Abstract. Organizations are nowadays immersed in the Big Data Era. Beyond the hype of the concept of Big Data, it is true that something in the way of doing business is really changing. Although some challenges keep being the same as for regular data, with big data, the focus has changed. The reason is due to Big Data is not only data, but also a complete framework including data themselves, storage, formats, and ways of provisioning, processing and analytics. A challenge that becomes even trickier is the one concerning to the management of the quality of big data. More than ever the need for assessing the quality-in-use of big datasets gains importance since the real contribution – business value - of a dataset to a business can be only estimated in its context of use. Although there exists different data quality models to assess the quality of data there still lacks of a quality-in-use model adapted to big data. To fill this gap, and based on ISO 25012 and ISO 25024, we propose the 3Cs model, which is composed of three data quality dimensions for assessing the quality-in-use of big datasets: Contextual Consistency, Operational Consistency and Temporal Consistency.

Keywords: Big Data, Data Quality, Quality-in-use model, 3Cs Model.

1 Introduction

If defining data quality was difficult, finding a sound definition for data quality for big data is even worse: there is still not an official definition for big data. Loshin in [1] gathers the definition given by Gartner’s IT Glossary: “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”. As [2] states, big data is an umbrella term that covers not only datasets themselves, but also problem space, technologies, and opportunities for enhancing business value.

And precisely, business value is the main reason for what big data can be used for, in fact, analytical projects aimed to extend the meaning of the facts currently happening in business processes in different business cases [1, 3]. In such business cases, regrettably, people in high management tend to think that the larger the big data project is (e.g. the largest amount of data involved in the project), the larger benefits (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 [2].
So, the first step in any big data project is to encourage high management to lead the project over to buy and deploy sophisticated technology that will not produce any suitable results for the business case at hand [4-6].

Once managers are convinced about the real need of kicking big data projects, they have to deal with the challenges that big data brings in order to achieve an alignment to the reality of the organizations [1]. The challenges have been identified in [7] as being 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.

In this paper, we are going to deal with data quality concerns in big data projects independent of which type of project the organization is facing. Our motivation is to define the quality characteristics that the different datasets being used for a specific use should present to fit for such use (quality in use).

Data quality is per se a real challenge although not a myth: all datasets included in big data projects do not necessarily have to lack of quality enough for the task at hand [8]. Anyway, we pose that studying how data quality management for big data has evolved is worthy to research since classical data quality principles cannot further be applied to big datasets due to its new very nature [1]. In this sense, data quality management for big data should prioritize those data quality dimensions really addressing the data quality requirements for the task at hand. Although dealing with internal or external data quality is actually important, sometimes the assessment of these types of quality for big datasets is almost impossible because there may be limited or multiples or even no restrictions on data [9], and consequently these studies are not producing results for developers as relevant as the results provided by the one addressing the levels of data quality-in-use. Unfortunately, not much research has still been conducted for quality-in-use for big data. Even when ISO 25010[10] deals with internal and external quality and quality in-use, and in spite that ISO 25012 [11], which addresses data quality concerns, does not specifically differentiate between internal, external or in-use quality for “traditional” data neither big data. So, there still lacks a quality-in-use model which can be used as a reference to manage data quality in big data.

Our contribution in this paper is precisely “The 3Cs data quality-in-use model for big data” which is to be specifically addressed and customized for big data projects. The 3Cs corresponds to the three kind of consistency required for most of the initiatives in which big datasets are to be integrated for part of a project: Contextual Consistency, Temporal Consistency and Operational Consistency. In the scope of this paper, the concept of consistency of 3C’s further specializes the concept of consistency in ISO 25012. For each one of the 3Cs, we will also analyze how the 3Cs are related to the 3Vs that characterize big data, namely, Volume, Velocity and Variety. To the best of our knowledge, the topic has not been previously tackled by others by addressing ISO 25012. The novelty of our proposal is the provision of predefined data quality dimensions sets to be used in specific situation as if they would be quality patterns. The proposal introduces an on-progress research whose results are to be applied to specific kind of big data quality projects, and validated against real case studies.

The remainder of the paper is structured as follows: Section 2 introduces a flash-back motivating how data quality management have evolved over the time. Section 3