Big Data Pre-Processing: A Quality Framework

Ikbal Taleb, Rachida Dssouli
CIISE, Concordia University
Montreal, Canada
e-mail: i_taleb@encs.concordia.ca,
rachida.dssouli@concordia.ca

Mohamed Adel Serhani
College of Information Technology
UAE University, Al-Ain, UAE
e-mail: serhanim@uaeu.ac.ae

***Abstract*— With the abundance of raw data generated from various sources, Big Data has become a preeminent approach in acquiring, processing, and analyzing large amounts of heterogeneous data to derive valuable evidences. The size, speed, and formats in which data is generated and processed affect the overall quality of information. Therefore, Quality of Big Data (QBD) has become an important factor to ensure that the quality of data is maintained at all Big data processing phases. This paper addresses the QBD at the pre-processing phase, which includes sub-processes like cleansing, integration, filtering, and normalization. We propose a QBD model incorporating processes to support Data quality profile selection and adaptation. In addition, it tracks and registers on a data provenance repository the effect of every data transformation happened in the pre-processing phase. We evaluate the data quality selection module using large EEG dataset. The obtained results illustrate the importance of addressing QBD at an early phase of Big Data processing lifecycle since it significantly save on costs and perform accurate data analysis.**

Keywords - Big Data, Data Quality, pre-processing.

I. INTRODUCTION

Big data is everywhere [1]. From analyzing larger volumes of data than was previously possible, to analyzing data in motion, whether the industry of concern is telecommunications, health-care or utilities, technologies for big data are needed. One of the most important expectation from big data analytics are the ability to reveal patterns, trends and associations, especially those affecting individuals and businesses with the goal of driving meaningful decisions. Extracting this information from extremely large data sets is not a trivial matter, a careful planning and dimensioning of Big Data systems is crucial in order to provide timely and meaningful inputs to decision layers. When it comes to adopting Big Data solutions, businesses and institutions are in dire need of new technologies and architectures optimized for the extremely large data sets.

The foundational characteristics of Big Data are often described as: Volume, Variety, Velocity and Veracity and commonly known as "4Vs" definition of Big Data [1]–[5]. In Big data systems [5], data is the ultimate source of knowledge. In its lifecycle, data travels through four different phases as shows in Figure 1: data generation, data acquisition, data storage, and data analytics. The data generation phase is where data is created, a large number of data sources are responsible for these data: Electrophysiology signals, sensors used to gather climate information, surveillance devices, posts to social media sites, videos and still images, transaction records, stock market indices, cell phone GPS location data to name a few.

The data acquisition phase [1],[5] consists of data collection, data transmission, and data pre-processing. With the exponential growth and availability of heterogeneous data production sources, an unprecedented amount of structured, semi-structured, and unstructured data is available.

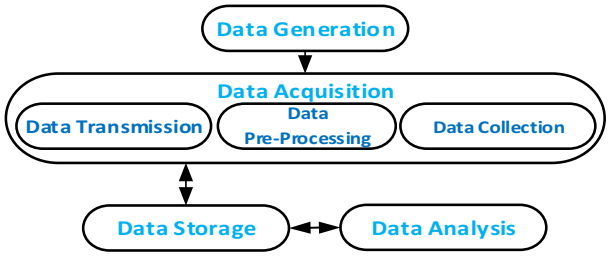

Figure 1. Big data Lifecycle

While it is well-known that, in theory, more data leads to better predictive power and overall insight, this raw data must be channeled through a pre-processing phase in which activities such as data cleansing, de-duplication, compression, filtering, and format conversion take place. This mandatory step is essential in order to refine and valuate the data. Other important pre-processing tasks such as data integration [6], [7] and fusion [8] of multiple heterogeneous sources are also taking place and have a considerable effect on the resulting transformed data and, consequently, the overall results analytics.

In order to keep track of data value and relevance as well as the severity of the impact of the aforementioned pre-processing transformations, a concept of data quality is needed. Moreover the nature of targeted data, such as those arising in social networks and which are characterized by unstructured data with no quality references, suggests that data must be profiled and provided with certain quality dimensions at inception phase. This also means that a quality of data attribute must be assessed, improved and controlled all along its life cycle as it directly impact the results of the analysis phase.

In this paper, we propose a big data pre-processing quality framework, which aims at solving the numerous data quality issues that occur when attempting to apply data quality concepts to large data sets. In Section II, data quality

978-1-4673-7278-7/15 $31.00 © 2015 IEEE
DOI 10.1109/BigDataCongress.2015.35

is introduced and the issues that arise when applied to Big Data are discussed. In Section III, a summary of related works that address quality in Big Data is presented. Section IV, introduces a framework for big data pre-processing quality which constitute the core of the paper and a platform for future works. In Section V, we conduct preliminary experiments to evaluate the quality profile selection key feature on EEG big data. To conclude the paper, Section VI highlights the strengths and potential weaknesses of the proposed architecture and provides insight into future works.

II. BIG DATA AND DATA QUALITY

Before addressing Big data quality, one needs to understand what data quality means in general. Most of the studies in this area are from the database and management communities. According to [11], Data quality is not an easy concept to define. Its definitions are data domain aware. In [12], Sidi *et al.* define data quality in quality management as the appropriateness for use or meeting user needs. In general, there is a consensus that data quality is always dependent on the quality of the source data [13]. Therefore, it is important to associate quality to data at its inception as data travelling through many phase processes will see its quality affected, either positively or negatively. Data quality issues take place when quality requirements are not met on data values [14].

In any data quality scheme, data is subjected to auditing, profiling and quality rules application with the objective of maintaining and/or increasing its quality. Data quality is a well-known concept in the database community and have been an active area of database research for many years [9],[10]. However, a direct application of these quality concepts to Big Data faces severe problems in terms of time and costs of data pre-processing. The problem is exacerbated by the fact that these techniques were developed in the context of well-structured data.

In the context of Big Data, any data quality application must be selected according to the origin, domain, nature, format, and type of data it is applied on. A proper management of these data quality schemes is essential to solve the many problems arising when dealing with large data sets. In addition, quality rules borrowed from the database research domain may not be relevant to new and emerging data types and formats.

A. Data Quality Dimensions

According to [15][12][16], a data quality dimension offers a way to measure and manage data quality. There are several quality dimensions each of which is associated with a specific metric. Data quality dimensions usually fall into two categories: intrinsic and contextual, see for instance [17] [16][18][19]. Figure 2 illustrates the contextual dimensions that are related to the information while the intrinsic dimensions refer to objective and native data attributes. Examples of intrinsic data quality dimensions include:

1. *Accuracy:* measures whether data was recorded correctly and reflect realistic values.
2. *Timeliness:* measures whether data is up to date. Sometimes represented as data currency and volatility [20].
3. *Consistency:* measures whether data agrees with its format and structure. Some works on Big data Quality refer to conditional functional dependencies as data quality rules to detect and capture semantic errors [18][21].
4. *Completeness:* measures whether all relevant data are recorded with no missing entries or missing values.

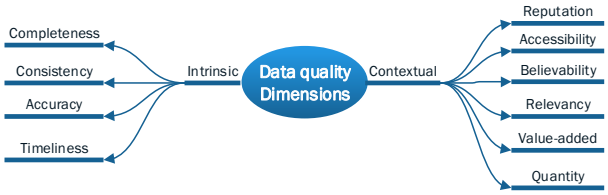

Figure 2. Data Quality Dimensions

B. Data Quality and data provenance

Data is transformed in many ways, many times and to many different formats, which induce reduction of size, re-representation from schema-less to structured, and including the removal of noisy data among other things. Transformations taking as input a set of data and producing a new set of data need to be recorded, saved and queried. Important information such as origin, sources, types of data, whether it is schema-less or structured, date of data production, creation, transformation among others, have to be described and attached to the data at its inception in order to have an efficient and effective data quality model in Big Data [22].

It is essential that low quality data is detected as early as possible in the data generation and acquisition phases in order to avoid wasting network bandwidth and storage space. These very same phases are themselves quality influencing and need to be modeled, managed and controlled.

C. Data quality rules, data cleansing rules

In [14] a data quality rule is defined as "*an externally given directive or a consensual agreement that defines the content and/or structure that constitute high quality data instances and values*". Each data quality rule targets a specific data quality issue. In the following some examples of data quality rules:

- *Duplicate instance rule*: define when a property of many instances represents the same state.
- *Illegal value rule*: specify what values that a property shouldn't have.
- *Functional dependency rule*: specifies legal value combinations for two or more properties that are allowed to occur within the same instance.

Moreover, there are other categories of data rules like cleansing rules, mapping rules, association rules, and Conditional Functional Dependencies (CFD) rules [9], [10], [18]. Each rule specifies what the data value should be. Whitespace removal and value substitution rules are examples of data cleansing rules. In [18], Saha and Srivastava suggest that data rules must be discovered from the data itself.

There are several data techniques and rules to improve data quality in each activity of the pre-processing phase. In this paper, we will consider data cleansing as the main big data pre-processing activity that directly affects data quality. In fact, our focus will be more on process management rather than the actual cleansing activity and techniques used therein.

D. Big data Quality

Data Quality in Big Data needs well-defined lightweight measurement processes that can run in parallel with each phase [23]. These processes include data quality management, monitoring and control whose main objective is to keep tracking any changes that would improve or degrade data quality.

Data in big data storage systems is distributed; this allows the provision of distributed computing and handling large data amounts without storage limits. Moreover, fault tolerance, geo-distribution and duplication of data provide high availability. However, this can result in data quality issues like consistency across many data centers [17].

E. Data Quality and Big data Pre-Processing

Due of the diversity of sources, the collected data sets 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. In the following a brief description of some typical data pre-processing activities:

a) Data integration

Data integration techniques aim to combine data residing in different sources and provide users with a unified view of the data and provide a coherent storage. Data integration has been tackled deeply in traditional database research.

Previously, two approaches prevailed, the data warehouse method ETL (Extract, Transform, and Load) [6] and the data federation method. In general, data integration methods are better intertwined with the streaming processing engines and search engines.

b) Data Enhancements and Enrichment

Define the processes of increasing the value of a pool of information by combining data from multiple sources and collections of data such as data integration and fusion [24],[25].

c) Data transformation

It is data normalization and aggregation by converting data from one format to another. It is considered as a data migration from a system to another.

d) Data reduction

Generate reduced views of the data with no impact on the analytics results. Data compression, clustering and dimension reduction belong to this category.

e) Data discretization

Divide the range of a continuous attribute into intervals when applying data mining algorithms that accept only categorical attributes.

f) Data Cleansing

Data cleansing refers to the process of searching, identifying, and correcting errors. It determines and identifies inaccurate, incomplete, or unreasonable data and then updates, repairs or deletes these data to improve quality [26][21]. In [27] and [28], several methodologies used in data cleansing process are listed: statistical, clustering, pattern-based, parsing, association rules and methods used for outliers identification. Moreover, methods for error detection, data format transformation, integrity constraint enforcement, duplicate elimination, and other statistical methods are used in data auditing. The cleansing process must follows a number of steps that are performed in a meaningful order [21][29][30].

Rahm *et al.* [31] and Mohamed *et al.* [27] describes several data cleansing phases summarized in the following:

1) Data analysis for data quality rules creation and discovery. It uses data auditing, data profiling to define and determine error types. Also descriptive data mining is applied to discover data patterns.
2) Definition of data transformation workflow and mapping rules (lead to the execution of a large number of data transformation and cleaning steps). Some examples of transformation types include: conflict resolution, validation, correction and standardization.
3) Verification of the correctness and effectiveness of a transformation.
4) Inspecting data formats, completeness and documenting error examples types.

Data cleansing is considered vital in keeping data consistent and updated. While it improves accuracy, data cleansing is computing and data intensive.

III. RELATED WORK ON QUALITY OF BIG DATA

Pre-processing data before performing any analytics is primeval. However, several challenges have been experienced at this essential phase of the big data value chain [5]. Data quality is one of them and it need to be highly considered in the context of big data.

In [15],[12] the authors point out, that to improve data quality, there are two strategies (1) data-driven and (2) process-driven. The data-driven strategy deals with the data as it is, using techniques and activities as cleansing to improve its quality. On the other hand, Process-driven attempts to identify the origin sources of poor data quality and redesign the process of the way data is created or recorded.

There have always been data quality problems even before the advent of Big Data. In [11], the author categorizes data quality issues and problems as follows: (1) errors correction, (2) conversion from unstructured to structured data and (3) data integration from multiple sources. In addition to the aforementioned problems, a number of specific big data issues, such as web 2.0 data generated at massive volumes, at an unusual speed with uncommon

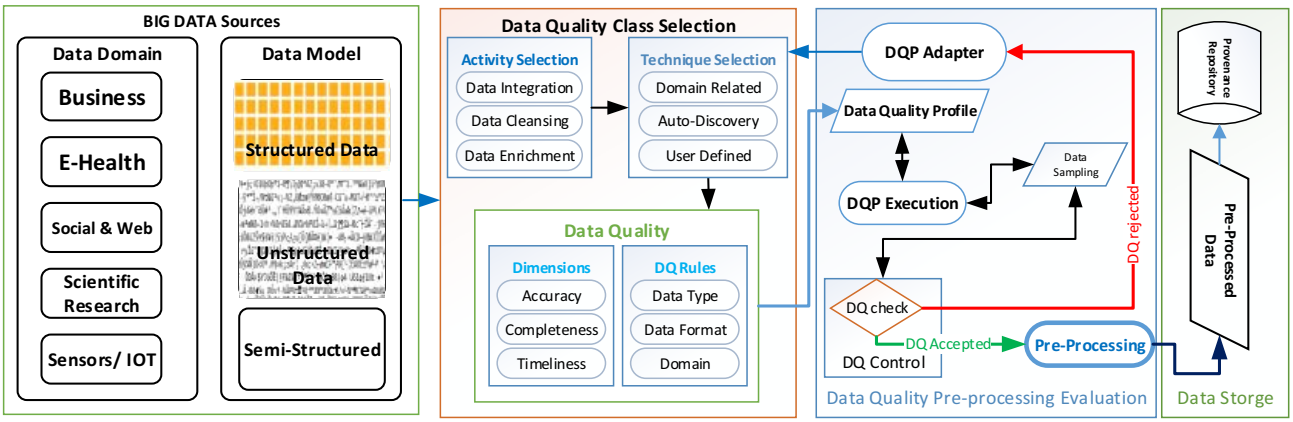

Figure 3. Big data Quality pre-Processing Framework

schema-less structures, are often encountered. As a result of these combined issues, big data cleansing and filtering processes are an important step to be carried out prior to analyzing data which has unknown quality.

As pointed out in [31], data quality problems arise when dealing with multiples data sources. This increases the data cleansing needs significantly. In addition, the large size of data sets that arrive at an uncontrolled speed generates an overhead on the cleansing processes.

Data Provenance, also called origin or lineage, has been studied in the database and distributed systems areas, dealing mostly with the scientific and business data and its provenance quality [22]. Recently an interest has gowned for big data provenance. As data provenance might be helpful when pre-processing big data to assess and improve its quality. The authors of [32] assert that provenance information, is an important source of data relevance and quality. In fact, provenance information can be used to track data from its creation and any transformations it went through. This reinforces our belief that the pre-processing phase in big data is the most important entity for which data quality management must be provided. In [23], Milieu a three-layer lightweight provenance collection framework for scientific data is presented. In [33] a business provenance management system solution is provided. The system was designed for provenance modeling and collection. It stores provenance events and share provenance data across organizations. As such, data quality could be part of this information.

In [34], data semantics to ensure data quality dimension consistency for big data management are proposed. They achieve data replication with consistency under an efficient network bandwidth optimization especially on geo-distributed data. In [29], NADEEF an extensible data cleaning system is proposed. The extension for big data cleaning based on NADEEF is presented in [21] for streaming data. The system deals with data quality from the data cleaning activity using data quality rules and conditional functional dependencies rules. Ramaswamy et al. [35] proposes a big data architecture platform with a pervasive data quality dedicated for sensor services. It uses a descriptive SDQ-ML language for services description and domain application sensor to feed requirements.

To the authors' best knowledge, providing a quality framework for pre-processing that is optimized for big data hasn't yet been fully addressed. Most of prior art addresses different components of data quality for big data, such as in data provenance for big data, big data cleansing using data auto-discovery of quality rules based on conditional functional dependency. In our framework, we use a combined approach, which consists of a data quality management system that deals with data quality rules for all the pre-processing activities prior to data analysis.

IV. System Architecture

The Big Data Pre-processing data Quality (BDPQ) framework is illustrated in Figure 3. It describes the framework's components and their interactions within the big data value chain. It also describes the main processes supporting end-to-end big data quality. The key components of our framework consist of the data quality profile selection, adaptation, and data quality control and monitoring.

Data quality-profiling component (data quality class selection) offers a list of data (quality) rules based on quality requirements, data domains, data types and quality dimensions. This profiler combines and composes a data quality profile that contains a set of rules to achieve a specific data quality requirement. However, data control component is responsible of verifying the quality of pre-processed data.

In the following sections, we introduce the main components of our framework namely data quality class selection, the architectural components together with their roles and the interactions between them.

A. Data quality class selection formulation and notations

Let us denote by ***Req*** *(DI, DO)* a request for pre-processing service. ***Req*** triggers the pre-processing selection phase. It contains the data sets *DI*, *DO* as input and output respectively:

$DI = \{I_1, I_2 \ldots I_i\}$ *Where i*: number of input data sources.
$DO = \{O_1, O_2 \ldots O_j\}$ *Where j:* number of output data destinations.

Here, I_p represents a p^{th} source in a multi-sources data input while O_q represents the q^{th} output in a multiple output data. In general, the numbers of inputs and outputs might be different as they depend on the type of the selected pre-processing activity, e.g. data integration. When i=j, it refers to data cleansing for example.

DQPList is the list of data quality profiles for each data source in the *DI* data set. Each ***DQPList*** consists of a set of Data Quality Profiles *{**DQP** $_1$, **DQP** $_{2...}$ **DQP**$_i$} **for*** each data source input. Formally, each ***DQP**$_p$* is denoted as:

DQP**$_p$*($<I_p, O_{q(p)}>$, ***RL($<r_1, o_1, c_1, a_1, t_1, TQDL_1>...<r_k, o_k, c_k, a_k, t_k, TQDL_k>$))

It contains a reference to the actual data tuple $<I_p, O_{q(p)}>$ where, in the general case *q(p),* is a multiple data sources to one and denote the output data affected, we call this many-to-one operation. ***DQP*** also contains a list of *k* data quality rules list ***RL.***

The s^{th} rule in the ***RL*** list is represented by a tuple $<r_s, o_s, c_s, a_s, t_s>$ where*:*

- r_s*:* represents the actual data quality rule.
- o_s*:* represents the rule priority and execution order.
- c_s*:* represents the category of the rule, and the targeted data (e.g. data type correction rules, and data attribute).
- a_s*:* represents the activity to which the rule is related.
- t_s*:* represents the selected technique.
- $TQDL_s$*:* represents a list of the targeted quality dimensions in the s^{th} rule.

A Request ***Req*** is updated with a ***DQPList*** (Data Quality Profile List). Typically at the end of the selection process a data quality profile is selected and the request ***Req*** is updated according to: ***Req*** *(DI, DO, DQPList)*

B. Architecture components

a) Pre-Processing Activity Selection

The entry point in the proposed framework is the selection of a Pre-Processing activity, e.g. data cleansing activity. This will trigger in return many automatic events. For data cleansing there are steps to be followed to start the cleansing process such as data auditing, and profiling. These steps are already included in our framework as data quality rules added to the request ***Req*** with a specific execution order.

b) Techniques Selection

After a pre-Processing Activity selection is performed a selection of which technique to be used is made under each activity. Techniques are grouped by the following:

1) *Auto- discovery:* uses heuristics to discover the best suitable techniques and rules to be applied on the input data. It consists on parsing and analyzing data samples chosen randomly from the large data set. The outcomes are auto-selected rules. This technique might be used in the case of new data types or an existing unsuitable activity, i.e. activities which do not fit user choices. It will create a set of rules to pre-process data under what is being discovered [10][18]. While the auto discovery is an automatic process, the user interaction is essential in confirming and tuning the discovered rules.
2) *Data domain:* Predefined rules specifically designed to address data quality problems occurring frequently in a specific domain. Domains may refer to data types, origins or fields of exploitation, e.g. Healthcare, Utilities, Business, Social networking, Seasonal and/or event related Data. In this case, data domain drives the technique selection.
3) *User Defined:* Users are given a set of parameters to define the rules that fulfill their requirements of data quality. Typically, user defined activity techniques are tailored to user data. The selection is based on user specification of data errors, anomalies and data irregularities not common to a domain or a data type.

c) Data quality selection

The data quality selection component is responsible for filtering and/or combining data quality rules sets selected in the previous steps. Skipping this step means that all the rules related to an activity are applied on the input data. For the sake of data quality management and cost optimization, a need of targeting specific data quality dimension(s) or rules that deals with specific characteristics of the data being pre-processed (data type, format, and domain).

Each data quality dimension is improved by targeting one or more data anomalies (e.g. bias data). If we are dealing with accuracy and it is selected only rules that deals with constraints to improve accuracy are kept in the request ***Req***. in the other side if the rules are correcting a specific data type to increase the consistency, accuracy or/and completeness are kept in the rules list.

d) Data quality Profile optimization

A data quality profile is generated with an execution flow for the selected activity. This is represented by a ***DQPList*** in the request ***Req.*** The ***DQPList*** contains the data quality profile for each data sources. The DQP selection for the pre-processing activity is based on many quality dimensions targets. The DQP is a composition of several quality profiles for multiple data quality dimensions improvement, related to many data anomalies. The running operations on these combinations are following the order of execution set by o_s in each tuple in the rules list ***RL***.

The data source tuple $<I_p, O_{q(p)}>$ is strongly linked to the multiplicity of sources and outputs and depends on the nature of the pre-processing activity. This is the case of data integration and fusion, which is considered as many-to-one operation. When a specific execution order is reached and established in a ***DQP**$_p$*, an integration process may begin and operations are done from many-to-one. Combined with the priority order, this could lead to an optimization problem where dealing with such cardinalities and parallel executions (certain processes might reside in wait states until others are executed or when certain processes require the outputs of concurrent processes). In pure cleansing pre-processing activity, i.e. one-to-one data flow $q(p)=p$ and such issues are easily dealt with. In the remainder of this paper, we will assume a strict respect of order of execution of rules to fulfill the request ***Req***.

When ***Req*** is ready for execution ***Req*** *(DI, DO, DQPList)*, it is essential, as part of the optimization process, to check for consistency of the selected rules prior to it's processing. In fact, due to the semi-automatic nature of the selection process, the same rule list may contain irrelevant and/or redundant intermediate executions.

The optimization process involves a set of modules and requires a couple of steps. The DQP analysis involves parsing the XML based profile and analyzes the DQ rules. The analysis consists of removing rule duplication and detecting anomalies based on assumptions included during DQ rule creation. This will lead to an optimized profile.

e) Data quality profile Execution

After ***Req*** data quality profiles are validated. The processing starts. In order to efficiently use the available resources, only a samples of data are used. This will allow testing the selected data quality rules even if the accuracy of the results is still uncertain. The samples might be chosen from different locations in the data set and using different sampling sizes or randomly generated. Moreover, the user interaction with the framework is a benefits and add value by directing the choice of samples. This process is done automatically and repeated until the resulted pre-processed data quality is accepted. The evaluation is done on the resulted pre-processed data.

f) Quality Control

At this point of the process, a data quality evaluation is done on the resulting sets of cleaned data samples. Iterations on several samples picked from different location from the data are advised. In fact, trying to cover the maximum possible data is dependent on how data patterns are distributed among the data set. Each iteration means a complete set of samples different from the others.

Once the data quality is accepted the full pre-processing can be applied on the whole data. On the other hand, if data quality didn't reach an acceptance level, after a sufficient number of iterations. The process is not started and the ***Req*** *(DI, DO, DQPList)* is fed back to the data quality profile adapter for revision and update. In addition, to ***Req*** a detailed report of the failed rules and their targeted quality dimensions is composed, namely:

For each, $\boldsymbol{DQP_p}$ $(<I_p, O_{q(p)}>, \boldsymbol{RL}\ (<r_1, o_1, c_1, a_1, t_1, TQDL_1>\ldots <r_k, o_k, c_k, a_k, t_k, TQDL_k>))$, fail or success information is added to each tuple of the rules listed under ***RL***. The new updated data quality profile with the error reports $\{\boldsymbol{e_1}, \boldsymbol{e_2}\ldots \boldsymbol{e_k}\}$ is illustrated as:

$\boldsymbol{DQP_p}\ (<I_p, O_{q(p)}>, \boldsymbol{RL}\ (<r_1, o_1, c_1, a_1, t_1, TQDL_1, \boldsymbol{e_1}>\ldots <r_k, o_k, c_k, a_k, t_k, TQDL_k, \boldsymbol{e_k}>))$

Each error report can be as simple as a binary value indicating success/fail or extended, including reasons of failure, e.g. expected output quality metrics and how they compare to their corresponding acceptance level. In general, they will dependent on the selected quality rule.

The control process start with execution of the DQP. This is done to ensure that all the process is controlled and to add more information's to the DQP which will be reused later.

g) Data Quality Profile Adapter

This component reassesses and updates the request whenever the pre-processed data did not reach the expected quality that the DQP was selected for. If it was originally user defined, the user is notified and a report about the failed rules is proposed with suggestions on quality profile rules for better results. Extended failure reports are therefore instrumental in guiding both user as well as automatic adaptation.

This process will update the $\boldsymbol{DQP_i}$ by removing irrelevant rules and reinsert the ***Req*** request into the activity techniques selection.

V. Evaluation and Discussion: A Case Study

In this section, we evaluate through a case study the most important module of our big data pre-processing framework: the data quality class selection module encompassing activity selection, technique selection, and data quality sub-processes. We consider EEG data set recorded from continuous monitoring of a number of patients. EEG data can be characterized as big data as it is collected from multiple sources (channels), continuous, for different patients, and is of big size. Quality of EEG preprocessing is of prime importance as it a discriminating factor in detecting accurately epileptic seizures.

A. Dataset

The data set we have used is from the CHB-MIT dataset [36]. This dataset consists of EEG recordings from pediatric subjects with intractable seizures collected from 22 subjects. Most files contain 24 EEG signals. Subjects were monitored for up to several days following withdrawal of anti-seizure medication in order to characterize their seizures and assess their candidacy for surgical intervention. The EEG dataset is considered as big data since it fulfills most of the Big data characteristics such as volume, velocity, and variety. In terms of volume for example, a data generated from 12 hours continuous monitoring episode of one patient can exceed 2 GB. This data is very representative, as it needs to be pre-processed to remove infiltrated noises, artifacts generated during signal acquisition. Therefore the quality of pre-processed EEG data is very important for seizure detection and analysis.

B. Data quality selection implementation: case of EEG monitoring

Figure 4 describes the implementation view of the data profile selection; it gets as input the raw EEG data and generates an EEG data quality profile in form of an XML file. This file represents all the information about the input data, the output data and the pre-processing rules to be applied. In the experiment below, we used a set of rules that removes EEG artifacts by applying a band pass filter and then a notch filter to improve the accuracy of pre-processed EEG signal. This is considered as a data cleansing activity.

We used Hadoop MapReduce to run our pre-processing algorithms on EEG data. We implemented digital filters in order to remove signal noises, artifacts such as eyes blinking artifact, and other signals interferences (e.g. power line). Such algorithms are listed under data cleansing activity of

health domain. The following are the steps of our data quality module implementation:

1. Data Input: Raw EEG Signals.
2. Data Output: EEG pre-processed Signals.
3. The selected DQP is generated as an XML file containing all the information related to the pre-processing activity; the cleansing EEG algorithms, and the targeted data quality.
4. The XML file is sent to the DQP execution component. Then algorithms are distributed across different Hadoop nodes.
5. The DQ controller components check whether the results meet the quality requirements. We use MATLAB code for detecting seizures on only some sample of data.
6. The results we have obtained meet the quality expectation and the quality of analysis matches the quality of pre-processed data.
7. EEG pre-processing will run now on a large dataset given the high quality resulted on small EEG data samples.

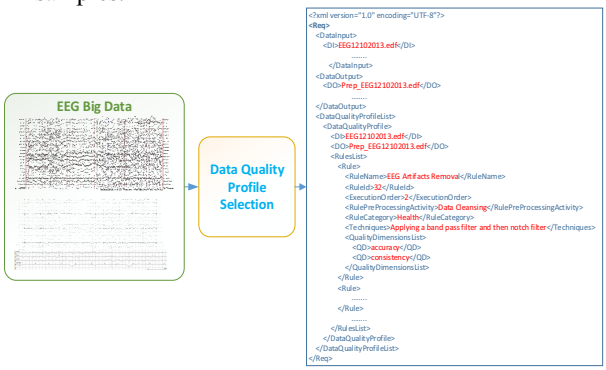

Figure 4. EEG Data quality Profile Selection

C. Discussion and future work

The implementations we have completed so far include one module of the framework, which is the data quality selection module. We are currently implementing the other modules of the framework, for instance, data control module, big data provenance module, and data quality profile adapter. We are eventually planning to extend the evaluation by implementing additional Hadoop MapReduce based algorithms for pre-processing using a large EEG dataset. This will definitely impact positively the quality of pre-processed data. Further implementations will tackle the data pre-processing evaluation and dynamic assessment of data quality rules. Quality adaptation will be possible through the DQP adapter taking the appropriate actions to adjust and reevaluate the quality profile. Yet, big data sampling is very important to maximize data quality, and minimize the processing time.

VI. CONCLUSION

Most of previous work addresses different components of data quality for big data, such as in data provenance for big data, big data cleansing using data auto-discovery of quality rules based on conditional functional dependency. In this paper, we proposed a big data pre-processing quality framework that aimed at solving the numerous data quality concerns that occur when attempting to apply data quality concepts to large data sets. In our framework, we used a combined approach, which consists of a data quality management system. It is used for data quality profile generation that deals with data quality rules selection. These rules are applied as pre-processing activities prior to data analysis. We evaluated the data quality profile selection for the pre-processing of an EEG dataset. We demonstrated that using our data quality selection framework to manage data quality in an earlier stage will not only provide high quality data but also helps gaining time and resources.

The generation of the DQP on data samples rather than the whole data sets provides faster data quality evaluation and an immediate update when new quality rules are inserted, or irrelevant ones deleted. In an ongoing work we are investigating big data sampling techniques to obtain an accurate DQP on large big dataset. Currently, the selection component lacks quality rules diversity. The paramount importance of having a DQP repository for data quality rules with mechanisms to populate, update, query and rate automatically these rules will be a value-added feature.

VII. REFERENCES

[1] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[2] C. L. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci.*, vol. 275, pp. 314–347, 2014.

[3] J. Wielki, "The Opportunities and Challenges Connected with Implementation of the Big Data Concept," in *Advances in ICT for Business, Industry and Public Sector*, M. Mach-Król, C. M. Olszak, and T. Pełech-Pilichowski, Eds. Springer International Publishing, 2015, pp. 171–189.

[4] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, 2015.

[5] 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.

[6] S. K. Bansal, "Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration," in *2014 IEEE International Congress on Big Data (BigData Congress)*, 2014, pp. 522–529.

[7] X. L. Dong and D. Srivastava, "Big data integration," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 2013, pp. 1245–1248.

[8] G.-Z. Yang, J. Andreu-Perez, X. Hu, and S. Thiemjarus, "Multi-sensor Fusion," in *Body Sensor Networks*, G.-Z. Yang, Ed. Springer London, 2014, pp. 301–354.

[9] P. Z. Yeh and C. A. Puri, "An Efficient and Robust Approach for Discovering Data Quality Rules," in *2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2010, vol. 1, pp. 248–255.
[10] F. Chiang and R. J. Miller, "Discovering data quality rules," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1166–1177, 2008.
[11] P. Oliveira, F. Rodrigues, and P. R. Henriques, "A Formal Definition of Data Quality Problems.," in *IQ*, 2005.
[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] M. Maier, A. Serebrenik, and I. T. P. Vanderfeesten, *Towards a Big Data Reference Architecture*. University of Eindhoven, 2013.
[14] C. Fürber and M. Hepp, "Towards a Vocabulary for Data Quality Management in Semantic Web Architectures," in *Proceedings of the 1st International Workshop on Linked Web Data Management*, New York, NY, USA, 2011, pp. 1–8.
[15] 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.
[16] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," *ACM Comput Surv*, vol. 41, no. 3, pp. 16:1–16:52, Jul. 2009.
[17] 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," *Int. J. Prod. Econ.*, vol. 154, pp. 72–80, 2014.
[18] B. Saha and D. Srivastava, "Data quality: The other face of Big Data," in *2014 IEEE 30th International Conference on Data Engineering (ICDE)*, 2014, pp. 1294–1297.
[19] C. Cappiello, A. Caro, A. Rodriguez, and I. Caballero, "An Approach To Design Business Processes Addressing Data Quality Issues," 2013.
[20] W. Fan, F. Geerts, and J. Wijsen, "Determining the currency of data," *ACM Trans. Database Syst. TODS*, vol. 37, no. 4, p. 25, 2012.
[21] N. Tang, "Big Data Cleaning," in *Web Technologies and Applications*, L. Chen, Y. Jia, T. Sellis, and G. Liu, Eds. Springer International Publishing, 2014, pp. 13–24.
[22] 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.
[23] 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, pp. 46–53.
[24] K. Holley, G. Sivakumar, and K. Kannan, "Enrichment Patterns for Big Data," in *2014 IEEE International Congress on Big Data (BigData Congress)*, 2014, pp. 796–799.
[25] D. Loshin, "16 - Data enrichment/enhancement," in *Enterprise Knowledge Management*, D. Loshin, Ed. San Diego: Academic Press, 2001, pp. 399–424.
[26] G. A. Liebchen and M. Shepperd, "Software productivity analysis of a large data set and issues of confidentiality and data quality," in *Software Metrics, 2005. 11th IEEE International Symposium*, 2005, p. 3 pp. –46.
[27] H. Hj Mohamed, T. L. Kheng, C. Collin, and O. S. Lee, "E-Clean: A Data Cleaning Framework for Patient Data," in *2011 First International Conference on Informatics and Computational Intelligence (ICI)*, 2011, pp. 63–68.
[28] J. I. Maletic and A. Marcus, "Data cleansing: A prelude to knowledge discovery," in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 19–32.
[29] A. Ebaid, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, J.-A. Quiane-Ruiz, N. Tang, and S. Yin, "NADEEF: A generalized data cleaning system," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1218–1221, 2013.
[30] H. Müller and J.-C. Freytag, *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.
[31] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng Bull*, vol. 23, no. 4, pp. 3–13, 2000.
[32] B. Glavic, "Big Data Provenance: Challenges and Implications for Benchmarking," in *Specifying Big Data Benchmarks*, Springer, 2014, pp. 72–80.
[33] R. Hammad and C.-S. Wu, "Provenance as a Service: A Data-centric Approach for Real-Time Monitoring," in *2014 IEEE International Congress on Big Data (BigData Congress)*, 2014, pp. 258–265.
[34] Á. G. Recuero, S. Esteves, and L. Veiga, "Towards quality-of-service driven consistency for Big Data management," *Int. J. Big Data Intell.*, vol. 1, no. 1/2, p. 74, 2014.
[35] 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, pp. 86–93.
[36] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23) 2000 (June 13).