An Hybrid Approach to Quality Evaluation Across Big Data Value Chain

Mohamed Adel Serhani
United Arab Emirates University

Hadeel Talaat El Kassabi
Concordia University Montreal

An Hybrid Approach to Quality Evaluation Across Big Data Value Chain

Mohamed Adel Serhani[1], Hadeel T. El Kassabi[2], Ikbal Taleb[2], Alramzana Nujum[1]

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

Abstract— **While the potential benefits of Big Data adoption are significant, and some initial successes have already been realized, there remain many research and technical challenges that must be addressed to fully realize this potential. The Big Data processing, storage and analytics, of course, are major challenges that are most easily recognized. However, there are additional challenges related for instance to Big Data collection, integration, and quality enforcement. This paper proposes a hybrid approach to Big Data quality evaluation across the Big Data value chain. It consists of assessing first the quality of Big Data itself, which involve processes such as cleansing, filtering and approximation. Then, assessing the quality of process handling this Big Data, which involve for example processing and analytics process. We conduct a set of experiments to evaluate Quality of Data prior and after its pre-processing, and the Quality of the pre-processing and processing on a large dataset. Quality metrics have been measured to access three Big Data quality dimensions: accuracy, completeness, and consistency. The results proved that combination of data-driven and process-driven quality evaluation lead to improved quality enforcement across the Big Data value chain. Hence, we recorded high prediction accuracy and low processing time after we evaluate 6 well-known classification algorithms as part of processing and analytics phase of Big Data value chain.**

Keywords—Big Data, Quality assessment, Metadata, Quality metrics, quality Metadata, Quality of process, Hybrid quality assessment.

I. INTRODUCTION

Data is exploding at rates that had never been experienced and perceived before. The data gathered from different sources such as healthcare has exceptionally grown in volume, velocity, variety, and veracity. These new trends characterize the phenomena known as "Big Data". This extensive data growth have urged organization's strategy to shift from traditional data management systems to Cloud enabled Big Data. The latter offers on-demand, scalable, and flexible data management proven to be efficient and cost-effective. In healthcare domain the increase of electronic medical data, advocated the use of sophisticated health systems to support health Big Data transfer, processing, storage, replication, analysis and retrieval. Big Data requires a high computation processing power to handle huge data labeled in petabytes.

Most of recent works [1] [2] [3] have proposed few initiatives to incorporate data quality, however, these initiatives remain premature and do not provide comprehensive solutions that guarantee quality in all Big Data processes. Building end-to-end quality enforcement in Big Data value chain is of vital importance. Cloud infrastructure and services help implementing QoS enforcement mechanisms for Big Data processes including Big Data storage, distribution, replication and retrieval. Such developments consider 1) Data provenance and annotation scheme to track the effect of data transformation occurred in each phase, 2) Cost optimization scheme for Big Data distribution, 3) QoS-aware Big Data resource allocation and scheduling, and 4) Extending Big Data technologies to incorporate QoS enforcement and management features. Across

In this paper, we propose an end-to-end quality enforcement solution to support health Big Data value chain. This solution tackle every Big Data process namely pre-processing, processing, analytics and visualization with the ultimate goal of guaranteeing efficient, and cost effective Big Data processes. It relies on extending Big Data technologies such as Hadoop MapReduce and develops new algorithms to incorporate QoS annotations and enforcement.

This paper is organized as follows: next section surveys existing work on Big Data quality evaluation. Section 3, identifies the key requirements of Big Data quality assessment and proposes a set of quality metrics along with quality metadata to evaluate Big Data quality. While section 4 proposes a hybrid model to quality assessment of Big Data value chain, section 5, evaluates this model using a large dataset collected on sleep-disordered breathing. Finally, section 6 concludes the paper and points to some future research directions.

II. BACKGROUND AND RELATED WORK

The quality of Big Data is affected by several characteristics like the variety of data and diversity of sources; as volume of data increases, the quality of data can increase or decrease, so a set of tradeoffs, concepts and data quality rules need to be considered when dealing with data in key phases Big Data value chain. The first phase is pre-processing; it is a crucial phase as it is the entry process of the Big Data value chain, thus many challenges can be encountered [1]. Evaluating the quality of data is one of these challenges that necessitate a considerable study. Two approaches have been identified to deal with data quality, which are: data-driven and process-driven as stated in [3], [2]. Process-driven aims to redesign or use best practices to handle data (e.g. processing, storage, and collection), while the data-driven strategy improves data quality using some pre-processing techniques such as filtering, and cleansing.

A classification model has been proposed in [4], and characterized pre-processing data quality issues into: (1) errors correction, (2) conversion from unstructured to structured data and (3) data integration from multiple sources. Not to mention other specific Big Data issues related to data volume, speed, and schema-less data, which makes Big Data cleansing an imperative activity to improve data quality. In [5], the authors highlighted that data quality issues increase once handling data from various sources. They also emphasized the cleansing processes overhead produced from uncontrolled speed of generated data.

Data provenance is a well-known concept used for distributed database with scientific and business data to evaluate its provenance quality [6]. With Big Data, data provenance is considered by [7] as relevant source of data to help evaluating its quality. Provenance data in Big Data can be used to trace data from its collection and any transformations it went through up to its visualization. A multi-layer framework for data provenance collection has been proposed in [8]. Data semantics is proposed in [9] to evaluate consistency of data quality dimension for Big Data management. NADEEF is a platform proposed in [10] to offer data cleaning tools for Big Data. NADEEF has been extended in [11] to cope with data streaming issues using different forms of quality rules. Likewise, the work of [12] developed a Big Data architecture platform for a pervasive sensory data quality.

In [13] and [14], data quality issues and challenges have been clearly identified and explained. Also, data quality dimensions, and their corresponding metrics have been explained and classified. Authors in [15] identified the key challenges of data quality in Big Data and proposed a comprehensive quality assessment process to evaluate quality for Big Data. However, [16] differentiated between subjective and objective quality assessment to identify quality discrepancies and propose actions for improvements.

Very few initiatives have been conducted to address Big Data quality assessment in Big Data value chain. In [17], a cross layer approach was proposed to assess data quality. This has been applied to the evaluation of the trustworthiness of sensory data. A generic framework has been proposed in [18] to integrate different areas of data quality and identify issues of assessing quality dimensions. A general matrix has been proposed for evaluating and comparing between different quality assessment tools.

In addition, to providing a comprehensive quality assessment the authors in [19] proposed a data repairing recommendation scheme to cope with data quality degradation. While in [20], authors proposed a framework to evaluate and manage social media data quality across the Big Data pipeline, the work of [21] proposed a quality evaluation model using five quality maturity levels across the Big Data lifecycle.

To address challenges related to Big Data quality management throughout the Big Data value chain, we propose in the next section a hybrid quality assessment model that incorporates and integrates both data-driven and process-driven quality assessment. Such processes include: pre-processing, processing, and analytics. The proposed model assesses Big data quality while involving quality management processes that deals manage quality of data since its inception till its analytics while enforcing quality for all Big Data activities.

III. Big Data value chain quality requirements

In this section, we describe Big Data quality specification, including description of quality dimensions, attributes, and the corresponding metrics. We also, introduce the quality of process and its metrics. In addition, we explain the quality of metadata, which is considered as an essential element for quality assessment of Big Data.

A. Big Data quality specification

1) Big Data dimensions and metrics

Data quality dimensions play important role in data quality assessment. There are multiple definitions of data quality dimensions in the literature, however, they are commonly classified into two categories: contextual and intrinsic [22]. Contextual dimensions are related to the data values while intrinsic deals with the intension of the data, which is the data schema [23] [24]. Standard quality dimensions discussed in the literature involve: timeliness, accuracy, completeness, and consistency [23]. The following are the agreed upon definitions of four well-reputed quality dimensions accepted in the literature:

- ***Timeliness:*** it is also referred to by currency and volatility. It is usually related to the age of the data and the degree of its validity in the system or to the real world. In other words, it describes how much the data is up-to-date (currency dimension). On the other hand, the frequency of the data value change occurrence defines the volatility dimension.
- ***Accuracy:*** measures how much the recorded data is correct and resembles real world values and hence is reliable.
- ***Completeness:*** mainly related to the existence of missing or null values.
- ***Consistency:*** means that the structure and semantics of data follows a set of rules and constrains [23].

Each quality dimension is characterized by one or more than one quality metrics as shown in the table below where some of quality dimensions are adapted from [23] such as timeliness, currency and consistency quality related metrics. While others are newly introduced or altered in order to integrate the quality framework proposed in Figure 1 such as volatility (VMb) and completeness metrics (CMPMc). In this work, we use the four quality dimensions defined above, as they are very relevant for the application domain we have considered in this paper, which is health monitoring. Data acquisition from sensors is very sensitive to such quality dimensions including precision, accuracy and timeliness. For example any timely EEG episode may reveal crucial information to disease monitoring, inaccurate collected data may lead to wrong diagnosis, thus may engenders incorrect clinical decision.

2) *Quality Metadata*

Metadata describes relevant information about the data such as provenance, quality, and other details. In other words Metadata is the "data about data" [20] that makes it easier and faster to process and extract data features. It is defined as "Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" [25]. Usually, it includes extra information about the quality to help evaluating the data [23]. These quality attributes of the data are referred to as the quality metadata [20]. Accuracy, timeliness, consistency are all attributes related to the data quality that comprises the quality metadata. There are many initiatives that studied metadata, in [20], authors provided a comprehensive classification of multiple quality metrics along with their description, purpose, target, evaluation technique, value range, constraints, and applicability.

Table 1. Data Quality dimensions and metrics

	Formula	**Description**
Timeliness metrics:		
TMa	= 1 - CMa / VMb	1 - Currency/Volatility
TMb	=numOfProcessedRecs/totalRecs/timePeriod	Percentage of the completed processed records within a time limit
Currency metrics:		
CMa	= currentTime – updateTime	Time of update
CMb	= updateTime - storageTime	Difference between time of update and time of storage
Volatility metrics:		
VMa	= ConstantTimePeriodValue	Time length for which data remains valid
VMb	= (storageTime - updateTime) / totalTime	Volatility: (time of data – time of update)/total time
Accuracy metrics:		
Ama	= numOfCorrectValues / totalValues	The ratio between the number of correct values stored and the total number of values.
Amb	= AvgUsrResponse	User questionnaire
Completeness metrics:		
CMPMa	= numOfEmptyValues / totalValues	The ratio of the number of empty of null values over the total number of values.
CMPMb	= AvgUsrResponse	User questionnaire
CMPMc	= actualTotalSize/ expectedTotalSize	The total size of the stored records over the expected size of the data
Consistency metrics:		
CNSMa	= numOfInconsistentValues/totalValues	The ratio of the total number of inconsistent values over the total number of values
CNSMb	= numOfViolations	The total number of values violating constraints and rules

a) Metadata in the Big Data value chain

Creating Metadata needs a domain expert knowledge to define quality policies and rules [20]. During data extraction stage, quality policies define the acceptable quality attributes and metrics of the data related to given quality dimensions. These quality attributes are evaluated and the results of evaluation are placed in the specified quality Metadata. The extracted data is saved in designated data storage while the corresponding Metadata can be stored in a different storage space (External) or together with the data (Internal) [26]. Another classification can describe Metadata as static or dynamic. The static Metadata is fixed and has to do with information that doesn't change about the data. On the other hand, the dynamic Metadata is continuously changing during runtime. [27].

Metadata description is represented using vocabularies that follow well-defined standards and models. The syntax of Metadata is defined as the set of rules that govern the structure of its basic elements [28]. Each metadata scheme can be represented using many markup or programming languages having different syntax notations. One of commonly used standard is the Dublin Core, which can be written using HTML, XML, and RDF [29]. Other domain related Metadata languages were proposed in literature like Ecological Metadata Language (EML) or the Federal Geographic Data Committee Biological Data Profile (FGDC BDP) which are languages that give formal description to information that describes ecological data [30]. A more general Metadata model is Open Information Model (OIM), which is a specialization of the UML related to a particular domain based on the UML, XML and SQL [31]. Also, JSON (Java Script Object Notation) [32] is a metadata standard used to represent massive data into a format based property graph models. It is a lightweight standard for Big data-interchange format.

3) *Quality of Service (Process)*

In order to evaluate the framework proposed in Figure 1, we need to evaluate the quality of the data and also the quality of service (process) of data handling at each stage. We therefore, describe the quality metrics related to data processing in Big Data value chain including pre-processing, processing and analytics. However, the quality evaluation of the visualization process is out of scope of this paper. The common processing quality dimensions discussed in literature include for instance:

- ***Capacity****:* is the maximum number of concurrent connections and/or processes.
- ***Performance:*** is the speed of data processing.
- ***Response time:*** is the maximum or average time to complete the processing of each record (or the total records)
- ***Latency****:* is the total time to receive the resulted processed data (delay)
- ***Throughput:*** is represented in terms of the number of processed records over a time period.
- ***Accuracy:*** is measured by the number of errors resulted after processing the data.

Additional quality attributes that are not considered in this work are reliability, availability, robustness, and scalability as they are more specific attributes that are related to the quality of the hardware, the infrastructure used [33].

IV. HYBRID BIG DATA QUALITY EVALUATION MODEL

Figure 1 describes a conceptual view of the main processes that constitute the hybrid Big Data quality assessment model. These processes include: data collection, pre and post Big Data quality evaluation, pre-processing quality evaluation, processing and analytics quality evaluation. These processes communicate and integrate seamlessly to achieve a complete quality assessment of the Big Data value chain.

The Big Data value chain incorporates a set of processes including Big Data pre-processing, processing, analytics and visualization. These processes generate data, process data, analyze and visualize data within the complete Big Data pipeline. In this work, we don't consider quality evaluation in the data collection phase as well as on the visualization phase. In the following, we describe each of these processes and how they handle Big Data quality assessment.

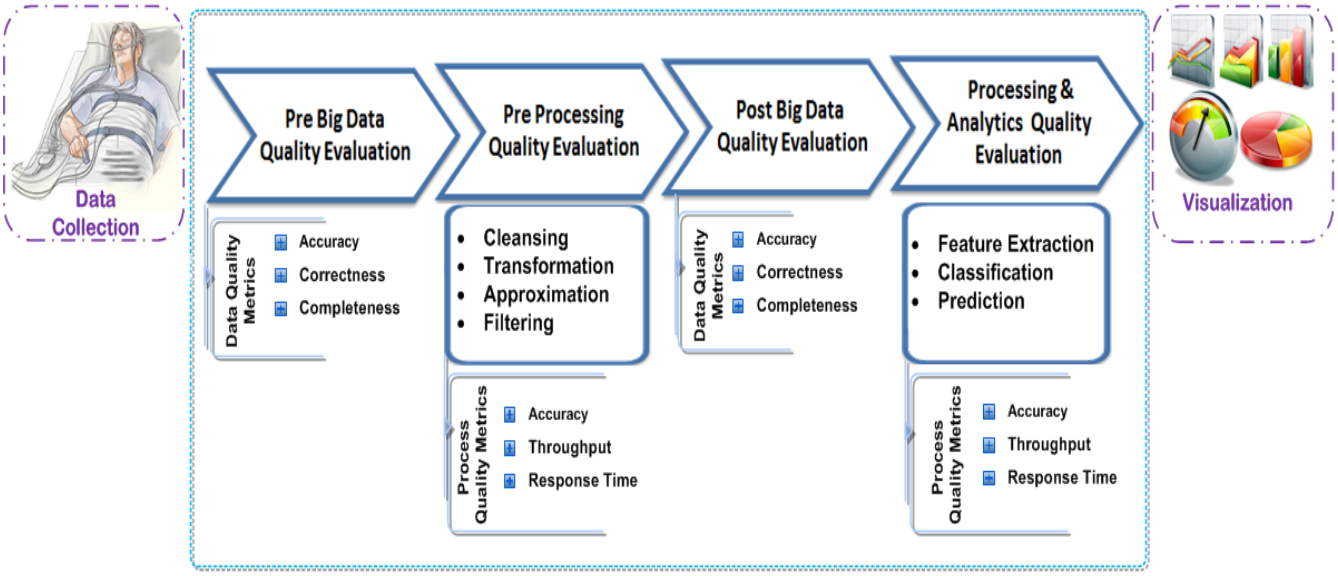

Fig. 1. A Hybrid model for assessing Quality of Big Data value chain

Data Collection: it handles data collection from its source, and relays it to the storage and processing location. Quality evaluation in this phase is important as for the next phases of the Big Data value chain. However, it is highly linked to the design of how data is generated from sources and the devices used, the data sampling technique used, the underplaying network, and the communication protocol used. All these might affect the quality of the data collection, including accuracy, timeless, latency, etc. We do not handle quality of data collection and we consider it for future work although it might influence the quality evaluation of the remaining phases of Big Data pipeline.

Pre & Post Big Data quality evaluation: it is a data-driven quality evaluation and it is conducted before and after pre-processing. It aims to measure the degree to which the quality of data got improved after pre-processing. In the pre-Big Data quality evaluation, we identify for instance the percentage of incomplete data, inconsistent data, and incorrect data in order to decide which pre-processing scheme (e.g. cleansing, transformation, and approximation) should be applied. Many data quality metrics can be measured and considered to be very important to access the overall data quality. Examples of these metrics include data accuracy, correctness, completeness, and consistency.

Pre-processing quality evaluation: it is a process-driven quality evaluation that consists of many activities related to data preparation for the next phase in Big Data chain (processing and analytics). Due of the diversity of sources, the collected datasets may have different levels of quality in terms of noise, redundancy, consistency, etc. Transferring and storing raw data would have necessary costs. On the consuming side, certain data analysis methods and applications might have strict requirements on data quality. As such, data pre-processing techniques that are designed to improve data quality should be used in Big Data systems. A set of quality metrics can be measured to evaluate the quality of pre-processing and include for instance, accuracy, throughput, and response time.

Processing & analytics quality evaluation: it consists of immediate exploitation of data after which a supervised pre-processing is done. Processing may involve the application of data mining methods and machine-learning procedures to lead to a set of target data results. Processing can be centralized, or distributed over a cluster or a data center, as it needs high powerful processing nodes. However, analytics consist of mining large amount of long-term period, heterogeneous data, and data from different sources to extract data knowledge, hidden patterns, unknown correlations, to other useful information to get insights for further decision-making. It is also the process of refining data processing and analytics algorithms by iterating the Big Data value chain processes many times. The analytical outcomes can lead to more effective clinical decisions, faster interventions, improved processes efficiency, and competitive advantages over traditional data analytics techniques. The same pre-processing quality metrics can be measured to evaluate the quality of the processing and analytics phase and include accuracy, throughput, and response time.

Visualization quality evaluation: despite the fact that this process quality is not evaluated in this work and was left for future consideration, the visualization process consists of viewing and validating data resulted from the Big Data value chain in order to support formulating decisions, to report on continuous updates about Big Data status. Data can be presented using different views including a summary of monitoring results, graphs, pattern of readings, and even report on discrepancies of measures, then generate automatic preventive actions. A set of quality metrics can be used to evaluate quality of this last process and are mostly linked to user satisfaction, and quality of data representation.

V. Evaluation

In this section, we describe the evaluation part of selected features exhibited by our hybrid quality evaluation and enforcement model in Big Data value chain described in Figure 1. Both quality of data and process have been evaluated for which two quality dimensions were considered and set of metrics were measured. This quality evaluation is studied for two Big Data processes in the value chain: Big Data pre-processing, processing and analytics.

A. Setup

The experiments have been executed on a cluster node having 14 GB of memory, equipped with an Intel i7 quad core of 2.66 GHz, running a virtual machine and the following software and tools:

- Oracle VM VirtualBox 5.0.16: it is a powerful x86 and AMD64/Intel64 virtualization product from Oracle.
- Talend: data preparation tool and Talend Open Studio for data preparation and Data Quality metrics measurements.
- Trifacta Wrangler -version 3.0.1-client1+push2: for data preparation, preprocessing and Data Quality metrics measurements.
- ML Vagrant: a 64 bits virtual machine for Machine Learning. The VM includes:
  - Apache SPARK 1.6.0
  - Python 2.7.5
  - MapReduce YARN version 2.
  - Scientific Python stack (scipy, numpy, matplotlib, pandas, statmodels, gensim, networkx, scikit-learn) plus IPython 4 + Jupyter notebook.
  - R 3.2.2 (rmarkdown, magrittr, dplyr, tidyr, data.table, ggplot2)
  - Spark notebook Kernels for Scala and R
  - Two spark external libraries: Kafka Spark Streaming, Spark CSV library.

B. Dataset description

We used the Sleep Heart Health Study (SHHS) dataset [34], which have been collected to determine the cardiovascular risks and other consequences of sleep-disordered breathing. The SHHS dataset represents data from the baseline and first follow-up visits, collected on 6441 patients between 1995 and 1998. Data consists of EEG, EOG, EMG, thoracic and abdominal excursions, nasal airflow, oxygen saturation, ECG, heart rate, body position, ambient light. These data were sampled at different sampling rates ranging from 1 Hz to 250 Hz. Each participant sleep study is represented in EDF format, which comes around 40 MB. Combined data of different patients is represented in CSV format. Collected data including polysomnograms were obtained in a homely atmosphere from the participants.

We used Talend data preparation tool to import the CSV files through its interface, the file is parsed and a couple suggestions are shown for data cleansing including tips to replace missing data as well as enrich it. Talend Open Studio for Data Quality offers advanced data profiling to improve the quality and integrity of data. It is a visual tool with graphical charts that also measures some data quality metrics.

C. Scenarios

In the following, we developed two scenarios: one is data-driven which evaluates the quality of data before and after pre-processing using a set of data related metrics, the second scenario is process-driven which evaluates the quality of processing and analytics along with its generated data quality.

1) Scenario 1: to evaluate the data quality before and after preprocessing we choose two Big Data quality dimensions, which are completeness and consistency for simple illustration. Specifically, we used the following metrics: CMPMa for the Completeness dimension and CNSMa for the Consistency dimension (see Table 1 for more description details). First, we evaluate the completeness metrics as follows:

$$\mathrm{CMPMa} = numOfEmptyValues \;/\; totalValues$$

According to the above equation we evaluated the percentage of the CMPMa for each attribute of our dataset. The average CMPMa for all attributes was 9%. Our pre-processing method fills the empty cells with valid values according to the attribute specifications. In other words, for BP, the empty values are filled with a value within the normal range. Other empty values of attributes are filled with the mean value. Empty values in other attributes are filled with zeros. Completing the empty values relies on the domain knowledge expertise and the validation of clinicians. After preprocessing we evaluate the completeness metrics to be 0% meaning no empty values exist in our preprocessed dataset. Figure 2 shows the improvement percentage of the completeness quality metric after preprocessing for each attribute named $T_1, T_2, \ldots, T_N$. In this graph, only the attributes that included empty values are represented. The degree of completeness ranges from 0.01% incomplete data to 60% incomplete data.

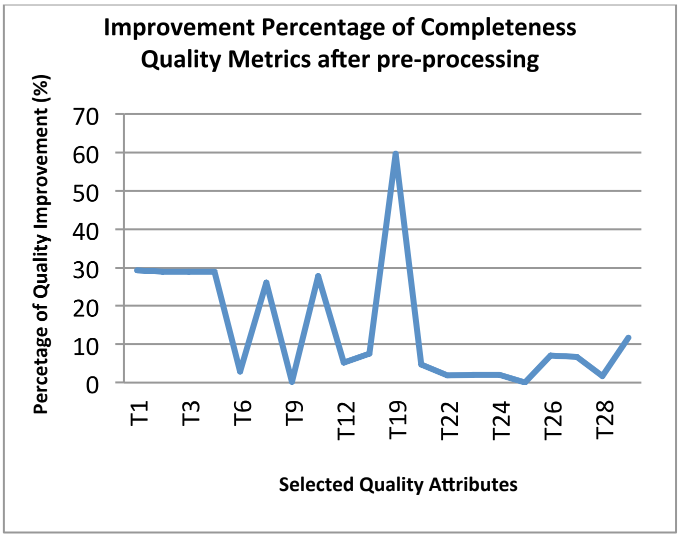

Fig. 2. Big Data Completeness evaluation after pre-processing

Second, we evaluate the Consistency metrics as follows:

$$CNSMa \ = numOfInconsistentValues \, / \, totalValues$$

According to the above equation we evaluated the percentage of the CNSMa for each attribute in our dataset. The inconsistency here means that the attribute values mismatch the type of the attribute. Our preprocessing method clears the inconsistency of decimal to integer by rounding the value to the nearest integer. Figure 3, shows the improvement percentage of consistency metric after pre-processing, which reached up to 100% for one attribute and the lowest value of 90%. In this graph only the attributes with inconsistent values are represented.

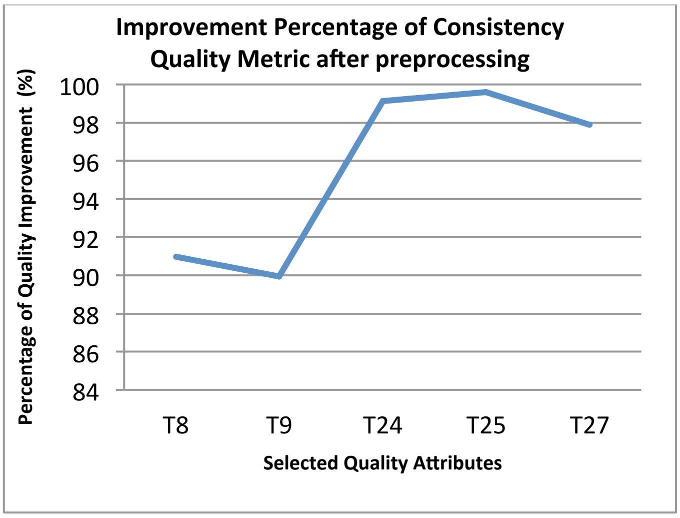

Fig. 3. Big Data Consistency evaluation after pre-processing

In this experiment, we can also refer to another quality dimension, which is accuracy. The accuracy of data is linked to the completeness metric because in order to reach 100% completeness, we need to fill the empty cells with calculated values that are not the actual real values. This will result in decreasing the accuracy quality dimension level to some extent. However, the selected data is carefully thought of and are set either to the mean value or to a normal value that falls within the normal range like in SYSTBP (normal BP) attribute that is filled with the value 100 falling within the normal range. Hence, in our experiments, the accuracy level degradation is not significant because of the aforementioned reason.

In the next scenario we will evaluate how the results of quality evaluation conducted in this process will affect positively the quality of the subsequent process in the Big Data value chain.

2) Scenario 2: in this secenario we evaluate the quality of processing and analytics in order to assess that the data used as input ensures quality of these two Big Data processes. The main activities conducted in this process is building the classification model on the supposed high quality data as input, then train the model using different classification algorithms. Two metrics were meausred; classification and prediction accuracy, and the processing time. We used Spark's machine learning library MLlib to build and train classification models. The aim was to predict cardiovascular death from sleep heart data with a given set of attributes. Classifications used were Naïve Bayes, Logistic Regression With LBFGS, Decision Tree, SVM, Random Forest and Gradient Boosted Trees. Time taken to build the model and train was recorded and the accuracy of the prediction was also recorded.

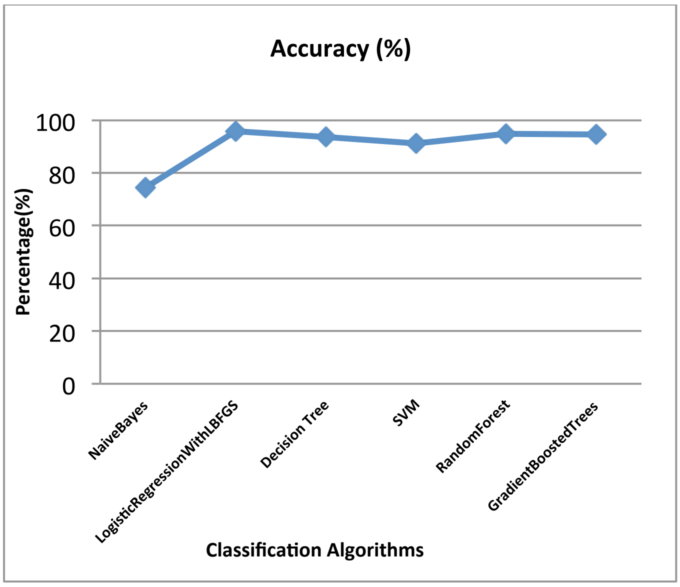

Fig. 4. Classification accuracy evaluation after processing and analytics

The results reported in Figure 4 showed an above 90% prediction accuracy for all the classification algorithms we have experimented except for Naïve Bayes, which reported a prediction accuracy of 75 %, this can be easily explained by the fact that classification is not affected by dependencies among different attributes [35].

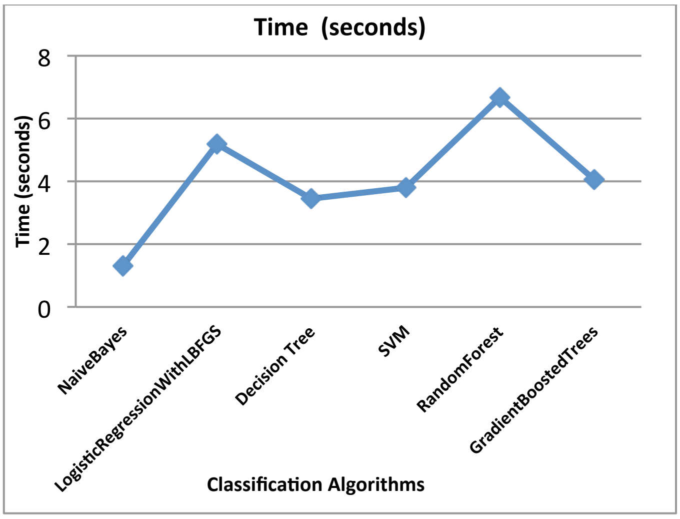

Fig. 5. Processing time evaluation after processing and analytics

Figure 5 reported the processing time we calculated for building the classification model and conduct training. The processing time remain low and in the range between 1 second and 6.5 seconds, this time is considered optimum given the Big Data size and its large number of attributes which reached 1500 attributes. Spark VM played a great role here to maintain a very low processing time and thus ensure a high quality of processing and analytics of Big Data.

In summary, based on the quality results we obtained from the evaluation of both quality of pre-processing data, which we conducted in the first scenario and the quality of process, which we conducted in the second scenario, we conclude the following:

- The earlier we address the quality of Big Data the more we enforce quality in the remaining phases of the Big Data value chain.
- Quality evaluation is a continuous process that involves both data-driven and process-driven quality evaluation.
- A loop back process needs to be implemented to allow any adjustment and/or re-evaluation of the quality metrics and processes.
- Quality should be addressed across the value chain at each single activity, data, and phase.

VI. Conclusion

Big Data quality evaluation has become an urgent concern for researchers in both academia and industry. There are very limited initiatives so far that tackled this important aspect in Big Data. Therefore, in this paper, we addressed data quality assessment in the Big Data value chain. We offered a hybrid Big Data quality assessment model that combined both data-driven and process-driven quality evaluation. The model evaluated quality of pre and post data pre-processing and used the resulted data to evaluate quality of successive processes: processing and analytics. The results of experiments we have conducted on a large health dataset proved that quality got improved and maintained across the Big Data chain when quality of data and quality of process are both implemented.

As future work we are planning to evaluate if the quality evaluation completed in earlier stages of Big Data pipeline will improve considerably the quality at final stages such as visualization of Big Data.

Acknowledgment

This work has been supported by UPAR grant # 31T064.

References

[1] H. Hu, Y. Wen, T.-S. Chua and X. Li, "Toward Scalable Systems for Big Data Analytics:A Technology Tutorial," *IEEE Access,* vol. 2, pp. 652-687, 2014.

[2] 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 *International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012.

[3] P. Glowalla, P. Balazy, D. Basten and A. Sunyae, "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.

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

[5] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng Bul,* vol. 23, no. 4, p. 3–13, 2000.

[6] Y. L. Simmhan, B. Plale and D. Gannon, "A Survey of Data Provenance in e-Science," *SIGMOD Rec,* vol. 34, no. 3, pp. 31-36, 2005.

[7] B. Glavic, "Big Data Provenance: Challenges and Implications for Benchmarking," *Specifying Big Data Benchmarks,* p. 72–80, 2014.

[8] Y.-W. Cheah, R. Canon, B. Plale and L. Ramakrishnan, "Milieu: Lightweight and Configurable Big Data Provenance for Science," in *2013 IEEE International Congress on Big Data (BigData Congress)*, 2013.

[9] A. G. Recuero, S. Esteves and L. Veiga, "Towards quality-of-service driven consistency for Big Data management," *International Journal of Big Data Intelligence,* vol. 1, no. 1/2, p. 74, 2014.

[10] A. Ebaid, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, J.-A. Quiane-Ruiz, N. Tang and S. Yin, "NADEEF: A generalized data cleaning system," *Proceedings of the VLDB Endowment,* vol. 6, no. 12, p. 1218–1221, 2013.

[11] N. Tang, "Big Data Cleaning," *Web Technologies and Applications,* p. 13–24, 2014.

[12] L. Ramaswamy, V. Lawson and S. V. Gogineni, "Towards a Quality-centric Big Data Architecture for Federated Sensor Services," in *2013 IEEE International Congress on Big Data (BigData Congress)*, 2013.

[13] D. Rao, V. N. Gudivada and V. V. Raghavan, "Data quality issues in big data," in *IEEE International Conference on Big Data (Big Data)*, 2015.

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

[15] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal ,* vol. 14, p. 2, 2015.

[16] L. L. Pipino, Y. W. Lee and R. Y. Wang, "Data Quality Assessment," *Communications of the ACM,* vol. 45, no. 4, pp. 211-218, 2002.

[17] M. Monga and S. Sicari, "Assessing Data Quality by a Cross-Layer Approach," in *IEEE International Conference on Ultra Modern Telecommunications & Workshops (ICUMT'09)*, 2009.

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

[19] X. Ding, H. Wang, D. Zhang, J. Li and H. Gao, "A Fair Data Market System with Data Quality Evaluation and Repairing Recommendation," *Web Technologies and Applications,* pp. 855-858, 2015.

[20] 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.

[21] I. Caballero and M. Piattini, "CALDEA: a data quality model based on maturity levels," in *IEEE Proceedings. Third International Conference on Quality Software*, 2003.

[22] B. T. Hazen, C. A. Boone, J. D. Ezell and L. A. Jones-Farmer, "Data quality for data science, predictive analytics,and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics,* vol. 154, pp. 72-80., 2014.

[23] C. Batini, C. Cappiello, C. Francalanc and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM computing surveys (CSUR),* vol. 41, no. 3, p. 16, 2009.

[24] I. Taleb, R. Dssouli and M. A. Serhani, "Big Data Pre-Processing: A Quality Framework," in *IEEE International Congress on Big Data*, 2015.

[25] "Understanding Metadata," NISO Press, Bethesda, MD, USA, 2004.

[26] D. O'Neill, "ID3.org," [Online]. Available: http://id3.org/. [Accessed Feb 2016].

[27] T. N. Huynh, O. Mangisengi and A. Min Tjoa, "Metadata for object-relational data warehouse," *DMDW,* p. 3, 2000.

[28] W. Cathro, "Metadata: an overview," National Library of Australia Staff Papers, 2009.

[29] DCMI, "Semantic Recommendation," 2009. [Online]. Available: http://dublincore.org/specifications/. [Accessed Feb 2016].

[30] E. H. Fegraus, S. Andelman, M. B. Jones and M. Schildhauer, "Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation," *Bulletin of the Ecological Society of America 86,* vol. 3, pp. 158-168, 2005.

[31] T. Vetterli, A. Vaduva and M. Staudt, "Metadata Standards for Data Warehousing:Open Information Model vs. Common Warehouse Metamodel," *ACM Sigmod Record 29,* vol. 29, no. 3, pp. 68-75, 2000.

[32] "Introducing JSON," [Online]. Available: http://www.json.org/. [Accessed 10 May 2016].

[33] S. Ran, "A model for web services discovery with QoS," *ACM Sigecom exchanges,* vol. 4, no. 1, pp. 1-10, 2003.

[34] S. Redline and et al., "Sleep Heart Health Study," National Sleep Research Resource, [Online]. Available: http://sleepdata.org/datasets/shhs. [Accessed 2016].

[35] R. Kohavi, "Kohavi, R. (1996, August). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," *KDD,* vol. 96, pp. 202-207, 1996.