The Data Quality Landscape - Q1 2022

Given the degree to which data drives much of our modern lives, it is a little troubling that quite a lot of it is suspect in quality. A 2019 Deloitte CEO survey found only a third of executives trusted their own organization's data, with 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. Data can be out of date, incomplete, unreliable, duplicated, unavailable, insufficiently granular or just plain wrong, with consequences in all manner of industries and situations. The issues caused can be relatively trivial, such as duplicated addresses in marketing lists, or very serious such as the unit of measure error that caused the loss of a NASA Mars space probe in 1999. One 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.

A 2017 Harvard Business Review case study that tested data quality in 75 companies found that just 3% of executives had data that met their own "acceptable" target levels of accuracy, with 47% of newly created data records having at least one work-impacting error. Various studies (from IBM, Experian and more) have estimated that companies spend 10-15% or more of their total revenue addressing data quality issues. There are many reasons for this sorry state of affairs, from duplication of systems, carelessness in data entry, lack of supervision and monitoring through to quirks of human nature, but the software industry has been offering data quality software to improve this situation for several decades.

Modern data quality tools usually offer a wide range of functionality, from basic data profiling through to sophisticated matching algorithms, de-duplication, workflow capabilities, business glossaries, data enrichment and monitoring. There are some stand-alone data quality tools, while similar capabilities may these days be embedded in broader solution suites, alongside master data management and data integration capabilities. Data quality was often delegated to the IT department, but these days most organizations have realised that it is a business responsibility, with data governance committees and data stewards embedded within business lines being responsible for the quality of data. Many offerings now use artificial intelligence techniques, such as the use of machine learning to detect anomalies or to improve data matching by being trained on data from human experts.
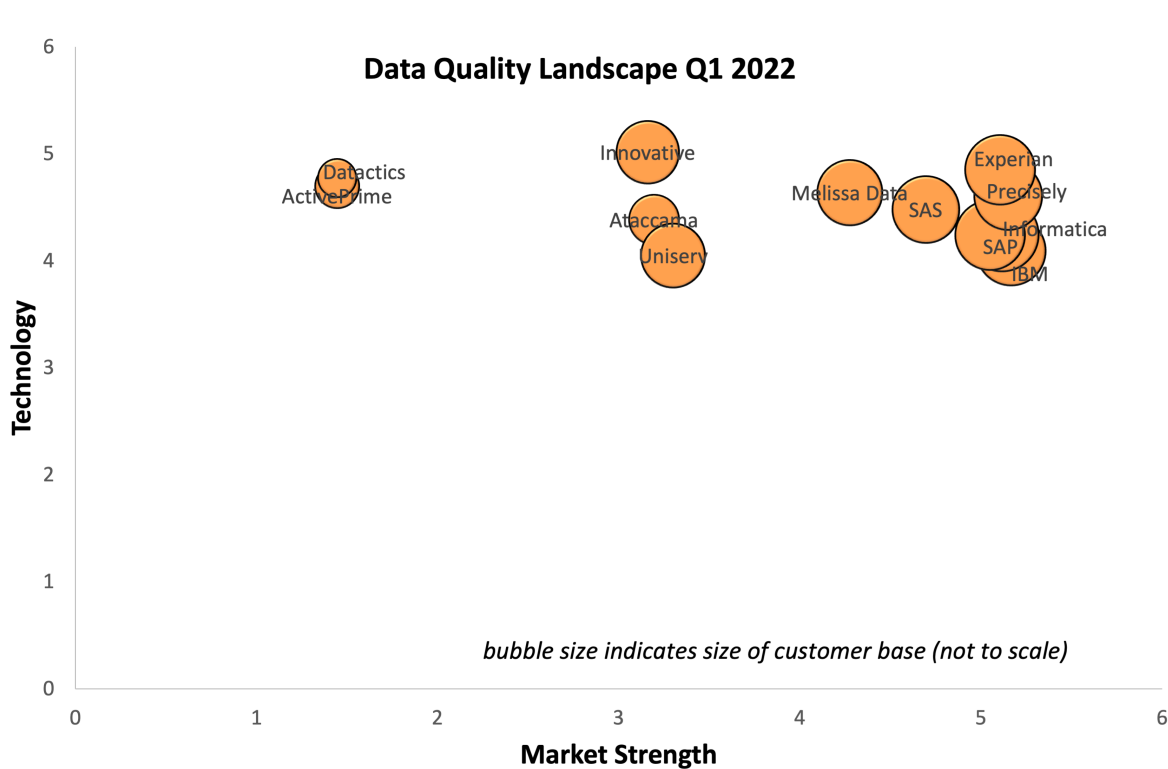
Much of the industry has focused on customer name and address records, since these are common to almost every organization, but data quality tools can be used for a range of other data, such as that for products, assets, location etc. These days a business name and address can not only be checked for accuracy, but tools can enhance that record with all manner of additional data, such as identifying the company at an address, its revenues, numbers of employees etc. A residential address can be enriched to see which voting district it is within, and if the address falls within a flood plain, which is handy if you are an insurer. Some data quality tools specialise in particular industry verticals, such as compliance for financial regulation, where stiff penalties can (and have been) levied on companies that have demonstrably poor data quality. In one case this resulted in a financial institution processing transactions from a body that was on a sanctions list, the error occurring due to a simple data quality error that placed the correct business address in the wrong data field.

Data quality tools are getting smarter at helping to spot duplicate records in corporate systems, and the use of machine learning is helping to increase the degree of automation that is practical in

resolving candidate records that may, or may not, be duplicated or in error. The more such cases that can be handled by software, the less need for human expert intervention, and the lower the cost of fixing underlying issues.

Data quality is probably always going to be an issue, but the software offerings available now can go a long way to improving things and to reducing the surprisingly high costs of poor data quality that persist within organizations to this day.

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

Data Quality Landscape Q1 2022

bubble size indicates size of customer base (not to scale)

Technology (vertical axis) vs Market Strength (horizontal axis). Vendors plotted: Datactics, ActivePrime, Innovative, Ataccama, Unisery, Melissa Data, SAS, Experian, Precisely, Informatica, SAP, IBM.

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 Experian followed by ActivePrime, then those of Innovative Systems followed by Datactics. 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 information; now owned by Informatica.	www.informatica.com/addressdoctor.html - fbid=-gz2yeRJkyH

Ataccama	Canada-based company 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 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 based on crowdsourced AI, including data profiling, data quality, address validation/geocoding, 360° view, 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	International vendor offering data quality solutions including verification, matching and enrichment for address, geocoding, identity, corporate, health and other data domains.	www.melissadata.com
Microsoft	DQS is the data quality offering of the Redmond software behemoth.	www.microsoft.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
Postcode Anywhere	UK vendor of web-based addressing software.	www.postcodeanywhere.co.uk
Precisely	This major player in the market owns what were the data quality offerings of both Pitney Bowes and Syncsort (Trillium).	www.precisely.com/
SAP	The software giant is a major data quality player.	www.sap.com
SAS	One of the leading players in data quality.	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

Data Quality Functionality Areas

Data profiling	Cleansing rule configuration
data type & format validation	Merging & survivorship
Data dependency validation	Data enrichment
Matching algorithm tunability	Language standardisation
Fuzzy matching	External source integration
Matching accuracy	Data type support
Match rule configuration	Data quality monitoring
Parsing	Industry-specific templates
Context-sensitive data cleansing	