

Reflections on BI

On the difficulty of keeping things simple

Katharina Wirtz

„The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.“ [Proust]

The genesis of the Business Intelligence (BI) concept lies in the desire to transfer the interrelations of data ‚trapped‘ in complex operating systems into a simply constructed analysis system.^[1] With this goal in mind, complex BI systems were created that, for all intents and purposes, rival the complexity of the operational systems they were designed to make more transparent. The original ambition behind these efforts, namely that of developing an easy and straightforward representation of company data to enable the recognition of relevant correlations, still lies in the distant future.

Why did the development of BI take this course?

BI concepts around data management, such as the star scheme or the cube, are conceptionally straightforward. This equation was made, however, without factoring in the user. Users do not leave the company’s BI world as pristine as it was initially modeled by the data warehouse architect. Instead, they create their own ‚individual‘ BI world by extending the existing system with external files and query results. Additionally, users exchange subsets of their BI worlds; these subsets in turn form the basis for further analysis and potential extensions of these worlds. This rank growth means that the originally ‚clean‘ Data Warehouse quickly morphs into yet another complex and opaque construct.

The contribution of the Data Warehouse towards a reduction in complexity should, however, not be belittled. Star schemes and cubes support a user’s typical operating methods by correctly recognizing the categories of dimensions and facts. It corresponds to human thought processes to think in terms of facts per dimension, i.e. revenue per branch per year per product. Why then this differentiation? Why do users create chaos out of a cleanly-structured warehouse? The answer becomes evident when regarding the typical usage scenario: queries are often not provided the necessary data by often pre-

^[1] Kimball R.: The Data Warehouse Toolkit, 2nd Edition, New York u.a. 1995, S. 310

defined data warehouse structures. To add to the complexity and depending on the question, data within different star schemes or cubes and aggregate levels as well as external data must be combined. In order to access data within different star schemes or clusters, they must first be linked to each other, which can be technically and semantically challenging. As the runtimes for such queries are often huge, it is advantageous to the user to save query results in order to re-use them during future queries. This will be approached in greater detail in the following example.

A user's thought processes usually move across the barriers of data silos. In analyzing buyer behavior, for example, it only makes sense to consider **all** customer transactions. However, many retail chains allow customer purchases through several channels – in store, by telephone, or through an internet portal. Analyzing the totality of customer transaction-relevant information generally requires insight into different information systems. However, integrating this data is an enormous problem alone due to the sheer size of those disparate data silos. Furthermore, data integration can be challenging due to semantically reasons.

Consider this example. An analyst in the healthcare industry wants to know what type of treatment general practitioners deliver when diagnosing depression and what type of medications they prescribe. This requires linking three data sets – ,treatment', ,diagnoses' and ,prescriptions', which can only occur when an attribute linking the data sets exists. This, however, is rarely the case. Treatment and diagnoses are available per doctor, day of treatment and patient. A treatment voucher will note which diagnoses were made and what types of treatment delivered, but will not correlate which diagnosis is connected to which treatment type. One can only ascertain that, in a particular case (one patient-doctor contact), a number of treatments and a number of diagnoses occurred. A similar set of problems occurs when considering pharmaceuticals. In order for a correlation to occur, those datasets must be aggregated at the case level. Aside from a depiction on the case level, analysts require results on the levels of individual doctor, patient, active pharmaceutical ingredients and treatment groups. The methods to establish these are time-intensive, so analysts by necessity will save interim results, methods and analyses paths. This increases the complexity of the base of information they structure their daily work around. In view of such scenarios, it is clear that the analysis needs of BI tool users are not adequately covered by traditional analytical data structures of stars and cubes.

Another reason for the complexity associated with BI systems is the fondness of both users and providers for attractive reports. The development of BI has increasingly become specialized in the depiction of what is already known, thereby moving further and further away from its original target – creating a user-friendly, complete and cross-organizationally integrated analytics system. In the meantime, the Business Intelligence community has realized that they are far from offering an integrated analytics system that encompasses the information gleaned from operational systems, and is now actively seeking to close this gap. Current integration solutions, however, generally don't offer

much more than the linking to each other of reports derived from isolated data silos. BI's actual charter, i.e. creating the logical and interpretative connections between an organization's various data sources, remains, in essence, outsourced to the recipient of those reports.

It is most likely only human to package the already known attractively and then desiring to recognize and understand that as something new. Discovering the truly new is most often challenging: the discovery process is vague and can be quite aggravating to the 'discoverer'. Progress, however, can only be made if one is willing to navigate through uncharted territory. Curiously, the act and process of discovery has not been adequately considered. Attractive graphics and nicely designed, interactive process steps are of course generally more attractive to users than the logarithms of data mining. Furthermore, data mining methods are usually complex and rigid, and have to date only provided unconvincing results. While the chimera of some type of computational logic that is fed with data and spits out interesting results is a stubborn one, this only really occurs in very narrowly defined applications. True insight remains a process driven by human intelligence and underpinned by iterative human-machine interaction, in which, based on assumptions and speculation, relevant data relationships are analyzed.

Instead of continuing to perfect the reporting around existing knowledge, we should develop and use systems that can enable new insight and therefore new courses of action.

In the BI arena, this requires, first and foremost, a fragmentation of the star scheme or cube. Analysts must be able to look at data that has been cross-linked. Traditional BI tools, however, force analysts to create such linkages themselves, assuming he or she is not relying entirely on pre-defined reports. This limits analysts in their ability to search intuitively, as they constantly need to ensure that their queries have also been depicted correctly from a technical point of view.

The linkages between data should be ensured by the analytical systems these analysts use. Traditional data structures cannot deliver this capacity, as correlation possibilities between data are too complex to adequately be represented by star schemas or cubes. When performing analyses, users relying on traditional BI tools therefore have no other choice but to predefine their queries as well as the solution path. This clearly represents a major constraint to the flexibility and spontaneity of the analysis process.

An entirely different approach is enabled by Munich-based software company Panoratio and its PANOsight for Lean Enterprise Data Warehousing Platform. Panoratio was founded in 2003 as a spin-off from Siemens' research lab. The company has developed multiply-patented technology that makes it

possible to perform in-memory analysis on enormous data sets. Automatic algorithms cluster data based on the similarities in their characteristics. This allows patients, for example, to be clustered according to traits such as age, gender and certain types of diagnoses, (i.e. women suffering from depression between the ages of 50 and 60, or obese 10-12-year old boys) who are highly likely to visit certain doctors or have received certain types of treatment. This lets analysts determine automatically what differentiates such a cluster from a reference cluster.

The creation of these clusters is determined algorithmically and provides probabilities from which the source data can be reconstructed at any time. These data clusters correspond to the categorization within cognitive recognition processes. According to Aristotle, categories are discrete entities characterized by a set of attributes that all of its elements have in common. By selectively perceiving the attributes that are meaningful for differentiation, people are capable of efficiently shaping their interactions and decision-making processes.^[2]

Because its technology emulates cognitive thought patterns, Panoratio is able to very efficiently store data (data compression within a PDI^[3] down to ca. 1% of the original data volume) as well as enabling in-memory access to that data in its compressed form. The compression process archives data in a tree structure within the PDI. Locating information within a tree structure is, by dint of its nature, significantly more efficient than in relational structures. This type of data structure can store a virtually unlimited number of dimensions and deliver answers to queries in seconds. Both factors are a prerequisite in the search for fresh insights – the large number of dimensions is necessary to comprehensively analyze correlations, the speed at which answers are provided crucial for maintaining a train of thought. In order to discover new correlations, the inquirer is dependent on his or her information flow not being interrupted too often or for periods of time that are too long, as disruptions in information flow impede the discovery process significantly. Generally, answering one question leads to a series of follow-up questions. Should these not be addressed quickly and interactively, and instead lead to database queries swallowing up hours, the discovery process is in danger of ceasing before the truly essential questions can be posed and answered. Should it, for example, be discovered during the course of an elaborate analysis effort that patients suffering from depression exhibit a higher number of diagnoses than those who do not, this would raise several questions. Do depressives have a higher number of diagnoses because depression makes them ‚sick‘? Do doctors not diagnose depression in these patients quickly enough? Is the depression not being treated effectively, causing depressives to seek out medical help in greater frequency? ... An analysis process abandoned too early because its

^[2] Jordan M. I., Russell S.: Categorization; In: The MIT Encyclopaedia of the Cognitive Sciences, The MIT Press, Cambridge, Massachusetts, 1999, S. 104-106.

^[3] The Portable Database Image (PDI) is a compact representation of the source information's complete statistical data. Using multiply-patented technology developed by Siemens Corporate Technology, Panoratio's software can generate and analyse PDIs.

expense in terms of time is too high most likely means that those insights truly necessary for effective decision-making in this regard will remain undiscovered.

Gaining insight is a challenging and often uncomfortable process. One navigates unfamiliar territory and cannot immediately surmise what advantages can be gained from those insights. Analysts must often fly solo when dealing with this challenge and are rarely duly recognized for either their work or its results. In addition to companies often not recognizing the challenges inherent in analysis work, there are hardly any effective tools to support it. The BI industry's initial promise – to show users previously unrecognized correlations within their organization's data through analytical means – remains unfulfilled to date. In reality, BI has limited itself to reporting around what is already known. We should pick up on the original concept; even it is a difficult one, and move forward towards undiscovered territory.