



Data Quality Tools: Horses for Courses

Tom Breur
August 2009

Introduction

Data quality programs have both technical and organizational aspects to them. Organizational facets like accountabilities and business alignment are crucial to make producing quality the default. The technical components in a data quality program deal with deduplication, data profiling, fuzzy matching of records, etc. A whole raft of skills are needed to support the “backroom process” of data creation. Data can be manipulated for a data warehouse (DWH), operational data store (ODS), or any other process that relies on integrated data.

Given the immature state of the market for data quality tools, and the lack of business results in an overcrowded market, vendor consolidation was a “natural” outcome. So in the recent past we’ve seen vendors come and go. Unfortunately, this has made offerings less transparent. One wonders whether vendors consciously pursue this, or whether lack of strategic vision has kept them from displaying their value proposition more clearly. In some cases following an acquisition, vendors may need some time to work out how to rationalize product overlaps.

Mapping out the product space

When looking at the data quality product space, it helps to apply some structure. First of all, although there are often (if not always) data quality issues involved, master data management (MDM) is something rather different from “ordinary” data quality tools. However, because MDM has come in vogue recently, many providers were quick to point to their offering there. Master data management is about installing processes and metadata to ensure that a system of record (SOR) be used and fed consistently throughout the enterprise. Data quality naturally plays a role in determining the business rules that guide these processes. But that’s about where the overlap ends. IBM

(acquired Ascential in 2005) is probably the only data quality player with a significant complementary MDM offering.

The majority of data quality applications are in the realm of cleaning or standardizing customer data. Some tools are specifically aimed at customer data, others are called "domain agnostic." Within the customer data field, a special place is reserved for address data. This area is even more complex for companies that operate on an international scale: although each country has its own set of rules for mapping address data, these rules vary considerably across countries.

From a technology perspective, it makes sense to migrate towards "domain agnostic" tools. This will make the architecture more versatile, flexible, and hence resilient to change. Domain agnostic technology also has the advantage that it can be scaled up to an enterprise class solution. The core data quality technology and most of the algorithms really are quite similar.

Some companies specialize in particular domains. Players like AddressDoctor, DQ Global, Melissa Data, Omikron or QAS, for instance, specialize in address data, sometimes for particular geographic regions. Others like Ciant, Stalworth or VedaAdvantage focus on customer data, etc. Because domain specific intricacies lead to elaborate and rather complex business rules that are exceedingly difficult to learn from the data alone (even if you have huge volumes), this focus makes sense.

A second reason for domain specific tools is that some vendors (e.g.: Human Inference, Innovative Systems, etc.) maintain reference data sets. These can significantly help to improve accuracy of matching, cleaning or deduplication. By bundling forces across their base, vendors can help their clients achieve better results than each individually could ever attain.

There's an important distinction between profiling and data cleansing functionality. The former "merely" assesses existing data, for the purpose of gathering meta data: you get an inventory of existing values, ranges, outliers, etc. This is the only area where most vendors (still) provide stand-alone tools that are available without acquiring the entire suite or platform. Data cleansing tools come in many varieties, depending on their proposed usage.

Since clients are recognizing the importance of continuity of their efforts, and the value of monitoring data quality, a relatively new

domain of quality scorecarding has emerged. This is sometimes offered stand-alone, but makes more sense when integrated in a suite or platform. The ability to drill down and visualize results adds significant value.

The importance of tooling

Data quality functionality can be embedded in many ways. Does one really need tools for this? Certainly in early stages, it might appear just as quick and easy to 'simply' program exceptions or necessary transformations and/or mappings into ETL logic. After known errors have been identified, the data quality work appears to be "done." There are some significant problems with this reasoning, though.

Tools help standardize backroom processing, and keep the ETL architecture flexible. The last thing you want is to create "new" legacy by allowing a monster of an ETL process to mushroom in your backroom. And the same reasoning applies to any area (besides ETL) where data quality tools add value.

As with any architectural consideration, you need to balance short-term and long-term considerations. Introducing professional tools later on invariably comes at the cost of scrap and rework of existing programs. The "pain" of inflexible architectures, and loss of agility is hard to quantify but very real nonetheless.

In short, we believe that tools are usually worth their investment; what is *difficult* is to make the business case. This holds for architecture in general, and certainly for losses as a result of poor data quality.

To SOA or not to SOA?

Several vendors have made the push towards web-based, service oriented architecture (SOA) versioning of their tools. Gartner appears to really like this, and seems to consider SOA the next best thing since sliced bread. We find the distinction between on-premise deployment versus software as a service (SaaS) not so interesting, and even trivial from a business perspective.

What appears to be getting much less attention from industry analysts is that SOA implies a radical change in *business model* for providers. There is a technical question of where the software resides. How the contracts are set up, and how that evolves as the business changes

(grows) is much more interesting. This is also a more relevant consideration for the client. If all this does is lower the barrier to entry, that is fine. But longer term consequences should be considered at the outset.

We have seen *some* successes, and several (huge) failures. The fact that a move to SOA involves (some) technology, but also people and processes is easily neglected, and not something sales people draw attention to. However, this oversight makes failure a near certainty. Quite apart from the fundamental question whether the business process at hand can be broken down in logical units at the right grain. We have found opportunities to “componentize” business processes a critical success factor. If that doesn’t come natural, there is little to gain from a SOA from a technological perspective.

Conclusion

The market for data quality products is not very transparent, at the moment. Due to consolidations and strategic changes, notably a move from domain specific to general purpose tools, we’ve seen many changes in the last few years. Most (large) vendors have bundled their offerings in suites or platforms. Data profiling tools remain the only niche where “stand alone” products are still widely available.

Add to that a surge in MDM and data integration, just when you thought you survived the CRM hype. It is clear that the market is not very “clear” at this moment. Several gurus have argued that there should be a “data quality profession” to counter balance the ubiquitous lack of background and insight at board levels into data quality related issues. This is particularly surprising when you consider the huge costs, and dramatic project overruns caused by poor quality data.

The recent trend of “packaging” functionality in suites or platforms is an important improvement. Now you can have profiling, parsing, matching, deduplication, etc., and also visualization and scorecards (data quality monitoring, dashboards) of a single platform. Data stewardship has never been more fun (and productive).

On-demand data quality services (SaaS) allow smallish customers to take advantage of these tools, at relatively low start up costs. Most data quality players are jumping on this band wagon, offering more options to their clients.