data scientist skills* (case 1, 2, 4, 5, 6, and 7).

### 4.9 Volatility, Viability, Value

No case mentioned volatility and viability characteristic of big data influence DQ. An explanation for this is that these factors are less essential for finance service organizations. Meanwhile, value is not coded from the investigated cases because it is conflicting with value-added dimension of DQ and 'value' is not big data specific.

## 5 Mapping big data and data quality

From the aforementioned DQ issues that were resulted by big data characteristics, each issue was then mapped into DQ dimension, as shown in Fig. 3. The corresponding case number either online or offline are put near the arrow.

The finding indicates there are no relationship between viability and volatility characteristic of big data with DQ in the investigated finance service organizations.

The most dominant correlation is Velocity-Timeliness that were found in all online cases. The relationship reflects that finance service organizations perceive the rapid generation of the data and real-time use of data, such as credit card transaction data or insurance holder's claim, plays an important role to create timely value of data, such as for fraud detection. The next dominant correlation is Variety-Ease of operations, interpreted as inclusion of data from multiple sources that may come with inconsistent formats and conflicting contents makes organizations difficult to process the data. Variety-Value added follows behind, which indicate that value creation is strongly influenced by number of data sources and complexity level of content (unstructured) residing in the data. Another most dominant pair is Visibility-Value which reflect the need of visualization to quickly discover the relationship among variables in the data. Vast resources-Value added is the next, which indicates the need of vast resources (hardware, software, data engineers, and data scientists) to retrieve, exploit, visualize and analyze the data so the value from the data could be derived.

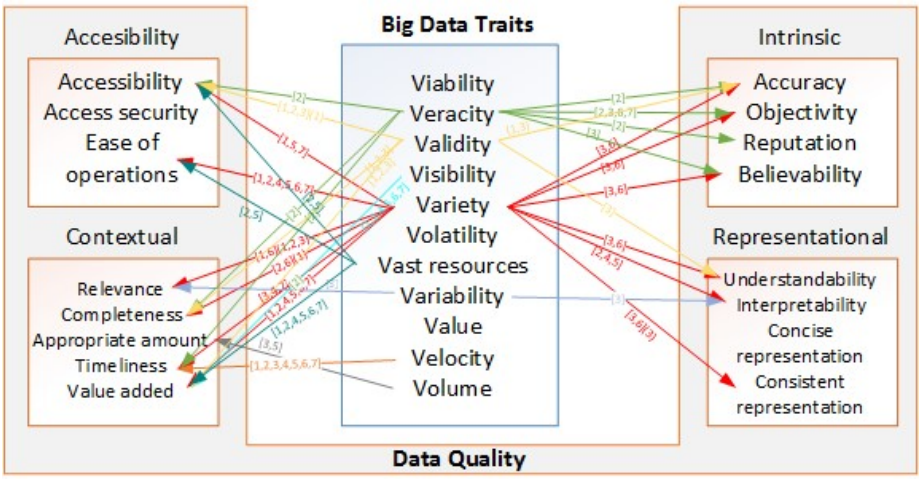


**Fig. 3.** Impact of big data characteristics on DQ dimensions ([x]: online case number, (x): offline case number)

The Table 4 was summarized from Fig. 3. It constructs a matrix that matches big data characteristics to DQ dimension. The number indicated in the pair represents the number of cases that mentioned the correlation.

**Table 4.** Number of cases from correlation pair between big data characteristics and DQ dimension

| Data Quality | Volume | Velocity | Variety | Variability | Veracity | Validity | Volatility | Visibility | Viability | Vast Resources | Value | SUM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Intrinsic** | | | | | | | | | | | | **17** |
| *Accuracy* | 0 | 0 | 2 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | **5** |
| *Believebility* | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **4** |
| *Reputation* | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | **2** |
| *Objectivity* | 0 | 0 | 2 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | **6** |
| **Representational** | | | | | | | | | | | | **10** |
| *Concise representation* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| *Consistent representation* | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **3** |
| *Understandability* | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | **3** |
| *Interpretability* | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** |
| **Accessibility** | | | | | | | | | | | | **19** |
| *Accessibility* | 0 | 0 | 3 | 0 | 2 | 4 | 0 | 0 | 0 | 2 | 0 | **11** |
| *Access security* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| *Ease of operations* | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | **8** |
| **Contextual** | | | | | | | | | | | | **47** |
| *Relevance* | 0 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **6** |
| *Completeness* | 0 | 0 | 3 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | **7** |
| *Timeliness* | 0 | 7 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **11** |
| *Approriate amount* | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **2** |
| *Value added* | 0 | 0 | 6 | 0 | 0 | 3 | 0 | 6 | 0 | 6 | 0 | **21** |
| **SUM** | **2** | **7** | **40** | **2** | **13** | **13** | **0** | **6** | **0** | **10** | **0** | |

From big data characteristics, variety is the most dominant one in our cases of the financial service organizations. It influences all categories of DQ, i.e. intrinsic, representational, accessibility, and contextual DQ. The reason for this is that nowadays organizations utilize multiple data sources, for example the ones that have formerly been ignored – namely "long tail" of big data, as well as new generated ones (Bean, 2016). The next most influential big data characteristic is Validity which reflects organization's compliance to regulation and procedures, for example about use of personal data (e.g. privacy law, untraceable requests, and confidentiality agreements). Compliance to privacy is very vital for service organizations (Yu, Mu, & Ateniese,
2015), especially bank and insurance companies (Breaux, Vail, & Antón, 2006; Karagiannis, Mylopoulos, & Schwab, 2007). Moreover, validity affects the accessibility to customer's data in the long run, meaning that one day organization may loss its right to access the personal data if the customer or regulator requests to disclose or remove personal data. As a result, completeness of the data drops and value creation process (e.g. analyzing data) becomes more complex if anonymous data is the only way organization can use. Another dominant big data characteristic is veracity. Veracity or trustworthiness of the data is inevitable when multiple data sources are utilized to discover more insights (Leboeuf, 2016). Since veracity includes authenticity, origin/reputation, availability, accountability of the data (Tee, 2013), unsurprisingly intrinsic quality which embodied the issues is mostly influenced by this characteristic.

As depicted in Table 4, the most correlated category of DQ dimension is contextual quality. It is unsurprising because every organization tries their best for extracting contexts from big data. Two dimensions from contextual quality are dominant in the finding, i.e. value-added and timeliness. Since today's organizations struggle creating business value from the data (Reid et al., 2015), the value from use of the data needs ample research. Another dominant correlated DQ dimension is accessibility which sounds the awareness of the financial service organizations to compliance.

## 6 Conclusion

The objective of this paper is to investigate the relation between big data and data quality. This study is among the first that investigated the complex relationship. To attain the objective, we conducted literature review, online and offline case studies in financial service organizations. Seven online case studies were initially performed to reveal the correlation, followed by three offline studies for cross-referencing and refining the findings. DQ issues raised from the case studies are then coded and mapped into the corresponding pair of big data characteristic and DQ dimension using content analysis. This provided detailed insight into the relationships between the V's of big data and dimensions of DQ. The Vs' take a blackbox perspective on the data. It characterizes the data form the outside. Meanwhile, DQ is about the actual data and can only be determined when investigating the data and by opening the blackbox. The V's characteristics and DQ are similar in the sense that they provide insight about the data. They are complementary as the V's take a look from the outside and at the possible usage, whereas, DQ look at the actual datasets.

The most related pair is Velocity-Timeliness, which indicates the more rapid the data being generated and processed, the better timely the data to use. This is followed by Variety-Ease of operations (more data sources and more varied structure of the data, the more complexity to retrieve, exploit, analyze and visualize the data), Variety-Value (the more data sources and more varied structure of the data resulting in more difficult to create value from the data), Visibility-Value (the more hidden relationship within the data, the more difficult to create value from the data) and Vast resources-Value (the more resources needed to process the data, the more difficult to create value from the data). Except for Viability and Volatility all Vs of big data influence DQ. Concise representation and access security were not found to be DQ issues in the cases. Variety is the most dominant factor impacting all categories of DQ, followed by Validity and Veracity. This suggest that term 'big data' is misleading as in our research we found that most of the time volume ('big') was not an issue and variety, validity and veracity is much more important.

Our findings suggest that organizations should take care of managing the variety of data and also ensure the validity and veracity of big data. The most correlated category of DQ dimension is contextual quality, which includes value and timeliness as the most dominant correlated DQ dimensions, followed by accessibility. These findings suggest that more effort should be spent on improving contextual use of the data as well as ensuring long-term accessibility to the data.

Further research recommendation is to cross-reference the findings with big data implementation in other information-intensive domains, such as telecommunication, government, and retail for generalization. This findings also open avenue to develop tools to improve and manage big DQ.

## References


Akerkar, R. (2013). *Big data computing*. CRC Press.

Bean, R. (2016). Variety, Not Volume, Is Driving Big Data Initiatives. *MIT Sloan Management Review*, 1–5.

Beattie, C., & Meara, B. (2013). *How big is "big data" in healthcare? Oliver Wieman*. Retrieved from http://blogs.sas.com/content/hls/2011/10/21/how-big-is-big-data-inhealthcare/

Bovee, M., Srivastava, R., and Mak, B. (2001). A conceptual framework and belief-function approach to assessing overall information quality. Paper presented at the Proceedings of the 6th International Conference on Information Quality, *18*, 51–74.

Breaux, T. D., Vail, M. W., & Antón, A. I. (2006). Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. *Proceedings of the IEEE International Conference on Requirements Engineering*, 46–55. https://doi.org/10.1109/RE.2006.68

Chemitiganti, V. (2016). Big Data Use Cases Across Financial Services. Retrieved from http://www.wallstreetandtech.com/data-management/big-data-use-cases-acrossfinancial-services/d/d-id/1268649?

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Science*. https://doi.org/10.1016/j.ins.2014.01.015

Dini, P. (2016). *Big Data : Processing Anatomy*.

Douglas, L. (2001). *3d data management: Controlling data volume, velocity and variety*. Gartner.

Eckerson, W. W. (2002). Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute*.

Eppler, M. J. (2001). A Generic Framework for Information Quality in Knowledge-intensive Processes. *Proceedings of the Sixth International Conference on Information Quality*, 329–346.

Fan, J. Q., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). Big Data with Cloud Computing: An insight on the computing environment, MapReduce, and programming frameworks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(5), 380–409. https://doi.org/10.1002/widm.1134

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19.

Glowalla, P., & Sunyaev, A. (2012). Process-driven data and information quality management in the financial service sector in the financial service sector.

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Huang, K.-T., Lee, Y. W., & Wang, R. Y. (1998). *Quality information and knowledge*. Prentice Hall PTR.

Hulstijn, J., Jagt, J., & Heijboer, P. (2011). Integrity of Electronic Patient Records. *Electronic Government: Proceedings of the 10th IFIP WG 8.5 International Conference, EGOV*

*2011*, *6846*, 378–391. Retrieved from http://www.springerlink.com.offcampus.lib.washington.edu/content/u4r1830566618753/

IBM. (2012). *Global Technology Outlook 2012*. Retrieved from http://www.research.ibm.com/careers/internships/index.shtml?lnk=intern-btn

Janssen, M., Van Der Voort, H., & Wahyudi, A. (2016). Factors influencing big data decisionmaking quality. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2016.08.007

Kahn, B. K., & Strong, D. M. (1998). Product and Service Performance Model for Information Quality: An Update. *Proceedings of the 1998 Conference on Information Quality*. Retrieved from http://mitiq.mit.edu/ICIQ/Documents/IQ Conference 1998/Papers/ProductServicePerformanceModelforIQ.pdf

Karagiannis, D., Mylopoulos, J., & Schwab, M. (2007). Business process-based regulation compliance: The case of the Sarbanes-Oxley Act. *Proceedings - 15th IEEE International Requirements Engineering Conference, RE 2007*, (October), 315–321. https://doi.org/10.1109/RE.2007.11

Klaus, K. (2011). On the Importance of Data Quality in Services: An Application in the Financial Industry. *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 148–152. https://doi.org/10.1109/EIDWT.2011.31

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *Mit Sloan Management Review*, *21*.

Leavitt, N. (2013). Storage Challenge: Where Will All That Big Data Go? *Computer*, *46*(9), 22–25.

Leboeuf, K. (2016). The 5 Vs of Big Data: Predictions for 2016. *Excelacom, Inc*, (1), 3–5. Retrieved from http://www.excelacom.com/resources/blog/the-5-vs-of-big-datapredictions-for-2016

m-Brain. (n.d.). Big Data Technology with 8 V´s. Retrieved from https://www.mbrain.com/home/technology/big-data-with-8-vs/

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Marx, V. (2013). THE BIG CHALLENGES OF BIG DATA. *Nature*, *498*(7453), 255–260.

Miller, H. (1996). The Multiple Dimensions of Information Quality. *Information Systems Management*, *13*(2), 79. https://doi.org/10.1080/10580539608906992

Mouzhi, G., & Helfert, M. (2007). A review of Information Quality research. *Proceedings of the International Conference of Information Quality (ICIQ) 2007*, 1–16. https://doi.org/10.1049/cp:20070800

Naumann, F. (2002). Quality-Driven Query Answering for Integrated Information Systems, *43*, 1–175.

Nelson, R. R., Todd, P. A., & Wixom, B. H. (2005). Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing. *Journal of Management Information Systems*, *21*(4), 199–235. https://doi.org/10.1362/026725705774538390

Owais, S. S., & Hussein, N. S. (2016). Extract Five Categories CPIVW from the 9 V ' s Characteristics of the Big Data, *7*(3), 254–258.

PricewaterhouseCoopers. (2013). *Where have you been all my life? How the financial services industry can unlock the value in Big Data*.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the Acm*, *41*(2), 79–82.

Reid, C., Petley, R., McClean, J., Jones, K., & Ruck, P. (2015). Seizing the Information Advantage. *PWC Iron Mountain*.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the Acm*, *41*(2), 54–57.

Tee, J. (2013). Handling the four 'V's of big data: volume, velocity, variety, and veracity. Retrieved from http://www.theserverside.com/feature/Handling-the-four-Vs-of-big-datavolume-velocity-variety-and-veracity

The Economist Intelligence Unit. (2012). *The Deciding Factor: Big Data & Decision Making*. *Capgemini*. Retrieved from http://www.capgemini.com/sites/default/files/resource/pdf/The_Deciding_Factor__Big_Data___Decision_Making.pdf

Verhoef, P. C., Kooge, E., & Walk, N. (2015). *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*. Routledge.

Wang, R. Y. (1998). A Product Perspective on Total Data Quality Management. *Communications of the ACM*, *41*(2), 58–65. https://doi.org/10.1145/269012.269022

Wang, R. Y., Kon, H. B., & Madnick, S. E. (1993). Data Quality Requirements Analysis And Modeling. *Proceedings of IEEE 9th International Conference on Data Engineering*, (April), 670–677. https://doi.org/10.1109/ICDE.1993.344012

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Source Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.2307/40398176

Ward, J. S., & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions.

Yu, Y., Mu, Y., & Ateniese, G. (2015). Recent advances in security and privacy in big data J.UCS special issue. *Journal of Universal Computer Science*, *21*(3), 365–368.

Zahay, D., Peltier, J., & Krishen, A. S. (2012). Building the foundation for customer data quality in CRM systems for financial services firms. *Journal of Database Marketing & Customer Strategy Management*, *19*(1), 5–16. https://doi.org/10.1057/dbm.2012.6

Zhou, Z. H., Chawla, N. V, Jin, Y. C., & Williams, G. J. (2014). Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives. *Ieee Computational Intelligence Magazine*, *9*(4), 62–74. https://doi.org/10.1109/mci.2014.2350953 Zicari,

R. V. (2014). Big data: Challenges and opportunities. *Big Data Computing*, 103–128.