Data Quality
A pragmatic approach

Data quality management is one of the greatest challenges of information technology. According to the experts, the **cost of poor data quality** can reach as high as **15 to 25%** of the operating profit (source: TDWI)

Data quality management can help companies in various aspects. A telecommunication company sending one million letters in a direct marketing campaign can significantly improve its customer satisfaction; next to that it can reduce its operational costs and even its environmental footprint by making sure that its addresses database is reaching a high level of quality. Every duplicated or bad address will lead to wasted paper, useless expense and maybe the lost of some customers who won't have received the correct information about the last offering made to them.

In some other even more dramatic cases, the company stability might be jeopardized due to data quality issues. There have been numerous examples in recent history where it is not rare to have company to be forced to republish there financial figures due to incorrect or missing information. Even if the impact on the figures is positive, which is not often the case, it always has a serious impact on the stock value and on the trust of the investors in the company. It can also lead to the departure of some key persons of the company.

On a Business Intelligence and Performance Management environment, poor data quality will have some visible and some less visible consequences.
The **visible** consequences will be that the reporting is not produced on time or does not produce the single version of the truth expected by the users; different reports will produce different figures. This will lead to a lack of trust by the users who will under use the BI platform and start creating their own version of the truth.

The **non-visible** consequences are even more critical for the company, insufficient data quality can lead to wrong decisions making, inaccurate planning or forecasting which later on while have impact on the revenue or profit and finally wrong data quality will always generate extra cost and rework when the data are used.

It is therefore obvious that every company needs to take an initiative to treat the data and the information that they represent as a key asset and to be able to measure, protect and monitor the quality of this information. But in order to be efficient that initiative has to be pragmatic and to be able to offer, depending on the needs, a time and cost effective solution using IBM solutions which will guarantee a continuous quality of the information. That is what we will describe in that paper.

How to approach Data Quality

Data Quality is a broad topic but it is often seen only as a technical matter which should be solved by applying technical solutions. This is partially true but not entirely. Within a company all systems must be setup with quality in mind but it is impossible to have a system which can have data with no data quality issues. By nature every system will have data quality issues due to incorrect data entry, data decay or simply because of data movements in it. A 100% mistake proof system does not exist and if it would the cost to implement and use such system would be incredibly high.

Most of the data quality issues are due to incorrect or incomplete data entry and processes, business or technical, which are incomplete, incorrect or not followed correctly and that at many different levels and departments within the organization. A key data element for reporting purposes might be completely useless for the person or the system which is supposed to generate it.

This is why Data Quality is an organizational challenge which requires a global approach, a Data Quality Program. This program must be setup with only one goal: to reach and maintain high levels of data quality within critical data stores.

The missions of that program will be:

- Improve Data Quality: to bring the existing data to the correct and expected level of quality for its usage.
- Prevent Data Quality problems: to help the developers and users of new systems to integrate data quality in their systems to get them to a higher level of quality.
- Monitor Data Quality: Trends and evolution of the Data Quality must be monitored to analyze if the remedies are improving the quality or to identify possible drop in quality.

The Data Quality program will then interact with all levels of the organization to improve data quality and to create the needed awareness. Just as is best practice in Business Intelligence, Data Quality is best implemented in an incremental approach, delivering subject area after subject area, rather than a Big Bang project.

How to measure Data Quality

Data Quality is a difficult topic to define. What is quality data and what isn't? From a pragmatic point of view, one could claim the following definition:

"Data has quality if it satisfies the requirements of its intended use".

There is a strong like between quality and usage. The same data set can be considered of a very high quality for a certain usage but of a very low quality for another. For example in a sales system, if all the transactions of the last two years are available in the system but not before the 10th of the next month for the previous month, the system will then be considered of high quality for the analyst when they have to do historical and trend reporting but of low quality to support the month closing or the bonuses calculation.

It is then important to be able to define and measure the data quality. Typical mistake is to base those measures on users' perception and not on facts but like any other KPI, data quality's metrics must be based on generated facts.

The key data quality dimensions to measure are:

- **Accuracy**: This is the main dimension. If data are of low accuracy the other dimensions are not relevant. Accuracy can be measured by looking at the number of missing, incorrect or duplicated records for example.
- **Timeliness**: This dimension measures if the data are available at the moment they are needed. This can be measured by checking for example if all the transactions of a certain period are available at a certain point in time.
- **Relevance**: This dimension measures the relevance of the information. The information available might be correct and current but not relevant for a certain usage.
- **Completeness**: This dimension measures how complete the data are. Usually this is a percentage of completeness which will permit to evaluate if the data are complete enough for the intended use.
- **Understood**: Good information of good quality is useless if not understood by its users. That's why it is important to know if the data are understood. This is more difficult to measure but can be done by measuring for example the number of training organized or the number of calls logged at the service desk for possible problems which are in fact due to a lack of understanding.
- **Trusted**: Maybe the most difficult dimension to measure. The best way to know if the data are trusted is to measure its usage. Typically if data are not trusted or start to be not trusted the system generating those data will see a drop in its usage, on the other end, trusted data will be intensively used.

Once defined within the organization, those dimensions will generate metrics which can be used to:

- Measure the problem: What is exactly the level of quality of data?
- Measure the trends: How is the level of quality evolving over time?
- Measure the effect of the remedy: When a remedy is implemented, to measure its impact and if it had the intended effect.

Finally, it is important to keep in mind that those metrics will evolve over time due to changes in the usage of the data. It is then very important to always compare metrics which have been calculated using the same definition in order not to give false impression of a sudden data quality drop or improvement. It is also important to start by defining metrics and targets on a limited list of key data elements and to focus on those. There is no point to look at other data elements as long as those ones are not correct.

How to improve Data Quality

The first mission of a data quality program is to get the data correct to improve the quality. This is true but shouldn't be seen as the end target. To improve the data quality means also to improve the processes or systems which are generating those data in order to get them right at the first place so no corrective actions are needed anymore.

So the source of the issue must be identified by doing an in-depth root cause analysis. This analysis should be done both at technical and business level and should lead to:

- System improvement: If the root cause is a system bug or a fault this must be addressed and corrected if possible.

- Process improvement: If the root cause is a process issue it must be addressed to the process owner to
- Users training: If the root cause is an incorrect usage of the system then a users training campaign should be considered and organized.

To significantly improve data quality will require a combination of those actions. But it is also very important to look at how realistic/pragmatic those solutions are. If the data need to be right next week but the implementation of the remedy will take months, or if the cost to implement the remedy is very high combined actions must be considered to get the data right when they are needed, but also to work on a more effective and future proof solution. And sometimes the corrective action will be the only option so they should also be foreseen and structured to keep track of those corrections.

How to maintain Data Quality

Once the data reaches the expected level of quality this is not the end of the story but the beginning of the last mission of the data quality program. This is to monitor and maintain the data quality level.

The reason is that data and the data needs evolve over time. The good data of today might turn into bad data tomorrow. This can be due to a new version of a system which is not acting the way it used to act, or which is suffering from new bugs or bad functionality. But the reason can also be that the data quality requirements evolved and some data which were good enough yesterday are maybe not of a sufficient quality to fill the business needs of today. Finally the natural data decay also affects the data quality level, a customer database which has not been updated for quite some time can be considered of a lower quality now than it used be when it has been loaded.

Data quality is also a growing process. It should always start with a limited set of data elements to improve but once the expected level of quality is reached it must be extended to other data elements to improve the general level of quality of the systems.

This is why it is important to have at all time a complete overview of the data quality level and of the data quality trends. That can be done by using sets of standard reports which will have been defined together with the metrics. Those reports are very useful to show what the size and reach of the problem; in a business case for a new initiative it is always more efficient to say that the data quality dropped by 20% in the last 6 months and to be able to prove it than to base it on feelings and perceptions. A constant monitoring of the data quality will also allow problem detections before the business really feels the pain of it.

A pragmatic approach

But if data quality is so important why is it a pain for so many organizations and why so many data quality initiatives are failing. The main reasons for that are that quite often those initiatives do not look at the right problem, are more corrective than preventive or that they want to tackle too many issues at the same time.

To guarantee the successes of such initiative in a time and cost effective way it is important to take a very pragmatic approach and to look at:

- What are the key objectives? The objective of the program must be clearly defined and measurable to be able to assess the produced results.

- What are the key data elements? Focus first on the elements which are mandatory and which make the quality of the other ones irrelevant if their own quality is not good.
- What is the current timeline? By when are the data expected to be right?
- What are the means available? What can be realistically achieved within the current infrastructure and budget?
- Which tools are available? Quite often a lot of tools are already available in house to help to support the data quality initiatives. There is not always a need to by extra tools to start with data quality.

That last point is very important. Even if a data quality program should not be considered as a technical program the tools used and how they are used to execute it will have a significant impact. Within the element61 IBM Competence Center we have extensive experience in using and deploying IBM software as building blocks for automating a data quality solution.

- Database: IBM DB2 to implement the Kimball Data Quality Audit principles in order to log and store the data quality facts.
- Profiling: IBM Information Analyzer for profiling and auditing the data in a systematic and repeatable way.
- Processing and cleansing: IBM Datastage and QualityStage to process and cleanse the data according to a flexible set of business rules.
- Reporting and Scorecarding: IBM Cognos BI to create and automate reporting on the top of the data quality framework that contains the actual data quality processing data.

The main objective when using all those tools should always be to automate as much possible the systematic tasks so the data analyst or steward can focus on analyzing, improving and expanding the data quality level. element61 has proven experiences and best practices for each of those tools to use them in a data quality program context.

Conclusion

Data quality is a very complex matters that companies find difficult to address.. Based on a pragmatic and realistic approach a data quality program can bring significant benefits to an organization in a very short amount of time.

The key success factor of such program is that data quality should be seen as a continuous improvement effort and that data should be considered as a key asset for the organization comparable to the financial or sales figures and therefore should be monitored as closely as those.

A data quality program should be ambitious but not too ambitious because it will always imply quite some changes within the organization. It should always start small to get bigger and to achieve its objectives.

Finally, a data quality program cannot be a man only project. It requires tools to automate each of its tasks so time can be spent on looking for the root cause of the problems and not on tracking the possible problems.

Contact information
Toon Puissant, toon.puisant@element61.be, +32(0)475474349
Stijn Vermeulen, stijn.vermeulen@element61.be, +32(0)477788033
Christophe Cabrera, christophe.cabrera@element61.be, +32(0)479980770