The Data Quality Landscape - Q1 2023

Data quality has been an issue ever since data started to be captured on computers. Despite huge investments in technology across industries, the level of trust in the quality of data remains consistently and depressingly low. A Deloitte survey showed that 67% of executives are not comfortable in using data from their own corporate systems, while a May 2022 survey of 500 companies by a market research firm Pollfish found 77% of respondents admitting to problems with their data quality. These are consistent with earlier surveys e.g. a 2021 survey by Precisely of over 300 executives finding that 82% of C level executives found data quality was a barrier to successful data integration projects. Such issues go across industries, often with serious consequences: a US government study found that up to 10% of patients in US hospitals were mis-identified, with duplicate patient records running at 12%. Prescription errors in the US healthcare system are reckoned to cost $21 billion and cause 7,000 deaths annually, according to the Network for Excellence in Health Innovation.

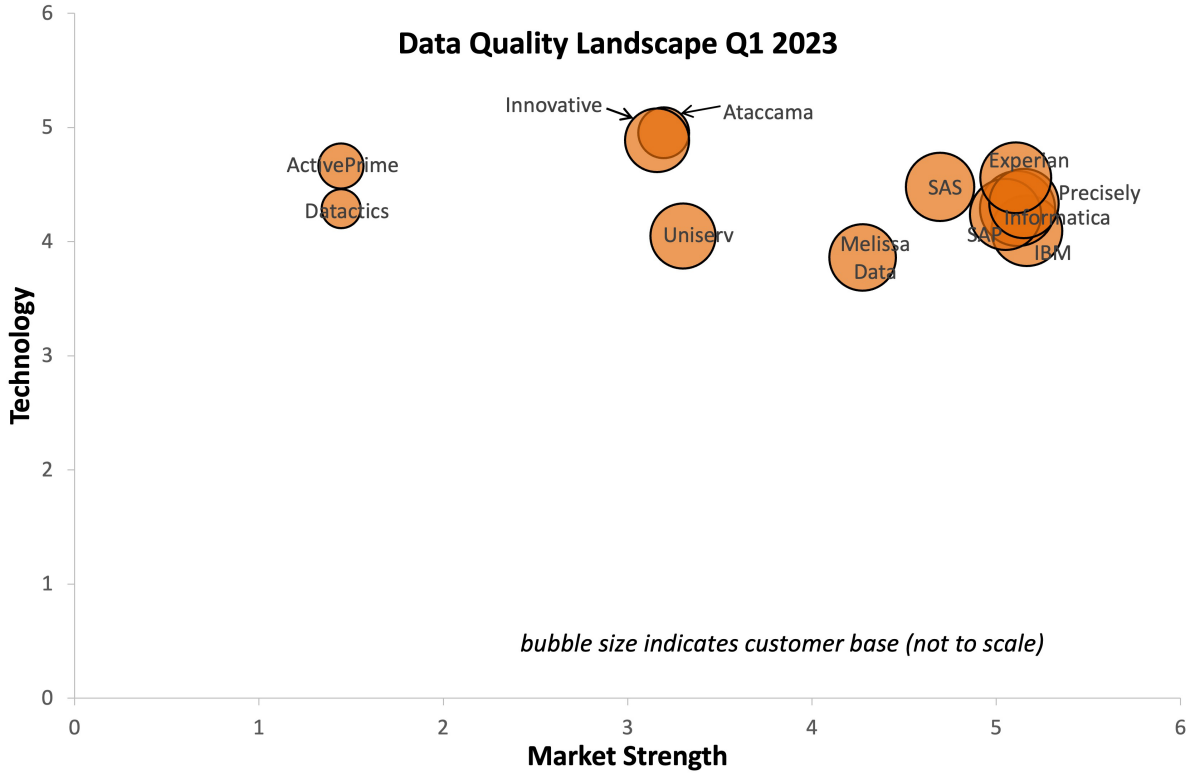
There are many reasons for this state of affairs, with human nature playing a major part: if an employee is asked to type in data to a computer system that they see no direct use for, they will inevitably be less careful about its accuracy than with something that impacts them directly. An employee will pay close attention to their payroll slip and check that their expenses have been paid on time, but filling out some general background information on a customer that only benefits an unknown person in another department is liable to involve less diligence. These days data quality is often considered an important part of broader data governance initiatives, with business people taking ownership of their data rather than just delegating this to IT departments, who often lack the knowledge (or the authority) to do this job effectively.

Data quality tools emerged to try and improve things, often trying to improve data capture at source, as well as in scanning large amounts of data for likely errors. Data quality software initially focused on customer name and address data that is common to virtually every industry, with clever algorithms that are designed to spot common misspellings and errors. A modern data quality suite can scan ("profile") data to spot likely errors based on statistics and examine data records to identify possible duplicates. Despite all the best efforts to ensure that customer or product records are unique, hard reality shows that duplication rates of 10% to 30% are common. One customer master system that this author examined some years ago had 80% duplicates. Good data quality software can help diagnose this issue, highlight likely errors and duplicates, and help combine duplicate records into a high-quality system of record. They can also suggest business rules that can be applied to help keep data quality high, and can monitor systems to check on progress over time. These days data quality software can be applied to different data domains such as product or material data, not just customer name and address records. Extensive 3rd party databases can be used to enrich name and address records. For example, an insurance company can check whether a house is built on a flood plain, or is in a high crime area, and adjust a quotation accordingly.

In recent times many vendors have adopted machine learning techniques to help with this process. Systems can observe human domain experts resolving possible duplicate records, and can then suggest more refined business rules and carry out more automation of common errors, freeing up human time for more useful activities. The use of machine learning in merging and matching records is now becoming common, with software getting much smarter at this than it was a few years ago.

Data quality is a persistent issue, and is not going to be magically resolved by a software fix. However, there is no doubt that the industry is evolving, and the use of machine learning in particular shows promise in spotting and resolving data quality problems with less need for costly human intervention.

The diagram that follows shows the major data quality vendors, displayed in three dimensions. See later for definitions of these.



It is important to understand that this is a high-level representation of the market, with vendors represented on the chart specialising in different areas and at very different price-points. If you are considering data quality software, it is important to tailor your selection process to the particular needs that you have rather than relying on high-level diagrams such as this. The Information Difference has various detailed models that can assist you in vendor selection and evaluation.

As part of the landscape process, each vendor was asked to provide at least ten reference customers (some vendors provided many times that number), which were surveyed to determine their satisfaction with the data quality software of the vendor. The happiest customers based on this survey were those of ActivePrime followed by Experian, then those of Ataccama and Innovative Systems. Congratulations to those vendors.

Below is a list of the main data quality vendors.

Vendor	Brief Description	Website
Address Doctor	Vendor that specialises in providing wide coverage of name and address	www.informatica.com/addressdoctor.html - fbid=-gz2yeRJkyH

	information; now owned by Informatica.	
Ataccama	Vendor with a modern data quality suite.	www.ataccama.com
ActivePrime	US-based vendor of data quality solutions for CRM systems.	www.activeprime.com
Capscan	London-based provider of address management and data integrity services, now owned by GB Group.	www.gbgplc.com/uk
Data Mentors	Long-established US data quality vendor	www.datamentors.com
Datactics	UK-based vendor of data quality and matching software to banking, finance, government, healthcare and industry.	www.datactics.com
Datiris	Colorado vendor of data profiling technology.	www.datiris.com
Datras	Munich-based vendor with wide ranging data quality functionality.	www.datras.de
DQ Global	UK data quality and address verification software.	www.dqglobal.com
Experian	UK-based vendor specialising in data quality, including name and address validation, data profiling and data enrichment.	www.edq.com/
Google	The search engine giant does data quality.	github.com/OpenRefine
360 Science/helpIT	US/UK vendor of integrated contact data quality solutions including matching and address validation.	www.helpit.com
Human Inference	Dutch data quality vendor.	www.humaninference.com
IBM	Data quality software from the industry giant.	www.ibm.com
Informatica	California-based data management vendor, a major player in data quality.	www.informatica.com
Infogix	Illinois-based vendor specialising in controls and compliance.	www.infogix.com

Infoglide	US vendor specialising in identity resolution.	www.infoglide.com
Infoshare	UK data quality specialising in the public sector market.	infoshare-is.com
Inquera	Israeli company with an approach to product data quality using machine-learning technology based on subject domain experts' knowledge.	www.inquera.com
Innovative Systems	Long established data management vendor with extensive offerings including data profiling, data quality, address validation/geocoding, and risk management solutions.	www.innovativesystems.com
Intelligent Search	Identity management company now with a more general data quality capability.	www.intelligentsearch.com
Irion	Italian data quality vendor specialising in financial services.	www.irion.it/index.php/en
Melissa Data	US/German global data quality vendor offering address verification, geocoding and matching solutions.	www.melissadata.com
Microsoft	DQS is the data quality offering of the Redmond software behemoth.	www.microsoft.com
MIOsoft	US data quality vendor.	miosoft.com
Netrics	New Jersey vendor of matching software. Now owned by Tibco.	www.tibco.com/products/automation/application-integration/pattern-matching
Oracle	The software giant's data quality offerings are based on the acquisitions of Datanomic and SilverCreek.	www.oracle.com
Precisely	Precisely is a rebranding of Syncsort, which bought Trillium, and which itself acquired Pitney Bowes data quality software.	www.precisely.com/product/data-integrity/precisely-data-integrity-suite/data-quality

Postcode Anywhere	UK vendor of web-based addressing software.	www.postcodeanywhere.co.uk
Redpoint	Data Integration software with a data quality component	www.redpointglobal.com
SAP	The software giant is a major data quality player.	www.sap.com
SAS	One of the leading players in data quality, now integrated within their broader data management suite.	www.sas.com/en_us/software/data-management/data-quality.html
Satori Software	Seattle-based provider of address management solutions.	www.satorisoftware.com
Talend	Open source vendor with wide range of quality functions that are tied to data integration and MDM.	www.talend.com
TAMR	Vendor that applies machine learning to the data quality problem.	www.tamr.com
Uniserv	Large German data quality vendor.	www.uniserv.com

Other vendors of data quality software include:

Ciant	www.ciant.com
Data Lever	www.redpoint.net
Data Mentors	www.datamentors.com
Infosolve	www.infosolvetech.com
Intervera	www.intervera.com
Ixsight	www.ixsight.com
MSI	www.msi.com.au
Rever	www.rever.eu
TIQ Solutions	www.tiq-solutions.com
Winpure	www.winpure.com
Wizsoft	www.wizsoft.com

Research Methodology

The Information Difference Landscape diagram shows three dimensions of a vendor:

- Market strength
- Technology
- Customer base.

“Market strength” is made up of a weighted set of five factors: revenues, growth, financial strength, geographic scope and partner network. Each of these individual elements is scored, the total producing the “market strength” figure. Similarly “technology” is made up of four factors: “technology breadth” (the coverage of the vendors in various data quality areas as illustrated below), the longevity of the software in the market, analyst perception of the product via briefings, and customer feedback from reference customers (this has a high weighting), which we surveyed. In each case the scoring is on a scale of 0 (worst) to 6 (best).

Vendors were asked to submit answers to various questions via a questionnaire. Vendors were interviewed directly by an analyst and their software demonstrated and assessed. Reference customers were surveyed to give their experience of the software of each vendor. The technology functions which the vendors were asked about are as shown below. These are drawn from the Information Difference vendor functionality model; if you are interested in more detail on this then please contact The Information Difference.

Functional Areas