# Data Quality Management - Tools and Techniques

Jigna Patel
Jigna4u@gmail.com

**Abstract:** In this era of big data, data mining and data warehousing, organizations are required to deal with very large amount of data i.e. in terabytes and petabytes. Storing this kind of huge data, manipulating and searching it becomes very tedious task, for this organizations has stated employing data quality management tools. This paper primarily focuses on what is data quality, how good quality of data can be achieved, which data quality tools are available in market and its performance. We evaluated each data quality tool on common framework which is specially and specifically designed for evaluation of data quality tools, this paper also takes consideration pricing, overall viability of product, customer services and experience. On the basis of this evaluation organization can decide which data quality management tool will be suitable for required tasks and they also come know more about various data quality tools available in the market and which of them is effective and convenient for the particular organization.

**Keywords:** Data Quality, Tools, Techniques.

## Introduction

In this era all the technical and strategic decision are data driven, to make perfect decisions and add values to organization quality of data must be standardized. Level Of efficiency and quality of data is called Data Quality. Data quality can be defined in many ways but data generally referred to high quality data, if "they are fit for their intended uses in operations, decision making and planning." (J. M. Juran) [1]. Data is considered to be of high quality if it effectively represents the real world hypothesis to which it refers. Moreover, as size of data increases, the question of internal consistency among data rises. And it is regardless how data can be perceived. Data Quality can also be defined in following ways:

(1) The measure of accuracy, completeness, consistency timeliness, interpretability and believability, these parameters makes data appropriate or of high standard.

(2) In actual scenario, degree of excellence and completeness exhibited by the data is data quality.

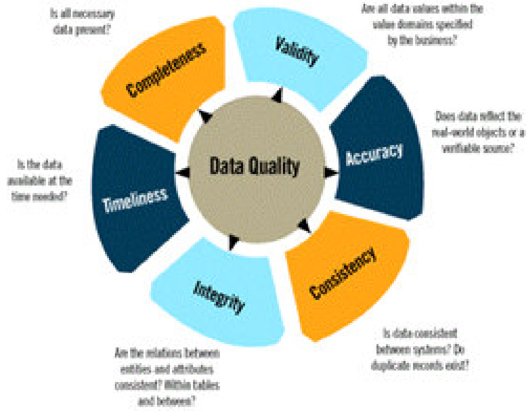

Is all necessary data present?
Completeness
Validity
Are all data values within the value domains specified by the business?
Does data reflect the real-world objects or a verifiable source?
Accuracy
Data Quality
Is the data available at the time needed?
Timeliness
Integrity
Consistency
Is data consistent between systems? Do duplicate records exist?
Are the relations between entities and attributes consistent? Within tables and between?

Fig. 1 Data Quality

Today organization leverages their data goldmine to improve customer satisfaction, improve sales and they can cut down amount of time which employee spends by taking these decision manually. The organization can make fruitful decisions and put their efforts on those decision are made on the basis of, reliable data. These processes can be much more easily automated and placed into workflows based on individual data elements Complete organizational identity management systems can be implemented that take the pain of out account

DOI: 10.090592/IJCSC.2016.029

provisioning and de-provisioning. Organizational reporting becomes as simple as pushing a button to run even the most complex reports. Of course, the reports will have to be developed, but the ongoing execution of those reports can be trusted because the underlying data is trusted. Maybe worst of all, if decisions are made with bad data or coarse data then this can lead to serious organizational problems down the line. So the data quality should be very good if decisions based on the given data are to be trusted by the organization.

## 3 HOW GOOD DATA QUALITY CAN BE ACHIEVED

Following function are required to be performed effectively in order to achieve good quality of data.

- Data Profiling: According to the definition from [2], "it is the process of examining the data available in an existing data store and collecting statistics and information about that data". Profiling is considered to be the process of collecting metadata. Following types of analysis are performed under data profiling.

▪ Completeness: This type of analysis performs checking of missing attributes and values.

▪ Uniqueness: How many distinct (unique) values are there for a given attribute across all data? Are there any duplicates values? Or should there be?

▪ Range: Statistical information for an attribute like, minimum, maximum, median and standard deviation.

▪ Pattern: Form of data across database or records, particular patterns etc.

- Data Visualization: Representation of data in graphical format is data visualization. An essential objective of information representation is to convey data effectively and productively by means of the measurable design, plots, data illustrations, tables, and graphs.

- Data Matching: Data matching ( also known as data integration, field matching or object identification) is the process of identifying, matching and merging data records that correspond to the same entities or attributes from several databases or even within one database. Following function are performed in order to achieve data integration entity identification, redundancy and correlation analysis, tuple duplication (duplication should be detected at tuple level) and data value conflict detection and resolution.

- Data Cleaning: Data cleansing (also known as data cleaning or data scrubbing) is the process of detecting and correcting incorrect, inaccurate or irrelevant records from a record set, table, or database. It also includes tasks of replacing, modifying, or deleting this dirty data or coarse data.

Data cleaning includes following tasks.

▪ Removing missing values (by ignoring tuple, filling missing value by mean or median etc.)

▪ Removing or correcting noisy data (by binning method, regression and outlier analysis.)

- Data Monitoring: This functionality includes continues monitoring of data according to user defined constraints.

## DATA QUALITY FRAMEWORK

Data Quality Framework can be defined as comprehensive checklist of all the basic practices and desirable requirements for an optimal management of data quality. This data quality framework will provide common base for evaluation of data quality tools.According data quality framework provided by [3], one should focus on following terms before proceeding for designing data quality framework.

- Data Quality Policy: It refers to the general goal and heading of an organization as for issues identified with the quality of data products. This strategy is formally communicated by top administration.

DOI: 10.090592/IJCSC.2016.029

- Data Quality Management: Determination and implementation is done by this management function.

- Data quality system: It includes the procedures, processes, organizational structure, resources and responsibilities for implementing data quality management.

- Data quality control: It is considered to be as the set of operational techniques and activities that are used to accomplish the quality required for a data product.

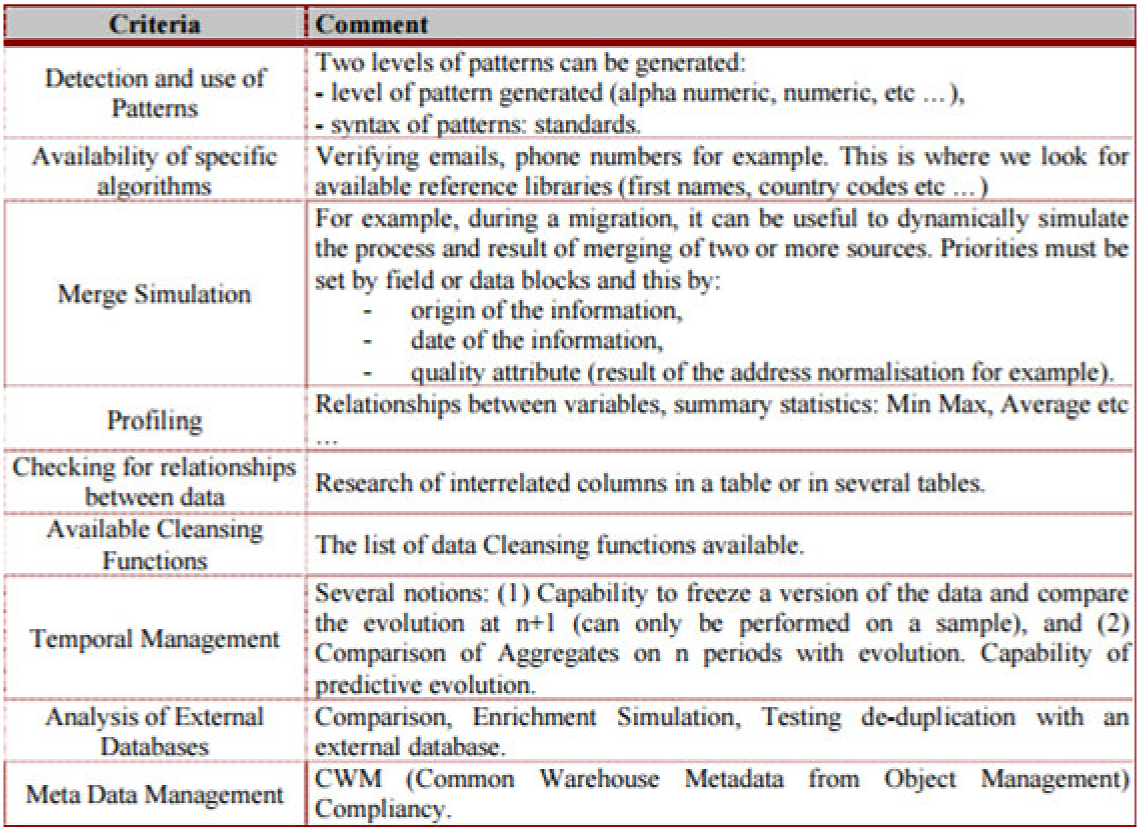
- Data quality assurance: It incorporates every one of those arranged and orderly activities important to give sufficient certainty that a data product will fulfill a given arrangement of prerequisites values.

## DESIGING DATA QUALITY FRAMEWORK

To design data quality framework first of all we will require set of rules and requirements which we will call the criteria's. There are many attempts done in order to design data quality framework by scholar of data quality field. This section gives idea about how to build effective data quality framework using prior work done in this field.

According to work of [4], data quality framework can be divided among the following criteria's.

- General Criteria: It is said that general criteria can be applied in order for choosing any type of software based on available languages, interface user friendliness, availability and language of help, error management etc.



| Criteria | Comment |
|---|---|
| Detection and use of Patterns | Two levels of patterns can be generated:<br>- level of pattern generated (alpha numeric, numeric, etc ...),<br>- syntax of patterns: standards. |
| Availability of specific algorithms | Verifying emails, phone numbers for example. This is where we look for available reference libraries (first names, country codes etc ... ) |
| Merge Simulation | For example, during a migration, it can be useful to dynamically simulate the process and result of merging of two or more sources. Priorities must be set by field or data blocks and this by:<br>- origin of the information,<br>- date of the information,<br>- quality attribute (result of the address normalisation for example). |
| Profiling | Relationships between variables, summary statistics: Min Max, Average etc ... |
| Checking for relationships between data | Research of interrelated columns in a table or in several tables. |
| Available Cleansing Functions | The list of data Cleansing functions available. |
| Temporal Management | Several notions: (1) Capability to freeze a version of the data and compare the evolution at n+1 (can only be performed on a sample), and (2) Comparison of Aggregates on n periods with evolution. Capability of predictive evolution. |
| Analysis of External Databases | Comparison, Enrichment Simulation, Testing de-duplication with an external database. |
| Meta Data Management | CWM (Common Warehouse Metadata from Object Management) Compliancy. |

Fig. 2 General Criteria

- DE duplication Criteria: De duplication criteria evaluates tool on the basis of object identification, duplication detection & resolution, rule matching etc.

- Rule Consistency Criteria: This criterion corresponds to how data quality tool able to find efficient correlation or consistency between two or more attributes.

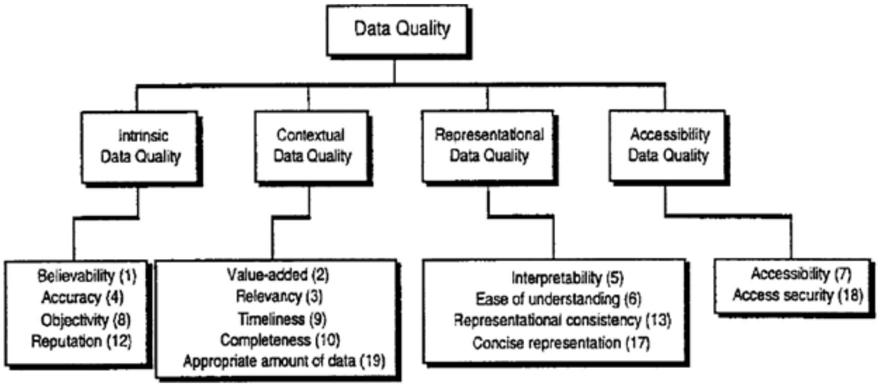


Fig. 3 Conceptual data quality framework

When this framework is applied to particular data quality tool, first a data instance is selected and hypothesis corresponding to particular criterion is tested and according to its result the score is provided for particular data quality tool.

In another work done on data quality framework by [5], data quality can be disintegrated into intrinsic data quality, conceptual data quality, representational data quality and accessibility data quality. This framework was only conceptual and designed for only study purpose.

## Conclusion

If all the strategic and technological decision for an organization has to be data driven then the data quality tool must ensure guarantee to provided quality data on which organizations can entrust their decisions. From midsize organization to large size organization we can see growing demand for data quality tools. In the area of business intelligence and analytical research data must be very accurate in order to make decision based on them. Today, there is no provision for checking which algorithm is used for specific functionality, how this algorithms can be evaluated on the basis of time and space. So we must work on the area of evaluating internal aspect of data quality tool and build such a framework that can provide criteria's for evaluating algorithms and we can design a system in which organization which want to employ data quality tool has to provide their requirement and functionalities and the system will make performance analysis according to given input. The system will generate list of data quality management tools as results which are suitable to particular customer.

## References

1. Definition of Data quality, https://en.wikipedia.org/wiki/Data_quality.
2. Pitney Bowes, "Data Profiling Best Practices".
3. VirginieGoasdoué, SylvaineNugier, Dominique Duquennoy, Brigitte Laboisse,"An Evaluation Framework For Data Quality Tools".
4. Richaed Y. Wang , Diane M. Storng, "Beyond Accuracy: What Data Quality Means to Data Consumer".
5. Richard Y. Wang, Veda C. Storey, and Christopher P. Firth," A Framework for Analysis of Data Quality Research".
6. "The Practitioner's Guide to Data Profiling", http://www.sas.com/en_us/whitepapers/practitioners-guide-data-profiling106046 .html
7. "Definition of Data Parsing", https://en.wikipedia.org/wiki/Parsing.