12821 East New Market Street, STE #310
Carmel, IN 46032
Phone: (317) 514-6172
Fax: (317) 534-3176
www.metamorsys.com

WHITE PAPER: AUTOMATED DATA QUALITY

Background and How to Start this Initiative

12821 East New Market Street, STE #310
Carmel, IN 46032
Phone: (317) 514-6172
Fax: (317) 534-3176
www.metamorsys.com

Background

As we discussed in our previous white paper on data quality, good data is imperative for any data-driven innovations to provide good insights. To further reinforce this, Information Age provides this article on the use and pitfalls of bad data in machine learning (https://www.information-age.com/automate-data-quality-machine-learning-123485510/). As noted before, even the most sophisticated algorithms will go wrong if the data behind it is poor. Going further, major negative consequences can happen if we rely too heavily on technology based on data with quality issues. As noted in the Information Age article, healthcare outcomes can be affected by bad data quality. Therefore, we know that data quality isn't just "a good idea", it is critical to the infrastructure of innovation.

As explained in the previous white paper, data quality can be investigated, quantified, and addressed by existing tools. These manual interventions are very good as a starting point but aren't the goal. This isn't to take anything away from humans addressing data quality issues themselves. Indeed, this is a needed step to understand the breadth and depth of any issues that may exist but were previously unknown (or known by only data analysts who didn't have a "seat at the table" when it comes to strategic initiatives). Metamor Systems has found that, as an organization, the manual step is indispensable for the collective "institutional knowledge". That is, many in the organization need to learn and embrace the data issues found and resolution has to be a shared goal of the organization. Without this, this tends to be only one person's (or a few) personal crusade that is often sidelined when the next "shiny" opportunity is presented. Improving data quality is not "sexy", but it is critical for the foundation of innovation moving forward.

Improving data quality must be a top-down priority. That is, someone with enough "political capital" must invest some of that in data quality improvements. That's because data quality improvements don't have a quick return on investment (ROI). The improvement in data will improve the bottom line. However, it isn't going to be the "end of month" revenue enhancer that marketing campaigns may be. Therefore, improving data quality is often a "hard sell" in that it takes some time for the benefits to show itself. Often the best visual of improvements can be seen in existing reporting. Presumably, existing reports will be affected negatively by bad data quality. To provide a "before and after" picture of the existing reports will be able to highlight the progress made by the data improvement team. If an organization has started machine learning initiatives, then focusing on the underlying data for that will also be a good test of data improvements. All-in-all, senior executives must buy-in to data improvement processes and enable enough time to see the true impacts. Given that is a longer timeline than some other cost-cutting or revenue enhancing measures, the senior person must provide the team with enough time and space to effectively do the work.

12821 East New Market Street, STE #310
Carmel, IN 46032
Phone: (317) 514-6172
Fax: (317) 534-3176
www.metamorsys.com

Additionally, tests must be set up at the beginning to evaluate the data quality results. Otherwise, it is difficult to quantify the improvements and even judge if the issue has been resolved. Creating test cases is - of course - normal course for any changes to a system. However, it is surprising how little attention is paid - sometimes - to testing. The idea of robust testing is often minimized because of its lack of clear ROI. It takes time and effort but the value is in what it prevents. Unfortunately, humans are not programmed to think about "what might have been". So successful testing is often the "unsung hero" because everything looks like it should when it's going well and many won't stop to think about what would have been the outcome if it weren't in place. Therefore, a key component to advocate for in implementing data quality improvements is ensuring that a robust testing regime is implemented. One way to highlight the value of testing is to publish the results of the test cases at a high-level: number of tests performed, initial success rate, number of bugs identified due to testing, and bugs resolved.

When test scenarios are in place and working, then that lays the foundation for automated testing of data quality. Running and re-running the test scenarios across the data sources will identify if the resolution to data quality issues is a one-time effort or if it needs to be recurring. While some data issues may be a one-time fix (e.g., ensuring that form entry adheres to certain data standards), other data issues may change and evolve over time. Metamor Systems has found that oftentimes, data quality issues morph and evolve over time. Therefore, we know that data quality isn't a one-time initiative. It requires constant vigilance and investment in an on-going monitoring process.

Automated Data Quality

Related to turning test cases into automation, the following example outlines one organization's effort to build a automated data quality system from the ground up: https://medium.com/people-ai-engineering/data-quality-automation-with-apache-spark-ac87cbbf3c37. While admirable, many organizations will find developing a completely new system from nothing to be a daunting task - especially when working with something as fluid as data quality checks. Therefore, Metamor Systems invite you to engage us in implementing automated data quality for your organization.

To reiterate, the details of the above test cases and results of previous data quality checks are absolutely critical. Metamor Systems can anticipate some data issues from our experience. However, every organization is different and and has different quality issues. Therefore, the first step in implementation (aside from executive buy-in and project kick-off, which is another paper altogether), is to work with the key data analysts and developers. The people who work with the data day-in and day-out will know the details and what to look for. Our task is to codify the "institutional knowledge" into repeatable and discrete tests. This will then make it possible for those same data analysts to get back to mining for insights rather than dealing with poor data.

12821 East New Market Street, STE #310
Carmel, IN 46032
Phone: (317) 514-6172
Fax: (317) 534-3176
www.metamorsys.com

Business-Facing Interface

Another key to the success of any automated quality system is for the "business rules" (those criteria to watch for) to be adjusted and added by business people. Oftentimes, the critical data analysts are not located in the IT department. Additionally, their technical skills often are in data manipulation and statistical analysis rather than programming. So expecting the key people who know the data to become programmers of a new system is unreasonable. Additionally, turning these people into programmers will eliminate the valuable jobs that they do: identifying business opportunities. So the interface for making the system run must be in an easy-to-use interface that must enable non-programmers (and non-IT staff) the ability to change the rules. Of course, a process needs to be in place to manage any changes. Just like any production system, only specific people must have the authority to update the rules of the system (presumably after testing in a test environment). Fortunately, the changes to this system simply become another "change request" submitted to a Change Approval Board (CAB) on a regular basis.

Alerts and Notifications

In addition to the interface, alerts and notifications must be sent to the right individuals. All too often, Metamor Systems have seen notifications go to someone at a higher level (due to the high-profile nature of an initiative) only to have those notifications ignored because the person either doesn't have time or is uncertain about what to do. Therefore, from the beginning, a notification and escalation process is needed to be developed. This is not technology and more about the human element of dealing with issues and resolving them.

If there's not one already, this is a good time to consider creating a Chief Data Officer (CDO) position (or similar). Having IT or a business person in charge of the data quality might work in some circumstances, but Metamor Systems has seen too many times that:

- **The IT manager/director/VP has too many other technology issues to deal with to thoughtfully determine what to do with a new data quality alert**
- **And a business person may have too many business-specific issues to deal with also**

Therefore, someone who is at a high-enough level to get the attention of senior management to be able to do something is key. Additionally, this cannot be "just another duty" added to someone's already busy day. If a CDO isn't feasible, then a matrix team of application and data analysis staff can be a workable solution. That is, there must be real collaboration between the application team (where data is created or ingested) and the data analysis staff (who analyze the output). This matrix approach can surface data issues and good solutions because there are

12821 East New Market Street, STE #310
Carmel, IN 46032
Phone: (317) 514-6172
Fax: (317) 534-3176
www.metamorsys.com

people who are involved from beginning to end. It is always best to create solutions when there is an understanding of the input and output of a process.

While there is much more to explore in implementing an automated data quality system, please contact Metamor Systems for more information. It's not a trivial matter, but we have been improving data quality for many years. We can help your organization too.