Data Validation as a Main Function of Business Intelligence

PhDr. M.A. Dipl.-Betriebswirt (FH)
Thomas H. Lenhard, PhD.
University of Applied Sciences Kaiserslautern
Campus Zweibruecken
Zweibruecken, Germany

Abstract

Thinking about Business Intelligence, Data Mining or Data Warehousing, it seems to be clear, that we generally use three levels to compute our data into information. These levels are validation of input data, transformation and making output data available. This study/paper will give a short overview about the importance of validation and the validation level of a Data Warehouse or any BI-System. Several examples shall prove that validation is not only important to get proper and useful output data. Some combinations of circumstances will show that validation might be the result that is requested. After an introduction concerning the levels of generating information and after explaining necessary basics about BI, different methods of validation will be discussed as well as it will be thought about different variations to use the validation level or the validated data. Examples from the practice of Business Intelligence in general and in the area of Health Care IT specifically will support the theory that the level of validation must be considered as the most important. Especially in the area of operational management, it will be proven that a profound use of data validation can cause benefits or even prevent losses. Therefore it seems to be a method to increase efficiency in business and production.

The Importance of Quality in Business Intelligence

Business Intelligence (BI) is a non-exclusive collection of methods that are used as a transformation process to generate information out of data.

As a hyperonym of data analysis, its usage is not limited to any special kind of business or appliance. Depending on the individual purpose of using BI, the requirements of quality may vary between wide ranges. But thinking about a (pseudo-) information like e.g. “the share

price will increase tomorrow – or not", it seems to be clear that a necessary minimum of information quality is expected when using BI. Business Intelligence should not end in it self and therefore, there must be a benefit when using it. Based on the amounts of data inside a company's data bases or data of any other institution like a ministry or a clinic, we can often face the problem that over years, all data is saved but no overview about state or trends is available [compare HKMW2001]. Such a missing overview results in a situation where we do not know, what we know [Bol1993, BG2006]. Therefore BI is needed to extract information and also knowledge out of an amount of data.

Thinking outside the box of Business Intelligence and looking at the scientific discipline of production planning and scheduling, Berning describes the interdependency between the aimed target, the parameters of a transformation process and the requirements about quality and quantity of its input [Ber2001].

Transferred to the field of BI, quantity can be understood in terms of the requirement of an average sample in statistics [see Puh1991]. In other words, in science and research it is commonly indefensible to calculate a trend if only two values are looked at. Therefore amount/quantity of data may also be a parameter of quality. Especially if raw data is taken out of a system that does not prevent input errors it seems to be absolutely essential to validate data before it is used in calculations and aggregations to be transformed to information. Figure one shows the transformation process in BI and that the data quality is improved on its way from raw data to information. In this figure the upper arrows describe the sequence that data is validated prior to be used for calculations and aggregations. The mid arrow comes directly from raw data towards the information. The reason for this process inside the transformation process is explained later in this paper.

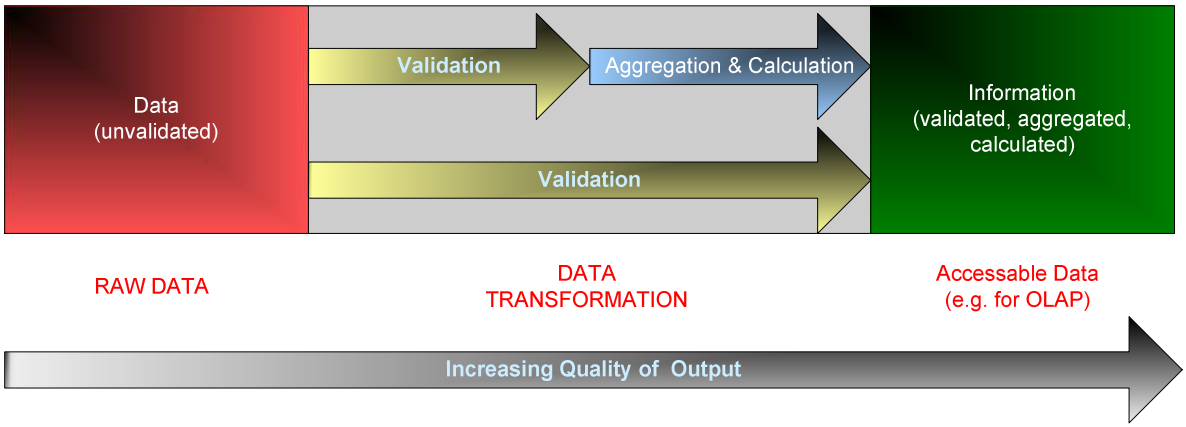

Figure 1: Abstraction of the Transformation Process inside a BI-System

Problems of Raw Data

Depending on the data base where raw data is extracted from to analyse it and the method used to record data in the system, there may be different sources of error. These might be especially [Len2012]:

- typographical errors
- transposed digits
- multiple data capture / redundancy
- incomplete data sets
- missing data sets

To further explain what may happen, if data errors are not eliminated before information is generated, let's consider a simple example, that is taken from daily practice: Due to an incompatibility between computer keyboards and soft drinks, a digit key stuck happened, so that as an amount which has to be registered for business statistics the value "22222" is stored in the data base instead of the real value "2". Normally, in this case, expected values that are used for a later analysis are found between "1" and "15". If we then analyse about 500 data sets and the average value without errors would be "3.5" we will get a calculated value of approximately "47.9" due to the key stuck error. This simple example proves how important it is to check the raw data before using it. Because even one statistical outlier, that is caused by a misentry can make the result of any analysis totally unusable. It's important to validate data prior to processing it. One method to prevent unusable results or information like shown in the example is described in a previous publication of the author. As a solution to such a problem a "range of security" can be defined [Len2010]. If any value does not hit this range the data set is taken into a table of exceptions, where an analyst has to decide if it will be deleted, changed or used without any alteration.
Other kinds of data validation can be used to check whether

- the amount of data sets is complete
- each data set is complete
- there are abnormalities
- there are trends (e.g. a accumulation of serious misentrys)
- there are misentry / key errors
- there are general data errors
- the data is plausible / makes sense

The validation method to be used depends on the parameters of each BI-project.

Operational Business Intelligence

Since Business Intelligence is not only limited to strategic data analysis like the future trends or analysing the difference between past fiscal years, it can also be used in the operational area of any business. Operational Business Intelligence (OBI) can be established as a system that goes hand in hand with Total Quality Management (TQM), because it can detect problems before they become a problem. At the first time this seems to be an anachronism but if we think about the seven basic tools (Q7) of Quality Management, we found one method that is called the quality control chart [TC2005]. By using OBI, the functional principle of such a quality control chart (QCC) can be implemented in the BI-System, so that manually or automatically measured values are analysed and trends will be identified. Especially if a trend seems to come closer to a warning threshold or even to an alarm threshold, the System can trigger warnings or even an emergency shutdown of a whole production line or in the context of health care, it can close an operation theatre, if any parameters of air conditioning seem to leave the allowed range in the near future. Therefore OBI can really be used to prevent waste of resources and to ensure a high degree of quality of services. Another real example on how OBI can improve business is shown in figure two. It shows a screen print from a clinical BI-system that is used to identify and eliminate data errors. As necessary background information it might be interesting, that hospitals in Germany get their payments by usage of diagnosis related groups as a kind of flat rate system. But if any highly expensive medicine is given to a patient, the clinic will get extra money for it. If the medical personal fails to enter that a patient was given an expensive drug, the hospital will loose a lot of money. To prevent this, the OBI on figure two reminds the controllers of the clinic, that there are still missing links between drug consumption and patients' data, so that they can track for the cause and so prevent the hospital from losses [Len2012].

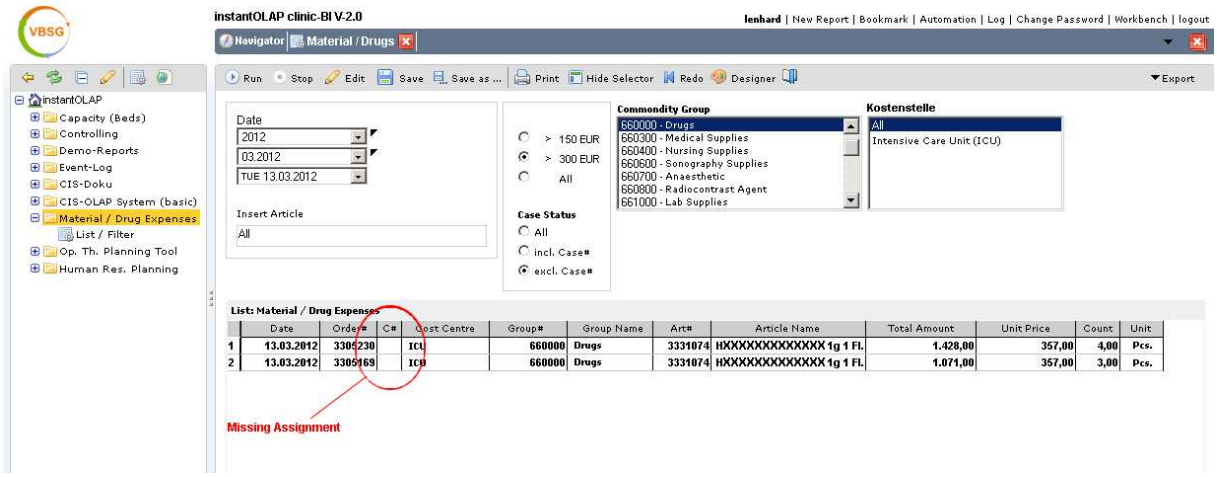

Figure 2: Operational BI in a Clinical BI-System (instantOLAP) [Len2012]

To list all possible fields of application for Operational Business Intelligence, it would fill more than one book. In this paper this idea is limited to the explained examples. Further benefits by use of OBI may be e.g. finding free resources, missing documentation, pending jobs or to optimise transportation and deadline monitoring. OBI seems to be as universally applicable as TQM. It can be used as a tool for TQM but in any case, it can save a lot of resources when used in operational business.

Importation of Validation

In the 3rd Millennium, some software companies still offer so called BI-systems that are not able to validate input data. That means nothing else than one basic function of BI is totally missing in such systems. Already hundreds of years ago, alchemists failed in making gold out of filth. In a transferred sense this would tell us, that the quality of output of any data transformation process depends directly on the quality of input data. As we have seen in a given example, information might be totally unusable if the validation of raw data is left. We use generated information for management decisions in operational business as well as in the field of strategic management. It is essential that we can trust this BI-generated information. Due to the danger of making wrong decisions and possible dire consequences caused by weak or unusable information, it is clear, that only validated data can be used for generating information or making decisions. The previous chapters proved that data validation is absolutely necessary for getting a proper output of a BI-System.

Conclusion

In this paper, it has been proved that making decisions require data or information on a sufficient level of quality. Such information can only be gathered inside a BI-system, if the processed data is also on an adequate level of quality. To ensure, that these requirements are fulfilled, it is necessary to validate raw data before computing any results. As shown, data validation can be the main intention of a transformation process inside a BI-system, especially in the field of Operational Business Intelligence. In this context BI can be used as a tool within the meaning of Total Quality Management that identifies problems before they become a real problem. In the beginning of this paper, transformation and validation were explained as levels of Business Intelligence. As a conclusion we can now state, that borders between these levels are not only floating. Validation has to be understood as a part of the transformation process and in some constellations it is the transformation process. But in any case, data validation is proved to be a main function of Business Intelligence.

Shortcuts

BI	Business Intelligence
OBI	Operational Business Intelligence
QCC	Quality Control Chart
Q7	Seven Basic Quality Tools
TQM	Total Quality Management

Literature

[BG2006] Beňová, E., Greguš, M.: Excel – Applied Software for Managers (in Slovak: Excel – Aplikačný softvér pre manažérov), Ed. Merkury sro., 2006, ISBN 978-80-89143-49-8

[Ber2001] Berning, R., Grundlagen der Produktion, Cornelsen Verlag, Berlin, 2001, ISBN-978-3-464-49513-1

[Bol1993] Bolz, N., Kann sich die Informationsgesellschaft eine Ethik leisten? – in Universitas No. 563, Wissenschaftliche Verlagsgesellschaft, Stuttgart, 1993

[HKMW2001] Hippner, H., Küsters, U., Meyer, M., Wilde, K. (editor), Data Mining im Marketing, Verlag Vieweg, Braunschweig/Wiesbaden, 2001, ISBN 3-528-05713-0

[Len2010] Lenhard, Th. A Possible Evolution from the Decision Support System to the Solution System for Business Usage, Verlag Kovac, Hamburg, 2010, ISBN 978-3-8300-5263-0

[Len2012] Lenhard, Th., Operative Business Intelligence in der Klinik, published in mdi – Forum für Medizin-Dokumentation und Medizin-Informatik 02/2012, Herausgeber DVMD e.V. und BVMI e.V., Heidelberg, 2012, ISSN 1438-0900

[Puh1991] Puhani, J., Statistik, Bayrische Verlagsanstalt, Bamberg, 1991, ISBN 3-87052-736-6

[TC2005] Theden, P., Colsman, H., Qualitätstechniken, Carl Hanser Verlag, München, 2005, ISBN 978-3-446-40044-3