ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

Survey on Data Cleaning

Prerna S.Kulkarni, Dr. J.W.Bakal

Abstract*— DATA warehouse of an enterprise consolidates the data from multiple sources of the organization/enterprise in order to support enterprise wide decision making, reporting, analyzing and planning. The processes performed on data warehouse are highly sensitive to maintain the quality of data. They depend on the accuracy and consistency of data. Degraded quality of data leads to wrong conclusions of these processes which ultimately lead to wastage of all kinds of resources and assets. Data received at the data warehouse from external sources usually contains errors, e.g. spelling mistakes, inconsistent conventions across data sources, and/or missing fields. Significant amounts of time and money are consequently spent on data cleaning, the task of detecting and correcting errors in data. We aim to increase the awareness by providing a summary of the impacts of poor data quality on a typical enterprise. These impacts include customer disappointment, increased working cost, less effective decision-making, and a reduced ability to make and execute the plan. More subtly perhaps, poor data quality hurts employee morale, breeds organizational mistrust, and makes it more difficult to align the enterprise.

Index Terms—**Data, Homonyms, Synonyms, Quality, Data Warehouse.**

I. INTRODUCTION

Recent literature proposes several state-of-the-art solutions for matching and merging data sources. Relevant work and approaches exist in the field of data integration [3], [4], [7], [8], data deduplication [5]. As it is known, creating awareness of a problem and its impact is a critical first step toward resolution of the problem [9]. The needed awareness of poor data quality, while growing, has not yet been achieved in many enterprises. After all, the typical executive is already besieged by too many problems, low customer satisfaction, high costs, a data warehouse project that is late, and so forth. This article aims to increase awareness by providing a summary of the impacts of poor data quality on a typical enterprise. These impacts include customer dissatisfaction, increased operational cost, less effective decision-making, and a reduced ability to make and execute strategy. More subtly perhaps, poor data quality hurts employee morale, breeds organizational mistrust, and makes it more difficult to align the enterprise. Poor data quality and its underlying causes are potent contributors to an "information ecology" [10] inappropriate for the Information Age. Further, leading enterprises have demonstrated that data quality can be dramatically improved and the impacts mitigated.

One study estimates this combined cost due to bad data to be over US$ 30 billion in year 2006 alone [17]. As business operations rely more and more on computerized systems, this cost is bound to increase at an alarming rate. Companies use up millions of dollars per year to detect the errors in the data. Data quality refers that data is exactly fit for the purpose of business use; that it is consistent, accurate, complete and uniform. Cleaning of data refers to an activity which determines and detects the unwanted, corrupt, inconsistent and faulty data to enhance the quality of data [15].

II. DATA WAREHOUSE

The term Data Warehouse (DW) was coined by Bill Inmon in 1990, which he defined in the following way: "A warehouse is a subject oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision making process". Ralph Kimball provided a much simpler definition of a data warehouse as "a copy of transaction data specifically structured for query and analysis" [23]. Fig.1 shows the data warehouse architecture.

A. Data Acquisition

Data extraction is one of the most time-consuming tasks of Data Warehouse development. Data consolidated from heterogeneous systems may have problems, and may need to be first transformed and cleaned before loaded into the Data Warehouse. Data gathered from operational systems may be incorrect, inconsistent, unreadable or incomplete. Data cleaning is an essential task in data warehousing process in order to get correct and qualitative data into the Data Warehouse. This process contains basically the following tasks [24]:

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

- Converting data from heterogeneous data sources with various external representations into a common structure suitable for the Data Warehouse.
- Identifying and eliminating redundant or irrelevant data.
- Transforming data to correct values (e.g., by looking up parameter usage and consolidating these values into a common format).
- Reconciling differences between multiple sources, due to the use of homonyms (same name for different things), synonyms (different names for same things) or different units of measurement

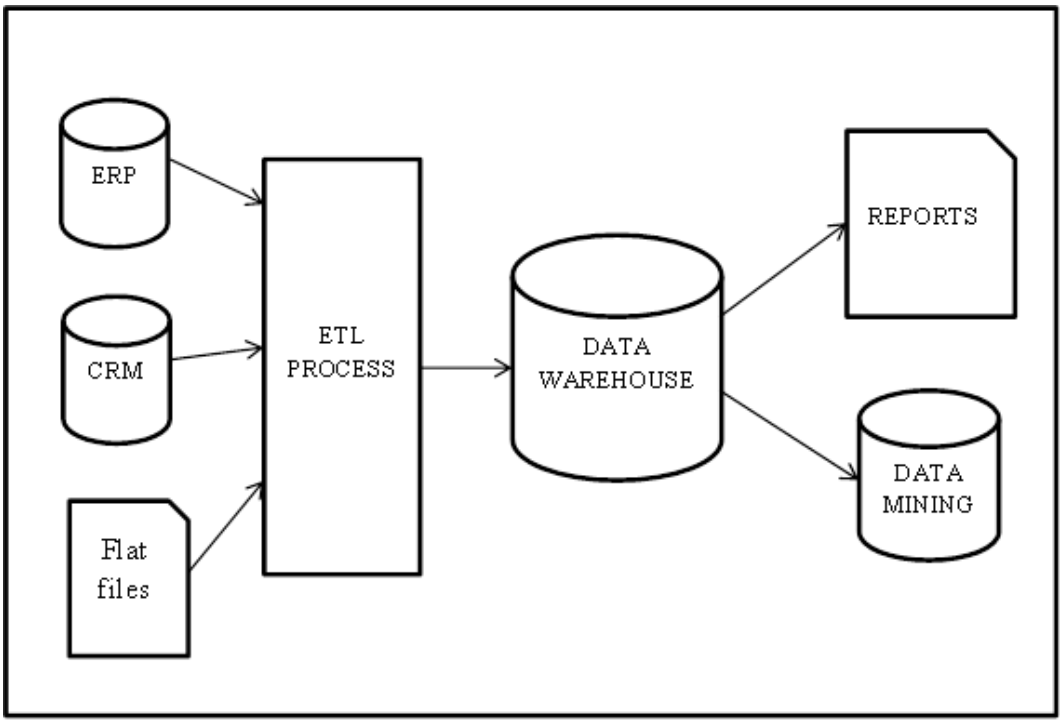

Fig. 1 Data warehouse [Source 26]

B. Extraction, Cleansing, and Transformation Tools

The tasks of capturing data from a source system, cleaning, and transforming the data and loading the consolidated data into a target system can be done either by separate products or by a single integrated solution.

III. DATA CLEANING FRAMEWORKS

Madnick et al. [19] analyze the TDQM framework, which advocates continuous data quality improvement by following the cycles of Define, Measure, Analyze, and Improve. Subsequent research developed theories, methods, and techniques for the four cycles of the TDQM framework. The strong points are: The TDQM is very easy to implement and manageable in enterprise environment for data cleansing. To define data quality from user perspective is also important that helps user to clean data that is fit for use in business. The short comings are: The proposed framework mainly focused for characterizing data quality research along the dimensions of topic and method rather than specifically on data cleansing framework.

Hao et al. [20] analyzes the framework as based on rules-base, rule scheduling and log management. Data cleaning process is divided into four parts: data access interface, data quality analysis, data transformation and results assessment. The strong points are: The framework design is unified as all data cleaning process performed at single place. The data access interface provides unified data extraction interface for single source and multi-source data. The process log management records the operation information of whole data cleaning process. The short comings are: The data quality analysis should be done only once and should not be a repetition work. The process should be sequential and in one go not iterative.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

As per Jebamalar et al. [21] the main objective of data cleaning is to reduce the time and complexity of the mining process and increase the quality of datum in the data warehouse. This new framework consists of six elements: Selection of attributes, Formation of tokens, Selection of clustering algorithm, Similarity computation for selected attributes, Selection of elimination function and Merge. As per Arora et al. [15] the algorithm proposed in the paper deals with the error of duplicity of data of string type in data warehouse in different data marts. The proposed algorithm was deduplication in the name field of a data warehouse. The strong points are: The proposed alliance rules for data cleansing can be easily changeable. The short comings are: This framework is only limited to duplicate records elimination of 'name' field. The manual intervention in this alliance rules for data cleaning is almost zero that makes this algorithm very less interactive from user point of view.

According to Yu et al. [22] the framework consists of access to database objects for user model, definition of user model and definition of quality model based on user model. The user model is a data model which is abstracted from the real model in perspective of the user. The strong points are: The proposed framework is simple and interactive as three steps process. The short comings are: This framework requires user model to work for cleansing that is basically lengthy process. This framework does not allow selection of attributes that actually makes processing very lengthy and wastage of time and resources.

IV. DATA QUALITY

Data quality is the degree to which data meet the specific needs of specific customers, which contains several dimensions. Poor data quality costs businesses vast amounts of money every year. Defective data lead to breakdowns in the supply chain, poor business decisions, and inferior customer relationship management. Data are the core business asset that needs to be managed if an organization is to generate a return from it. The following are characteristics and measures of data quality [23]:
Definition Conformance: The chosen object is of most important and its definition should have complete details and meaning of the real world object.

- *Completeness (of values):* It is the characteristic of having all required values for the data fields.
- *Validity (Business rule conformance):* It is a measure of degree of conformance of data values to its domain and business rules. This includes Domain values, Ranges, reasonability tests, Primary key uniqueness, Referential Integrity.
- *Accuracy (to the Source):* It is a measure of the degree to which data agrees with data contained in an original source.
- *Precision:* The domain value which specifies business should have correct precisions as per specifications.
- *Non-duplication (of occurrences):* It is the degree to which there is a one-to-one correlation between records and the real world object or events being represented.
- *Derivation Integrity:* It is the correctness with which two or more pieces of data are combined to create new data.
- *Accessibility:* Is the characteristic of being able to access data on demand.
- *Timeliness:* It is the relative availability of data to support a given process within the timetable required to perform the process.

V. DATA CLEANING

Data cleaning is the process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. This process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors [25].

A. Data Cleaning Problems

There exist severe data quality problems that can be solved by data cleaning and transformation. These problems are closely related and should thus be treated in a uniform way. Data transformations are needed to support any changes in the structure, representation or content of data. These transformations become necessary in many

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

situations, e.g., to deal with schema evolution migrating a legacy system to a new information system, or when multiple data sources are to be integrated.

B. Sources of Error in Data

Data errors can creep in at every step of the process from initial data acquisition to archival storage. An understanding of the sources of data errors can be useful in designing data collection, and in developing appropriate post-hoc data cleaning techniques to detect and ameliorate errors. Many of the sources of error in data fall into one or more of the following categories [27]:

- Data entry errors: It remains common in many environments for data entry to be done by humans, who typically extract information from speech or copying data from notebook, document or other written or printed sources. In these environments, data errors are often happened at entry time by typographic errors or misunderstanding of the data source.
- Measurement errors: In many cases data is intended to measure some physical process in the world: the depth of a bore well, the density of the mud, the value of the pressure, etc. In some cases these measurements are undertaken by human processes that can have errors in their design (e.g., improper surveys or sampling strategies) and execution (e.g., misuse of instruments). In the measurement of physical properties, the increasing proliferation of sensor technology has led to large volumes of data that is never manipulated via human intervention. While this avoids various human errors in data acquisition and entry, data errors are still quite common: the human design of a sensor deployment (e.g., selection and placement of sensors) often affects data quality, and many sensors are always worked in complicated environment, they were influenced by noise signals.
- Distillation errors: In many conditions, raw data are preprocessed before they entered into the database. This data distillation is done for a variety of reasons: to reduce the complexity or noise in the raw data. All these processes have the potential to produce errors in the distilled data, or in the way that the distillation technique interacts with the final analysis.
- Data integration errors: In almost all conditions, a database contains information collected from multiple sources via many ways over time. Any procedure that integrates data from multiple sources may lead to errors. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

VI. DATA QUALITY ISSUE AND DATA ACCURACY

Over the last several years, more and more references to poor data quality and its impact have appeared in the news media, general-readership publications, and technical literature [7, 11]. An enterprise may have a wide array of data quality problems. One way to categorize these issues is as follows [7]:

- Issues associated with data "views" (the models of the real world captured in the data), such as relevancy, granularity, and level of detail.
- Issues associated with data values, such as accuracy, consistency, currency, and completeness.
- Issues associated with the presentation of data, such as the appropriateness of the format, ease of interpretation, and so forth.
- Other issues such as privacy, security, and ownership.

The science of data quality has not yet advanced to the point where there are standard measurement methods for any of these issues, and few enterprises routinely measure data quality. But many case studies feature accuracy measures. Measured at the field level, error rates range wildly, with reported error rates of 0.5–30%. Naturally there are difficulties in comparing these error rates. For the purposes of this article, the following statements may be useful:

- Unless an enterprise has made extraordinary efforts, it should expect data (field) error rates of approximately 1–5%. (Error rate = number of erred fields/number of total fields.)
- That which doesn't get measured, doesn't get managed, so the enterprise should expect that it has other serious data quality problems as well. Specifically, enterprises have redundant and inconsistent databases and they do not have the data they really need.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

VII. CONCLUSION

Poor data quality impacts the typical enterprise in many ways. At the operational level, poor data leads directly to customer dissatisfaction, increased cost, and lowered employee job satisfaction. It also increases operational cost because time and other resources are spent detecting and correcting errors. With the wrong data the organizations morale can be lost.

ACKNOWLEDGMENT

We thank to all the authors for the information provided.

REFERENCES

[1] T. Redman, "The impact of poor data quality of typical enterprise", Communications of ACM, vol. 41, no. 3, pp.79-82, 1998.

[2] Rajiv Arora, Payal Pahwa and Shubha Bansal,"Alliance Rules for Data Warehouse Cleansing", 2009.IEEE Press, Pages 743-747.

[3] M. Hernandez and S. Stolfo, "The merge/purge problem for large databases," Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 127–138, 1995.

[4] M. Lenzerini, "Data integration: A theoretical perspective," in Proceedings of the ACM SIGMOD Symposium on Principles of Database Systems, 2002, pp. 233–246.

[5] R. Ananthakrishna, S. Chaudhuri, and V. Ganti, "Eliminating fuzzy duplicates in data warehouses," in Proceedings of the International Conference on Very Large Data Bases, 2002, pp. 586– 597.

[6] J.Jebamalar Tamilselvi, Dr.V.Saravanan,"Handling Noisy Data using Attribute Selection and Smart Tokens", 2008.IEEE Press, Pages 770-774.

[7] I. Bhattacharya and L. Getoor, "Iterative record linkage for cleaning and integration," in Proceedings of the ACM SIGKDD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004, pp. 11–18.

[8] W. W. Cohen, "Data integration using similarity joins and a word-based information representation language," ACM Transactions on Information Systems, vol. 18, no. 3, pp. 288–321, 2000.

[9] Kotter, J.P. Leading Change. Harvard Business School Press, Boston, MA, 1996.

[10] Davenport, T.H. Information Ecology. Oxford University Press, New York, 1997.

[11] Redman, T.C. Data Quality for the Information Age. Artech House, Boston, MA, 1996.

[12] E.Rahm, H.H.Do, "Data cleaning: problems and current approaches." IEEE Data Engineering Bulletin, 23(4), 2000, pp.3-13.

[13] J. Hipp, and U. Grimmer, "Data Quality Mining - Making a Virtue of Necessity" In Proc of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2001), Santa Barbara, California, USA, 200 1 pp.52-57.

[14] P.Vassiliadis, Z.Vagena, and S.Skiadopoulos "ARKTOS: A Tool for Data Cleaning and Transformation in Data Warehouse Environments" , Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 28, no. 4, pp. 42-47, December 2000.

[15] R. Arora, P. Pahwa, S. Bansal, "Alliance Rules of Data Warehouse Cleansing", IEEE , International Conference on Signal Processing Systems, Singapore, May 2009, Page(s): 743 – 747.

[16] S. Chaudhuri, K. Ganjam, V. Ganti, "Data Cleaning in Microsoft SQL Server 2005", In Proceedings of the ACM SIGMOD Conference, Baltimore, MD, 2005.

[17] Sang-goo Lee, Seoul Nat, Univ Seoul, "Challenges and Opportunities in Information Quality", E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2007, Tokyo, Jul 2007, Page(s): 481 – 481.

[18] Deaton, Thao Doan, T. Schweiger, "Semantic Data Matching Principles and Performance", Data Engineering - International Series in Operations Research & Management Science, Springer US, vol. 132, pp. 77-90, 2010.

[19] S.E. Madnick, Y.W. Lee, R.Y. Wang, and H. Zhu, "Overview and framework for data and information quality research", ACM, Journal of Data and Information Quality, Vol. 1, No. 1, Article 2, June 2009.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 4, July 2014

[20] Hao Yan, Xing-chun Diao, Kai-qi Li, "Research on Information Quality Driven Data Cleaning Framework", IEEE, FITME '08. International Seminar - Future Information Technology and Management Engineering, China, Nov 2008, Page(s): 537 – 539.

[21] J. Jebamalar Tamilselvi and Dr. V. Saravanan, "A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse", ACM, IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.5, May 2008, Page(s): 117 – 121.

[22] Yu Huang, Xiao-yi Zhang, Zhen Yuan, Guo-quan Jiang, "A universal data cleaning framework based on user model", IEEE, ISECS International - Computing, Communication, Control, and Management, Sanya, China, Aug 2009, Page(s): 200 – 202.

[23] T. Manjunath, S. Ravindra, and G. Ravikumar, "Analysis of data quality aspects in data warehouse systems," International Journal of Computer Science and Information Technologies, Vol. 2, No. 1, 2010, pp. 477- 485.

[24] B. Pinar, A Comparison of Data Warehouse Design Models, Master Thesis, Atilim University, Jan. 2005.

[25] I. Ahmed and A. Aziz, "Dynamic approach for data scrubbing process", (IJCSE) International Journal on Computer Science and Engineering, ISSN: 0975-3397., 2010.

[26] E. Rahm and H. Do, "Data cleaning: Problems and current approaches," IEEE Bulletin of the Technical -Committee on Data Engineering, Vol. 23, No. 4, December 2000.

[27] J. M. Hellerstein, Quantitative Data Cleaning for Large Databases, United Nations Economic Commission for Europe, 2008.

[28] Manning CD, Schutze H (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.