Future Generation Computer Systems 63 (2016) 123–130

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A Data Quality in Use model for Big Data

Merino Jorge *, Caballero Ismael, Rivas Bibiano, Serrano Manuel, Piattini Mario

Alarcos Research Group, Escuela Superior de Informtica. Universidad de Castilla-La Mancha, Paseo de la Universidad 4, 13071, Ciudad Real, Spain

HIGHLIGHTS

- Data Quality is basic to decide about the suitability of data for intended uses.
- A Data Quality-in-Use Model based on ISO/IEC 25012, 25024 is proposed for Big Data.
- The main concern when assessing the Data Quality-in-Use in Big Data is Adequacy.
- The model accomplishes all the challenges of a Data Quality program for Big Data.
- The results obtained must be understood in the context of each Big Data project.

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form
18 November 2015
Accepted 25 November 2015
Available online 3 December 2015

Keywords:
Data Quality
Big Data
Measurement
Quality-in-Use
Model

ABSTRACT

Beyond the hype of Big Data, something within business intelligence projects is indeed changing. This is mainly because Big Data is not only about data, but also about a complete conceptual and technological stack including raw and processed data, storage, ways of managing data, processing and analytics. A challenge that becomes even trickier is the management of the quality of the data in Big Data environments. More than ever before the need for assessing the Quality-in-Use gains importance since the real contribution – business value – of data can be only estimated in its context of use. Although there exists different Data Quality models for assessing the quality of regular data, none of them has been adapted to Big Data. To fill this gap, we propose the "3As Data Quality-in-Use model", which is composed of three Data Quality characteristics for assessing the levels of Data Quality-in-Use in Big Data projects: Contextual Adequacy, Operational Adequacy and Temporal Adequacy. The model can be integrated into any sort of Big Data project, as it is independent of any pre-conditions or technologies. The paper shows the way to use the model with a working example. The model accomplishes every challenge related to Data Quality program aimed for Big Data. The main conclusion is that the model can be used as an appropriate way to obtain the Quality-in-Use levels of the input data of the Big Data analysis, and those levels can be understood as indicators of trustworthiness and soundness of the results of the Big Data analysis.

1. Introduction

Traditionally, organizations realized that the insights of owned data could largely benefit their business performance by means of Business Intelligence techniques [1–3]. These insights are new ways to make business by leveraging new types of analytics over new types of data. Organizations are now being challenged to create new business actions based on the benefits brought by these types of analysis [4]. The ability of classics (e.g., those based on relational databases) to process structured data is not sufficient (in terms of performance and latency) when data comes at certain volumes, in different formats and/or at different rates of speed [5]. Furthermore, the rise of unstructured data, in particular, means that captured data has to move beyond merely rows and tables [6,7].

Big Data is rising as a new solution to the common problems found when processing large amounts of data, that might be also diverse, and likely to be processed with massive parallelism as well. Depending on the type of analysis to be performed, some specific data must be gathered and arranged in a particular way, to tackle the new challenges from various natures (technological, conceptual and methodological) [8]. The gathered data must be related to the domain of interest or the context of the analysis, in other words, data must be valuable for the analysis.

* Corresponding author.
E-mail addresses: jorge.merino@uclm.es (J. Merino), ismael.caballero@uclm.es (I. Caballero), bibiano.rivas@uclm.es (B. Rivas), manuel.serrano@uclm.es (M. Serrano), mario.piattini@uclm.es (M. Piattini).
URL: http://alarcos.esi.uclm.es/ (J. Merino,I. Caballero,B. Rivas,M. Serrano,M. Piattini).

http://dx.doi.org/10.1016/j.future.2015.11.024
0167-739X/

Taking into account that both, raw data and the results of data analytics, are worthy for organizations and bearing in mind that their organizational value is so high, some authors and practitioners consider data as a business asset [9,10]. This fact highlights the concern and the need for a special focus on the quality of the data [11,12].

The author of [13] declares that whilst classic Data Quality foundations works fine in old challenges, – commonly based on relational models – they are not meant to yield properly in Big Data environments. Loshin in [14] states that it is naive to assert that is possible to adopt the traditional approaches to Data Quality on Big Data projects. On top of that, the regular ways to oversee Data Quality with classic models are generally intended to detect and fix defects in data from known sources based on a limited set of rules. Instead, in Big Data surroundings, the number of rules might be huge, and fixing found defects might be neither feasible nor appropriate (e.g., the huge volume of data, or volatility of streaming data). In these circumstances, it is necessary to redefine the ways of supervising Data Quality and put them within the context of Big Data. Unfortunately, to the best of our knowledge, not much research has still been conducted related to Data Quality Management in Big Data, beyond cleansing incoming data indiscriminately. Thus, we pose that there is a lack of a Data Quality model which can be used as a reference to manage Data Quality in Big Data.

Our proposal is a model that can be used to assess the level of Quality-in-Use of the data in Big Data. We consider that it is paramount to align the investigation with the best practices in the industry in order to produce repeatable and usable research results. Taking advantage of the benefits of using international standards is one of those best practices. In this sense, – among the different Data Quality models for regular data that might influence our solution – bringing to the arena standards like ISO/IEC 25012 and ISO/IEC 25024 may be very convenient. According to ISO/IEC 25010 [15], the Quality-in-Use depends on the external quality, and the external quality depends on the internal quality. ISO/IEC 25012 [16] contains a Data Quality model with a set of characteristics that data from any information system must fulfill to attain adequate levels of external Data Quality. ISO/IEC 25024 [17] provides general measures to quantify the external and internal quality of the data with compliance to the characteristics from ISO/IEC 25012. Albeit, these standards cannot be applied straight into Big Data projects, because they are devised for classic environments. Rather, they must be understood as "reference guides" that must be tailored and implemented accordingly to the particular technological environment to analyze Data Quality.

The structure of the rest of the paper is depicted below. Section 2 describe the foundations and the most significant aspects of Data Quality that can be used in a Big Data scenario. Section 3 shows our proposal. Section 4 presents a working example of the application of the Data Quality in Use model. Finally, in Section 5 some conclusions are reached.

2. Foundations

2.1. Data Quality in Big Data

If defining Data Quality was difficult, finding a sound definition of Data Quality for Big Data is even more challenging as there is not a regulated definition for Big Data yet. Gartner's definition is the most widely used: "*Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making*" [18]. Loshin states in [14] that "*Big Data is fundamentally about applying innovative and cost-effective techniques for solving existing and future business problems whose resources requirements exceed the capabilities of traditional computing environments*".

Then, Big Data is an "umbrella term" that covers not only datasets themselves, but also space problems, technologies and opportunities for enhancing business value [19]. Precisely, achieving a large business value from data is the main reason for what Big Data can be used for. Regrettably, High Management tends to think that the larger the Big Data project is (e.g., the largest amount of data involved in the project), the larger benefits (e.g., the soundest knowledge) can be obtained; unfortunately this happens even when they do not know exactly how to address Big Data concerns nor how to get the maximum benefits from the projects [1]. Hence, the very first step in any Big Data project is to encourage High Management to lead the project over acquiring and deploying sophisticated technology that will not produce any suitable results for the business case at hand [20,2].

Once High Management is convinced about the real need of undertaking Big Data projects, they have to be willing to deal with the challenges that Big Data brings in order to achieve an alignment to the reality of the organizations [14]. The challenges have been identified in [21]: Data Quality, adequate characterization of data, right interpretation of results, data visualization, real-time view of data vs. retrospective view and determining the relevance of results of projects. Among these hurdles, Data Quality takes a decisive part in the sense of addressing the trustworthiness of input data. Considering if data, – which is to be processed by the Big Data solution – has inadequate levels of quality, errors are likely to appear and they can accidentally and unknowingly be spread throughout the Big Data becoming even harmful for the organization [14].

Generally speaking, Data Quality Management is focused on the assessment of datasets and the application of corrective actions to data to ensure that the datasets fit for the purposes for which they were originally intended [14]. In other words, the input data is useful and appropriate for the Big Data analysis. Big Data introduces new technological and managerial challenges that makes the application of Data Quality Management principles a bit different than in regular data [21]. Table 1 gathers some of these facts.

2.2. International standards addressing Data Quality concerns

ISO/IEC 25000 is the family of standards addressing Systems and Software Quality Requirements and Evaluation (SQuaRE). It provides several divisions: ISO/IEC 2500n—Quality Management, ISO/IEC 2501n—Quality Model, ISO/IEC 2502n—Quality Measurement, ISO/IEC 2503n—Quality Requirements, and ISO/IEC 2504n—Quality Evaluation.

An interpretation of **Quality** provided by ISO/IEC 25010 [15] allows the classification of the concept in three categories: **Internal** quality, **External** quality and **Quality-in-Use**. The manufacturing process generates a specific configuration for the internal and static properties of a product, which are assessed by means of **internal quality**. This internal quality influences the dynamic properties of the product, which represent the **external quality**. This latter influences the **Quality-in-Use**, that is the sort of quality perceived by the final user.

2.2.1. ISO/IEC 25012

ISO/IEC 25012 [16] gathers the main desirable Data Quality characteristics for any dataset. In [16], Data Quality is described using a defined external Data Quality model. The Data Quality model defined in [16] categorizes quality attributes into fifteen characteristics considered by two points of view: