DATA QUALITY IN BUSINESS INTELLIGENCE APPLICATIONS

Gorgan Vasile

Academy of Economic Studies Bucharest, Faculty of Accounting and Management Information Systems, Academia de Studii Economice, Catedra de Informatica de Gestiune, Piata Romana, nr. 6, Bucuresti, 0724643651, gorganvasile@yahoo.com

Oancea Mirela

Academy of Economic Studies Bucharest, Faculty of Accounting and Management Information Systems, Academia de Studii Economice, Catedra de Informatica de Gestiune, Piata Romana, nr. 6, Bucuresti, 0722380309, mirelaoancea@gmail.com

To survive an organization must develop a strategy. To develop a successful strategy it must be capable to forecast the future circumstance. This is why nowadays business intelligence applications are essential for the success of a business. The decision support system is the eye through which the business strategist can look out on the organization's environment and detect behavior trends. Making decision on poor quality data can dramatically affect the strategy of the organization. This white paper addresses issues concerning data quality from business intelligence applications, the sources of poor quality data and possible ways to overcome these problems.

Keywords: business intelligence, data warehouse, data quality, ETL, data mining

Introduction

Business Intelligence is a broad category of applications and technologies used to collect, archive, analyze and access data, which helps users in decision making at an economic entity level. In fact Business Intelligence is an environment in which decision makers get reliable, consistent, comprehensible, easy to use and timeliness data. Using this data, decision makers can perform analysis that offers a broader view of the entity position in the past, present and the near future.

Therefore, why does an organization need Business Intelligence? To survive an organization must develop a strategy. To develop a successful strategy it must be capable to forecast the future circumstances. Understanding the past is the best method in trying to predict the future. This is the reason why information is considered the main ingredient of a strategy. The decision support system is the eye through which the business strategist can look out on the organization's environment and detect behavior trends.

Today the central element of business intelligence architecture is represented by the data warehouse, although there are people that use the terms "business intelligence" and "data warehousing" on an interchangeable basis. According to **Larissa Moss** business intelligence is a framework of *cross-organizational disciplines* and an *enterprise architecture* for the construction and management of an integrated pool of operational as well as decision support applications and databases that provides the business community *easy access* to their business data and allows them to make *accurate business decisions"* while "data warehousing is a subcomponent of and a vehicle for delivering business intelligence".

Issues concerning data quality in Business Intelligence applications

However the participants at the economic activity, the analysts and the clients set up an alarm signal about the decisions that are frequently made on the basis of data of low quality, data that is not set up to date because of the failure of the process of data cleaning. Knowing the impact of the poor quality of data it is "tormenting" to see the careless way in which most companies manage critical resources. Most of the companies don't create programs that produce quality data in a proactive, systematic and regular manner. According to a TDWI study, more than half of the companies don't have any plan to manage quality of the data.

The sources for the low quality data are uncountable. An important source is the process of data introduction that produces most of the problems and the systems interfaces. There is no wonder that the

employees at the introduction of data are blamed for most errors. In general the source of data errors falls into the following categories:

The lack of validation routines is a source responsible for bad data introduced on the Web or in the operational systems

Valid, but incorrect data: validation routines can miss the typing mistakes that respect the validation rules. A value can be valid but it doesn't mean it is also correct

Wrong syntax format and structure. The organizations try to introduce the data from more systems. In these cases, the ETL systems have to map these differences to a standard format before starting to clean the data.

Unexpected system changes. This situation occurs for instance when the database administrator adds a new field or a new code in the database and forgets to notify these changes to the systems administrator that makes the import

The multitude of interfaces. Complex architecture of nowadays systems leads to a multitude of interfaces which are difficult to update.

The lack of referential integrity check. In order to increase performance many administrators deactivate the check of the referential integrity when importing data.

Errors of data conversion. The programmers do not allocate enough time to understand the source and destination data model and, consequently, they write code that generates errors. One change in the migration program or in the interface systems can create thousands of wrong entries.

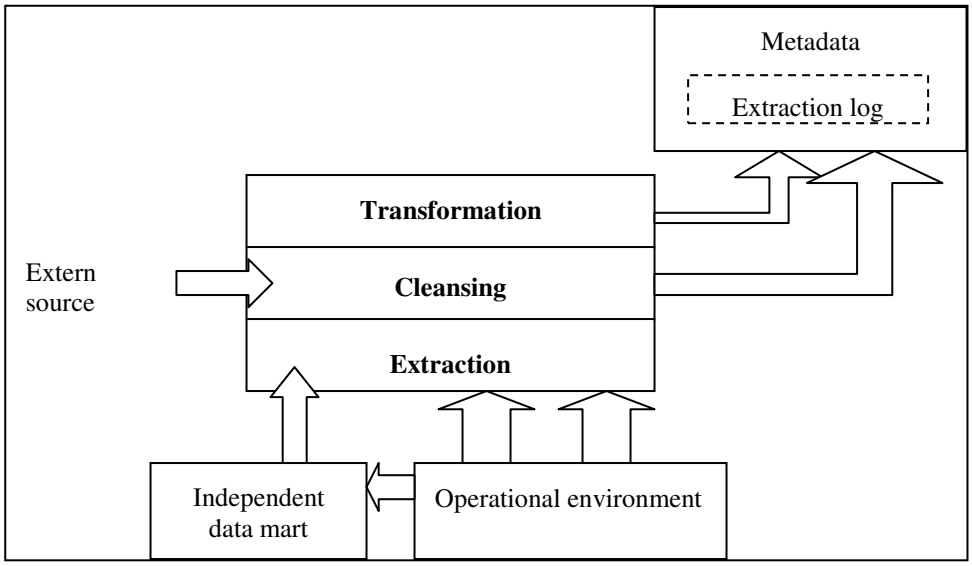
Fragmenting the definitions and the rules.A bigger problem comes from splitting the company into departments, divisions and operational groups, each of them with a different business process managed by distinct system. Slowly and unavoidable, each group starts to use slightly different definitions for common entities – "clients" or "suppliers" – and apply different rules for the computation of the same things – "net sales" and "profit before tax".

Slowly changing dimensions. Slow changes in the dimensions can create data quality problems according to the expectations of the users that watch the data. For instance, if an analyst wishes to compute the total of the fixed assets bought from one company in the last year, but this company merged with another one from which we also bought goods, problems may appear.

The ETL process and its role in data quality

ETL processes in a data warehouse environment extract data from operational systems, transform the data in accordance with defined business rules, and load the data into the target tables in the data warehouse. There are two different types of ETL processes: *initial load* and *refresh*.

The initial load is executed once and often handles data for multiple years. The refresh populates the warehouse with new data and can, for example, be executed once a month. The requirements on the initial load and the refresh may differ in terms of volumes, available batch window, and requirements on end user availability.



Extracting data from operational sources can be achieved in many different ways. Some examples are: total

Figure 1The ETL process

extract of the operational data, incremental extract of data (for instance, extract of all data that is changed after a certain point in time).

Data integration is the process of collecting data from the operational system in a central repository for analysis. Operational databases are the main source of a data warehouse and the integration process must assure a coherent environment for data used in analysis. The integration process consists of two operations: data transformation and data cleansing.

Data cleansing is the process in which errors are removed from the input data and it is a part of the integration process. It is probably one of the most critical steps of a data warehouse project. If the cleansing process is faulty, the analyst may not confide in the data warehouse and even a more serious scenario involve taking decision using bad data.

An efficient cleansing process can improve not only the quality of data from the data warehouse but also from the operational environment. The data administrator can verify the extraction log in order to identify the source of errors. Sometimes it is possible to detect errors that originate in the operational environment. Some errors can be generated by the original operational application or they can be simply typing errors. In both cases the data administrator must report these errors to the person responsible for data quality from the operational environment. Some errors can be generated by metadata, when the cleansing process doesn't intercept a metadata transformation or metadata used in the cleansing process are incomplete or incorrect.

There are debates concerning the actions that must be undertaken when input data mistakes are detected. Some consider this data must be returned to the operational environment in order to be corrected and send back to the data warehouse once the correction process is complete. Other thinks these errors should be corrected and integrated in the data warehouse whenever it is possible. Errors should be still reported to the operational environment. As a conclusion the data administrator must assure that there is a correspondence between data warehouse and operational environment. Otherwise a lack of confidence concerning the data warehouse can appear.

Data cleansing process cannot detect all errors. Some errors are simply typing errors. Other errors are more serious and put to test the data administrator competence. An example of such errors is the one in which the sales representative, instead of using each client unique identifier they use some generic identifiers that are accepted by the system.

Data cleansing is an important premise in a successful data warehouse project. The data administrator must have an active role in detecting and removing errors. While there is no ingredient that can guarantee the data warehouse success there are for sure some that can assure its failure. A faulty data cleansing process or a not very careful data administrator are certain premises of the failure.

Data transformation is the process in which data from operational systems are transformed into one consistent format. Each operational system contributing to the data warehouse must be analyzed to

understand data and their formats. Once these elements have been selected and defined, an integration process must be defined that will generate consistent data. Data transformation mainly concerns data description, data econding, the units of measure and data format.

The ETL process is one of the most expensive and time consuming component of a data warehouse development process. If a decade ago the majority of ETL were hand coded, market for ETL software has steadily grown and the majority of practitioners now use ETL tools instead.

Joy Mundy identifies a series of advantages and drawbacks of ETL tools. The main advantages are:

Structured system design. ETL tools provide a metadata-driven structure to the development team and this is valuable for teams that build their first ETL system

Operational resilience. Unlike home-grown ETL which present many operational problems, ETL tools provide functionality and practices for operating and monitoring the ETL system in production

Data-lineage and data-dependency functionality. Most analysts expect to be able to see the way a certain value from a report was obtained (its source, its transformation etc). Unfortunately few ETL tools supply this functionality.

Advanced data cleansing functionality. Most ETL tools offer either advanced cleansing and de-duplication modules (usually for a substantial additional price) or they integrate smoothly with other specialized tools.

Performance. It is not proven that using an ETL tool leads to increased performance. It's possible to build a high-performance ETL system whether you use a tool or not

There are also presented some **disadvantages** of ETL tools: Software licensing cost which varies from several thousand dollars to hundreds of thousands of dollars Uncertainty seen as misinformed ETL teams that sometimes are uncertain about what an ETL tool will do for them and reduced flexibility.

The key element of a successful ETL system is the practice of designing ETL system before development begins. Efficient ETL system implements standard solutions to common problems but also offer enough flexibility to deviate from those standards where necessary.

Conclusions

In order to make decisions that support the strategy of a business, the decision makers must confide in data they analyze. A series of studies discovered that many business intelligence projects failed because the poor quality of data in the phase of data warehouse's feeding. In order to increase the quality of data a series of steps must be followed. Among that we remember: **the launch of a program of data quality, the creation of a team for the quality of data, the reviewing of the business processes and the data architecture, continuous monitoring of the data, the use of intelligent application that integrate data mining for the validation of data.** We insist on the last one of the steps presented because it presents a series of advantages. First, the validation decisions are taken without needing any code to be written. The data mining algorithms learn the functioning rules of the entity directly from the data, setting the user free from the concern of discovering these rules and developing specific code for their description. On the other hand the data validation is made in different ways for each client. Using the data mining the rules are deducted from the client's data generating logic of validity that is automatically specialized for each particular client. Also the data mining process allows the application logic to be automatically updated in a simple processing process. The rewrite, recompilation and deployment of applications are not necessary because they are always available, even in the moment of processing.

REFERENCES

1. Frawley, W.; Piatetsky-Shapiro, G. and Matheus, C. 1992. *Knowledge Discovery in Databases: An Overview*. AI Magazine, 13(3):57-70.
2. De D. J. Hand, Heikki Mannila, Padhraic Smyth, 2001 *Principles of Data Mining*
3. Inmon W. H. *Building the data warehouse*, John Wiley and Sons USA,1996
4. William Giovinazzo *Object Oriented Data Warehouse Design*, Prentice Hall PTR, 2000
5. Larissa Moss, Shaku Atre *Business Intelligence Roadmap,* The Complete Project Lifecycle for Decision-Support Applications, 2003

6. Bob Becker *Kimball University: Data Stewardship 101: First Step to Quality and Consistency,* http://www.intelligententerprise.com/showArticle.jhtml?articleID=188101650
7. Andy McCue *Poor quality data is biggest CIO headache Making decisions on bad business intelligence data is recipe for disaster,* http://software.silicon.com/applications/0,39024653,39158648,00.htm
8. Jonathan G. Geiger *Ensuring Quality Data,* http://www.dmreview.com/issues/20070102/1072496-1.html
9. Michael L. Gonzales *Data Quality Discipline,* http://www.intelligententerprise.com/showArticle.jhtml?articleID=50500765
10. Michael L. Gonzales *Data Quality Audit, http://www.intelligententerprise.com/showArticle.jhtml?articleID=22102198*
11. Joy Mundy *Kimball University: Should You Use An ETL Tool?,* http://www.intelligententerprise.com/showArticle.jhtml?articleID=207002081