Towards a Powerful Solution for Data Accuracy Assessment in the Big Data Context

Mohamed TALHA[1], Nabil ELMARZOUQI[2], Anas ABOU EL KALAM[3]
ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco[1]
ENSET Rabat, Mohammed V University in Rabat, Rabat, Morocco[2]
ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco[3]

Abstract—**Data Accuracy is one of the main dimensions of Data Quality; it measures the degree to which data are correct. Knowing the accuracy of an organization's data reflects the level of reliability it can assign to them in decision-making processes. Measuring data accuracy in Big Data environment is a process that involves comparing data to assess with some "reference data" considered by the system to be correct. However, such a process can be complex or even impossible in the absence of appropriate reference data. In this paper, we focus on this problem and propose an approach to obtain the reference data thanks to the emergence of Big Data technologies. Our approach is based on the upstream selection of a set of criteria that we define as "Accuracy Criteria". We use furthermore a set of techniques such as Big Data Sampling, Schema Matching, Record Linkage, and Similarity Measurement. The proposed model and experiment results allow us to be more confident in the importance of data quality assessment solution and the configuration of the accuracy criteria to automate the selection of reference data in a Data Lake.**

Keywords—Big data; data quality; data accuracy assessment; big data sampling; schema matching; record linkage; similarity measurement

I. Introduction

Data Quality is an essential topic for any organization to get accurate data and to make good decisions accordingly. Concerns about this topic can be addressed in three ways: data quality evaluation, data quality improvement and data protection [1]. In this work, we will focus on the evaluation of data quality. Different dimensions characterize the quality of the data but there is no general agreement on the complete list of these dimensions and their exact meanings [2]. The most studied dimensions in literature are Accuracy, Completeness, Currency and Consistency [3], [4]. Several studies have identified Accuracy as the key dimension of Data Quality [1], [5], [6], [7] and the hardest to assess [8]. Knowing beforehand the Accuracy of their data brings many benefits to organizations. Indeed, the decision-making process is improved when we know the degree of confidence that can be attributed to the data. On the other hand, inaccurate data can lead to erroneous conclusions and can significantly compromise a range of decision-making processes which can lead to lost opportunities, lost revenue, strategic mistakes, etc. Thus, evaluating data accuracy is a process that requires planning before any data exploitation. It is with these elements in mind that we will devote this work to the evaluation of data accuracy.

Many works in literature attempt to find solutions to evaluate the accuracy of data. Not all, but most require a key step of comparing the data to evaluate with the correct data without giving sufficient details on how to obtain this reference data. Today, thanks to the emergence of Big Data, Cloud Computing and IoT, organizations can more easily collect, store and manage very large volumes of data. In this article, we will offer a solution to obtain reference data in a Big Data environment. After exposing an overview of the related work, we will present a state of the art about data accuracy in order to deduce a clear definition that will be our basis for this work. We will then present a set of techniques and concepts necessary for the implementation of our solution. Next, we will detail our model that we will apply on a case study in order to experiment our approach. The analysis of the results will allow us to deduce a set of very interesting findings to successfully implement our solution for any other use case. We will end this paper with a conclusion and a glimpse into future work.

II. Related Work

Many studies are interested in evaluating the quality of data, particularly looking into the accuracy of the data. The authors of [8] propose a data accuracy assessment tool based on a collection of datasets and three different phases: training, record linkage and accuracy assessment. The idea is to be able to identify the reference data which will then serve as a basis for the comparison process. In their approach, they use machine learning techniques to choose, from the datasets already present in the lake of an organization, the closest dataset that they deem correct. This assumption can be correct if it is guaranteed that the data present in the lake are correct and up to date, which is generally not the case; the new data collected may even be of better quality in some cases. In addition, the correct information can exist in more than one dataset, not just one as they assume. In this case, it will be necessary to manage many aspects such as the heterogeneity of the schemas, the choice of the dataset offering the best quality for a particular data perimeter, the lack of correspondence for certain data, etc. Moreover, the use of Google’s Word2Vec word embedding as a basis for data training can slow down the assessment process. The idea of using word embedding to determine the closest dataset remains logical but it will be necessary to test its performance in an environment that hosts very large volumes of data.

Another work done by Taleb et al. [9] in which they propose a system for assessing the accuracy, completeness and consistency of Big Data based primarily on data sampling. The adopted principle is to create a set of samples, without replacement, from the original dataset. Then, from each sample created, generate a set of samples using the BLB (Bag of Little Bootstraps) resampling technique. For each sample thus generated, a data profiling process is applied to extract descriptive information from the data such as the description of the data format, the different attributes, their types and their values, the possible constraints, the ranges of the authorized values, etc. All this information obtained through the profiling process is then used to select the appropriate metrics for each dimension before proceeding with their evaluation. For the accuracy dimension, a metric can be defined to satisfy a certain number of constraints related to the type of data such as a zip code, an email, a social security number, or an address. For example, an attribute can be defined as a range of values between 0 and 100, otherwise it is incorrect. The accuracy of the attribute is then calculated based on the number of correct values divided by the number of observations or rows. The authors of this article have only dealt with syntactic accuracy which is much simpler to verify than semantic accuracy as we will see a little further.

One last interesting work we wish to present concerns the evaluation of the quality of unstructured data. In [19], the authors are interested in evaluating the quality of data collected from social networks through the integration of a metadata management system into Big Data architecture. In their approach, the authors distinguish five groups of metadata:

- Navigational metadata used to identify the location of each dataset.
- Process metadata used to describe the source and the processing performed on each dataset.
- Descriptive metadata consists of business metadata that describes the meaning of a dataset from a business perspective, and technical metadata that provides technical information about the dataset such as data size, content description, data creator, data type and format of content, etc.
- Quality metadata including dimensions and metrics used to describe the quality of the data.
- Administrative metadata used to describe the data provider, applicable licenses and access rights on the datasets, the copyright holder and the data privacy level indicator, etc.

The use of metadata to evaluate the quality of unstructured data seems to be a good solution especially when it is difficult, if not impossible, to compare these data with data that represent the real world. However, managing metadata could be a very expensive and complex process especially for quality metadata. The high volume and velocity of Big Data are real challenges to overcome. The use of metadata in conjunction with other techniques of comparison with correct data seems to us to be more efficient.

There are many other works in literature that are concerned with assessing the accuracy of data. For example, in [1], Motro and Rakov present a solution for assessing the accuracy and completeness of databases using Set Theory. Redman, for its part, provides in [6] a framework for assessing the accuracy of data based on four factors, namely where to take measurements, the choice of data to include, the measurement device and the scales on which the results are reported.

III. Data Accuracy in Literature

A. Definition

A multitude of definitions for data accuracy exist in literature. Each definition involves aspects of the context in which it was given. Generally, there is a reasonable consensus that the accuracy of the data is linked to a specific concept, namely the magnitude of an error [10]. For Ballou and Pazer [11], data are accurate when their values stored in a database correspond to the real values. Authors in [12], [13], [14] and [15] link the accuracy of data to the percentage of objects that do not contain errors in the data such as misspellings, values outside the allowed range, and so on. Several works [4], [5], [7], [16] define accuracy as a measure of the proximity of a given value ν to another value ν' considered to be correct.

Furthermore, ISO [17] and many studies [2], [5], [16] distinguish between syntactic accuracy and semantic accuracy. Each of them presents a particular aspect of the accuracy and has its own metrics. Syntactic accuracy is defined as the proximity of data values to a set of defined values in a domain considered to be syntactically correct. It concerns the structure of the data [18] and expresses the degree of syntactic error-freeness [14]. Semantic accuracy, as for it, is defined as the proximity of data values to a set of defined values in a domain considered to be semantically correct. It represents the correctness and the degree of validity of the data [12], [14]. It describes the extent to which data represent real-world conditions [18].

Although they diverge on some particularities, all of these definitions implicitly or explicitly imply a comparison between the data of a system and the real world. Therefore, we adopt the following definition: Accuracy reflects the degree of correctness at which the data in an information system represents the real world. More formally, let v be the value of a datum in an information system and ν' the corresponding reference value considered as correct; the accuracy of ν represents the degree of similarity between ν and ν'.

Whether for structured or semi-structured data, comparing the values of the data with those of the real world allows us to deduce the degree of accuracy. However, this definition does not seem well-suited to unstructured data such as files with free text (tweets, studies, personal reports, etc.), multimedia files (image, audio or video), etc. Certainly unstructured data may contain information that can be compared with the real world (a person's photo, information about an object, information in a story, a mathematical formula, etc.) but the problem lies in the information that only concerns the people who gave it (personal impressions, opinions about a subject, intentions, desires, etc.). Unstructured data is more complex to evaluate and cannot be evaluated in the same way as structured or semi-structured data. If we take the example of a scientific paper, we cannot evaluate its quality by analyzing its content or by comparing the information it contains with reference data.

Instead, we must analyze some information attached to it, such as the opinions of the reviewers, the importance of the magazine in which it was published, the reputation of the authors and their institutional affiliations, the type of publisher (academic, commercial ...), the source of the article (peer-reviewed journals, unpublished articles …), etc. and anything else that is relative to it. Thus, the evaluation of unstructured data quality will be more relevant if it concerns the information relating to the data rather than the data themselves. In [19], confirming our hypothesis, the authors present a solution to evaluate the quality of data collected from social networks by integrating a metadata management system in the Big Data life cycle. For this reason, and in order to limit the scope of this work, we will remain focused on structured and semi-structured data.

B. Reference Data

According to Redman [6], it is impossible to tell by direct examination if a data value is correct; all measurements of data accuracy must refer to human knowledge, other reference data or the real world. Comparing the data values with real-world values makes the measurement of their accuracy complex and costly because, very often, these real values are unknown [3], [20] or are hypothetical, really unavailable [1]. The degree of complexity changes according to the type of accuracy to assess; syntactic accuracy is usually simpler than semantic accuracy. Indeed, it can be verified by comparing the data values with reference dictionaries such as name dictionaries, domain dictionaries (list of product categories, commercial categories ...), address list, range of values, etc. On the other hand, semantic accuracy is more complex to measure because the terms of comparison must be derived from the real world, which is almost always costly [6].

One systematic way to verify semantic accuracy when multiple data sources are available is to compare information about the same instance stored in different locations. According to [3], a typical process for checking semantic accuracy consists of two phases: a searching phase and a matching phase. The first one is to identify the matching instances, while the second one is to make a decision on correspondence, non-correspondence or possible correspondence. Different criteria can be applied to make the comparisons. Generally, the values are considered correct if they come from a reliable source. In some cases, a data expert may be required to estimate the accuracy.

Moreover, when collecting data, there is no guarantee that the information collected is accurate. Today, to check the accuracy of the information provided by the consumers of a service, new methods are used such as automatically sending a secret code by mail to check a postal address, by email to check an email address, by SMS or voice call to check a phone number, etc. These methods, while effective in data collection, cannot be applied to estimate the accuracy of data already collected because of their high cost and the time it may take. They are therefore not suitable for checking the timeliness of information. In practice, the comparison is made with data collected from a reference source considered sufficiently reliable [21]. When multiple data sources provide the same types of information, the most reliable ones can be considered as data references for comparisons. The reliability of data sources can be determined through the trust and the reputation of the information provider. Other strategies include considering other quality factors to determine the most reliable source of data, for example, data consistency [4].

Reference data can be obtained through different mechanisms such as:

- Identification: we can assign to each data set available in an organization's information system a reliability level. Data sets with a high level of reliability can be used as reference data during the accuracy evaluation process.
- Collection: if reference data do not exist, some reference dictionaries such as addresses (postal codes, city names and codes, streets, ...), product catalogs, lists of names and surnames, the possible values for certain fields (diplomas, professional activities, ...), etc. can be collected independently and serve as reference data for checking syntactic inaccuracies. Business information can also be collected from external reliable providers to update data or fill in missing values.
- Correction: obtaining reference data is possible thanks to the improvement of the data quality of an organization. This can be done by implementing different techniques such as data cleaning, updating obsolete data values, correcting incorrect values, etc.

Reference data represent then the reality and make it possible to evaluate the accuracy of others data by calculating their degree of similarity. We consider this as the basic element for any process of evaluating the accuracy of structured and semi-structured data. In the next section, we outline the main accuracy criteria that allow us to identify reference data.

C. Accuracy Criteria

In this section, through examples, we present some of the criteria that data and their providers must meet in order to consider them as reference data.

The very first criterion one can think of is reliability in the source of the data. If the data are collected from a competent source that we are sure will provide correct data, we can consider them as reference data. As an example, to verify someone’s personal identity information, the best solution is to compare them with the data of the civil registry office of his/her country. However, this operation can be more complex for other cases. For example, to verify the validity of the diplomas declared by a person, the ideal is to validate this information with the institutions having issued these diplomas which can be a very difficult or impossible task. It would then be easier to verify it with an organization providing this service and having a good reputation for doing so. We deduce from these two examples that the Trust and the Reputation of data providers are two key criteria for identifying reference data.

Now let us take another illustrative example using financial market data which has instrument values that change continuously. When making a financial decision, traders usually rely on market data prepared by the internal departments of their organizations. Financial institutions, during orders validation process, require a crucial step of validating the data upon which traders make their decisions, referring to up-to-date data acquired from third-party organizations such as Bloomberg, Thomson Reuters, CQG, etc. who are specialized in this field and guarantee a real-time update of their data. The third criterion to add to our list is the Timeliness of the data. This time it is a criterion that relates to the data itself and not the data providers.

Lastly, in Big Data environments, many data providers can feed an organization's data lake with information about subjects that may be redundant or contradictory. To determine the most reliable source, we can of course refer to the criteria mentioned in the examples above, but if this is not possible, we can take into account other quality factors such as data Consistency [4].

We understand that different criteria related to the data providers or the data themselves can be used to identify the reference data. Among others, these criteria include:

- Trust: this criterion plays a central role in assessing the quality of information [22]. It reflects the reliability and the trustworthiness of the provider. We define a trusted source as a competent data source in a particular field that can provide accurate data.
- Reputation: in the field of IT security, different approaches exist to build trust models, among which are those based on reputation. These models consider interactions and past experiences between entities [23]. We define reputation as a measure that reflects public opinion about the data reliability of a source of information. This value evolves over time based on people's experiences with the source.
- Timeliness: data timeliness can be a fundamental criterion in some contexts as illustrated in the previous example. For Fox et al. [24], a datum is said to be current or up-to-date at time t if it is correct at time t and is out-of-date at time t if it is incorrect at t but was correct at some moment preceding t. So to be up-to-date is to be correct right now and to be out-of-date is a special case of inaccuracy; an inaccuracy caused by a change in time. Timeliness reflects the mechanisms and processes put in place by the data provider to refresh and update their data in real time.
- Consistency: for Rafique et al. [25], consistency represents the degree to which information has attributes that are free from contradiction and are coherent with other information in a specific context of use. A consistency error would be that a 5-year-old child has a "married" marital status (semantic error) or postal codes that are not within an allowed range (syntactic error). Consistency of data indicates whether the logical relationship between correlated data is correct and complete [26]. We can then use consistency as a criterion that justifies the accuracy of the data [4].

The accuracy criteria presented so far can be used as indicators to determine reference data. Obviously, the list is not exhaustive and the choice of accuracy criteria is strongly dependent on the context of application.

In many cases, a single criterion would not be sufficient and the combination of two or more accuracy criteria would be necessary. If we take the example of financial market data, it may happen that two providers diverge on a particular datum, which requires, in addition to the timeliness, to take into account others accuracy criteria such as trust and/or reputation.

In the case where several accuracy criteria are pooled together to identify the reference data, the ranking of these criteria in order of importance would be mandatory. For this, it will be necessary to assign to each criterion a weight which represents its importance during the resolution of the possible conflicts. But before that, the values should be normalized so that all the criteria are represented by the same unit of measure. For example, if for a given case, three accuracy criteria are used, the trust represented by a binary value (0 or 1), the reputation represented on a scale (from 1 to 5) and the consistency represented by a percentage (between 0 and 100), the pooling together of the three criteria requires the normalization of the values using a mathematical technique such as Min-Max Scaling defined by the following formula:

$$X_{normal} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

With:

- X: represents the current value of the criterion.
- X_{min}: represents the minimum value that the criterion can have.
- X_{max}: represents the maximum value that the criterion can have.

This formula allows us to transform values with heterogeneous units into values in a range $[0, 1]$ which then enables us to apply comparative and analytical studies to different accuracy criteria.

D. Metrics

Quality evaluation can be quantitative or qualitative [19]. Quantitative evaluation is a systematic and formal process. It relies on the existing knowledge of an organization and applies computational methods as the result of a condition, a mathematical equation, an aggregation formula, etc. to reach the values of objective metrics. The results of the quantitative evaluation are therefore objective and more concrete than in the case of a qualitative evaluation. The latter is based on subjective metrics, measuring the perceptions and experiences of the stakeholders. They are generally carried out by data administrators or users through satisfaction questionnaires, user surveys, etc. [2], [14].

We denote by "data unit" the set of data belonging to a level of granularity. This is the basic element on which the accuracy assessment operations are applied: the finer the level of granularity, the longer is the calculation time, but the more precise the evaluation of the accuracy. Different metrics exist in literature to measure the accuracy of the data, among which we find:

- Boolean measure: this type of metrics takes Boolean value to indicate if a data unit is accurate or not.

- Degree measure: this metric, used to express the degree of confidence in data, is calculated by dividing the number of correct data units by the total number of data units.
- Distance measure: this is a numeric value that captures the distance between a data unit in the system and a reference data. Generally, this metric is calculated by the distance between the objects. The smaller the distance, the more similar the objects are.

In practice, determining what constitutes a unit of data and what is an error requires a set of clearly defined criteria [14] that depend on the context of each project. As an example, it is possible for an incorrect character in a text string to be tolerable in one circumstance but not in another.

Moreover, to assess the quality of data in a Big Data context, we often use the data sampling technique, which, as will be explained later, is extremely useful in circumventing the problem of large volumes of data. Quantitative measurement of accuracy requires the establishment of a set of aggregation functions. Let S be a sample of n data units to be evaluated and A_i the value of the accuracy of the i^{th} element in S. Inspired by [2], [14], [27] and [28], the following two typical aggregation functions are the most used:

- Ratio: this function measures the ratio of the number of correct data units in the sample, divided by the cardinality of the sample.

If we consider that the accuracy of the data units is expressed using a boolean measure $\mathrm{A_i} \in \{0,1\}$ with $1 \leq \mathrm{i} \leq \mathrm{n}$, then the accuracy of S is calculated as follows:

$$\text{Accuracy }(S) = \frac{|\{\mathrm{A_i}\}, \mathrm{A_i} = 1|}{\mathrm{n}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A_i}}{\mathrm{n}} \qquad (2)$$

$|\{\mathrm{A_i}\}, \mathrm{A_i} = 1|$ denotes the cardinality of data units with correct accuracy.

If we consider that the accuracy of the data units is expressed using measurements in degree or distance $A_i \in [0, 1]$ with $1 \leq i \leq n$, then it will be necessary to consider a threshold θ from which the data is considered as correct. In this case, the accuracy of S is calculated as follows:

$$Accuracy\ (S) = \frac{|\{A_i\}, A_i \geq \theta|}{n} \qquad (3)$$

$|\{A_i\}, A_i \geq \theta|$ denotes the cardinality of data units having accuracy greater than or equal to θ.

- Average: this function measures the average of the correct data units. Whatever the metric used, the accuracy of S is calculated as follows:

$$Accuracy\ (S) = \frac{\sum_{i=1}^{n} A_i}{n} \qquad (4)$$

If we consider that the data units do not all have the same importance, we can assign each unit a weight w_i and calculate the accuracy of S with the following method:

$$Accuracy\ (S) = \frac{\sum_{i=1}^{n} w_i A_i}{\sum_{i=1}^{n} w_i} \qquad (5)$$

IV. Comparison Techniques

In this section, we briefly present some techniques and concepts needed to understand our study, especially in performing data comparison in a Big Data environment such as Big Data Sampling, Schema Matching, Record Linkage and Similarity Measurement.

A. Big Data Sampling

The need for a quick response is sometimes more important than a precise answer, especially in the case of evaluating the accuracy of the data. Data Sampling is extremely useful for making Big Data usable for analysis [29]. To analyze large sets of data in order to assess their quality, one can be satisfied with the selection and analysis of a representative sample of all data units. For certain types of problems, sampling gives results as good as performing the same analysis using all the data [30], but for particular cases, especially in the analysis of large volumes of data, sampling seems to be the most appropriate solution [1], [20], [31].

As we presented in [32], to create a sample of a dataset, different techniques exist such as Simple Random Sampling, Stratified Sampling, Cluster Sampling, Multistage Sampling, Systematic Sampling, etc. Several techniques can be used together to create an effective sample, the main rules are that the sample must be representative of all data and all data units must have the same chance of being selected in the sample. Moreover, to know the size of the sample, it will be necessary to know in advance the size of the data to be sampled which is not easy to obtain in a Big Data project. To meet these constraints, there is an effective approach called "Reservoir Sampling" initially introduced by Vitter [33]. It’s a family of randomized algorithms that randomly select a sample of k elements from a large set of n-sized data or from a data stream of size n, where n is unknown or difficult to know. All elements have the same probability to be selected in the sample. The principle is: Let S be the set or the stream of data to be sampled. We start by creating a sample of the first k elements that will be called the Reservoir R and then, by sequential access on the rest of the elements of S, we randomly replace elements in R. Algorithm 1 is a typical example.

Algorithm 1: Example of Reservoir Sampling Algorithm

```
1. Let S be the data set or data stream to be sampled;
2. Create an empty array R of maximum size k;
3. Fill the array R by the first k elements of S;
4. For each element from position k+1 to the last element in S, repeat the following process:
4.1. Let i be the position of the current element in S;
4.2. Let j be a digit generated randomly between 0 and i;
4.3. If j < k, then replace R [j] by S [i];
```

The advantage of this algorithm is that it makes it possible to create a sample by crossing the data only once, as the sample is created, by sequential access, without having to know the size of the data to be sampled and guarantees that all elements have the same chance of being selected.

B. Schema Matching

The heterogeneity of data sources is a challenge that makes data manipulation processes such as data integration, data fusion, application interoperability, software reuse, etc. complex. Data accuracy assessment also experiences this challenge, especially when comparing the data to be evaluated with those representing the real world. The heterogeneity between these data requires matching their schemas.

Bernstein et al. [34] define a schema as a formal structure that represents an engineered artifact, such as a SQL schema, XML schema, entity-relationship diagram, ontology description, interface definition, or form definition. They also define a correspondence as a relationship between one or more elements of one schema and one or more elements of another. In relational databases, Schema Matching consists in linking the tables and columns of a database to those that represent the same concepts in another database. The authors in [34] and [35] present and detail a taxonomy of techniques and methods used to achieve the schema matching, among which we find:

- Linguistic matching based on an element’s name or description, using stemming, tokenization, string and substrings matching, and information retrieval techniques.
- Matching based on auxiliary information such as thesauri, acronyms, dictionaries, and lists of mismatches.
- Instance-based matching which considers that the elements of two schemas are similar if their instances are similar based on statistics, metadata, or trained classifiers.
- Structure-based matching that considers elements in two schemas to be similar if they appear in similarly-structured groups, have similar relationships, or have relationships to similar elements.

Sutanta et al. [36], with reference to others research work, carried out a comparative study based on 34 prototype models and diagrams corresponding to different aspects such as the input type, the methods used, the field of use as well as the existence of a graphical interface allowing users to adapt the results of the prototype. According to this study, one of the most successful schema mapping prototypes is COMA 3.0 [37], which is an evolution of COMA++ [38]. This prototype accepts different types of input data (XSD, XDR, OWL, CSV, SQL), uses different matching algorithms (Linguistic based, Structure based, Instance based), is not specific to a particular field of use and interactive via a GUI. As part of our work, we used this prototype to implement our case study.

C. Record Linkage

Record linkage consists of gathering information from two records that are assumed to be related to the same entity. This involves linking records within a single file or between two or more files to identify similar records. The challenge is to collect the records of the same individual entities by searching for exact matches [39]. Record linkage can be used when assessing data quality (to detect similarities between data), when improving data quality (including data cleaning and the deletion of duplicates processes), when merging data sets, etc. Two main types of record linkage exist: deterministic and probabilistic. Deterministic record linkage is a relatively straightforward method, which usually requires exact agreement on a match key, which may be a unique identifier (e.g. national identity number, social security number, etc.) or a collection of partial identifiers (e.g. a key consisting of full name, year of birth and the postal code of the city of birth). A record pair is considered as a link only if the match keys (unique identifier or partial identifier collection) are identical. Deterministic linkage is unfortunately not always obvious. Errors or lack of information in the records may exist. To overcome these limitations, probabilistic models have been proposed to determine the linkage in the presence of recording errors and/or without using the matching keys. Newcombe et al. [40] were the first to propose probabilistic methods, suggesting that a matching weight could be created to represent the probability that two records actually correspond given the agreement or disagreement on a set of partial identifiers.

D. Similarity Measurement

Different methods exist to measure the similarity between data. The choice of a method depends largely on the type of data that need to be compared (characters, strings, numbers, binary values, etc.) and the context of how it will be used (for example, it could be considered that two strings of characters are similar even if they have one or two different characters which cannot be the case for other situations).

To compare strings, many methods exist such as Levenshtein and Jaro-Winkler distances. The Levenshtein distance is defined as the minimum number of changes needed to convert one string to another. However, depending on the context of each project, adjustments may be necessary to adapt this method to specific needs (case sensitivity, accented and special characters, use of acronyms, etc.). Jaro-Winkler distance, for its part, measures the similarity between two strings by calculating the number of characters that they have in common. It’s a variant proposed by Winkler derived from the Jaro distance used in the field of record linkage for duplicates detection. Many other methods exist in these topics such as Cosine similarity, q-Gram, Damerau-Levenshtein, etc.

For numbers, the similarity can be calculated as the difference between values [41] taking into consideration a threshold from which one can consider that two numbers are similar. The choice of the threshold depends on the context of comparison and the order of magnitude of the numbers (a difference between two small numbers does not have the same impact as that between two very large numbers). Other types, such as dates and geographical coordinates, can follow the same principle since they are convertible into numbers by retrieving the timestamp dates and latitude/longitude geographical positions.

V. Big Data Accuracy Assessment

In this section, we will demonstrate our solution to evaluate the accuracy of data in a Big Data context. We will first present our model to understand the main steps of the evaluation process. Then, to prove our concept, we will present our case study as well as its implementation in a Big Data environment.

Finally, to go back to the objective of this research work, we will analyze the results of the study and present our findings in the form of a conclusion.

A. Assessment Data Accuracy Process

Our model, as shown in Fig. 1, consists of five steps:

1) Master data set: the first step is to continuously collect data from different data providers and store them in their raw state in the data lake of the Information System. Data providers can be external or internal services of the organization.

2) Golden data set: before implementing a data accuracy assessment solution, as explained in the state of the art, an organization will need to determine the accuracy criteria that will enable it to determine the quality of its source data. In this step, each data set of the Master Data Set should be assigned a value for each of the accuracy criteria. Since these values are likely to change over time, this pre-processing step will have to be recurrent.

3) Mapped data: this step corresponds to the schema matching between the data to be evaluated (Input Data) and those present in the Golden Data Set. If the desired level of granularity is finer (values or objects for example), this step will consist in linking the columns of the data to evaluate with their correspondents in the Golden Data Set. If $X\,\{x_1, \dots, x_n\}$ represents the set of columns to be evaluated and $Y\,\{y_1, \dots, y_m\}$ represents the set of columns of all datasets in the Golden Data Set, then we will have 3 scenarios:

- Simple scenario: for each column x_i corresponds one, and exactly one, column y_j.
- Conflict scenario: for each column x_i corresponds a set of columns $\{y_1, y_2, \dots, y_k\}$ of cardinality $k \in [2, p]$ where p is the number of datasets of the Golden Data Set. In case of conflict, it is the column that belongs to the dataset that best meets the required accuracy criteria that will be considered for the mapping operation.
- Incomplete scenario: there exists a set of columns $\{x_1, \dots, x_l\}$ of cardinality $l \in [1, \mathrm{n}]$ of which no element has a correspondence in Y. Note that the larger the l, the more accuracy calculation loses its reliability. If $l = n$ the calculation of the accuracy cannot be carried out because no reference data will be found.

4) Reference data: for each accuracy criterion, the organization will need to determine a threshold for a set of data to be considered sufficiently correct. This step consists firstly in eliminating all the data sets that do not meet the levels of accuracy criteria required by the organization and in resolving the various possible conflicts from the previous step. Then, and in order to get around the problem of the large volume of data hosted in the lake, a process of sampling the data may prove necessary. Finally comes the step of extracting records that need to be evaluated via a Record Linkage process. In this way, we will have dynamically constructed reference data whenever an assessment process is launched.

5) Data accuracy: the last step is to calculate the similarity between the related records. Depending on whether the granularity is about the objects or the values, it will be necessary to determine the good processes of computation of similarity. It will also be necessary to determine if all the columns are involved in the similarity calculation or only particular columns.

B. Proof of Concept

As part of this work, and to demonstrate the feasibility and reliability of our solution, we have put in place a proof of concept. For our case study, we are interested in evaluating the accuracy of the data concerning the railway stations in Paris and its suburbs. Our approach is to prepare a Data Lake from open data sources found on the Internet. Our data are collected from three open databases:

- The website data.iledefrance.fr: an open platform of public data concerning the Ile-de-France region. The platform is managed by the communication department of the Regional Council of Ile-de-France.
- The website data.ratp.fr: an open platform of public data concerning public transportation in the Paris region. This platform is managed by the RATP (Régie Autonome des Transports Parisiens) which is a public establishment of an industrial and commercial nature fully owned by the French government. It is a control unit that ensures the operation, maintenance and engineering of networks of part of public transport in Paris and its suburbs.
- The website data.sncf.com: an open platform for public data on railway transport in France. This platform is managed by the SNCF (Société Nationale des Chemins de Fer) which is a public establishment of industrial and commercial character fully owned by the French government. It is a board that manages the transport of passengers and goods and carries out the management, operation and maintenance of the railway network in France.

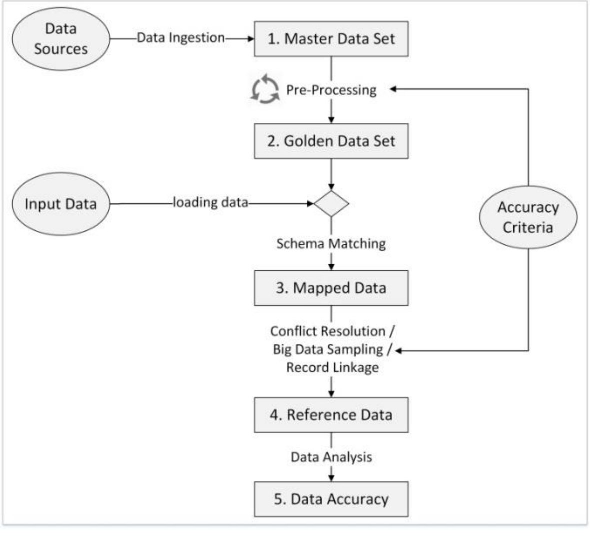

Fig. 1. Data Accuracy Assessment Process.

The very first step in implementing our solution is to define the accuracy criteria. Each organization is free to choose and define the criteria that suit it according to its activity, its projects and the nature of its data. For this case study, we will work on three criteria: Trust (T) represented by a percentage, Reputation (R) measured on a scale of 1 to 5 and Consistency (C) also represented by a percentage. The heterogeneity of the units of measurement is solved thanks to the normalization of the values by applying the Min-Max Scaling formula (1) as explained above.

C. Solution Setting up

We implemented our solution in six steps:

1) Big data platform: we have developed our solution on Hadoop 2.7.7 installed on Ubuntu 18.04.2 LTS 64-bit with an 8 GB RAM and 100 GB SATA disk. We have developed the different modules in python (version 3.6.7) and spark (version 2.3.1). We used HBase database (version 1.1.0) for metadata management.

2) Master data set: our Data Lake is composed of 3 Data Sets downloaded from the websites data.iledefrance.fr, data.ratp.fr and data.sncf.com. Data are stored in their raw state on HDFS (Hadoop Distributed File System).

3) Golden dataset: for each dataset, we assigned values to each of the three accuracy criteria (T, R, and C) as metadata. In a real life setting, these values must be calculated according to a well-defined approach. We will see at the end of this study that this stage is crucial. For this case study, the objective is to demonstrate the impact of accuracy criteria on the selection of reference data and, consequently, on the accuracy calculation reliability. We will then study several scenarios and in each scenario we will assign, for each of the three datasets, hypothetical values for each criterion to cover all possible cases.

4) Mapped data: this step consists in loading the data and selecting the columns to be evaluated and then matching each of them to those existing in the Golden Data Set. To achieve this step, we used the prototype COMA 3.0 [37].

5) Reference data: from the previous step, we were confronted with different situations. To perform schema matching, the evaluated column can be mapped to zero, one or many columns in the Golden Data Set. In case of mapping with multiple columns, a conflict exists and requires its resolution. For this, we have implemented a conflict resolution algorithm that consists in assigning a weight to each accuracy criterion and, in the case of a conflict, the Data Set with the highest weighted sum of the values of the accuracy criteria will be considered for the mapping. Once the mapping between the columns to evaluate and those of Data Lake is determined and all conflicts are resolved, we can select reference data through a record linkage process. We have implemented this mechanism using a Python's library called RecordLinkage [42], which provides indexing methods, similarity measurements, and classifiers.

6) Data accuracy: the record linkage result is a mapping table between records to be evaluated and reference records. All that remains now is the comparison of values. For this, the RecordLinkage library adopted for this study presents a class named *Compare* that compares the attributes of records while choosing the appropriate method for each type of data (character strings, numbers, geographic positions, etc.). By calculating the similarities of all the values of all records, we can deduce the accuracy for each column as well as for the entire table we want to evaluate.

D. Experiments

To justify the performance of our solution and the reliability of the results, we need to study the relationship between the accuracy criteria assigned to each dataset and the accuracy calculation result. The objective is to study the impact of the solution parameters to determine the best configuration that guarantees the most reliable result. For this, we have to execute a set of scenarios whose results are known beforehand and compare them with those calculated. To better understand our analysis approach, here is a use case:

"data.csv is a file that contains information about railway stations in Paris and its suburbs. We want to calculate the accuracy of this file by referring to data whose sources have a level of Trust ≥ 95%, with a Reputation ≥ 4.4/5 and whose Consistency ≥ 90 %."

To answer this use case, we must retrieve from our Data Lake all datasets that meet all the required accuracy criteria (i.e., after normalization, $T \geq 0.95$, $R \geq 0.85$ and $C \geq 0.90$). To get the reference data, the columns of the data.csv file must be matched with those of the selected data sets. A conflict resolution stage may be necessary. Then, the records will need to be matched through a Record Linkage process. Finally, to obtain the accuracy, it only remains to calculate the similarity between the matched records.

The reference data are extracted from datasets having at least the required values for each of the accuracy criteria. Since we have in our Data Lake three Data Sets; DS1 whose source is data.iledefrance.fr, DS2 whose source is data.ratp.fr and DS3 whose source is data.sncf.com, we can then distinguish between three groups of scenarios:

- Group A: for each scenario in this group (Scenario 1 – Scenario 3 of Table I), a single Data Set holds the maximum values for the three accuracy criteria (T, R, and C).
- Group B: for each scenario in this group (Scenario 4 – Scenario 21 of Table I), a Date Set holds the maximum values for two accuracy criteria and the maximum value of the third criterion is held by another Data Set. In this group we have 18 possible scenarios.
- Group C: for each scenario in this group (Scenario 22 – Scenario 27 of Table I), a Data Set can only have one maximum value for one of the three accuracy criteria. In this group we have 6 possible scenarios.

TABLE. I. CASE STUDY AND SCENARIOS

	Scenarios	***DS1***			***DS2***			***DS3***		
		T	***R***	***C***	***T***	***R***	***C***	***T***	***R***	***C***
Group A	Scenario 1	X	X	X	-	-	-	-	-	-
	Scenario 2	-	-	-	X	X	X	-	-	-
	Scenario 3	-	-	-	-	-	-	X	X	X
Group B	Scenario 4	X	X	-	-	-	X	-	-	-
	Scenario 5	X	X	-	-	-	-	-	-	X
	Scenario 6	-	-	X	X	X	-	-	-	-
	Scenario 7	-	-	-	X	X	-	-	-	X
	Scenario 8	-	-	X	-	-	-	X	X	-
	Scenario 9	-	-	-	-	-	X	X	X	-
	Scenario 10	X	-	X	-	X	-	-	-	-
	Scenario 11	X	-	X	-	-	-	-	X	-
	Scenario 12	-	X	-	X	-	X	-	-	-
	Scenario 13	-	-	-	X	-	X	-	X	-
	Scenario 14	-	X	-	-	-	-	X	-	X
	Scenario 15	-	-	-	-	X	-	X	-	X
	Scenario 16	-	X	X	X	-	-	-	-	-
	Scenario 17	-	X	X	-	-	-	X	-	-
	Scenario 18	X	-	-	-	X	X	-	-	-
	Scenario 19	-	-	-	-	X	X	X	-	-
	Scenario 20	X	-	-	-	-	-	-	X	X
	Scenario 21	-	-	-	X	-	-	-	X	X
Group C	Scenario 22	X	-	-	-	X	-	-	-	X
	Scenario 23	X	-	-	-	-	X	-	X	-
	Scenario 24	-	X	-	X	-	-	-	-	X
	Scenario 25	-	X	-	-	-	X	X	-	-
	Scenario 26	-	-	X	X	-	-	-	X	-
	Scenario 27	-	-	X	-	X	-	X	-	-

A special case not covered by any of the previous scenarios is the case where no Data Set satisfies all the accuracy criteria required by a use case. To be able to study the behavior for this particular case, we will assign to each criterion values close to, but less than, the maximum possible value.

For each scenario, we calculated the accuracy for all possible cases for the accuracy criteria, that is, for {T, R, C} ranging from {0.00, 0.00, 0.00} to {1.00, 1.00, 1.00}. Fig. 2, Fig. 3 and Fig. 4 show respectively the results of Scenario 1 of Group A, Scenario 4 of Group B and Scenario 22 of Group C (the first scenario of each group). The results of the other scenarios follow the same logic of those in the same group. For each scenario, we have 9261 iterations (for each variable T, R and C, from 0.00 to 1.00 with a step of 0.05, we have 21 iterations, and $21^3 = 9261$). By analyzing all the iterations of all the scenarios, we find that the accuracy can take on one of four values:

- 50%: when reference data are extracted from DS1.
- 83.14%: when reference data are extracted from DS2.
- 33.33%: when reference data are extracted from DS3.
- None: If none of the three data sets meets the criteria required for a given iteration, the accuracy value cannot be calculated for this iteration; our implementation returns then the value "None".

The analysis of the results of the different scenarios allows us to deduce that:

- For Group A, the value of the accuracy depends on the reference data set which holds the maximum of the values of the accuracy criteria. If for certain iteration no data set satisfies the criteria required, the value of the accuracy is None.
- For Group B, two values of the accuracy are possible and depend on the two data sets sharing the maximum values of the three accuracy criteria. If however no data set meets the criteria required by certain iteration, the value of the accuracy is None.
- For Group C, three values of accuracy are possible. For each iteration, the conflict resolution mechanism determines the data set that will be the source of the reference data. If no data set meets the criteria required by the current iteration, the value of the accuracy is None.

Fig. 5, Fig. 6 and Fig. 7 show the distribution of values as well as the execution time for each group. The analysis of these diagrams allows us to deduce that the smaller the number of Data Sets holding the maximum values of the accuracy criteria, the less we have None values, and the more the calculation of the accuracy is reliable. On the other hand, the execution time is longer. This is explained by the large number of data involved in the process of record linkage, reference data extraction and similarity measurement.

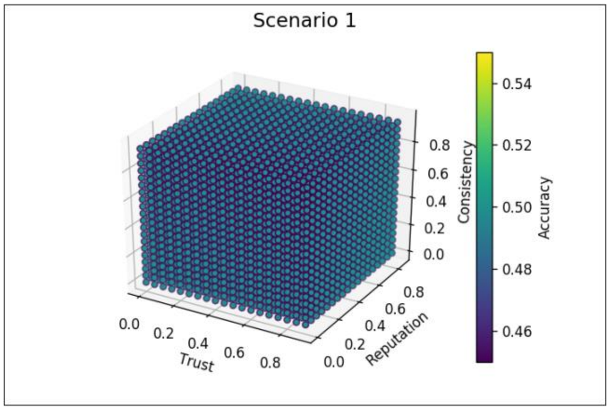

Fig. 2. Accuracy Calculation for Scenario 1 (Group A).

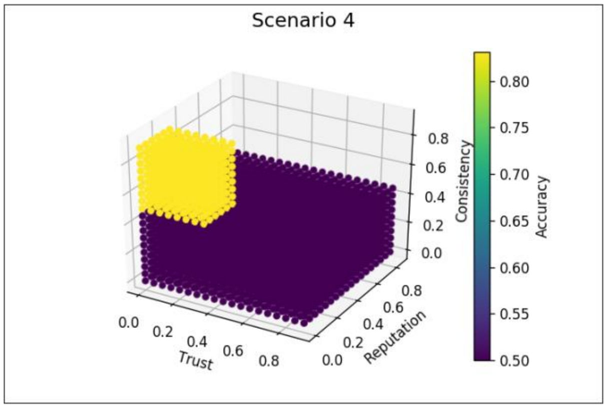

Fig. 3. Accuracy Calculation for Scenario 4 (Group B).

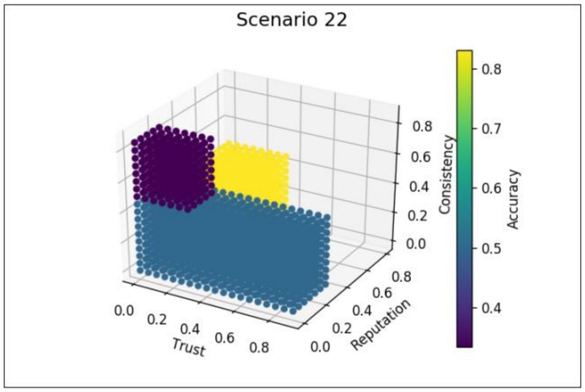

Fig. 4. Accuracy Calculation for Scenario 22 (Group C).

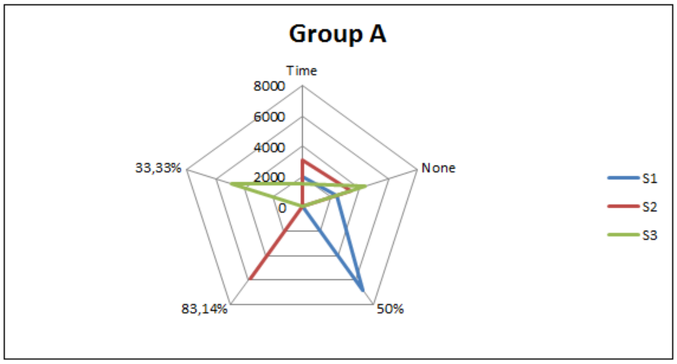

Fig. 5. Accuracy Assessment Metrics for Group A.

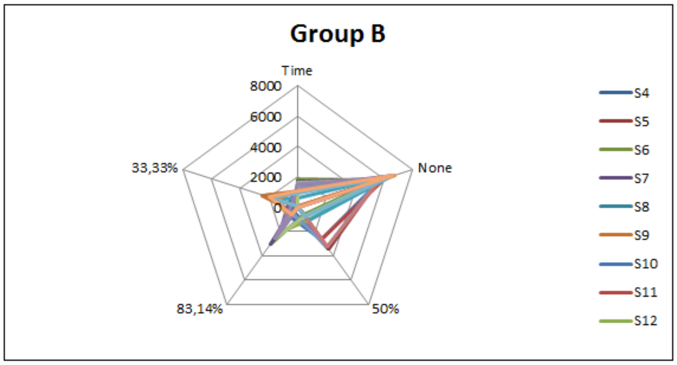

Fig. 6. Accuracy Assessment Metrics for Group B.

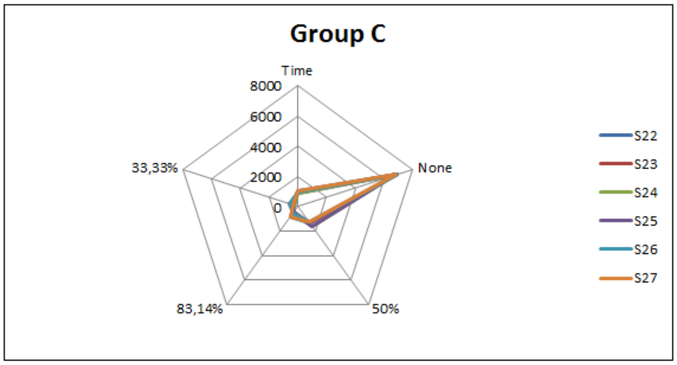

Fig. 7. Accuracy Assessment Metrics for Group C.

VI. Conclusions

In this work, we have highlighted a topical problem, namely the quality of data in Big Data through the evaluation of Data Accuracy. Whether syntactic or semantic, the evaluation of data accuracy requires their comparison with correct data called reference data. Obtaining such reference data is a very complex process and requires the establishment of a prior study to identify the quality criteria that measure the reliability of the data and their sources. We have proposed a solution allowing the configuration of the accuracy criteria in order to automate the selection of reference data in a Data Lake. Our study allows us to deduce that the implementation of a Big Data Accuracy Assessment System depends on several elements mainly related to the context of each project. The main steps to set up such a system are:

1) Having a data lake with data of good quality: The organization's Data Lake is the only source of reference data. The better the data lake, the more accurate the reference data.

2) Defining the right accuracy criteria that best characterize the notion of "data of good quality": For our case, we considered Trust, Reputation and Consistency of data sources as accuracy criteria. For other projects, other criteria may be more relevant such as Timeliness of the data. These criteria must be clearly measurable and assigned to each Data Set before initiating the assessment process.

3) Implementing the solution: For our case study, we have developed a demonstrator that exactly meets our needs in order to justify the reliability of our model. It is quite possible to develop a more generic application to define and manage the accuracy criteria used, to automate the mapping, to model the conflict resolution rules, etc.

VII. Future Work

As we have detailed in [43], confidentiality involves setting up a set of rules and restrictions to limit access to confidential data. It is generally handled with access control and cryptographic mechanisms. However, data quality assessment requires read access to the whole data. As for improving data quality, it requires write access to the data. We can therefore deduce that data security can make data quality management processes slower, more complex or even impossible. For our data quality assessment solution, we assumed that all datasets in the Data Lake are accessible, which cannot be the case in a professional setting in which data are often protected by different mechanisms and security policies even if they are hosted within the same system. Our next work will focus on this issue. We will work on implementing an effective solution to access all data without compromising their security. Our goal is to implement a data quality assessment solution in a Big Data context without compromising data security and without it being a barrier.

References

[1] A. Motro and I. Rakov, "Estimating the Quality of Databases," Springer-Verlag Berlin Heidelberg, 1998, 298-307, 10.1007/BFb0056011.

[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," ACM Computing Surveys 2009, 41, 10.1145/1541880.1541883.

[3] M. Fugini, M. Mecella, P. Plebani, and M. Scannapieco, “Data Quality in Cooperative Web Information Systems,” Kluwer Academic Publishers, Netherlands (2002).

[4] T. Redman, “Data Quality for the Information Age,” Artech House 1996, isbn:0890068836.

[5] C. Batini and M. Scannapieco, “Data Quality: Concepts, Methodologies and Techniques,” Springer-Verlag Berlin Heidelberg, 2006, isbn:978-3-540-33172-8, 10.1007/3-540-33173-5.

[6] T. Redman, “Measuring data accuracy: A framework and review, Information Quality,” London and New York: Taylor & francis group 2005, 0-7656-1133-3, 21-36, isbn:9781317467991.

[7] Y. Wand and R. Wang, “Anchoring Data Quality Dimensions in Ontological Foundations,” Commun, ACM 1996, 39, 86-95, 10.1145/240455.240479.

[8] G. Mylavarapu, J.P. Thomas, and K.A. Viswanathan, “An Automated Big Data Accuracy Assessment Tool,” IEEE 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, 2019, 193-197, 10.1109/ICBDA.2019.8713218.

[9] I. Taleb, H. El Kassabi, M. Serhani, R. Dssouli, and C. Bouhaddioui, “Big Data Quality: A Quality Dimensions Evaluation,” Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress 2016, 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122.

[10] T. Haegemans, M. Snoeck, and W. Lemahieu, “Towards a Precise Definition of Data Accuracy and a Justification for its Measure,” Proceedings of the International Conference on Information Quality (ICIQ), 2016, Article 16.

[11] D. Ballou and H. Pazer, “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” Management Science, February 1985, 31(2):150, doi:10.1287/mnsc.31.2.150.

[12] R. Wang and D. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” Journal of Management Information Systems 1996, 12, 5-33, 10.1080/07421222.1996.11518099.

[13] F. Naumann, U. Leser, and J.C. Freytag, “Quality-driven Integration of Heterogenous Information Systems,” Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999, 447-458.

[14] L. Pipino, Y. Lee, and R. Wang, “Data Quality Assessment,” Communications of the ACM 2003, 45, 10.1145/505248.506010.

[15] D. Loshin, “Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph,” San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2013, isbn:9780124186644.

[16] M. Scannapieco, P. Missier, and C. Batini, “Data Quality at a Glance,” Datenbank-Spektrum 2005, 14, 6-14.

[17] International Organization for Standarization ISO/IEC 25012, “Report: Software product quality requirements and evaluation (SQuaRE) - Data quality model,” 2006, Source : JTC 1/SC7 WG06.

[18] G. Shanks and B. Corbitt, “Understanding data quality: Social and cultural aspects,” Proceedings of the 10th Australasian Conference on Information Systems 1999.

[19] A. Immonen, P. Pääkkönen, and E. Ovaska, “Evaluating the Quality of Social Media Data in Big Data Architecture,” IEEE Access 2015, 3, 1-1, 10.1109/ACCESS.2015.2490723.

[20] T. Dasu and T. Johnson, “Exploratory Data Mining and Data Cleaning,” John Wiley & Sons 2003, Canada, isbn:0471268518, 10.1002/0471448354.

[21] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, and P. Angeletti, “Improving Data Quality in Practice: A Case Study in the Italian Public Administration,” Distributed and Parallel Databases (2003), 13, 135-160, 10.1023/A:1021548024224.

[22] M. Gamble and C. Goble, “Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model,” Proceedings of the 3rd International Web Science Conference, WebSci 2011, 10.1145/2527031.2527048.

[23] H. Hussain, O. Hussain, and E. Chang, “An overview of the interpretations of trust and reputation,” IEEE International Conference on Emerging Technologies and Factory Automation, 2007, ETFA, 826-830, 10.1109/EFTA.2007.4416865.

[24] C. Fox, A. Levitin, and T. Redman, “The notion of data and its quality dimensions,” Information Processing & Management 1994, 30, 9-19, 10.1016/0306-4573(94)90020-5.

[25] I. Rafique, P. Lew, M.Q. Abbasi, and Z. Li, “Information quality evaluation framework: Extending ISO 25012 data quality model,” World academy of science, Engineering and Technology 2012, 65, 523-528.

[26] L. Cai, and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” Data Science Journal 2015, 14, 10.5334/dsj-2015-002.

[27] V. Peralta, “Data Quality Evaluation in Data Integration Systems,” Human-Computer Interaction [cs.HC], 2006, Université de Versailles-Saint Quentin en Yvelines - Université de la République d’Uruguay, HAL Id: tel-00325139.

[28] D. Loshin, “Enterprise Knowledge Management: The Data Quality Approach,” San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001, isbn:0-12-455840-2.

[29] S. Pyne, B.L.S. Prakasa Rao, and S.B. Rao, “Big data analytics: Methods and applications,” Springer India 2016, isbn:978-81-322-3626-9, 10.1007/978-81-322-3628-3.

[30] J. Dean, “Big Data, Data Mining, and Machine Learning,” John Wiley & Sons 2014, Canada, isbn:9781118618042.

[31] V. Prajapati, “Big Data Analytics with R and Hadoop,” Birmingham: Packt Publishing 2013, isbn:978-1782163282.

[32] M. Talha, N. Elmarzouqi, and A. Abou El Kalam, “Quality and Security in Big Data: Challenges as opportunities to build a powerful wrap-up solution,” Journal of Ubiquitous Systems and Pervasive Networks 2019, 12, 09-15, 10.5383/JUSPN.12.01.002.

[33] J. Vitter, “Random Sampling with a Reservoir,” ACM Transactions on Mathematical Software 1985, 11, 37-57, 10.1145/3147.3165.

[34] P. Bernstein, J. Madhavan, and E. Rahm, “Generic Schema Matching, Ten Years Later,” 2011, PVLDB, 4, 695-701.

[35] E. Rahm and P. Bernstein, “A Survey of Approaches to Automatic Schema Matching,” VLDB J. 2001, 10, 334-350, 10.1007/s007780100057.

[36] E. Sutanta, R. Wardoyo, K. Mustofa, and E. Winarko, “Survey: Models and Prototypes of Schema Matching,” International Journal of Electrical and Computer Engineering, 2016, 6:3, 1011-1022, issn:2088-8708, 10.11591/ijece.v6i3.9789.

[37] E. Rahm, “Towards Large-Scale Schema and Ontology Matching,” Springer-Verlag Berlin Heidelberg 2011, 10.1007/978-3-642-16518-4_1.

[38] D. Aumueller, H. Do, S. Massmann, and E. Rahm, “Schema and ontology matching with COMA++,” Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005, 906-908, 10.1145/1066157.1066283.

[39] T.N. Herzog, F.J. Scheuren, and W.E. Winkler, “Data Quality and Record Linkage,” Springer Science+Business Media, LLC, 2007, isbn:978-0-387-69502-0, 10.1007/0-387-69505-2.

[40] H. Newcombe and J. Kennedy, “Record linkage: making maximum Use of the discriminating power of identifying information,” Commun, ACM, 1962, 5:11, 563-566, 10.1145/368996.369026.

[41] G. Shankaranarayanan, M. Ziad, and R. Wang, “Managing Data Quality in Dynamic Decision Environments,” Journal of Database Management 2005, 14, 14-32, 10.4018/jdm.2003100102.

[42] J. de Bruin, “Python Record Linkage Toolkit,” 2018, https://recordlinkage.readthedocs.io/en/latest/index.html.

[43] M. Talha, A. Abou El Kalam, and N. Elmarzouqi, “Big Data: Trade-off between Data Quality and Data Security,” Procedia Computer Science 2019, 151, 916-922, 10.1016/j.procs.2019.04.127.