IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 1 (Jan. 2013), ||V1|| PP 75-76

Data Quality Tools for Data Warehousing: Enterprise Case Study

Er. Muheet Ahmed Butt[1], Er. Majid Zaman[2]

[1]*Scientist, PG Department of Computer Science, University of Kashmir, Srinagar, J&K, India*
[2]*Scientist, Directorate of Information Technology & Support Systems, University of Kashmir, Srinagar, J&K, India*

Abstract: Ensuring Data Quality for an enterprise data repository various data quality tools are used that focus on this issue. The scope of these tools is moving from specific applications to a more global perspective so as to ensure data quality at every level. A more organized framework is needed to help managers to choose these tools so that that the data repositories or data warehouses could be maintained in a very efficient way. Data quality tools are used in data warehousing to ready the data and ensure that clean data populates the warehouse, thus enhancing usability of the warehouse. This research focuses on the on the various data quality tools which have been used and implemented successfully in the preparation of examination data of University of Kashmir for the preparation of results. This paper also proposes the mapping of data quality tools with the process which are involved for efficient data migration to data warehouse.

I. INTRODUCTION

Data quality has two distinct aspects: one is the “correctness” of data (such as accuracy and consistency), and the other involves the appropriateness of data for some intended purposes. Data producers and users generally assume that the purpose of data quality assurance is to provide the best data possible. However, this obscures the need to evaluate data. The implication is that if a data set is the best available and is as good as it can be made, and then there are no other options than to use it. In this case, there is no point in worrying about just how good it can be made. The flaw in this is that merely saying that a data set is as good as it can be made does not tell us how good it is or whether it is any good at all. What may be considered good data in one case may not be sufficient in another case.

Data warehousing is now considered as the foundation of an enterprise information infrastructure. It is the repository where data of the enterprise is stored. It is imperative that the issue of data quality be addressed if the data warehouse is to prove beneficial to an enterprise. Corporations, government agencies (public or private) and not-for-profit groups are all flooded with enormous amounts of data. The desire to use this data as a resource for the enterprise has increased the move towards data warehouses. This information has the potential to be used by an enterprise to generate smarter and efficient understanding of their customers, processes, and the enterprise itself.

There combining of data with other sources prospectivestep towards increasing of the usefulness of the utilization of information in a proper manner. But, if the underlying data is not accurate, any relationships found in the data warehouse will be obviously misleading. For example, most student registration system requires a Registration Number of the student when setting up student information. If no or invalid number is available an invalid or no output is generated. If the student registration numbers are not changed, then some relationship may exist in the database, but the relationship would be misleading because the underlying data is inaccurate.

The steps for building a data warehouse or repository are well understood. The data flows from one or more source databases into an intermediate staging area, and finally into the data warehouse or repository. At each stage there are data quality tools available to massage, clean and transform the data, thus enhancing the usability of the data once it resides in the data warehouse which could be easily mined at later stage.The proposed research tries to address the various issues regarding the association between data quality tools and the data enabled processes so that quality data resides in a Data Warehouse.

II. DATA QUALITY IN DATA WAREHOUSES: REVIEW

Estimates as high as 75% of the effort spent on a data warehouse is attributed to backend issues, such as readying the data and transporting it into the data warehouse. Data cleansing activities account for nearly half of that time [1]. Hundreds of tools are available to automate portions of the tasks associated with auditing, cleansing, extracting, and loading data into data warehouses. Most of these tools fall into the data extracting and loading classification while only a small number would be considered auditing or cleansing tools. Historically, IT personnel have developed their own routines for cleansing data. For example, data is validated on data entry based for students evaluation module both internally as well as externally. Internally means data is validated at every step say whether the marks that are entered do not exceed the maximum or minimum marks. The external one deals with the double data entry which are to be matched later to ensures that data entered is correct and complete. Data quality tools are emerging as a way to correct and clean data at many stages in building and maintaining a data warehouse [3]. These tools are used to audit the data at the source, transform the data so that it is consistent throughout the warehouse, segment the data into atomic units, and ensure the data matches the business rules. The tools can be stand-alone packages, or can be integrated with data warehouse packages.Data flows from the source database into an intermediate staging area, and then into a data warehouse.

III. DATA QUALITY TOOLS

Data quality tools generally fall into one of three categories: auditing, cleansing and migration. The focus of this research is on tools that clean, audit and migrate data into a data warehouse.

1. *Data auditing* tools enhance the accuracy and correctness of the data at the source. These tools generally compare the data in the source database to a set of business rules. (Williams, 1997). When using a source external to the enterprise, business rules can be determined by using data mining techniques to uncover patterns in the data. Business rules that are

internal to the enterprise should be entered in the early stages of evaluating data sources. The data that does not adhere to the business rules could then be modified as necessary.

2. *Data cleansing* tools are used in the intermediate staging area. A data cleansing tool cleans those attributes of a database that can be compared to an independent source. These tools are responsible for parsing, standardizing, and verifying data against known lists. The data cleansing tools contain features which perform the following functions:
   - *Data parsing* breaks a record into atomic units that can be used in subsequent steps. Parsing includes placing elements of a record into the correct fields.
   - *Data standardization* converts the data elements to forms that are standard throughout the data warehouse.
   - *Data correction and verification* matches data against know lists.
   - *Record matching* determines whether two records represent data on the same subject.
   - *Data transformation*- ensures consistent mapping between source systems and datawarehouse.
   - *Householding* – combining individual records that have the same address.
   - *Documenting* the results of the data cleansing steps in the metadata.
3. *Data Migration* is used in extracting data from a source database, and migrating the data into an intermediate storage area. The migration tools also transfer data from the staging area into the data warehouse. The data migration tool is responsible for converting the data from one platform to another.

IV. FOLLOWED METHODOLOGY

The Examination Wing of Kashmir University declares around 900 odd results every year for around 376,000 Students appearing in both under and post graduate courses. Out of these 900 odd results around 30 results have the student count more than 25000[2][4]. Every student has to appear in around 9 to 15 Examinations for completing his under and post graduate program. A huge data warehouse has been designed and maintained for storing such big information. The list of the data warehousing tools which are used at every step of preparation of University results which are to then migrated to the warehouse is discussed below.

Data Entry Process: The first Step for preparation of the result is the data entry process. This process is carried out by around 60 data entry operators working on day and night shifts. In this process the data which are the subject marks is entered directly from award rolls which have been acquired from the concerned evaluators and is entered into the system. The data quality tools which are used in this process are Data Auditing, Data Cleansing, Data Standardization, Data Correction and Verification and Record Matching. Double data entry procedure is implemented so that proper accuracy is maintained.

Temporary Data Migration and Mapping: In this process the enrollment of the students is acquired from the warehouse who have applied for the said examinations with the courses they have to appear. Migration of this data entered in the Data Entry Process is carried out and mapped with the courses. The tools which are used to ensure data quality are Data Transformation, Data parsing, Documenting, Migration.

Statute Verification and Result Declaration: In this process the accumulated data from the last two processes of data entry and Data Migration and Mapping is analyzed with the existing business rules and metadata so that final result of the student is declared. The data quality tools which are used in this process are Data Auditing, Data Transformation, Householding, Data Standardization and Documenting. The result notification is the output of this process. Now the data is ready for final migration to the warehouse where it is stored permanently unless there is no modification done on the migrated data.

Final Migration to the Warehouse: In this processthe data which is in the form of a result notification is migrated to the main warehouse where it is stored permanent unless some modifications or changes are not required.
These processes are done for every result which university declares ensuring data quality at every level. Keeping the sanity of the information these tools play a very important role in maintaining data and ensuring data integrity.

V. CONCLUSION

Data quality tools are available to enhance the quality of the data at several stagesin the process of developing a data warehouse. Data Quality tools can be useful inautomating many of the activities that are involved in cleansing the data-parsing,standardizing, correction, matching, transformation ect. Many of the toolsspecialize in auditing the data, detecting patterns in the data, and comparing the data tobusiness rules. Once theseproblems have been isolated, the warehouse builder could determine which features ofthe data quality tools address the specific needs of the data sources to be used. The association of the data quality tool at every step of the data cleansing process plays a very important role in ensuring that sanity of the data is protected at every manner.

REFERENCES

[1] “Toxic Data”, Haggerty, N., DM Review Magazine, June 1998

[2] “DataWarehouse Implementation of Examination Databases” MA Butt, SMK Quadri, M Zaman, International Journal of Computer Applications 44 (5), 18-23, 2012

[3] Horowitz, A. (1998). “Ensuring the Integrity of Your Data”, Beyond Computing, May1998

[4] “Star Schema Implementation for Automation of Examination Records”, International , Conference on Computer Science, Computer Engineering and Applied Computing Las Vegas, USA, July 16-19, 2012 ISBN 1-60132-050-7O’

[5] Neill, P. (1998). “It’s a Dirty Job: Cleaning Data in the Warehouse”, Gartner Group,January 12, 1998