Tom Breur
runs consulting firm XLNT Consulting (www.xlntconsulting.com) dedicated to helping companies make more money with their data. His fields of interest span data mining, analytics, data quality, IT governance, data warehousing, and business models.

Data quality is everyone's business — Designing quality into your data warehouse — Part 1

Tom Breur

Received (in revised form): 14 April 2009

Abstract

In this information age with dramatically growing data volumes, data quality management is proving an avenue for creating sustainable competitive advantage. Combining data from disparate sources provides an opportunity to create new and valuable information. However, it also tends to surface previously existing, but so far unnoticed, data quality issues. To manage these challenges, we propose a data modelling paradigm (Data Vault) and a system development method (Agile), which provide the best alignment among stakeholders.

Journal of Direct, Data and Digital Marketing Practice (2009) **11,** 20–29.
doi:10.1057/dddmp.2009.14

Keywords: data quality, business intelligence, data warehousing, data models, system development methods

Introduction

The importance of data quality is gaining widespread attention. Reporting obligations for regulatory compliance, Sarbanes-Oxley act of 2002 (SOX), and the general need for better, more timely reporting has raised attention for data quality. Data quality plays a major role in the success of data warehouse (DWH) projects, and once your DWH is put to use, managing data quality requires ongoing attention.

We will address these aspects in a series of two papers. The first paper deals with consolidating disparate data sources in your DWH, and the second paper deals with ensuring ongoing data quality once your Business Intelligence (BI) solution is in place.

Data quality denotes its fitness for use. 'Fitness' refers to how suited data are for their intended use. 'Use' implies that how good data quality is, depends entirely on what you need the data for.

Data quality emerges from effective *usage* of data

DWH projects are only as successful as the *use* of data they are enabling. Reporting, Online Transaction Processing (OLAP), data mining, Customer Relationship Management (CRM), and Supply Chain Management (SCM) are some of the purposes for which data warehouses are used. Note that quality is not *inherent* in data, but follows when someone can *use* them successfully. Thus, 'quality' follows from the value data create after being put to use.

In two consecutive papers, the author will discuss how data quality can be achieved. The first paper is on designing DWH and Extract,

Tom Breur
XLNT Consulting,
Langestraat 8–03,
SE 5038 Tilburg, The Netherlands
Tel: +31–6–463 468 75
E-mail: tombreur@xlntconsulting.com

Transform, Load (ETL) processes to set up BI projects for success. We provide recommendations for software development methodologies that enable learning and success (Agile methods), and for a DWH modelling paradigm (Data Vault) that best supports initiatives to improve data quality. In the second paper, we assume that a DWH is in place and we focus on maintaining and improving data quality. This can be a one-time 'clean-up' exercise, but this could also be an ongoing data quality program.

Increasing competitive pressures

Markets become ever more dynamic and competitive, and there has been a slew of business books published that allude to this trend. In this regard, the books such as 'Blue Ocean Strategy',[1] 'The World is Flat',[2] and 'Good to Great'[3] come to mind. Some recent bestsellers that deal explicitly with the use of data to gain competitive advantage have been 'Competing on Analytics',[4] 'Supercrunchers',[5] and 'The Numerati'.[6]

Awareness of this information trend is rising, and corporations all over the world are reconsidering their information strategies. This is both out of necessity to stay competitive, and also because the changing information landscape offers opportunities they want to capitalize on. A recent study by Accenture showed a powerful link between organizations with pronounced analytical orientations and market out-performance. High performers were much more likely to value fact-based decision making and to have the skills and capabilities to effectively use analytics across their organizations.[7]

Proprietary data provide a means to sustainable competitive advantage

As companies are competing in ever more crowded market spaces, they seek ways to find a sustainable competitive advantage. Many sources of data are either publicly available (such as the census) or can be purchased (such as zipcode or lifestyle databases). This makes it very hard to build up and maintain a competitive edge: every time a 'new' solution is discovered and put in place, the competition will or can learn about this, and they will catch up. However, there is *one* source of data that the competition will never ever get a hold of, and that is data that are registered as a byproduct of dealing with or servicing your customers.

Proprietary data are therefore a source of sustainable competitive advantage: any value that can be extracted from such data can never be copied by the competition, and therefore corporations are particularly keen to maximize the use of proprietary data. It is for these reasons that utmost care must be taken to extract as much value from such data sources as possible, and to ensure good data quality. In this information age, building up and maintaining high-quality data are a strategic imperative.

Data quality and data warehousing

Data warehouse projects often fall victim to poor data quality

What are some typical quality issues that corporations are facing? DWH projects have a nasty habit of delivering late, and going over budget. Why is that? What challenges are you likely to face when trying to consolidate data streams from multiple legacy systems, each collecting data in different ways and for different purposes?

In some cases, (senior) managers may want to avoid involvement in such messy, non-linear projects such as data warehousing. We have observed, at times, that keeping IT at an arms' length provides a measure of safety in one's comfort zone. But just as often, nitty gritty details, such as record consolidation, name and address deduplication, etc, are just not perceived as the most 'sexy' topics to get involved with. Yet, it is exactly at this level that data quality is produced, and where really exciting issues around business alignment require attention and involvement.

What is data quality, really?

Data quality means relevance, timeliness, completeness, trust and accessibility besides accuracy

Olson[8] calls accuracy 'the mother of all data quality'. Although there are many aspects to data quality, Olson mentions relevance, timeliness, completeness, trust, and accessibility. All of these are useless if data are not accurate (enough)! But the other factors are important, too.

Data are only of value if they are *relevant*. The days are gone when DW projects could decide to read in all the data, 'just in case'. Oftentimes you see that wholesale copying of source systems brings in unnecessary data. ETL programmers may not fully understand the data model, and are afraid to miss any of the columns that might turn out to be vital. After the data warehouse has gone 'live', a Database Administrator (DBA) can monitor access to tables and even individual columns. Then, you often find that many parts of the DWH are rarely or never used. More could be said here, but the cost trade-off is clear.

Timeliness refers to having the data available as soon as possible, at least at a point in time when adequate value can be extracted. In online interactions, for instance, this applies when up-to-date customer details are required to support a flawless customer dialogue. When customers update their details, or pass on new address data, they expect a company to process this. It becomes even more confusing for the customer when the changed address is used in *some* communication, but this same change has not yet been processed by other parts of the company.

Completeness of data refers to the fact that unless one can be reasonably sure that all data have been assimilated when accessing a customer record, there could be serious issues. Imagine assessing the credit worthiness of a customer without taking into account the fact that he holds significant assets in stocks or savings. When you decline a demand for credit to this high net worth client, your customer might get very upset.

Only data that are *trusted* will get used for maximum benefit, and trust needs to be earned over time. One client we worked with had a corporate DWH with an interface to insurance data that were known to be of dubious quality. Any findings that pointed to a significant relation with insurances were always questioned, and therefore had little impact. By the same token, many other important relations might never even have been discovered for the same reason.

Accessibility has to do with not only having data, but also making them quick and easy to access, so that users can actually get their

hands on them. The principal difference between the On-Line Transaction Processing (OLTP) systems and a Decision Support System is that the former is made for getting data *in* quickly. The latter is designed to optimize getting data *out*. Too many DWHs are more like a data jailhouse: the data goes in but then is impossible to get out.

Why is data quality central to strategic success?

Trends that have amplified the importance of data quality are as follows:

— The total *volume* of data that corporations have at their disposal is growing rapidly.
— *Change* appears to be the only constant, and BI solutions have been challenged to keep up with this; *Agile* methods are adaptive, embrace change, and are therefore better geared for producing quality in rapidly changing environments.
— Increasing disappointment in Return On Investment (ROI) from DW and BI investments.

Volume

IDC (www.idc.com) conducted a study in 2007 and estimated the total volume of data worldwide at 281 exabytes (281 billion gigabytes). It is growing at a compounded rate of 60 per cent, which is projected to be nearly 1.8 zettabytes (1.800 exabytes) by 2011, a tenfold increase in 5 years. The time window for loading your DW is constant, yet the volume of data going in continues to grow. Many DWs are loaded every day. To avoid interference with users' access, this typically means that you have about a 6-hour loading window overnight. Loading volumes (and complexity!) grow, but the time to accomplish this remains the same. This trend clearly cannot continue so something has to give. Traditional DW models do not cope well with this stress. This phenomenon has been one of the development drivers of the Data Vault model.

Volume and proliferation of data sources is growing rapidly

Not only is the total *volume* of data growing, *new sources* of data continue to become available. New systems and new ways of interacting with customers (increasingly through digital channels) give rise to additional data feeds. Many companies are expanding the number of channels to interact through. Thus, besides face-to-face, mail, and telephony, you now have the web and mobile services. If you want to get a complete view on your customer's contact history, you need to take all possible channels into account.

The promise of CRM was that we would know the customer and would show we do in the way we interact with him. When you miss part of that contact history, you risk alienating your customers. Creating a Single Customer View has become more complex, and data needed to create it have grown in size and complexity.

Advances in data warehousing make connecting all these systems now feasible.

More data are also *needed* to support the increasing use of digital and interactive customer dialogue. Interactive digital connectivity is more economical as this decreases the reliance on (expensive) branch networks. It also adds value for the customer because the dialogue is available whenever and wherever the customer chooses to interact. Of course, all these data need to be of sufficient quality to support a satisfactory customer experience.

Change

Change appears to be the only constant, not only in business but, certainly, also in IT and BI systems. What was lacking, until fairly recently, was the technology and know-how to design BI systems and data structures that could change gracefully over time.

Star schemas struggle with changes in the grain of data

Dimensional modelling (Star schemas) has become the de facto standard for delivering data to end-users. However, Star schemas do not cope well with changes in the data structures and source systems. When multiple stars are connected through conformed dimensions, for instance, it becomes nearly impossible to change the grain without breaking the model. This inherent rigidity in the ETL process does not work well in a changing environment. This is why an Enterprise Data Warehouse (EDW) calls for an appropriate approach for data modelling: you need to be *prepared* for change.

Enormous strides have been made in this area, modelling EDWs that cope well with (un)expected changes. Data architecture is the number one problem in today's integration environments. Poor design and poor business practice have led to costly changes in the past, which contributed to poor data quality, and failed or abandoned DWH projects.

An IDC and DM Review (www.dmreview.com) survey revealed the implementation time of an average BI project to be 17 months, and the average project success rate of only 31 per cent. A Cutter survey found that 20 per cent of DWHs contributed no value, and only 15 per cent were called a complete success. This study found that only 27 per cent of respondents had confidence in DWH technology.

Waterfall methods are not well suited for data warehouse projects

Traditional 'fixed' requirements have proven unrealistic and ineffective for DWH projects. They are *unrealistic* because it is simply not possible for business users to imagine using a system that they have never seen, and to dream of what it would mean for their everyday practice of supporting ongoing BI needs. It has proven *ineffective* because time and time again the 'waterfall' design method delivers systems not only too late, but also imperfectly according to specifications. By the time these systems become ready for testing or use, the environment and requirements are guaranteed to have changed.

Given the uncertainty of the planning and the investments required for DWHs, management will be keen to put effective measures for risk management in place. There has to be a better way than has been achieved with waterfall methods, and there is. Scott Ambler, Practice Leader, Agile Development, IBM: ' … if you want to truly achieve a high-quality DWH that is responsive to the changing needs of its stakeholders, then you need to move away from traditional techniques and adopt an agile, Rational Unified Process (RUP)-based approach'.

Agile system development leads to better data quality

The history of system development paradigms shows some interesting shifts. We began with 'code as we go' in the 1960s until the first more formal system development methods appeared. As more and more structure was applied to development, this allowed for an approach that has similarities with 'ordinary' engineering in which the architect is not directly involved in the actual construction work. This structure allows relatively unskilled labour to be involved in the execution of the project.

From the 1980s and 1990s on this trend began to reverse: decoupling of architecture and building (programming) turned out not to work very well for development of complex systems such as DWHs. The trend evolved from early unstructured approaches to prescriptive methods, which stifled creativity, and currently we have moved to a period in which 'agile' methods are proving very effective — certainly for knowledge intensive development such as data warehousing. In the United Kingdom, DSDM[9] is one such flavour (www.dsdm.org), but there are many others. Extreme Programming (XP)[10] and SCRUM[11–13] are probably the largest.

Agile project methods embrace change

Agile methods (eg: SDL, DSDM, BDL, RAD, XP, Crystal, RUP, SCRUM, Lean Development, Adaptive Software Development, etc) share a number of traits. They are more 'people' centred (rather than process- or design-driven), adaptive, code-oriented (instead of document-oriented like technical specifications), there is no separation of design and build, testing is integrated in the development rather than following it, high-frequency feedback loops, and time and budget are fixed rather than the requirements as in the traditional methods. Thus, when in the old days we would build something according to plan, a good agile project will deliver something different — but better.

One of the outcomes of this shift is that much faster turnaround times between iterations (weeks or a month at the most) entice business users to provide more and better input. Iterations are smaller, and much, much shorter. And feedback from business users is dearly needed to produce quality: fitness for its intended use.

But there is no such thing as a free lunch. Agile methods require the business user's full-time involvement in building and designing; highly qualified and experienced programmers are needed, and at the outset the deliverables remain somewhat ambiguous, which can scare some people. It will lead to more dynamic changes in the business environment, and the cooperative model between business and IT that is required means that programmers need the mandate to make technical decisions within clearly defined boundaries. It also means that developers must really fancy this mode of working, because it places different demands on programmers as well.

ROI

Data warehouse and BI projects have a dubious track record of delivering on promises. There is increasing disillusionment in the business community about ROI that can be expected from such initiatives. In many cases, unforeseen (!) data quality issues caused

significant delays and budget overruns. Isn't it curious how this recurring pattern can continue to be 'unforeseen'?

Poor data quality often causes project overruns

Gartner suggests that more than 50 per cent of data warehouse projects will fail, and companies not using BI properly lose market share to those that implement and leverage BI correctly. Ted Friedman, their principal analyst, claims that this is due to poor data quality. 'Success in BI can be defined as the ability to add real insight to the business and enhance the decision making process', said Howard Dresner, VP and research director for Gartner.

Total quality management

In his landmark achievement, 'Improving Data Warehouse and Business Information Quality',[14] Larry English draws on the TQM framework and extends this to data warehousing, which he refers to as Total data Quality Management (currently he uses the term 'Total information Quality Management'). A fundamental notion in the information age is that data are no longer seen as a byproduct of business processes, but as a resource that has value beyond its immediate use. English stated, 'To deliver quality information, you must actively involve the information customers in the data modelling and design phase prior to implementing information systems', a thought that is central to agile software development methods.

English's main contribution, however, is a comprehensive model to categorize diverse sorts of (downstream) information non-quality costs. It turns out, time and time again, that preventing errors upstream is *overall* the least expensive way to fix problems. Data warehouse projects, however, rarely have the luxury of waiting until all data quality issues are resolved. This reality is dealt with by using the Data Vault modelling technology, which we further elaborate on in a later section.

As you are developing your data warehouse, you will most likely run into errors and inconsistencies. There might be customers in the source systems, for instance, who have paid a negative amount for their order. Or there will be customers with no orders, or orders with no customer, etc. Usually, these errors originate as a byproduct of either operating errors or dysfunctional systems. When an interface does not support the primary process, people will 'rig' the system, find workarounds to do their job *despite* some missing functionality. This then leads to incorrect entries.

The data warehouse team can work productively with business partners to identify and resolve data errors

Business owners know and hold the rules that data should conform to. They are also familiar with the interpretation of all or most of the 'illegal' codes in the database. The DWH team can add value by identifying the errors and reporting on them. They can point to the orders with negative amounts, they can say when these data were recorded, by whom, using which systems, etc. 'There is something about human nature that when we look at errors it makes us wonder: "Why did that happen?" and then delve into the causes of such faulty data', states Dan Linstedt, the inventor of the Data Vault.

One of the characteristics of the Data Vault is that data are always loaded 'as is' in the historical data store. 'So the good, the bad, and the

ugly data get loaded', states Dan Linstedt, 'and business rules are applied downstream as to what data can or cannot be shown in the "official" corporate reports'. This not only has the advantage that data remain auditable and traceable, but also that the business maintains ownership of 'their' data, errors and all.

'Because the Data Vault contains data "as is", you can perform gap analysis between the data and the business rules', states Dan Linstedt. In other words, you see the difference between the data and what they are expected to be. Any discrepancy between what the data are and what they are supposed to be can be surfaced.

This enables and facilitates investigations into the root causes of the errors, which in turn helps to change and improve business processes that caused these errors in the first place. In this way, the Data Vault helps to improve information and business process quality.

Value of data increases by merging disparate sources

Initially, the motivation for data warehousing was primarily to create a consolidated view of the business, and a 360-degree view of the customer. To this end, it is necessary to consolidate data sources from throughout the organisation. It turns out that merging data produces *new* data of potentially higher value, as properties that are merged can be related with new types of aggregations, analyses, and correlations.[15] Hence, the quality of data *increases* by compiling disparate data sources.

For example, when you combine data from finance or sales with logistics, you will be able to drive out process inefficiencies and optimise your procurement strategies. Or you can combine address data with zip-code databases, and merge them with a detailed recording of surf behaviour on your website. Together these can generate a lifestyle/psychographic micro-segmentation that allows behavioural targeting, aimed at precisely hitting niche audiences. The possibilities are endless. Once you begin to develop these analytic skills, new opportunities continue to emerge in unexpected places.

Data warehousing: A *joint* effort

Data have become a major source of strategic advantage. The arduous task of consolidating data can no longer be left to IT, at least not to IT *alone*. Not only is ongoing input on strategic priorities required, governance of data warehouse projects is an extremely important success factor to create maximum value from all available data throughout the corporation.

Demands on IT have gone up and we are finally learning how to model and architect the systems to meet the demands for more value and faster delivery. These demands have led to a new data-modelling paradigm, the Data Vault. Bill Inmon, a long-time data warehouse guru states that 'The Data Vault is the optimal choice for modelling the EDW in the DW 2.0 framework'.

The Data Vault defers data transformations downstream

Using the Data Vault modelling paradigm, you defer most data transformations downstream, when moving data from the historical data store to the data mart. One of the benefits from this is that the business still owns the data as it is loaded more-or-less unchanged into your

historical corporate data store. In addition, auditability and traceability are greatly improved because business users can 'see' their own data, and business rules are applied as close as possible to where they originate: in the business. The business has greater control, but, therefore, also greater responsibility to ensure that data are entered properly the first time around.

This new way of modelling 'naturally' gravitates toward better business alignment between IT and business partners. The business retains ownership of the data they supply and stays much more in control of the development and application of business rules. Because of increased participation and ownership, they will also become a more valuable partner and more critical user. The end result is higher quality data models, leading to better systems and data.

Conclusion

Data quality means fitness for its intended use. Oftentimes, data quality issues arise because data from disparate sources are merged and recombined. Although the data may (apparently) have been of adequate quality in the underlying legacy systems, the fact that data are put to use for *new* purposes can surface non-quality, which was never apparent before.

Recombining existing data sources creates new information

When source data are (re)combined in the data warehouse, fundamentally new information is produced, which adds value to the enterprise. We have moved from seeing data as a byproduct of business processes to the notion of attributing value to information *beyond* its immediate use.

Data quality is, or should be, everyone's concern. This all-important topic cannot be left to IT or your BI department. Data volumes around us are exploding and corporations rely on the intelligent use of data to create sustainable competitive advantage. This makes extracting the most possible value from (in particular) proprietary data a strategic imperative.

To ensure data warehousing success, as well as maintaining quality data throughout the lifecycle of the BI systems, Data Vault modelling has proven effective. According to Bill Inmon, it is the method of choice for an EDW in the DW 2.0 framework.

An important innovation for data modelling is the Data Vault. This model can cope much better with changing requirements and source systems, and ever-increasing volumes of data that enter BI applications. An added benefit is better alignment between business and IT because business partners maintain ownership of 'their' data, thus naturally enforcing better quality data.

Agile development methods have proven their value for data warehousing and BI applications. This is in part because of the dynamic nature of markets, and hence the expected (and ongoing) change in requirements. Equally important is the fact that they enable a much closer and more fruitful cooperation among business partners.

BI provides most value in between IT and business users

BI plays an interesting role, smack in the middle between 'pure' IT professionals and business users. BI tends to be very familiar with

caveats in data provision, and is usually aware of 'known' data quality issues. It is their responsibility as an impartial player inside the business to expose the truth as it appears within the data. They need to help business users who own the information systems and business processes, to understand where their systems may be broken or non-compliant. This can pose a considerable tension, as business owners may have a different idea about what is really happening.

References

1. Chan, K. W. and Mauborgne, R. (2005) *Blue Ocean Strategy*, Harvard Business School Press, Boston.
2. Friedman, T. (2005) *The World is Flat*, Farrar, Straus and Giroux, New York.
3. Collins, J. (2001) *Good to Great*, Harper Collins, New York.
4. Davenport, T. and Harris, J. (2007) *Competing on Analytics*, Harvard Business School Press, Boston.
5. Ayres, I. (2007) *Supercrunchers*, Bantam, New York.
6. Baker, S. (2008) *The Numerati*, Houghton Mifflin Harcourt, New York.
7. Todd. (2008), DM Review, June.
8. Olson, J. (2003), ref. 2 above.
9. Stapleton, J. (1997) *DSDM — Dynamic Systems Development Method*, Addison-Wesley, London.
10. Beck, K. (2003) *Extreme Programming*, Addison-Wesley Verlag, London.
11. Schwaber, K. and Beedle, M. (2001) *Agile Software Development with SCRUM*, Prentice Hall, NY.
12. Schwaber, K. (2004) *Agile Project Management with SCRUM*, Microsoft Press, Redmond, WA.
13. Schwaber, K. (2007) *The Enterprise and SCRUM*, Microsoft Press, Redmond, WA.
14. English, L. (1999) *Improving Data Warehouse and Business Information Quality*, Wiley, New York.
15. Batini, C. and Scannapieco, M. (2006) *Data Quality — Concepts, Methodologies and Techniques*, Springer-Verlag, Berlin.