4 KEY FACTORS FOR DATA QUALITY ON A DATA LAKE

(OR: HOW TO AVOID THE DATA SWAMP)

JOSH HERRITZ | MIOSOFT CORPORATION

The trends in digital business promise that the future holds an unprecedented volume, variety, and velocity of data. For organizations that already struggle to manage their existing data systems, that promise can seem more like a threat.

But the basic solution to an overwhelming number of separate systems seems clear: empty all of the systems into a single place.

Forget trying to extract meaning from dozens, or hundreds, or thousands, of separate systems. Forget your analysts wasting time while they wait for IT to get them the data they need. Forget your operational systems being brought to their knees because someone made a typo in their query.

Just pour everything—all the data, from everywhere you have—into one place and work with that. No IT bottlenecks. No working directly with the source system. Anyone can just walk up and pull out the data they're looking for.

That's the philosophy of the data lake. But it hasn't lived up to its hype.

Instead, people are disillusioned with the data lake. On-premise solutions like Hadoop and Spark were the site of the first attempts, but they didn't deliver. So now data lakes are moving to the cloud, in the form of Amazon's AWS S3, Google's Cloud Storage, and Microsoft's Azure Blob Storage and their associated analytics engines.

The data lake isn't unreasonable. It's just not as simple as it seems.

But acting like it is that simple, then wondering why the value never materializes, is why Hadoop- and Spark- based lakes failed. If you want your cloud-based data lake to avoid the same fate, you need to take proactive action to assure that your data lake doesn't become a data swamp.

The good news is, this proactive action is completely achievable. This whitepaper will look at the four key ways that you can use data quality to keep your data lake clean and ready to deliver value—and avoid the data swamps of the past.

THE SWAMP PROBLEM

A big data lake will take whatever data you give it, and that's going to be a problem.

Unless your current systems all contain nothing but the purest, highest-quality data imaginable: no formatting differences, no missing values, no typos, nothing ever entered in the wrong field. Also, all of your current systems are perfectly consistent with one another and their information never conflicts. Then there's no problem.

But if your organization operates in the real world, all of your systems contain data that is a little bit dirty. Depending on your current efforts, your data could be anywhere from “slightly grimy” to “in dire need of a bleach power wash.” Most likely it’s somewhere in between, and different for different systems.

But when you pour data from all of your systems together, all of that “slight grime”—or worse—starts to build up. And more and more dirty data keeps coming: a whole Lake Superior’s worth of dirty data.

It doesn’t take long before all that dirty data grime makes your data lake look more like a swamp.

Think of the insights you’re hoping to get out of your data. Do you want that insight to come out of a crystal-clear lake full of accurate, consistent, high-quality data? Do you want your insight to reflect the real entities your data describes?

Or do you want to take your chances with “insight” based on whatever misshapen swamp creature emerges from your dirty lake?

THE SOLUTION

No business leader wants to make decisions based on a swamp creature. Especially not if that swamp creature is a metaphor for “potentially-erroneous conclusions derived from dirty data in the organization’s data lake.”

To avoid the swamp creature, your organization needs data quality for the data lake. There’s no way around it.

Of course, no business leader wants to give up another of the data lake’s essential promises: that IT won’t be a bottleneck between the data and the analysts anymore.

But that isn’t a problem. Yes, IT is traditionally in charge of data quality. But a true data quality tool for big data doesn’t stick to the “IT does this” paradigm.

A true data quality tool for big data knows that it’s going to be primarily—maybe even only—used by non-IT, non-programmer users. It doesn’t need you to learn how to code before you can get anything done. It shows users clearly what’s being done to the data, and how. It makes it easy to do complex things.

A true data quality tool allows you to have all the benefits of the data lake, and all the benefits of data quality.

DATA QUALITY FOR A (BIG) DATA LAKE

The point of a data lake isn't just to have a data lake.

The point is to turn data into information, information into knowledge, and knowledge into action—all informed by the people who know your data best.

In other words, trustworthy data is essential to achieving your data lake goals.

You just need to be sure that you're getting a data quality tool that is truly suitable for big data. What you don't want is to get stuck with your standard, IT-centric data quality tool with a new label on its box.

Here are the 4 key factors you can use to identify a true data quality for big data solution.

KEY #1: ACCEPT THAT PEOPLE AREN'T ROBOTS

Your end users are people—and people are always going to make mistakes. A big data quality solution can work with that: it doesn't demand impossible, robotic perfection.

The data in your data lake ultimately originates from human beings, entering data into a front-line system somewhere. And because to err is human, the people entering your data are going to make mistakes, no matter how careful they are.

An effective data quality tool for big data accepts this, and can work with the data as-is.

Yes, part of a data quality initiative is going to include encouraging and supporting accuracy in data entry. But you can only do so much, and so can they. Mistakes are inevitable.

Mistakes are a bump in the road, but they don't have to derail the whole train.

The point of a data lake was to bring a lot of data from a lot of sources together. That means the right answer is in there somewhere, because reasonably careful people should come up with the right answer more often than not.

A data quality tool for big data can use validation by consensus. In other words, it should use the wisdom of the crowd: when values conflict, it should work *across* all the data sources in the data lake to automatically determine which data is correct. It should know when a misspelling can be overridden and when slight mismatch can be ignored.

Mandating perfection is doomed to fail. Mandating manual correction puts people to work doing something that software can do, while preventing them from doing what software *can't* do: innovating and experimenting to invent the new ideas that will help you succeed as a digital, data-driven business of the future.

KEY #2: EMBRACE ITERATION

Iteration in big data quality is inevitable and desirable. A big data quality solution needs to embrace that.

A data lake will contain data of all shapes and sizes, from countless source systems. Anyone working with this data will encounter a wide variety of data quality problems.

No one can discover all those problems, and figure out the best way to solve them, on the first try. That's true no matter how deeply they understand the data, or how experienced they are.

Each system is different, and its data quality problems will have their own quirks and learning curves. Some mistakes front-end users will tend make are obvious and predictable, and some aren't. An inflexible, pre-planned approach is bound to encounter conditions it didn't anticipate.

An effective data quality tool for big data doesn't just tolerate this fact: it embraces it. It prepares you to succeed when you encounter the unexpected. It doesn't demand that you stick to the original plan at any cost.

Users preparing data must be able to rapidly try operations, see their results, change something, and try again. They must be able to side-by-side compare their results with the original data. They must be able to easily see what has already been done to the data, and to review what was tried in the past.

But the need for iteration doesn't stop once data is prepared. At every level of the solution, from creating the solution's shared framework for the data to preparing match and merge rules, users must be able to iterate, creating feedback loops that tell them whether they're on the right path.

A data quality tool for big data provides capabilities that make all this possible. It encourages you not to settle: it encourages you to keep going, keep iterating, until you find the best solution.

KEY #3: GO FOR COLLABORATIVE USABILITY

Big data quality needs experts from all over your company, not just IT. A big data quality solution is usable without programming experience.

Effective data quality and data preparation depends on a deep understanding of the data.

What system did the data come from, and how was it used? What is the domain that the data originated from? How is that data defined and thought of in that domain? What is the intended business use of the project that will use the data?

In a data lake, issues of complexity and cross-domain data are virtually guaranteed to arise for any project beyond the most basic analytics. That makes it absolutely imperative to involve expert business knowledge—even when that doesn't come from an IT department.

Business analysts, data scientists, system end users, and others must all be able to use the tool. System end users should be able to help with data preparation; subject matter experts should be able to help configure match and merge rules for maximum meaning.

An effective data quality tool for big data in the data lake empowers all of the organization's users to make meaningful contributions, even when they have no programming experience at all.

This requires the solution to have usable, visual tools that don't rely on an understanding of any particular programming language.

With these tools, the organization can distribute responsibility for data quality and preparation based on the user's knowledge of the relevant business areas and the organization's operational efficiency.

KEY #4: CREATE AN EVOLVING SHARED BIG PICTURE

Your experts, systems, and projects are fragmented—but your data lake isn't. A big data quality solution uses a shared vision of the organization's data to unify data quality work into something that makes sense.

It's nearly inevitable that your organization will have to distribute responsibility for data quality and involve experts from all areas of the organization. And once distributed, each group's responsibility will only cover part of your organization's overall picture.

You can only succeed if everyone has a shared vision of what that picture actually *is*. How else can

everyone work to put the pieces of the puzzle together?

But the phrase “a model” evokes all the worst IT bottlenecks: static, difficult to iterate, hard to understand, with long waits for new versions.

An effective data quality tool for big data relies on the idea of a unifying, evolving vision and uses a model that works for you, not against you.

An effective data quality model is:

- Flexible, easy to iterate, and easy to view—for anyone, not just IT—while still representing complex entities and relationships.
- Driven by the data quality work, not the other way around. Concurrent work on the model and on the data is possible; changes to the model don’t delay data quality efforts.
- Independent of the project’s output. It’s a shared vision of the data, not a mandate about how that data is sent to other systems.

A data quality model helps you make sure all the puzzle pieces are showing the same picture, without doing too much.

Deciding whether two particular puzzle pieces really fit together is part of the data quality process for your organization. That means it has to be your choice, not the tool developer’s.

An effective model for data quality gives your data quality efforts direction so all your efforts are based on the same big-picture understanding. It reduces complexity without restricting your iteration and evolution as your understanding of your data improves.

Finally, an effective model for data quality doesn’t dictate the shape of your project’s output. When you’re ready to deliver data to your destination systems, you should be able to take only the data that you need—even if that’s not the whole model—reorganized and formatted in the way that you need.

The promise of a data lake is the promise of analytics, insight, and knowledge. To get that, you need to be able to see the real entities that your data describes.

Bad data quality turns your data lake into a swamp. At best, you might be able to get the right general shape of the entities from your swamp’s dirty data. But the details are wrong, obscured and distorted

by data quality problems. Any so-called insight based on these disfigured entities is simply going to perpetuate those problems—with results that can be downright scary.

True data quality doesn't just look at your data and tell you how big your swamp monster problem is. It makes it possible for your data lake to deliver on its promises.

With an effective data quality tool for big data, you don't just see your data quality problems, you *fix* them, so that your lake really becomes a lake, not a swamp. You can make sure your lake's data is trustworthy. You can see the real entities that your data describes, in clear and reliable detail.

With data quality for a data lake, you can get useful, trusted insight that helps you connect the dots and know where your business needs to go.

A true data quality tool for big data will:

#1: Accept That People Aren't Robots
#2: Embrace Iteration
#3: Go for Collaborative Usability
#4: Create an Evolving Shared Big Picture

When you're ready to get data quality on your data lake, keep these four key factors in mind. Ask a prospective vendor how their software will deliver on each of these points, and make sure you get a solution that will help you see the true entities in your data lake.

RELATED READING

The Trailblazer Approach to Data Quality: An Introduction & Definition

Data Quality Vendor Evaluation Checklist

The 3 Cs of Data Quality

MIOsoft®

READY TO HELP

If you're in the market for a data quality solution for a data lake, be sure to check out MIOsoft. Our software supports an iterative approach to deploying our cutting-edge data quality capabilities, including a flexible model, so you can start getting meaning out of your data lake as quickly as possible.

Is your data lake looking more like an ocean? No worries: MIOvantage is designed to scale, so no much how data you're working with, MIOvantage has the capabilities you need.

STANDING BY

Interested in working with MIOsoft to get data quality for your data lake? Get in touch!

sales@miosoft.com