Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 1, January 2015, pg.64 – 73

RESEARCH ARTICLE

REVOLUTION IN DW BY SOLVING CAUSES OF DATA QUALITY PROBLEMS IN DW AND ETL

Mayuri Kirange
R.K.Makhijani
Department of Computer Engineering Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal
North Maharashtra University, Jalgaon
M.Tech IV semester, Dept. of CSE, S.S.G.B.C.O.E.T., Bhusawal
Associate Professor, Dept. of CSE, S.S.G.B.C.O.E.T., Bhusawal
mayurikirange2121@gmail.com, richa_makhijani@yahoo.co.in

Abstract— Data warehousing is gaining in eminence as organizations become awake of the benefits of decision oriented and business intelligence oriented data bases. Informed decision-making is required for competitive success in the new global marketplace, which is fraught with uncertainty and rapid technology changes. Decision makers must adjust operational processes, corporate strategies, and business models at lightning speed and must be able to leverage business intelligence instantly and take immediate action. Over the period of time many researchers have contributed to the data quality issues, but no research has collectively gathered all the causes of data quality problems at all the phases of data warehousing Viz. 1) data sources, 2) data integration & data profiling, 3) Data staging and ETL, 4) data warehouse modelling & schema design.

Keywords— Data Warehouse, ETL, Data Quality

I. INTRODUCTION

Data Warehouse plays an important part in the process of knowledge engineering and decision-making for Enterprise, as a key component of the data warehouse architecture, the tool that support data extraction, transformation, loading (ETL) is a critical success factor for any data warehouse projects Traditional methods of ETL development The existence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken. The one definition of data quality is that it's about bad data - data that is missing or incorrect or invalid in some context. A broader definition is that data quality is achieved when organization uses data that is comprehensive, understandable, consistent, relevant and

timely. Understanding the key data quality dimensions is the first step to data quality improvement. To be process able and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. Abundant attempts have been made to define data quality and to identify its dimensions. Dimensions of data quality typically include accuracy, reliability, importance, consistency, precision, timeliness, fineness; understand ability, conciseness and usefulness. To solve heterogeneity problems of different data sources in the processes, this seminar review framework models for optimization of the ETL processes by using semantic web technologies and discusses how ontologies are used to support the data integration. A metadata management System with good design can highly improve the ETL efficiency.

II. DW ARCHITECTURE AND END-TO-END PROCESS

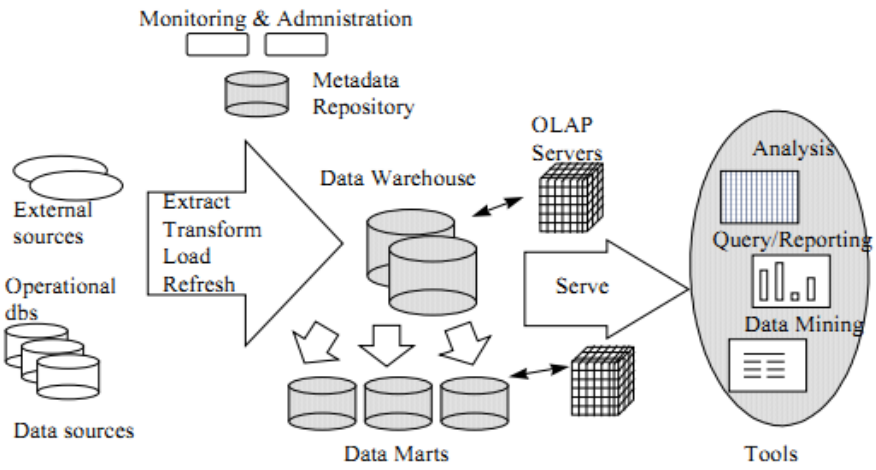

Figure 1. Data Warehousing Architecture

In the process of data-warehouse, the data moves from the external source systems through the ETL process using the capabilities provided by the ETL services layer. This flow is driven by metadata that describes the locations and definitions of the sources and targets, data transformations, timings, and dependencies. It includes tools for extracting data from multiple operational databases and external sources; for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage.

In cleaning phase Once the data is cleaned and aligned in the ETL process, the ETL system selects, aggregates, and restructures the data into business process dimensional models. These datasets are loaded onto the presentation server platforms in below fig and tied together via the conformed dimensions and conformed facts specified in the enterprise bus architecture.

Define all data accessed in any way by a business user, query tool, report writer, or software application as belonging to one or more business process dimensional models. In any case, it helps to separate the issues of ETL from data presentation because they have very different dynamics. In addition to the main warehouse, there may be several departmental data marts. Data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of front end tools: query tools, report writers, analysis tools, and data mining tools. Finally, there is a repository for storing and managing metadata and tools for monitoring and administering the warehousing system.

A. The Main Module of ETL

Data warehouses are one of the foundations of the Decision Support Systems of many IS operations. Data Warehouses can contain detailed data, lightly summarized data and highly summarized data, all formatted for analysis and decision support. Accurate way to design ETL process as in fig. to make it efficient, flexible and maintainable, ETL can be divided into 5 modules: data extraction, data validation, data cleaning, data conversion and data loading.

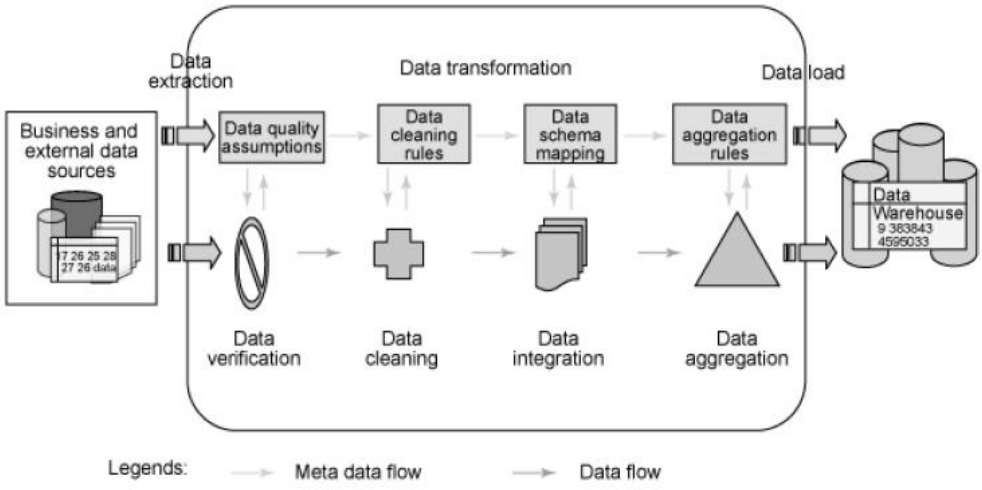

Figure 2. The Data Extraction, Transformation and Loading (ETL) process

1) Data extraction:

This step requires a great deal, first of all, need to find out which business system does the business data come from, what kind of database management system the business database server runs. Secondly, need to know what kind of table structure the database has and the corresponding meaning of each table structure. Thirdly, need to check whether there exist the manual data and the quantity of the data. Fourthly, define whether there exist the unstructured data. After collecting all of the information, give out he data explanation file.

2) Data validation:

Data validation involves a lot of checking work, including the effective value of the property, foreign key checking and so on. As for the low-quantity data, refuse them firstly, and then these data will be stored in order to be fixed in the field of the amendment.

3) Data cleaning:

The task of data cleaning is to filter out the undesirable data, and then send them to the business operation department. These undesirable data include: incomplete data, wrong data, duplicated data and so on.

4) Data conversion:

From the micro-details perspective, data conversion involves the following types: direct mapping, field operations, character string processing, null value determination, date conversion, date operation, and assemble operations and so on.

5) Data loading:

Data will be moved to the center of the target dataware house table, it is usually the last step in the process of ETL. As to the best way to load data, the implementation depends on the type of operation and the quantity of data. There are two ways to insert or update data in the database table: SQL insert/update/delete or batch loading application program.

B. Stages of Data Warehousing Susceptible to Data Quality Problems

A descriptive taxonomy of all the issues at all the stages of Data Warehousing. The phases are:

- Data Source
- Data Integration and Data Profiling
- Data Staging and ETL
- Database Scheme (Modeling)

Quality of data can be compromised depending upon how data is received, entered, integrated, maintained, processed (Extracted, Transformed and Cleansed) and loaded. Data is impacted by numerous processes that bring data into your data environment, most of which affect its quality to some extent. All these phases of data warehousing are responsible for data quality in the data warehouse. Despite all the efforts, there still exists a certain percentage of dirty data. This residual dirty data should be reported, stating the reasons for the failure in data cleansing for the same.

Data quality problems can occur in many different ways. The most common include:

- Poor data handling procedures and processes.
- Failure to stick on to data entry and maintenance procedures.
- Errors in the migration process from one system to another.
- External and third-party data that may not fit

With your company data standards or may otherwise be of unconvinced quality.

The assumptions undertaken are that data quality issues can arise at any stage of data warehousing viz. in data sources, in data integration & profiling, in data staging, in ETL and database modelling. The model is depicting the possible stages which are vulnerable of getting data quality problems.

III. ETL AGGREGATION POLICY DESIGN OF DATA WAREHOUSE

Aggregation is to operate with the data in fact table and store the results as aggregation table in a period, such as a month, one year. The goal of aggregation is the convergence of most important information, so as to decrease the amount of data and improve the query performance. Sometimes aggregation aims not at decreasing data amount, but introducing new items to make the data new process, and then to satisfy the higher requirement as response speed in complex query.

In data warehouse, not all table need aggregation. Generally the fact table, which is used frequently and data scale is large or the relations in aggregation are complex. Aggregation algorithm relies on query requirement and actual data. As a rule, in selecting dimensions of aggregation, some principles need be in line:

A. Transformation process of aggregation: Recalculation method

- Clear the content which need refresh in aggregation table
- Specified fact tables and all dimensions need to aggregate
- Link the dimension tables and fact tables
- Build a view, which includes: All external key used by common dimensions related with aggregation tale; All hierarchical properties in higher dimensions;
- Measurement need to aggregate in all fact tables
- Define the formula of aggregation
- Execute the aggregation and store the result in aggregation table.

B. Transformation process of aggregation:

Increment calculation method

Generally, to design the increment aggregation is relying on the following factors:

- The logic of aggregation calculation
- Calculation time
- Loading policy of star model

IV. CLASSIFICATION OF DATA QUALITY ISSUES

To determine the scope of the underlying root causes of data quality issues and to plan the design the tools which can be used to address data quality issues, it is valuable to understand these common data quality issues.

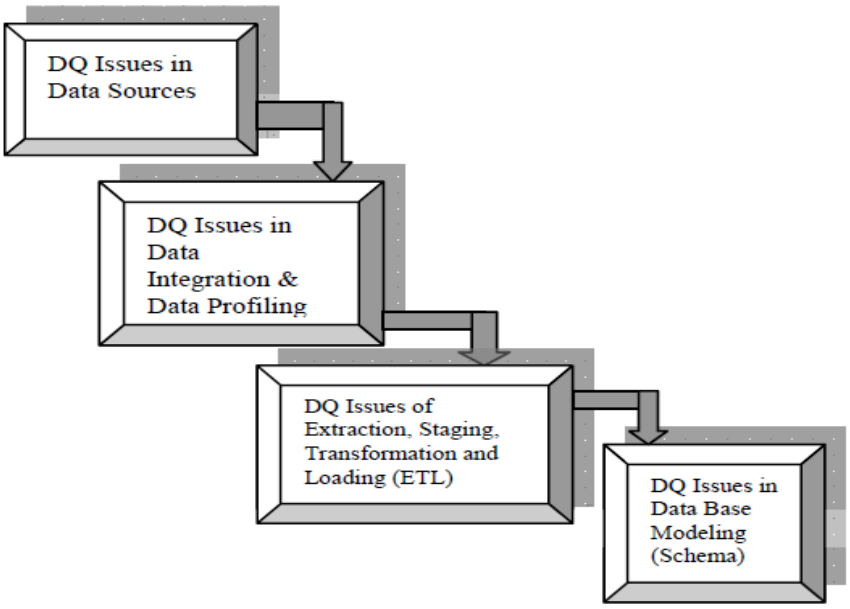

Figure 3: Stages of Data Warehouse Susceptible for DQ Problems

C. Data Quality Issues at Data Sources

A leading cause of data warehousing and business intelligence project failures is to obtain the wrong or poor quality data. Eventually data in the data warehouse is fed from various sources as depicted in the figure 4. The source system consists of all those 'transaction/Production' raw data providers, from where the details are pulled out for making it suitable for Data Warehousing. All these data sources are having their own methods of storing data. Some of the data sources are cooperative and some might be non cooperative sources.

Because of this diversity several reasons are present which may contribute to data quality problems, if not properly taken care of. A source that offers any kind of unsecured access can become unreliable-and ultimately contributing to poor data quality. Different data Sources have different kind of problems associated with it such as data from legacy data sources (e.g., mainframe-based COBOL programs) do not even have metadata that describe them. The sources of dirty data include data entry error by a human or computer system, data update error by a human or computer system. Part of the data comes from text files, part from MS Excel files and some of the data is direct ODBC connection to the source database. Some files are result of manual consolidation of multiple files as a result of which data quality might be compromised at any step.

D. Causes of Data Quality Issues at Data Profiling Stage

When possible candidate data sources are identified and finalized data profiling comes in play immediately. Data profiling is the examination and assessment of your source systems' data quality, integrity and consistency sometimes also called as source systems Analysis. Data profiling is a fundamental, yet often ignored or given less attention as result of which data quality of the data warehouse is compromised At the beginning of a data warehouse project, as soon as a candidate data source is identified, a quick data profiling assessment should be made to provide a go/no-go decision about proceeding with the project.

E. Data Quality issues at Data Staging ETL

One consideration is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse. A data cleaning process is executed in the data staging area in order to improve the accuracy of the data warehouse. The data staging area is the place where all 'grooming' is done on data after it is culled from the source systems. Staging and ETL phase is considered to be most crucial stage of data warehousing where maximum responsibility of data quality efforts resides. It is a prime location for validating data quality from source or auditing and tracking down data issues.

V. INVESTIGATION AND HANDLING ON TEMPORAL DATA WAREHOUSING

As a consequence of the fact that the decisional process typically relies on computing historical trends and on comparing snapshots of the enterprise taken at different moments, one of the main characterizations of data warehousing systems is that of storing historical, non volatile data. Thus, time and its management acquire a huge importance. In this discuss the variety of issues, often grouped under term temporal data warehousing, implied by the need for accurately describing how information changes over time.

These issues, arising by the never ending evolution of the application domains, are even more pressing today, as several mature implementations of data warehousing systems are fully operational within medium to large business contexts. In comparison with operational databases, temporal issues are more critical in data warehousing systems since queries frequently span long periods of time; thus, it is very common that they are required to cross

the boundaries of different versions of data and/or schema. Besides, the criticality of the problem is obviously higher for systems that have been established for a long time, since unhandled evolutions will determine a stronger gap between the reality and its representation within the database.

A. Handling changes in the data warehouse

This mainly has to do with maintaining the data warehouse in sync with the data sources when changes on either of these two levels occur. When considering temporal data, it is first of all necessary to understand how time is reflected in the database, and how a new piece of information affects existing data. From this point of view proposes the following proposes the following classification,

1) Transient data: alterations and deletions of existing records physically destroy the previous data content.
2) Periodic data: once a record is added to a database, it is never physically deleted, nor is its content ever modified. Rather, new records are added to reflect updates or deletions. Periodic data thus represent a complete record of the changes that have occurred in the data.
3) Semi-periodic data: in some situations, due to performance and/or storage constraints, only the more recent history of data changes is kept.
4)Snapshot data: a data snapshot is a stable view of data as it exists at some point in time, not containing any record of the changes that determined it. A series of snapshots can provide an overall view of the history of an organization.

B. Handling data changes in the data mart

Events are continuously added to data marts; while recorded events are typically not subject to further changes, in some cases they can be modified to accommodate errors or late notifications of up-to-date values for measures. Besides, the instances of dimensions and hierarchies are not entirely static.
Content changes result from user activities that perform their day-to-day work on data sources these changes are reflected in the data warehouse and then in the data marts fed from it. The multidimensional model provides direct support for representing the sequence of events that constitute the history of a fact: by including a temporal dimension (say, with date granularity) in the fact, each event is associated to its date. For instance, if we consider an ORDER fact representing the quantities in the lines of orders received by a company selling PC consumables, the dimensions would probably be product, order Number, and order Date.
Thus, each event (i.e., each line of order) would be associated to the ordered product, to The number of the order it belongs to, and to the order date. On the other hand, the multidimensional model implicitly assumes that the dimensions and the related levels are entirely static. This assumption is clearly unrealistic in most cases; for instance, considering again the order domain, a company may add new categories of products to its catalog while others can be dropped, or the category of a product may change in response to the marketing policy.
Another common assumption is that, once an event has been registered in a data mart, it is never modified so that the only possible writing operation consists in appending new events as they occur. While this is acceptable for a wide variety of domains, some applications call for a different behavior; for example the quantity of a product ordered in a given day could be wrongly registered or could be communicated after the ETL process has run. These few examples emphasize the need for a correct handling of changes in the data mart content. Differently from the problem of handling schema changes, the issues related to data changes

have been widely addressed by researchers and practitioners, even because in several cases they can be directly managed in commercial DBMSs. In the following subsections we separately discuss the issues related to changes in dimensional data and factual data, i.e. events.

1) Changes in Dimensional Data:

The study of changes in dimensional data has been pioneered by Kimball (1996), who coined the term slowly-changing dimension to point out that, differently from data in fact tables, changes within the dimension tables occur less frequently. Here proposed three basic modeling solutions for a ROLAP implementation of the multidimensional model, each inducing a different capability of tracking the history of data. The solutions discussed so far have different querying capabilities; with reference to the terminology proposed by SAP (2000), three main querying scenarios can be distinguished:

a) Today is yesterday: all events are related to the current value of the hierarchy. This scenario is supported by all the discussed solutions.

b) Today or Yesterday: each event is related to the hierarchy value that was valid when the event occurred. This scenario, that reconstructs the historical truth, is supported by Type II and VI solutions.

2) Changes in the Factual Data:

The transactional model, where each increase and decrease in the inventory level is recorded as an event, and the snapshot model, where the current inventory level is periodically recorded. A similar characterization is proposed by Bliujute et al. (1998), who distinguish between event-oriented data, like sales, inventory transfers, and financial transactions, and state-oriented data, like unit prices, account balances, and inventory levels. This has been later generalized to define a classification of facts based on the conceptual role given to events.

a) Flow facts: record a single transaction or summarize a set of transactions that occur during the same time interval; they are monitored by collecting their occurrences during a time interval and are cumulatively measured at the end of that period. Examples of flow facts are orders and enrolments.

b) Stock facts: refer to an instant in time and are evaluated at that instant; they are monitored by periodically sampling and measuring their state. Examples are the price of a share and the level of a river.

C. Handling schema changes in the data mart

The data mart structure may change in response to the evolving business requirements. New levels and measures may become necessary, while others may become obsolete. Even the set of dimensions characterizing a fact may be required to change. Schema changes in the data mart may be caused by different factors:

• Subsequent design iterations in the context of an incremental approach to data mart design.

• Changes in the user requirements, triggered for instance by the need for producing more sophisticated reports, or by new categories of users that subscribe to the data mart.

• Changes in the application domain, i.e., arising from modifications in the business world, such as a change in the way a business is done, or a changing in the organizational structure of the company.

• New versions of software components being installed.

• System tuning activities.

D. Querying temporal data

Querying in presence of data and schema hangers require specific attention, especially if the user is interested in formulating queries whose temporal range covers different versions of data and/or schema. The development of a model for temporal data warehousing is of little use without an appropriate query language capable of the DW effectively handling time. In principle, a temporal query could be directly formulated on a relational schema using standard SQL, but this would be exceedingly long and complex even for a skilled user. Golfarelli & Rizzi distinguish three querying scenarios in presence of late measurements:

• Up-to-date queries, that require the most recent measurement for each event;

• Rollback queries, that require a past version measurement for each event;

• Historical queries, that require multiple measurements for events, i.e., are aimed at reconstructing the history of event changes.

To cope with schema changes, Mendelzon and Vaisman (2000) proposed the Temporal OLAP(TOLAP) query language. TOLAP, based on the temporal multidimensional model proposed by Hurtado et al, fully supports schema evolution and versioning.

E. Designing temporal data warehousing

It is widely recognized that designing a data warehousing system requires techniques that are radically different from those normally adopted for designing operational databases (Golfarelli & Rizzi, 1999). On the other hand, though the literature reports several attempts to devise design methodologies for data warehouses, very few attention has been posed on the specific design issues related to time.

VI. Conclusions

Here mentioned all possible causes of data quality problems that may exist at all the phases of data warehouse. Our objective was to put forth such a descriptive classification

Which covers all the phases of data warehousing which can impact the data quality? It has attempted to think on near possible set of causes of data quality problems at all the phases at one attempt. The classification of causes will really help the data warehouse practitioners, implementers and researchers for taking care of these issues before moving ahead with each phase of data warehousing. It would also be helpful for the vendors and those who are involved in development of data quality tools so as to incorporate changes in their tools to overcome the problems highlighted in classifications.

It also reviewed the various optimization rules for ETL processes in order to solve the data integration problem of structural and semantic heterogeneity. Utilizing ontology makes ETL more flexible and efficient and the whole processes can be partially automated with the ETL rules in the framework.

ACKNOWLEDGEMENT

The completion of this paper would not have been possible without the support and guidance of Prof. R.K.Makhijani. She has been a constant of motivation and has encouraged me to pursue this seminar. With my deep sense of gratitude, I thank my respected teachers for supporting this of my seminar. This paper provides me with an opportunity to put in the knowledge of advanced technology. I thereby take the privilege opportunity to thank my guide and my friends whose helps and guidance made this study possibility.

REFERENCES

[1] Wayne W. E. (2004) "*Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data ",*The Data warehouse Institute (TDWI) report , a vailable at www.dw-institute.com .

[2] Tech Notes (2008), Why Data Warehouse Projects Fail: Using Schema Examination Tools to Ensure Information Quality, Schema Compliance, and Project Success. Embarcadero Technologies.

[3] Simitsis, A.; Vassiliadis, P.; Sellis, T.; "State-space optimization of ETL workflows", Knowledge and Data Engineering, IEEE Transactions on Volume:17 , Issue: 10 Digital Object Identifier: 10.1109/ TKDE.2005.169 Publication Year: 2005, Page(s): 1404 –1419.

[4] Zhuolun Zhang; Sufen Wang; "A Framework Model Study for Ontology-Driven ETL Processes", Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on Digital Object Identifier: 10.1109/ WiCom.2008.2651 Publication Year: 2008, Page(s): 1 – 4

[5] Wayne Eckerson, *"Data Profiling: A Tool Worth Buying (Really!)" ,Information Management Magazine, June 1, 2004 available http://wwwInformationmanagement.com/issues/20040601/1003990-1.html* Erhard Rahm & Hong Hai Do (2003) *"Data*

[6] *Cleaning: Problems and Current Approaches ",* available at: http://homepages.inf.ed.ac.uk/wenfei/tdd/reading /cleaning.pdf.

[7] Wayne W. E. (2004) "*Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data ",* The Data warehouse Institute (TDWI) report, available at www.dw-institute.com .

[8] Simitsis, A.; Wilkinson, K.; Dayal, U.; Castellanos, M.; "Optimizing ETL workflows for fault-tolerance",Data Engineering (ICDE), 2010 IEEE 26th International Conference on Digital Object Identifier: 10.1109/ICDE.2010.5447816 Publication Year: 2010 ,Page(s): 385 – 396

[9] Xiaoyun Liu; Jianhua Chen; Yingchun Zha; "Design of ETL Aggregation Policy in Local Tax System (LTS)", Management and Service Science, 2009. MASS '09. International Conference on Digital Object Identifier:10.1109/ICMSS.2009.5300877 Publication Year: 2009, Page(s): 1 – 4

[10] Amit Rudra and Emilie Yeo (1999) "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

