Make the case for a predictive and self-service data quality tool
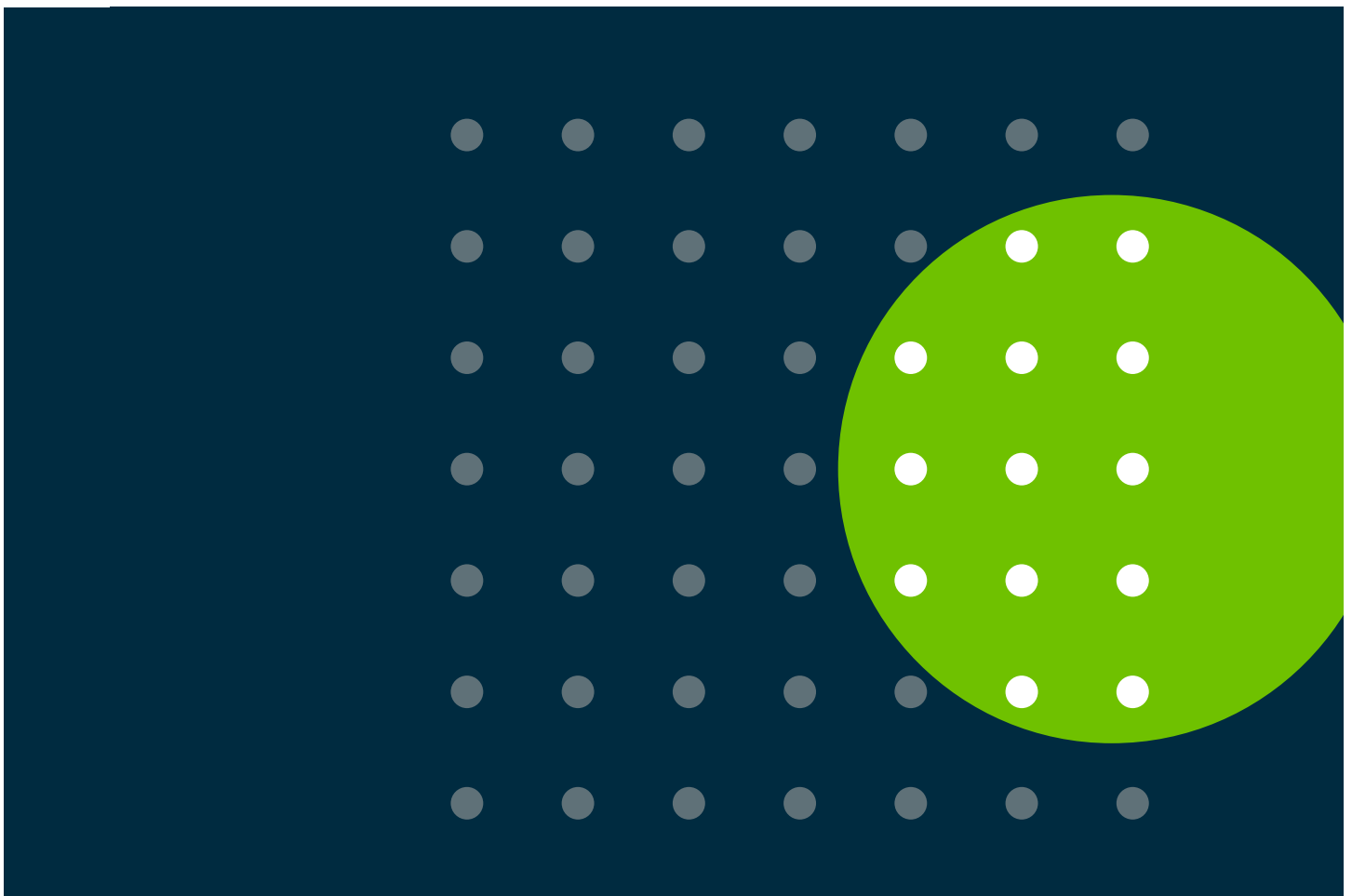
Create an enterprise vision for data quality

Executive summary

As organizations turn their focus to better leveraging their growing volumes of data, key business and technical stakeholders are working through the long, arduous process of making the case to formalize various data capabilities and investing in the related technologies.

A common reason to invest in a data strategy is an overall need for better data understanding and easier access to quality and trusted data to support operational and analytical activities. Given that, why is an enterprise data quality tool commonly an after thought or put at the end of a wish list?

It's likely because talking about data quality issues or quality management can be overwhelming or too theoretical. Those seeking to establish a successful enterprise data quality function need a concrete path to build confidence and enthusiasm when making the case to invest in resourcing and a new tool.

Prepare to make the case for quality

To convince others of a vision of high-quality data and its many benefits, the visionary themselves need to have a thorough knowledge of what the data quality function is comprised of and be able to relay the values of a modern predictive and self-service data quality tool.

This white paper outlines the critical components of a successful data quality function and considerations on how to get there.

Data quality:

the extent to which data represents what it purports to represent and the extent to which data satisfies a specific requirement

Determine your why

> Improving data quality can benefit the organization in a myriad of ways, and it is important to prioritize those benefits rather than try to "boil the ocean."

When getting started in a data capability (or improving an existing one), it is important to articulate your vision of success and why that success is important to the organization as a whole. Improving data quality can benefit the organization in a myriad of ways, and it is important to prioritize those benefits rather than try to "boil the ocean."

Here are examples of ways that data quality can add value to an organization:

- Address regulatory issues around data quality, access, and sharing

- Ensure KPIs are correct and trusted

- Support financial reporting needs and a more efficient close process

- Facilitate new CRM and ERP implementation by addressing data quality issues from legacy systems

- Optimize machine learning model risk management

- Expedite cloud migrations and new advanced analytics technologies

By identifying the impact that the data capability can make, you can create a plan to deliver value to the company as quickly as possible while building a sustainable practice.

The following questions can help to elucidate that value:

- What are your companies' strategic initiatives?

- Why are they critical, and what are their expected outcomes?

- What data do they need to be successful?

- Where is that data located?

- Is that data "fit for use" and able to support the outcome? If not, why?

- What is the impact of data quality on those strategic initiatives?

- Can you determine the impact of poor data quality on other key processes?

Any organization's goals are going to change over time, and this line of questioning should be frequently revisited at an enterprise, department and data domain level to ensure alignment and progress is being made.

Establish your data quality function

> A defined data quality function in your organization would mean there is a mechanism to set expectations and support data citizens when quality issues arise.

Data citizens (data creators and consumers) should be accountable for reporting or correcting data quality issues when identified. A defined data quality function in your organization would mean there is a mechanism to set expectations and support data citizens when quality issues arise. This function would provide guidance and best practices to solve common data quality challenges. For example, sharing a simple standard that a reference data code set dropdown list (or pick list) could be leveraged as opposed to a free-form text field pays dividends across the data landscape and for years to come.

A data quality functional area's primary accountability is establishing the mechanism to define and communicate data standards and practices (like the above), work to incorporate them into standard processes, such as an SDLC, and oversee they are being followed. This area is also in place to educate and inform the organization of data quality practices and ensure tools are in place to deliver the best outcome (see Data Quality Steps section).

Following are the building blocks for establishing a data quality function:

- Documented charter, including mission, vision, objectives and scope of activities

- A value-focused roadmap to establish and scale data quality capabilities

- Defined roles and responsibilities, including a technology administration model

- Functional impact and progress metric reporting

- Data quality function engagement model aligned to support stewardship

- An organizational change management plan

This service area is typically a small-but-mighty functional team that sits within an enterprise data governance/management organization, with often a dotted line to IT when the corresponding data quality technology needs administration. The team's purpose is to promote data quality practices and ensure data owners, stewards (business and technical) and other data citizens have what is needed to inform data quality in an efficient and effective means across data's lifecycle.

Team members usually include:

- **Data Quality Lead/Manager,** who sets the vision and strategy for the data quality function, executes the strategy and is accountable for measurable enterprise-level impact and progress. An organizational change management background is required for this individual to be successful. This role is focused on promoting awareness, focusing on adoption and expansion.

- **Data Quality Analyst(s),** the experts who understand all aspects of data quality and typically never want to do anything but profile data, talk about data dimensions, translate the findings, or find new rules or more efficient ways to apply the rules broadly across the data landscape.

- **Data/DataOps Engineers,** the developers who implement the data design based on data requirements and typically hold an abundant amount of knowledge about the enterprise's data, how it is used and where there are quality and trust issues.

Depending on the size and complexity of your organization, there may be analysts assigned to support enterprise-level quality initiatives, as well as those aligned to a specific functional area or data domain. The goal of any data quality analyst should be to help data stewards become the experts of their data and drive positive outcomes.

> **Data quality management:**
> Practice of defining expectations of data, monitoring for conformance to expectations and correcting non-conformance.

The number of data quality analysts also depends on the maturity of this functional area and the volume of use cases and data. At least one skilled data quality analyst is required to get the function going, either through cross-training or as a new hire. These individuals also monitor results and help to identify quality issues and ensure remediation occurs.

In the past, a leading measurement of data quality function growth was expanding the team by bringing in another quality analyst. Collibra's automated machine learning rule identification is helping data analysts scale monitoring of data sources, as well as bring data stewards to the decision-making table, to review and approve the automated rule recommendation. Collibra's new-generation data quality tool has had a positive impact on productivity through self-service and embedding quality rules more broadly.

Without designated resources and a supporting tool, it is important to recognize much of the burden to enforce and monitor data quality falls on DataOps engineers. Due to lack of visibility into the data pipeline and no means to continuously profile and monitor, their day-to-day activities are reactive, manual and inefficient. DataOps is becoming mainstream in data management, and data engineers are critical data citizens. It is imperative to provide a means to support data development and automatically validate production data quality across a data pipeline.

Data quality steps

As the visionary trying to make the case to bring more rigor around data quality, it is critical to know the various data quality-related activities and their order. Knowing these aspects can help you help the organization recognize they are already doing data quality management, but likely not in the most efficient or effective way.

When developing a new report or testing out a new machine learning model, many hours go into knowing the data needed to be successful and introducing quality controls within the code, as well as standardizing, storing and creating access paths to the data to deliver the anticipated results. Getting to the right results typically requires collaboration between business and IT, with individuals of varying skill sets and managing multiple responsibilities.

For example, when a new report or model is migrated into production, the initial team remains dedicated to a successful launch. Output is monitored closely and, often, the IT project team is still available to help find and fix any production issues. The cracks start to appear when the data asset (e.g., report or model) goes into maintenance mode and is handed off to a production support team.

The seven steps below outline activities needed to ensure data is assessed, monitored and measured against expectations for use. A project team (such as the example above) typically performs the first four activities very well with or without an automated data quality technology, resulting in a successful implementation. Once in maintenance and without a predictive technology, such as Collibra Data Quality & Observability, many of these steps are highly manual and require individuals to watch over the data flow. It is likely that quality issues are not caught.

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Data Profiling	Data Quality Assessment	Data Quality Cleansing and Standardization	Data Quality Monitoring	Data Issue Management	Data Issue Remediation	Data Quality Performance and Impact Reporting
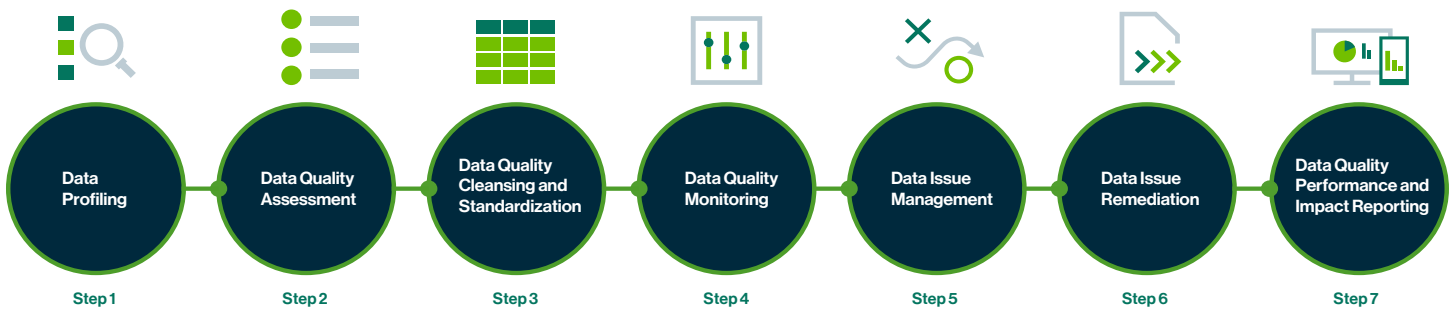
First San Francisco Partners' recommended seven steps to maintain quality results

This section explores the typical roles engaged and activities in this work — with and without an enterprise-level data quality tool — to show how a formalized data quality function results in better-established data governance within your organization.

1. Data Profiling – Acquire and assemble data quality dimension measurements and classification results of targeted data sets and data sources.

Traditionally, we defined metrics by how many nulls a column could have, for example. The newer generation tools profile the data and generate a baseline for all technical metrics and forecast the expected rules automatically.

No enterprise data quality tool

Most technologies used by analysts or developers have profiling utilities or the ability to query the database for results (e.g., completeness, uniqueness, validity, integrity). Data quality analysis usually requires custom code or SQL to validate specific columns in a dataset. Developers and data analysts commonly use Excel for presenting results, limiting the ability to scale.

There is no user interface to inform the findings, and the consumers need to have an extensive technical background to decipher the outcome.

Collibra Data Quality & Observability

Connecting to a source and scanning the data to generate profiling results is easy. Once set up, generating results for the same data source later is a button-click.

The ease of operation opens Collibra Data Quality & Observability up to a broader group of people with a stake in data quality. The profile created is used on a later step to provide insight and automatically identify data quality issues.

All Collibra Data Quality & Observability generated checks and rules are adaptive and explainable, constantly learning from new data and making predictions for typos, formatting issues, outliers and relationships.

Example of detailed profile for each dataset up and under management

Column	Sensitive	Semantic	Data Type	Baseline Profile Discovered	Quality		Uniques	Run Profile Discovered	Quality		Uniques	Ru Un %
activity			String	String	100.0%		58	String	100.0%		58	
activity_id			Integer	Int	100.0%		15978	Int	100.0%		15978	
audit_ts			Timestamp	Timestamp	100.0%		15978	Timestamp	100.0%		15978	
log_string			String	String	100.0%		1362	String	100.0%		1362	
session_end_ts			Timestamp	Timestamp	86.3%		1462	Timestamp	86.3%		1462	
source_ip		IP ADDRESS	String	String	99.0%		279	String	99.0%		279	
source_session			String	String	99.0%		1783	String	99.0%		1783	
username		NAME	String	String	100.0%		167	String	100.0%		167	

2. Data Quality Assessment – Leverage profiling results in determining if the dataset will meet the intended need and any additional quality considerations. This is where data quality rules are identified, and agreement is made on how best to apply them — once implemented.

No enterprise data quality tool

Engaging and valuable stakeholder conversations are rare without showing scorecards within a UI or profiling metrics.

Organizations frequently miss quality issues, and decisions about setting up rules and monitoring tends to be done by IT.

Collibra Data Quality & Observability

Results can be quickly shared across various roles, including those responsible for sourcing the data. With a standard user experience for profiling results, an enterprise can promote the collaboration of business and technical stewards with IT, data analysts and consumers and potential owners, which is a model for data governance.
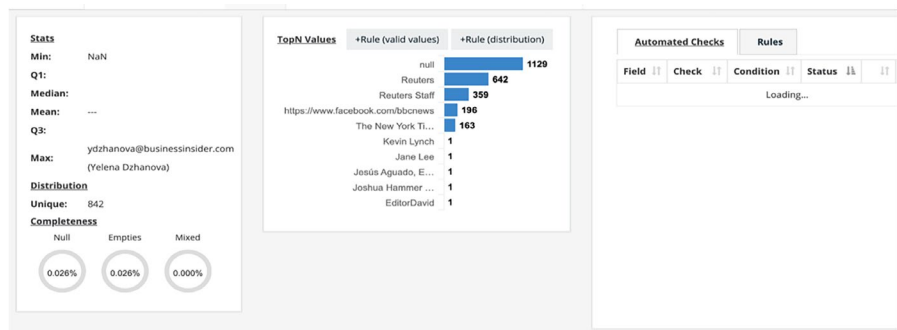
Output view of the profiling step, which can be valuable to determine results of the assessment.

Users can also view a sample dataset with results per column, with the option to mask the data in the case of sensitive columns.

3. Data Quality Cleansing and Standardization – Ensure the data is in the expected format and data quality rules are applied based on analysis in the previous steps. Cleansing and standardization to a dataset should be done by coding within a system, ETL or reference/master data management (MDM) technologies instead of using a data quality tool such as Collibra. Recognize there are hundreds, if not thousands, of ways to make dataset modifications, but leveraging other enterprise technologies promotes reuse and consistency across your data landscape. Where a data quality technology does support this step is implementing data quality rules/checks in the data ecosystem to ensure the requirements for cleansing and standardization were implemented as intended and consistently.

No enterprise data quality tool

The value of a standard rule library is one of the most overlooked components of an enterprise data quality tool. Without a central location to share data quality rules and apply them consistently, the rules are interpreted from initial requirements and are applied using various technologies.

When quality issues arise, it takes extensive resources to triage, and it's likely the quality issue has already wreaked havoc across the enterprise.

Collibra Data Quality & Observability

Collibra Data Quality & Observability has a standard data quality rules library, owned by the business (if you are running your data governance program correctly), and accessible to other enterprise data management technologies to apply common treatment to the data, no matter where you are in the data landscape (e.g., operational to analytical data stores).

A sample listing of rules in Collibra Data Quality & Observability

4. Data Quality Monitoring – Perform quality checks, executing the profiling and data rules. Monitoring is typically scheduled and implements the requirements a data owner or steward identifies, including how best to handle the data when quality issues arise (e.g., bypass records or stop processing and then notify).

No enterprise data quality tool

Similar to data profiling, several mechanisms monitor quality across the data landscape and generate alerts. But as the technology varies between data stores, so do the notifications and formatting, how they are sent and to whom.

This often means reading into what an alert is trying to inform and inconsistent roles handle the situation.

Collibra Data Quality & Observability

There is a unified scoring system to report across all data sources, including personal alerts, which a broader group can also access.

Also, prior profiling results of monitoring are automatically baselined, allowing Collibra's machine learning to identify additional data rules.

Horizontal and vertical scalability helps establish enterprise-wide trust in data with the ability to scan large and diverse databases, files and streaming data.
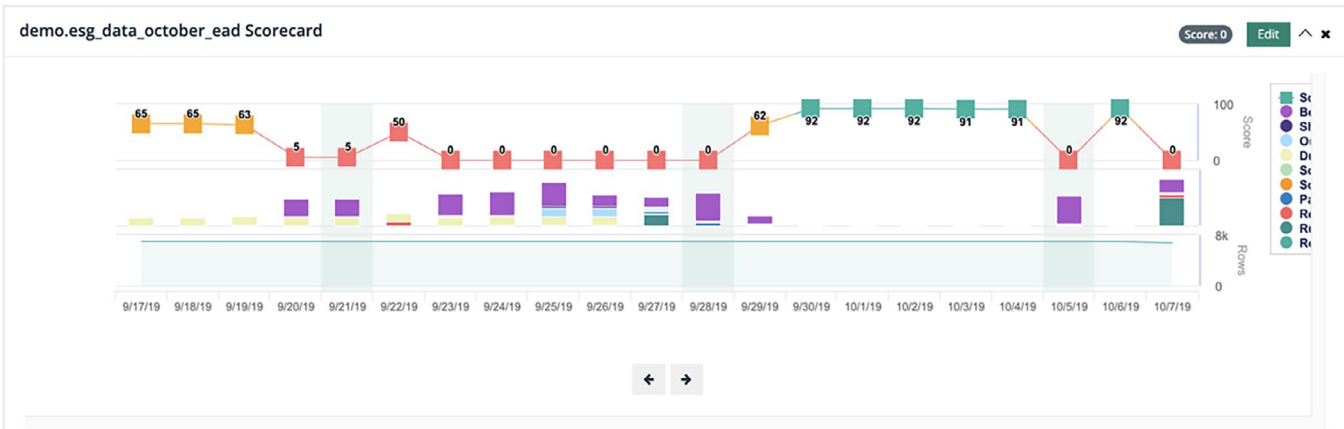
Scorecard of a specific job that runs daily and the number of issues caught by type (outlier, dupe, rule breaks)

5. Data Issue Management – Perform oversight whenever a data exception is identified; common activities include:

• Assign a data issue criticality rating to ensure high-priority issues are addressed first.

• Notify the business owner/data owner when an issue in their area of responsibility is logged.

• Assign responsibility for a data exception to a data steward and custodian.

• Track and report status.

• Notify impacted data consumers of the exception and plan of action. (Once resolved, share this with them, as well.)

No enterprise data quality tool

There is no standard way to share issues and updates with a broader stakeholder group, nor is there an automated, consistent way to alert those potentially impacted, both upstream and downstream of the data exception.

This means missed opportunities to engage stewards and stakeholders.

Collibra Data Quality & Observability

Business-friendly scorecards can be integrated into the catalog for broad communication of the impact, and workflows are used for notifications.

6. Data Issue Remediation – Identify and correct what caused the data issue; for example:

• Conduct a root-cause analysis to determine the reason for the data exception.

• Determine corrective action, which may include correcting a software defect, implementing a business process change and providing business-user training, correcting an input file, or potentially adding a new data quality rule as part of monitoring.

• Test and implement the corrective action.

• Remediate the erroneous data in the impacted databases.

No enterprise data quality tool

Like the statement in the Data Profiling section, manually creating profiling results to support root cause analysis is complex and is not dynamic to accommodate remediation. Additionally, because it is a separate tool, the content is stored separately in a static platform, leaving the whole process inefficient.

Collibra Data Quality & Observability

Users review the business-friendly scorecard and historical results as the starting point of root-cause analysis. The unified scoring system can also be used to view row, column, conformity and value checks between the source and target datasets.

Data owners and stewards can be notified via a configurable DQ workflow to initiate remediation when data quality scores drop below the target threshold. In addition, the end-to-end automated data lineage helps stewards and data quality teams narrow the focus of root cause investigations and can prioritize issues.

7. Data Quality Performance and Impact Reporting – Report various measures to assess the quality of an organization's data quality function, such as the number of identified issues for each data asset, number of assigned data owners, stewards and custodians, or average or maximum resolution time from detection to resolution.

No enterprise data quality tool

A custom reporting solution needs to be created to collect and store various data related to data quality management, profiling, issue identification and resolution.

This would be costly and likely something difficult to maintain.

Collibra Data Quality & Observability

An enterprise data quality tool is set up to support the work of an enterprise data function, which includes associating data stewards and quality analysts to particular datasets, running processes via workflows and capturing data quality-related results in a central location. This shared, unified location can be leveraged to report out the overall health/maturity of data quality for an organization.

For example, it is common for the Chief Data Officer to require a report showing data quality coverage across all technical systems or business units.

Sample day-by-day monitoring scorecard set up for a business group
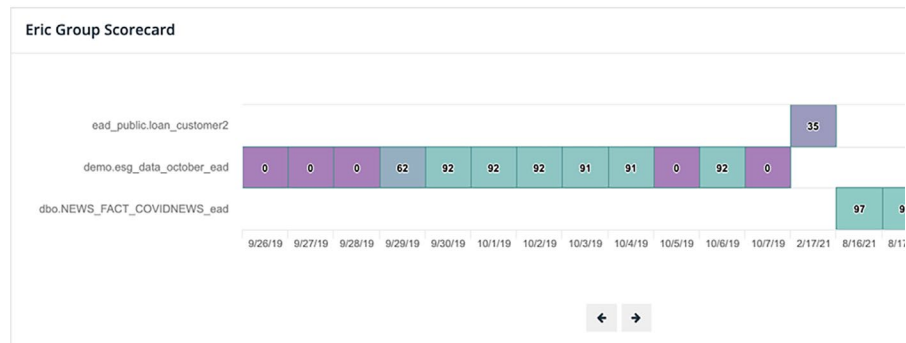
Using Pulse View, a Chief Data Officer can monitor quality across multiple systems
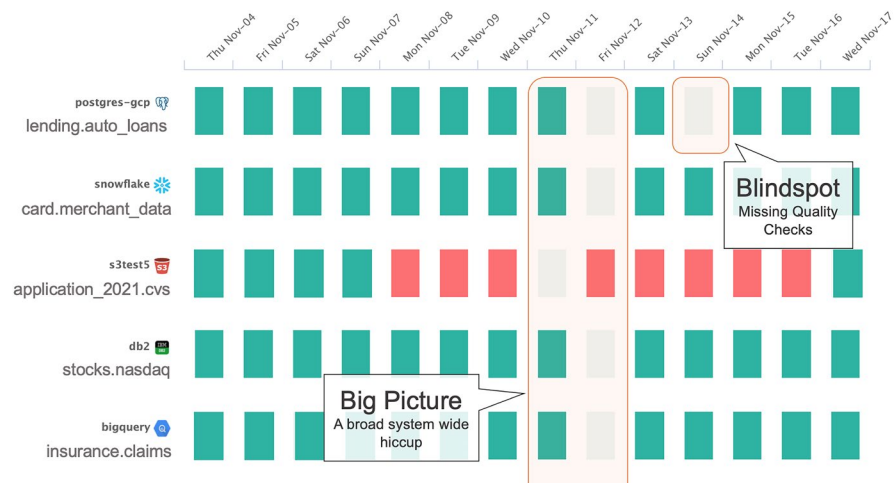
Expand data quality across the enterprise

As organizations invest heavily in advanced analytics, cloud migration, digital experience, CRM/ERP, etc., effective results depend on trusted and understood data. An established data quality function will help reduce the risk and ensure success criteria is met. It is also advantageous to leverage the momentum of building out other capabilities that leverage data quality, such as master data and metadata management or data governance. Data stewards (business and technical) are the glue between all these capabilities, yet each capability is quite distinct, with different supporting roles and technologies.

Another place to embed data quality steps is during project development. Organizations typically follow a software development lifecycle as part of a build-out. What better place to embed the steps of data quality management than in a data-heavy project, such as a CRM or ERP implementation?

Measuring your progress and impact

Up until this point, we have not discussed Collibra Data Governance Center (DGC) and Data Catalog, and where Collibra Data Quality & Observability fits in. It is critical to understand the rich connection between these two products, and what that connection enables. These technologies allow transparency into stewardship activities as well as the ability for stewardship to expand across the enterprise. The work of metadata and data quality management are done in these platforms. Data quality standards and rules are captured and governed in Collibra DGC, which are then executed consistently using Collibra Data Quality & Observability. Data profiling results and scorecards can be shared in the Data Catalog to create trust in the catalog content. Additional modern functionality between Collibra Data Quality & Observability and Collibra is to help identify where sensitive data is being used and then capture this detail in the catalog for better enforcement. Workflows engage the right people at the right time to be proactive in handling issues identified in the data quality process. The simple act of capturing metadata in Collibra will inform quality results; consumers will better know about the data and how best to use it.

Collibra Data Quality & Observability self-service functionality also allows organizations to report key metrics which demonstrate what is going well and what needs additional oversight or governance. User-friendly dashboards are easy to create and provide a view into how the implementation of data capabilities creates tangible results.

Conclusion

Now that you know the critical components of a successful data quality function, create the vision for your organization. First, write the story of your organization's data quality journey. What is its current state? Where do you need to get to and why? Seize the data-focused momentum and make your case for data quality. Operationalize the function with the seven steps to ensure efficiency and consistency across the organization.

**Schedule a demo of Collibra
Data Quality & Observability**

About the author

Sarah Rasmussen

FSFP Collibra Practice Lead and Engagement Partner

Sarah's interest in data began when her IT career evolved from developer to systems analyst and managing and learning from a team of enterprise data architects and modelers. Soon, she was developing strategy and implementation for enterprise information management capabilities. Sarah is a thought leader and featured speaker at industry conferences, such as Enterprise Data World, and is actively involved with many organizations, including Women in IT and Girls Who Code.

About First San Francisco Partners

First San Francisco Partners helps data-driven organizations navigate change to make information actionable. Founded by Kelle O'Neal in 2007, FSFP focuses on implementing sustainable solutions to transform data value into measurable business value.

With an average of 20 years each of data-centered experience FSFP senior consultants know how to shape and put into action highly customized information management, data governance, metadata management, master data management, data architecture and data quality solutions that work for some of the world's most notable companies.

FSFP has been a trusted Collibra Partner since 2012. In 2020, FSFP received Collibra's Honorable Mention Partner of the Year commendation.

For more information about FSFP, visit firstsanfranciscopartners.com or call 1-888-499-DATA (3282).

About Collibra

Since 2008, Collibra has been uniting organizations by delivering trusted data for every use, for every user, and across every source. Our Data Intelligence Cloud brings flexible governance, continuous quality and built-in privacy to all types of data. The Global 2000 relies on Collibra to create the critical alignment that accelerates workflows and delivers better results faster. We have a diverse global footprint, with offices in the U.S., Belgium, Australia, Czech Republic, France, Poland and the U.K. To learn more, visit collibra.com, follow @Collibra on Twitter or follow us on LinkedIn.