certus
PERSPECTIVES

The Importance of Data Quality

The late American politician Daniel Moynihan once famously said, "Everyone is entitled to his own opinion, but not to his own facts." While Moynihan was, of course, bemoaning the level of political discourse, his aphorism can be applied with equal relevance to the question of decision-making in business. Are we relying on facts or on our gut feeling? Is it somewhere in between, a mixture of data and intuition?

There's no right or wrong answer about the degree to which we should rely on data versus intuition. Steve Jobs vehemently rejected the market research approach to product development, invoking the famous (but perhaps apocryphal) quotation from Henry Ford, "If I'd asked my customers what they wanted, they'd have said a faster horse." But it's clear that some decisions, from the strategic to the daily and mundane, should be at least in part based on evidence.

This is one of the reasons organisations have invested heavily in business analytics, in an effort to ensure people have the right data at hand to support daily operational decisions and even long-term strategic decisions. Whether it's a government agency approving a pharmaceutical for sale, a retailer deciding on the location for a new store, or a bank giving a customer a credit score, decisions depend not only on data, but also on the quality of that data.

Poor quality data can mean poor quality decisions.

Poor quality data can arise from the most trivial of causes. One of our consultants tells of a client whose staff were using the quantity and unit cost fields of their new ERP system interchangeably, reasoning that A times B is the same as B times A, so "where's the harm?" The harm was felt by senior management, whose reports were telling them that volumes were increasing, while the average price of goods sold was falling sharply. Significant work was undertaken to analyse why this was, until the chance discovery that sales order entry staff were taking liberties.

Every day, organisations are making decisions, relying on data whose quality may not merit the importance placed on it. Not all of these decisions are made by people. Some are made by software, such as an airline that uses revenue management software to determine the price of every seat on a given flight, or a job scheduling and optimisation system that determines which of a telco's field technicians will be assigned to repair a residential telephone fault. A person can make a mistake, and no algorithm is perfect, but whether the decision is made by a person or a machine, it is instantly flawed when the data on which it is based is itself flawed.

But it's not all doom and gloom. Data quality initiatives can have significant, substantial and measurable positive impacts on a business, and drive revenue or profit. Recently, a US bank told of a data quality project which looked at the data being used to make decisions on capital assets. By implementing a targeted process to improve the quality of the information they held on these assets, they were able to more accurately assess each asset's value. This enabled the bank to hold only the capital required rather than a default position and in the process released billions of dollars that were previously held as contingency.

Data Quality and Information Governance

There are ways to ensure data can be relied upon so that decisions, big and small, are based on trusted information. This is achieved through the process of Information Governance, of which data quality is only one aspect.

The problem with data quality is two-fold. First, it's just part of the solution. Second, it's an overused phrase that, without further definition, has lost meaning. The IBM Data Governance Unified Process defines data quality as including data's validity, uniqueness, completeness, consistency, timeliness, and accuracy, and its adherence to business rules:

Validity – Data values are stored in the appropriate type, format and syntax.

Uniqueness – There are no duplicate records within a data set or across several data sets.

Completeness – A measure of the total information stored for an individual record, within a single system or across multiple systems.

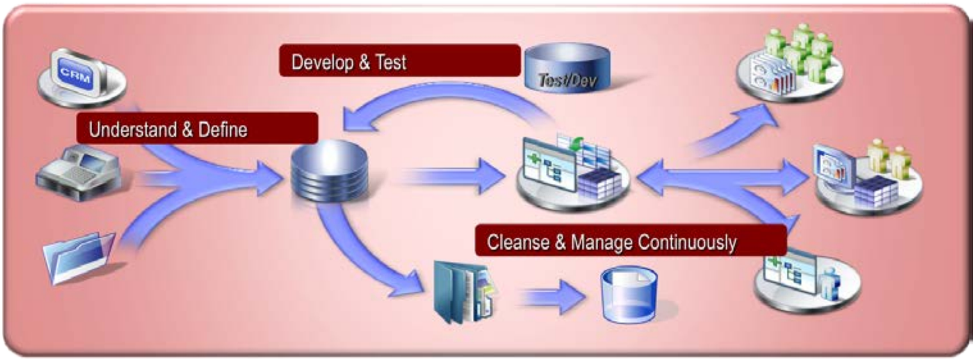
Consistency – All data adheres to a set of business and technical rules that describe the appropriate classification and use of the information.

Timeliness – All relevant data required for a decision is presented in time to be considered.

Accuracy – The data is a true representation of information recorded. For example, employee job codes are accurate to ensure that an employee does not receive the wrong type of training.

Adherence to business rules – The data attribute or combinations of data attributes adhere to specified business rules. For example, a business rule might check whether a birth year is prior to 1/1/1900 or whether the effective date of an insurance policy is prior to the birth date on the policy.

Having defined what data quality includes, the next step is to improve it with a three-phase methodology, as below.



Understand and Define – Understanding and defining data is key to delivering trusted data. Data quality issues are identified and classified and can be captured in a business glossary along with data definitions. Once specific information assets have been defined, classified and their current data quality status documented, decisions can be made. Do you fix all data quality issues regardless of their potential impact? How do you decide where to spend your data quality budget? An Information Portfolio Management approach can help here. Subjective information is gathered from stakeholders and used in combination with your objective technical information to rank information assets across dimensions such as cost, risk, compliance, etc. This provides a 360 degree view for prioritising limited data quality resources. And it can be adjusted in real time to changing business directives.

Develop and Test – A set of data quality rules is developed to validate source system data and provide metrics through a reporting tool. This can be a daunting task but with a metadata driven approach and a good data quality framework, development and testing time can be significantly reduced. The IBM InfoSphere Information Analyser can define a single data quality rule that can be used in many validation and ETL jobs yet be managed as a single asset.

Cleanse and Manage Continuously – Cleansing and managing data quality over time requires the capability to audit data quality, report on data quality metrics, and provide the required information to data stewards for remediation of issues at the source systems. It is essential that quality issues are fixed at their source, to avoid band-aid solutions and wasted time and money.

Other Information Governance Essentials

There are three other elements vital to ensuring that decisions are fact-based.

Data Stewardship – All aspects of a specific information asset require ownership, or data stewardship. This includes terms, definitions, classification, use, security, retention and, of course, quality.

Accessibility – It matters little if data is of the highest quality, if people can't access it. Ensuring people have access to a complete view of relevant and trustable data when they need it is the fundamental goal of all decision support systems.

Understanding – Assuming people have access to good quality data, a good decision is predicated on their understanding the data. Understanding reports and other analytic outputs requires a central and up-to-date set of business terms and definitions delivered to the users where they need it.

Data Lineage – The Key to Trusting your Data

Knowing a data element's value is useful. Knowing why it has that value can be even more useful. Anyone who remembers the Windows-based query tools in the late 1980s and early 1990s will remember one of the key selling points was the ability for the user to "drill down" from aggregate data to the data beneath, to see how the value was calculated. Going a few steps further is the concept of data lineage. Data lineage provides the "answer to basic questions such as 'where did this data come from?', 'where does this data go?', and 'what happened to it along the way?'"[1]

So if a business glossary enables us to understand the definition of gross margin, data lineage takes us on the journey to discover how it was calculated, and from what source systems the underlying data came. A financial report can tell us that gross margin is 12%. Data lineage allows us to drill deeper and start to understand where the data used to calculate the value of 12% came from.

Return on Investment

The data quality ROI can be measured by looking at the business impact of failing to act. It is possible to estimate the costs arising from poor data quality by looking at the purpose and value of the data, and assessing the business impact should the wrong decision or action be taken. Alternatively, you can consider the potential for additional revenue or cost savings through greater understanding and efficiencies in the business, as in the example of the US bank.

The consequences of poor data can go much further than a lost opportunity to sell something. Poor data around equipment, work environments, products or a range of other areas can have significant impacts on the health and safety of your employees, on your customers, and on your bottom line.

The question is how many of your business decisions are currently being undermined by faulty data?

To find out more please contact Julien Redmond, General Manager - Information Management on +61 401 716 427.

[1] IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices, Sunil Soares, MC Press, 2010