Representational Quality Challenges of Big Data: Insights from Comparative Case Studies

Agung Wahyudi, Samuli Pekkola, Marijn Janssen

Representational Quality Challenges of Big Data: Insights from comparative case studies

Agung Wahyudi[1], Samuli Pekkola[2] , and Marijn Janssen[1]

[1] Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands
{a.wahyudi,M.F.W.H.A.Janssen}@tudelft.nl

[2] Tampere University of Technology, PO Box 541, 33101 Tampere, Finland
samuli.pekkola@tut.fi

Abstract. Big data is said to provide many benefits. However, as data originates from multiple sources with different quality, big data is not easy to use. Representational quality refers to the concise and consistent representation of data to allow ease of understanding of the data and interpretability. In this paper, we investigate the challenges in creating representational quality of big data. Two case studies are investigated to understand the challenges emerging from big data. Our findings suggest that the veracity and velocity of big data makes interpretation more difficult. Our findings also suggest that decisions are made ad-hoc and decision-makers often are not able to understand the ins and outs. Sense-making is one of the main challenges in big data. Taking a naturalistic decision-making view can be used to understand the challenges of big data processing, interpretation and use in decision-making better. We recommend that big data research should focus more on easy interpretation of the data.

Keywords: big data, interpretation, sense-making, naturalistic decision making

1 Introduction

Big data can provide a number of benefits such as better understanding the customers, effective and efficient the marketing effort and fraud prevention & detection, all aiming for creating a competitive advantage (Beattie & Meara, 2013; LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2013). Gaining these advantages is not easy as big data often originates from multiple sources, each having different and varying data quality (DQ). The diversity of formats and difference in quality makes big data complex to manage (Wahyudi & Janssen, 2016).

DQ is defined "as its scale of fitness for use by data consumers" (Wang & Strong, 1996, p. 6) which implies broad characteristics, i.e. not only intrinsic properties of the data but also other aspects, such as representation, accessibility and value-creation

context (Wang & Strong, 1996; Wang, Ziad, & Lee, 2002). Representational Data Quality (RDQ), a subset of DQ, emphasizes how the data is structured and comprehended by the data consumers (Wang & Strong, 1996, p. 21). It is represented by dimensions such as understandability, interpretability, consistent representation, and concise representation.

RDQ varies due to veracity and variety of big data. Inclusion of multiple sources brings varied levels of trustworthiness, interpretability, and representation. RDQ issues such as lack of metadata, heterogenous sampling periods, different meaning of terminologies, unstructured or semi-structured representation, variety of data format, and lack of a primary key on data are reported in literature as challenges for extracting values from big data (Hilbert, 2016; Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, 2012).

Such issues create a complex endeavor to any organization for interpreting big data for subsequent use like a decision making (Janssen, Van Der Voort, & Wahyudi, 2016). First interpretation problem is that stakeholders' interpretation often unfit to the senses so that it inhibits further action (Weick, Sutcliffe, & Obstfeld, 2005). Moreover, multiple interpretations on the same dataset are often occurred in a multiactor environment.

Many studies often take one viewpoint, such as focus on e data (Leavitt, 2013; Qiu, 2016; Yaqoob et al., 2016; Zhou, Chawla, Jin, & Williams, 2014), the process or solution (Geerdink, 2013; Merelli, Pérez-Sánchez, Gesing, & D'Agostino, 2014; Scarf, 2015; Wahyudi, Kuk, & Janssen, 2018), or the action (Hofmann, 2015; Osuszek, Stanek, & Twardowski, 2016; Power, 2014; The Economist Intelligence Unit, 2012). An perspective including all phases from raw data to the decision-making and including data, process and organizational aspects is missing.. Such a perspective should help us to arrive at a broader, socio-tech view of big data challenges

This paper looks at the process from big data to a decision making in an integrated manner. We start by providing the background of big data representational quality. Next, the literature is surveyed to identify the challenges of big data RDQ according to the big data usage cycle. In Section 4, we discuss our research approaches. Two cases regarding big data interpretation are presented in Section 5, i.e. a case in a telecom and a case in a manufacturing company. We elaborate the findings in more details in Section 6. Finally, the conclusions are drawn in Section 7.

2 Big Data Representational Quality

Turning big data into a decision is influenced by contractual governance, relational governance, big data analytics capability, knowledge exchange, collaboration, process integration and standardization, flexible infrastructure, staff, decision maker quality, and data quality of the big data sources (Janssen et al., 2016). Low quality big data provides little value, hinders people to interpret it, and results in questionable decisions.

DQ is defined “as its scale of fitness for use by data consumers” (Wang & Strong, 1996, p. 6). This implies that DQ ranges from internal dimensions such as accuracy and completeness to wider properties such as relevance to the task in hand, how secured the method of data retrieval, and how easy the data to be operated.

Representational Data Quality (RDQ) emphasizes how the data is structured and comprehended by the data consumers. It includes “aspects related to the format of the data (i.e. concise representation and consistent representation) and the meaning of the data (understandability and interpretability)” (Wang & Strong, 1996, p. 21). Concise and consistent representation describe how well and persistent is the content structure of the data. The extent of data consumers’ acceptance of the data is specified by understandability and interpretability.

The RDQ dimensions are defined as follow. Understandability is the extent to which data are clear, unambiguity and easily comprehendible (Wang & Strong, 1996). Interpretability is the extent to which language, units and data definitions are clear (Wang & Strong, 1996). Concise representation is the extent to which data are compactly represented without being overwhelming (i.e., brief in presentation, yet complete and to the point) (Wang & Strong, 1996). Consistent representation is the extent to which data are always presented in the same format and are compatible with previous data (Lee, Strong, Kahn, & Wang, 2002; Wand & Wang, 1996).

RDQ covers many aspects and requires that data is structured and fully comprehended by the data consumers so that it is interpretable, easy to understand, and represented in a concise and consistent manner. Yet, big data is often not structured. Velocity and veracity often prevent this. RDQ can be decomposed of a number of dimensions, including understandability, interpretability, concise representation, and consistent representation.

3 RDQ Challenges in Literature

In this section, we discuss the big data usage lifecycle. In each step, the main challenges from the literature are discussed.

3.1 Big data usage lifecycle: From Big Data to Decision

To analyze big data necessitates understanding of overall journey of big data value creation, i.e. from big data to action. literature shows several steps for the big data processes (Bizer, Boncz, Brodie, & Erling, 2012) (Chen, Mao, & Liu, 2014). In this research we use four interactive and iterative steps as depicted in Fig. 1. We refer to these steps as the big data usage lifecycle, as the use of big data is a continuous learning process.

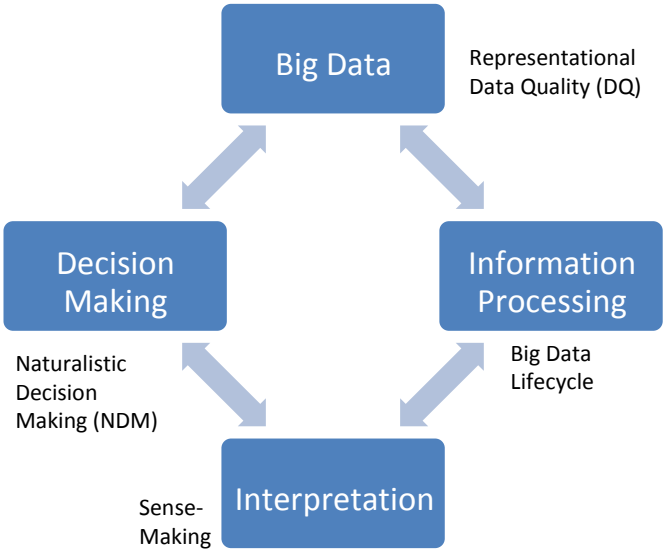

Fig. 1. Big data usage lifecycle

3.2 Challenges in Big Data step

Big data is commonly described by its characteristics. Initial big data characteristics consist of three V's, i.e. Volume, Velocity, and Variety (Gartner 2001). Over time, the number of V's is expanding. Currently 11 V's are identified, including Variability, Veracity, Validity, Volatility, Visibility, Viability, Vast resources, Value (Fernández et al., 2014; Leboeuf, 2016; m-Brain, n.d.).

From all characteristics of big data, RDQ is strongly related with the veracity and variety (Katal, Wazid, & Goudar, 2013). Veracity is the characteristic of big data that conveys questionable trustworthiness of the data (e.g. authenticity, origin/reputation, availability, accountability) (Tee, 2013). Meanwhile, variety imposes the use of various data sources and diverse format (i.e. structured, semi-structured, unstructured data) (Douglas, 2001).

Internal big data (i.e. data sourced internally) is owned and managed by the organization itself. As a consequence, the dataset usually has high RDQ, i.e. well-presented, in a standardized format, concise and structured. Meanwhile, other datasets, especially external large datasets (e.g. open data or social media), often have insufficient RDQ due to lack of proper description (e.g. unit) on observation records, unstructured content, lack of metadata, missing a primary key, among many others.

Variety of data sources introduces various levels of understandability and interpretability. Human-made image data, such as doctors' handwritings in medical records, is difficult for computers to interpret and analyze due to its unique style and the quality of personal handwriting (Strong, Lee, & Wang, 1997). The doctors' coding system may also differ from one hospital to another. Another example is the low resolution document scans, which is also difficult to read and understand. Lack of metadata and the observational unit creates an interpretability problem. Although the

variable name is usually known, the unit of measurement is not. This makes it difficult to interpret the value of the variable. Multiple interpretation could occur without metadata since terminologies may differ. Inclusion of unstructured contents is also problematic for interpretation. Ermolayev (2013) mentions 42% of respondents find unstructured content e.g. social media and email, too difficult to interpret. Furthermore, 40% of respondents believe that they have too much unstructured data (Zicari, 2012).

Due to variety of big data, some data may not comply with the organization's quality standards, such as conciseness. Data may have multiple variables merged together (e.g. variable income and population into variable per-capita) (Wickham, 2012). Sawant (2013) mentions that 90% of data are noise and Gantz (2011b) that only 0.5% of all data are analyzable.

Consistency is inevitably met with a variety of data sources and heterogeneous systems. For example, currency in a U.S. database is mostly in dollar, whereas those in a Japanese database are in yen. The format of date may also become an issue.

3.3 Challenges in the Information Processing Step

The V's characterizing big data complicate information processing. Insights are mostly derived from multiple diverse datasets (Wahyudi et al., 2018). Interpretation may begin directly by looking at the content of observation on the raw dataset, such as rain precipitation in a dataset from a weather station. However, most interpretations are resulted from multiple datasets combination, e.g. combination of weather data and flight data to determine the impact of bad weather to the departure/arrival delay. Combining many datasets needs a big data lifecycle that compromises of a number of stages, such as discover, access, exploit, analyze, and manage (Wahyudi & Janssen, 2016).

Discover involves activities for looking for the right datasets for the task at hand. In discover stage, activities like search, quality assessment, and making an agreement with data providers are important in matching relevant and required datasets.

Those data are retrieved in the *Access* stage. In the access stage, the stakeholders retrieve datasets from the providers and pool them in a central repository, e.g. a data lake. *Exploit* includes activities for data preparation and data transformation. In this stage, the datasets are prepared, cleansed, combined, transformed, and aggregated using different operations such as conditioning, filtering, manipulating, partitioning, reformatting, sorting, joining, merging, and grouping. In *Analyze* stage, the relationships within the data are investigated by using a model or a set of hypotheses. Various analytical methods are employed, e.g. predictive analytics, text mining, time series, trade-off analytics, machine/deep learning, and natural language processing. Once the data is processed, the result are disseminated and communicated to corresponding stakeholders. A number of media are introduced for such purpose, e.g. a dashboard, reports, alerts, or notifications (Matheus, Janssen, & Maheshwari, 2018). This keeps the stakeholders informed about the situation under concern.

Manage stage is the stage where all activities are orchestrated and managed to ensure smooth data processing sequences. The functions include data catalogs, metadata, process integration, and security.

There are a number of RDQ challenges in this process. First, substantial efforts are made to overcome multiple levels of conciseness and consistency. To get the data tidy and homogenous, a number of activities, such as cleansing, normalizing, preparing, transforming, and aggregating the data, need to be done in the Exploit stage. Interpretability causes difficulties to analyze the data. The lack of primary key in the data need more efforts to query and combine. Second, data with low interpretability is difficult to analyze and integrate with other data. For example, combining multiple datasets with a different sampling period and incorporating a dataset without a (clearly explained) metadata is challenging in Analyze step.

3.4 Challenges in the Interpretation Step

Interpretation can be conducted on the raw dataset or on the processed information from a data lifecycle which is further disseminated using different media (Matheus et al., 2018; Wahyudi et al., 2018). Interpretation of the data or processed data (i.e. information) relies on sense-making. Sense or meaning is defined as "mental representation of possible relationships among things, events, and relationships" (Baumeister, 1991, p. 15). Sense of data is made by using the sense-making process that relies on individual sense-making capacity, carved out through work, experiences, and training.

Making sense of the data or information is referred to as sense-making. "Sensemaking is a way station on the road to a consensually constructed, coordinated system of action" (Taylor & Van Every, 1999, p. 275). Sense-making is a process that has the seven identifiable characteristics (Weick (2005). First, it is grounded in identity construction, meaning that sense-making is a subjective process where individual change is derived from three fundamental needs: the need for self-enhancement, the selfefficacy motive, and the need for self-consistency. Second, sense-making is a retrospective process. People understand what they are doing only after the completion of the action. Sense-making is a backward process that the future action is determined by what the actor has learned in the past. Third, sense-making is enactive of sensible environments. Sense-making is shaped by the context of the environment in which people interact. Fourth, sense-making is a social process. People develop sensemaking in an organizational network of collectively shared meanings and agreed vocabularies. Fifth, sense-making is an ongoing process which neither starts nor stops. People chop certain moments out of continuous flows and extract cues for these moments. Sixth, sense-making is focused on and by extracted cues process. Seventh, sense-making is driven by plausibility rather than accuracy. For individual perspective, plausibility is more important than accuracy. People in any given situation are exposed to multiple cues, with multiple meanings, often intended for multiple audiences. They should make sense of it based on their capacity. The interpretation does not have to be accurate, merely plausible and acceptable.

Sense-making occurs along the data lifecycle (i.e. from discover to analyze). If the data fit in the stakeholder's sense, he proceeds to the subsequent actions, i.e. make a data-driven decision. However, it is not always the case. Misinterpretation or multiinterpretation phenomena are trivial in a multi-actor environment comprising stakeholders with different requirements and different levels of capacity such as knowledge, expertise, and skill. Different persons on the same job probably have different meanings on the same data. For example, an experienced sales manager may doubt information about top sales area that is out of his prediction and sense that has been carved through guidance, pattern, skill, and knowledge during his lifetime work. His initial prediction that is the area where most of his clients reside may not come up on the dashboard. As a result, he probably needs certain steps (e.g. validation of the information) prior to taking further actions. A competent engineer may recognize a cause of certain machine trouble just looking at few symptoms (i.e. fewer data). This is because he has trained his sense through guidance, pattern, skill, training and knowledge that have been encountered during his work. Meanwhile, others may need more datasets (e.g. environment's measures, machine logs) to understand the situation better. Their initial prediction may not align with the information on the dashboard or report. Consequently, validation of the information (e.g. more datasets that support the finding, other hypotheses' testing) prior to taking further actions.

Multi-interpretation occurs in a multiple-stakeholder situation when same information is interpreted differently. Front-end units (e.g. customer retention department) and back-end units may have a different interpretation of the same performance figure, e.g. customer handling time. The front-end unit may not be satisfied with the figure because the number of customers increases their dissatisfaction towards current service delivery. On the other hand, as long as the figure complies with the performance indicator or SLA targets, the back-end unit remains satisfied.

Sense-making relies on RDQ. Low interpretability data such as the absence or unclarity of metadata and the lack of convention of vocabularies or metrics within the organization may cause misinterpretations. Unavailability of metadata and unclear metadata descriptions create different interpretations among the data consumers. Multi-interpretation could be a result of the lack of definition of vocabularies or metrics. For example, throughput can be perceived differently by various stakeholders; some may consider it as the end-to-end transfer rate while others may refer it to the actual connection speed between the end user and the nearest point of the providers' equipment.

3.5 Challenges in the Decision Making Step

People take actions by their interpretations of the situation. The sense-making interpretation plays also an important role in the decision-making. Decision is primarily not driven by a set of choices but subjectively. This is the area of naturalistic decision making (NDM), which is "an attempt to understand how people make

decisions in real-world contexts that are meaningful and familiar to them" (Lipshitz & Klein, 2001, p. 332).

In a NDM four aspects are relevant. e.g. 1) process-oriented, 2) situation-action matching decision rules, 3) context-bound information modeling and 4) empiricalbased prescription. First, NDM views decision-making as a process-oriented activity that focuses on the cognitive process of proficient decision makers rather than predicting which options will be implemented (Lipshitz & Klein, 2001). This view is complementary to big data which is often used to show various options and insights. To be valid, NDM has to describe what information decision makers actually seek, how they interpret it, and which decision rules they actually use. Second, NDM follows situation-action matching decision rules. Proficient decision makers make a decision on various forms of matching on the situation in hand and not by concurrent choices. Appropriateness is more important than outcome superiority. Third, NDM is contextbound informal modeling. Proficient decision making is driven by experience-tied knowledge. Last, NDM offers empirical-based prescription, namely deriving prescriptions from descriptive models of expert performance.

There are a number of challenges in this step. First, very often the data or extracted information unfits to stakeholder's sense. Consequently, decision makers face a dilemmatic situation when the information is different with their sense.

Another challenge is that sense-making and naturalistic decision making are difficult to standardize since they are subjective. They may work perfectly in individual level but to code them as a standard in an organization and internalize the standard are impossible.

3.6 Overview of RDQ challenges

We summarize the RDQ challenges in every step of big data usage lifecycle in Table 1.

Table 1. Summary of RDQ challenges

Big data usage lifecycle step	**RDQ Challenges**	**Descriptions**	**References**
Big data	• Lack of proper description (e.g. unit) on variables • Too much unstructured content • Lack of metadata • Missing a primary key • Unconcise variables (e.g. date and hour in a single variable) • Inconsistency sampling periods • Inconsistent format of observations due to use of heterogeneous systems • Understandability issue for machine (e.g. human-made image and low resolution document)	Variety and veracity of big data sources introduced various level of understandability, interpretability, conciseness, and consistency.	(Ermolayev et al., 2013; Gantz & Reinsel, 2011a; Sawant & Shah, 2013; Wickham, 2012; Zicari, 2012; Zuiderwijk et al., 2012)

Information processing	• Need substantial efforts to overcome multiple levels of conciseness and consistency of big data • Data with low interpretability is difficult to analyze and join with other data	Most insights need to be extracted from multiple big datasets with different levels of RDQ which require a big data platform to process with in a big data lifecycle	(CHANGQING JI et al., 2012; Matheus et al., 2018; Wahyudi et al., 2018)
Interpretation	• Low interpretability data causes misinterpretation • People with different levels of senses (due to experiences and capabilities) may have different interpretation on a data • Multi-interpretation occurs in a multiple-stakeholder situation	Wrong or multiple interpretations on the same data might occur among data consumers	(Baumeister, 1991; Taylor & Van Every, 1999; Weick et al., 2005)
Decisionmaking	• Frequently the extracted information unfits to stakeholder's sense; Decision makers faced a dilemmatic situation when the information is different with their sense • Sense-making and naturalistic decision making is difficult to be standardized in an organization since they introduces subjectivity	Decision is hardly made naturalistically in case the information is out of sense	(Lipshitz & Klein, 2001)

4 Research Approach

This study investigates the challenges of RDQ of big data. To understand the process of turning big data into the decision, we investigate real-life scenarios. This necessitates deep understanding of the context (Dale et al., 1992; Davenport, Harris, & Morison, 2010). Then we can construct and decompose the decision-making process and divided it into notable steps and concepts, described earlier.

We conduct a qualitative and comparative case study (Yin, 2013) to gain deep understanding about the process of big data decision making. Case study research is a widely used in information systems research, and is well suited for investigating organizational issues (Benbasat, Goldstein, & Mead, 1987). We compared a telecommunication case and a manufacturing case. They were selected because of variety of challenges and aspects of representational quality.

The first case consists of six interviews with the people in charge of network performance management in the operation department. They included the network performance analyst/scientists, radio network specialists, big data platform engineer, and network performance manager. The interviews were conducted through a video call and lasted for 30 to 60 minutes. We also investigated various documents; servicelevel agreements, quarter reports (i.e. InfoMemo), network performance reports, SOPs, manuals, and the network configuration document.

For the second case, twelve open-ended interviews were conducted. The interviewees include Vice president of product management; Global product manager, Condition monitoring engineer, Business development manager, PLM manager,

Manager conceptual design & analysis, Chief mechanical engineer, Quality engineer, Unit manager, Finland; Sales manager, Northern Europe; Customer service engineer; Development engineer. The open-ended interviews focused on different themes related to service development and information usage. The themes included information needs, managerial practices, knowledge concepts, information technology and information systems, and knowledge and network dynamics. The interviews were conducted face-to-face in the company premises. They lasted for 30 to 90 minutes.

5 Case Studies

5.1 Case 1: From network performance data to network optimization decision (Telecom case)

Background.

A mobile telecom company in Indonesia is the first case. As of September 2017, 73% of the population of Indonesia are their subscribers. Around 152 thousand Base Transceiver Stations are serving the massive number of subscribers. In 2014, the company has been utilizing big data platform to support them in operation and maintenance. Operational effectiveness and cost efficiency are strived to achieve using the platform.

Findings.

The telecom company utilizes big data for a multi-vendor scheme in their business, but usually clusters the same brand in one region. However, in a specific region, multi-vendor approach may occur as it supports faster service deployment and keeps competitiveness among vendors high, e.g. in terms of price, performance, or service. Inevitably, the company has to deal with the complexity raised by multi-vendor environment.

Due to large customer base, the company collects vast amounts of data related to connection activity, for example call detail records, connection records, cell handover logs, authentication, authorization, and accounting records, and network performance data. Every network device collects the data and pools it into a centralized operational system support (OSS).

The first case study focuses on the network performance data, specifically the dataset related to drop call rate (DCR). DCR is the fraction of the calls, which were cut off before the speaking parties had finished their conversation and hung up. This fraction is measured as a ratio to all calls, and usually denoted in percentage. The bigger the ratio, the worse performance the network indicates. The company internally aimed at DCR rate of less than 2%.

Each vendor has different terminologies and mechanisms to derive DCR. For example, some vendors report DCR value directly in their proprietary applications while the others did not directly report DCR but its components, i.e. the number of drop calls and the number of total calls.

Regarding accessibility, also there are different data collection methods. Some vendors allow data collection directly from their databases while some provide only a streaming text file or a certain port for streaming the data. The network data is usually at an aggregated level. Mostly the level of granularity is on base transceiver station level although some providers may grant access to raw data, i.e. the network performance counter record.

Representation quality of DCR-related data was good in the case. This is understandable as the variables have defined names (e.g. “dropped call rate”, “drop call”, “total call”), and their values are generated digitally. The data is easily interpretable since its metadata is explicitly articulated in the vendors’ technical guidebook. Also, as the variables (i.e. the number of calls) has no unit, its interpretation is easy. The data is represented concisely as variables are not merged, and there is no consistent representation issue since the values are integers and unitless. The level of granularity of the data from every vendor is consistent.

The vendors have different naming conventions and data generation policies (such as sampling period and granularity), so the DCR-related data need to be prepared, tidied, filtered, and transformed before being aggregated on the level of base stations, clusters, regions, or national.

Data delivery is managed at the vendors’ OSS from where the data is moved to the data analytics cloud. Batch and real-time processing are separated in the vendors’ data processing platform (Marz & Warren, 2015). The batch layer retrieves the dataset and stores it in the master copy. Then it pre-computes the batch views on updated master datasets. The data is also forwarded to speed layer for real-time operation. This ensures new data being represented in query functions quickly. Once new batch views are available, the serving layer automatically swaps those in so that the most up-todate results are available. The DCR figure is visualized in a Tableau dashboard. A number of reports are also built by using Tableau application.

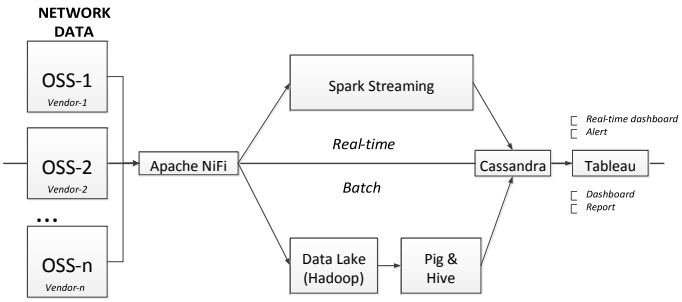

Fig. 2. Big data platform (for processing DCR-related datasets)

Interpreting DCR values is subjective, i.e. it varies between the analysts. The values may also contradict with the experienced field analysts’ presumptions, carved out from everyday complaints (i.e. social and sensible to the environment). The analysts have tacit understanding about problematic sites (i.e. retrospective aspect), so

they perceived it strange to have some clusters repeatedly harvesting complaints with a good DCR value. On the other hand, areas with fewer complaint sometimes have higher DCR than the more complaint areas (i.e. extracting specific cues aspect). Therefore, some areas are plausible with the corresponding DCR, but others are not (i.e. plausibility aspect).

The interpretation problems hinder the operation managers to make decisions on resource allocation for network optimization. He usually delegates the network performance review process to the analysts (i.e. process-orientation aspect). Based on his/her input, the manager may allocate resources to improve DCR of certain areas (i.e. situation-matching decision rules aspect), and follow them up (i.e. appropriateness aspect). He may also have experiences from previous network optimizations (i.e. empirical-based prescription aspect) before deciding to proceed with allocating resources (i.e. context-informal modeling aspect). As a response to the situation, the manager postponed the network optimization. Instead he initiated an investigation on finding how drop call value is derived or generated in every vendor's OSS.

The investigation resulted that every vendor employs different definition for a drop call. Every flow in a call flow could end the call, resulting different classes of call terminations. Every vendor has their own interpretation of which termination classes could be categorized as drop calls. Some vendors only include radio frequency termination while others include terminations that occurred in their own premises and neglect drops in interconnected equipment, or even include all abnormal terminations.

Retrospectively, the findings indicate the interpretation quality of the DCR-related data varies from initially good to poor due to different references used to define a drop call. Standardization (i.e. which termination classes belong to a drop call) might restore the interpretation quality of the data.

5.2 Case 2: From maintenance data to decision (manufacturing case)

Background.

The second case is a globally operating Finnish manufacturing company of about 1000 employees and service partners, and operations in 15 countries. About 90% of their products are exported worldwide. Service business plays a minimal role as only 1/3 of the turnover (total 105M€ in 2016) origins from services. In 2015, the company wanted to increase this share by better utilizing internal and external data for product maintenance and for advanced telemonitoring services.

Findings.

The service development began with mapping the needs and current maintenance processes and data available. In principle, the maintenance process was adequate and well defined, with a limited number of actors; engineers and repairmen. The products, when getting broken, are either fixed on-site (seldom) or disassembled from the larger

machine and shipped to Finland for maintenance work. However, there is no standardized information system but the data was recorded on spreadsheets or text documents, or on a very primitive database. The copies of these records were sent to the customers, but not used for any other purpose or data analysis. There is evidently a lot of room for improvements, new services, and new businesses.

When analyzing the data and the process in details, the situation turned out to be worse than initially expected. Although the process was quite straightforward, it was not documented, defined in details, or unambiguously supported. This meant that when the engineers did their work, they did it in their own way, used the document template (spreadsheets, text documents), the tool (paper, computer) and the style (database entry, text entry, picture) they preferred, and documented the details as they found appropriate. All this made the latter use of data difficult.

Let us illustrate this with an example. One of the ideas for new services was the ability to predict the machine (product) breaks. It was assumed that certain weather conditions increase the failure likelihood. As every maintenance report included a timestamp, testing this assumption was considered easy. It turned out to be impossible. The date information was ambiguous. It could indicate the time the machine breaks, the time the broken part is unmounted, the time the part is sent to Finland to the company's premises, the time it arrives there, the time maintenance work begins – or the time it is finished. The interpretation of the time stamp depended on the person filling in the report, making the merge of date and weather impossible.

Service development was thus problematic from several points of views. Representative data quality varies. Understandability was low as different data entries were named differently or placed in different cells in the spreadsheet. Data interpretations were impossible to make from the data and their consistency varied. On the other hand, no variables were merged which kept the concise representation high. The processes to record data, analyze it, and provide services were simple, but these problems prevented smooth information processing and the development of different systems. Data quality problems resulted that information needs and requirements for an information system could not be generated. Initial assumptions about what information could and should be used in the analysis and forthcoming services could not be tested.

The company had a hunch what could be interesting. They knew some variables would correlate, but not how. Our trials resulted in 54 issues that the engineers considered making sense. Thus, from the sense-making perspective already partially incomplete and poor quality data assisted in systems requirements specification, and generally, what is needed in building big data analytics capabilities and new services. This was the basis for decision-making in the development project, not for making decisions by the data.

6 Discussion

We identified a number of challenges in interpreting big data for a data-driven action from our cases. Those challenges are summarized in Table 1.

Table 2. Summary of the challenges per step in the big data use cycle

PHASE	DESCRIPTION	CASE-1	CASE-2
DATA	Representational Data Quality Understandibility: *extent to understand the content* Interpretability: *extent of interpretation of the content* Concise representation: *how concise the data is represented* Consistent representation: *how consistent the representation*	Understandibility Having proper variable names, digitally generated (high) Interpretability Initial quality is high (i.e. the metadata was easy to understand); but after first cycle, it is found out that the data had low interpretability due the use of multiple definitions Concise representation No variables were merged (high) Consistent representation The variables had a consistent format, consistent granularity level (high)	Understandibility Different data entries were named differently or placed in different cells in the spreadsheet (low) Interpretability Multi-interpretation on the timestamp of the data (low) Concise representation No variables were merged (high) Consistent representation Consistency is varied depending on the reporting unit (low-high) **Problems with data integrity, conformity, and accuracy**
INTERPRETATION	Sense-making Interpretation of raw datasets or processed information using the individual sense that has properties: 1) subjective, 2) retrospective, 3) enactive of sensible environments, 4) social contextual, 5) always ongoing, 6) targeted cues focused, and 7) plausibility-oriented	Some DCRs are appropriate, but some are out of sense, i.e. some cluster that harvests many complaints have a good DCR, but fewer complaint areas sometimes have higher DCR than the more complaint areas	The engineers' sense indicated correlations among some variables. **Problem with data quality prevented the testing of this assumption.**
INFORMATION PROCESSING	Big Data Lifecycle Includes cyclic stages of processing big data to better interpret the data, i.e. discover ➔ access ➔ exploit ➔ analyze ➔ manage applications. Due to low representation quality, it is difficult to define requirements for an information system.	Deployed an architecture for big data platform: batch, serving, and speed layer using various	**Problems with technological support and undefined data collection practices**

DECISION	Naturalistic Decision Making Decision making occurred at individual level relying on sense, that is characterized by: 1) process-oriented, 2) situation-action matching, 3) appropriateness seeking, 4) context-bound informal modeling, 5) empirical-based	Decision made is whether to allocate resources (i.e. worker, cost, time) for network optimization project. An investigation was conducted on inappropriate DCRs, indicating the different definition of drop call among vendors. After standardization, the interpretation becomes the same.	Decision made is in the development project, i.e. what is needed in building analytics capabilities and new services **Problems with the processes and understanding what the big data and say and is actually wanted or possible. This resulted in poor decision making.**

We summarize our findings from the two cases in Table 2. Different organizations perceive the representational quality of big data in a different way due to the inclusion of a variety of data sources. In Case 1, the telecom initially had a good representational data quality of the relevant data, i.e. easy to understand, interpretable, concise represented, and consistently represented. Later on, they found out that they had a multiple interpretation problem. In Case 2, the relevant data had low understandability, low interpretability, and varied consistency.

To deal with varying data quality and to serve the organizations with a ready-tocombine data that eventually leads to better user's interpretation, big data platform is required. As organizations may have the different legacy technology, business & IT strategy, data objectives, resources availability, and environmental landscape, they may have different requirements of the information system. The telecom in Case 1 had a clear list of functionalities that a big data platform should possess to attain all big data objectives within the organization. On the other hand, the requirements seem to be not straightforward in the manufacturing company in Case 2 that a number of follow-up actions need to be taken. The choice of big data solution, i.e. off-the-shelf commercial or open sourced, may differ across organizations. In Case 1, the organization preferred to use open-sourced big data solution because they did not want to depend on certain providers and build their internal capabilities with customized solutions.

The processed information may be perceived in various ways by multiple stakeholders within the organization. In Case 1, more experienced analysts perceived that some DCRs did not make sense, while some juniors just took the data for granted. The out-of-sense DCR represented some clusters that often harvest major complaints and according to their sense should have bad network performance, unfortunately turn out to have a good DCR. Meanwhile in Case 2, although the data quality is insufficient for further interpretation, the engineers already had a sense of the data, e.g. some correlations might exist among variables in the data. The properties of sense-making are clearly indicated in the cases, such as subjectivity (e.g. multi-interpretation occurred), retrospectiveness (e.g. interpretation quality that seems initially good turns out to be problematic later on), enactiveness of sensible environments (e.g. customer perception about the services did not match with the network performance data), social context (e.g. dealing with perception of multiple actors and reaching multistakeholder consensus), always ongoing process (e.g. DCR

consensus may change over time responding customer need and aligning organization's dynamic goal), targeted cues focus (e.g. only DCR-related information is relevant), and plausibilityorientation (e.g. the evaluation on DCR is to validate the appropriateness).

Certainly every organization has different types of decision that should be made. For example, in Case 1 the operational manager had to decide on allocating resources (i.e. worker, cost, time) for network optimization project based on his or his subordinates' interpretation of DCR information. Meanwhile, in Case 2, the decision was made about what is needed in building analytics capabilities and new services in the development project.

The way how the decision was taken in both cases suggested naturalistic manner. It is indicated by a number of NDM properties such as process-oriented (e.g. operation manager decision was based on analysts' reviewing process and his interpretation), situation-action matching (e.g. allocating resources for network optimization if the DCR is low or giving good points in SLA for vendors if the DCR is high), appropriateness seeking (e.g. further investigation was required to validate the inappropriate DCR), context-bound informal modelling (e.g. the sense of analysts automatically mind modeled the DCR under concern), and empirical-based (e.g. determining which network to be improved was carved out by working experiences).

Some recursive or retrospective actions most probably occurred. For example, in Case 1, the multi-interpretation on the DCR further leads to a thorough investigation about the appropriateness of the DCR. The investigation then revealed multiinterpretation problem on each vendor's data, i.e. they used a different definition of drop call in an entire call flow. A countermeasure initiative such as employing a standardization could help the organization to have the same interpretation.

7 Conclusion

Representational quality aspects affects how the decision-makers interpret the data. Big data that are coming from multiple sources result in challenges representational data quality (RDQ), which comprises understandability, interpretability, concise representation, and consistent representation of the data. One of the underlying cause is the involvement of various stakeholders resulted in fragmentation as activities are conducted by different people not being aware of what is happening in the whole process. Studies focusing on this whole process are rare.

RDQ challenges in every phase of big data usage lifecycle. Variety and veracity of big data sources introduced various level of understandability, interpretability, conciseness, and consistency. Challenges like lack of proper description (e.g. unit) on variables, too much unstructured content, lack of metadata, and among others are encountered. Most insights need to be extracted from multiple big datasets with different levels of RDQ which require a big data platform to process with in a big data lifecycle. The big data processes in the platform spans from *discover* to *analyze*. Data consumers face a number of challenges to process big data such as requiring substantial efforts to overcome multiple levels of conciseness and consistency of big

data and dealing with low interpretability data. The resulted information is subsequently interpreted by the data consumers by aligning the data with their sense. Different data consumers may have wrong or multiple interpretations on the same data due to low interpretability on the data, different levels of senses (due to experiences and capabilities), and multiple-stakeholder environment. Interpretation supports data consumers for performing a data-driven action such as decision making. A number of challenges need to tackled, e.g. out of sense data hinders decision making and sense-making & naturalistic decision making is difficult to be standardized in an organization since they introduces subjectivity. As suggested by the literature and confirmed by the case studies, the sense-making is identified by a number of properties, e.g. subjectivity, retrospective, enactive to sensible environments, social contextual, and plausibility driven. Sense-making can be challenged by the data quality. If the data quality is poor or poorly represented, its interpretations are consequently wrong. This may have significant impacts on the big data and big data projects since bad experiences and war stories get distributed.

The cases shows that RDQ challenges could occur in any stage of big data usage cycle. In case 1, the organization intially perceived that the data has good RDQ. However, their sense conflicted with the extracted information. Interpretability of the data was then found out to be problematic due to the use of different definitions of a certain terminology in a multi-vendor environment. On the other hand, in Case 2 were no terminology issues. There the challenges largely origin from poor quality data, which could not be intergrated. In both cases, there were RDQ challenges at the very beginning of the lifecycle. They just emerged differently, and for different reasons.

Acknowledgment
Part of the research was funded and supported by PT. Telekomunikasi Indonesia, Tbk. in the context of the Global Education Program 2015.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Baumeister, R. F. (1991). *Meanings of life*. Guilford Press.

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-bigis-big-data-in-healthcare/

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, 369–386.

CHANGQING JI, YU LI, WENMING QIU, YINGWEI JIN, YUJIE XU, UCHECHUKWU AWADA, … WENYU QU. (2012). *Big Data Processing: Big Challenges and Opportunities. Journal of Interconnection Networks* (Vol. 13). https://doi.org/10.1142/S0219265912500090

Dale, L., Michael, D., Laurie, J., Goodhue, B. D. L., Wybo, M. D., G, C. H. a, & Kirsch, L. J. (1992). The Impact of Data Integration on the Costs and Benefits of The Impact of Data Integration on the Costs and Benefits of Information Systems Introduction. *Misq*, *16*(September), 293–311. https://doi.org/10.2307/249530

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at work: Smarter decisions, better results*. Harvard Business Press.

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

Ermolayev, V., Akerkar, R., Terziyan, V., & Cochez, M. (2013). Towards evolving knowledge ecosystems for big data understanding. *Big Data Computing*, 3–55.

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

Gantz, J., & Reinsel, D. (2011a). Extracting value from chaos. *IDC Iview*, (1142), 9–10.

Gantz, J., & Reinsel, D. (2011b). Extracting Value from Chaos State of the Universe : An Executive Summary. *IDC IView*, (June), 1–12. Retrieved from http://idcdocserv.com/1142

Geerdink, B. (2013). A reference architecture for big data solutions: Introducing a model to perform predictive analytics using big data technology. *2013 8th International Conference for Internet Technology and Secured Transactions, ICITST 2013*, 71–76. https://doi.org/10.1109/ICITST.2013.6750165

Hilbert, M. (2016). Title: Big Data for Development: A Review of Promises and Challenges Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, *34*(1), 135–174. Retrieved from http://escholarship.org/reader_feedback.html%5Cnhttp://escholarship.org/uc/item/4nq8z7dn%5Cnhttp://www.escholarship.org/help_copyright.html#reuse%5Cnhttp://doi.org/10.1111/dpr.12142

Hofmann, E. (2015). Big data and supply chain decisions: the impact of volume, variety and velocity properties on the bullwhip effect. *International Journal of Production Research*, *7543*(December 2015), 1–19. https://doi.org/10.1080/00207543.2015.1061222

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

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and Good practices. *2013 6th International Conference on Contemporary Computing, IC3 2013*, 404–409. https://doi.org/10.1109/IC3.2013.6612229

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-bigdata-predictions-for-2016

Lipshitz, R., & Klein, G. (2001). Taking stock of naturalistic decision making. … *Decision Making*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/bdm.381/full

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

Marz, N., & Warren, J. (2015). *Big Data: PRINCIPLES AND BEST PRACTICES OF SCALABLE REAL-TIME DATA SYSTEMS*. *Big Data - Principles and best practices of scalable real-time data systems* (Vol. 37). Manning Publications Co. https://doi.org/10.1073/pnas.0703993104

Matheus, R., Janssen, M., & Maheshwari, D. (2018). Data science empowering the public: Data-driven dashboards for transparent and accountable decisionmaking in smart cities. *Government Information Quarterly*, (November 2016), 0–1. https://doi.org/10.1016/j.giq.2018.01.006

Merelli, I., Pérez-Sánchez, H., Gesing, S., & D'Agostino, D. (2014). Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/134023

Osuszek, L., Stanek, S., & Twardowski, Z. (2016). Leverage big data analytics for dynamic informed decisions with advanced case management. *Journal of Decision Systems*, *25*(sup1), 436–449. https://doi.org/10.1080/12460125.2016.1187401

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211. https://doi.org/10.1145/505248.506010

Power, D. J. (2014). Using 'Big Data' for analytics and decision support. *Journal of Decision Systems*, *23*(2), 222–228. https://doi.org/10.1080/12460125.2014.888848

Qiu, P. (2016). Big Data? More Challenges! *Technometrics*, *58*(3), 283–284. https://doi.org/10.1080/00401706.2016.1196946

Sawant, N., & Shah, H. (2013). *Big Data Application Architecture Q&A: A ProblemSolution Approach*. Apress.

Scarf, A. (2015). Modeling and Processing for Next-Generation Big-Data Technologies. *Modeling and Processing for Next-Generation Big-Data Technologies*, *4*(January), 283–317. https://doi.org/10.1007/978-3-319-09177-8

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data Quality in Context. *Communications of the ACM*, *40*(5), 103–110.

Taylor, J. R., & Van Every, E. J. (1999). *The emergent organization: Communication as its site and surface*. Routledge.

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

Wahyudi, A., & Janssen, M. (2016). Towards Process Patterns for Processing Data Having Various Qualities. In *Conference on e-Business, e-Services and e-Society* (Vol. 9844, pp. 493–504). https://doi.org/10.1007/978-3-319-45234-0
Wahyudi, A., Kuk, G., & Janssen, M. (2018). A Process Pattern Model for Tackling and Improving Big Data Quality. *Information Systems Frontiers*, 1–13. https://doi.org/10.1007/s10796-017-9822-7

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

Wang, R. Y., Ziad, M., & Lee, Y. W. (2002). Data Quality. Advances in Database Systems, vol. 23. Kluwer Academic Publishers.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science*, *16*(4), 409–421. https://doi.org/10.1287/orsc.1050.0133

Wickham, H. (2012). Tidy data. *Journal of Statistical Software*, *46*(10). https://doi.org/10.18637/jss.v059.i10

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231–1247.

https://doi.org/10.1016/j.ijinfomgt.2016.07.009

Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.

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. (2012). Managing Big Data. An interview with David Gorbet. Retrieved from http://www.odbms.org/blog/2012/07/managing-big-data-an-interviewwith-david-gorbet/

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Sociotechnical impediments of open data. *Electronic Journal of E-Government*, *10*(2), 156–172.

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2012). The meaningful use of big data: four perspectives -- four challenges. *SIGMOD Rec.*, *40*(4), 56-60. doi:10.1145/2094114.2094129

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171-209.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133-146.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the Acm, 39*(11), 86-95.