Big Data Pre-Processing: Closing the Data Quality Enforcement Loop

Ikbal Taleb[1], Mohamed Adel Serhani[2]

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

***Abstract*— In the Big Data Era, data is the core for any governmental, institutional, and private organization. Efforts were geared towards extracting highly valuable insights that cannot happen if data is of poor quality. Therefore, data quality (DQ) is considered as a key element in Big data processing phase. In this stage, low quality data is not penetrated to the Big Data value chain. This paper, addresses the data quality rules discovery (DQR) after the evaluation of quality and prior to Big Data pre-processing. We propose a DQR discovery model to enhance and accurately target the pre-processing activities based on quality requirements. We defined, a set of pre-processing activities associated with data quality dimensions (DQD's) to automatize the DQR generation process. Rules optimization are applied on validated rules to avoid multi-passes pre-processing activities and eliminates duplicate rules. Conducted experiments showed an increased quality scores after applying the discovered and optimized DQR's on data.**

Keywords - Big Data; Data Quality Evaluation; Data Quality Rules Discovery; Big Data Pre-Processing;

I. INTRODUCTION

Nowadays, most of companies consider data as an asset in an era where almost all business strategic decisions are based on insights collected from the 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 technical and human factor. Big Data travels through all phases of its lifecycle, such phases include data processing, analytics, and visiualization. However, without clean, consistent, and complete data these phases will not prevail. Yet, any data processing remains very sensitive when data is not suitable and ready to be processed. This result in an unusable data and analysis caused by factors such as bad data preparation, nature of data, its format, its origin, and its type.

In Big Data, the crucial problem resides in the data itself and its quality. Big Data reveals a set of characteristics that have a direct impact on Data Quality (DQ) [9]. Thus, a data preparation is required to build confidence on data to assure a certain level of data quality.

The most important questions to be raised are: (1) How can we get benefit from data quality evaluation performed on Big Data samples? (2) How to discover quality rules from data to improve data quality? (3) finally, how these strengthen data pre-processing activities?

The answers are: (a) we need to redefine and personalize the pre-processing activities for each data set based on specific quality requirements and evaluations, (b) the data quality requirements and specifications must include targeted DQD's and their related data attributes, and (c) use the quality evaluation results, to discover, generate, test, validate and optimize the DQRs.

In this paper, we propose a Big Data quality rules generation and discovery model from the quality evaluation results. The quality is estimated prior to any pre-processing task. This phase provides a well-constructed data quality rules to be used in the pre-processing phase. These target data attributes for specific data quality dimensions based on quality evaluation scores. This information provides a way to build quality rules for attributes with low quality score. The rule set can be refined by a user-expert and applied to improve quality dimension.

The paper is organized as follows: next section discusses related works around data quality enhancement and Quality rules. In Section III, we describe our data quality rules discovery model and its modules. However, section IV, highlights the evaluation results and discusses the quality rules generation algorithm we have developed around a set of quality dimensions and pre-processing activities. Section V concludes the paper and points to some future directions.

II. RELATED WORKS

In this paper, we investigate the discovery of data quality rules from data quality evaluation scores. These rules will be used in Big Data pre-processing activities in roder to improve quality of data. This process is characterized by many challenges that should consider different factors including data attributes, data quality dimensions, data quality rules discovery, and their connection with pre-processing activities.

There are two main strategies to improve data quality according to [1], [2]; data-driven and process-driven. The first strategy handles the data quality in the pre-processing phase by applying some pre-processing activities (PPA) like cleansing, filtering, and normalization. These PPA are important, and take place prior to the data processing stage, preferably as early as possible. However, the process-driven tackles the quality at each Big data value chain process.

978-1-5386-1996-4/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataCongress.2017.73

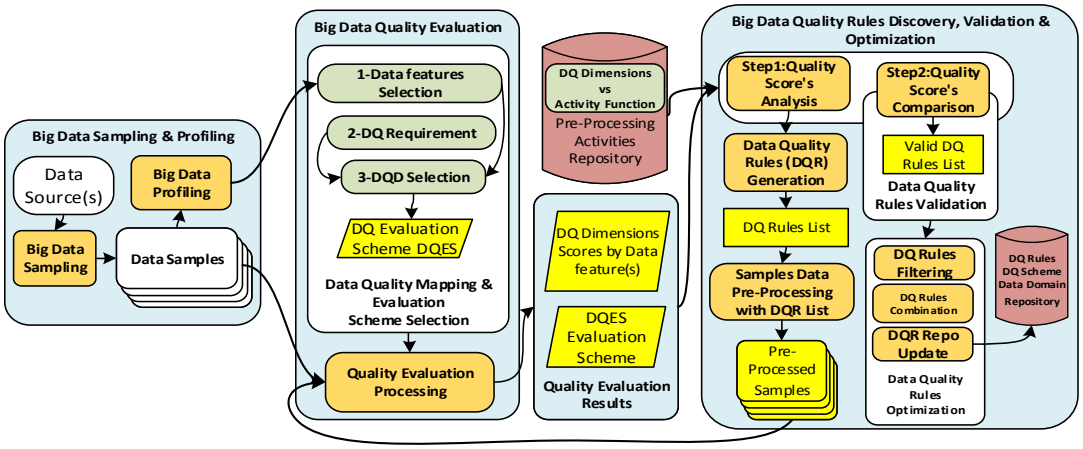

Figure 1. Big Data: Quality Rules Discovery Framework

In [3], the authors concluded that the data quality problems are data, time, and context dependent. Quality rules are applied on data to solve and/or avoid quality problems. Accordingly, the quality rules must be continuously assessed, updated, and optimized.

Most works on data quality rules discovery comes from database community, they are often based on conditional functional dependencies (CFDs) to detect inconsistencies in data. CFDs are used to formulate data Quality rule, generally expressed manually, and discovered automatically using several CFD's approaches [4], [6].

Data quality assessment in Big Data has been addressed in several works. In [7], a Data Quality-in-Use model is proposed to assess the Quality of Big data. Business rules for data quality were used to decide which data must meet a defined constraints or requirements. In [8], a new quality assessment approach was introduced and involve both provider, and consumer of the data. The assessment is mostly based on data consistency rules provided as metadata.

All the aforementioned works on data quality and data quality rules discovery are based on CFD's. In Big Data quality assessment, the size, variety and veracity of data are key characteristics that have to be considered as they reduce the quality assessment time and resources since they are handled in the pre-pre-processing stage. In this paper, we propose a Big Data Quality rules discovery model that analyze quality evaluation scores and generate data quality rules (DQR). The DQR's are pre-processing activities that target and correct a specific poor quality dimension. The introduction of pre-processing activity repository (PPA REPO) to define PPA functions (e.g. remove missing data) indexed by DQD's (e.g. accuracy) and by activity (e.g. data cleansing) is considered a value added feature of Big Data Pre-Processing. These DQR estimation is done before and after the intermediate pre-processing phase.

III. DATA QUALITY RULES DISCOVERY

The purpose of a Data Quality Rule (DQRP) Process is to discover, optimize and generate a set of data quality rules taking into account many parameters: (1) DQ requirements, (2) data attributes or features, (3) targeted DQD's and (4) DQD evaluation results.

In this work, we are dealing with data quality before the pre-processing phase. These DQR's are essential to correct and improve the data quality while setting the best pre-processing activities. The DQR component is illustrated in Figure 1 where the key components of our DQRP consist of: (a) Big Data sampling and profiling, (b) Big data quality mapping and evaluation, (c) Big data quality rules discovery (e) DQR validation and (f) DQR optimization.

In the following sections, we describe each module, its input(s) and output(s), the main functions, and its roles and interactions with the other modules.

A. Big Data Sampling and Profiling

Since profiling is an activity to discover data characteristics from one or more data sources. It is considered as data assessment process that provides a first impression on the data quality reported in its data profile. In our previous work [9], we used the BLB bootstrap for Big data sampling to efficiently sample Big data while not losing precision and reducing evaluation time. Let $\boldsymbol{S}$ a set of data samples from the data source: $\boldsymbol{S}= \{s_0,\ldots, s_i,\ldots, s_n\}$ and $\boldsymbol{P}$ the respective profiles.

B. Quality Mapping / Evaluation Processing

A mapping must be done between data quality dimensions and the targeted data features/attributes. Each DQD is measured for each attributes and for each sample. A quality score is calculated for each DQD and for each attributes or set of attributes depending on the DQD itself. A data quality dimension may target one or more data attributes. The quality evaluation generates quality scores, a quality score model is used to analyze these results. This model is provided as quality requirements to understand the scores expressed as quality level of acceptance. The quality requirements can be a set of values, an interval in which values are accepted or rejected, or a single score ratio.

Let's note by $\boldsymbol{A}$, a set of data attributes, $\boldsymbol{D}$ a set of data quality dimensions, and $\boldsymbol{R}$ a set of Quality requirements. The quality mapping produce a Data Quality Evaluation

Scheme set ***DQES*** $(Q_0(a_j, d_k, r_t),\ldots Q_x(a_a, d_b, r_c),...)$ each elements is a quality score for a specific attribute, DQD, and a quality requirement. At the processing stage the ***DQES*** is applied on a set of samples ***S***, which result in a Quality scores represented by ***QScore*** containing the DQD quality scores for each attribute. (Lines 5 to 10 in Table 1.)

Table 1. Big Data Quality Rules Discovery Algorithm

Algorithm: Big Data Quality Rules Discovery

```
1  Input: (S, A, D, R): Quality Mapping Selection
2      A={a_0,...,a_i,..., a_n}, R = {r_0,...,r_l,..., r_t}
3      D={d_0,...,d_k,..., d_a }, S= {s_0,...,s_i,..., s_n}
4  Output: DQES(Q_0(a_i, d_k, r_t),...Q_x)
5  Quality Evaluation Processing: QEP (DQES, S, QScore)
6      For each Mapped tuple Q_x(a_i, d_k, r_t) in DQES
7          For each s_i in S
8              QualityEval ( Q_x(a_i, d_k, r_t), s_i ) → Q_xScore(a_i,
9          End s_i
10     End Q_x
11 Quality Rules Discovery
12 Input: QScores, DQES, PPA(d_k af_k, v) af_k.v:Activity fonction for
13 Output: DQ Rules List : DQRL(Q_xR(a_i, d_k, PPA(d_k, af_k.v))
14     For each Score tuple Q_xScore(a_i, d_k, r_t)
15         Analyze (Q_xScore(a_i, d_k, r_t ), PA(d_k, af_k.v))
16         GenerateQRules () -->Q_xR(Q_x(a_i, d_k, r_t ), PA(d_k,
17     End Q_xScore
18 Samples Pre-Processing: Input (S, DQRL) Output (S')
19     For each Rule Q_xR(a_i, d_k, r_t ), PA(d_k, af_k.v)) in DQRL
20         For each s_i in S
21             For each a_i , dk
22                 Pre-Processing( Q_xR( a_i,d_k,af_k.v), s_i)
23             End a_i , d_k
24         End s_i
25         Output: s'_i Pre-Processed samples
26     End Q_xR
27 Quality Evaluation Processing: QEP (Q_x, S', QScore')
28 Quality Scores Validation: Input (QScores, QScores', DQRL)
29 Output: DRQLV, DRQLN (V:valid N: not valid Quality rules)
30     For each Q_xScore(a_i, d_k, r_t)
31         For each s_i in S
32             if (ValidScore( Q_xScore(a_i, d_k, r_t),Q_xScore'(...)))
33                 Add Q_xR( a_i,d_k,af_k.v) to DRQLV
34             else Add Q_xR( a_i,d_k,af_k.v) to DRQLN
35         End s_i
36     End Q_xScore
37 Data Quality rules Optimization (DQRLV)
```

C. *Quality rules discovery, validation, and optimization*

1) Quality scores results analysis

Each DQD evaluation $Q_x(a_j, d_k, r_t)$, in ***DQES*** generate a quality score $Q_xScore(a_j, d_k, r_t)$. These scores are analyzed against quality requirements. The quality rules are generated, and attributes fully violate these rules might be discarded.

2) Pre-Processing activities repository (PPA) and quality rules generation

The PPA repository is organized as a tuple $PPA(d_k, af_{k,y})$. Each data quality dimension d_k is associated with an activity function. For example: the DQD completeness of a set of attributes is evaluated to give the ratio of complete data observation within a set of selected attributes or features of the data. One of the corresponding pre-processing activity is to eliminate the data that didn't satisfy this DQD. In another hand, the activity function can also fill the missing data with specific range of values (e.g. repeated value or the mean). All these possibilities are expressed in the requirements set ***R*** in the Quality mapping stage.

For the failed evaluation score $Q_xScore(a_j, d_k, r_t)$ a rule is generated based on the pre-processing activity repository. Each rule Q_xR is represented by the tuple: $Q_xR(Q_x (a_j, d_k, r_t), PPA(d_k, af_{k,y}))$ with Q_x the Quality mapping element from DQES and ***PPA*** is the pre-processing activity selected from the repository.

3) Quality Rules Validation

To validate the discovered rules a pre-processing is applied on a set of samples ***S*** and a reevaluation of quality based on the same evaluation scheme ***QEP (DQES, S', QScore')*** resulted a set of samples ***S'.*** A direct ccomparison of resulting scores from both evaluations results ***QScore and QScore'*** is conducted to filter the set of valid rules (***DRQLV***) from the original set (***DRQLN***). There are two types of unproductive rules: rules that didn't improve the quality score when applied on data, and the rules that decrease the original quality scores.

4) Quality rules optimization

In the final stage, the ***DRQLV*** rules are optimized under several situations which may depend on the selected features/attributes. In the following are some optimization scheme that can be applied on the set of rules.

a) The rules are grouped per attributes, dimensions, or pre-processing activity to detect duplicate **PPA**s.

b) Remove duplicated rules per attributes by grouping all the activities or same activity for multiple attributes..

c) Combining rules targeting same attribute(s) or a set of. Then ordering the activities per execution priority.

d) Prioritizing the activity function that replaces data attributes in the case of missing values and also for fulfilling data completeness quality dimension. In this case, some activities eliminate the whole feature from the data and automatically any other related activity where the attribute exist alone or within a set of attributes should be canceled.

e) Combining same (DQD, PPA) tuple for many attributes/features in one rule to avoid multi-passes pre-processing.

IV. QUALITY RULES EXPERIMENTS AND ANALYSIS

In this section, we describe some experiments we have conducted using a generated Big Data set injected with noisy data. The quality evaluation scores are analyzed before and after discovering and applying quality rules on a samples set. Quality dimensions were measured using a set of quality metrics including Accuracy (***Acc***), Completeness (***Comp***) and Consistency (***Cons***). Figure 2 describes how we close the loop of quality assessment from data quality evaluation, rule discovery, validation, and optimization.

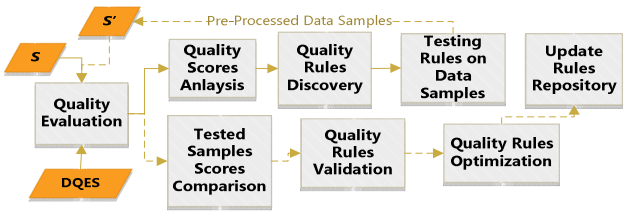

Figure 2. Big Data Quality Rules evaluation processes

In the following, we describe the above processes:

1) The **DQES** scheme represents a set of $\mathbf{Q_x}$ scores to evaluate a DQD, for a set of data attributes describing a structured Big data's samples set **S**. For example, $\mathbf{Q_0}$ evaluates the attribute $\mathbf{Att_{10}}$ of DQD completeness (**Comp**) with a requirement $\mathbf{r_0}$ (e.g. Minimum DQD score 80%). The quality evaluation result indicate a $\mathbf{Q_0Score}$ of 50%.

Quality Scores: $Q_x(a_j, d_k, r_t) \rightarrow Q_xScore(a_j, d_k, r_t)$
- $Q_0(Att_{10}, Comp, 80\%) \rightarrow Q_0Score = 50\%$
- $Q_1(Att_1,Att_2,Att_5,Att_{10},Att_{50},Att_{95},Comp,75\%) \rightarrow Q_1Score = 70\%$
- $Q_2(Att_{10}, Acc, 90\%) \rightarrow Q_2Score = 83\%$

2) The resulted $\mathbf{Q_0Score}$ (50%) is analyzed and compared to the targeted requirement $\mathbf{r_0} >= 80\%$. Accordingly, a Quality rule is generated to pre-process the data and achieve the 80% target score. To enhance the $\mathbf{Q_0Score}$ a $\mathbf{Q_0R}$ quality rules is generated mapping the DQD with a pre-processing activity **PPA.** For this instance the **PPA** is data cleansing with an activity function ***af*** to remove incomplete observations for $\mathbf{Att_{10}}$. Moreover, an ***af*** replaces missing values with the $\mathbf{Att_{10}}$ mean value.

Discovered Rules with their mapped *PPA af* functions
- $Q_0R(Q_0(Att_{10}, Comp, 80\%),50\%)$, *PPA(Comp, delete incomplete rows))*
- $Q_1R(Q_1(Att_1,Att_2,Att_5,Att_{10},Att_{50},Att_{95}, Comp, 75\%), 70\%)$, *PPA(Comp, delete incomplete rows for missing Attributes)) or*
- $Q_1R(Q_1(Att_1,Att_2,Att_5,Att_{15},Att_{50},Att_{95},Comp,75\%),70\%)$, *PPA(Comp, replace incomplete rows for missing attributes))*

3) A validation is conducted after the Quality rules are tested on new data samples set producing a new preprocessed samples set ***S'***. A reassessment of the ***DQES*** on ***S'*** is done to confirm the validity and the effectiveness of quality rules.

Quality Scores after pre-processing discovered rules
- $Q_0(Att_{10}, Comp, 80\%) \rightarrow Q_0Score(Att_{10}, Comp, 80\%) = 93\%$ **Valid**
- $Q_1(Att_1,Att_2,Att_5,Att_{10},Att_{50},Att_{95},Comp, 75\%) \rightarrow Q_1Score = 80\%$ **Valid**
- $Q_2(Att_{10}, Acc, 90\%) \rightarrow Q_2Score = 93\%$

4) The optimization can be applied in different situations such as combining redundant rules targeting the same DQD or the same attribute, for example:

- $Q_2R(Q_2(Att_{10}, Acc, 90\%), 83\%)$, *PPA(Acc, remove inaccurate rows)):* the DQD **Comp** is merged with **Acc (Accuracy)** when considering missing values.
- The user can decide to keep missing data rows, and replace them with the mean or the most repeated value.
- Finally, a rule execution priority is set to $\mathbf{Q_2R}$ before $\mathbf{Q_1R}$ (since the $\mathbf{Att_{10}}$ is targeted in $\mathbf{Q_1R}$ for **completeness**) to avoid discarding rows with $\mathbf{Att_{10}}$ missing values.

V. CONCLUSION

In this paper, we proposed a quality based rule model to support the Big Data pre-processing. The model relies on extracting quality rules from Big data quality evaluation while considering a set of quality requirements. We applied generated rules on Big data samples, then we re-evaluated the quality to validate these rules. The value-added feature of our model is the process of quality rule optimization, and the mapping between the pre-processing activities and the targeted DQD. The experiments we have conducted on a set of Big data samples prove that quality rules are discovered, validated, and then optimized to significantly improve the quality in Big data pre-processing stage.

VI. REFERENCES

[1] 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.

[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 *2012 International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012, pp. 300–304.

[3] Y. W. Lee, "Crafting rules: context-reflective data quality problem solving," *J. Manag. Inf. Syst.*, vol. 20, no. 3, pp. 93–119, 2003.

[4] P. Z. Yeh and C. A. Puri, "An Efficient and Robust Approach for Discovering Data Quality Rules," *22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2010, vol. 1, pp. 248–255.

[5] F. Chiang and R. J. Miller, "Discovering data quality rules," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1166–1177, 2008.

[6] W. Fan, "Dependencies revisited for improving data quality," in *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART,* 2008, pp. 159–170.

[7] I. Caballero, M. Serrano, and M. Piattini, "A Data Quality in Use Model for Big Data," in *Advances in Conceptual Modeling*, Springer Pubs, 2014, pp. 65–74.

[8] M. Kläs, W. Putz, and T. Lutz, "Quality Evaluation for Big Data: A Scalable Assessment Approach and First Evaluation Results," in *2016 Joint Conference (IWSM-MENSURA)*, 2016, pp. 115–124.

[9] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation," in *2016 Intl IEEE Conf. on (UIC/ATC/ScalCom/CBDCom/)*, 2016, pp. 759–765.