SODA

Achieving World-Class Data Reliability

Today, data touches every part of your business. With better data, companies can more accurately forecast demand, manage inventories, understand risk, and deliver better customer experiences. And increasingly, data is the business--driving new business strategies, new technologies, and new products. Data, in short, is ubiquitous. But it is far from perfect. And bad data – whether it's incomplete, inconsistent, or inaccurate – has consequences. Business decisions suffer. Silent errors go undetected. And data teams spend hours putting out fires that should never have started in the first place. At Soda, we believe that world-class data reliability can be achieved by every organization. For most organizations, this will require a cultural shift. Below, we've identified five steps organizations can take to master data quality so that everyone who relies on your data can trust that it will deliver great results.

01
Reactive

What it looks like.

You know you need data to drive business decisions. But you have very limited visibility into what data you have, how reliable that data is, or who might be able to answer questions about its purpose, use, and quality. You have no clear understanding of data lineage or how your data is being transformed and used across the business. Roles and responsibilities haven't been clearly established. When something breaks, it's often an executive (or worse, a customer) who brings it to your attention. And you have no clear idea about where to turn for answers. That leaves your data teams scrambling for answers and firefighting issues with inadequate tools. The result is often a long email thread involving data engineers, analysts, data scientists, data owners, and operations managers – all of whom use data in different ways and have a different understanding of "what counts" when it comes to reliable data. It's a dire scenario, but not at all an uncommon one.

Notable pitfalls in this phase:

- A lack of visibility into the data you use.
- Dashboards that provide limited insight.
- Lengthy discussions about root cause, terminology, and business rules.
- Data teams working as data janitors.
- A lack of trust in your data that pervades the organization.

How to move on.

First, determine the scope or your data quality initiative. In our experience, an organization's first attempts at overhauling their data quality practices often fail because they're too broad. (We believe this is a primary reason that data governance hasn't seen greater uptake in businesses today). If your organization is in the business of building data products (the end goal of your product or feature is facilitated by data), focus on one and determine the critical datasets and features driving that product, from ingestion to analytics-ready views. (If you're using data to drive better analytics, narrow your focus to a single dataset.) For each dataset, identify the data owners and subject matter experts who have real expertise in why the data is relevant, how it will be used, and how "quality" will be defined for that use.

02
Proactive

What it looks like.

Data teams begin to organize themselves by role, with an understanding of who is responsible for what. Together, data engineers and data experts have determined a set of observability metrics to identify commonly occurring data issues in the scoped data product or dataset. They are determining who needs to be informed of data issues and how. And they are sharing basic observability metrics with data consumers across the organization so that they can begin to work collaboratively to identify any additional metrics that will increase transparency and build trust.

In a proactive organization, data teams have, at a minimum, a process to report on the following metrics:

- Data freshness (arrival times and delays compared over time).
- Data volume (row counts, ideally at multiple points across the data lineage).
- Data schemas (column count and transparency into how well records and events adhere to predefined schema).
- Data validity (adherence to allowed values and value ranges, format, etc).
- Data completeness (identification of any missing values).

Notable pitfalls in this phase:

- Not all data issues have a clear owner or remediation plan.
- SMEs and data consumers are asking for additional reliability metrics based on their understanding of the data.
- Nothing fails like success: everyone wants data reliability metrics for their datasets and data products.

How to move on:

Build on your success. As you get ahead of these early data issues, you'll find more people will trust the reliability of your data products. Communicate these successes. Data SMEs will be more than willing to define additional reliability metrics. Data owners will be more likely to take on additional responsibility for reviewing new data requirements and tracking issues. Begin to link datasets to the data consumers in your organization so that data teams can understand more nuanced metrics such as usage dependencies. Begin to identify and communicate best practices.

03
Collaborative

What it looks like:

By now, data reliability testing is well on its way to being embedded in the organization's culture and processes. Data owners and consumers understand the criticality of observability metrics and want a more convenient way to monitor the metrics that matter. Data teams have started automating their data operations, and testing data has become ubiquitous. Now, hundreds to thousands of metrics per dataset are monitored in a fully automated fashion. Self-service tools allow data owners and consumers to easily define additional metrics and validations. Producers have transparency about what data consumers expect. Critical notifications are sent to the right data owners, SMEs, and data product teams at the right intervals. And every member of a data team can access the data, adjust thresholds, and annotate metrics and results.

Notable pitfalls in this phase:

- Alert fatigue (monitored metrics increase, and often metrics have other dependencies).
- Information overload (it's hard to understand how alerts across datasets are related).
- Root cause identification is still difficult (lack of contextual information slows analysis).
- Disconnected data teams (some teams working independently have different reliability definitions).

How to move on:

At this point, your data teams consider data monitoring as a foundational tool, but a full understanding of data quality definitions and their dependencies is not always available to everyone working with the data. To move to the next phase of data reliability excellence, consider publishing data reliability metrics to your data catalog so that everyone across the organization has easy access to definitions. Ensure that those using your data monitoring systems have comprehensive and relevant information for root cause analysis. And make sure that your reliability metrics are aligned across the organization (e.g. create a shared understanding of data freshness).

04
Contextual

What it looks like.

Data operations teams that have established data reliability metrics, implemented data monitoring tools, and mastered communications and collaboration are perfectly poised to address the next step toward achieving world-class data reliability: contextual alerting.

In this phase, organizations incorporate additional information (typically metadata) to make data alerts more meaningful and impactful. This contextual information allows teams to better describe what data the alert is about, intelligently group alerts, and display related alerts. It also allows them to apply the same logical monitor across different technical manifestations of that data. In our experience, data teams need to correlate historical data across tables to, for example, infer data lineage. They also need to link columns that contain very similar data, and even customize diagnostics information based on the expected

issue type. With a fully contextualized understanding of the data issues presented to them, data teams are less likely to be overwhelmed by alerts and better equipped to solve issues.

Notable pitfalls in this phase:

- As data points and data tables grow in number and transactions reach the millions, alerts, even contextualized, will be hard to manage.
- Data teams have no way to predict what data issues are most likely to occur.
- Data teams have no easy way to evaluate which data issues to prioritize.
- Data teams are not sure who analyzed what alert

How to move on.

While contextual information is key to solving data issues, the demand for high-quality data will only continue to grow. Data consumers need to understand quickly what datasets are reliable. If you have published reliability data to your data catalog, consider assigning it a reliability score. When new data consumers find a dataset that they need, they can easily view a reliability score as well as any recently

detected issues. For data teams, the challenges are operational. Consider leveraging statistical modelling to tie together different pieces of data quality and reliability information. Use historical patterns as a baseline, and start simplifying the multi-dimensional nature of your data quality metrics. Condense this multi-dimensional, historical data into a single anomaly score, using techniques like Principal Component Analysis (PCA).

05
Predictive

What it looks like.

Predicting data quality issues before they occur? It's possible. And today, it's becoming imperative. Organizations that have tens of thousands of tables and hundreds of data points and data owners require an additional, predictive layer for data reliability and quality management. To manage that volume, forward-thinking data operations teams are implementing forecasting systems that analyze leading indicators to predict which issues are about to occur. You can, for example, predict that data needed for a critical business process is going to arrive late. Data teams, guided to the set of metrics that caused the alert to fire, can analyze and prioritize these predictions and take appropriate action.

Pitfalls?

Not many. At this stage in your data reliability journey your data teams are able to easily anticipate and solve data issues. Your data teams have a shared understanding about what data quality means and their roles and responsibilities are clearly defined. Data scientists, data analysts, and data owners collaborate more enthusiastically to resolve issues quickly. Everyone across your organization knows that processes are in place to deliver trustworthy data and they believe in the ability of your data teams to help the business make better decisions, build better data products, and drive innovation for a more sustainable future.

Want to learn more?

The last couple of years, so much has changed in this area. It will be very interesting to see what the future holds. Soda is eager to help you on your journey to modern data management.

Let's Talk