AI-Enabled Data Quality: Improve Data Quality Across Your Enterprise.

Pradyumna S. Upadrashta, Chief Science Officer, Mastech InfoTrellis

Data Quality: A Rising Problem

Data quality is a huge problem, with a minimal estimated impact of $3 trillion per year to the economy. Very few companies meet basic data quality standards, or even know what those standards are. Poor data quality can cost an organization upwards of $15 million per year. Small data quality issues can have disproportionally large impact on the efficiency of your organization, think 5x. This means you not only have to work harder for your dollars, but that you keep fewer dollars after working so hard. The underlying scalability problem, induced by data supply-demand forces, along with a general vacuum of talent, will force organizations to adopt more automated approaches to assessing and managing data quality.

Contents

Gartner

"Poor data quality costs organizations an average of $15 million per year."

IBM

"Bad data costs the U.S. economy $3.1 trillion per year."

Harvard Business Review

"Only 3% of Companies' Data Meets Basic Quality Standards."

KPMG

"Benchmarks show that 90% data quality leads to 53% operational efficiency."

IDC

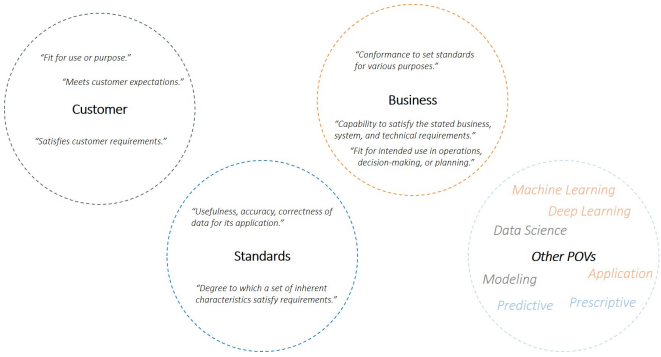
"A data quality initiative can raise an organization's sales bar as high as 20-40% and can cut down IT cost by 40-50%."

Gartner

"By 2022, 60% of organizations will leverage machine-learning-enabled data quality technology to reduce manual tasks for data quality improvement."

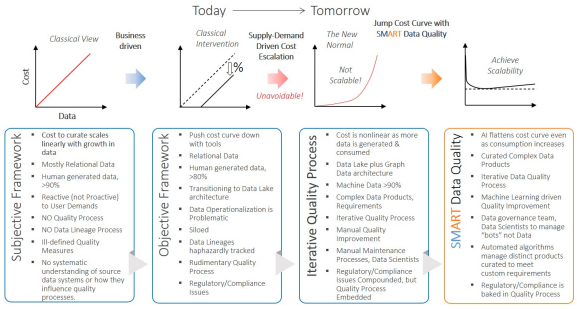
'What is your Data Quality', depends on 'Who is asking?'

One of the most basic problems when we talk about "Data Quality" is that we are not being clear about who is asking the question. We can approach data quality from a Customer, Business, or Regulatory/Compliance/Standards based perspective, and all of these will have different definitions. Also, this does not include various "futuristic" application-centric perspectives, for instance, the impact of machine learning and AI on data consumption.



The Macro View: Curation costs will grow out of control

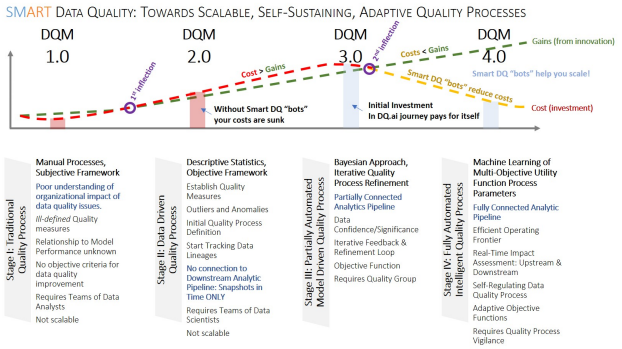
The data quality problem may be viewed through a maturity lens in 4 stages. Most of our clients are somewhere between stage 1 and 2, as they begin to understand the impact of data quality on their organizations. The four stages are characterized by the degree to which AI has been implemented directly/indirectly to solve the scalability problem. The fundamental reality is changing on both ends: Supply and Demand. On the supply side, we see a rising number and heterogeneity of sources, from traditional relational data stores (SAP, CRM, ERP, etc.), to data lake architectures, to peripheral IOT/IIOT sensor-based machine data. I suspect that the latter will overtake the former by orders of magnitude in the coming decades. On the demand side, we see an ever-increasing requirement for high-quality unbiased data as inputs to complex deep learning applications, and these in turn feed into ever more sophisticated software that serve complex ends. These trends will only continue. The future of data quality is one where a handful of people (data scientists and/or machine learning engineers) will manage a portfolio of data curation "bots" which will in-turn precisely control the data quality process (from source to application) for thousands (maybe millions) of enterprise applications, on a flattened-cost curve basis. There is simply no other alternative, making this journey inevitable. Organizations which "eat the frog" early will be more prepared for an AI-centric future, and will begin to see the benefits of AI more rapidly. Those who don't, will start to see their margins erode by 1000 cuts, or worse, become outdated and uncompetitive as they wither away into obscurity. The problem is very real and immediate: One of our clients, a large multinational bank, spends nearly $19 million a year on data quality issues, which is what inspired our solution. As pioneers in this space, we developed some of these ideas as early as 2016 in response to client challenges. The market is now beginning to catch on to the need for ML-driven data quality approaches, as highlighted in the latest Gartner report. Our data quality automation offering is more relevant than ever today.



Bold Investment: Bold Outcomes

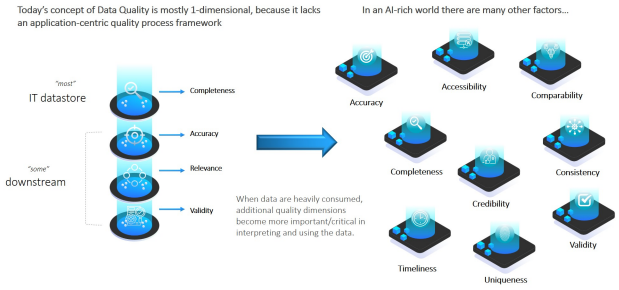
By investing in AI solutions in response to the rising data quality problem, we can bring transparency to the quality process, understand its impact on our organization, and effectively manage the inevitable challenge of exploding data sources and consumption. In DQM 2.0 we begin to implement a statistically rooted data quality framework, which allows us to quantify the problem. This sets us up to implement DQM 3.0 where we are able to develop an iterative statistical framework for tactically resolving data quality process issues, with a ROI based approach. In DQM 4.0, we are implementing machine learning based "AI bots" to pro-actively manage thousands of enterprise applications and processes, and the payoff continues to accelerate as our investment is gradually tapered off per the flattening cost-curve. If organizations DON'T invest, they risk seriously eroding their profits through quality leaks, which cause a host of inefficiencies to show up in downstream processes that depend on data. Imagine that you are a logistics company, delivering packages to millions of customers, and you discover that many of them have moved, divorced, or had other life changing events

where their data are either irrelevant or incorrect – this could mean re-routing your trucks, initiating 100s of data quality improvement processes, and hiring/firing 100s of employees, contractors and consultants over an extended period – imagine either eating those costs or passing them onto your customers –such scenarios represent the real world cost and implication of bad data quality in today's data hungry world! Data quality at first blush sounds harmless, but is one of the major problems to be addressed going forward – or it is almost certain to implode your organization from the inside out – death by 1000 cuts, as we like to say. Studies (e.g., by KPMG) suggest that even 10% erosion in data quality can have a 50% impact on your process efficiency. That goes straight to your bottom line. On the other hand, imagine the reverse, a 10% improvement in data quality implies up to a 50% enhancement of efficiency, or a 5x return on your investment, out of the gate. The amplified impact of data quality means that data quality is to process costs, as leverage is to institutional risk. It will make or break your business.



Understanding Data Quality Dimensions

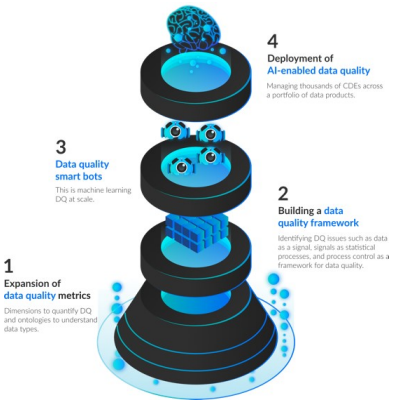
Today, when we think of data quality, we are often very 1-dimensional in our understanding; we think it is an "IT" problem. We don't yet have the nuance or bottom-line impact understanding of how quality is hurting us. We might cite "completeness" as a data quality issue (e.g., missing values in on-boarding processes); but it is important to recognize that downstream applications often see data quality as at least 3-4 other dimensions (accuracy, relevance, validity, etc.). In the future, in an AI empowered data hungry world, we may be talking about 9-10 data quality dimensions, which we aren't even measuring today. So, no organization that cannot express data quality issues along these multiple dimensions can truly call itself "AI ready". For instance, data that is not delivered to an application in a timely manner could have tremendous physical consequences. Likewise, if data are not credible, or lack consistency, then this can lead to outright disasters, for instance, what happened with the Macondo disaster – an entire oil & gas platform blew up as the engineers couldn't react in time to conflicting data from two different well-head sensor readings that indicated problems in poorly capping the well. Data that is not accessible is no good to anyone.



Deploying DQ.ai: A 4-Phase Solution

We are proposing a 4-phased AI-empowered data quality deployment process that build consecutively on top of one another towards a DQM 4.0 model, as we discussed earlier. Our phase 1 starts with rigorously defining and expanding on the set of data quality metrics we use to measure impact, across the organization. This allows us to size the problem and tackle it using an ROI based approach, that is not only cost effective, but incremental and allows us to see the impact right away, creating a self-funding virtuous cycle for driving further innovation across the organization. In phase 2 we build a statistical framework to monitor, track, quantify, and influence data quality processes. A data quality process looks at data quality in an end-to-end manner: from source to bottom-line. By connecting to various upstream parts of the data quality process, we can influence the downstream impact of data quality, and measure that impact based on data-driven benchmarks. In stage 3, we prototype and test what we call "data quality smart bots" which are designed to influence the data quality process, by measuring downstream impact, as a function of upstream levers. These bots "learn" to adjust data quality and adapt to changes in the data stream due to irregularities in the underlying process. By custom curating the data, the bots will demonstrate the ability to successfully reduce quality issues over time, through an iterative machine learning process. Depending on the volume and velocity of the data, and the nature of the underlying volatility in the data, the bots will learn faster or slower on different data quality processes, but over time they will reduce the noise in the data in a measured way. When the prototypes have been successfully tested "in the field", they can be deployed across the organization,

across market areas, across geographies, across lines of business, and in Stage 4 they will effectively operate as a "portfolio" solution: 100s or 1000s of "bots" will control 1000s or millions of processes, managed by a handful of data scientists on staff. A portfolio of data products means, the organization can start to monetize their data in other creative ways, outside their application domain. For instance, some organizations are looking to buy/sell their data to other organizations that will consume that data to drive value. Here, data products are built on a rigorous foundation and can be priced and delivered with exacting precision to customers, to satisfy their particular needs. That customized self-serve curation carries a premium in terms of the value of the data, as a function of its utility. These data are now ready to be traded in a data marketplace, where demand will set the fair economic price for data, as judged by its informational value to the market.



Author

Pradyumna S. Upadrashta, Chief Science Officer, Mastech InfoTrellis.

About

Mastech InfoTrellis partners with enterprises to help them achieve their business objectives by leveraging the power of data to derive deep, analytical insights about their business and its operations. We accelerate business velocity, minimize costs, and drastically improve corporate resiliency through personalized, process-oriented programs, consisting of strategy, data management (including master data management), business intelligence and reporting, data engineering, predictive analytics, and advanced analytics. Part of the NYSE-listed,$177.2M, digital transformation IT services company, Mastech Digital; we drive businesses forward around the world, with offices spread across the US, Canada, India, and Singapore.