ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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:

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024

Table 1
Traditional Data Quality programs vs. Data Quality programs for Big Data [8, Ch 9, pp. 102].

Technological and managerial challenge	Data Quality on Regular (Relational) Data	Data Quality on Big Data
Frequency of processing	Processing is batch-oriented	Processing is both real-time and batch-oriented
Variety of data	Data format is largely structured	Data format may be structured, semi-structured, or unstructured
Confidence levels	Data needs to be in pristine conditions for analytics in the data warehouse	Noise needs to be filtered out, but data needs to be good enough. Poor data quality might or might not impede analytics to glean business insights
Timing of data cleasing	Data is cleansed prior to loading into the data warehouse	Data may be loaded as-is because the critical data elements and relationships might not be fully understood. The volume and velocity of data might require streaming, in-memory analytics to clease data, thus reducing storage requirements.
Critical data elements	Data quality is assessed for critical data elements such as customer address	Data may be quasi- or ill-defined and subject to further exploration, hence critical data elements may change iteratively
Location of analysis	Data moves to the data quality and analytics engines	Data Quality and analytics engines may move to the data, to ensure an acceptable processing speed
Stewardship	Stewards can manage a high percentage of the data	Stewards can manage a smaller percentage of data, due to high volumes and/or velocity.

Table 2
Data Quality characteristics categorization in accordance with the inherent and system dependent points of view [16].

Characteristic	Inherent	System dependent
Accuracy	×	
Completeness	×	
Consistency	×	
Credibility	×	
Currentness	×	
Accessibility	×	×
Compliance	×	×
Confidentiality	×	×
Efficiency	×	×
Precision	×	×
Traceability	×	×
Understandability	×	×
Availability		×
Portability		×
Recoverability		×

- **Inherent** Data Quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.
- **System dependent** Data Quality refers to the degree to which Data Quality is reached and preserved within a computer system when data is used under specified conditions.

The categorization of the Data Quality characteristics according to these points of view is depicted in Table 2.

2.2.2. ISO/IEC 25024

ISO/IEC 25024 [17] defines some basics and concepts that allow to perform objective and unbiased Data Quality measurements. Fig. 1 shows the main concepts of ISO/IEC 25024 and their relationships. For a deeper knowledge about those concepts see [17].

In this way, one or more *Quality Measures* can be used for the measurement of a Data Quality characteristic. [17] establishes *Quality Measures* for the characteristics included in the Data Quality model defined in [16].

The *Quality Measures* described in [17] are quantifications of the Data Quality characteristics, and those concerning data can be used over all the stages of the Data-Life-Cycle and for other processes, for example:

- To establish Data Quality requirements.
- To evaluate Data Quality.
- To support and implement data governance, data management, data documentation process.
- To support and implement IT services management processes.
- To support improvement of Data Quality and effectiveness of business decisions process.
- To benchmark Data Quality of different data management solutions during investigation process.
- To evaluate the quality of system and/or software components that produce data as an outcome.

3. Data Quality-in-Use in Big Data

As aforementioned, the viewpoint of traditional Data Quality must be stretched to meet the new technological and managerial challenges that Big Data introduces.

Based on Loshin's claim [14] of the need for "*identifying the critical Data Quality dimensions that are important for the data processed by the Big Data project*", our proposal is focused on depicting a Data Quality-in-use model for Big Data that allows companies to assess the extent to which their data is good enough for the Big Data purposes and goals. To address the best practices of the industry, the presented Data Quality-in-Use model for Big Data is largely grounded on the standards ISO/IEC 25012 [16] and ISO/IEC 25024 [17].

3.1. The 3As Data Quality-in-Use model

Big Data solutions can be comprehended as full information systems. We are interested in transactional and analytic data that will play the role of the input of the Big Data. Under no circumstances are we aiming for the results of the Big Data analysis—whose quality might be assessed through other models. The measurement of the Data Quality levels of this input data is within the scope of the standard ISO/IEC 25012 [16].

This proposal relates the dependencies between all types of the qualities depicted in ISO/IEC 25010 (see Section 2.2). The interpretation of **Quality** provided by ISO/IEC 25010 [15], can be applied to data – understanding data as a product –: the extent to which data meets the defined requirements is the internal quality of data; the relationships and the appropriateness of representation of data is the external quality of data; the extent of fulfillment of the goals set for data is the Quality-in-Use.

The Data Quality model from ISO/IEC 25012 [16] would help as a reference for the study of the internal and external quality of the input data of Big Data solutions, but not for the study of the Quality-in-Use. The *3As Data Quality-in-Use model* introduced in this work is defined to fill the gap for a Data Quality-in-Use, enabling the assessment of the Quality-in-Use of the data in Big Data projects. That is, this new model is designed to provide a way to obtain the extent to which the data is sound and appropriate from the quality point of view for the intended use (i.e., to produce trustworthy results through the Big Data analysis).

According to this perspective based on the idea of quality, we pose that the main Data Quality concern when assessing

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024

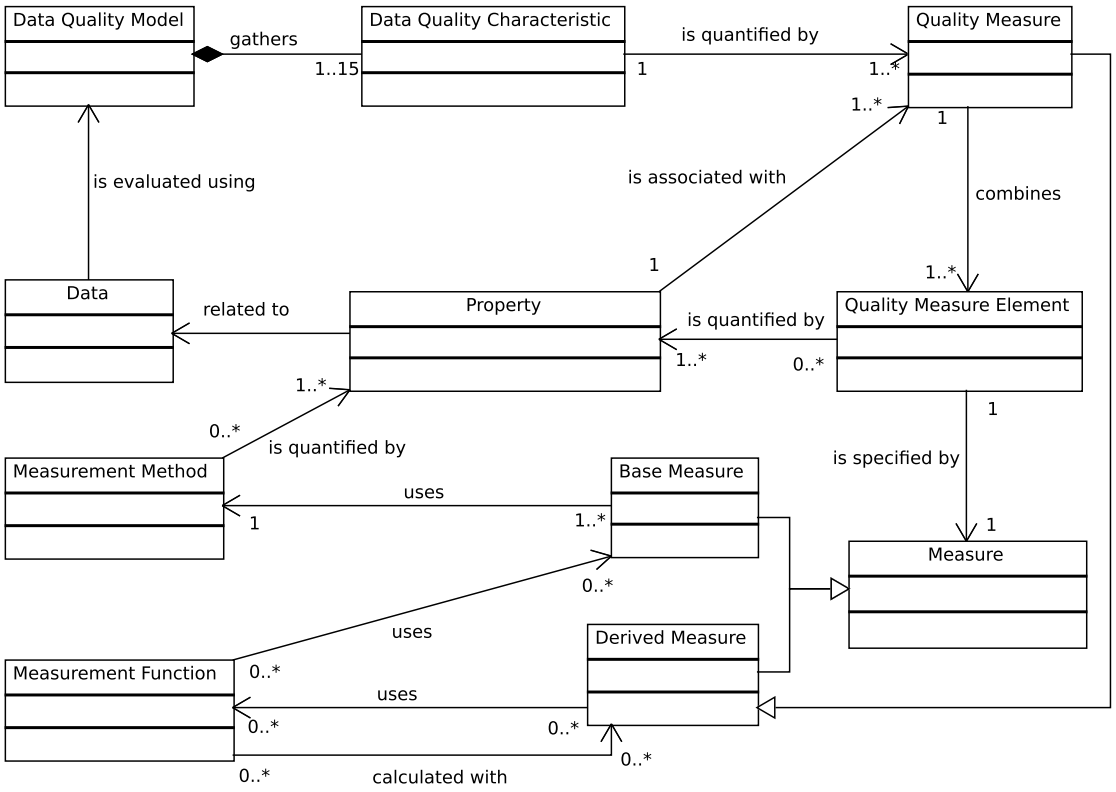

Fig. 1. ISO/IEC 25024 main concepts.

the level of Quality-in-Use in Big Data projects is the **Adequacy** of data to the purposes of the analysis. According to Merriam-Webster dictionary, Adequacy can be defined as "*the state or ability of being good enough or satisfactory for some need, purpose or requirement*". Adapting this definition, Adequacy of data is "*the state or ability of data of being good enough to fulfill the goals and purposes of the analysis*". In other terms, to be compliant to the specific characteristics of the analysis, which is conducted through a particular Big Data solution.

When we began to identify the main characteristics of the Data Quality-in-Use for our model, the classification of [22] was taken as basis. This work groups the characteristics in four categories: Accessibility, Contextual, Representational and Intrinsic. As part of the research process, these four categories were re-grouped in two characteristics for the context of Big Data: *Contextual Adequacy* and *Operational Adequacy*—addressing the characteristics: Representational, Accessibility and Intrinsic. The main reason of this simplification is based on the fact that data has to be processable with the resources and technologies available for the Big Data analysis, and these three categories fit into the definition of a single characteristic, that we called *Operational Adequacy*. With respect to *Contextual Adequacy*, we acknowledge the temporal aspects as part of the context. Notwithstanding, due to the growing significance of the real-time analysis, an isolated assessment of the temporal aspects was considered as required. Whence, we decided to split the Contextual category up into *Contextual Adequacy* and *Temporal Adequacy*. Consequently, we identify three critical Data Quality characteristics that are important for the data within the context of the Big Data analysis: *Contextual Adequacy*, *Temporal Adequacy* and *Operational Adequacy*. Subsequently, the definition of each characteristic from the 3As Data Quality-Use model is provided:

Contextual Adequacy refers to "the capability of datasets to be used within the same domain of interest of the analysis independently of any format (e.g., structured vs. unstructured), any size or the velocity of inflow". In this sense, it is important that data is:

1. *relevant and complete*: the amount of used data is appropriate and it is within the context of the task at hand (e.g., the Big Data analysis);
2. *unique and semantically interoperable*: so the data must be understandable taking into account the given context and free of inconsistencies due to duplicates;
3. *semantically accurate*: data must represent real entities in the context of the Big Data analysis;
4. *credible*: analysts should find the levels of credibility in data good enough for the context (e.g., all the data sources must be credible);
5. *confidential*: data must be accessed by the same group of people allowed to develop the analysis.
6. *compliant* to the stated regulations and requirements.

Temporal Adequacy refers to the fact that "data is within an appropriate time slot for the analysis (e.g., similar age, or throughout a specific duration for historical data, or coetaneous data, meaning that data refers to a similar period of time,…)". It is important to notice that the temporal aspects of the operation of data through the analysis are not included in this definition, instead, just the temporal aspects of data itself. This has several perceptions, so that, it is important that data being processed should be:

1. *time-concurrent*: referring to facts happened in similar or appropriate time slots (e.g., if an analysis is focused on a past event, then data must correspond to related and coetaneous things);
2. *current*: data must be similar in age. In some cases, merging data having different levels of *currentness* may not lead to sound analysis;
3. *timely* updated: data must be properly updated for the task at hand, so it has a convenient age for the analysis;
4. *frequent*: used data for producing results related to required future time slots (required *frequencies*), when performing some sort of tends analysis.

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024

Table 3
Data Quality-in-Use model for Big Data based on ISO/IEC 25012.

Data Quality characteristic	Contextual Adequacy	Temporal Adequacy	Operational Adequacy
Accuracy	×	×	
Completeness	×		
Consistency	×	×	
Credibility	×		
Currentness		×	
Accessibility			×
Compliance	×		
Confidentiality	×		×
Efficiency			×
Precision			×
Traceability			×
Understandability	×		
Availability			×
Portability			×
Recoverability			×

Table 4
3Vs affecting the measurement of the 3As based on ISO/IEC 25012 characteristics.

	Velocity	Volume	Variety
Contextual Adequacy	Completeness	Completeness Consistency Confidenciality	Accuracy Consistency Credibility Compliance Confidenciality Understandability
Temporal Adequacy	Accuracy Currentness	Currentness	Consistency Currentness
Operational Adequacy	Confidenciality Efficiency	Efficiency	Accessibility Confidenciality Efficiency

5. *time-consistent*: data must not include any incoherence related to the represented time (e.g., disordered events, impossible dates,…).

Operational Adequacy refers to "the extent to which data can be processed in the intended analysis by an adequate set of technologies without leaving any piece of data outside the analysis". This means, that there are sufficient and appropriate resources available to perform the analysis (e.g., similar data types, equivalently expressed data attributes,…). The cost-effectiveness and the performance related to the 3Vs must be taken into account as well. Therefore, data in the various datasets should:

1. be *available*, easily *recoverable* and *accessible* for the analysis;
2. be *authorized* for the intended purposes.
3. be expressed by using *similar data types* and with the same amount of *precision* and it can be also *portable*;
4. have an *efficient* representation in order to avoid wasting resources;
5. provide an audit trail that allows to *trace* the accesses and changes;

Table 3 shows the Data Quality characteristics from ISO/IEC 25012 [16] that may have influence over the 3As.

Table 4 shows the way the external Data Quality characteristics from ISO/IEC 25012 [16] can be included when assessing the Data Quality-in-Use levels through the 3As, taking into account the prevailing Vs of the particular Big Data project at hand.

3.2. Measuring the Data Quality-in-Use in Big Data environments

To reckon the extent to which the characteristics are fulfilled, some measures must be arranged. To make operative the 3As Data Quality-in-Use model, we have decided to ground the measurement on the **Quality Measures** from ISO/IEC 25024 [17]. These Quality Measures allow to calculate the level of fulfillment of the external Data Quality characteristics from ISO/IEC 25012 [16]. Those levels of external Data Quality are combined in a meaningful and representative way to obtain the levels of Data Quality-in-Use. Using the words "meaningful" and "representative" we mean that the combination of the characteristics is specific on each case, according to the nature of the Big Data project (e.g., Web and Social Media, Machine-to-Machine, Big Transaction data, Bio-Metrics, Human Generated,…). For instance, a Healthcare project would concern about the precision and the completeness of data to perform trustworthy diagnostics; while a Big Data solution used for Marketing purposes would care more about the currentness and the availability in order to predict customers needs.

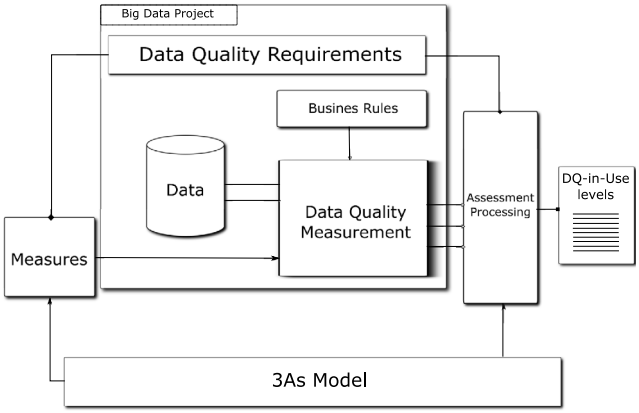

Fig. 2. Analyzing Data Quality-in-Use levels of the data from a Big Data environment with the Data Quality measures.

The Quality Measures from ISO/IEC 25024 [17] may be used as "reference guides" to define measures that quantifies the degree of compliance of the data with the characteristics from the external Data Quality model from ISO/IEC 25012 [16], in the scope of Big Data projects. Those Quality Measures must be implemented within the specific features of the particular Big Data project at hand.

The following steps shall be followed to measure the levels of Data Quality-in-Use of the data in a Big Data project (see Fig. 2):

1. Establish the Data Quality Requirements delimited by the scope of the Big Data environment. A Data Quality Requirement represents the concerns about Data Quality of the specific analysis of a particular Big Data project.
2. Select the types of Adequacy that better describe the Data Quality Requirements identified in step 1. These types can be deduced from the Data Quality Requirements using the definition of the three types of Adequacy provided in Section 3.1.
3. Identify those Data Quality characteristics that are critical to assess the level of the chosen Adequacy type(s) using Table 3.
4. Gather the business rules that data must meet. The business rules are the constraints or requirements defined over data. These business rules may come from different sources (e.g., the data itself, from the organization that owns the data, stakeholders associated with the analytics purposes,…). The external Data Quality characteristics can be used as a categorization of those business rules. The extent to which these business rules are met can be understood as the external Data Quality levels.
5. To reckon the external Data Quality levels some measures must be defined within the context of the specific Big Data project at hand. Each business rule will be used as the input to these measures. The defined measures shall be grounded

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024

on the *Quality Measures* from ISO/IEC 25024 [17] and must be developed within the specific Big Data solution scope.
6. Calculate the values of the Data Quality-in-use Measures, using the 3As Data Quality-in-Use model. For this purpose, the external Data Quality levels must be combined in a meaningful and representative way—see above. To do this, combination must take into account some decision criteria based on the scope of the particular Big Data project at hand. These decision criteria is the basis of the assessment of the Data Quality-in-Use.
7. Either generate a report with the results of the measurement process or attach the values of the Quality-in-Use to data. These levels represent the degree of trustworthiness, reliability and even the validity of data to be used for the specific analysis of the particular Big Data project. The data stewards shall use the information about the Quality-in-Use to decide whether or not to warn about the soundness of the analysis performed with the assessed data.

4. Working example

A working example is introduced below to show the way to use the *3As Data Quality-in-Use model*. The working example was performed within the financial domain. The dataset contains information about the clients of different banks, their financial movements and also some public information about demographic data. This dataset is from 1999, but it was selected due to its simplicity and its understandability. We are selecting a specific file with 1 056 320 records describing the *transactions* on accounts, that can be understood as high volume data.

The steps that must be followed to use the *3As Data Quality-in-Use model* are defined in 3.2:

1. We are setting up just one Data Quality Requirement for this working example, that is: "*The soundness and the validity of data for the analysis must be the highest*". This Data Quality Requirement represents a deep concern about the extent to which data is good enough to perform reliable analysis of the *transactions*.
2. According to that Data Quality Requirement, the most important characteristic from the *3As Data Quality-in-Use model* is *Contextual Adequacy*, because of the deep concern about the validity of the data for the domain of interest.
3. The following Data Quality characteristics from ISO/IEC 25012 [16] are used to get the external quality related to *Contextual Adequacy* according to Table 3: *Accuracy*, *Completeness*, *Adequacy*, *Credibility*, *Currentness*, *Confidentiality* and *Understandability*.
4. The business rules (i.e., constraints) about the data were extracted from the data itself (e.g., relationships, data types,…) and documentation (e.g., syntactic constraints, semantic values,…) [23].
5. The measures to obtain the external Data Quality levels were gathered from ISO/IEC 25024 [17]. The selected measures – those associated with the chosen characteristics – were implemented using the Map-Reduce paradigm. An example of the definition of the measures is given in Appendix.
6. The assessment was performed over the *transactions* file. To combine the measurement results in a "meaningful" and "representative" way, we used some combination functions called "*functions by profiles*". Those functions are defined in [24]. The "*functions by profiles*" are devised to perform the role of the evaluators and thence the decision criteria to set the levels of Quality-in-Use from the measurement results is defined through them.
   (a) The measures implemented in the step 5 were executed over the *transactions* file.
   (b) All the results of those measures were gathered and combined using the "*functions by profiles*" as decision criteria. The results of applying the "*functions by profiles*" are indicators of the levels of external data quality.
   (c) In a final step, the levels of external Data Quality were gathered and combined again using again the "*functions by profiles*". This time these functions are used as decision criteria to obtain indicators about the levels of Quality-in-Use of the *transactions*, according to the 3As Data Quality-in-Use model. The results of the assessment – see Table 5 – are provided for the *Contextual Adequacy*, as this characteristic is the main concern obtained from the step 2. The scale of the results is from 1 to 5 due to the definition of the "*functions by profiles*".
7. These Data Quality-in-Use levels, compared to the Data Quality Requirement, were reported to warn about the risks – from the Data Quality perspective – of using the *transactions* file to make any decision.

Table 5
Quality-in-Use levels of the *transactions* file.

3As model	DQiU level (over 5)
Contextual Adequacy	3
Temporal Adequacy	N/A
Operational Adequacy	N/A

According to these results the main conclusions can be drawn are:

1. From the point of view of an acceptable result – good enough – some would assert that a level 3 in the Contextual Adequacy means that data has appropriate levels of quality for the intended uses. In other terms, the analysis performed on this data would be sound enough and reliable.
2. Nevertheless, the Data Quality Requirement demands "*the* **highest** *soundness and validity*" of data for the analysis of the data about the *transactions*. This can be translated into a request for at least level 5 of Contextual Adequacy—the maximum level of the Contextual Adequacy for this specific working example. As a result of this specific case, the Data Quality Requirement is telling us that data is not good enough for the purpose of the analysis.

5. Conclusions

Currently, we are immersed in the Big Data era. Beyond the hype of Big Data, organizations need to keep on taking care about the data they use in their business processes. Nonetheless, the way to manage Data Quality is influenced not only by the nature of the data itself, but also by the analytical processes, and especially by the new technologies that allow new ways to support those analytical processes.

In this paper, a Data Quality-in-Use model has been presented for the assessment of the Quality-in-Use of the data from Big Data solutions. This model is composed of three characteristics: *Contextual Adequacy*, *Temporal Adequacy* and *Operational Adequacy*.

- The Contextual Adequacy refers to the closeness of the data to the domain of the analysis.
- The Temporal Adequacy refers to the coherence of data with the analyzed period of time, as well as the timeliness with which data takes part in the analysis.
- The Operational Adequacy refers to the extent to which it is possible to perform different operations into the processed data, with efficiency and effectiveness using the available resources and definitions of the Big Data solution.

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024

The model appropriateness for Big Data solutions can be decomposed using the technological and managerial challenges of Big Data presented in Table 1:

- The 3As Data Quality-in-Use model can be applied into any Big Data specific implementation, as its measures are independent of any situation, requirement or technology. An example of an implementation of a measure is given in Appendix, but this implementation can be re-programmed for any other Big Data specific solution. This allows to assert that the challenges "Frequency of processing" and "Variety of Data", relative to the different implementations of the Big Data solutions, are accomplished by the model. The challenge "Timing of data cleansing" is partially accomplished since it also concerns about performance.
- The specific analysis of the Big Data solution at hand can be performed independently of the assessment of the Data Quality-in-Use levels. Those levels provided through the 3As Quality-in-Use model are indicators that must be used to raise consciousness about the soundness of the analysis result. This allows to assert that the challenge "Confidence levels" is accomplished because it is just a complementary process, owing to the assessment does not impede to perform the analysis. In this sense, the performance concerns of the challenge "Timing of data cleansing" are accomplished as well.
- As far as we are concern, the dynamics of data will always be part of its external quality. The critical elements will be identified through the business rules that set the constraints over data. This is taken into account using the characteristics from ISO/IEC 25012 [16] when measuring the external Data Quality. This allows to assert that the challenge "Critical elements" is accomplished.
- The assessment is performed in a way that it is not necessary to move the data. Instead, the 3As Data Quality-in-Use model is implemented within the Big Data solution. Then, the data stewards are able to decide whether to assess a subset of data or full datasets. This has been shown in the Working Example (see Section 4) and allows to assert that the challenges "Location of analysis" and "Stewardship" are accomplished by the model.

As all the common challenges of a Data Quality program for Big Data are accomplished, the 3As Data Quality-in-Use model can be trusted as an appropriate solution to assess the Data Quality in Big Data projects.

In this paper, a working example to show the way to use the 3As Data Quality-in-Use model has been conducted. The results obtained through this model must be seen in the context of each Big Data project. In this particular working example, even though the results can be seen as acceptable (level 3 out of 5 of Contextual Adequacy), the Data Quality Requirement demands the highest level (level 5 out of 5 of Contextual Adequacy), so data is not good enough for its use for analyzing the transactions.

From our perspective, it is imperative to highlight that is essential to be aware of the quality of data in order to decide whether data is sound for the intended uses.

Acknowledgments

This work has been funded by the GEODAS-BC project (Ministerio de Economa y Competitividad and Fondo Europeo de Desarrollo Regional FEDER, TIN2012-37493-C03-01) and is also part of the SERENIDAD project (Consejera de Educacin, Ciencia y Cultura de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII11-0327-7035).

Appendix. Data Quality measures implementation details

As aforementioned, the *3As Data Quality-in-Use model* must be used alongside the international standard ISO/IEC 25024 in order to provide measures to obtain the Data Quality-in-Use levels. An example of the implementation of one of the measures will be shown. The selected measure is *Record Completeness* that takes part when measuring the Data Quality characteristic called *Completeness*. In listing 1, the pseudo-code of the measure *Record Completeness* is shown, and in listings 2 and 3 the final code (in Python) used to develop the measure using the Map-Reduce paradigm from Big Data solutions.

```
1  mapper(FILE file, int nAttributes, int []
       indexesNecessaryAttributes){
2     for (record in file){
3        reset(result);
4        attributes <- getAttributes(record);
5        if (length(attributes)==nAttributes){
6           result <- hasLostValues(attributes,
                 indexesNecessaryAttributes);
7        }
8        print(result);
9     }
10 }
11
12 hasLostValues(attributes, indexes){
13    res <- false;
14    for(i in indexes){ // Note that indexes does not have to
          be the typical sequence (0..n)
15    if(attributes[i] and length(attributes[i])>0){
16       res <- true;
17       break; // leave for loop
18       }
19    }
20    return res;
21 }
22
23 reducer(FILE mapperResults){
24    n <- 0;
25    total <- 0;
26    for (line in mapperResults){
27       result <- getResultFrom(line);
28       if (result){
29          total <- total + 1;
30       }
31       n <- n + 1;
32    }
33    recordCompleteness <- total/n;
34    print(recordCompleteness);
35 }
```

Listing 1: Pseudo-code of the measure Record Completeness

```
1  def mapper(args):
2     nline = 0
3     nAttributes = args[1]
4          indexes=args[2:len(sys.argv)]
5          for line in sys.stdin:
6             nline+=1
7        result = False
8        data = line.strip().split(";")
9        nrecords=len(data)
10
11       if nrecords == int(nAttributes):
12                         result = hasLostValues(data,indexes)
13       print (result)
14
15 def hasLostValues(data, indexes):
16    res =  True
17    i = 0
18    for i in indexes:
19       actual = int(i)
20       if data[actual] and len(data[actual]) == 0 or data[
             actual]=='None' or data[actual]=='' or data[
             actual]=='null' or data[actual]=='\'\'':
21          res= False
22          break
23    return res
```

Listing 2: Mapper of the measure Record Completeness

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024

```
def reducer(args):
    total = 0.0
    n = 0
    for line in sys.stdin:
        data = line.strip()
        if data=="True":
            total += 1.0
        n +=1

    print ('total records: {0}'.format(float(n)))
    print ('valid records: {0}'.format(total))
    print ('RecordCompleteness results: {0}%'.format(total/
        float(n)*100))
```

Listing 3: Reducer of the measure Record Completeness

The measures accept different values into the input fields in order to be applicable to different data repositories, using the business rules associated with them. Each business rule is aligned to a specific Data Quality measure. Thereby, it is possible to measure the extent of compliance from the point of view of Data Quality.

References

[1] B. Mantha, Five guiding principles for realizing the promise of big data, Bus. Intell. J. 19 (1) (2014) 8–11. URL: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=95066192&lang=es&site=ehost-live.
[2] A. McAfee, E. Brynjolfsson, Big data: the management revolution, Harv. Bus. Rev. 90 (10) (2012) 60–68.
[3] V. Chang, The business intelligence as a service in the cloud, Future Gener. Comput. Syst. 37 (2014) 512–534. special Section: Innovative Methods and Algorithms for Advanced Data-Intensive ComputingSpecial Section: Semantics, Intelligent processing and services for big dataSpecial Section: Advances in Data-Intensive Modelling and SimulationSpecial Section: Hybrid Intelligence for Growing Internet and its Applications. http://dx.doi.org/10.1016/j.future.2013.12.028. URL: http://www.sciencedirect.com/science/article/pii/S0167739X13002926.
[4] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, Int. J. Inf. Manage. 35 (2015) 137–144.
[5] N. I. of Standards, Draft NIST big data interoperability framework: Volume 4, security and privacy, Report, US. Department of Commerce, 2015.
[6] N. I. of Standards, Draft NIST big data interoperability framework: Volume 6, reference architecture, Report, US. Department of Commerce, 2015.
[7] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of things (IoT): A vision, architectural elements, and future directions, Future Gener. Comput. Syst. 29 (7) (2013) 1645–1660. including Special sections: Cyber-enabled Distributed Computing for Ubiquitous Cloud and Network Services and Cloud Computing and Scientific Applications Big Data, Scalable Analytics, and Beyond. http://dx.doi.org/10.1016/j.future.2013.01.010. URL: http://www.sciencedirect.com/science/article/pii/S0167739X13000241.
[8] S. Soares, Big Data Governance: An Emerging Imperative, MC Press, 2012.
[9] E. Lesser, R. Shockley, Analytics: The new path to value. 2014. URL: http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-embedding-analytics.html.
[10] G. Finch, S. Davidson, C. Kirschniak, M. Weikersheimer, C. Rodenbeck Reese, R. Shockley, Analytics: The speed advantage. why data-driven organizations are winning the race in today's marketplace. 2014. URL: http://www-935.ibm.com/services/us/gbs/thoughtleadership/2014analytics/?cm_mc_uid=48208801620614296137181&cm_mc_sid_50200000=1429613718.
[11] E. Lundquist, Data quality is first step toward reliable data analysis. 2013/07/22/ 2013. URL: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=89867448&lang=es&site=ehost-live.
[12] O. Kwon, N. Lee, B. Shin, Data quality management, data usage experience and acquisition intention of big data analytics, Int. J. Inf. Manage. 34 (2014) 387–394. URL: http://www.sciencedirect.com/science/article/pii/S0268401214000127.
[13] J. Becla, D.L. Wang, K.T. Lim, Report from the 5th workshop on extremely large databases, Data Sci. J. 11 (2012) 37–45. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84859721986&partnerID=40&md5=fc683361d4e5427bd6fe1780713b0c51.
[14] D. Loshin, Big Data Analytics: From strategic planning to enterprise integration with tools, in: Techniques, NoSQL, and Graph, Elsevier, Walthman, MA, USA, 2013.
[15] ISO, ISO/IEC 25010, Systems and software engineering— Systems and software quality requirements and evaluation (square)—System and software quality models, 2011.
[16] ISO, ISO/IEC 25012:2008—Software engineering. Software product quality requirements and evaluation (SQuaRE). Data quality model, Report, International Organization for Standarization, 2009.
[17] ISO, ISO/IEC CD 25024—Systems and software engineering—Systems and software quality requirements and evaluation (square)—Measurement of data quality, Report, International Organization for Standarization, 2014.
[18] G. Inc., Gartner's IT glossary. URL: http://www.gartner.com/it-glossary/big-data 2015.
[19] R. Mahanti, Critical success factor for implementing data profiling: The first step toward data quality, Softw. Qual. Prof. 16 (2) (2014) 13–26.
[20] T.C. Redman, Data's credibility problem, Harv. Bus. Rev. 91 (12) (2013) 84–88. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=92545713&lang=es&site=ehost-live.
[21] J. Parkinson, Six big data challenges. CIO INSIGHT. 2012. URL: http://www.cioinsight.com/c/a/Expert-Voices/Managing-Big-Data-Six-Operational-Challenges-484979.
[22] D.M. Strong, Y.W. Lee, R.Y. Wang, Data quality in context, Commun. ACM 40 (5) (1997) 103–110. http://dx.doi.org/10.1145/253769.253804. URL: http://doi.acm.org/10.1145/253769.253804.
[23] P. Berka, M. Sochorova, A collaborative effort in knowledge discovery from databases. guide to the financial data set. 1999. URL: http://lisp.vse.cz/pkdd99/Challenge/chall.htm.
[24] J. Merino, Functions per profiles in the 3as data quality model. 2015. URL: http://www.alarcos.esi.uclm.es/download/mobid15/fpp.pdf.

Merino Jorge is M.Sc. in Computer Science from the University of Castilla-La Mancha and Research assistant in the same university. His research interests are Data Quality, Quality Assessment, Standardization and Big Data.

Caballero Ismael works as associate professor at the University of Castilla-La Mancha, Spain. His main Research interests are on data and Information Quality management, and Data Governance.

Rivas Bibiano is B.Sc. in Computer Science from the University of Castilla-La Mancha and Research assistant in the same university. His research interests are Data Quality, Quality Assessment and Big Data.

Serrano Manuel is M.Sc. and Ph.D. in Computer Science from the University of Castilla-La Mancha. Assistant Professor at the Escuela Superior de Informtica of the Castilla-La Mancha University in Ciudad Real. His research interests are Data and Software Quality, Software measurement, DataWarehouses Quality and Measures and Big Data.

Piattini Mario is a full professor of Computer Science at the University of Castilla-La Mancha, Spain. His research interests include Global Software Development and Green Software. He received a Ph.D. in Computer Science from the Universidad Politcnica de Madrid.

Please cite this article in press as: J. Merino, et al., A Data Quality in Use model for Big Data, Future Generation Computer Systems (2015), http://dx.doi.org/10.1016/j.future.2015.11.024