Data Profiling Model for Assessing the Quality Traits of Master Data Management

Dilbag Singh, Dupinder Kaur

Abstract: ***Enterprise Resource Planning (ERP) and Business Intelligence (BI) system demand progressive rules for maintaining the valuable information about customers, products, suppliers and vendors as data captured through different sources may not be of high quality due to human errors, in many cases. The problem encounters when this information is accessible across multiple systems, within same organization. Providing adequacy to this scattered data is a top agenda for any organization as maintaining the data is complicated, as having high quality data. Master Data Management (MDM) provides a solution to these problems by maintaining "a single reference of truth" with authoritative source of master data (Customer, products, employees etc). Master Data Management (MDM) is a highlighted concern now a day as valid data is the demand for strategic, tactical and operational steering of every organization. The lane to MDM initiates with the quality of data which demands for discovery of master data, profiling and analysis. As inadequacy of data may leads to adverse effects such as wrong decision, loss of time, bad results and unnecessary risk. Thus there is a need to deal with master data and quality of this specific data in a successful and efficient manner. For ensuring this purpose, an approach is proposed in this paper. The research focuses on development of a Model for Data Profiling to assess the level of Quality Traits for Master Data Management. Results are shown by executing the defined steps on TALEND tool over collected dataset. Thus, level of quality traits processes directly correlates with an organization's ability to make the proper decisions and better outcomes.***

Keywords: Data Quality, Data Profiling, Master Data Management, Quality Traits.

I. INTRODUCTION

In the changing dimensions of business, major constituent of any organization's day-to-day business is the data that is used in functional, operational and progressive activities.
The quantity and complexity of storing this data becomes hard with the growth of organizations. If this data is incorrect, wrong or out of date, the organization may suffer from delays, bad results or financial losses. As data is scattered and stored in different types of locations, it becomes difficult to provide data consistency, accuracy, uniqueness, completeness etc. There must be proper organization and management with massive growth of data [2]. For example if a company creates a service which provides communication through XML messages that have a single view of customer data. But if the record of same customer is stored in four databases with two different addresses, it becomes a tedious task to provide service to that customer.

Thus it is crucial to addresses these critical data entities (golden record) also known as master data, which is a single truth version of record among different copies that belongs to same person having different values. Master data is the complete record of a human, like the data is about customers, patients, suppliers, partners, employees, and other critical entities. For this purpose, a high quality and consistent copy of master data becomes a primary task for any organization. The processes and systems which maintain this data are known as Master Data Management [5].
Master Data Management (MDM) is a holistic framework consisting of processes, tools and technologies which provides coordinated master data in the enterprise. When data is collected from multiple, conflicting sources of information, MDM initiates by determining, collecting and transforming(cleaning and standardizing)the data thus ultimately providing repair to that data [12]. Quality of master data is determined by the extent to which master data is satisfied to defined quality traits. Data Quality is defined as the process of making data fit for use and meeting the requirements of user [3][4]. Only qualitative data can provide useful, resilient, functional and applicable results. It allows a profound understanding of the research data that is always up-to-date. The completeness, uniqueness, correctness and timeliness of the data are crucial for successful operational processes [5].
Data profiling is a technique of exploring data from existing information sources and collecting informative summaries about that data. The intention is to discover data quality as a component of a large project by determining the accuracy, completeness, and consistency of data. This initial analysis ensures at an early stage if the correct data is available for use thus anomalies can be handled subsequently. Thus the objective of this research is to define and categorize the problems related to quality of data and to provide a systematic approach for assessing the level of quality traits in Master Data Management.

A. Data Quality Traits

Data Quality trait is a measure for qualitative or quantitative analysis in order to determine the quality of data [4].The major issue is that data is initially taken either from single source or multiple sources. The scattered data over different locations make the data unfit for use as it leads to quality issues such as duplicity of information inconsistency and incompleteness [6].

Revised Manuscript Received on February 10, 2020.
***** Correspondence Author
Dr. Dilbag Singh: Computer Science and Applications**,** Chaudhary Devi Lal University, Sirsa (HR), India. Email: dscdlu7@gmail.com
Dupinder Kaur*: Computer Science and Applications**,** Chaudhary Devi Lal University, Sirsa (HR), India. Email: dupinder.pahwa@gmail.com

Table- I: Data Quality Traits

S.No	Quality Traits	Definition	Example
1	Consistency	Assessment of equivalent data values across multiple systems.	The address of a person has the same value and format as in person's ID card as that stored within the database.
2	Completeness	Measurement of presence of non-blank values.	Student's first name and last name are mandatory; the record is incomplete if both fields are not provided.
3	Accuracy	Degree to which data matches with "real-life" objects.	Data entered into different data formats as DD/MM/YY and MM/DD/YY may affect accuracy of data.
4	Redundancy	Same data is stored at different fields or tables.	Student id must be unique. If two students have same Id no, it will make redundant information.
5	Domain Integrity Constraint	Defined valid set of rules for attributes.	In "AGE" column, only integer value is allowed rather than character, string or time.

Table I describes the necessary traits for adequate data quality in creating master data. It is suggested that these traits and standards should be adopted by data quality professionals as these designed methods are useful for accessing and maintaining the level of data quality [6].

II. LITERATURE REVIEW

It presents the review of work done by various researchers that provides a useful insight for improving data quality and managing the master data management. This work done is very helpful in understanding what has been done in the field of analyzing data quality issues.

Alex Gamero (2019) focused on processes of manual data cleaning and the problem of information quality in microfinance sector. MDM is proposed as the basis for interaction between different segments of data governance and infrastructure levels. The proposed model comprises of four sub microfinance organization levels including 8 phases. Among these phases only 2 are supportable, 40 used for evaluation criteria and 5 levels are kept for maturity[1].

Fajar Gumelar (2018) gave an advice to the organization to upgrade the master data maturity at different levels. MD3M model is used to access these levels. The result showed that, in MD3M only 50% of capabilities had been implemented by organization. High maturity level can be achieved within an organization by further implementing the proposed idea[2].

Lu Bai (2018) discussed a prototype for data quality with the use of multidimensional model. MySQL relational database management system has been used to access the data. The finding shows that these artifact will be assessed for functional usefulness and applicability of data[3].

M. Izham Jay (2017) identified four critical data quality parameters including data completeness, consistency, accuracy and timeliness. These parameters are critical and should be given highest priority in managing data quality within the organization. By adopting data quality management model like TDQM, IIM and AIMQ and methodologies may enhance data quality from various data sources especially in unstructured data [4].

Otmane Azeroual (2017) stated that effective adequacy level of the data quality can be achieved at various stages of any research information system. In this respect, improvement of data quality of RIS is targeted. It suggested a quality workflow procedure and a use case in order to better evaluate the data quality in RIS [5].

III. RESEARCH METHODOLOGY

A research is about establishing facts and reaching new conclusions by systematic study of a system. Research is conducted by taking some objectives in mind. The desired goals of an organization cannot be achieved, if their business processes are irregular, inefficient and do not meet their client needs. Thus the ultimate solution is to adopt MDM strategies to resolve the problem of inconsistent data. Thus the objective of this research is to access the levels of data quality traits and to provide a model for data profiling that an organization follows to achieve an effective adequacy level of data. In this study, Quantitative and Qualitative research methodologies are applied which provide insight information to know the extent of data consistency, accuracy and redundancy. The research is performed on TOSDQ (TALEND Open Studio for Data Quality) software. It is an ETL (Extract, Transform and Load) tool that provides a software solution for data preparation, data quality, data integration and data Management.

IV. IMPACT OF DATA QUALITY ON MASTER DATA

If master data is not consistent and accurate, organizations will be less reactive to new systematic compliance which ultimately affects the reporting processes. The loss due to poor data is mostly in financial terms [10]. Gartner conducted a research on the set of organizations and found that the annual cost due to dirty data was on average of $13.3 million dollars, in 2014 [8]. Thus, in the process of data source handling, data quality is of utmost importance. The quality of data is subjective term which varies as per the perspective of the system that means the data which is of good quality for one system might not be good for another system. The requirement varies as per the needs of organizations. Inadequate data quality may leads to followings problems: [8]

- Customer dissatisfaction and loss of customers.
- Delays in project deployment
- Wrong business decisions.
- Misguided opportunities.
- Legal and monetary penalties.
- The cost of populating master data will go waste if the data is of not good quality [7].

V. CONTRIBUTIONS

Based on the study and analysis, following are the major contributions of this paper

- Overview of Master Data management and analyzing the necessary quality traits across different sets of database.

- Describing the impact of dirty data and thus the role of adequate data quality in functional, operational and effective decision making.
- Proposing a Model for Data profiling by identifying these traits for valuable and better outcomes in Master Data Management.

VI. DATASET

The datasets for the research have been collected from the university campus in the raw form and a database is created for the same in MYSQL 5.5. The sample screenshots of few datasets are given below.

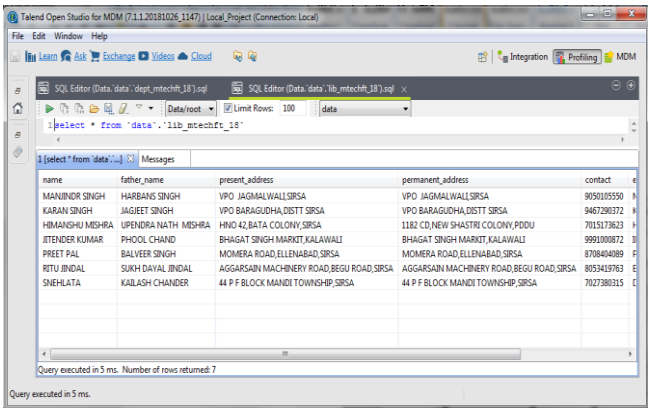

Fig 1: Screenshot of table "lib_mtechft_18".

Fig 1 depicts the screenshot of TALEND window containing record of MTech Full time students admitted in 2018, collected from university library.

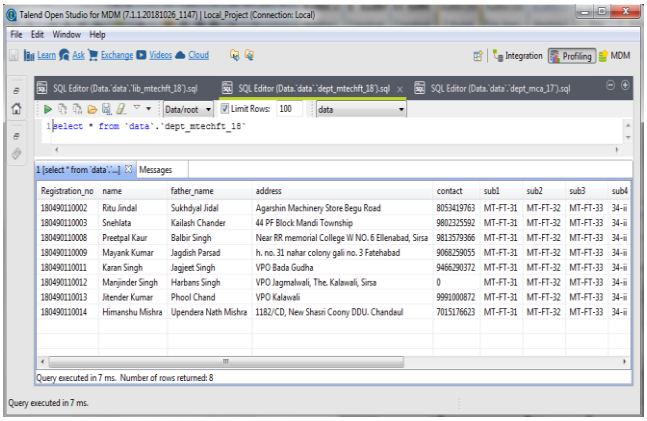

Fig 2: Screenshot of table dept_mtechft_18

Fig 2 depicts the screenshot of TALEND window containing the record of M.Tech. Full time students admitted in 2018, collected from university department.

VII. DATA PROFILING MODEL FOR ASSESSING QUALITY TRAITS

In many cases, maximum of data projects fail due to complexity of data, over scheduling, or over budgeting. This is because organization starts data projects with an inconsistent, incomplete or incorrect picture of data which may result into unexpected delays. Qualitative data, meeting the requirements of authors, users, and administrators, helps to improve the business performance. Hence, to ensure the quality of data, a profiling process is required.

A. Proposed Model

Data profiling is the practice of data analyzing, examining and reviewing data to collect statistics that surrounds the quality of the dataset. It is a continuous activity for defining acceptance level of data quality to meet business requirements. To achieve this target, a model for data profiling by considering quality traits is proposed in order to improve the worth of the source data.

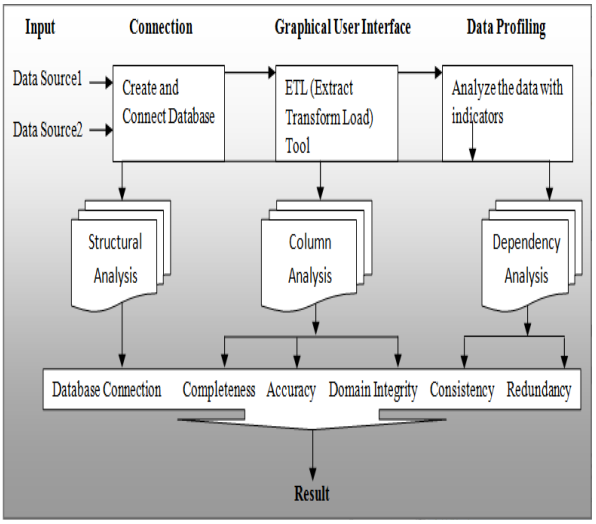

Fig 3: Model for Assessing Data Quality Traits

Fig 3 represents the proposed model for assessing data quality traits. It describes the overall steps involved in statistical analysis of quality traits. The complete model is divided into five sub modules:

- **Input**: Initially, the data is collected as the records of MCA and MTECH students admitted in 2017 onwards in the university. Source data 1consist of the records of students available in CSA department and Source data 2 is taken form library (The students who have taken library membership) of university.
- **Connection**: The collected database is firstly converted into CSV (Comma Separated Value) files and stored in MYSQL named as "DATA". This stored database is then connected to interface under metadata section.
- **Graphical user Interface:** The interface used for this research is "TOS-DQ" (TALEND Open Studio for Data Quality).It provides data profiling, advance statistics and analytics with graphical charts and data.
- **Data profiling:** Data Profiling involves the identification of required dimensions of adequate data quality preferable in master data management. The quality traits are categorized into three types of analysis under this section. Predefined and customized indicators are used here for analyzing quality traits on the created database.
- **Result:** The analysis is graphically represented showing the various levels of considered data quality traits.

VIII. RESULTS

The results of conducted research are represented here:

A. Database Connection

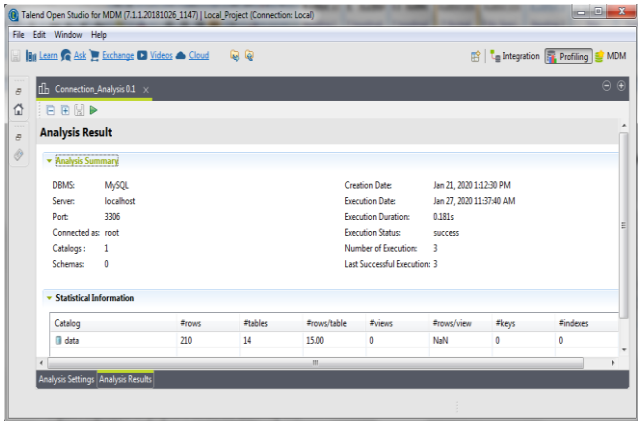

Fig 4: Screenshot of Database Connection Analysis

Fig 4 illustrates the screenshot of database connection with "TALEND Open Studio for Data Quality" tool. It shows that the database named as "Data" consists of 14 tables, 210 total rows and average 15 rows per table.

B. Analysis of Database Consistency

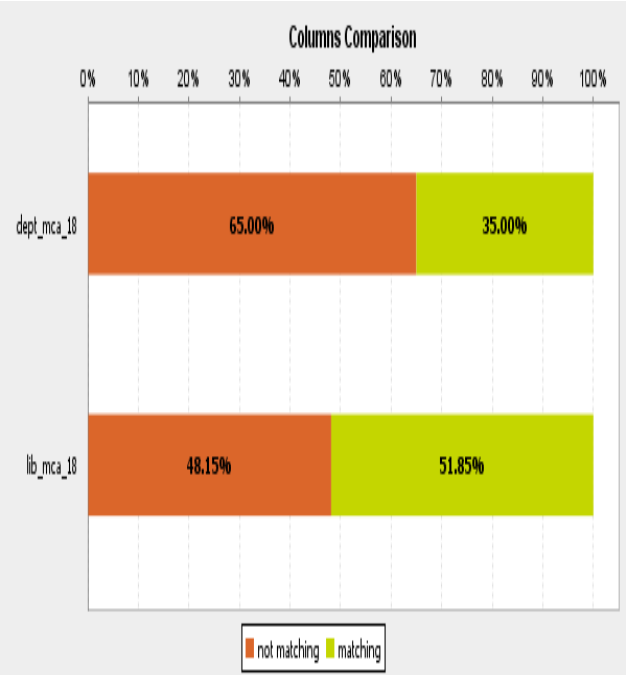

Fig 5: Level of Database Consistency

Fig 5 represents the level of database consistency. The analysis is on two tables: Table1: dept_mca_18 having records of MCA students admitted in 2018 available in department and Table 2: lib_mca_18 having records of MCA students admitted in 2018 who have taken library membership. Only three attributes which are common in both tables: Name, Father Name and Contact are considered. As the data must be consistent at both sides but it is found that out of 100% records of lib_mca_18 table only 51.85% of the records are matching with dept_mca_18 which must be 100%. Thus there is lack of consistency.

C. Analysis of Completeness and Redundancy

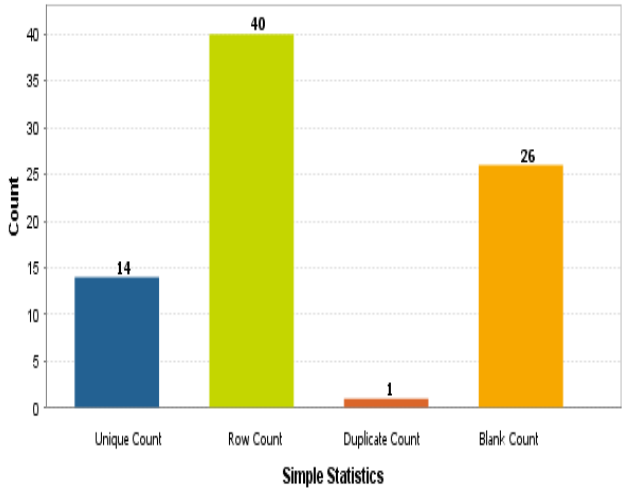

Fig 6: Level of Database Completeness and Redundancy

Fig 6 shows the statistics for incomplete and redundant records. This Analysis is made on column "Registration No" of MCA students admitted in 2018. As Registration no must be unique for each entry but it shows that there are total 40 rows out of which only 14 entries are unique, 1 record is duplicate and 26 entries are blank. Thus database redundancy and incompleteness exist here.

D. Analysis of Database Accuracy and Domain Integrity Constraint

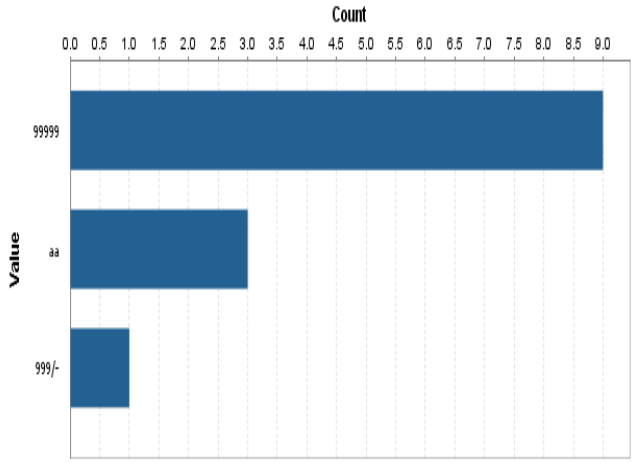

Fig 7: Level of Database Accuracy and Domain Integrity Constraint

Fig 7 describes the violation of accuracy and domain integrity constraint in database. This analysis is performed on "FEE" attribute. The fees submitted by students of MTECH class. It can be clearly seen that for one class there are three different formats for data representation. Also the format used for numerical value is not accurate for each entry. As Fee section must contain only numerical value, no character or string entry is allowed here. Thus it violates the accuracy and domain integrity constraints. As per the proposed model, research is conducted over the database collected from the university. An Analysis is made for assessing the level of quality traits in Master Data Management.

The output shows that there is violation of Data Consistency, Completeness, Accuracy and Domain Integrity constraints. Also Redundancy exits in the database table which makes data of bad quality. It is important to check and validate the data before decision making so that possibility of risk and time loss can be avoided at early stage.

IX. CONCLUSION AND FUTURE WORK

The paper presents a Data Profiling Model for assessing the level of Data Quality Traits in order to create Master Data. The Purpose is to determine essential traits and their impact on creating master data. MDM without high quality data is like trying to assemble a car without standard nuts and bolts. The proposed model describes the way of assessing the levels of data quality traits. The dataset is the record of students admitted in 2017 onwards, in the university. This raw data is firstly inbuilt into MYSQL. Thereafter, with the help of model, implemented on TALEND tool, it is justified that data collected from university is inadequate as it is Redundant and violates Domain Integrity Constraint. It is found that the data is inconsistence, incomplete and inaccurate also. These quality traits are base form for overall good quality of master data. Thus on the basis of poor quality data, wrong interpretation can be made which may ultimately affect the organization's outcomes. Data profiling involves early identification of duplicate records, incomplete and inaccurate data, different standard data format and inconsistent records. Proper Data Quality Management process should be followed in order to resolve these issues and adding new information to existing data for a continuous vision on maintaining updated and consistent data. This will result in minimizing cost, timeframes and risk in order to provide an effective level of adequate Data quality.

REFERENCES

1. Alex Gamero, "Reference Model with a Lean Approach of Master Data Management in the Peruvian Microfinance Sector", International Conference on Industrial Technology and Management, IEEE,201, pp:56-60.
2. Fajar Gumelar Pratama, "Master Data Management Maturity Assessment: A Case Study of Organization in Ministry of Education and Culture" International Conference on Computer, Control, Informatics and its Applications, IEEE,2018,, pp:1-6.
3. Lu Bai , "A Data Quality Framework, methods and tools for managing data quality in health care setting: an action case study", Journal of Decision Systems ,Tailor & francis Group, 2018.
4. M. Izham Jay, "A Review of Data Quality Research in achieving high data quality within organization", Journal of Theoretical and Applied Information Technology, Vol.95. Issue No 12, 2017.
5. Otmane Azeroual, "Improving Data Quality in Research Information System", International Journal of Computer Science and Information Security, Vol. 15, Issue No. 11, 2017.
6. Faizura Haneem, "Resolving Data duplication, inaccuracy and inconsistency issues using Master Data Manegement", IEEE, 2017.
7. Nuno Laranjeiro, "A Survey n Data Quality: Classifying poor data", 21st Pacific Rim International Symposium on Dependable Computing", IEEE, 2015.
8. Gartner, "Magic Quadrant for Data Quality Tools," available on "http://www.gartner.com/technology/reprints.do?id= 1-259U63Q&ct=141126&st=sb",2014.
9. E. Gomede, "Master Data Management and Data Warehouse: An architectural approach for improved decisionmaking," in Information Systems and Technologies (CISTI), 8th Iberian Conference on IEEE, 2013.
10. Otto B, "Functional architecture for company-wide master data Management", 2013.
11. https://profisee.com/master-data-management-what-why-how-who/,2019
12. https://www.informatica.com/in/services-and-training/glossary-of-terms/master-data-management-definition.html#fbid=xJIeflsjFKG, 2019

AUTHORS PROFILE

Dr. Dilbag singh, Professor, Department of Computer Science and Applications, Chaudhary Devi Lal University, Sirsa(HR.)
Research areas: Simulation, Data Mining, and Networking.

Ms. Dupinder kaur, PhD Research scholar, Department of Computer Science and Applications, Chaudhary Devi lal University, Sirsa(HR.). Gold Medalist, UGC-NET, GATE(Computer Science)
Research areas: Master Data Management, Data Profiling, Data Quality and integration Tools.