






# Big Data Quality: A Data Quality Profiling Model

Ikbal Taleb<sup>1</sup> , Mohamed Adel Serhani<sup>2</sup> ,  
and Rachida Dssouli<sup>1</sup> 

<sup>1</sup> Concordia University, Montreal, QC H3G 2W1, Canada  
i\_taleb@live.concordia.ca, rachida.dssouli@concordia.ca

<sup>2</sup> UAE University, Al Ain, Abu Dhabi, United Arab Emirates  
serhanim@uaeu.ac.ae

**Abstract.** Big Data is becoming a standard data model, and it is gaining wide adoption in the digital universe. Estimating the Quality of Big Data is recognized to be essential for data management and data governance. To ensure a fast and efficient data quality assessment represented by its dimensions, we need to extend the data profiling model to incorporate also quality profiling. The latter encompasses more value-added quality processes that go beyond data and its corresponding metadata. In this paper, we propose a Data Quality Profiling Model (BDQPM) for Big Data that involves several modules such as sampling, profiling, exploratory quality profiling, quality profile repository (QPREPO), and the data quality profile (DQP). Thus, the QPREPO plays an important role in managing many quality-related elements such as data quality dimensions and their related metrics, pre-defined quality actions scenarios, pre-processing activities (PPA), their related functions (PPAF), and the data quality profile. Our exploratory quality profiling method discovers a set of PPAF from systematic predefined quality actions scenarios to leverage the quality trends of any data set and show the cause and effects of such a process on the data. Such a quality overview is considered as a preliminary quality profile of the data. We conducted a series of experiments to test different features of the BDQPM including sampling and profiling, quality evaluation, and exploratory quality profiling for Big Data quality enhancement. The results prove that quality profiling tracks quality at the earlier stage of Big data life cycle leading to quality improvement and enforcement insights from exploratory quality profiling methodology.

**Keywords:** Big Data quality · Data Quality Profile · Profile repository · Data quality profiling

## 1 Introduction and Background

The big data ecosystem is defined as the way we gather, store, manipulate, analyze and get insight from a fast-increasing heterogeneous data. According to IBM [1], every day huge amounts of data are generated; this data represents 2.5 quintillion bytes (Exabyte (EB) =  $10^{18}$  bytes) [2]. In 2000, 800,000 Petabyte (1 PB =  $10^{15}$  bytes) of data were stored. In 2020, the worldwide storage will reach 35 Zettabytes (1 ZB =  $10^{21}$

bytes = 1 Trillion gigabytes). Which urge the need to automatically profile, characterize and categorize the quality of such data. These classifications are strongly coupled with the semantic meaning of what the data represents. In many cases, the data comes in a format and a quality state in which it is impossible to process immediately as it is, and if so, the results cannot guarantee a trustable analysis and insights. The importance of estimating and profiling the quality of Big Data is paramount and has priority over the other Big Data stages.

In this paper, we introduce a data quality profiling model (BDQPM) for Big Data that acts as a preliminary quality discovery in the Big data lifecycle before engaging with any analytics of a data source. Therefore, quality profiling will intervene before the pre-processing stage of Big Data. The exploratory quality profiling module is considered as the core of our model in which a pre-defined pre-processing quality scenarios actions are applied on data samples. These scenarios target specific Data Quality Dimensions (DQD's) while varying the DQD acceptance ratio level set (from min to max ratio). The resulted pre-processed samples will have their DQD evaluated in the quality evaluation module while the max ratio is not reached. Once the quality results are aggregated, an analysis is done and the set of pre-processing activity functions (PPAF's) that affected the DQD ratio are selected to be used as quality enhancements rules.

Our model relies on a quality profile repository (QPRepo) that handles all the related quality tables, from Pre-Processing Activities (PPA), Pre-Processing Activity Functions (PPAF), DQD's and metrics. The most important is the Data Quality Profile (DQP) that plays the role of a record book of quality that tracks all the data and quality profiling results, metadata, the pre-defined quality scenarios, the quality scores, and the quality reports. Finally, a set of experimentations to validate out model modules: (a) sampling and profiling, (b) sampling and quality evaluation and (c) exploratory quality profiling for Big Data quality improvement PPAF extraction.

The rest of paper is organized as follows: next section introduces Big Data and data quality fundamentals, definition, characteristics, and lifecycle. Section 3 surveys the most important research on quality profiling for Big Data. Section 4 introduces our Big Data quality profiling approach. Section 5 presents our Big Data quality profile repository. Section 6, analyzes and discusses our experimentations. Finally, the last section concludes the paper and highlights some ongoing and challenging research directions.

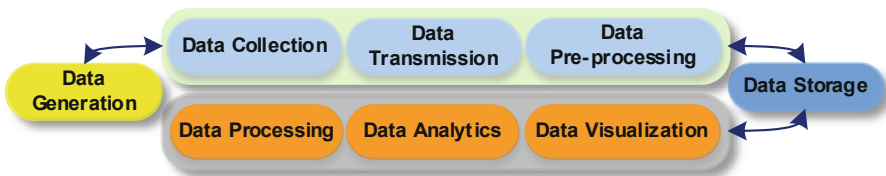


Fig. 1. Big Data key processes

## 1.1 Big Data Value Chain

As illustrated in Fig. 1, Big Data is handled through a lifecycle (also called value chain) that tracks and handles the data from its inception to insights generation. There are many stages where the data goes through to achieve a specific goal. In the following, a brief description of these stages:

- **Data Generation:** this is considered as the inception stage about the data sources, where the data is being engendered.
- **Data Acquisition:** it involves data collection, transmission and pre-processing. In the pre-processing, the data might be combined from different sources and tailored into a pre-defined format for the purpose of processing.
- **Data Processing and Analytics:** it consists of processing data using several analytics approaches and tools (e.g. Data Mining techniques, Machine Learning algorithms, Deep Learning).
- **Data Transmission and Storage:** consist of transmission of huge data over a network and the distribution and replication of storage.

**Big Data Characteristics (V's).** Big Data is also described by its characteristics that brand the Data as “*Big Data*”. The initial Big Data characteristics are volume, velocity, variety [3, 8, 19, 22]. However, these got extended to cover other 9 extra Big data characteristics that we compiled with the illustrated Fig. 2. Moreover, V's represents the key elements affecting traditional data to become Big Data.

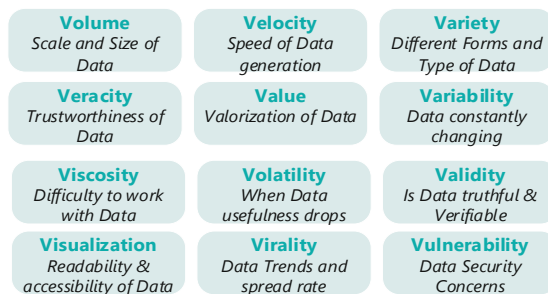


Fig. 2. Big Data characteristics

## 1.2 Data Quality

According to [18], data quality is not easy to define, its meanings are data domain dependent and context-aware. Data quality is continuously related to the quality of its data source [15]. It is also defined by its quality dimensions, metrics and assessments.

**Data Quality Dimensions:** To measure and manage data quality the concepts of a Data Quality Dimension (DQD) is presented in [4, 9, 20]. There are many quality dimensions classified under categories that define them. Some essential DQD categories are (a) the

contextual dimensions that are associated with the information and (b) intrinsic dimensions that refer to the objective and native data attributes. Examples of intrinsic data quality dimensions include Accuracy, Timeliness, Consistency, and Completeness.

**Data Quality Metrics:** Each DQD are associated with specific metrics. A metric is an equation, or a formula developed to compute a score or the ratio of the data by quantifying its quality dimensions. The metrics provide a way to evaluate a DQD from simple formulas to more complex multivariate expressions.

**Data Quality Assessment:** Using a set of metrics, it is feasible to evaluate quantitatively the quality when following a data-driven strategy on existing data. For structured data, a quantitative evaluation is not possible since it is not expressed in the form of attributes with columns or rows filled with values. Therefore, unstructured data needs a different evaluation approach given the fact that we don't know how it is organized, and what are we going to assess. The introduction of a module that extracts, discover, or define attributes and features with specific DQD mapping are mandatory to proceed with quality exploration.

### 1.3 Data Profiling and Data Quality Profiling

**Data Profiling:** Data profiling can be applied at different stages of the data lifecycle. It is defined as the process of verifying data in its different types and formats such as structured, semi-structured, and unstructured data. Then collect and visualize various information about data, including: structure, patterns, statistics, metadata, data attributes or features. Moreover, all these assembled information's are used or requested for data governance, data management, and data quality control [6]. There are many data profiling analysis schemes used for this purpose, such as attribute analysis, referential analysis and functional dependency analysis [1]. In the following, we summarize the main benefits of data profiling that helps in:

1. Finding irregularities in data in the earlier stages and takes correction actions.
2. Understand content, structure, and relationships about the data in the data source.
3. Assess, validate and analyze metadata.
4. Making a statistical analysis of the data at its source.

A typical use case of data profiling is in a data cleansing process. In many commercial tools, profiling is always bundled with a data quality cleansing software. Profiling using data quality assessment tool detects data errors, such as inconsistent formatting within a column, missing values, or outliers. Therefore, profiling results might be used to measure and monitor certain quality dimensions of a dataset such as the ratios of observations that are not satisfying data constraints [2, 13].

**Data Quality profiling:** is the process of analyzing a dataset in the context of a quality domain defined by a set of quality and data requirements, to detect quality issues. The results of data quality profiling may include:

- Summaries describing: (a) completeness of datasets and data records (b) data problems (e.g. Wrong entries, inconsistent data) (c) problem's distribution in a dataset.
- Details about: (a) missing data records (b) data problems in existing records.

## 2 Literature Review

By surveying the literature, we haven't found comprehensive research that emphasizes Big Data quality profiling in the context of Big data. Indeed, very few works addressed the profiling process in Big Data levelling some aspects, like in [11] where the authors addressed the challenges of user profiling in Big Data using techniques focusing essentially on privacy. On the same path, the authors in [5, 10] showed that web user profiling techniques are used based on cookies for real-time profiling and marketing. They presented a framework for web user profiling based on leveraging the redundant information on the web. On the other side, Jamil and al. presented a set of guidelines for planning, conducting and reporting a systematic review and provided a review of the literature on profiling digital news for Big Data veracity [12].

In [1], Abedjan et al. stressed that data profiling is important to discover metadata, and further profiling work is required in the context of new types of data such as Big Data. Especially for profiling results visualization and interpretation that is still challenging. In [7], Naumann presented Big data profiling following the same methodology as traditional data profiling with a taste of measuring the Big Data Characteristics V's and providing some summaries. The same author in [17] revisited data profiling and stressed the need to level up to a newly modernized data profiling by developing a framework to support data profiling and motivate the need to develop new profiling techniques for Big Data.

In [6], the authors emphasized that data profiling for Big Data is very important in Data Governance. They presented various data quality metrics formulas and calculation along with the commercial and free software profiling tools used for this purpose. However, there is no detailed information on how their profiling system architecture works and process data to provide more accurate metadata. In [14, 16], the authors considered that data profiling might be used as a fast quality assessment and quality issues detection.

Most of the investigated literature hasn't addressed directly or indirectly the Big Data quality profiling. Generally, data profiling is considered as preliminary of data quality and can be considered as an introductory to data quality profiling. Also, many papers addressed the use of subsets of data to discover metadata as tradeoffs to minimize processing time and costs in the case of Big Data characteristics such as volume.

## 3 Big Data Quality Profiling Model

We propose a quality profiling model that inspects the quality of a dataset following different processes depending on the quality dimension to be inspected.

### 3.1 Big Data Quality Profiling Model Description

As illustrated in Fig. 3, The Big Data quality profiling model is architected around many modules essentially, data sampling, data profiling, the exploratory quality profiling scenarios (EQP), the quality evaluation and PPAF discovery, the data quality profile (DQP) and the QP repository.

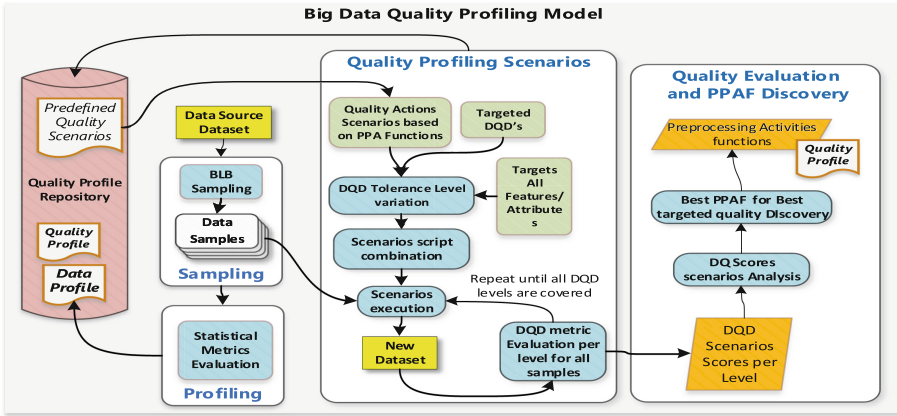


Fig. 3. Big Data Quality Profiling Model (BDQPM)

### 3.2 Sampling and Profiling: Statistical Metric Evaluation

We used the Bag of Little Bootstrap (BLB) [21], which combines the results of bootstrapping multiple small subsets of a Big Data dataset. The BLB algorithm uses Big Data Set to generate small samples without replacements. For each generated sample another set of samples is created by re-sampling with replacements. For data quality evaluation, we used the bootstrap since it is a re-sampling method used to gather the subsequent distribution of the whole data rather than assessing the quality of some estimators.

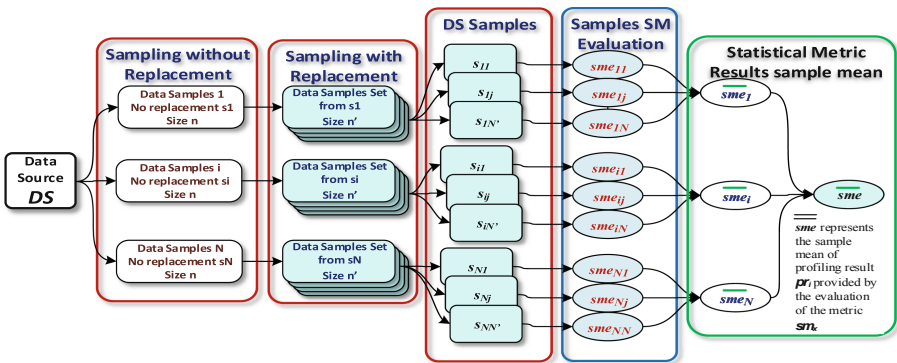


Fig. 4. Big Data sampling & statistical metric evaluation

**Sampling-Profiling Description.** The BDQPM has two modules handling the sampling and the profiling process. The sampling process is invoked also through BDQPM in other modules such as EQP and quality evaluation, In Fig. 4, we describe how data is sampled and profiled for a data statistic metric  $sm_x$ .

**The process sequence of steps:** It follows the 5 steps below:

1. Sampling of the data set  $DS$ ,  $N$  bootstrap samples of  $n$  size without replacement  $s_i$ . ( $i = 1 \dots N$ )
2. Each sample generated from step 1 is re-sampled into  $N'$  samples of size  $n'$  ( $n' < n$ ) with replacements  $s_{ij}$ . ( $i = 1 \dots N, j = 1 \dots N'$ )
3. For each sample  $s_{ij}$  generated in step 2, evaluate the data statistic metric  $sm_x$  as  $sme_{ij}$ .
4. For all the samples  $s_i$ , evaluate the sample mean of all  $N'$  samples  $sm_x$  the evaluation scores are expressed as  $\overline{sme}_{ij}$
5. For the data set  $DS$ , evaluate the data statistic metric score  $\overline{sme}$  which represents the mean of all  $N$  samples evaluation scores  $sm_x$ .

**The Big Data Sampling-Profiling Algorithm (SPA).** The sampling and profiling processes are portrayed in Algorithm 1 (SPA). The procedure *Sample\_Profiling* () is responsible for generation Big Data samples, then profiling these samples simultaneously by calculating the statistical metrics representing a certain data profile.

---

**Algorithm 1: Big Data Sampling and Profiling(SPA)**

---

```

1  Input:  $DS$  Dataset size  $ss$ ,  $A = \{a_1, \dots, a_k, a_r\}$  Attributes
2   $N$  samples  $s_i$  of size  $n$  from  $DS$ 
3   $SM = \{sm_0, \dots, sm_c\}$  Statistical metrics set
4  Output:  $sme$  : Samples Statistical Metric Evaluation
5  procedure Sample_Profiling( $DS, A, SM, N, n$  )
6  //  $s_i$  a sample without and  $s_{ij}$  with replacement
7  for each  $i$  from 1 to  $N$ 
8  |    $s_i \leftarrow$  Generate_Sample( $DS, N, n, no\ rep$  )
9  |   for each  $j$  from 1 to  $N$ 
10 |   |    $s_{ij} \leftarrow$  Generate_Sample( $s_i, N, n, rep$  )
11 |   |   for each  $k$  from 1 to  $R$ 
12 |   |   |   //  $sme_{ij}(k)$  Evaluate  $sm_x$  for  $s_{ij}$  sample attribute
13 |   |   |    $sme_{ij}(k) =$  Eval( $s_{ij}, sm_x, a_k, n$ )
14 |   |   end  $k$ 
15 |   |   for each  $k$  from 1 to  $R$ 
16 |   |   |    $sme_{ij} = sem_{ij} + sme_{ij}(k)$ 
17 |   |   end  $k$ 
18 |   |    $sme_{ij} = sme_{ij} / R$ 
19 |   |   end  $j$ 
20 |   |    $sme_i = sum(sme_{ij}) / N'$ 
21 |   end  $i$ 
22 |    $sme = sum(sme_i) / N$ 
23 return  $sme$ 
24 end procedure

```

---

### 3.3 Quality Evaluation Module

For the quality evaluation module, a DQD is evaluated through all the data attributes using its related metric. For example, for DQD completeness, the metric will compute per attributes, per observation, the ratio of non-empty values divided by total values. It is based on the same algorithm skeleton used in Fig. 4.

### 3.4 Exploratory Quality Profiling for PPAF Discovery

The exploratory quality profiling component executes a series of quality scenarios based on specific DQDs, then generates quality scores. These scenarios, target attributes DQD's at each iteration based on incremental DQD acceptance ratios and applies a set of actions accordingly. In Table 1, we present some Exploratory Data Profiling scenarios, the first scenario executes the following:

**Table 1.** Exploratory quality profiling scenarios

DQD Tolerance Levels %			DQD	PreDefined Scenarios	Execution
min	max	step			
5	95	5	Completeness	DeleteCols(dqd)	1
5	95	5		DeleteRows(dqd)	1
5	95	5		DeleteCols(dqd)	1
5	95	5		DeleteRows(dqd)	2
				DeleteRows(dqd)	1
5	95	5		newdq=Re-Evaluate()	2
				DeleteCols(newdq)	3
				DeleteCols(newdq)	1
5	95	5		newdq=Re-Evaluate()	2
				DeleteRows(dqd)	3

*“For Each iteration of tolerance level Repeat (Delete Columns with DQD ratio less than the tolerance level and recalculate (the resulted dataset DQD ratio))”.*

A pre-defined scenario represents systematic quality actions applied on data attributes or observations or entities when the DQD ratios don't meet the iterative tolerance level. For each DQD required level, if the ratio is not met, the action is applied. The actions are represented as PPAF. The previous example scenario is as a script added to the QPREPO pre-defined quality scenarios table. The script might be written in R, Python, Scala, Java based on the platform used in the experimentations.

A pre-defined scenario is identified by the following information in the QPREPO:

- Scenario ID #: 001 (Key Id in the Scenarios Table)
- Target DQD ID #: 001 (Key Id in the DQD Table)
- Target DQD Name: Completeness (DQD Description)
- Target Data: [A (C), O (R), AO (CR), E] where A/C: Attribute/Column, O/R: Observation/Row, AO/CR for both, E: Entity defined as a chunk of the Dataset.
- (Scenario Script, Language) tuple for each implemented platform.



In the following, we depict the Exploratory Quality Profiling Algorithm.

---

**Algorithm 2: Quality Profiling for Preprocessing Activity Function Discovery**

---

```

1 : Input: DS: Samples Set,
2 :   QAScenarios: predefined Quality Actions Scenarios (PPA,PPF)
3 :   DQDL: DDQD Levels ranging from min to max
4 : Output: QResults: Quality Results
5 :   RRD: Ratio of Remaining Data from the Original Data-Set DS
6 : PPAF_Proposals_Set: Preprocessing Activity Function Set
7 : DQDS_DS(): DQD Scores % per Data-Set
8 : procedure Extract_PPAF_Set(DS, QAScenarios, DQDL)
9 :   DQDS_R(): DQD Scores % per Row, DQDS_C(): per Column
10 :  for each(DQDL  $k$ : K from 1 to N)
11 :    procedure Compute_DQD_Scores(DS)
12 :      DQD_Scores_List ← (DQDS-R, DQDS-C, DQDS_DS)
13 :      return DQD_Scores_List
14 :    end procedure
15 :    procedure Filter_Failed_Rows_Cols(DQD_Scores_List)
16 :      for all  $r_i$  in DQDS-R
17 :        if (DQDS-R( $r_i$ ) >= DQDL  $k$ ) Add ( $r_i$ ,  $R'$ )
18 :      end for
19 :      for all  $c_j$  in , DQDS-C
20 :        if (DQDS-C( $c_j$ ) >= DQDL  $k$ ) Add ( $c_j$ ,  $C'$ )
21 :      end for
22 :      Failed_Scores_List ← ( $R'$ ,  $C'$ )
23 :      return Failed_Scores_List
24 :    end procedure
25 :    procedure Apply_Scenarios_Scores(Failed_Scores_List, QAScenarios)
26 :      PPAF_k ← Generate_PPAF(QAScenarios, DQDL_k)
27 :      DS' ← Execute_PPAF(PPAF_k)
28 :      DQD_Scores_List' ← Compute_DQD_Scores(DS')
29 :    end procedure
30 :    procedure Compute_RR_Ratio(DS, DS')
31 :      RR_k =  $\text{size}(DS') / \text{size}(DS) * 100$  %
32 :      return RR_k
33 :    end procedure
34 :    DQP ← QPrepo_Update(DQDL_k, DQDS_DS_k, DQDS_DS'_k, RR_k, PPAF_k)
35 :  end for  $k$ 
36 :  for each(DQDL  $k$ : K from 1 to N)
37 :    PPAF_Proposals_Set ← KNN_Best_PPAF_k per Targeted Ratios
38 :  end for  $k$ 
39 :  return PPAF_Proposals_Set
40 : end procedure

```

---

Algorithm 2 describes the quality profiling that evaluates first the targeted DQD, then applies a set of scenarios on data that fails the DQD acceptance level. After the PPAF actions on data, another re-evaluation of DQD's is done to check new quality scores. The process is repeated until all iterations on the acceptance levels are completed (e.g. From DQD level = 5 to 95 in step 5). After all the results are gathered and listed by ratio scores, for each type of PPAF actions, a query for the best PPAF actions that leveraged the DQD scores (or lower depending on the DQD description) is performed using the KNN algorithm. The latter ranks the best PPAF combinations that can achieve the goal of the query (refer to Table 4 in the last section for more details).

## 4 Big Data Quality Profile Repository

The Big Data Quality Profile Repository (QPrepo) is considered as an important component of our model. All the Quality information about the data is recorded across the different modules of the framework, including the simple data profile gathered from metadata, the newly data summary, profiled data. Consequently, the QPrepo is created with the aforementioned information for structured data. The more we get into the

quality profiling modules, additional information is added to the QPRepo like the quality evaluation scores of the exploratory quality profiling, the quality scores, the PPAF, the data sources, the data quality dimensions and their metrics, the pre-processing activities and their related activity functions. All this information is recorded in a repository database for Big Data quality profile management. An illustration of the QPRepo data and tables is showed in Fig. 5.

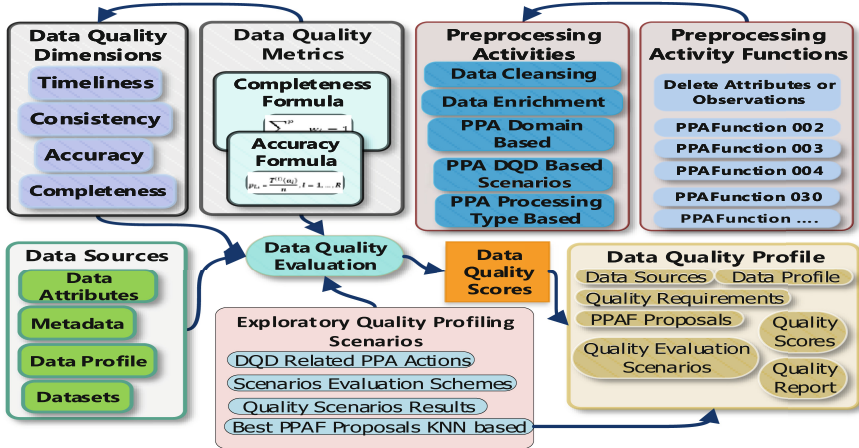


Fig. 5. Big Data quality profile repository

### 4.1 Quality Related Repository Components

The Data Quality Profile holds all the information about the Big Data Quality, from data and its sources, its contents (schema, metadata), its quality requirements, its quality dimensions scores, and the discovered quality pre-processing activities to enhance the quality per attributes, per DQD, or per dataset.

### 4.2 Data Quality Profile

As illustrated in Fig. 6, the DQP is specified in an XML document to store all the modification added through the different processes of the framework. At each module, a version of the DQP is recorded and named a data quality profile level, where the level indicates the module where the DQP was created, updated or upgraded with its processes related quality information and quality reports.

### 4.3 Pre-processing Activities and Related Functions (PPA+PPAF)

The PPA repository is organized as a tuple PPA (DQD, PPAF), where each data quality dimension DQD is associated with an activity function. One of the corresponding pre-processing activities is to eliminate the data that didn't satisfy this DQD or replace it with calculated values based on the data.

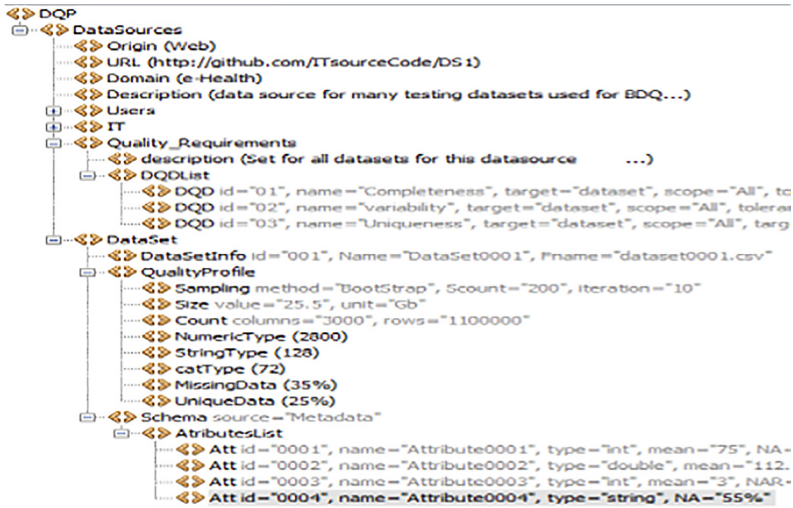


Fig. 6. Big Data quality profile

The QPREPO is responsible for defining and managing pre-processing activities and their related functions. We illustrate QPREPO\_ examples of both Preprocessing Activities (PPA) and their related PPA Functions (PPAF) in detail in Table 2 below. Each DQD is linked to a pre-processing activity which in itself has many functions that deal with the DQD issues and enhance it by removing them. For example, the DQD completeness in the table is described with the methods used to compute it, then the results formulas calculation, and its PPA category (data cleansing) with many functions (PPAF's) actions like data correction or data removal to deal with the completeness issues. The PPAF's vary from replace the missing values with several methods based on the existing data itself or drop and remove the missing data attributes, observations or a balanced combination of them as illustrated in the pre-defined scenarios in Table 2.

Table 2. Preprocessing activities and PPA functions

#	DQD	Metric	Data Type	Methods	Results*100 (%)	PPA	PPAF	PPF Related Actions or Proposals
11	Accuracy Consistency Validity	Outliers Detection	Num	Rule based	(Outliers	<i>Data Cleansing</i>	Retention	Use robust classification methods
12				Linear Regression	Count)/(Total		Winsorizing	Replace Outliers with closest values
13				High Dimensional	List of Outliers		Exclusion, Truncation	Remove Related Rows
				Outlier Detection	(Anomaly , Novelty)			
				Methods				
21	Completeness	Recorded Obs	All	Count the number of	(Not NA Count)	<i>Data Enrichment</i>	Data	Replace with Mean
22				Not (NA, Null or	(Not NA Count)		Correction	Replace with Mode
23				any other values that	/(Total observations			Replace with Median
24				express the No	(Rows))			Remove Rows
25				Available data values)			Data Removal	Remove Columns
26							Remove Rows and Cols	

## 5 Experimentations and Discussions

### 5.1 Experimental Setup

The following is the hardware, software, and dataset we have used:

- **Hardware:** Two Databrick Spark Clusters hosted on Amazon WS with 4 nodes each of 16 GB of RAM and auto-scaling storage (Max of 100 TB).
- **Software:** Spark Run-time 4.2 (Apache Spark 2.3.1) with pySpark (Python) and SparkR (for R).
- **Dataset:** Experimental Synthetic Dataset with 1300 attributes, 2000000 observations, and 26% of missing data.

### 5.2 Sampling and Profiling (Statistical Metric Evaluation)

As illustrated in Table 3, we conducted experiments to evaluate the Missing Values (MV) ratio for our dataset ( $MV\% = 100\% - \text{Completeness } \%$ ), where completeness is a Data Quality Dimension that can be computed for any type of tabulated attributes.

**Table 3.** Missing values mean using BLB sampling approach

Sample #	Num of Samples	NA Mean	Real NA	Std	p0	p25	p50	p75	p100	hist
1	25	25.91	25.65	0.56	24.77	25.56	25.89	26.25	27.14	
2	49	25.50	25.65	0.73	23.88	25.15	25.53	25.90	27.14	
3	100	25.71	25.65	0.86	23.73	25.17	25.60	26.42	27.36	
4	225	25.56	25.65	0.87	23.39	24.97	25.62	26.07	27.79	
5	400	25.68	25.65	0.78	22.97	25.20	25.73	26.18	27.67	
6	625	25.61	25.65	0.66	23.64	25.17	25.61	26.07	27.67	
7	900	25.55	25.65	0.79	23.32	25.04	25.55	26.07	28.32	
8	1225	25.71	25.65	0.61	23.57	25.32	25.70	26.10	27.73	
9	1600	25.76	25.65	0.70	23.49	25.28	25.78	26.25	28.63	
10	2025	25.68	25.65	0.82	22.88	25.15	25.68	26.25	28.14	

As illustrated in Fig. 7, the number of samples (sub-samples re-sampling) increases to the 10<sup>th</sup> iteration from 25 to 2025 samples with a sample mean that range between 25.50% to 25.91% of Missing Value (MV), the real missing value from the whole data is  $MV = 25.65\%$  making the missing values of samples ranging between  $MV - 0.15$  and  $MV + 0.41$ . This demonstrates the approximations based on Bootstrap sampling.

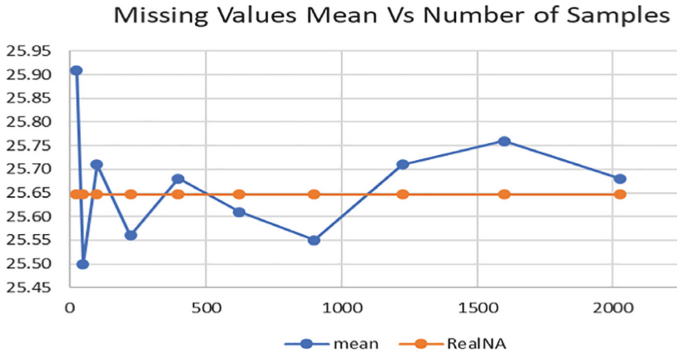


Fig. 7. Missing data representation

### 5.3 Missing Data Row and Column Wise

The completeness quality scores for all attributes are computed per rows and columns. An analysis of completeness of both rows and columns evaluation should highlight what is the most influential completeness acceptance level measured on rows or on columns.

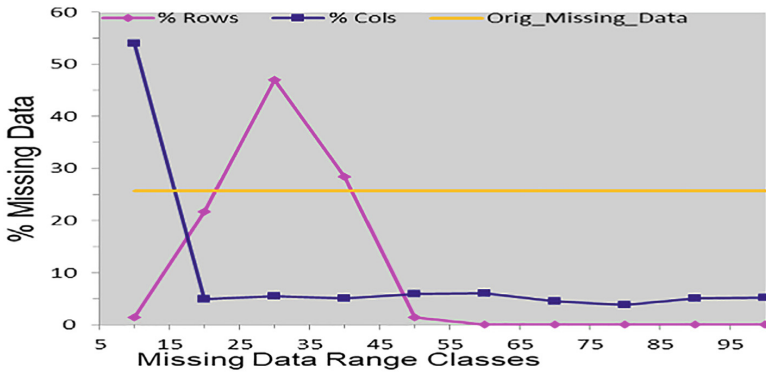


Fig. 8. Missing data ration row and column wise

Figure 8 shows the differences between the column and row-wise calculations (Attributes vs Observations). This is very important to decide which PPAF actions should be applied on data: remove first, columns or rows that have the Missing Values Threshold (MVT: Acceptance Level) less than the MVT fixed in the iterations. The possible PPAF scenarios actions that might enhance the targeted quality are: (1) Remove the objects with the percentage of missing values greater than MVT level. Either, the attributes (Columns), the instances (Observations, Rows), or a combination of both for an optimal acceptable missing value ratio and (2) Replace the missing data with the mean, median, or mode to make the qualitative assessment.

The results reported in Fig. 8 demonstrate that mostly 47% of rows has 30% of missing data. Thus, more than half of the columns (54%) have at least 10% of missing data. The ratio of MV (NA or Null) in the dataset is 26%, therefore removing rows and columns that have MV higher than 26% is not advisable for improving the completeness. In Big Data, removing observations will be prioritized before dropping attributes, thanks to the data volume sufficient to extract insights without losing some hidden features that might hold good insights.

### 5.4 Exploratory Quality Profiling for PPA Function Discovery

The results of this experiment are illustrated in Fig. 9, it shows that the higher MVT amputation percentage is the higher we gain in completeness improvements. However, there is a cost associated with this situation, which is the ratio of removal for both observations and features. Therefore, a reduction ratio between 100% and 70% will not be acceptable even for Big Data. For this reason, we decided that removing threshold must be higher than 50% (Redline) for the PPAF discovery and proposals as depicted in both Figs. 9, and 10.

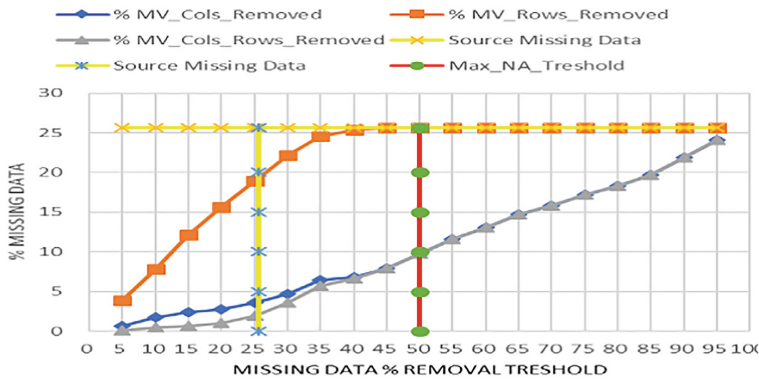


Fig. 9. Impact of removal threshold on completeness

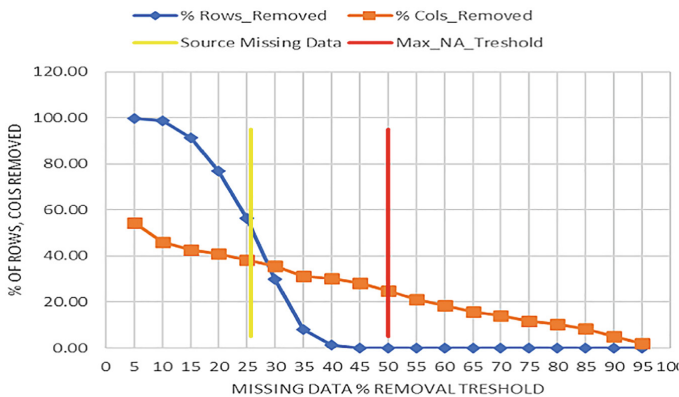


Fig. 10. Impact of removal threshold on Rows-Columns reduction

The pre-processing activity functions PPAF proposals are extracted only when a removing threshold is of 50%. Removing columns rules are more effective in completeness enforcement than rows. The removal of columns that have missing data higher than 50% achieves enhancement of completeness as the missing data drops to 10% while removing 25% of columns. The peak in Fig. 10, shows that removing rows that have more than 25% of missing values, will reduce the dataset size from 26% to 18% with 60% of rows removed, respectively. Therefore, changing the removing threshold to be higher than 30% and more will only achieve the original missing data ratio with 0% of rows reduction.

After the application of the quality scenarios and the DQD results for each iteration of acceptance levels, the next step is to select and extract the best PPAF to be used for quality improvements. The following Table 4 represents the best PPAF selection KNN Based on a set of targeted ratios parameters such as NA% ratio after columns or rows removal, and the ratios % of columns or rows Drop (might be identified as data size reduction caused by quality enhancements actions). An Example of these targeted quality ratios (in Red in Table 4) with 0% of missing values to achieve 100% of DQD Completeness in both columns and rows, an acceptance of 50% drops in rows as we are in a Big Data context letting us sacrifice observations and 10% of attributes.

The purpose is to extract the best pre-processing activity functions already selected from a general pre-defined scenario. The experimentations showed that quality profiling of big data samples is very conclusive since the best PPAF have already been enhancing the quality levels (MVT Level) with proven ratio results as illustrated in Table 4. Moreover, the extracted PPAF actions will be stored in the data quality profile (DQP) as potential functions to be applied in the pre-processing stage on the whole Big Data dataset with a certain level of confidence in the resulted data according to the targeted DQD's. The more DQD's involved the more PPAF are selected and combined to produce high-quality data for the analytics.

**Table 4.** KNN based best PPAF selection targeted ratios

Target %	0		50		10		PPAF Param Rank	Euclid Dist
	NA% After				%Rows Dropped	%Cols Dropped		
ID	MVT Level	Cols Drop	Rows Drop					
1	5	0.66	3.82	99.90	54.46	19	66.94	
2	10	1.73	7.80	98.59	46.01	17	61.00	
3	15	2.38	12.13	91.49	42.72	5	54.27	
4	20	2.79	15.64	76.91	41.00	3	44.02	
5	25	3.61	18.94	56.27	38.42	1	34.91	
6	30	4.74	22.19	29.89	35.45	2	39.58	
7	35	6.40	24.59	7.94	31.30	4	53.56	
8	40	6.81	25.40	1.46	30.36	11	58.84	
9	45	7.95	25.62	0.14	28.09	14	59.44	
10	50	9.72	25.65	0.00	24.73	12	58.90	
11	55	11.56	25.65	0.00	21.28	8	58.47	
12	60	13.06	25.65	0.00	18.62	6	58.33	
13	65	14.67	25.65	0.00	15.88	7	58.38	
14	70	15.78	25.65	0.00	14.08	9	58.51	
15	75	17.19	25.65	0.00	11.89	10	58.79	
16	80	18.29	25.65	0.00	10.25	13	59.10	
17	85	19.68	25.65	0.00	8.29	15	59.57	
18	90	21.92	25.65	0.00	5.16	16	60.51	
19	95	24.12	25.65	0.00	2.11	18	61.66	

## 6 Conclusion

Big Data has emerged as a paradigm for extracting insights from huge amounts of data. However, data quality is considered a key for its acceptance and adoption as a poor data quality might severely affect the Big data analysis results. In this paper, we identified the key research challenges in evaluating Big data quality. We proposed a Big data profiling model to cope with data quality in an early stage of the Big Data lifecycle by providing a set of actions to be implemented in the pre-processing phase to ensure a high-quality related dataset. The set of experimentations we have conducted have validated key features of our BDQPM model including sampling and profiling, sampling and quality evaluation, and exploratory quality profiling. The results we have obtained demonstrated that quality profiling strengthens the efficiency of the pre-processing and processing phases of Big Data. Finally, we are planning to extend our BDQPM to cover and discover quality profiling rules for unstructured data, that represent almost 80% of the overall Big data, where businesses and companies are highly interested to explore and get valuable insights.

## References

1. Abedjan, Z.: An introduction to data profiling. In: Zimányi, E. (ed.) eBISS 2017. LNBP, vol. 324, pp. 1–20. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-96655-7\\_1](https://doi.org/10.1007/978-3-319-96655-7_1)
2. Abedjan, Z.: Data profiling. In: Sakr, S., Zomaya, A. (eds.) Encyclopedia of Big Data Technologies, pp. 563–568. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-77525-8\\_8](https://doi.org/10.1007/978-3-319-77525-8_8)
3. Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A.S., Buyya, R.: Big data computing and clouds: Trends and future directions. *J. Parallel Distrib. Comput.* **79**(C), 3–15 (2015). <https://doi.org/10.1016/j.jpdc.2014.08.003>
4. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **41**, 1–52 (2009)
5. Chester, J.: Cookie wars: how new data profiling and targeting techniques threaten citizens and consumers in the “Big Data” era. In: Gutwirth, S., Leenes, R., De Hert, P., Poullet, Y. (eds.) European Data Protection: in Good Health, pp. 53–77. Springer, Dordrecht (2012). [https://doi.org/10.1007/978-94-007-2903-2\\_4](https://doi.org/10.1007/978-94-007-2903-2_4)
6. Dai, W., Wardlaw, I., Cui, Yu., Mehdi, K., Li, Y., Long, J.: Data profiling technology of data governance regarding big data: review and rethinking. *Information Technology: New Generations. AISC*, vol. 448, pp. 439–450. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-32467-8\\_39](https://doi.org/10.1007/978-3-319-32467-8_39)
7. Naumann, F.: Big Data Profiling (2014)
8. Géczy, P.: Big data characteristics. *The Macrotheme Review* **3**, 94–104 (2014)
9. Glowalla, P., Balazy, P., Basten, D., Sunyaev, A.: Process-driven data quality management – an application of the combined conceptual life cycle model. Presented at the 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 4700–4709 (2014). <https://doi.org/10.1109/HICSS.2014.575>
10. Gu, X., et al.: Profiling Web users using big data. *Soc. Netw. Anal. Min.* **8**, 24 (2018). <https://doi.org/10.1007/s13278-018-0495-0>



11. Hasan, O., Habegger, B., Brunie, L., Bennani, N., Damiani, E.: A discussion of privacy challenges in user profiling with big data techniques: the EEXCESS use case. In: *BigDataCongress*, pp. 25–30 (2013)
12. Eembi, N.B.C., Ishak, I.B., Sidi, F., Affendey, L.S., Mamat, A.: A systematic review on the profiling of digital news portal for big data veracity. *Proc. Comput. Sci.* **72**, 390–397 (2015)
13. Johnson, T.: Data profiling. In: Liu, L., Özsu, M.T. (eds.) *Encyclopedia of Database Systems*, pp. 808–812. Springer, New York (2018). <https://doi.org/10.1007/978-1-4614-8265-9>
14. Loshin, D.: *Rapid Data Quality Assessment Using Data Profiling*, vol. 15 (2010)
15. Maier, M., Serebrenik, A., Vanderfeesten, I.T.P.: *Towards a Big Data Reference Architecture*. University of Eindhoven (2013)
16. McNeil, B.J., Pedersen, S.H., Gatsonis, C.: Current issues in profiling quality of care. *Inquiry* **29**, 298–307 (1992)
17. Naumann, F.: Data profiling revisited. *ACM SIGMOD Rec.* **42**, 40–49 (2014)
18. Oliveira, P., Rodrigues, F., Henriques, P.R.: A formal definition of data quality problems. In: *IQ* (2005)
19. Prabha, M.S., Sarojini, B.: Survey on Big Data and Cloud Computing, pp. 119–122. *IEEE* (2017)
20. Sidi, F., Shariat Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A.: Data quality: a survey of data quality dimensions. In: *CAMP 2012*, pp 300–304 (2012)
21. Talwalkar AKA *The Big Data Bootstrap*. 20
22. Sun, Z.: 10 Bigs: Big Data and Its Ten Big Characteristics (2018). <https://doi.org/10.13140/rg.2.2.31449.62566>