A Process Pattern Model for Tackling and Improving Big Data Quality

Wahyudi, Agung; Kuk, George; Janssen, Marijn

DOI
10.1007/s10796-017-9822-7

Publication date
2018

Document Version
Final published version

Published in
Information Systems Frontiers: a journal of research and innovation

Information Systems Frontiers
https://doi.org/10.1007/s10796-017-9822-7

A Process Pattern Model for Tackling and Improving Big Data Quality

Agung Wahyudi[1] · George Kuk[2] · Marijn Janssen[1]

Abstract
Data seldom create value by themselves. They need to be linked and combined from multiple sources, which can often come with variable data quality. The task of improving data quality is a recurring challenge. In this paper, we use a case study of a large telecom company to develop a generic process pattern model for improving data quality. The process pattern model is defined as a proven series of activities, aimed at improving the data quality given a certain context, a particular objective, and a specific set of initial conditions. Four different patterns are derived to deal with the variations in data quality of datasets. Instead of having to find the way to improve the quality of big data for each situation, the process model provides data users with generic patterns, which can be used as a reference model to improve big data quality.

Keywords Big data · Data quality · Information quality · Data processing · Process patterns · Reference model telecom

1 Introduction

Today's organizations collect an unprecedented amount of data as a result of datafication, which involves digitalization of business activities and objects as part of the organizations' processing chains (Bauer and Kaltenbock 2011; Mayer-Schönberger and Cukier 2013). Datafication covers a range of conventional routine tasks such as sensor reading and contract digitalization. In addition, the pervasiveness of recent technologies such as internet-of-things, mobile computing, social media, and parallel computation have enabled organizations to amass data from their infrastructures and their customers (Akerkar 2013). However, with data come from multiple sources, data quality often varies, and this makes it difficult for organizations to control, in particular when data are not uniformly cleaned or corrupted. Some present a clean set of data whereas others may be corrupted due to missing attributes, specification errors and so forth.

Data quality is generally measured by its degree of fitness for use by data users or consumers, capturing a broad perspective of the extent to which the intrinsic and the use value of big data can be realized and further harnessed (Wang and Strong 1996; Wang et al. 2002). Many studies suggest that organizations can gain benefits from the data if they succeed in unlocking value from the data (Huang et al. 2017). These benefits include: greater efficiency and profits (Dwivedi et al. 2017; Gantz and Reinsel 2011), and competitive advantages (LaValle et al. 2013; Manyika et al. 2011; Popovič et al. 2016; Zikopoulos et al. 2012). However, the question remains, how organizations can realize the potential value from data (Janssen et al. 2016; LaValle et al. 2013). Central to this value creation is the data user's perspective of how to ensure high-quality datasets can be correctly processed (e.g. Green and Kent 2002, Michener and Jones 2012, and Burton and Treloar 2009). Although the significance of data processing seems to be intuitive, many organizations have failed to implement this. A recent study (Reid et al. 2015) suggests that two-thirds of businesses across Europe and North America have been unable to unlock value from big data.

In this paper,[1] we seek to develop a process pattern model that an organization can use to deal with data of variable

✉ Agung Wahyudi
A.Wahyudi@tudelft.nl

George Kuk
george.kuk@ntu.ac.uk

Marijn Janssen
M.F.W.H.A.Janssen@tudelft.nl

[1] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[2] Nottingham Business School, Nottingham Trent University, Nottingham, UK

[1] An earlier version of this paper appeared in I3E conference (Wahyudi and Janssen 2016). This current version has extended the earlier version formulating a process pattern model, identifying data quality deficit patterns, and providing a resolution strategy to reduce the deficit.

Published online: 25 January 2018

quality. The model will provide a systematic approach to identify, assess quality, curate and combine data. Which processing should be followed will depend on the context, data quality, and operational goals. Some of the variable quality of data are generic to a majority of data-driven operations and are not unique to a specific type of organization. The data quality provides the initial set of conditions for selecting the process steps that are necessary to prepare the data for use. Such use patterns can be viewed as a practice, which can be reused or from which others can learn. We define a process pattern model as a recurring sequence of steps that results in attaining the specific operational goal. Given the context and certain starting conditions, the models can be followed to create value from the data. The process pattern model should be independent of the implemented technology and should enable organizations to create value from the data.

The objective of this research is to develop a process pattern model for tackling the variable quality of data. A process pattern model will be defined in our study as a proven series of activities to overcome a recurring problem in a particular context against a set of objectives, and under a specific set of initial conditions. We used a case study approach to examine the everyday practice in a data-driven company. In the following sections, we first present the literature review, followed by the research approach and the case study. Then we discuss the process pattern models and conclude with the implications of our findings.

2 Literature Background

To derive a process pattern model for tackling variable data quality, we review a number of concepts from the extant literature on data quality, data processing, and process patterns. These concepts are central to attaining data-driven operations and objectives. In particular, data quality is the kernel factor, which affects the data processing activities in this research. A thorough description of data quality is required not only to understand the concept better but also to provide a baseline reference model for present and future research. And with the variable data quality, although the required processes may differ from one another, they may share certain identifiable steps in the emerging patterns. Hence, we also include the literature on process patterns and models in our review.

2.1 Data Quality (DQ)

2.1.1 DQ Concepts

Data quality (DQ) has been widely acknowledged to be a prominent challenge in the big data literature (Chen and Zhang 2014; LaValle et al. 2013; Umar et al. 1999; Zuiderwijk et al. 2012). As described by Redman (1998), low DQ can have an adverse impact on operational, tactical, and strategic levels of organizations. They include high cost (to 8–12% of revenue), poor decision making, and increased difficulties in formulating a strategy. Wang and Strong (1996) define DQ as "*data that are fit for use by data users or data consumers*" (p. 6). This definition underlines the view that DQ is not only related to the data that it conveys but also the use of the data. Wang and Strong (1996) classify DQ into four types based on the perspective of a data user. They are: intrinsic DQ, which denotes that data have quality in their own right (e.g. accuracy); contextual DQ, which highlights the requirement that DQ must be considered within the context of the task at hand (e.g. value-added); representational DQ, which describes DQ in relation to data representation (e.g. interpretability); and accessibility DQ, which emphasizes the importance of computer systems that provide access to data (e.g. accessibility). High DQ is instrumental to value creation as "high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer" (Wand & Wang, 1996, p. 22). Table 1 provides the definition of the dimensions of each DQ type.

2.1.2 DQ Assessment

In many business processes and operations, organizations often have to combine datasets from various internal and external data sources. Dataset from each source is likely to vary in terms of DQ. Some data especially data from external providers may have low DQ, such as missing attributes, incorrect labels, among others. This makes DQ assessment central to managing and improving DQ. The extent of DQ assessment goes beyond improving the quality of data and can mitigate the unintended consequences of poor quality data on organizational decision making, strategy setting, and organizational performance (Lee et al. 2009).

DQ assessment is often performed by establishing a baseline and periodically monitoring the status across databases, stakeholders, and organizations. The status is represented by a quantifiable parameter from a certain DQ dimension, such as free-of-error rating as a metric of accuracy. For the baseline, organizations employ "what is right" value that can be derived from internal goals, standards, customer demands, de facto specifications, or benchmark with others. For example, the baseline of timeliness rating (Tayi and Ballou 1998) can be set to nearly 1.0 for financial organizations that emphasize fraud reduction as their strategy.

According to Lee et al. (2009), three major techniques are available for assessing DQ. They are: using a data quality survey, using quantifiable data quality metrics, and embedding data integrity analysis in the Total Data Quality Management (TDQM) cycle. The data quality survey elicits evaluations of multiple data quality dimensions from a number of stakeholders in the organization. The resulting assessment reflects subjectivity, i.e. the perceptions of stakeholders.

Table 1 Definition of Data Quality Dimensions (Taken from Wang and Strong 1996)

Dimension	Definition	Illustrative examples from the case study
Intrinsic		
Accuracy	Conformity to the real-world fact or value (Fisher and Kingma 2001)	False readings from uncalibrated sensor (IoT)
Believability	The extent to which information is complete, consistent, credible source, and accurate	Complaints from social media that were unverifiable
Reputation	The extent to which information is highly regarded in terms of its source or content	False news from unreliable media and data sources
Objectivity	The extent to which information is unbiased, unprejudiced, and impartial	Nonrepresentational and biased Twitter data of certain demographics of the population
Representational		
Understandability	The extent to which data are clear without ambiguity and easily comprehended	The lack of metadata in network performance data
Interpretability	The extent to which data are inappropriate language and units and the data definitions are clear	Non explicit defined units from sensor readings
Concise representation	The extent to which data are compactly represented without being overwhelming	Datasets from DBpedia may contain overwhelmed information for the task in hand (e.g. demography data)
Consistent representation	The extent to which data are always presented in the same format and are compatible with previous data	Inconsistent and incompatible data format from dropped call measurement albeit from the same data provider
Accessibility		
Accessibility	The extent to which data are available or easily and quickly retrievable	The lack of API access to retrieve the data from the office of statistics
Access Security	The extent to which access to data can be restricted and hence kept secure	Restricted access to SAP data within organization networks
Ease of operations	The extent to which data are easily managed and manipulated	Different representations of network dataset provided by different vendors
Contextual		
Relevance	The extent to which data are applicable and helpful for the task at hand	The inclusions of irrelevant network data logs for marketing purpose
Completeness	The extent to which data are of sufficient breadth, depth, and scope for the task at hand	Missing observations
Appropriate amount	The extent to which the quantity or volume of available data is appropriate	The frequency of IoT recordings in every 2 s than every 5 s
Timeliness	The extent to which information is highly regarded in terms of its source or content	The retention of outdated statistical data in the office of statistics
Value(−added)	The extent to which data are beneficial and provides advantages from their use	The deluge of data stored in the data lake but very minimum usage

The quantifiable data quality metrics are objective measurement formulas to assess data quality. The organization develops a collectively agreed-on metric for each data quality dimension. These metrics are then repeatedly applied. Data integrity analysis focuses on the direct assessment of adherence to integrity constraints in the database. These assessments are performed within the context of the TDQM cycle. Appropriate application of database principles in practice includes conformance to all data integrity rules, including user-defined data integrity. This technique is less intrusive and initially may not require direct involvement of data consumers.

2.2 Data Processing

Due to the variability of DQ in datasets, there is no uniform way to process them. As such, which process should be followed depends on the DQ. Normally, data are processed sequentially in data lifecycles, which encompass all facets of data generation to knowledge creation (Michener and Jones 2012).

There are many models of data lifecycles in the literature. Some prominent ones include: Data Documentation Initiative (DDI) Combined Lifecycle Model (Green and Kent 2002), DataOne Data Lifecycle (Michener and Jones 2012), and ANDS Data Sharing Verbs (Burton and Treloar 2009). The DDI Combined Life Cycle Model has eight activities in a data lifecycle, namely 1) study concept; 2) data collection; 3) data processing; 4) data archiving; 5) data distribution; 6) data discovery; 7) data analysis; and 8) repurposing. In a similar token, there are eight activities defined in the DataOne Data Lifecycle: 1) planning, 2) collecting, 3) assuring, 4) describing, 5) preserving, 6) discovering, 7) integrating, and 8) analyzing. Also, the ANDS Data Sharing Verbs is comprised of 1)

create, 2) store, 3) describe, 4) identify, 5) register, 6) discover, 7) access, and 8) exploit.

Although these models use different terminologies, all models of the data lifecycles share common activities, which reflect a data user's (or data client's) standpoint. Our study seeks to examine variant processes for improving DQ from the perspective of data users.

We summarize data processing lifecycle from the literature in Fig. 1. From a data user's perspective, the first step of data processing is to *discover* relevant data from data providers. It uses searchable interfaces to locate the data or by making agreements with data providers. This step may require user registration and sign in.

The next step is to *access* the data. Data can be accessed either through an automated system (i.e. using a Web link, perhaps passing through an authentication barrier and/or licensing agreement), or by an application to a data user.

Third, data need to be *exploited*. Data exploitation requires good technical metadata (fields, descriptions, metrics, etc.), which provide contextual information about the way the data are created. Cleansing, parsing, and other functions to prepare the data to be fit for analysis are also involved in this step. Moreover, it also includes the transformation of several different datasets into a common representation (format, coding scheme, and ontology), accounting for methodological and semantic differences while preserving a provenance trail. In addition, the dataset that varies frequently needs to be linked and combined with other datasets on a regular basis so that continual updates and insights or knowledge can be obtained.

The final step is to *analyze* the data. The goal of this step is to extract meaningful insight from the data using certain methods of data analytics such as statistical analysis, machine learning, predictive analytics, etc. According to Leek (2015), the data analytics fall into five major methods, i.e. descriptive, exploratory, predictive, causal and mechanistic. Descriptive methods tackle the questions relating to population characteristics from a data sample such as central tendency, variability, and proportion. If the findings also interpret the characteristics

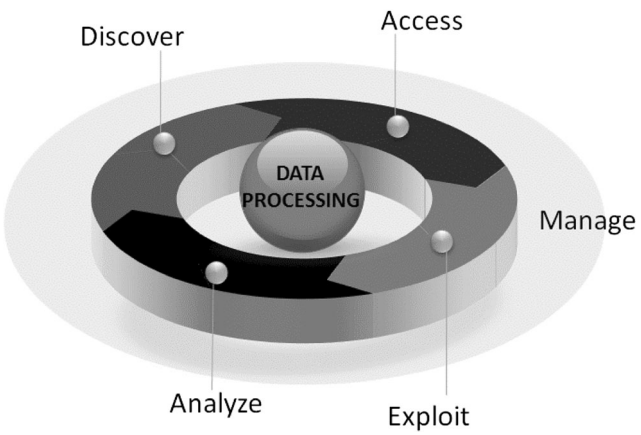

Fig. 1 Data processing lifecycle

and possibly held in a new sample, exploratory methods are then used. A data analyst can deploy predictive methods to predict measurement for individuals and inferential methods to predict measurement for the population. To investigate the causality among variables, causal methods are used if the investigation includes average measurement. Finally, mechanistic methods are used for deterministic measurement. These methods benefit users from the analysis in various media. A number of reports can be generated to communicate the findings. Moreover, a dashboard can be developed to display the real-time results and an alarm/warning system may be built to notify users about the findings (e.g. fraud detection) early. The results can also be used as an input from a decision support system, e.g. input of new product development.

2.3 Process Patterns and Models

The aforementioned data lifecycle provides the bases for data processing, which may vary based on the DQ of the data. For example, internal data of high DQ will require less processing whereas the DQ of external data (e.g. Twitter data) will need to be assessed and often require cleaning before they can be used and combined with internal data. The variable quality of internal and external data will result in the use of separate sets of protocols. The variant use of protocols in relation to DQ forms the basis of our process pattern model. The model is comprised of "process" and "pattern". According to Davenport (1993), a process is "a specific ordering of work activities across time and place with a beginning and an end, and also with clearly identified inputs and outputs: a structure for action" (p. 21). In line with this, Ambler (1999) defines a process as "a series of action to produce one or more outputs from one or more inputs" (p. 2), and a pattern as "a general solution to a common problem, one from which a specific solution may be derived" (p. 4). Patterns have been applied in various domains, e.g. architecture, economics, telecommunication, business, and software engineering (Becker et al. 2016; Buschmann et al. 1996; Yuan and Hsu 2017). Patterns in software engineering come in many forms including (but are not limited to) analysis patterns, design patterns, and process patterns. Hagen and Gruhn (2004) define process patterns as "patterns that represent proven process which solves a frequently recurring problem in a pattern like way" (p. 1). Process patterns provide flexibility in their use since one can select and apply a suitable process pattern according to the situation under study.

In the literature, there is no consensus about what should be included in a process pattern. Buschmann et al. (1996) mentions that a pattern must consist of contexts, problems, and solutions. A context of a pattern describes a design situation that gives rise to a design problem. The problem describes a concrete situation, which may emerge in the contextual application. A pattern should mention internal and external forces,

e.g. influences of customers, competitors, component vendors, time and money constraints, and requirements. The solutions describe the process that consists of a set of activities that are supposed to solve the problem if they are executed. Process patterns of overcoming DQ challenges assist organizations in creating value from the data. They also serve as catalogues and repositories to the organizations for future use. For our purpose, we define a process pattern as "*proven series of activities which are supposed to overcome a recurring problem in a certain context, particular objective, and specific initial condition*". Whereas a process pattern shows the actual steps that need to be executed, a *process pattern model* can be defined as "a recurring process pattern that results in the accomplishment of a certain operational goal".

3 Research Approach

Our research aim is to derive process patterns of how an organization can create value by tackling the variable quality of big data. We used a case study approach to inductively derive our process pattern model (Yin 2013). Qualitative case study research is widely used in information systems research and is well suited to understand the interactions between information technology-related innovations and organizational contexts (Nag et al. 2007). Such approach allows us to examine the everyday practice of tacking data quality in real-life contexts and explore contemporary problems in-situ. According to Yin (2013), the case study includes a variety of data collection instruments to ensure construct validity.

The following criteria were used in the selection of our case study organization: 1) it was a data-driven organization both in terms of its operations and business strategy; 2) it employed and combined data from multiple sources to attain its goals; and 3) provided invaluable insights into tackling the variable quality of data as a generic problem. Our case study organization was PT Telekomunikasi Indonesia Tbk., which was the biggest telecom in Indonesia. In addition to access to internal documents, the company gave the researchers permission to interview and shadow senior members of staff over nine months. The selection criteria and access allowed data triangulation and explored a contemporary problem. This gave us an opportunity of developing a deeper insight into the everyday practice, ensuring the validities of our constructs in our pattern model (Yin 2013). The personnel that we shadowed included five people who were in charge of the development and evaluation of the marketing programs in the mobile telecommunication unit, namely CDMA (Code Division Multiple Access) Division. They were: a marketing program planner, a business performance evaluator, a network performance evaluator, a data engineer/scientist, and a senior marketing manager. The shadowing involved participative observations of the implementation of several major data-driven marketing programs in various regions of Indonesia. We also followed and observed a number of data-driven activities including: the initial meeting to communicate the program plan with all stakeholders, the agreement between data providers and data users, the exploration of multiple data (e.g. billing, network, sales, customer, competitive intelligence, etc.), the program through the co-creation between marketing unit and local office, and the evaluation meeting of the programs. An example of the programs was free on-net call within Greater Jakarta region. The criteria used to select which product to be discounted and promoted in a certain region was described in Section 5.

Our focus was on the primary processes, which involved data handling and processing within the CDMA marketing department. We followed the processing steps that the department undertook to curate data and notably create value for its marketing programs. In total, we observed and analyzed the creation of seven marketing programs. Our analysis involved a detailed examination of all of the core documents related to the marketing programs including: business plan; evaluation and approval of the business plan; business processes to commercialize the marketing programs; evaluation plan of the marketing programs; final test report of production development of the marketing programs; and evaluation report of the marketing programs.

4 Process Patterns Model for Tackling Big Data of Variable Data Quality

Based on the concepts from the previous section, we propose a process pattern model as depicted in Fig. 2. Creating value from big data requires a fine-grained description of the underlying generic data processing activities with clear pathways to attaining certain goals and objectives.

The design and build stage of the data processing starts by considering the contexts and tasks that represent the objectives of data reuse. They include variant processes that use the same

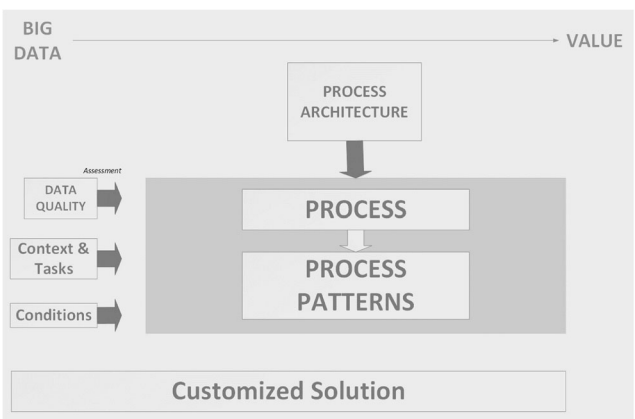

Fig. 2 Process pattern model for tackling big data quality

data, albeit with different objectives and outcomes. For example, the processes that support fraud detection and customer relationship management may often deploy the same data but rely on different data analytic methods. However, it is still possible to identify recurring process patterns that are generic to any organization despite differences in terms of the organization's type, sector, size, or stakeholders. For example, at the metadata level, the context and task conditions can be used to facilitate identification of reusable processes between organizations with a dual goal in detecting fraud and enhancing customer relationship.

Based on the assessed data quality, the data processing can differ from one dataset to another. For example, a dataset containing inaccurate observations needs to be cleaned prior to being exploited and combined with other datasets. With a good quality dataset, the processes will be straightforward with minimal or unnecessary steps in pre-processing.. However, the recurring pattern in each process may vary among cases where there exists a substantial difference in data quality. Take inaccurate observations as an example. Users not only have to clean the data but also need to solve an intrinsic quality problem of tackling biased (subjective) observations and untrusted data sources. The initial condition of an organization also determines the level of data processing when it involves legacy data and processes. For example, the legacy processes that processes data mostly in a batch manner in an organization will affect the target architecture. The organization may approach a hybrid architecture (e.g. lambda architecture) that combines batch and real-time data processes. Whilst the legacy processes are kept operating to maintain the current static reports or dashboard, the development and implementation of newly required processes are integrated to generate real-time information and output.

Exceptional handling can take place in some circumstances. In particular, in a critical situation, a specific process can be bypassed, and a bespoke and localized solution may be required as a special case of process pattern of data processing.

In the following case study, the context being investigated was the development of data-driven marketing programs in a large telecom. We observed how data quality influenced the data processing and proposed a number of patterns based on the observation. It will be discussed in the next section.

5 Case Study

The aim of the case study was to derive a process pattern model. For this reason, the case study involved multiple methods for data collection, including interviews, ethnography and document analysis. In the case study organization, the primary processes of PT Telekomunikasi Indonesia Tbk were selected for analysis. We analysed the primary processes of the CDMA marketing department in tackling with various data sources of variable data quality.

Historically, the programs from the marketing department were mostly driven by an intuition leading practice rather than evidence informed practice. This often resulted in ineffective targeting, segmenting, and positioning of products and also unsatisfactory returns on their investments. Some of the marketing programs did not have sufficient justification and were hard to measure and evaluate its outcomes. The situation was further compounded by a recent, unexpected increase of customer complaints, customer churn, and financial loss. The tight competition in the market and the increasing power of customers kept forcing them to respond to competitor moves and customer voices with attractive programs. In a recent relaunch of the marketing department, the aims of the programs were designed to reduce customers' satisfaction and to increase long-term profitability, e.g. discounting cash-flow products for which customers were willing to pay or giving massive national-wide promotion, which would result in network overload and congestion in densely populated areas.

Many data generated by the telecom including internal transaction data, customer data, machine logs, network performance data, and also external data such as social media, crowd-sourced maps, were exploited to target customers and markets better. By combining these data, they designed an attractive discount program. As mentioned by Verhoef et al. (2015), organizations can obtain value from the data in a bi-directional way, i.e. value to consumers and value to the organization (data user). From the program, customers of the telecom positively influenced by the benefits of budget tariff and perceived quality. Also, as the impact on customer experience, the telecom benefited by increasing its market share, improved brand recognition, and high return from the marketing investment. The way the telecom turned the data into value is illustrated in Fig. 3.

The company built an information system that had a number of functionalities to process big data. Prior to running the program, an initial kick-off meeting that included data providers and related departments was held. The marketing and IT department proposed a model that described how to turn the data into decision. From the model, they listed all the required datasets and made agreements with the data providers on access, metadata, cut-off time, etc. The IT department built a data lake to create a data pool as a means to accommodate data sources with very restricted, limited access, and other concurrency issues.

They also employed a number of tools to cleanse any low-quality data and parse out data that were unfit for further processing. Syncsort DMX-h Hadoop application was utilized to exploit the data. The application provided extracting, transforming, aggregating, and loading functionalities. Many datasets were combined and transformed based on the task at hand. The processes involved a number of execution

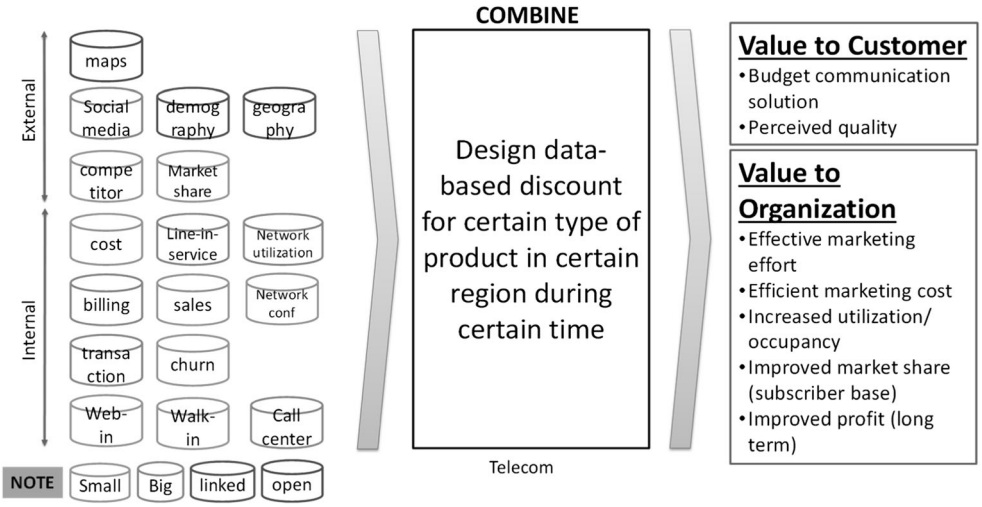

Fig. 3 Case Study: from data to value

activities, which included one or more datasets. They included: joining, aggregating, then manipulating fields, rejoining, etc. Furthermore, the data were analyzed using trade-off analytics and visualized using Microsoft Excel. The program was then proposed to the board of executives for decision. Sometimes, iterations occurred between the aforementioned processes.

Initially, extracting value from the data seemed straightforward using the functions provided in the information system. However, it was found to be complex in terms of tackling variable data quality. Since the telecom incorporated many datasets, their varied greatly. Internal big data and partners' data usually came with high intrinsic DQ because these data were self-managed (e.g. by periodic calibration of data-generating sensors, quality control, or using service-level agreements with partners). However, the datasets also had low accessibility, contextual and representational DQ. For example, call centre's recordings, which were mostly unstructured, could cause difficulties for technical staffs to process (ease of operation); many data were just thrown into the data lake but never used and not adding any value; machine logs had varied representations depending on the machine's manufacturer (consistent representation).

Unlike internal big data, external big data such as social media often had low intrinsic DQ. For example, Twitter data might be biased and over represented certain demographics, for example, opinions of the younger generation. The inherent bias in the data could lead to inaccurate outcomes if deployed to generalize to the assessment to the whole population. External big data were also reported to have low accessibility DQ (e.g. license/subscription fee which led to no/limited access to the data), low representational DQ (e.g. no metadata which caused a problem in understanding and interpreting the data), and low contextual DQ (e.g. outdated statistical data which was not fit in the task).

There were many ways that data processing could be deployed to tackle the various DQ issues. For example, accuracy problem could be resolved by cleansing the data before use. Some cleansing routines were indicated in the case study to underline a specific generic solution for a particular problem of DQ. From these, process patterns that were recurrent could be identified with each data processing for a particular DQ problem. The link between process patterns and DQ problem provided a basic building block of the architecture for our process pattern models.

6 Process Patterns

In identifying a process pattern model for tackling variable data quality, our case study organization showed that although it developed an information system with various functionalities, the data were processed in a sequence, following the data lifecycle from a high level of abstraction. From the literature and the case study, we derived the following typical data lifecycles. We used similar steps as found in the literature, but we extended this by including a managing process in any step of the data lifecycle. It comprised functions of connecting, controlling, and integration such that the data processing was executed sequentially. We also listed all functionalities related to every step of the nominal data lifecycle. The nominal data lifecycle together with the functionalities used to process the data is shown in Fig. 4.

The first step in the nominal data lifecycle was the *discover data* step. In this step, some functions such as search, assess quality and make agreement were employed. Search functions assisted them to quickly find relevant data from many datasets in the data lake. Assessing quality was important to determine whether actions to improve the quality were needed in the subsequent steps. In order to use the

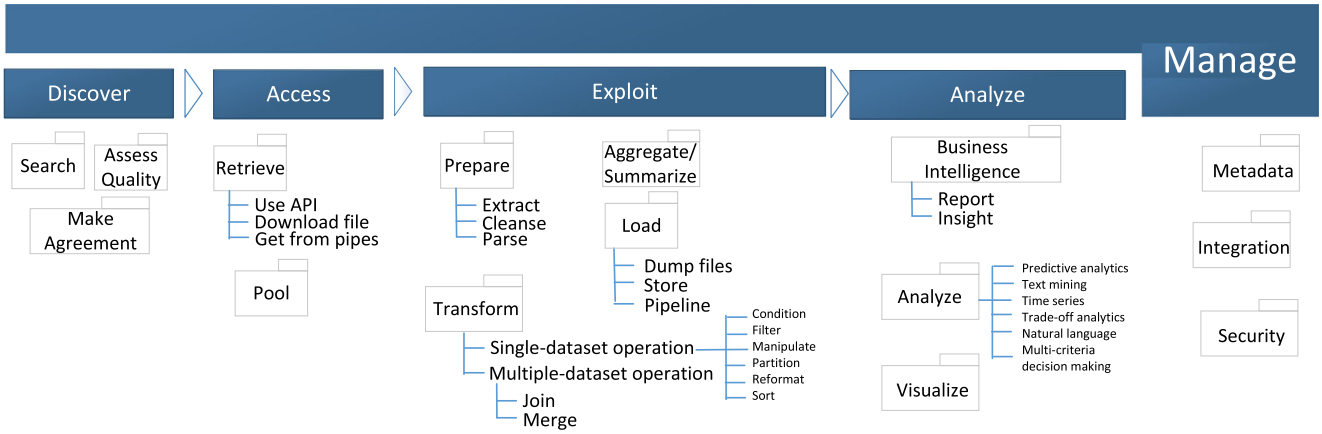

Fig. 4 Process architecture for processing big data in the case study

data properly, the organization made agreements with the data providers on:

1) what data should be included in the process?
2) how to retrieve the data?
3) when was the cut-off time or the retrieval time?
4) how to read the data? and
5) what if the data were not intrinsically good (e.g. corrupted)?

The *access data* step was consisted of retrieving and pooling the data. Retrieving the data was strongly related to accessibility. A number of activities were used, such as query, flat file transfer, or process pipeline. Sometimes organizations created a data pool in the data lake for several reasons, such as limited/restricted access, concurrency issue, etc.

The third step, *exploit data* was one of the most challenging steps in terms of the application complexity. Because there was seldom a single application that encompassed all the functionalities, various applications having separate functions were composed together to perform data exploitation. Interoperability and standardization were key success factors to get all applications working together. Some functions in this step were preparing, transforming, aggregating, and loading the data. In the "preparing step", some data might need to be extracted because they were retrieved as compressed flat files, cleansed because they contained low intrinsic quality (e.g. low accuracy), or excluded because their original representation was not fit for further processing. The organization transformed the data using single-dataset and multi-dataset operations. Functions such as conditioning, filtering, manipulating, partitioning, reformatting, sorting, joining, and merging were selected based on the task at hand. The combination and iteration of those functions were found very often. Aggregating the data was supposed to reduce the data based on certain fields. The outputs were then loaded either to dumb flat files, stored in the relational databases, passed to HDFS, or put into the pipeline for the next process.

The next step was to *analyze* the data. Functions included in this step were business intelligence, analyzing, and visualizing the data. Business intelligence was used extensively to generate reports. Analyzing the data was the most difficult task because of potentials of creating value from the data. The data were analyzed using various analytical methods such as predictive analytics, text mining, time series, trade-off analytics, and natural language, depending on the task at hand. In the case study, the telecom exhibited trade-off analytics between the projected revenue (from existing customer and new subscribers) and projected cost (from revenue opportunity loss and marketing campaign expense). Visualizing data was important to quickly grasp insights (e.g. trend, relationship) between datasets.

The step of *manage* data was not part of the sequential process but managed all the of aforementioned data processing steps. It ensured the data pressing sequence run smoothly. It was conducted through metadata, integration, and security. Metadata was important in order to understand and interpret the data so that they could be reused. Integration ensured the involvement of many actors and the utilization of many applications could run smoothly.

From the case study, we found that every dataset had a variety of data processes depending on the DQ as illustrated in Fig. 5. The "ideal" situation occurred when the dataset had high DQ and was followed by four basic patterns for processing data.

Organizations often have to deal with datasets with low quality. This occurs mostly with datasets from external sources. The gap due to low intrinsic quality causes the *internal deficit*, i.e. a condition where the internal users perceive the internal properties of the data of low quality, such as biased, inaccurate, untrusted, etc. To overcome this problem, organizations can improve the authenticity of the datasets by conducting activities such as assessing their accuracy and representativeness, rating the credibility of the data providers, pre-processing the data, and so forth.

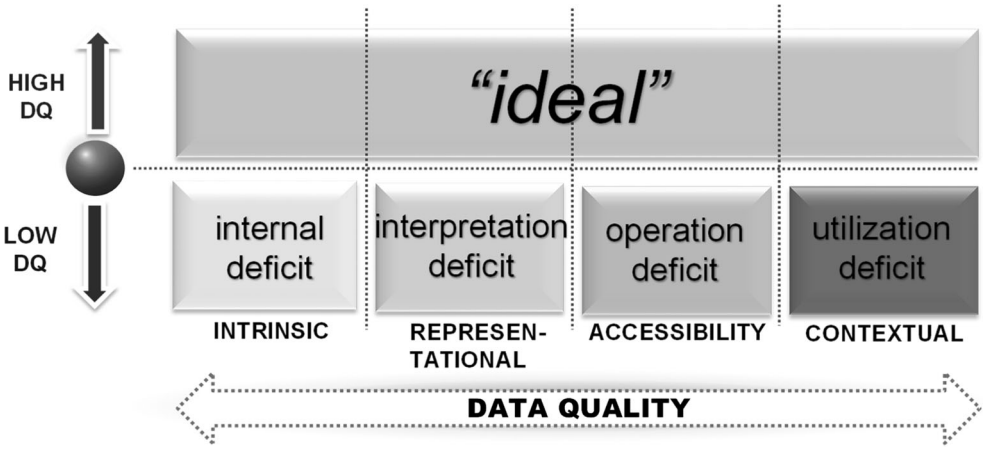

Fig. 5 Classification of patterns

If the datasets have low representational quality such as the inconsistent format of the observations, lack of metadata, etc., the organizations face the *interpretation deficit*, i.e. the gap between the actual understanding of the data and the correct interpretation. Interpreting precisely the observations is substantial for understanding the value and then further exploiting the data using data analytics. The strategies that can reduce the interpretation deficit include: 1) incorporating standardization (e.g. format of observation, metadata, etc.) and 2) providing repositories for the entire organization (e.g. terminology, semantics, sentiment library).

Organizations face difficulties to operate the datasets if they have low accessibility quality datasets, namely *operation deficit*, such as limited access (e.g. limited to aggregated data) may hinder them to create insights that may be revealed only using data with the highest granularity (e.g. individual level). The strategy to overcome this problem entails data providers/owners conduct certain activities such as privacy-by-design, service level agreement, and compliance with regulation (e.g. GPDR).

The *utilization deficit* occurs because of the gap between expected context and actual outcome from the use of the data,

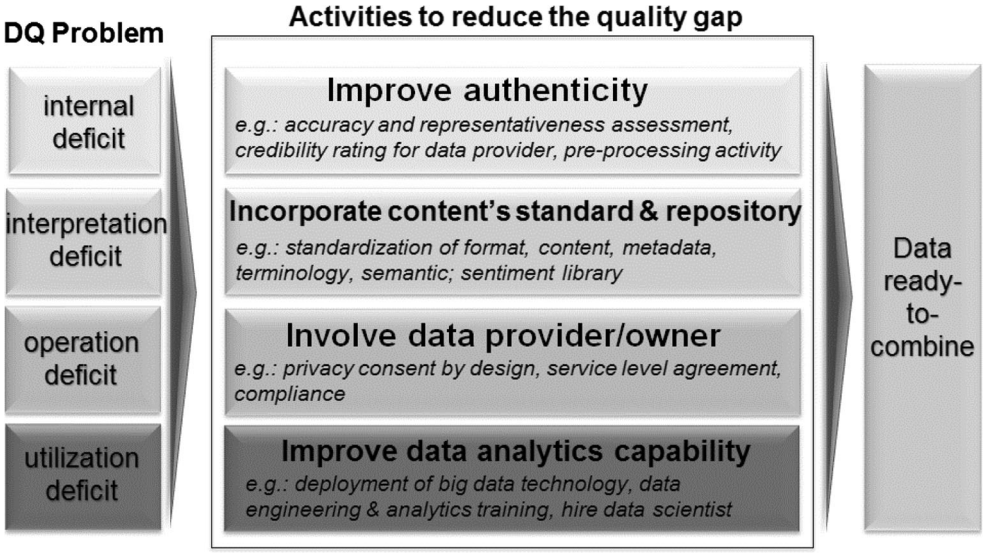

Fig. 6 Process pattern model

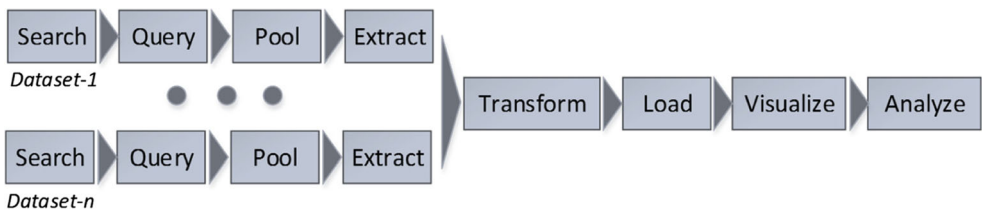

Fig. 7 Typical data process pattern when all datasets have high DQ

such as the data are not timely (or too late) for the task in hand (e.g. real-time fraud detection) or the data is too complex to process and find meaningful insight. The strategy to overcome this problem is improving data analytics capability in the organization, such as adoption of big data technology and enhancement of data engineering and analytics skills (Fig. 6).

Table 2 Process patterns of DQ challenge: the context, problem, and solution (the red words indicating different patterns from the typical process)

DQ dimension	Dataset example in the case	Problems in the case	Process Pattern
Intrinsic - Accuracy - Objectivity	Customer complaints from social media	Some data were from real customers; other data may be from black campaigners	1 Search ➔ *Assess accuracy* ➔ Query ➔ Pool ➔ Extract ➔ *Cleanse* ➔ Transform ➔ Load ➔ Analyze ➔ Visualize
Representational - Interpretability - Consistent representation	Network performance data	Varied terminologies and data representation across vendors' machines	2 Search ➔ *Make agreement* ➔ *Metadata* ➔ *Integration* ➔ Query ➔ Pool ➔ Extract ➔ Transform: *Manipulate* ➔ Load ➔ Analyze ➔ Visualize
Accessibility - Accessibility - Security	Transaction data	- Existing machines were not capable of handling much concurrent access (scalability) - Very restricted access - Privacy issue	3 Search ➔ *Access securely* ➔ Query ➔ Pool: *Data lake* ➔ Extract ➔ Transform: *Manipulate* ➔ Load ➔ Analyze ➔ Visualize
Contextual - Value-added	Many datasets	Lack of knowledge of how to derive value	4 Search ➔ *Metadata* ➔ Query ➔ Pool ➔ Extract ➔ Transform ➔ Load ➔ Analyze: *Model* ➔ Visualize

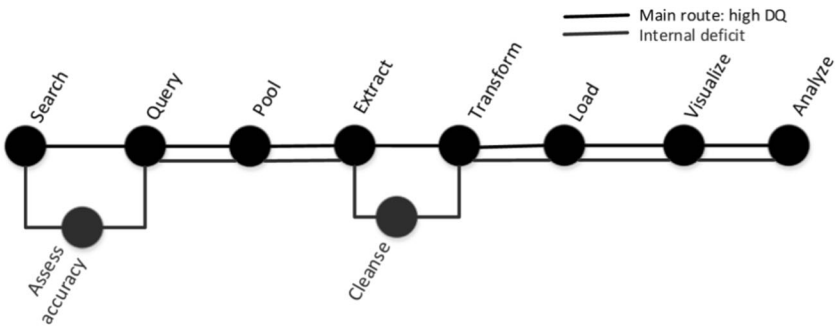

Fig. 8 Process pattern 1: Internal deficit solution

7 Process Pattern Example

The organization usually takes the typical process, comprised of a number of functions from the data processing lifecycle. Such typical process is arranged in a data pipeline, which includes searching the identified dataset, query the data using certain methods (e.g. API, flat file transfer), pooling the data to the organization's data lake, extracting the data for further processing, transforming the variables in the dataset based on the task at hand, loading the result to output container (e.g. memory, storage), visualizing the result to identify relationships among variables, and analyzing the results using certain analytic functions (Fig. 7).

However, if any dataset has a low DQ, the data processing takes different paths, different from the typical process. Combining the concept of DQ from Wang and Strong (1996) and the case study, we derive four process patterns as described in Table 2. The process patterns consist of DQ context, problem, and the solution that reflect the modification of the typical data process.

Process pattern 1 (Fig. 8) represents the change of typical data processing to take low intrinsic DQ into account. It is supposed to solve internal deficit issue. The example of the dataset from the case study is customer complaint from social media (e.g. Twitter). The data had low accuracy because some conversations were not generated by real customers but driven by fake accounts and black campaigners. Also, with the possibility that older generation was under-represented in the social media, it is important that prior to exploitation, the data need to be assessed for their accuracy. Because it is unlikely that we can improve their accuracy at the data sources, cleansing is the only way to exclude data with low accuracy.

Process pattern 2 (Fig 9) is aimed at solving interpretation deficit, i.e. low representational DQ. From the case study, the network performance data generated by machines from many vendors were hard to interpret because varied terminologies

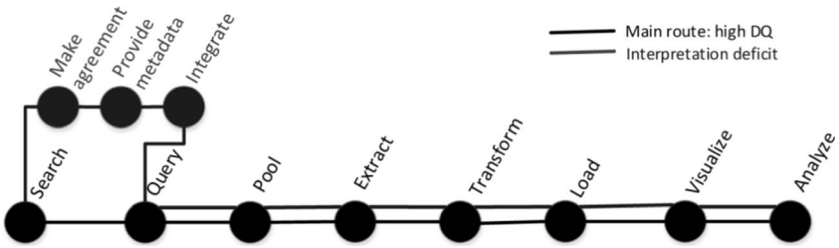

Fig. 9 Process pattern 2: Interpretation deficit solution

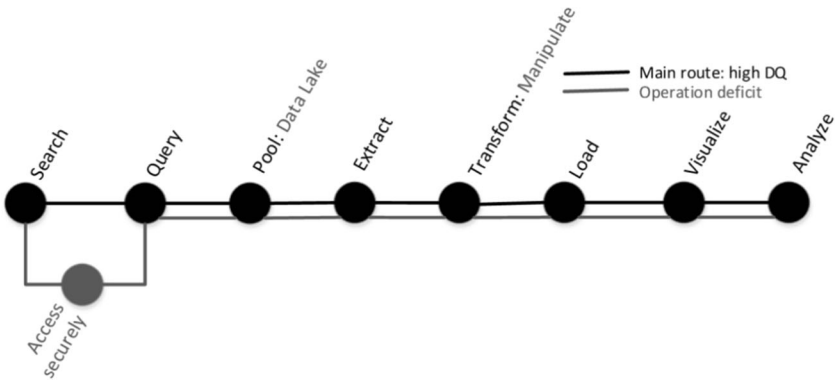

Fig. 10 Process pattern 3: Operation deficit solution

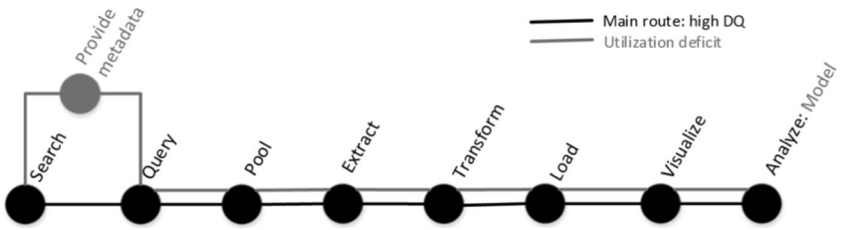

Fig. 11 Process pattern 4: Utilization deficit solution

were used across vendors' machines. Hence, metadata were important to facilitate understanding of the context and task conditions of data reuse. It was also found that the data were inconsistently represented. Each vendor used different formulations of performance indicators (e.g. drop call). The solution entails organizations agree on performance indicators (e.g. standardization) that can be applied across vendors' machines. In the exploitation step, the fields containing performance indicators need to be manipulated so that they represent consistently in the subsequent process. In a multi-software vendor environment, often different methods of access are used, e.g. direct query to databases, file transfer, query from the vendor's application, SNMP (Simple Network Management Protocol) logs, etc. Therefore, integration ensures that the information system can handle multiple ways of access.

The operation deficit caused by low accessibility DQ is solved by *process pattern 3* (Fig 10) . From our case study, the transaction data were generated by the machines that were not designed to process many concurrent connections. Hence, this led to the creation of a data lake to facilitate access to the source and also data reuse. In everyday practice, organizations may have strict regulations about access to data. Hence, accessing data from the data providers in a secure way is important. Privacy issue can concern organizations. Therefore, in the exploitation process, sensitive data including personally identifiable information need to be obfuscated prior to use.

Process pattern 4 (Fig. 11) addresses low contextual DQ. Most of the datasets have unknown value prior to the use. Therefore, the model to use the data in the analysis step is important for the organizations to create value from the data. Metadata are also important so that the data can be put into a contextual use.

The process patterns describe the recurring problems of big data quality together with the solutions that contain certain functions to solve the problem. This makes the process patterns reusable for any organization to reduce cost and create value from the data.

8 Conclusion

Organizations are struggling with using big data. Each time, they have to find ways of tackling data with variable quality. Instead of having to start from scratch, we develop a generalized framework of process pattern models to support creating value from data in data-driven organizations. Our model encompasses four variants of process patterns. Each pattern is related to the variable quality of the data.

We define a process pattern as "proven series of activities which are supposed to overcome a recurring problem in a certain context, particular objective, and specific initial condition". The patterns are comprised of data processing activities using data from multiple sources to attain certain goals and objectives from one context to another. These patterns are determined by the variable quality of the data being used. We proposed four process patterns that map big data quality problems in data processing, together with the following solutions. Process pattern 1 deals with low intrinsic data quality, e.g. inaccurate and biased. Functionalities such as accessing accuracy and cleansing are added to the typical data processing pattern. Low representational data quality is encountered by process pattern 2. Challenges like interpretability and consistent representation are solved by the functionalities such as metadata, making agreements, integration, and manipulation. Process pattern 3 considers low accessibility data quality. Secure access, building a data lake, and data manipulation are needed from dealing with restricted access, concurrency, and privacy. Process pattern 4 copes with low contextual data quality. To turn the data into value, models on data use and metadata are two important elements in this pattern.

Besides the patterns, a model to overcome the DQ problem was also proposed. Any organizations can benefit from the patterns and use the model to solve DQ issue. In this way, organizations can use the patterns as 'best practices', and save time and resources and avoid omitting steps by reusing the patterns.

The patterns and the strategy may enrich the repository of big data process patterns and big data strategy in the literature. Furthermore, they can be used as a starting point to further refine the patterns and generalize to other situations.

A limitation of this study is that the patterns are derived using a single case study in a particular field. Empirical research in other fields can be used to test and refine the proposed process patterns as well as to evaluate the process patterns and the significance of their elements.

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

References

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

Ambler, S. W. (1999). *More Process Patterns: Delivering Large-Scale Systems Using Object Technology*. Cambridge: Cambridge University Press.

Bauer, F., & Kaltenbock, M. (2011). Linked Open Data: The Essentials. Edition mono/monochrom:Vienna.

Becker, J., Delfmann, P., Dietrich, H.-A., Steinhorst, M., & Eggert, M. (2016). Business process compliance checking–applying and evaluating a generic pattern matching approach for conceptual models in the financial sector. *Information Systems Frontiers, 18*(2), 359–405.

Burton, A., & Treloar, A. (2009). Designing for discovery and re-use: the "ANDS data sharing verbs" approach to service decomposition. *International Journal of Digital Curation, 4*(3), 44–56.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., & Stal, M. (1996). A system of patterns: Pattern-oriented software architecture. Chichester: Wiley Publishing.

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

Davenport, T. H. (1993). Process innovation: reengineering work through information technology. Boston: Harvard Business Press.

Dwivedi, Y. K., Janssen, M., Slade, E. L., Rana, N. P., Weerakkody, V., Millard, J., et al. (2017). Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modelling. *Information Systems Frontiers, 19*(2), 197–212.

Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information Management, 39*(2), 109–116. https://doi.org/10.1016/S0378-7206(01)00083-0.

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

Green, A., & Kent, J. P. (2002). The metadata life cycle. In J. P. Kent (Ed.), MetaNet work package 1: Methodology and tools, Chap. 2.2 (pp. 29–34). http://www.epros.ed.ac.uk/metanet/deliverables/D4/IST_1999_29093_D4.pdf. Accessed 3 March 2016.

Hagen, M., & Gruhn, V. (2004). Towards flexible software processes by using process patterns. In *IASTED Conf. on Software Engineering and Applications* (pp. 436–441). Rome: IEEE.

Huang, S.-C., McIntosh, S., Sobolevsky, S., & Hung, P. C. K. (2017). Big Data Analytics and Business Intelligence in Industry. *Information Systems Frontiers, 19*(6), 1229–1232.

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.

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*, *52*(2), 21.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2009). *Journey to data quality*. Cambridge: The MIT Press.

Leek, J. (2015). The Elements of Data Analytics Style: A guide for people who want to analyze data. Retrieved from https://leanpub.com/datastyle.

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. http://abesit.in/wpcontent/uploads/2014/07/big-data-frontier.pdf. Accessed 8 May 2015.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution, 27*(2), 85–93.

Nag, R., Hambrick, D. C., & Chen, M. (2007). What is strategic management, really? Inductive derivation of a consensus definition of the field. *Strategic Management Journal, 28*(9), 935–955.

Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2016). The impact of big data analytics on firms' high value business performance. *Information Systems Frontiers*, 1–14. https://doi.org/10.1007/s10796-016-9720-4.

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: How organizations can unlock value and insight from the information they hold. PwC. https://www.pwc.es/es/publicaciones/tecnologia/assets/Seizing-The-Information-Advantage.pdf. Accessed 12 January 2016.

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

Umar, A., Karabatis, G., Ness, L., Horowitz, B., & Elmagardmid, A. (1999). Enterprise data quality: A pragmatic approach. *Information Systems Frontiers, 1*(3), 279–301.

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

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). Cham: Springer International Publishing.

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. Dordrecht, Kluwer Academic Publishers.

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

Yuan, S.-T. D., & Hsu, S.-T. (2017). Enhancing service system design: An entity interaction pattern approach. *Information Systems Frontiers*, 1–27. https://doi.org/10.1007/s10796-015-9604-z.

Zikopoulos, P. C., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding big data*. New York et Al: McGraw-Hill.

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

Agung Wahyudi is a PhD candidate in ICT section of Technology, Policy and Management Faculty of Delft University of Technology starting from April 2015. He is working on value creation from big data, open data, and linked data using a reference architecture. Before that, he has spent almost 10 years as a data scientist in the telecom industry.

George Kuk is professor of innovation and entrepreneurship at the Nottingham Business School. His research focuses on open innovation and strategy in software, data, design and platform within the creative industries. He examines how companies can attract creative resources for digital and service innovation.

Prof.dr. Marijn Janssen is a full Professor in ICT & Governance and chair of the Information and Communication Technology section of the Technology, Policy and Management Faculty of the Delft University of Technology. He is Co-Editor-in-Chief of Government Information Quarterly, conference chair of IFIP EGOV series and is chairing mini-tracks at the DG.o, ICEGOV, HICCS and AMCIS conferences. He was ranked as one of the leading e-government researchers in surveys in 2009, 2014 and 2016, and has published over 400 refereed publications. More information: www.tbm.tudelft.nl/marijnj.