2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress

Big Data Quality: A Quality Dimensions Evaluation

Ikbal Taleb[1], Hadeel T. El Kassabi[1], Mohamed Adel Serhani[2], Rachida Dssouli[1], Chafik Bouhaddioui[3].

[1]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC, Canada
{i_taleb, h_elkass}@encs.concordia.ca, rachida.dssouli@concordia.ca
[2]College of Information Technology, [3]College of Business and Economics, UAE University, Al Ain, UAE
{serhanim, chafikb}@uaeu.ac.ae

Abstract— **Data is the most valuable asset companies are proud of. When its quality degrades, the consequences are unpredictable and can lead to complete wrong insights. In Big Data context, evaluating the data quality is challenging and must be done prior to any Big data analytics by providing some data quality confidence. Given the huge data size and its fast generation, it requires mechanisms and strategies to evaluate and assess data quality in a fast and efficient way. However, checking the Quality of Big Data is a very costly process if it is applied on the entire data. In this paper, we propose an efficient data quality evaluation scheme by applying sampling strategies on Big data sets. The Sampling will reduce the data size to a representative population samples for fast quality evaluation. The evaluation targeted some data quality dimensions like completeness and consistency. The experimentations have been conducted on Sleep disorder's data set by applying Big data bootstrap sampling techniques. The results showed that the mean quality score of samples is representative for the original data and illustrate the importance of sampling to reduce computing costs when Big data quality evaluation is concerned. We applied the Quality results generated as quality proposals on the original data to increase its quality.**

Keywords - Big Data; data quality dimensions; data quality evaluation; Big data sampling.

I. INTRODUCTION

Nowadays, most of all small and big companies consider data as an asset in an era where almost all business and politics strategic decisions are based on insights from data. Originally, data is incomplete and might contain a lot of discrepancies and inconsistencies such as poor, missing and incomplete data. These data anomalies are caused by many factors including human factor. In Big Data environments, data is the most vital element that travels through all phases of its lifecycle. Such phases include data processing and analytics. However, without a ready-to-go data these phases will not prevail. Yet, any data processing remains very sensitive when data is not suitable, clean and ready to be processed. Improper data can generate biased analytics caused essentially by factors such as bad preparation, nature of data, including format, origin, and type.

To define data quality we must first define quality and its characteristics. Since Quality is complex, multi-dimensional, and continuous process, it usually refers to different aspects ranging from quality of service, quality of software to quality of data. Additionally, quality is (1) domain related, (2) defined through a set of attribute(s), (3) relies on measurement and assessment methods. In other words, a deep knowledge of the data domain, a well-defined data attributes and a targeted quality dimensions are major requirements for any quality assessment. Therefore, data quality can be captured using a multitude of measures and assessment tools for several different areas and domain activity.

In the context of Big Data, a crucial problem resides in the data itself and consequently in its quality. There are many Big Data characteristics that have a direct impact on Data Quality (DQ). Data Variety is one of the four characteristics of Big Data. It describes the diversity of data sources and its multiple formats. The variety of data gives an intuitive idea about data quality. For example, a data warehouse is a structured schema-based while social media data is unstructured and schema-less. Data velocity is another quality characteristic where higher volumes of data are being speedily generated, this involves more quality parameters to be considered for quality evaluation such as timeliness. Consequently, all these parameters have a direct impact on data quality. Thus, the data require a preparation phase to build some confidence and insure somehow its quality.

In this paper, we propose a fast big data quality evaluation scheme by applying sampling strategies on large data sets. The Sampling will reduce the data size to a representative population samples, for a fast quality evaluation. We are looking into the data quality of Big Data using data-driven approach. Each data source, which gets into Big Data, is profiled and its quality is estimated prior to any inclusion in Big Data lifecycle processes. This evaluation provides well-constructed data quality information about data attributes and their statistics within some selected quality dimensions. This information provides a good strong start when planning a big data analytics project, by targeting the best attributes and data sets that their quality evaluation achieves a sufficient confidence level.

The paper is organized as follows: next section presents and discusses related works around data quality evaluation in Big Data. In Section III, we briefly describe and discuss data quality issues and quality dimensions in the context of Big Data. Section IV, introduces the Big Data quality evaluation

978-1-5090-2771-2/16 $31.00 © 2016 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.145

based on BLB Big data Bootstrap sampling algorithm. Section V describes the experimentations and discusses the data quality estimation algorithm developed based on some quality metrics. Section VII concludes the paper and proposes possible extensions related to data quality dimensions.

II. Related Works

In this paper, we investigate the evaluation of Big data quality. It is characterized by many challenges that need to be tackled from different angles, the most important ones are: data size, speed of generation, data attributes, Data Quality Dimensions (DQD) and their measurement metric. Very few works have been done on Big Data quality evaluation. However, these research initiatives have different point of views and address quality from different perspectives. Some attempted to provide a solid general definition for data quality [1] others defined quality from dynamic viewpoint and based on the domain of the data [2]. Most of works have agreed that data quality is related to the phases or processes of data life cycle [3]. Specifically, data quality is highly coupled with data generation phases and/or with its origin. Hereafter, we describe example of approaches used to assess the quality of data based on traditional data strategies that were adopted and adapted to Big Data quality assessment. The type of data to be evaluated affects the quality evaluation metrics. It is Content-based, Context-based, or Rating-based. In Content-based metrics, the information itself is used as quality indicators, while in Context-based metrics meta-data is used as quality indicators. On the other hand, Rating-based metrics use explicit ratings of both the information, and the sources of information [4].

Authors in [5] classified the data quality issues for any data (Big Data or not) to the following types: data error correction, unstructured data conversion, and integrating data from multiple sources. More issues are also discussed for Big Data specifically like large volumes of data, vast speed and schema-less structures. In [1], [4], [6], [7] they identify also some of Big Data quality problems correlated to some Big Data characteristics.

Data quality assessment was discussed early in literature as in [8] where they divide data quality assessment into two main categories: subjective and objective. Furthermore, they provide an approach that combines these two categories to provide organizations with usable data quality metrics to evaluate their data. However, their approach was not meant to deal with Big Data. More recently, authors in [4] propose a framework to evaluate and manage Big Data quality in the domain of social media during each phase along the Big Data pipeline. This solution is limited to a specific domain of Big Data and introduced limited quality attributes and did not consider some data sources like feedback data from the customer, data about the product and market analysis.

Another approach was suggested in [9] where their quality metrics are based on categorizing the purpose for which the data to be produced for or consumed by.

In [10], the authors presented a comprehensive studies on Big Data quality issues related to computing infrastructure like hardware faults, code defects, human errors, configuration and their possible solutions. On the same matter, In [11], only the big data computations under restricted resources are targeted. They designed an elastic mining algorithm to approximate quality results when varying cost, time and resources allocations.

Finally, most of the related works on Big data quality missed the main problem of Big data quality, which consist of how to evaluate this quality, what to evaluate, and what is the purpose of this evaluation. We believe that Big Data quality has to be addressed and evaluated as early as possible before engaging in any Big data quality evaluation Project. Specific mechanisms need to take place to achieve this perception. The results of such process leads to a specific tasks that increase the quality.

In this paper, we propose Quality of Big Data evaluation scheme to gather important insight about data attributes quality and profile. This information is used to suggest for the Big data evaluation some quality rules that must be taken into consideration when preparing the data analytics plan. These quality rules are extracted from the evaluation of quality dimensions results and will help improving the Big Data sets by correcting and eliminating data or attributes that most probably hurts any data analytics.

III. Data Quality

The need to evaluate Big Data Quality is justified by the high impact poor data has on analytics results. All companies from different domains rely on data when planning their short and long terms strategies. But before any of the aforementioned we need to get an outlook of the Big data quality by estimating and evaluating what data quality is made of?

In the following, we briefly describe important elements to handle data quality evaluation on classical data and eventually on Big data.

Data and Data Types: according to [12] and [13] the data is always recorded using a schema providing a well-organized structure. With the emergence of social media, data is unstructured and semi-structured.

Data Quality (DQ) Definition: In [14], data quality was summarized from ISO 25012 Standard as "*the capability of data to satisfy stated and implied needs when used under specified conditions*". In the literature: "*fitness for use*".

Poor data, DQ issues and problems: Data is always altered due to many factors. When it needs a quality evaluation and improvement, these factors must be known

and classified under the data quality dimensions (DQD). Several factors or processes generated bad data: human data entry, sensors devices readings, social media, unstructured data, and missing values. The authors in [15], [14] enumerate many reasons of poor data which affect its quality elements and its related dimensions. In Table 1, a shortlist of the well-known data issues vs DQD.

Table 1 Data Quality Issues vs. DQD

		Data Quality Dimensions Related		
	Data Quality Issues	Accuracy	Completeness	Consistency
Instance level	*Missing data*	X	X	
	Incorrect data, Data entry errors	X		
	Irrelevant data			X
	Outdated data	X		
	Misfielded and Contradictory values	X	X	X
Schema Level	*Uniqueness constrains, Functional dependency violation*	X		
	Wrong data type, poor schema design			X
	Lack of integrity constraints	X	X	X

DQ Dimensions (DQD): many initiatives addressed data quality dimensions [1], [13], [16], the DQ is classified into four categories (Intrinsic, Contextual, Representational, Accessibility). A DQD offers a way to measure and manage data quality [17] [12]. Some popular DQD's are commonly cited in the literature, the following are the most used:

- Accuracy is defined as the closeness the data is represented from real-life event for which an attribute data value is assigned.
- Completeness measures the missing values.
- Consistency refers to the respect of data constraints.

DQ Evaluation, Metrics, and Measurement: any data can have its quality measured. Using a data driven strategy, the measurements acts on the data itself to quantify the DQD. As mentioned before, our work is based on structured data represented in a set of attributes, columns, and rows with their values. Any data quality metric should specify whether the values of data respect or not the quality attributes. In [1], the author quoted that data quality measurement metrics tend to evaluate a binary results correct or incorrect or a value between 0 and 100 (100% is the best), and use universal formulas to compute these attributes. This will apply to many quality dimensions such as accuracy, completeness, and consistency.

The DQDs must be relevant to the DQ problems as identified in Table 1. Therefore, DQ Metrics are designed for each DQD to measure if the attributes respect the previously defined DQD. These measures are done for each attribute given its type, data ranges values, and if it is collected from data profiling.

For example a metric that calculates the accuracy of a data attribute is defined as follows:

- The data type of an attribute and its values.
- For numerical attributes, a range or sets of acceptable values (Textual also) are defined. Any other values are incorrect.
- The accuracy of an attribute is calculated based on the number of correct values divided by number of observations or rows. Table 2 lists the metric used to calculate the DQD's scores.
- For another data types/formats like images, videos, audio files, another type of metrics must be defined to evaluate accuracy or any other quality dimensions. The authors of [13] describe usefulness as an aspect of data quality for images. For this kind of data, features extraction functions are defined on the data and extracted for each data item. These features have constraints that characterize the goodness or badness of data values. Some of the quality metrics functions are designed based on the extracted features such as, usefulness, accuracy, completeness (based on many features) and any other data quality dimensions judged by domain experts to be candidate for such data type (e.g. video, image, or audio).

DQ issues and Big Data characteristics: The main Big Data characteristics commonly named V's are initially, Volume, Velocity, Variety and Veracity. Since the Big Data inception, we reached now 7 V's and probably we will keep going [18]. The veracity tends more to express and describe trust and certainty of data that can be expressed mostly as quality of the data. The DQD accuracy is often related to precision, reliability and veracity [19]. A mapping tentative between these characteristics, data and data quality is compiled in [6], [13], [16]. The authors attempted to link the V's to the quality dimensions. In another study, the authors of [20] addressed the DQD "Accuracy" versus Big Data characteristic "Volume". They conclude, that the increase in data size has high impact on DQ improvements.

Table 2. DQD metric functions

DQ Dimensions	Metric functions
Accuracy	*Acc = (Ncv / N)*
Completeness	*Comp = (Nmv / N)*
Consistency	*Cons = (Nvrc / N)*
Ncv	*Number of correct values*
Nmv	*Number of missing values*
Nvrc	*Number of values that respects the constraints*
N	*Total number of values (rows) of the sample Dataset*

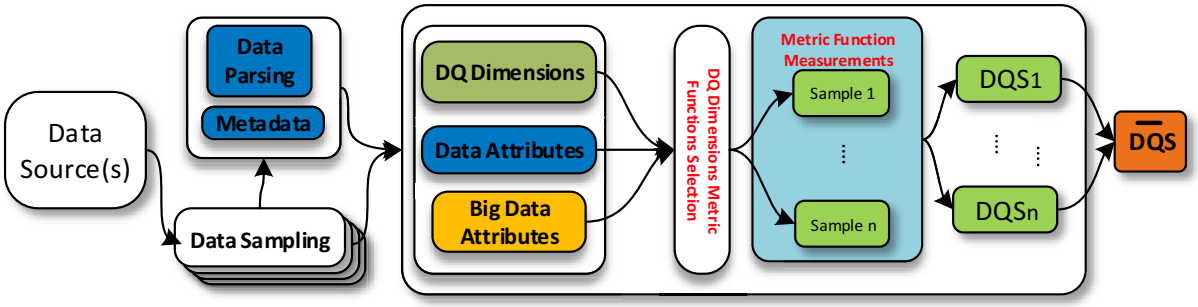

Figure 2. Big Data Quality Evaluation Scheme

IV. BIG DATA QUALITY EVALUATION SCHEME

The purpose of Big Data Quality Evaluation (BDQ) Scheme is to address the data quality before starting data analytics. This is done by estimating the quality of data attributes or features by applying a DQD metric to measure the quality characterized by its accuracy, completeness or/and consistency. The expected result is data quality assessment suggestions indicating the quality constraints that will increase or decrease the data quality. We believe also that data quality must be handled at many other phases of data lifecycle. However, it is out of scope of this work.

In this paper, we are dealing with data quality of data source, more precisely in its dataset(s). This evaluation is essential to assure a certain quality levels for any related processes with an optimal costs. Here, we should highlight that Big Data Quality is essential since we cannot produce strong estimates of the cost of our analytics.

The BDQ Evaluation scheme is illustrated in Figure 2 where the data goes through many module to estimate its quality. The key modules of our scheme consist of: (a) data sampling, and data profiling, (b) DQD vs attributes selection, (c) data quality Metric selection, (d) samples data quality evaluation. In the following sections, we describe each module, its input(s), output(s), and the main functions.

A. *Big Data Sampling*

There are several sampling strategies that can be applied on Big Data as expressed in [21], [22]. They evaluated the effect of sampling methods on Big Data and believed that sampling large datasets reduces run time and computational footprint of link prediction algorithms though maintaining sufficient prediction performance. In statistics, Bootstrap sampling technique evaluates the sampling distribution of an estimator by sampling with replacement from the original sample. In the context of Big Data, Bootstrap sampling has been addressed in many works [23]–[25]. In our data quality evaluation scheme, we decided to use the Bag of Little Bootstrap (BLB) [25], which combines the results of bootstrapping multiple small subsets of a Big data dataset. The BLB algorithm use an original Big dataset used to generate small samples without replacements. For each generated sample another set of samples are created by resampling with replacements.

B. *Data Profiling*

Data profiling module performs screening of data quality based on statistics and information summary. Since profiling is meant to discover data characteristics from data sources. It is considered as data assessment process that provides a first summary of the data quality. Such information include: data format description, different attributes, their types and values. data constraints (if any), data range, max and min. More precisely information about the data are presented in two types; technical and functional. This information can be extracted from the data itself without any additional representation using it metadata or any descriptive header file, or by parsing the data using any analysis tools. This task may become very costly in Big Data. To avoid costs generated due the data size we will use the same sampling process BLB to reduce the data into a representative population sample, in addition to the combination of profiling results.

C. *Data Quality Evaluation*

The data profiling provides information about the dataset:

- Data attributes (e.g. type, format)
- Data summary (e.g. max, min)
- Big data attributes: size, number of sources, speed of data generation (e.g. data streams)
- What DQDs to evaluate.

The previous information's are used to select the appropriate quality metrics functions $\boldsymbol{F}$ to evaluate a data quality dimensions $\boldsymbol{d_k}$ for an attribute $\boldsymbol{a_i}$ with a weight $\boldsymbol{w_j}$.

In the Figure 3, we describe how data quality is evaluated using bootstrap sampling for Big data. The process follows 5 steps:

1) Sampling from the data set **S** $\boldsymbol{n}$ bootstrap samples of **ss** size without replacement $\mathbf{DS_i}$.
2) Each sample generated from step 1 is sampled into n' samples of size **SS** with replacements $\mathbf{DS_{ij}}$.
3) For Each sample $\mathbf{DS_{ij}}$ generated in step 2, evaluate the data quality score $\mathbf{Q_{ij}}$

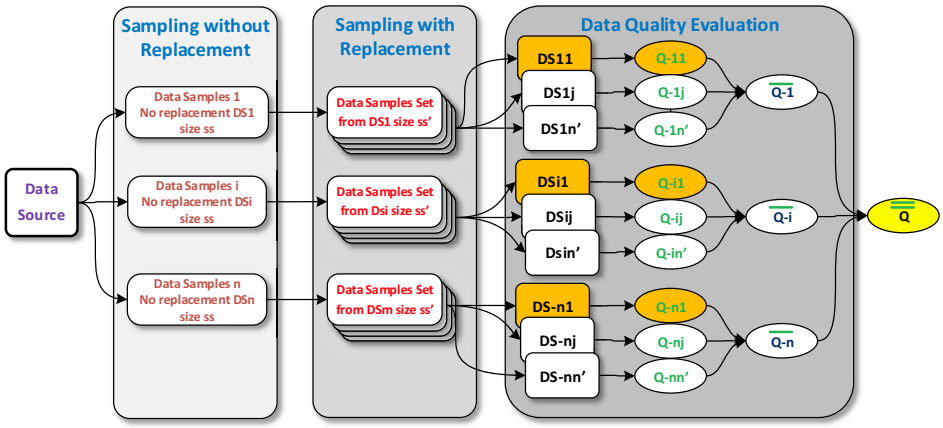

Figure 3. Big Data Quality Sampling Evaluation

4) For all the samples **DS$_i$**, evaluate the data quality score **Q$_i$** which represents the mean of all **n'** samples quality scores Q$_{ij}$.
5) For the data set **S**, evaluate the quality score **Q** which represents the mean of all **n** samples quality scores **Q$_i$**.

Table 3. Big Data Quality Evaluation Algorithm

Algorithm: Big Data Quality Evaluation

```
Let ds a Original Data Set with size SS and N Observation (N~SS)
Let ss ( b(SS) )the samples size with ss < SS
Let n samples s_i of size ss and M Observation ( M~ss)
Let D a set of DQD D= {d_0,…,d_k,…, d_q}
Let F a metric functions F (completeness, accuracy,...)
Let cc ←0 counter of correct, valid attribute value(when F is true cc=cc+1)
Let S = {DS_0,…,DS_i,…, DS_n} without replacement
For each iteration i from 0 to n
    Generate a sample s_i of size SS from ds
    For each iteration j from 0 to n'
        //Generate a sample s_ij of size SS from sample s_i
        For each DQD metric Function tuple (d_k, F)
            For each attribute a_ij
                For each a_ij(x) ss values
                    If ( F(a_ij(x), value)== 1) // measure metric
                        cc ← cc+1
                End a_ij(x)
            Calculate the scores vector DQD (F, d_k, a_ij, DS_i) = cc/N
            cc ←0 // counter of correct valid attribute value (d_k, F)
        End a_ij
        // DQD d_k computed for all attributes for a sample ds_ij
    End (d_k, F)
    // DQS_ijk is the d_k scores for an attribute a_ij for sample DS_ij
    Q_ijk sum of all d_k scores for attribute a_ij for DS_ij
End j
Q_ik += 1/n' (Q_ijk)
End i
// Q_k is the mean of all Q_ik for a specif d_k
Q_k += 1/n (Q_ik)
```

D. *BDQ Evaluation Algorithm*

Let ***F*** represents a set of data quality metrics**,** ***F*** = ***{f$_0$,…,f$_i$,…, f$_m$}*** where ***f$_i$*** a quality metric function that will measure and evaluate a DQD ***d$_k$*** for each value of an attribute ***a$_i$*** in the sample ***s$_i$*** and returns 1 if correct, 0 if not. Each ***f$_i$*** function will compute if the value of the attribute reflects the ***d$_k$*** constraints. For example, the metric accuracy of an attribute is defined as a range of values between 0 and 100, otherwise it is incorrect. Similarly, it can be defined to satisfy a certain number of constraints related to the type of data such as: a zip code, email, social security number, or an address. If we are evaluating the same DQD ***d$_k$*** for a set of attributes, if the weights are all equal, a simple mean is computed. The metric ***f$_i$*** will be evaluated to measure if all the attributes individually have their ***f$_i$*** correct. This is done for each instance (cell or row) of the sample ***s$_i$***.

In Table 3, we describe the detail of BDQ Evaluation Algorithm. The **Q$_k$** represents the mean quality score for a DQD **d$_k$** for measurable attributes. For the data set let note ***A*** as a set of attributes or features. The **Q$_k$** values respectively for each attribute are represented by a set of quality scores:

V = {Q$_k$a$_1$… Q$_k$a$_m$} **where** ***A*** is a set of ***m*** attributes. With this evaluation, we have more insights, statistics and benefits about the Big data quality to ensure a well-refined analytics that targets the best precision.

E. *After evaluation Analysis*

The data evaluation process done on Big data set provides data quality information and scores of quality dimensions of each attributes or features. These scores are used to identify the data that must be targeted and omitted. A set of proposals actions is generated based on many parameters, like DQD, or data quality issue. If a data attribute got a lower score than the required level (%) of accuracy or completeness the following actions are proposed:

- Discard it from the dataset.
- Tune, reformat, and normalize its values.

- Replace values, as in missing data.

Whatever the Quality evaluation results, it always contains actions to be taken on the dataset to remove any irregularities using techniques like cleaning, filtering and pre-processing based on the quality assessment.

V. EXPERIMENTATIONS, RESULTS AND ANALYSIS

In this section, we describe the experimentations we have conducted to evaluate the DQ of big data. DQD were measured using a set of quality metrics.

A. *Setup*

For our experimentations, we used a computers equipped with 16 GB of RAM, an Intel i7 quad-core (2.66 GHz) running a 64 bits virtual machine Vagrant-VM as a Spark cluster, running Apache SPARK 1.6.1 with Spark R (support for R language) and Jupyter Notebook with Kernels (PySpark, Python 2.7.5, Scala, R).

B. *Dataset description*

A Sleep Heart Health Study (SHHS) dataset [26] have been used for our experiments, it is used to assess effects of sleep-disordered breathing. The SHHS dataset is collected from 6441 people. It contains data attributes such as ECG, EEG, EOG, EMG, thoracic and abdominal excursions, nasal airflow, oxygen saturation, ECG, and heart rate. Each patient's data is represented in EDF format with 40 MB of size. The data set is represented by 1278 attributes.

C. *Scenarios:*

Two scenarios have been developed to evaluate the quality of Big Data set. The first scenario evaluate the completeness of the data set, the second scenario evaluates its consistency.

1) Scenario 1: the evaluation of DQD completeness is calculated by measuring if an attribute has a recorded value of the data in all observations (rows). By looking for missing values in the data set represented by NA or no data. The result is the percentage of missing data in a dataset.

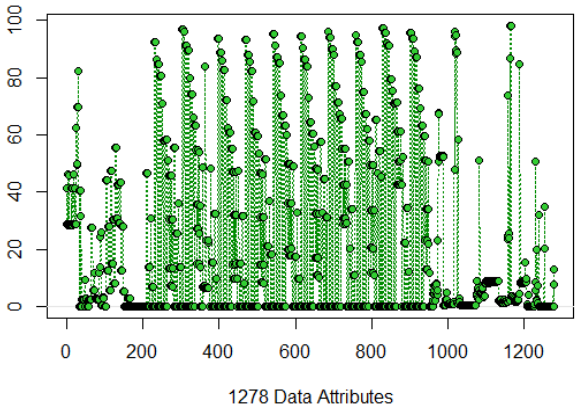

Figure 4. Missing data % vs attributes.

From Figure 4 and 5 we can infer that almost 80% of the attributes have less than 60% of missing data. This information provides a set of steps to takes to get rid of these missing data. Many proposals are highlighted after the evaluation process in a set of actions to improve the data quality. In the following, a sample of proposed actions that the experts must refine and use. The suggested ratio values are only used to explain the results proposals and valid them with real scores:

1. Discard rows with attributes >= 80% of missing data.
2. Discard attributes (columns) >= 80% missing data.
3. Replace missing data (rows) with the attribute mean for the attributes that have 20% of missing data. (The expert will judge that 80% mean is a representative value).
4. A combination of the above actions will optimize the data improvement process by keeping the most important attributes. The latter are targeted from the analytics experts; this is done by applying priority weight on attributes quality.

2) Scenario 2: the evaluation of DQD consistency is done by checking if an attribute or a set of attributes respects some data constraints in all observations (rows). Here, the constraint is completeness, which is applied to all attributes. Only the complete observations are scored correct, and if any attribute has a missing data, its consistency decreases. Consistency is defined as the conformance of data values to other values in the Data Set.

Based on the completeness experimentations, and with the hypothesis that we are considering all the 1280 attributes of the data set; the consistency evaluation gave subsequently the following results to achieve high consistency:

1. A 5% (65) attributes have more than 90% missing data.
2. A 29.1% of attributes have 0% missing data. If we keep only these attributes we will achieve 100% consistency.
3. We achieve only 29.1% consistency when using all the attributes.

The design of a metric is imperative, since we can combine many constraints scores gathered from others DQD evaluation to compose a new specific understanding.

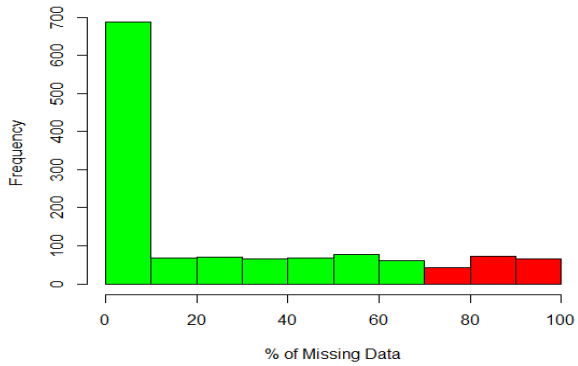

Figure 5. Number of attributes and their % of missing data.

VI. CONCLUSION

In this paper, we proposed a Quality of Big Data evaluation scheme to generate a set of actions to increase the data quality of Big data set. We developed a Big Data quality evaluation algorithm based on BLB, a bootstrap sampling for Big data. The BLB sampling helped achieving an efficient DQ evaluation by reducing computing time and resources. The experimentations we conducted on large sleep-disordered dataset, showed that the data quality of a large data set can be restricted to a small representative data samples. The results are data quality scores and a set of based generated proposals. Each proposal targets a DQD for a dataset attributes. These proposed actions are applied on the source data set to enforce and increase its quality. As future work, we are planning to develop an automatic optimizations and discovery of quality proposals based on DQD evaluation results. Also, build a DQD context metric and/or model for Big data and use it as a reference for automatic generation of DQD metric.

VII. REFERENCES

[1] S. Juddoo, "Overview of data quality challenges in the context of Big Data," in *2015 International Conference on Computing, Communication and Security (ICCCS)*, 2015, pp. 1–9.

[2] H. M. Sneed and K. Erdoes, "Testing big data (Assuring the quality of large databases)," in *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 2015, pp. 1–6.

[3] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, "Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model," in *2014 47th Hawaii International Conference on System Sciences (HICSS)*, 2014, pp. 4700–4709.

[4] A. Immonen, P. Paakkonen, and E. Ovaska, "Evaluating the Quality of Social Media Data in Big Data Architecture," *IEEE Access*, vol. 3, pp. 2028–2043, 2015.

[5] P. Oliveira, F. Rodrigues, and P. R. Henriques, "A Formal Definition of Data Quality Problems.," in *IQ*, 2005.

[6] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Sci. J.*, vol. 14, no. 0, p. 2, May 2015.

[7] J. Krogstie and S. Gao, "A Semiotic Approach to Investigate Quality Issues of Open Big Data Ecosystems," in *Information and Knowledge Management in Complex Systems*, K. Liu, K. Nakata, W. Li, and D. Galarreta, Eds. Springer International Publishing, 2015, pp. 41–50.

[8] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[9] L. Floridi, "Big Data and Information Quality," in *The Philosophy of Information Quality*, L. Floridi and P. Illari, Eds. Springer International Publishing, 2014, pp. 303–315.

[10] H. Zhou, J. G. Lou, H. Zhang, H. Lin, H. Lin, and T. Qin, "An Empirical Study on Quality Issues of Production Big Data Platform," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE)*, 2015, vol. 2, pp. 17–26.

[11] R. Han, L. Nie, M. M. Ghanem, and Y. Guo, "Elastic algorithms for guaranteeing quality monotonicity in big data mining," in *2013 IEEE International Conference on Big Data*, 2013, pp. 45–50.

[12] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in *2012 International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012, pp. 300–304.

[13] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the Meaningfulness of 'Big Data Quality' (Invited Paper)," in *Data Science and Engineering*, Springer Berlin Heidelberg, 2015, pp. 1–15.

[14] M. Chen, M. Song, J. Han, and E. Haihong, "Survey on data quality," in *2012 World Congress on Information and Communication Technologies (WICT)*, 2012, pp. 1009–1013.

[15] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A Survey on Data Quality: Classifying Poor Data," in *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2015, pp. 179–188.

[16] I. Caballero, M. Serrano, and M. Piattini, "A Data Quality in Use Model for Big Data," in *Advances in Conceptual Modeling*, M. Indulska and S. Purao, Eds. Springer International Publishing, 2014, pp. 65–74.

[17] I. Taleb, R. Dssouli, and M. A. Serhani, "Big Data Pre-processing: A Quality Framework," in *2015 IEEE International Congress on Big Data (BigData Congress)*, 2015, pp. 191–198.

[18] M. Ali-ud-din Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," in *American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the*, 2014, pp. 1–5.

[19] V. Goasdoué, S. Nugier, D. Duquennoy, and B. Laboisse, "An Evaluation Framework For Data Quality Tools.," in *ICIQ*, 2007, pp. 280–294.

[20] A. B. Philip Woodall, "An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics," 2014.

[21] V. Gadepally, T. Herr, L. Johnson, L. Milechin, M. Milosavljevic, and B. A. Miller, "Sampling operations on big data," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 1515–1519.

[22] G. Cormode and N. Duffield, "Sampling for Big Data: A Tutorial," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 1975–1975.

[23] F. Liang, J. Kim, and Q. Song, "A Bootstrap Metropolis-Hastings Algorithm for Bayesian Analysis of Big Data," *Technometrics*, vol. 0, no. ja, pp. 0–0, Jan. 2016.

[24] A. Satyanarayana, "Intelligent sampling for big data using bootstrap sampling and chebyshev inequality," in *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2014, pp. 1–6.

[25] A. Kleiner, A. Talwalkar, P. Sarkar, and M. Jordan, "The big data bootstrap," *ArXiv Prepr. ArXiv12066415*, 2012.

[26] S. Redline and et al., "Sleep Heart Health Study - National Sleep Research Resource." [Online]. Available: https://sleepdata.org/datasets/shhs. [Accessed: 14-Mar-2016].