DATA QUALITY EVALUATION FRAMEWORK FOR BIG DATA

By

GRACE AMINA ONYEABOR *

AZMAN TA'A **

* Lecturer, Department of Information Science, University of Ibadan, Nigeria.
** Senior Lecturer, Department of Information Science, University of Ibadan, Nigeria.

ABSTRACT

Data is an important asset in all business organizations of today. Thus the results of its poor quality can be very grievous leading to erroneous insights. Therefore, Data Quality (DQ) needs to be evaluated before the analysis of any Big Data (BD). The evaluation of DQ in BD is challenging. Given the enormous datasets that are of varied format fashioned at a rapid speed, it is impossible to use the traditional methods of evaluating DQ in BD. Rather, there is a requirement of strategies and devices for the assessment and evaluation of DQ in BD in a rapid and more efficient manner. However, assessing the quality of data on the whole BD can be very expensive. In addition, there is also a need for improvement in data transformation activities of BD. This paper proposes a framework for DQ evaluation with the application of data sampling technique on BD sets from different data sources reducing the size of the data to samples representing the population of the BD sets. The Bag of Little Bootstrap (BLB) sampling technique will be used. The target Data Quality Dimensions (DQDs) to be used in this paper are completeness, consistency, and accuracy. In addition, the DQDs will be measured using different metric functions relevant to the DQDs. This will be done before and after an improved data transformation techniques to check the improvement of DQ in BD.

Keywords: Big Data, Data Sampling, Data Transformation, Data Quality Evaluation.

INTRODUCTION

Data is considered today as an asset to various organizations. This is due to the fact that business strategic decisions are founded on the insights from the generated data. Data from wherever originally holds irregularities and discrepancies, such as dirty, incomplete, or missing data triggered by several elements like readings from sensory devices, data entry by humans just to mention a few. The term BD can be described as vast datasets with variety of formats generated at a fast speed, which is almost impossible to be managed using traditional or the classical database management systems. Nowadays, there are a lot of commercial businesses and organizations producing huge datasets in same way, where vast number of data is being acquired, shared, and stored in various sources (Maier, Serebrenik, & Vanderfeesten, 2013). Quality in itself is complex, multidimensional and with continuous process, referring to diverse facets ranging from service quality to software quality to DQ (Taleb, El Kassabi, Serhani, Dssouli, & Bouhaddioui, 2016). The authors went on to say that quality is related to domain, it is well-defined via a set of attributes and depends on dimension and methods of assessment. This shows that a good knowledge of the domain of the data and DQDs are some of the main requirements for a DQ evaluation. Consequently, DQ can be captured by making use of multiple measures and assessment tools for numerous domain activities. In BD context, a vital problem exist within the data itself and thus in its quality.

Going by the wide acceptance and usefulness of BD by several organizations, it comes with some challenges. These challenges are recognized in (Merino, Caballero, Rivas, Serrano, & Piattini, 2016) namely: DQ, adequate characterization of data, interpreting correctly the results, visualization of data, real time data view versus retrospective view, and determination of the importance of projects and results. The challenge that is fiddler is DQ,