White Paper

Bloor

The business case for Data Quality

A White Paper by Bloor Research
Author : Philip Howard
Publish date : March 2012

... there is much to be said in favour of a platform-based approach to data quality

Philip Howard

The importance of good data quality

Forbes Insights published a report in 2010 stating that "data-related problems cost the majority of companies more than $5 million annually. One-fifth estimate losses in excess of $20 million per year." The question is: where do those losses accrue? In practice, the value of having accurate information is important in four areas: for operational purposes, for decision making, in order to support regulatory compliance and for a variety of technical reasons. We shall consider each of these in turn.

Operational significance

In 2010 the British Army won the Data Governance Best Practice Award (presented by DebTech International, Wilshire Conferences and TDAN.com) for its use of data quality with respect to personnel records. As an illus tration of why this is important, consider that a private might actually be designated as a trooper, artilleryman, driver, guardsman, marine and so on, depending on their regiment and skillset. If this data is not accurately recorded individuals could easily be wrongly assigned to tasks, both in the barracks and on the battlefield, with implications that are all too obvious.

Another example relates to a data quality survey conducted by GS1 (UK) in conjunction with IBM and published in "The Data Crunch Report" published in 2009. This research compared the product records of the UK's four largest supermarkets with the corresponding records of their four largest suppliers. More than 60% of the supermarkets' records had errors in them. GS1 estimated that the costs associated with this inaccurate data would amount to a total of £1bn for the four companies over a 5 year period, based on the costs of remediation, overstocking, under-stocking and lost sales opportunities. Note that the allocation of products to supermarket shelves is exactly parallel to the assignment of troops in the army. A similar scenario would apply in other context such as the allocation of parts to an assembly.

Decision making

One of the difficulties with business intelligence is that the relationship between decision making and known facts is a nebulous one. It is entirely possible to have all the facts to hand and still draw an erroneous conclusion. On the other hand, we are all familiar with the concept of making "the right choice for the wrong reasons".

To illustrate the problems associated with decision making due to poor quality data consider the position that the Germans were in during the spring of 1944. They concluded, on the basis of the information that they possessed, that the allied invasion would take place at the Pas de Calais. As a result of this, they placed the bulk of their forces in that region. In fact, Operation Overlord was launched against the beaches of Normandy. The Germans, of course, were not to know that they had been fed false intelligence but, had they had some measure of the quality of the data upon which this decision was made, they might well have come to a different decision and changed the course of history.

So, the quality of the data that you have can profoundly affect your judgement. However, that is not all. The degree of inaccuracy that exists can also be important. Suppose that you are trying to reduce the number of payment defaulters that arise in the course of your mail-order service. On a turnover of, say, $500m per annum, an improvement from 95% to 96% of customers who complete their payments, can be very significant. So you implement an appropriate data mining technique to try to establish a correlation between defaulters and other known facts about these people. However, if you have a 10% error rate in the quality of your data, this will dwarf the possible improvements that you can achieve. On the other hand, imagine that the Germans had had rather better intelligence and that they could estimate the chances of a Normandy landing at, say, 80%. What difference would an error rate of 10% make? In this case: not much.

A more recent example of how better information can lead to additional profits: Sallie Mae won the 2011 Data Governance Best Practice Award, having implemented a data quality initiative in 2010 aimed at better targeting of potential clients, based on improved data quality. It estimates that this enabled an increase in revenue of over $2m based on increased loan volumes of around $50m. At the same time the company saved between $4m and $5m by switching from postal to email-based marketing, which was similarly enabled by its use of data quality. The converse of this sort of example, of course, is that poor data quality can negatively impact marketing. For example, if you are running an email campaign and you

The importance of good data quality

have a 0 instead on an O in an email address then your marketing will not reach your intended recipient. A bad impression can also be created if multiple mailings (of whatever sort) are directed to the same person, because of duplicated data in your marketing database.

Finally, good data quality can make it possible to accurately report on, and analyse, information where that was not previously the case. For example, Emerson Power has manufacturing plants all around the world, which historically ran independently. When the company started to implement data quality processes, it discovered that it had multiple plants manufacturing the same product(s), even though these had different product codes and different descriptions (in different languages). After implementing a data quality programme it was able to match these products and, for the first time, actually quantify the sales of the said product(s). As corollaries:

- Once you have this information you can start to consider manufacturing plant consolidation. Multi-sourcing may be fine but seven or more plants making the same product (which was the case with some Emerson products) may be too many. A similar argument would apply to supply chain consolidation.

- Emerson reported a significant reduction in staff turnover after cleaning up their data. This is logical: if you have to make business decisions (which you will be blamed for if they go wrong) based on data that you don't trust then that causes stress. Stress results in low morale and low morale leads to high staff turnover.

Compliance

The EU's Solvency II regulation for the insurance sector requires that data be "accurate, complete and appropriate". In effect, the legislation requires that data quality procedures be in place. The forthcoming MiFID II regulation for capital markets uses the same terminology. We expect more regulators to follow suit. The Financial Services Compensation Scheme in the UK also requires the provision of accurate information from deposit holders. Of course, compliance for its own sake is one thing but good data quality can save you money. For example, one European bank was able to save Ð50m on its capital adequacy requirements under the Basel II regulations when it was able to prove the reliability of its data.

A completely different example is, again, the British Army. It discovered, as part of its data quality exercise described previously, that it had a foreign national in the army who was not legally allowed to be a member of the British Army. The discovery of this fact through the use of data quality techniques meant that the issue could be resolved without causing a diplomatic incident.

A third compliance requirement arises with respect to data privacy and protection laws. It is important to be able to redact or mask such things as credit card and social security numbers so that unauthorised personnel cannot see this information. Applying appropriate techniques to hide this information means first discovering where it is. If a credit card number has been entered into an address field that fact needs to be identified and it is one of the roles of data quality tools to identify cases such as this. Note that this sort of requirement is not limited to regulated data but also applies to intellectual property and other details that you may wish to keep from prying eyes.

Technical

There are a variety of IT functions and processes that may not or will not work (or will not work effectively) without due care being paid to data quality. Perhaps the most well-known of these is with respect to data warehousing, wherein data quality has been a perennial problem. As long ago as 1999, DCI and the Meta Group (now part of Gartner) published the results of a survey into issues affecting the ongoing maintenance of a data warehouse and more than 45% of respondents cited "managing data quality" as their number one issue; and data quality remains the number one issue today.

However, it is not just in maintaining a data warehouse that data quality is significant, it is also vital in data migrations (whether migrating from one database to another, consolidating databases or migrating from one version of an ERP or CRM package to another), when archiving data and when implementing master data management (MDM). In Bloor Research's 2011 survey of the Data Migration market the most frequently cited reason for overrunning and failed projects was the combination of either "poor data quality" or "lack of visibility into data quality issues".

Issues around data quality

There are a number of major issues that need to be addressed when we consider data quality. The first applies to any environment where data quality has not previously been tackled such as Sallie Mae, the British Army, Emerson Power and British supermarkets. Here a significant exercise is required in identifying faulty records and remediating them. Note that 100% accuracy should not be your goal—that would be prohibitively expensive—something like 90% or 95%, depending on the data in question and what it is being used for, would be more appropriate. However, what do you do once that task has been completed? If the answer is nothing then you will shortly be back where you started. While figures vary it is estimated that data, left to its own devices, deteriorates by between 1.25% and 1.5% per month. So the corollary is that simply clearing up today's mess is not enough: you need to prevent it recurring and, when it does occur, you need to be able to detect that an error has arisen and remediate it.

The second issue is that data quality problems are effectively of three types:

- Invalid data, which cannot be processed correctly by your software. Typical examples would be a null field where application software is expecting a value or a numeric field that has alphabetic characters in it. Errors of this type can be automatically detected.

- Incorrect data that is valid but erroneous. That is, it can be processed but the results will be faulty. Examples include duplicate records, incorrect delivery addresses and so forth. While appropriate software can suggest likely duplicates on an automated basis it will do so based on using probabilistic or statistical methods that cannot be guaranteed and will require visual inspection.

- Business rule violations. There are two types of business rule in this context. The first relates specifically to data quality. For example, you might have a rule that insists on a valid postcode in an address or that a product code must have a specific format. It is also possible to have business rules that are exactly that. For example, you might have a business rule that checked that credit limits do not exceed a specific figure or that internal departments were not over-spending their budgets. We should say that while the use of data quality processes can be used for these types of business rule many authorities do not recommend this approach. Nevertheless, our experience suggests that it is commonplace, perhaps as a short-term measure, when application software does not (currently) provide the required functionality.

While data quality software may be able to detect issues with the data it should be clear that the remediation of the data requires human intervention of some sort. For non-data quality business rule violations it may be that you simply require an alert to be raised, but, in most other instances, correction will be required and this will involve not IT, but relevant business analysts or data stewards, who are familiar with the data in question. This means that whatever software is used to support data quality initiatives will also need collaborative capabilities that will allow those business analysts or stewards to work in conjunction with IT to remediate problems.

You might think this collaborative capability would not apply where data quality is important for technical reasons. Nothing could be further from the truth. Whenever you are moving data, whether to archive it, migrate it or set up an MDM hub, it is important that related data remains related during the movement process. In order to do that you need to understand those relationships. However, while it is certainly important to understand these at a technical level—tables, columns and so forth—it is also necessary to understand them at a business level: you need to understand a customer, with his orders, with his delivery addresses, with his service history, with his invoices and with anything else that is a part of the business entity that makes up a customer. This can only be done, once again, by relevant personnel with business domain expertise. Thus collaborative capability is needed here also.

The third issue is that, once you have taken on board the principle that a one-off data cleansing exercise is not enough, the obvious approach is to try to prevent poor quality data entering your system in the first place. This is simply on the basis that prevention is better (more cost effective) than cure. There are a couple of requirements for this. The first is that it will be helpful to have plug-ins that work directly with popular ERP, CRM and other applications that check the validity of data and check whether it complies with business rules at the point of entry. Secondly, we must bear in mind that a) we will not have 100% accurate data, as discussed previously, and b) that

Issues around data quality

applying data quality will typically be done on a domain by domain basis so that some data that we want to migrate or put into a data warehouse will not have been thoroughly cleansed. It will therefore be useful to be able to use data quality rules and validation as a part of the data integration processes that move this data. In other words, data quality modules should integrate with whatever ETL/ELT (extract, transform and load/load and transform) processes you are using for that purpose.

As a result of these deliberations we can conclude that to support data quality initiatives there are various functions that will be required from any comprehensive solution to data quality issues. These include the ability to discover errors, remediate those errors, monitor those errors on an ongoing basis, preventative capabilities, the ability to discover relationships and particular forms of data, integration with data movement functions and collaborative capabilities.

Finally, we should say that if you are undertaking all the suggestions made here then you are effectively implementing a data governance programme, even if you do not specifically call it that. So, a final requirement will be the ability to interact with any specific data governance capabilities that may be in use, although this could equally fall under the heading of collaborative capabilities.

Required technology

What does all of the foregoing mean in terms of specific technologies? At a minimum, a complete solution would include:

1. Business glossary. A business glossary provides a definition of business terms that users can track back to see where they derive from in IT terms. Conversely, the IT professional can track forward to see what business processes are dependent on particular data elements. Thus a business glossary is fundamental to support collaborative working between the business and IT. It is for this reason that we put it first in our list of requirements: we regard collaboration as being of paramount importance and a business glossary is essential to collaborative working.

2. Data profiling. This provides various capabilities:

   a. The ability to measure the quality of your data and provide statistics that relate to said quality.

   b. The ability to monitor data quality on an ongoing basis and present the results in a dashboard. This should include the ability to monitor trend information (is data quality improving and, if so, how fast?) as well as web-based interfaces that business managers and others can easily see the quality of the data they are using without having to install any software.

   c. The ability to discover data that is subject to data privacy or intellectual property requirements prior to masking or redaction.

3. Data discovery. This is about the discovery of relationships between and across data elements, which are often located in multiple (heterogeneous) data sources. These allow the creation of business entities (a business level view of these relationships) discussed previously. Most data profiling tools have some limited data discovery capabilities but they tend to be poor at working across data sources, do not have the ability to visualise the data as a business entity, and lack the automation of true data discovery tools. In addition to relationships per se, data discovery tools will also discover matching keys, perform precedence analysis (which source is more trustworthy), discover cross-source business and transformation rules and their violations, and so on and so forth.

4. Data cleansing. These tools provide the ability to correct errors and de-duplicate records as well as to define data quality rules and support data enrichment (that is, to augment data from outside sources such as D&B or to add geospatial data). Depending on the environment you may also require your data cleansing tool to be able to do such things as check watch lists. In some cases, where dealing with potentially shady characters is a part of the business, then you may need specialised identity resolution software that has been designed to recognise such things as aliases.

5. Workflow. This may be an independent capability or it may be embedded within the environment. It is important because if your data quality monitoring recognises an error in a record or a data quality alert is raised because of a contravention of a business rule, then you need to be able to assign remediation to the appropriate data steward or business analyst and track the progress of that remediation.

6. Integration. All of the above need to be based on a common repository so that metadata can be shared. This is important for data lineage and to support where-used and impact analyses. In addition, as previously mentioned, it will be useful to be able to embed data quality processes into ETL workflows and to integrate, at the front-end, with relevant application packages. Further, it will be useful if data discovery, in particular, is integrated with any archival and MDM applications that may be available. As an aside, one would also like to see ETL capabilities integrated with both of these.

Conclusion

It should be clear from the preceding discussions that there is much to be said in favour of a platform-based approach to data quality. While we do not denigrate the advantages of a best-of-breed approach there are certain requirements that are not available as best-of-breed products. For example, there are no stand-alone business glossaries in this space. We therefore believe that you should at least start with the premise of a platform and then, if you want to replace an independent component of the solution with a third-party product, then that is always a possibility though there may (probably will) be issues with integration; for example, the platform's business glossary may not support the third party product, thereby lessening its value. There may also be similar issues with respect to the platform's repository and such things as data lineage.

Further Information

Further information about this subject is available from http://www.BloorResearch.com/update/2124

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

About the author

Philip Howard
Research Director - Data Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director focused on Data Management.

Data management refers to the management, movement, governance and storage of data and involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration (including ETL, data migration and data federation), data quality, master data management, metadata management and log and event management. Philip also tracks spreadsheet management and complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to IT-Director.com and IT-Analysis.com and was previously editor of both "Application Development News" and "Operating System News" on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), dining out and walking Benji the dog.

Copyright & disclaimer

2nd Floor,
145–157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com