National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Survey of Data Quality Parameters and system for verification Data Quality

Trushna S. Duddalwar **Pooja K. Kale** **Prof.Deepa P. Vaidya**

***Abstract:*Data Warehousing is a blend of technologies aimed at effective integration of operational database into an environment that enable strategies use of data. These technologies include relational and multidimensional database management system, client/server architecture, metadata modeling and repositories, GUI and much more.**

In Data Warehouse, data quality which implies the data is accurate, timely and consistent for makingdecisions.

During this study we have identified the different data quality parameters. These parameters play a very vital role in the information generation. A good quality data can produce quality and more authentic information which can be effectively used for decision making in the business environment. Manual verification of quality of data is much difficult and thus we have proposed a system for automatic data quality verification mechanism in this paper. TheData quality parameterslike: accuracy, consistency, completeness, uniqueness, validity are some of the selected one which affect a quality index with great extent. We have to proposed a system which check whether the data under examination isaccurate or notby using syntactic accuracy and semantic accuracy and data is complete or not by using value completeness, tuple completeness, attribute completeness, relation completeness etc.

***Keyword*: Data Warehouse, data quality, data quality parameters.**

I. INTRODUCTION

Data quality has gained more and more importance due to an extended use of data warehouse systems, management support systems and a higher relevance of customer relationship management as well as multichannel management. This refers – for decision makers – the benefit of data depends heavily on their completeness, correctness, and timeliness, respectively. These properties are known as data quality dimension. [1]

A high degree of data quality is required, for organizations to be best served by their information systems, and the need to ensure this quality has been defined by both re- searchers and practitioners. The objective of this paper is to develop data by using data quality parameter.[2]

In data warehousing, its success depends on the quality of the data stored in it. [3]

II. RELATED WORK

The Researchers introduces analyze how data quality can be quantified with respect to particular dimension. [1] Similarly data quality management program is not an easy task, but the rewards are enormous.[4]The proposed definition for accuracy, completeness, consistency and time related dimensions are applicable in many contents, including e-Business and e-Government. [7] Similarly, the approach is used to analyze the quality of data warehouse system by checking the expected value of quality parameters which that of actual values.[3]

III. DATA QUALITY MANAGEMENT

Data quality management entails the establishment and deployment of roles, responsibilities, policies andprocedures concerning the acquisition, maintenance, dissemination, and disposition of data. A partnership betweenthe business and technology groups is essential for any data quality management effort to succeed. The businessareas are responsible for establishing the business rules that govern the data and are ultimately responsible for verifying the data quality. The Information Technology (IT) group is responsible for establishing and managing theoverall environment – architecture, technical facilities, systems, and databases – that acquire, maintain, disseminate,and dispose of the electronic data assets of the organization.[4]

Quality management can be based on the ISO 9000, self-assessments using quality award criteria e.g. the Excellence Model of the European Foundation for Quality Management (EFQM) , the Malcolm Baldrige National Quality Award in the USA, practical experiences within the company or a combination of those. The ISO 9000 and quality award criteria require the quality system to be documented. [5]

IV. DATA QUALITY

Data quality is an essential characteristic that determines the reliability of data for making decision. Data quality is not linear and it has many dimensions.

The meaning of 'quality' depends on the context in which it is applied. The term is commonly used to indicate the superiority of a manufactured good or attest to a high degree of craftsmanship or artistry. In manufacturing industries, quality is viewed as a desirable goal to be achieved through management of the production process. Unlike manufactured products, data do not have physical characteristics that allow quality to be easily assessed. Quality is thus a function of

intangible properties such as 'completeness' and 'consistency'. [6]
In the following we discuss how metrics for selected DQ dimensions can be designed with regard to two objectives: (1) Enabling the measurement of DQ, (2) Analyzing the economic consequences of DQ measures taken. [1]
In selecting the appropriate body of research for examination, two primary criteria were used:
1) The authors of the articles specifically recognize a data quality problem, which they attempt to address in their work. That is, the research is motivated by a data quality issue.
2) The researchers address a problem that, although not specifically described as a data quality issue, is comprised of components that are related to data quality management.[2]

V. DATA QUALITY PARAMETER

- *Accuracy:*

Accuracy can be evaluated for disparate granularity levels of a data model, ranging from single values to entire databases. For single data values, accuracy measures the distance between a value v and a value v' which is considered correct. Two kinds of accuracy can be identified, namely a syntactic accuracy and a semantic accuracy.
Syntactic accuracy is measured by means of comparison functions that evaluate the distance between v and v'. Edit distance is a simple example of comparison function, taking into account the cost of converting a string s to a string s' through a sequence of character insertions, deletions, and replacements. More complex comparison functions exist which take into account similar sound, transpositions etc.

As an example, let us consider again the relation Movies, shown in Fig. The accuracy error of movie 3 on the Title value is a syntactic accuracy problem. As the correct value for Rman Holidays is Roman Holidays, the edit distance between the two values is equal to 1 and simply corresponds to the insertion of the char "o" in the string Rman Holidays.Semantic accuracy captures the case in which v is a syntactically correct value, but it is different from v'.

In the same Movies relation, swapping the directors' names for tuples 1 and 2 results in a semantic accuracy error, because although a director named Weir would be syntactically correct, he is not the director of Casablanca, therefore the association between movie and director is semantically inaccurate.

Id	Title	Director	Year	#Remake	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poet Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	Null	1964	0	1985

Table1:- Relation Movies with Data Quality Problems

From these examples, it is clear that detecting semantic accuracy is typically more involved than detecting syntactic accuracy. It is often possible to detect semantic inaccuracy in a record, and to provide an accurate value, by comparing the record with equivalent data in different sources. This requires the ability to recognize that two records refer to the same real world entity, a task often referred to as the object identification problem. As an example, if two records store J.E. Miller and John Edward Miller as a person's name, the object identification problem aims to realize if the two records represent the same person or not. There are two aspects to this problem:

- *Identification:*records in different sources have typically different identifiers. Either it is possible to map identification codes, or matching keys must be introduced to link the same records in different sources.
- *Decision:*once records are linked on the basis of a matching key, a decision must be made as to whether or not the records represent the same real world entity.

The version of accuracy discussed above, both syntactic and semantic, refers to a single value, for instance of a relation attribute. [7]

- *Completeness:*

In a model with null values, the presence of a null value has the general meaning of a missing value. In order to characterize completeness, it is important to understand why the value is missing. Indeed, a value can be missing either because it exists but is unknown, or because it does not exist at all, or because its existence is unknown.

Let us consider, as an example, a Person relation, with the attributes Name, Surname, Birth date, and Email. The relation is shown in Figure.
For tuples with ID 2, 3 and 4, the email value is null. Let us suppose that the person represented by tuple 2 has no email; in this case, there is no incompleteness. If the person represented by tuple 3 has an email but it is not known which is the value, than tuple 3 is incomplete. Finally, if it is not knownwhether the person represented by tuple 4 has an email or not, we cannot determine whether the tuple is incomplete.

Id	Name	Surname	Birth Date	Email
1	John	Smith	3/17/1974	Smith@abc.it
2	Edward	Monroe	2/3/1967	Null
3	Anthony	White	1/1/1936	Null
4	Marianne	Collins	11/20/1955	Null

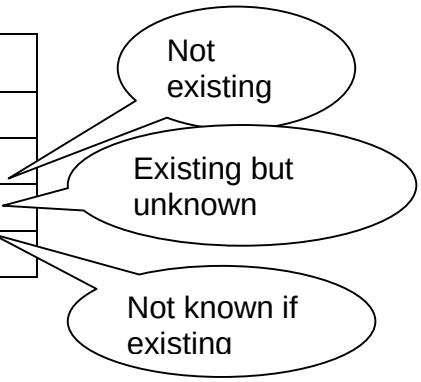


Table 2: Example of different Null Value Meaning

Besides null values meaning, precise definitions for completeness can be provided by considering the granularity of the model elements, i.e., value, tuple, attribute and relations.
Specifically, it is possible to define:
- avalue completeness to capture the presence of null values for some attributes of tuples;
- atuple completeness to characterize the completeness of a whole tuple with respect to the values of all attributes;
- anattribute completeness to measure the number of null values of a specific attribute in a relation;
- arelation completeness that captures the presence of null values in the whole relation. [7]

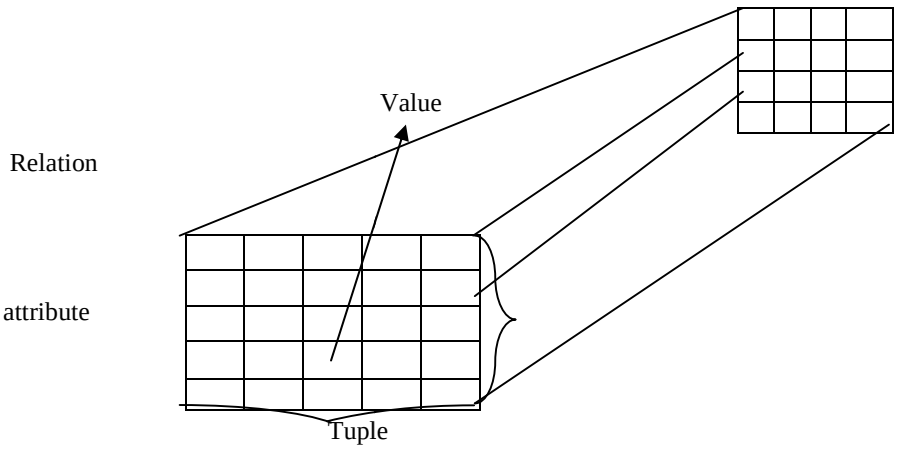


Fig:- Completeness of different element of the Relational Model.

VI. PROPOSED SYSTEM

From this study it is learn that there are different data quality parameters which overall defined the purity and usability of the data. The data quality measures are upto some extent implemented at the data entry level by using technique called validation. The validation check found to be a effective tool to chance the data quality but it has many limitation. The data extend through the interface having validation check enable quality data entry but it assure only few parameters like completeness and uniqueness. The accuracy, timeliness, consistency, availability are some other parameter required to be addressed and pertained for overall quality of data. In this paper, we suggested few methods to validate the accuracy and consistency and completeness.

- Methods for evaluating the accuracy of data:- the syntactic accuracy of data is effectively maintained using the attribute parameter like data type, data

format for semantic accuracy methodology suggested as follows Range value, vocabulary maintained.

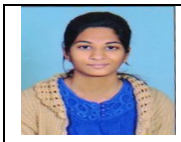
CONCLUSION

In this paper, we can check how the data can be accurate, complete by using data quality parameter methods. Data quality play important role for data warehouse and data mining so in this paper, parameter like accuracy and completeness are explained. We can also further extend consider remaining parameter in future.

REFERENCES

[1].Bernd Heinrich, Marcus Kaiser, Mathias Klier "How to measure data quality" A Metric Based Approach.
[2].Richard Y.Wang, Veda C. Storey, and Christopher P. Firth" A Framework for Analysis of Data Quality Research." IEEE Transaction on knowledge and data engineering, vol. 7, no.4, august 1995.
[3].Vinay Kumar and Reema Thareja "A Simplified Approach for Quality Management in Data Warehouse." International Journal of Data Mining and Knowledge Management Process (IJDKDP) Vol.3, no.5, September 2013
[4].Jonathan G.Geiger, Intelligent solution, Inc, Boulder, CO "Data Quality Management"
[5].Antti Jakobsson "Data Quality and Quality Management- example of Quality Evaluation Procedure and Quality Management in European National Mapping Agencies. "
[6].H Veregin "Data Quality Parameter".
[7]. Monica Scannapieco, Paolo Missier, Carlo Batini "Data Quality at a Glance."

AUTHOR'S PROFILE



Trushna S. Duddalwarreceived her BCA from RTM Nagpur University. Presently student of MSc in Computer Science from HVPM an autonomous college.

Pooja K. Kale received her BSc in Computer Science from HVPM an autonomouscollege. Presently student of MSc in Computer Science from same college.

Prof. Deepa P. Vaidya
H.O.D of BBA, BCA, B.Sc A Prof. in PG Department of computer science and technology in HVPM autonomous college.