Implementation and Substantiation of Source Data Quality Issues in Data Warehouse of Educational Institutions

Ranjit Singh[1] Ms. Jyoti Arora[2]
[1,2]Assistant Professor
[1,2]Department of Computer Engineering
[1]APEEJAY-IMTC Jalandhar Punjab India [2]Deshbhagat University Mandi Gobindgarh Punjab India

Abstract— Data quality has been discussed at length in literature since last two decades. Many researchers have highlighted the data quality from various points of views. Some of them have discussed data quality from view point of defining the and some have highlighted the data quality in terms of issues and problems of the data present in the organizations. . Typically in an organization there may be as many applications and methods of data storage as many number of departments are existing in that organization and is existing in different formats. And hence data is becoming less useful or in harsh words it is becoming "dirty". The purpose of this research paper is to verify the data quality issues present in the data sources of educational institutions. An in house developed web based data entry system of campus management system was examined to view the pattern of data quality issues.

Key words: Data Warehouse, Source Data Quality Issues in Data Warehouse

I. INTRODUCTION

Most of the enterprises deal with in three categories: master data, transactional data, and historical data. Master data are defined as the basic characteristics of business entities, i.e. customers, products, employees, suppliers, etc. Thus, typically, master data are created once, used many times and do not change frequently. Transaction data describe the relevant events in a company, i.e. orders, invoices, payments, deliveries, storage records etc. Since transactions are based on master data, erroneous master data can have significant costs, e.g. an incorrect priced item may imply that money is lost. Transactional data sources are more prone to the errors of various types [1]. Poor quality data can, therefore, have significantly negative impacts on the efficiency of an organization, while high quality data are often crucial to a company's success. However, several industry expert surveys indicate that data quality is an area, to which many companies seem not to give sufficient attention or know how to deal with efficiently [2]. We have taken into consideration about one of the data entry form of in-house developed campus management system and tried to observe the data quality issues which are propagated while data entry into the system. This is later on going to serve the data source for data warehouse projects of the educational institutions. If the data quality issues are not identified at the sources level then these issues will propagate to the next level and hence producing *'dirty data'* for poor decision making.

II. PROBLEM STATEMENT

On Average, Data Warehouses harvest Data from 10 Distinct Sources.

In educational institutional domain we are not fetching data from all possible more than 10 data sources as shown in figure1. In educational institutions either the data is stored in flat files or OLTP (Online Transaction Processing System) systems. In our research we have taken one of the data source for the examination and observation of data quality issues. This data source is a web based OLTP application which is in house developed for the purpose of admission data storage. Our aim is also to verify both schema level and record level data quality problems present in the data sources of institutions.

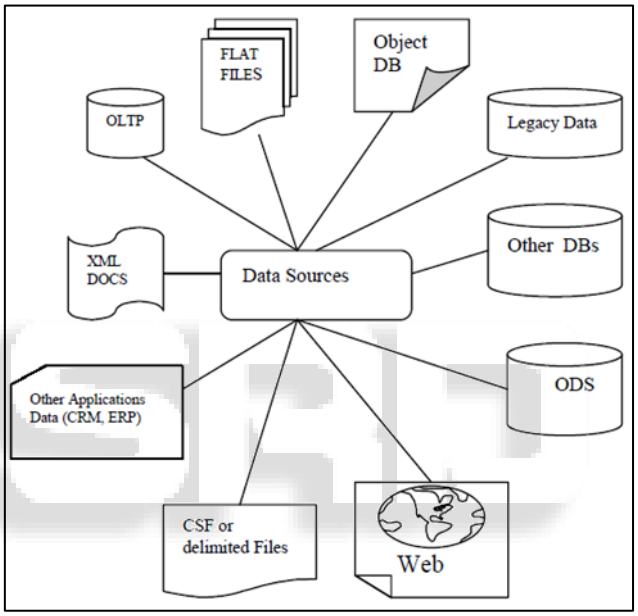

Fig. 1: Possible Data Sources for Data Warehouse [3].

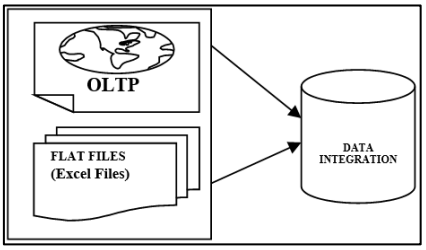

Fig. 2: Data Sources in Educational Institutions

III. RESEARCH METHODOLOGY

The most common source of data inaccuracy is that the person manually entering the data just makes a mistake. Convoluted and inconvenient data entry forms often further complicate the data entry challenge. The same applies to data entry windows and web-based interfaces. Frustration in using a form will lead to exponential increase in the number of errors. Users often tend to find the easiest way to complete the form, even if that means making deliberate mistakes.

We are examining a web based OLTP for entering admission data. This is a web based data entry form designed to enter every incoming query for admission. Various teachers and admission councilors are using this portal during

their duty times. They have to record every incoming admission query in to this portal.

For the purpose of testing and verification of data quality issues in data entered into the data source we have taken 30 records as sample to observe the following data quality issues shown in Table 1.

DQI	Name of the Data Quality Issue
DQI-1	Outliers (Out of Range) values
DQI-2	Missing Attribute/Column Values
DQI-3	Misfielded Values
DQI-4	Misspellings
DQI-5	Unique Violations
DQI-6	Synonyms
DQI-7	Homonyms

Table 1: Data Quality Issues under observation

Following figure shows the data entry form used by admission counselors to enter the data

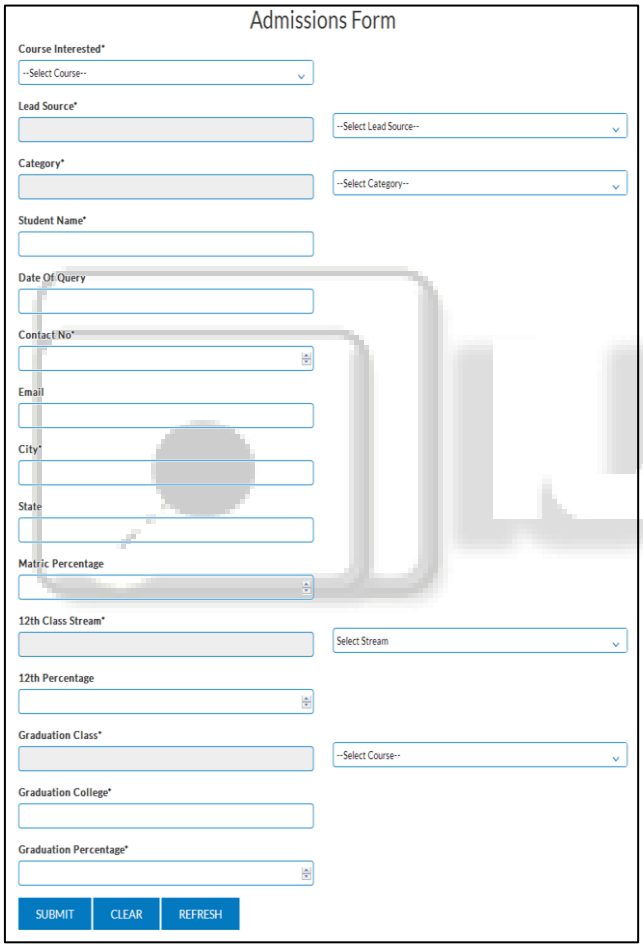

Fig. 3: Data Entry form

We have developed wab based data analysis tool in PHP which is performing column wise data quality analysis and is showing the counts/frequencies of data quality issues present in the source data.

IV. RESULTS & DISCUSSIONS

In our data source we have single table to be analyzed. This table is in backend MYSQL. Broadly we have focused on main data quality issues which may encounter in single source systems. Table 2 shows the observed count of issues in each column of data source.

Table 2 Column Wise Count of Data Quality Issues Observed

DQ Issues → / Column Name	column No	Missing Column Values	Outliers	Misfielded Information	Misspellings	Synonyms	Homonyms	Unique Violations/ Duplicates	Invalid Data
Source	1	0	0	0	0	0	0	0	0
Course	2	0	0	0	0	0	0	0	0
Ref_id	3	0	0	0	0	0	0	0	0
Student_cat	4	0	0	0	0	0	0	0	0
Student_name	5	10	0	0	0	0	0	0	3
Matric %age	6	20	3	0	0	0	0	0	3
12th Stream	7	0	0	0	0	0	0	0	0
12th %age	8	10	2	0	0	0	0	0	2
Graduation	9	0	0	0	0	0	0	0	0
Graduation College	10	19	0	2	0	0	0	0	5
Graduation %age	11	10	3	0	0	0	0	0	0
Contact	12	0	6	6	0	0	0	2	0
Email	13	20	5	2	5	0	0	0	5
date	14	5	5	0	0	0	0	0	3
city	15	0	0	5	15	15	0	0	0
entered_by	16	0	0	0	0	0	0	0	0
status	17	0	0	0	0	0	0	0	0
State	18	10	2	2	3	0	0	0	3

Fig. 4: DQ Issues observed in Data collected through Web Form

A. *DQI -1 Outliers*

Human errors such as errors caused during data collection, recording, or entry can cause outliers in data. For example: Annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher. Evidently, this will be the outlier value when compared with rest of the population. Another cause of outliers is experimental error. Measurement Error also gives rise to the outliers. This is caused when the measurement instrument used turns out to be faulty

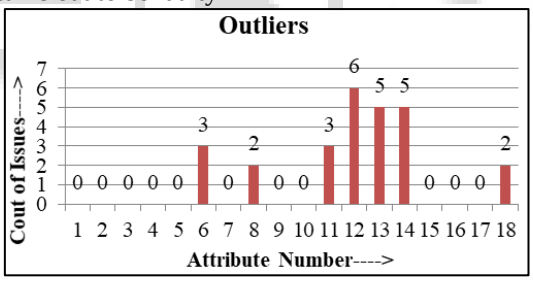

Fig. 5: Statistics of Outliers Observed

B. *DQI -2 Missing Values*

The reasons for missing values are clearly the data entry operators have left these values just because the data entry forms are not having proper checks and validations enforced on them. Missing Values and its problems are very common in the data cleaning process. Several methods have been proposed so as to process missing data in datasets and avoid problems caused by it.

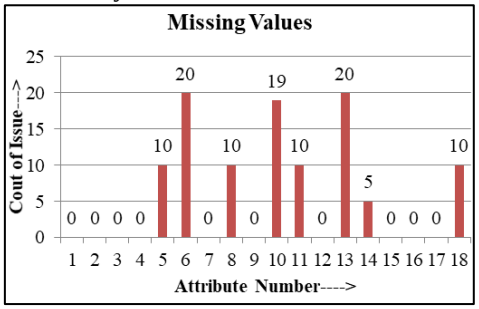

Fig. 6: Statistics of Missing Values Observed

C. DQI -3 Misfielded Values

Misfielded value problem occurs when the values entered are correct as far format is concerned but does not belong to the field. Example in field of city, value recorded is "PUNJAB". Primary reasons of these errors are because of free form fields designed in data entry forms where data is following rules of domain but values do not belong to these columns.

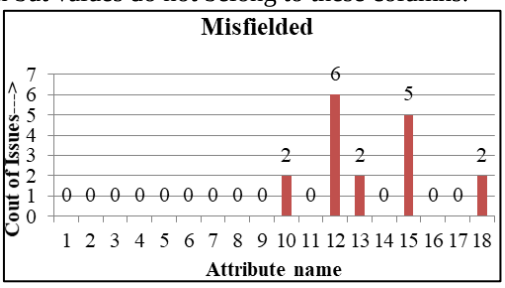

Fig. 7: Statistics of Misfielded Values Observed

D. DQI -4 Misspellings

Misspellings can occur in any of the column of data source. For example principle instead of principal, effect instead of affect, Jullandar instead of Jalandhar. Wrong data due to misspelling is obvious.

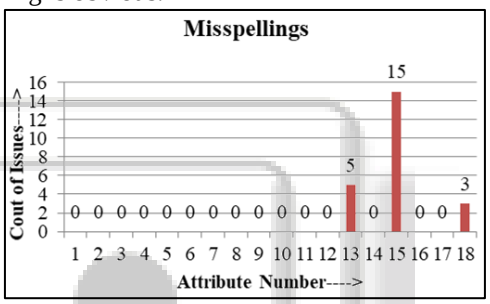

Fig. 8: Statistics of Misspelled data

E. DQI -5 Unique Violations

In our data source of admissions reference id is having unique constraint but no other column is having unique constraint because of the requirements of the system. But there is a column CONTACT_NO on which physically we can't put unique constraint but logically we don't want the values to be repeated in it because if one student is coming for enquiring for two or more courses he has to give different phone numbers for registration in different courses. But as there is no mechanism to check this ambiguity of repeated phone numbers therefore we have made this provision of our data quality tool to identify such repeated values in CONTACT_ NO attribute.

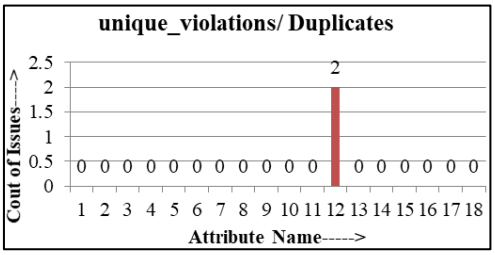

Fig. 9: Statistics of Unique Violations

F. DQI -6 Invalid Data

In our data Column no 10 and 13 are reported to have high number of invalid values. Students may have given wrong numbers which are not in use and may have told wrong name of the college from where they have done graduation or there may be typo errors as well or operators have typed email id which are not valid such as missing '@'or '.' Or having started with special character.

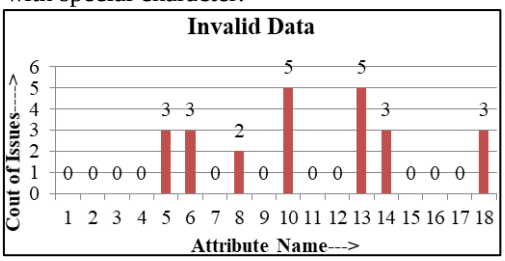

Fig. 10: Statistics of Unique Violations

V. CONCLUSIONS

The entire analysis is indicating towards the fact that organizations are maintaining their data sources very poorly. They prefer to record the transactions without paying attention to the quality of data captured during transaction recording. In educational institutions data sources data quality is highly required but managements are spending very little effort towards data quality. Primarily the crux of the research has highlighted following major reasons of poor data quality in data sources of the educational organizations

- Deliberate Data entry errors
- Poor training of data entry operators
- Poorly or wrongly designed data entry systems of data sources
- Manual entry of data into data sources.
- Lack of commitment of management for spending on data quality.

REFERENCES

[1] Eppler and Helfert, "A Classification and Analysis Of Data Quality Costs", Proceedings of the Ninth International Conference on Information Quality (ICIQ-04), Available: https://pdfs.semanticscholar.org/02ef/a0fb30d72a587d5531ddeb360f71e02c5704.pdf

[2] Richard Marsh "Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management", Journal of Database Marketing & Customer Strategy Management, January 2005, Volume 12, Issue 2, pp 105–112. Available: link.springer.com/content/pdf/10.1057%2Fpalgrave.dbm.3240247.pdf

[3] K. Singh & R. Singh "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010.

[4] Shaweta "Critical Need of the Data Warehouse for an Educational Institution and Its Challenges", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 4556-4559, ISSN No. 0975-9646. Available at: http://ijcsit.com/docs/Volume%205/vol5issue03/ijcsit20140503412.pdf

[5] A. Haug et. al. "The Cost of Poor Data Quality." Journal of Industrial Engineering and Management, 4(2), 168-193. Available: http://www.jiem.org/index.php/jiem/article/view/232/130

[6] Amit Rudra and Emilie Yeo, "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences –1999. Available: https://www.computer.org/csdl/proceedings/hicss/1999/0001/07/00017012.pdf
[7] Thomas C. Redman (1998). "The impact of Poor Data Quality on the Typical Enterprise." Communications of the ACM, Vol. 41, No. 2, 1998, 79-82
[8] Kim et. al. – Taxonomy of Dirty Data". Data Mining and Knowledge Discovery, 7. 2003. pp. 81-99.

