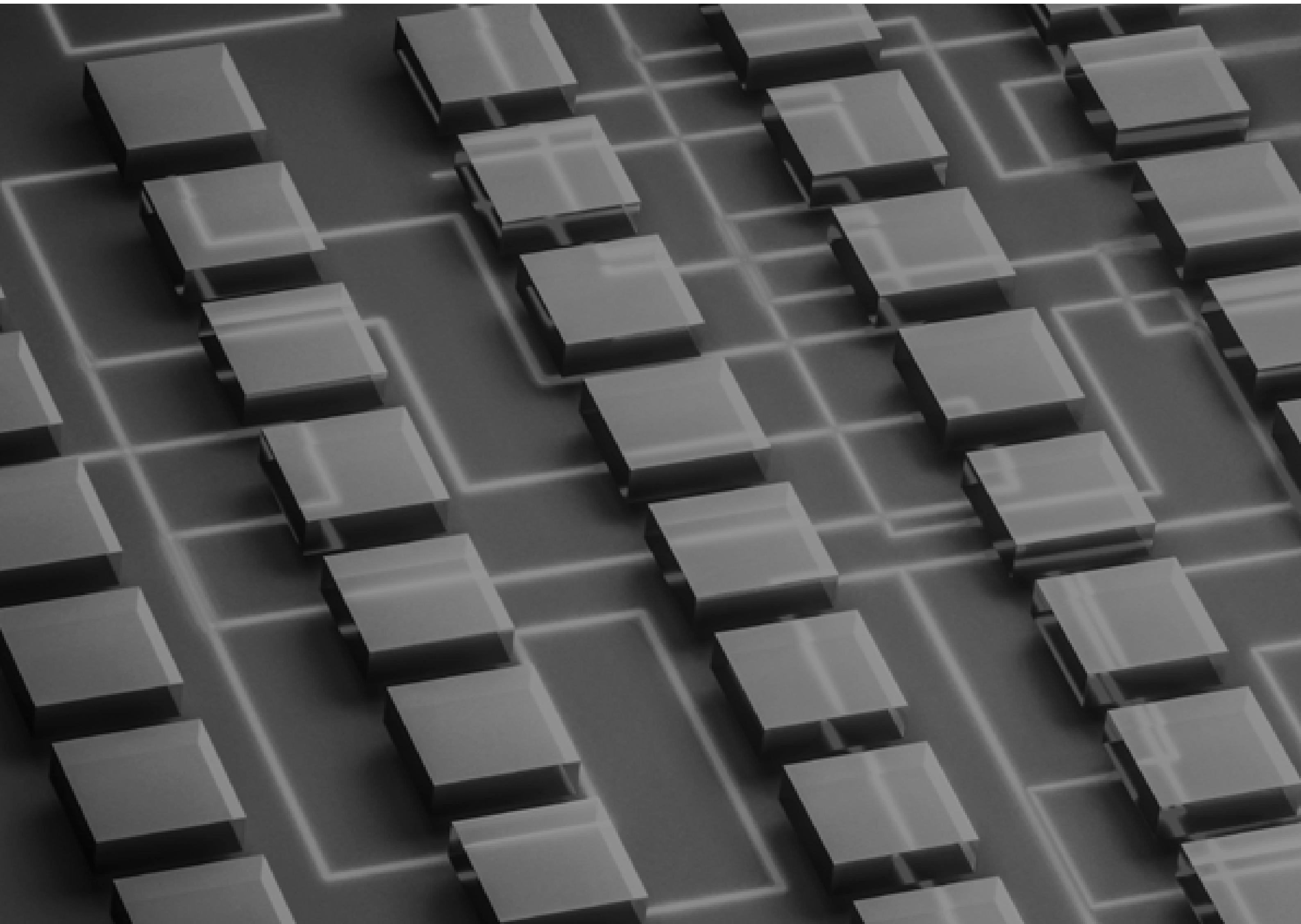
CAPCO

PRO-ACTIVE DATA QUALITY MANAGEMENT: A LOOK THROUGH THE JOHARI WINDOW



SPEED READ:

- Poor data quality can lead to a lack of business agility and sub-standard customer service, which can cause late and inaccurate reporting, in addition to over-allocation of capital and loss allowance provisions. Significant regulatory fines have also been levied in the past year

- Many financial institutions are taking reactive approaches to data quality management, with poor data quality primarily identified through issues, leading to consumer-run compensating controls and adjustments, which can make the architecture more complex over time and add significant operational costs

- This is typically caused by a lack of communication and transparency between data producers and consumers, underpinned by inadequate mechanisms to communicate data requirements and data quality issues. Institutional inertia and complex organizational structures can also compound these challenges

- Applying the Johari Window to the discipline outlines tangible actions to increase information sharing in a structured, coordinated manner, driving pro-active data quality management

- This can result in substantial cost-saving opportunities through streamlined issue remediation and rationalization of data controls, amongst other factors. The open lines of producer-consumer communication can also result in a multitude of other benefits

- Organizations can quickly demonstrate benefits through a series of low-risk and scalable proof-of-concepts, leveraging techniques pioneered in other industries.

THE COST OF REACTIVITY

Data quality management is not a new discipline; it is considered one of the core building blocks of any data management framework. Financial institutions know all too well the need to trust the data driving their organization – and the huge regulatory fines for not being able to demonstrate this trust. The oft-cited maxim 'Garbage in, garbage out' can be dated to the late 1950s. But can organizations now better determine whether their data is garbage?

Although no-longer the gold standard, reactive (or defensive) data quality management is the current foundation from which many organizations maintain a grip on data quality. Reactive data quality management responds to data risk events which have already materialized (i.e. issues) and focuses primarily on data quality issue management (DQIM). This provides a mechanism to raise, track and remediate data quality issues after business impact, typically to data consumers, such as Risk or Finance functions. These impacts can include misreporting, untimely delivery of regulatory reporting and over-allocation of capital and loss allowance provisions. More advanced organizations will accompany DQIM with data quality monitoring (DQM), the systematic monitoring and measurement of the data control environment, to identify breaches against data quality rules. However, a lack of consistency in DQM standards, limited control comprehensiveness and the concentration of controls at consuming systems and processes can limit the effectiveness of this approach. The inherent complexity in financial institutions, such as legacy architecture and vast organizational structures can make this hard to achieve.

It is this identification of poor data quality through either issues (DQIM) or consumer-focused data controls that leads to compensating controls – local consumer-driven cleansing efforts and "sticking plasters", such as extensive journal adjustments or risk calculation adjustments, which can make the architecture more complex over time and add significant operational costs. Research has shown that one-third of analysts spend more than 40 percent of their time vetting and validating their analytics data before it can be used for strategic decision-making[1]. The price of re-activity therefore goes far beyond substantial misreporting and regulatory fines. Research indicates that cleansing data downstream costs 10 times more than verifying and remediating at source[2].

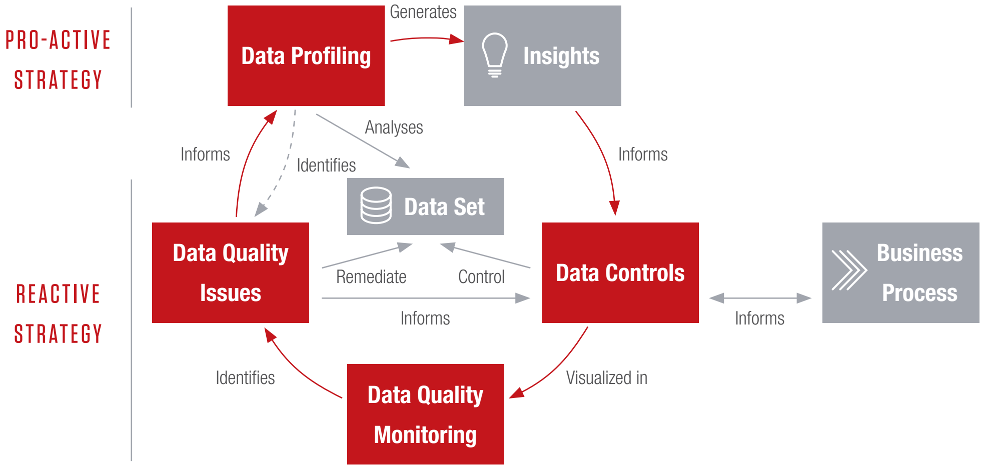


Figure 1 – A typical data quality management framework shown in red, highlighting the difference between reactive and pro-active components

[1] https://www.forrester.com/report/Build+Trusted+Data+With+Data+Quality/-/E-RES83344
[2] https://www.destinationcrm.com/Articles/ReadArticle.aspx?ArticleID=52324

> “*The cost of bad data is an astonishing 15% to 25% of revenue for most companies*[3]”

MIT Sloan Management Review

The pro-active component of the data quality management framework introduces periodic data profiling, the detailed analysis of the structure, content, and relationships in data[4], to both identify anomalies and data quality issues, and define new data controls. This targeted generation of insights provides the ability to challenge and improve the data control environment over time, and reduce bias resulting from data controls defined purely on subject matter expert-defined rules – typically based on previously-identified data quality issues.

True pro-activity, however, goes beyond periodically executing data profiling. Organizations must also understand and re-consider how, where, and by who their data quality is being managed – and bake this into processes and future architecture changes. Achieving this requires the systematic capture of data quality information (‘metadata’) in a sustainable manner, made transparent to both data producers and consumers across the organization.

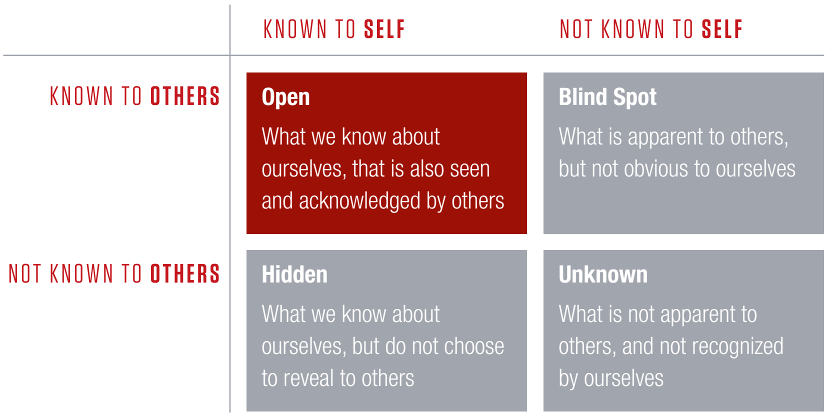
So, what does pro-active data quality management really mean, and how can organizations begin to drive towards this state?

[3] https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/
[4] https://www.capco.com/Intelligence/Capco-Intelligence/Data-Management-Strategy-Launching-The-Journey-Toward-Value-Generation

THROUGH THE JOHARI WINDOW

A well-known psychological tool, the Johari Window, named after its creators Joseph Luft and Harrington Ingham, aims to improve self-awareness and mutual understanding between individuals under the premise that a better understanding of oneself can be acquired by revealing information and receiving feedback.

The subject is represented by the Johari Window where four quadrants separate information, such as motivations and feelings, that are known or unknown to oneself or others. The findings enable an improved knowledge of oneself by broadening the 'Open' quadrant, which can be actioned to make constructive changes.

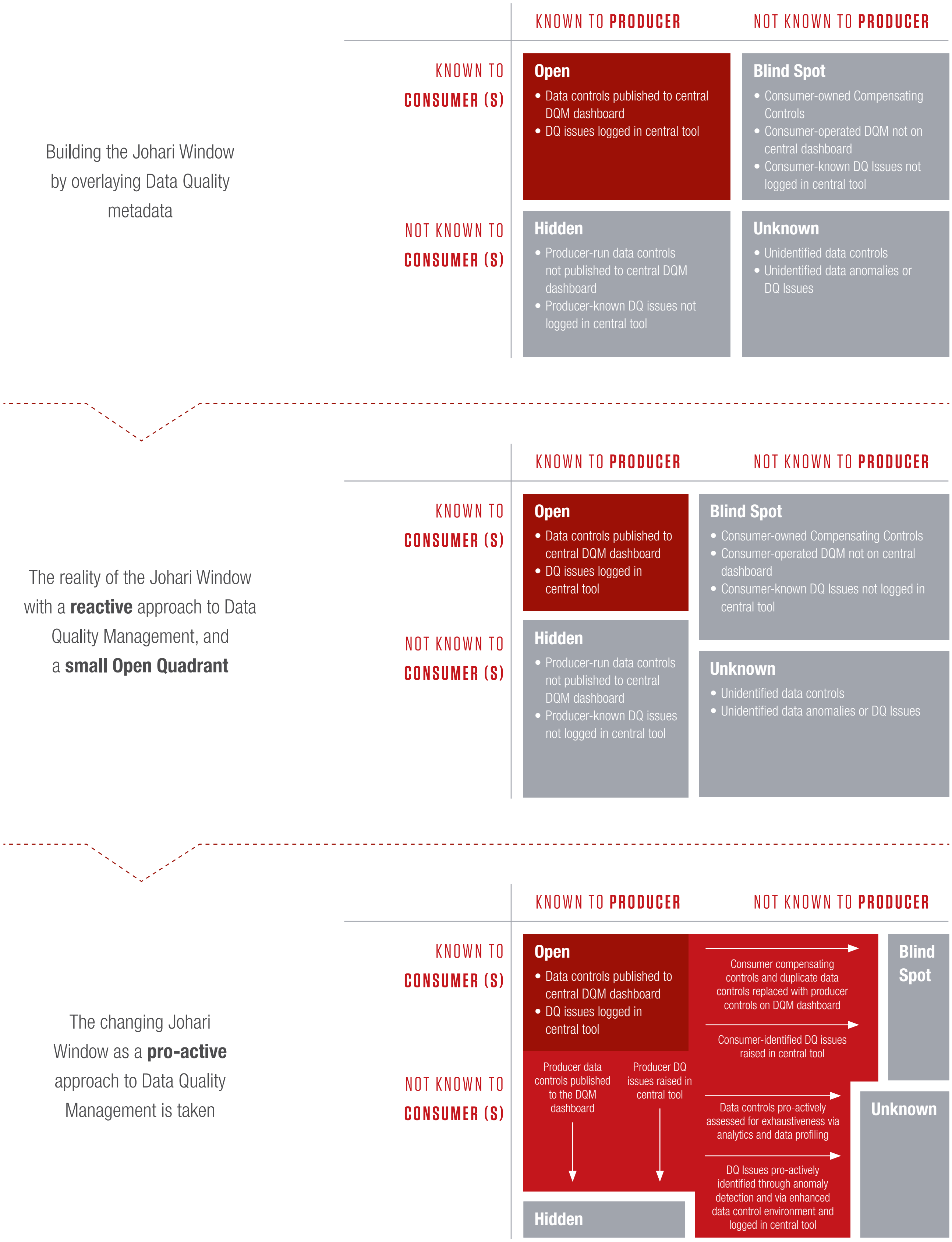


	KNOWN TO **SELF**	NOT KNOWN TO **SELF**
KNOWN TO **OTHERS**	**Open** What we know about ourselves, that is also seen and acknowledged by others	**Blind Spot** What is apparent to others, but not obvious to ourselves
NOT KNOWN TO **OTHERS**	**Hidden** What we know about ourselves, but do not choose to reveal to others	**Unknown** What is not apparent to others, and not recognized by ourselves

Figure 2 - The Johari Window

> *"Improved understanding of oneself by broadening the 'Open' quadrant"*

To best explain how to pro-actively manage data quality, we look at the discipline through the Johari Window, and replace 'Self' and 'Others' with 'Producers' and 'Consumers', representing the producers and consumers of data within an organization. Substituting personal information, motivations, and feelings for metadata on the quality of the producer's data (such as DQ Issues and data controls), aligned to the four quadrants, we start to fill out the Johari Window. From this, the actions required to shrink the Blind Spot, Hidden and Unknown quadrants become evident, reducing what is hidden from view.



Building the Johari Window by overlaying Data Quality metadata

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls published to central DQM dashboard • DQ issues logged in central tool	**Blind Spot** • Consumer-owned Compensating Controls • Consumer-operated DQM not on central dashboard • Consumer-known DQ Issues not logged in central tool
NOT KNOWN TO **CONSUMER (S)**	**Hidden** • Producer-run data controls not published to central DQM dashboard • Producer-known DQ issues not logged in central tool	**Unknown** • Unidentified data controls • Unidentified data anomalies or DQ Issues

The reality of the Johari Window with a **reactive** approach to Data Quality Management, and a **small Open Quadrant**

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls published to central DQM dashboard • DQ issues logged in central tool	**Blind Spot** • Consumer-owned Compensating Controls • Consumer-operated DQM not on central dashboard • Consumer-known DQ Issues not logged in central tool
NOT KNOWN TO **CONSUMER (S)**	**Hidden** • Producer-run data controls not published to central DQM dashboard • Producer-known DQ issues not logged in central tool	**Unknown** • Unidentified data controls • Unidentified data anomalies or DQ Issues

The changing Johari Window as a **pro-active** approach to Data Quality Management is taken

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls published to central DQM dashboard • DQ issues logged in central tool	Consumer compensating controls and duplicate data controls replaced with producer controls on DQM dashboard Consumer-identified DQ issues raised in central tool **Blind Spot**
NOT KNOWN TO **CONSUMER (S)**	Producer data controls published to the DQM dashboard Producer DQ issues raised in central tool **Hidden**	Data controls pro-actively assessed for exhaustiveness via analytics and data profiling DQ Issues pro-actively identified through anomaly detection and via enhanced data control environment and logged in central tool **Unknown**

Figure 3 - Applying the Johari Window to Data Quality

In taking these actions, the Open quadrant grows, placing the onus on data producers for control definition, measurement, and remediation. The producer is best placed to identify and remediate poor data quality at source before an impact to business operations is made. This approach also ensures that all consumers of the data benefit from increased data quality, rather than the self-remediating consumers in isolation. The characteristic of producer-driven data quality management (and a large Open quadrant) is the guiding principle for pro-active data quality management.

OPENING THE JOHARI WINDOW

To understand how to progress from a reactive to pro-active Data Quality Management approach, it is important to truly understand what each quadrant represents, and the reasons why the Blind Spot, Hidden and Unknown quadrants typically dominate. This enables the definition of targeted action plans to shrink each of these quadrants, holistically broadening the Open quadrant.

Although every organisation is unique, we have typically seen many of the same challenges across the Financial Services domain. We have summarised some of the key challenges and ways to overcome these, including how we can learn from innovations in other industries to drive improvements.

OPEN – KNOWN TO BOTH PRODUCERS AND CONSUMERS

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls published to central DQM dashboard • DQ issues logged in central tool	**Blind Spot** • Consumer-owned Compensating Controls • Consumer-operated DQM not on central dashboard • Consumer-known DQ Issues not logged in central tool
NOT KNOWN TO **CONSUMER (S)**	**Hidden** • Producer-run data controls not published to central DQM dashboard • Producer-known DQ issues not logged in central tool	**Unknown** • Unidentified data controls • Unidentified data anomalies or DQ Issues

WHAT DOES IT REPRESENT?

The nirvana state – data quality knowledge which is transparent across the organization, structured, easily accessible and well understood.

WHAT ARE COMMON EXAMPLES?

A bank's client reference data team performing daily monitoring on the completeness of client data and publishing results to the central DQM dashboard. The team pro-actively raise data quality issues in the central DQIM tool when identified, auto-notifying impacted consumers. Exceptions (clients with incomplete reference data) are flagged when propagated to downstream applications.

WHY IS IT IMPORTANT?

The Open quadrant is the most powerful place for knowledge to exist. Common knowledge encourages action and enables effective identification and remediation of root-causes. The Open quadrant is the key driver of pro-active DQ, where producers can rapidly mobilize remediation efforts where required, and consumers are aware of data weaknesses at the earliest opportunity.

WHAT DOES IT ENABLE?

Intelligent Issue Root Cause Analysis: Through integration of data lineage, data controls and data quality issue management, a "map" (modelled as a Knowledge Graph) of data quality issues across the organization can be built. By leveraging novel techniques, such as graph inference, root causes of issues across the application landscape can be intelligently identified, identifying factors such as control weaknesses or gaps. Such techniques are commonly applied in the manufacturing industry leveraging 'digital twins', forming a central facet of the "Age of Industry 4.0"[5].The clear value gained from such techniques doesn't leave us surprised to see Gartner claim that by 2021, half of large industrial companies will use digital twins, resulting in those organizations gaining a 10% improvement in effectiveness[6]. Their increased adoption in financial services should also not come as a surprise given their transformational power[7].

Improved Remediation Efficiency: With data quality issues logged by both producers and consumers on a unified platform, common themes across disparate issues can be identified and issues can be clustered accordingly. This can be performed using machine learning and natural language processing techniques, which even enable insights to be derived from free text fields. Such abilities are particularly powerful when we consider that five different consumers may identify the same one issue as five different symptoms, raising five different issues. Through smart grouping, one producer would instead see one issue impacting five different consumers, enabling rapid and powerful business case development and remediation prioritization. This ability to identify macro trends maximizes return on investment for a single remediation activity, resulting in significant cost savings compared to five local tactical fixes.

Control De-duplication and Rationalization: By overlaying data controls on data lineage, inference techniques can also be used to identify opportunities to remove duplicate data controls from the application landscape, such as those which are operated in multiple business divisions. These same techniques can identify compensating controls which can be removed and replaced with strategic preventative data controls at source. This increased control coverage at source means that multiple consumers can benefit from the same control, providing significant cost saving opportunities.

HOW CAN IT BE INCREASED?

In addition to the actions outlined to reduce the other three quadrants, organizations should actively undertake the following activities:

Improve Data Literacy: Key to recognizing and actioning poor data quality is a base level of data literacy across the organization, with employees understanding data quality and its impact on the organization. A structured and concerted effort to design and provide training and education should be driven from the top-down, supported by data management SMEs in each business division.

Promote the Possibilities: It can be challenging to drive the adoption of data quality management above simple data quality issue management – often ignorance is bliss until something goes wrong. Chief Data Officers or similar roles should define and communicate a strong vision, articulating the business benefits and improvements to employee satisfaction enabled through pro-active data quality management. Understanding and tracking the benefits associated with the Open quadrant will drive adoption of best practices and guidelines and drive a virtuous cycle of data improvements.

[5] https://medium.com/datadriveninvestor/root-cause-analysis-in-the-age-of-industry-4-0-9516af5fb1d0
[6] https://www.gartner.com/smarterwithgartner/prepare-for-the-impact-of-digital-twins/
[7] https://www.capco.com/intelligence/capco-intelligence/knowledge-graphs-building-smarter-financial-services

BLIND SPOT – NOT KNOWN TO PRODUCERS, KNOWN TO CONSUMERS

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls published to central DQM dashboard • DQ issues logged in central tool	**Blind Spot** • Consumer-owned Compensating Controls • Consumer-operated DQM not on central dashboard • Consumer-known DQ Issues not logged in central tool
NOT KNOWN TO **CONSUMER (S)**	**Hidden** • Producer-run data controls not published to central DQM dashboard • Producer-known DQ issues not logged in central tool	**Unknown** • Unidentified data controls • Unidentified data anomalies or DQ Issues

WHAT DOES IT REPRESENT?

Perceived data quality weaknesses known by (multiple) consumers, but not transparent to producers. A common example is data quality issues that are identified but not communicated to upstream producers or logged in central data quality issue management repositories. Additionally, compensating controls (such as adjustments or cleansing routines, including imputation of nulls with default values) are a common example, particularly in Risk or Finance functions. Local monitoring of inbound data for incomplete or incorrect values is also common, particularly by teams producing regulatory reports.

WHAT ARE COMMON EXAMPLES?

A bank's Finance function repeatedly posting adjustments to a financial instrument's market price due to inaccuracies, without informing the instrument reference data team.

WHAT CAN IT CAUSE?

Proliferation of Compensating Controls: The concentration of data quality issue identification responsibility at the consumer can place significant strain on consumers to communicate, manage and resolve issues which result from upstream causes. This results in the formation of large teams operating costly compensating controls and local tactical "sticking plaster" fixes, repeated reporting period after reporting period. This harmful driver of inefficiency is exacerbated when considering how similar compensating controls will often be executed by other consumers of the same data, who may be blind to problems which are only visible to one consumer group. Furthermore, the local remediation only benefits the single consumer, so remediation efforts are duplicated across multiple consumers, acting as a significant cost multiplier.

Limited Ability to Perform Root Cause Analysis: These tactical controls prevent the ability to perform strategic root cause analysis and remediation, as root causes cannot be 'pin-pointed' based on data quality monitoring upstream. This ultimately means that the lack of an ability to strategically remediate issues is institutionalized.

WHY DOES IT PERSIST?

Institutional Inertia: Long-standing organizational culture, past failures and inherent complexity can act as change barriers. The structures put in place purely to service the Blind Spot have resulted in cost, time, and effort – and it can be hard for organizations to move forward from what should have been a temporary measure.

Insufficient Issue Communication Mechanisms: Mechanisms to communicate issues to producers are often non-existent, insufficient, or overly complex. Furthermore, consumers may not even know where their data is consumed from, and who to communicate issues to.

Unclear and Changing Requirements: Often, the Blind Spot persists where consumer requirements (such as data coverage) are either not effectively communicated to or implemented by producers, resulting in data which may not be fit-for-purpose. This is compounded by continuously changing requirements to meet business or regulatory needs, which can mean producers and consumers become further misaligned on requirements.

Concentration of Regulatory Monitoring: The pressure for quality data is generally felt more by consumers than producers, who are often required to demonstrate compliance and conformity to regulatory instructions within tight reporting timelines. These can often be delivered with ease via data quality rules embedded in common regulatory reporting engines. Although such engines can provide better control coverage, they can also compound the concentration of monitoring and remediation in such consuming functions downstream of the data's Master or Authoritative Source, due to the need to submit sufficient data within tight timelines.

HOW CAN IT BE REDUCED?

Documentation of Data Lineage: Key to reducing blind spots is the ability for consumers to understand the flow of data into their applications, and the relevant producers of this data. As such, easily accessible data lineage is advantageous to consumers to enable targeted producer engagement.

Documenting Data Controls and Monitoring: Documenting and overlaying data controls on the data lineage provides transparency to producers on where their data is controlled by downstream consumers, including by manual compensating controls and adjustments. This ability to share control information with upstream producers in a structured and transparent manner enables producers to design and implement controls at source which can replace controls operated by (multiple) consumers. These controls should be monitored, with metrics and exceptions made transparent to downstream consumers via an organization-wide data quality monitoring dashboard.

Consumer Requirements Elicitation: Mechanisms should also be implemented to allow consumers to document and communicate data quality requirements to producers in a structured, consistent manner, in-line with data control standards outlining standard DQ dimension and DQ rule taxonomies. The emerging trend of augmented data quality, which refers to the application of AI and ML across DQ products, has been recognized by Gartner as a high business benefit enabler two to five years away from mainstream adoption[8]. Augmented data quality can support structured requirements definition in alignment with standards through AI-driven analysis of existing control documentation.

Consistent Data Quality Issue Management: A consistent organization-wide approach to data quality issue management should be employed, with a single, easy-to-use DQIM tool adopted. When linked to data lineage and the data glossary, such a tool and process enables transparency of DQ issues to producers and group data management functions, where in place, enabling prioritization and remediation at source. This 'big picture' outlook enables strategic architecture decisions to be made more easily, rather than fragmented approaches based on disparate, disconnected issue management methods.

[8] https://www.gartner.com/doc/reprints?id=1-1ZNVTRHP&ct=200812&st=sb

HIDDEN – KNOWN TO PRODUCERS, NOT KNOWN TO CONSUMERS

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls published to central DQM dashboard • DQ issues logged in central tool	**Blind Spot** • Consumer-owned Compensating Controls • Consumer-operated DQM not on central dashboard • Consumer-known DQ Issues not logged in central tool
NOT KNOWN TO **CONSUMER (S)**	**Hidden** • Producer-run data controls not published to central DQM dashboard • Producer-known DQ issues not logged in central tool	**Unknown** • Unidentified data controls • Unidentified data anomalies or DQ Issues

WHAT DOES IT REPRESENT?

Data quality weaknesses which are known by data producers, but not communicated to downstream consumers. This can include the metrics and exceptions identified by producer-run data controls, or other known issues which consumers do not have visibility of. A large Hidden quadrant is the most problematic quadrant of the Johari Window, as consumers may not have compensating controls in place to rectify issues, resulting in significant impacts on business outcomes borne out of a lack of transparency and communication.

WHAT ARE COMMON EXAMPLES?

A client reference data team not communicating to consumers that address information is only partially complete for Canadian clients, due to a fault in the client onboarding process in that locality.

WHAT CAN IT CAUSE?

Poor Visibility of Consumer Impact and Misguided Remediation Prioritization: Impact, such as operational costs or potential fines due to misreporting, cannot be determined without providing transparency to consumers of data quality weaknesses. As such, producers are left with limited information on which to make holistic prioritization calls for remediation.

Audit and Regulatory Scrutiny: Firms in-scope for BCBS-239, CCAR and GDPR regulations know all too well the audit and regulatory scrutiny around data management. Data producers will be expected by audit and regulators to pro-actively inform consumers of potential issues impacting their business processes or reporting.

WHY DOES IT PERSIST?

Poor Visibility of Consumer Impact: Not just a byproduct of the Hidden quadrant, but also a driver, which rarely persists because of an intent to hide the problem. Without a clear view of all potentially impacted consumers, it can be challenging for producers to identify relevant stakeholders who will be impacted by their poor data quality. As such, key stakeholders may be missed off communications. Furthermore, producers may be inclined to only make the most severe issues transparent to reduce operational overheads and 'noise'. Such approaches can however be problematic, as seemingly low volume issues may have a significant impact depending on the usage of the data by consumers. Furthermore, the impact of several smaller issues in aggregate may be severe to consumers.

Limited Tooling: A lack of consistent tooling to log and inform consumers of data quality issues, or publish data quality monitoring metrics, can leave producers unable to effectively communicate poor data quality.

HOW CAN IT BE REDUCED?

Transparency of Data Quality Monitoring: By publishing data control metrics to a central DQM dashboard, or implementing monitoring of data controls where not already existing, producers can pro-actively engage consumers and highlight any potential data weaknesses, enabling consumers to determine if they are impacted by such weaknesses.

Pro-Active Issue Raising: As with the Blind Spot, the pro-active raising of issues in a central tool by producers is key to engaging consumers. When linked to data lineage, the tool will enable pro-active notifications to impacted consumers, based on the documented data flows. This removes the onus on producers to manually understand and engage all data consumers, and enables the knowledge captured in data lineage to do the hard work.

Data Contracts: Establishing formal data contracts between data producers and data consumers clearly defines and sets accountability for data quality rules and other data delivery expectations. Data contracts provide transparency to any rule breaches and outline the producer's response process when breaches are identified.

UNKNOWN – NOT KNOWN TO PRODUCERS OR CONSUMERS

	KNOWN TO **PRODUCER**	NOT KNOWN TO **PRODUCER**
KNOWN TO **CONSUMER (S)**	**Open** • Data controls onboarded to central DQM dashboard • DQ issues logged in central tool	**Blind Spot** • Consumer-owned Compensating Controls • Consumer-operated DQM not on central dashboard • Consumer-known DQ Issues not logged in central tool
NOT KNOWN TO **CONSUMER (S)**	**Hidden** • Producer-run Data controls not onboarded to central DQM dashboard • Producer-known DQ issues not logged in central tool	**Unknown** • Unidentified Data Controls • Unidentified data anomalies or DQ Issues

WHAT DOES IT REPRESENT?

Neither producers nor consumers are aware of these problems. What you can't see can't hurt you, right? The Unknown quadrant is the hardest to conceptualize, but by no means the hardest to action. This quadrant covers inadequate data control environments and anomalous behavior representing issues waiting to happen. The advent of advanced, easy-to-use tooling with embedded Artificial Intelligence has the potential to rapidly decrease the Unknown quadrant.

WHAT ARE COMMON EXAMPLES?

Large statistical fluctuations in foreign exchange trade amounts, due to the incorrect currency being selected for the trade (E.g. Japanese Yen instead of US Dollar), caused by a lack of warning / control mechanism in the trade capture system.

WHAT CAN IT CAUSE?

False Confidence: Statements like "Our controls will pick it up" or "Consumers will tell us when they have problems" are often cited by data owners creating a sense of false confidence. Vulnerabilities resulting from a limited control environment can quickly become apparent when assessing data controls against standards, guidelines, and best-practices.

Reactive Issue Identification: A lack of pro-activity in assessing the effectiveness and exhaustiveness of the data control environment and detecting data anomalies can result in data quality issues only being identified after business impact.

WHY DOES IT PERSIST?

Lack of Defined Data Control Standards: Data control standards, guidelines and best-practices provide a benchmark of “what good looks like”. Without this clearly defined yardstick, organizations may be unknowingly operating without a fit-for-purpose data control environment.

Data Quality Rule Bias: Typically, data quality rules are biased to SMEs’ experiences, or knowledge of where things have previously gone wrong. This bias can result in a reliance on overly simple data controls, such as simple validity rule checks (e.g. conformance to a valid value set). Reliance on such data controls can prevent the identification of certain kinds of data quality issues, making the Unknown further out of reach.

Under-Utilized Tooling: Most modern data quality platforms have in-built data profiling capabilities, enabling the identification of anomalies and definition of new data controls. IT teams and database engineers typically use the most basic functionality, such as data type and field length validation, at the initial point of development, but often do not leverage their full potential with input from business users.

Limited Understanding of Emerging Risks: IT, people or process changes can cause risk due to control or process disruption. Furthermore, systems integration issues can result in new data quality issues which may go unnoticed. Without proper change impact analysis, emerging risks may not be identified and mitigated appropriately.

Issue Swamps: Most organizations are already overburdened with data quality issues, often described unclearly and with many duplicates. These Issue Swamps make life difficult for the limited pool of SMEs who understand how to analyze and remediate the issues. As such, without first clustering, rationalizing and prioritizing remediation of known issues in an intelligent way, the appetite to identify further issues can be lower. This means that significant risks may not be identified pro-actively and can result in even greater issues toppling the pile.

HOW CAN IT BE REDUCED?

Data Control Standards, Guidelines and Assessments: Data control standards, guidelines and generic rule libraries should be defined, codified, and rolled out to provide an objective assessment of the data control environment. Integration with the operational risk framework is also key to ensuring appropriate scrutiny as part of standard control effectiveness testing. Codified rule libraries also drive the standardization of data control definitions across multiple business divisions, enabling a consistent language and understanding across the organization.

Robust Change Impact Analysis: Pro-actively understanding the impact of change on systems, processes and people is key to mitigating risk associated with such change. Data lineage is a key enabler for such analysis and can quickly articulate downstream and upstream impacts of changes. Organizations embarking on a digital twin journey are also well-positioned to leverage their investment and perform change impact analysis at speed and scale.

Power Profiling and Anomaly Detection: Many best-in-class data profiling and data discovery tools include AI capabilities. Such capabilities can intelligently run rules based on the context of data, and auto-identify relationships in disparate datasets. This functionality can quickly identify potential data quality issues and propose new data controls at scale, with minimal SME input. We have also seen increased adoption of AI-driven anomaly detection algorithms for data quality, such as those used in fraud detection. Outside of the financial services industry, we have seen significant benefits gained by companies like Uber[9],and even as part of particle physics experiments at CERN[10].Whilst these techniques are not yet at the stage of widespread adoption, early adopters will reap benefits by carefully selecting anomaly detection use cases and pilots (such as on Risk, Finance or Payments data) to demonstrate value.

[9] https://eng.uber.com/monitoring-data-quality-at-scale/
[10] https://indico.cern.ch/event/635481/contributions/2685066/attachments/1507350/2350634/CERN_PRESENTATION_Autosaved.pdf

THE JOURNEY AHEAD: PRACTICAL NEXT STEPS

Reaching this state is not without its challenges. Legacy architecture, severed lines of communication and a limited data culture makes the journey particularly challenging. There is, however, significant value to be gained – and we recommend five key principles for organizations embarking on the evolution from reactive to pro-active data quality management:

1. **Set the Tone from the Top:** Define and communicate a strong vision articulating the benefits of pro-active data quality management, driven from the top down.

2. **Leverage Regulatory Delivery:** Utilize, enhance, and consolidate regulatory data management artefacts[11],such as those produced to meet BCBS-239, CCAR or GDPR requirements, to accelerate the adoption of pro-active data quality management capabilities. Start small, prove value on key data, and scale to a broader data scope.

3. **Align with Wider Data Management Capabilities and Standards:** Integrate data quality management with other data management capabilities, such as data lineage and data governance. Define a consistent set of standards and language to describe data quality.

4. **Unite Producers and Consumers:** Open communication channels between producers and consumers and make their lives easier through adopting a single data quality issue management platform and central data quality monitoring dashboard, enabling the opening of the 'Open' Quadrant.

5. **Adopt Intelligent Tooling:** Understand and leverage the full capabilities of existing data quality toolsets to reap their full benefits and consider piloting and adopting modern tooling with in-built AI. Explore the use of digital twins as a key enabler to answer more complex questions, accelerating root cause analysis.

To quickly get started, organizations of any size can perform low-risk and scalable proof-of-concepts and recognize their benefits:

1. **Data Quality Issue Clustering:** Perform some simple text mining and clustering of data quality issues to identify common themes and related issues and present the results to key data stakeholders from both producing and consuming functions. Where possible, aggregate their quantitative impact to aid prioritization. Demonstrate how seemingly disparate data duality issues have similar characteristics and potential root causes. Where data quality issues are not logged in a central tool or widely known, perform a data discovery exercise against operational risk issue repositories or similar to tag data quality issues before performing clustering.

2. **Digital Twin:** Produce a simple knowledge graph showing data lineage across key applications and overlay data quality issues and data controls, in-line with appropriate knowledge graph standards. Use this to identify potential root causes for known data quality issues and show other downstream impacts of these issues.

3. **Power Profiling:** Perform a short proof-of-concept with business and IT stakeholders to profile a key dataset. Utilize existing tooling where available or identify appropriate vendor or open source toolsets. Capture and present findings to key stakeholders and validate anomalies and potential data quality issues, demonstrating how these could be missed via traditional rule-based techniques.

Look out for more upcoming insights on how organizations can be transformed via adoption of pro-active data quality management.

Capco has extensive experience of designing, mobilizing and operationalizing sustainable enterprise-wide data quality management capabilities across global financial institutions. Speak to us today to find out how we can maximize your data quality return on investment.

[11] https://www.capco.com/Intelligence/Capco-Intelligence/Data-Management-Strategy-Launching-The-Journey-Toward-Value-Generation

REFERENCES

1. https://www.forrester.com/report/Build+Trusted+Data+With+Data+Quality/-/E-RES83344

2. https://www.destinationcrm.com/Articles/ReadArticle.aspx?ArticleID=52324

3. https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

4. https://www.capco.com/Intelligence/Capco-Intelligence/Data-Management-Strategy-Launching-The-Journey-Toward-Value-Generation

5. https://medium.com/datadriveninvestor/root-cause-analysis-in-the-age-of-industry-4-0-9516af5fb1d0

6. https://www.gartner.com/smarterwithgartner/prepare-for-the-impact-of-digital-twins/

7. https://www.capco.com/intelligence/capco-intelligence/knowledge-graphs-building-smarter-financial-services

8. https://www.gartner.com/doc/reprints?id=1-1ZNVTRHP&ct=200812&st=sb

9. https://eng.uber.com/monitoring-data-quality-at-scale/

10. https://indico.cern.ch/event/635481/contributions/2685066/attachments/1507350/2350634/CERN_PRESENTATION_Autosaved.pdf

11. https://www.capco.com/Intelligence/Capco-Intelligence/Data-Management-Strategy-Launching-The-Journey-Toward-Value-Generation

AUTHORS

Joseph Forooghian
Principal Consultant & UK Data Quality Lead

Teddy Lee
Associate Consultant

CONTACT

Chris Probert
Partner & Head of UK Data Practice
chris.probert@capco.com

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

To learn more, visit our web site at www.capco.com, or follow us on Twitter, Facebook, YouTube, LinkedIn and Instagram.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Gurgaon
Hong Kong
Kuala Lumpur
Mumbai
Pune
Singapore

EUROPE
Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Munich
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM