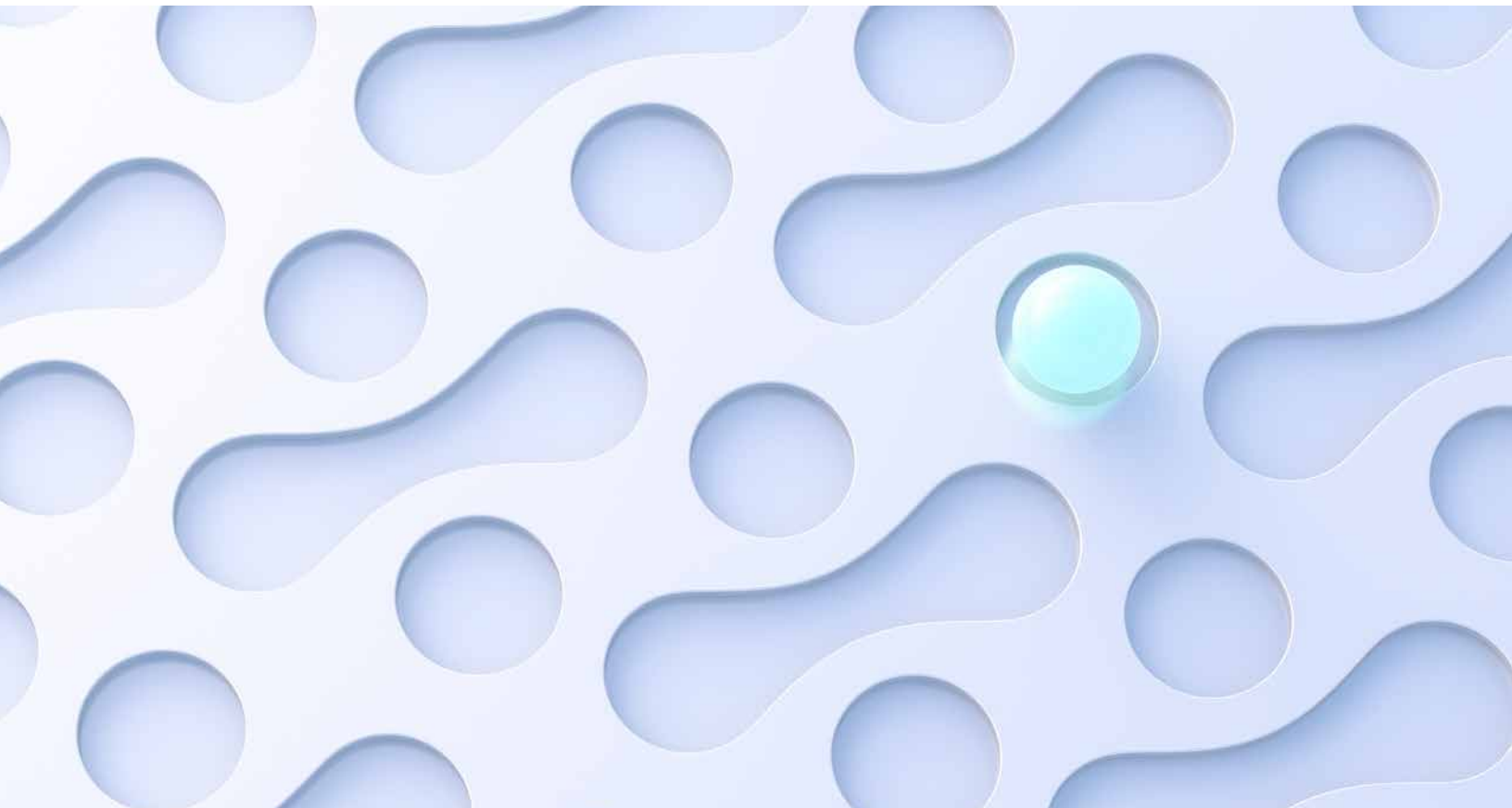


Risk Practice

Optimizing data controls in banking

Banks need to do more in four important areas of data culture to build the risk-related data-control capabilities they will need in the coming decade.

by Tony Ho, Jorge Machado, Satya Parekh, Kayvaun Rowshankish, and John Walsh



Over the past decade, banks across the globe have made considerable progress in building risk-related data-control capabilities, prompted in large part by regulatory demands. The starting point was the Basel Committee's BCBS 239 principles, issued in 2013 to strengthen banks' risk-related data-aggregation and reporting capabilities. Progress, however, has not been uniform, and most institutions are not fully compliant. In fact, many banks are still struggling with major deficiencies, particularly when it comes to data architecture and technology.

One major reason for this limited progress is that the Basel Committee called for effective implementation of BCBS 239 principles without clearly explaining what that means or how to implement them. This ambiguity has led to a wide range of interpretations, which vary from institution to institution, country to country, and even regulator to regulator. At the same time, a host of other regulations with substantial data implications have emerged, particularly those involving stress testing (CCAR in the United States), data privacy (CCPA in the US, GDPR in Europe), BSA/AML, and CECL.¹ As might be expected, banks have a monumental task in analyzing the layers of data requirements across all these regulations and building common and reusable capabilities that meet regulatory expectations.

In response, the industry has adopted some common, workable solutions in a few key areas. These include data-aggregation capabilities to support regulatory reporting requirements, such as automating some of the reporting required by the Federal Reserve in the US and the European Banking Authority (EBA) in Europe,² preparing to collect evidence for regulatory examinations, and deploying a federated data operating model

with central capabilities under a chief data officer. Industry leaders are clear, however, that they struggle in four areas: the scope of data programs, data lineage, data quality, and transaction testing.³

There is considerable variation within the industry on how to address these four challenging areas, in investment, degree of risk mitigation, sustainability, and automation. A few institutions, however, are leading the way in improving their data programs and management and have made great strides toward regulatory compliance.

Scope of data programs

Banks need to define the scope of their data programs clearly enough to create a basis for easily conversing with regulators and identifying additional actions necessary for regulatory compliance. Most banks have defined the scope of their data programs to include pertinent reports, the metrics used in them, and their corresponding input-data elements. Thus a credit-risk report or a report on strategic decision making might be covered, as well as risk-weighted assets as a metric and the principal loan amounts as an input. Unfortunately, the industry has no set rules for how broadly or narrowly to define the scope of a data program or what standard metrics or data elements to include.

As a result, many banks are trying to identify industry best practices for the number of reports and types of data to include in their data programs. Our industry benchmarking indicates that the average bank's data program includes 50 reports, 90 metrics, and 1,100 data elements. Interestingly, over time, we have seen the number of reports in data programs increase while the number of metrics and data elements decreased (Exhibit 1). We believe

¹ BSA/AML refers to the US Bank Secrecy Act (anti-money laundering law) of 1970; CECL is the Current Expected Credit Losses standard issued by the US Financial Accounting Standards Board in 2016; GDPR is the EU's General Data Protection Regulation, which came into force in 2018; CCPA is the California Consumer Privacy Act of 2018; and CCAR is a regulatory framework for comprehensive capital analysis and review introduced by the US Federal Reserve in 2011.

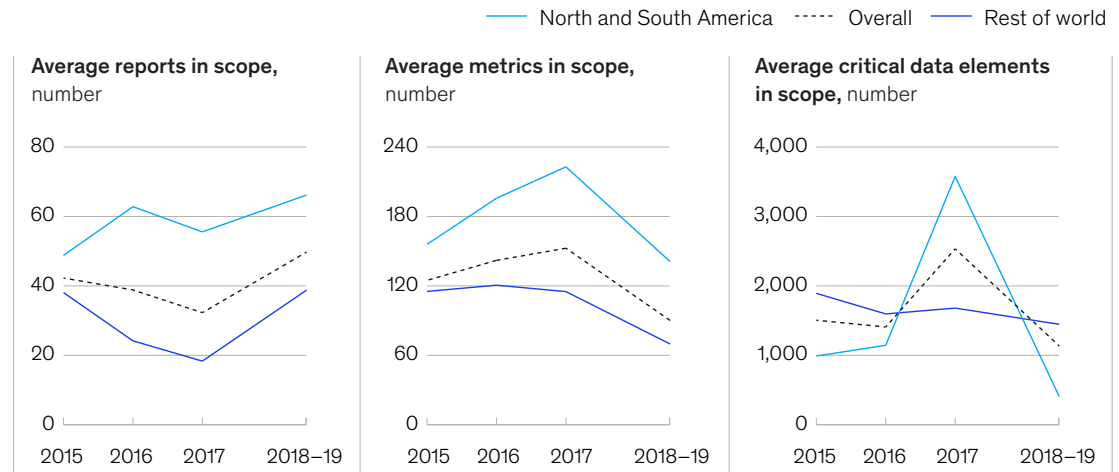
² For example, Federal Reserve form FR Y-14M reports monthly data on the loan portfolios of bank holding companies, savings and loan holding companies, and intermediate holding companies; FR Y-14Q reports quarterly data for the same kinds of institutions on various asset classes, capital components, and categories of preprovision net revenue. The EBA issued the Common Reporting (COREP) framework as the standard for capital-requirements reporting; the EBA's standard for financial reporting is known as FINREP.

³ McKinsey benchmarking survey on data programs with 60 banks, 2020.

Exhibit 1

The scope of banks' data programs has varied considerably over time.

Reports, metrics, and data elements in scope



the increase in reports reflects the inclusion of different nonfinancial risk types, such as operational or compliance risk. The reduction in metrics and data elements is the result of banks' attempts to reduce management costs and efforts and focus only on the most critical metrics and data.

More important than the number of reports, metrics, and data elements is a bank's ability to demonstrate to regulators and other stakeholders that the scope of its data program covers the major risks it faces. With this in mind, leading banks have established principles to define the scope and demonstrate its suitability to regulators. Leading institutions usually define the scope of their data programs broadly (Exhibit 2).

For all banks, the application of the principles illustrated in Exhibit 2 ranges from narrow to broad. However, supervisors are increasingly advocating for a broader scope, and many banks are complying. Best-in-class institutions periodically expand the scope of their data programs as their needs shift. From purely meeting regulatory objectives, these banks seek to meet business objectives as well. After all, the same data support

business decisions and client interactions as well as regulatory processes.

Data lineage




Of all data-management capabilities in banking, data lineage often generates the most debate. Data-lineage documents how data flow throughout the organization—from the point of capture or origination to consumption by an end user or application, often including the transformations performed along the way. Little guidance has been provided on how far upstream banks should go when providing documentation, nor how detailed the documentation should be for each "hop" or step in the data flow. As a result of the lack of regulatory clarity, banks have taken almost every feasible approach to data-lineage documentation.

In some organizations, data-lineage standards are overengineered, making them costly and time consuming to document and maintain. For instance, one global bank spent about \$100 million in just a few months to document the data lineage for a handful of models. But increasingly, overspending is more the exception than the rule. Most banks

Exhibit 2

Best-in-class institutions usually define the scope of their data programs broadly.

Dimensions and characterizations in data program scope

Dimensions	Characterization		
	 Narrow	 Targeted	 Broad
Risk types	<ul style="list-style-type: none"> • Main risk types: market, credit, operational 	<ul style="list-style-type: none"> • All quantitative risk types: market, credit, operational, liquidity 	<ul style="list-style-type: none"> • All risk types: market, credit, operational, liquidity, compliance, audit, reputational, strategic
Legal-entity data sources ¹	<ul style="list-style-type: none"> • Material legal entities 	<ul style="list-style-type: none"> • Material legal entities 	<ul style="list-style-type: none"> • All legal entities
Legal-entity alignment ²	<ul style="list-style-type: none"> • Bank holding company 	<ul style="list-style-type: none"> • Bank holding company and material legal entities 	<ul style="list-style-type: none"> • Bank holding company and all legal entities
Functions and business units	<ul style="list-style-type: none"> • Risk 	<ul style="list-style-type: none"> • Risk and all material business units (but not support functions) 	<ul style="list-style-type: none"> • Risk + all business units + all support functions
Audience for reports	<ul style="list-style-type: none"> • Board and/or senior management 	<ul style="list-style-type: none"> • Board and senior management, heads of business units and functions, regulators 	<ul style="list-style-type: none"> • Board, senior management, heads of business units and functions, managers of functions and businesses, regulators

¹Refers to legal entities whose data feeds into reports within the scope of BCBS 239.

²Refers to legal entities that are independently aligned with BCBS 239 principles including controls, governance, and reporting.

are working hard to extract some business value from data lineage; for example, by using it as a basis to simplify their data architecture or to spot unauthorized data-access points, or even to identify inconsistencies among data in different reports.

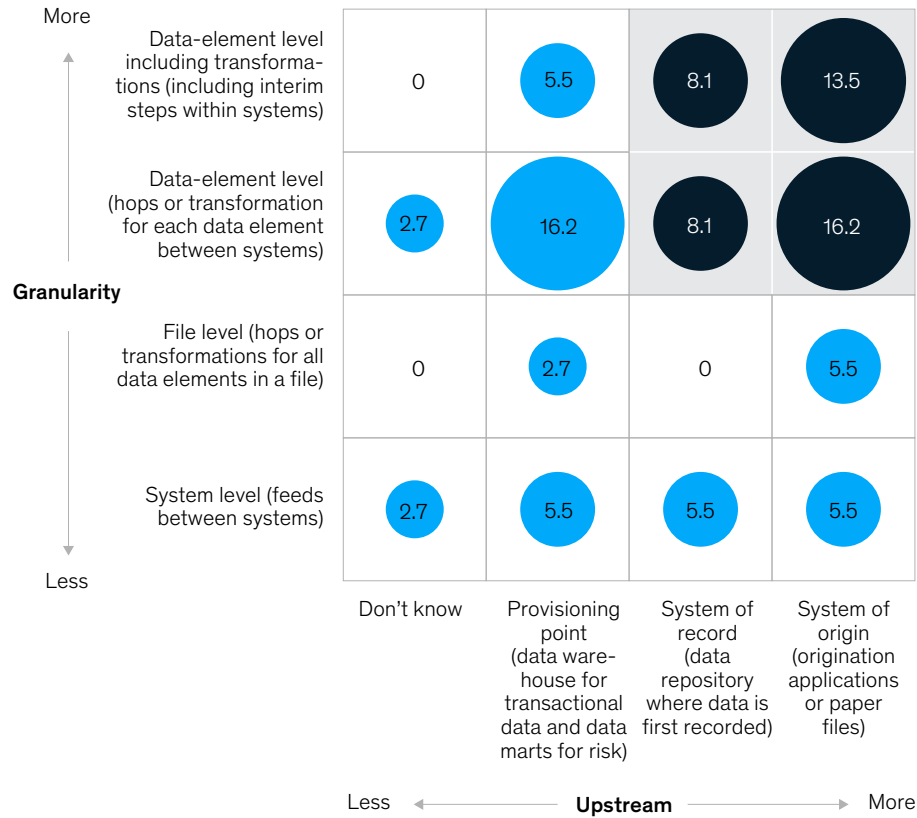
Our benchmarking revealed that more than half of banks are opting for the strictest data-lineage standards possible, tracing back to the system of record at the data-element level (Exhibit 3). We also found that leading institutions do not take a one-size-fits-all approach to data. The data-lineage standards they apply are more or less rigorous depending on the data elements involved. For example, they capture the full end-to-end data lineage (including depth and granularity) for critical data elements, while data lineage for less critical data elements extends only as far as systems of record or provisioning points.

Most institutions are looking to reduce the expense and effort required to document data lineage by utilizing increasingly sophisticated technology. Data-lineage tools have traditionally been platform specific, obliging banks to use a tool from the same vendor that provided their data warehouse or their ETL tools (extract, transform, and load). However, newer tools are becoming available that can partly automate the data-lineage effort and operate across several platforms. They also offer autodiscovery and integration capabilities based on machine-learning techniques for creating and updating metadata and building interactive data-lineage flows. These tools are not yet widely available and have no proven market leaders, so some banks are experimenting with more than one solution or are developing proprietary solutions.

Exhibit 3

Nearly half of a responding sample of banks produce data lineage at the data-element level back to the system of record.

Data lineage, % share of respondents (n = 37)



Other ways to reduce the data-lineage effort include simplifying the data architecture. For example, by establishing an enterprise data lake, a global bank reduced the number of data hops for a specific report from more than a hundred to just three. Some institutions also use random sampling to determine when full lineage is needed, especially for upstream flows that are especially manual in nature and costly to trace. Another possibility is to adjust the operating model. For instance, banking systems change quickly, so element-level lineages go out of date just as fast. To tackle this issue, some banks are embedding tollgates on change processes to ensure that the documented lineage

is maintained and usable through IT upgrades. Report owners are expected to periodically review and certify the lineage documentation to identify necessary updates.

Data quality

Improving data quality is often considered one of the primary objectives of data management. Most banks have programs for measuring data quality and for analyzing, prioritizing, and remediating issues that are detected. They face two common challenges. First, thresholds and rules are specific to each bank, with little or no consistency across the industry.

Banks are pushing for more sophisticated controls, such as those involving machine learning, as well as greater levels of automation throughout the end-to-end data life cycle.

Although some jurisdictions have attempted to define standards for data-quality rules, these failed to gain traction. Second, remediation efforts often consume significant time and resources, creating massive backlogs at some banks. Some institutions have resorted to establishing vast data-remediation programs with hundreds of dedicated staff involved in mostly manual data-scrubbing activities.

Banks are starting to implement better processes for prioritizing and remediating issues at scale. To this end, some are setting up dedicated funds to remediate data-quality issues more rapidly, rather than relying on the standard, much slower IT prioritization processes. This approach is especially helpful for low- or medium-priority issues that might not otherwise receive enough attention or funding.

As data-quality programs mature, three levels of sophistication in data-quality controls are emerging among banks. The first and most common uses standard reconciliations to measure data quality in completeness, consistency, and validity. At the second level, banks apply statistical analysis to detect anomalies that might indicate accuracy issues. These could be values beyond three standard deviations, or values that change by more than 50 percent in a month. At the third and most sophisticated level, programs use artificial intelligence and machine learning–based

techniques to identify existing and emerging data-quality issues and accelerate remediation efforts (Exhibit 4).

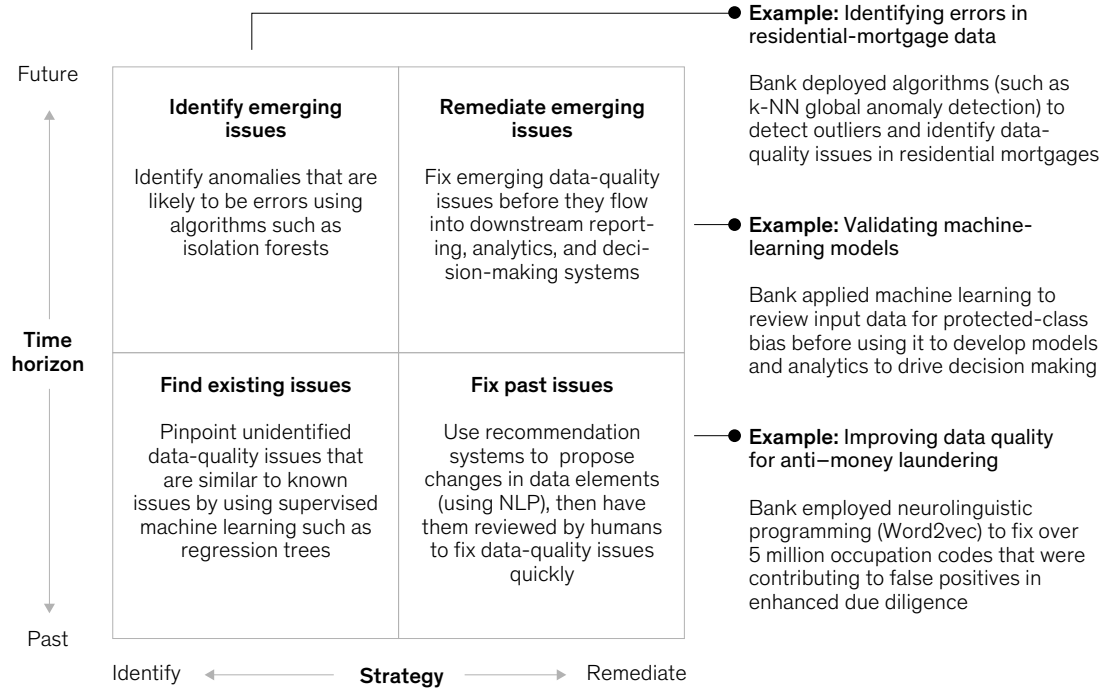
One institution identified accuracy issues by using machine-learning clustering algorithms to analyze a population of loans and spot contextual anomalies, such as when the value of one attribute is incongruent with that of other attributes. Another bank applied artificial intelligence and natural-language processing to hundreds of thousands of records to predict accurately a customer's missing occupation. To do this the program used information captured in free-form text during onboarding and integrated this with third-party data sources.

Leading institutions are revising and enhancing their entire data-control framework. They are developing holistic risk taxonomies that identify all types of data risks, including for accuracy, timeliness, or completeness. They are choosing what control types to use, such as rules, reconciliation, or data-capture drop-downs, and they are also setting the minimum standards for each control type—when the control should be applied and who shall define the threshold, for example. Banks are furthermore pushing for more sophisticated controls, such as those involving machine learning, as well as greater levels of automation throughout the end-to-end data life cycle.

Exhibit 4

Banks are starting to use artificial intelligence to manage data quality.

Finding errors, improving quality, and validating models



Transaction testing

Transaction testing, also referred to as data tracing or account testing, involves checking whether the reported value of data at the end of the journey matches the value at the start of the journey (the source). Banks use transaction testing to assess the validity and accuracy of data used in key reports and to determine if “black box” rules have been implemented correctly. Banks utilize a spectrum of different transaction-testing approaches, with single testing cycles taking between a few weeks and nine months to complete.

Regulators are putting pressure on banks to strengthen their transaction-testing capabilities through direct regulatory feedback and by conducting their own transaction tests at several large banks. At the same time, many banks are inclined to focus more on transaction testing

because they increasingly recognize that maintaining high-quality data can lead to better strategic decision making, permit more accurate modeling, and improve confidence among customers and shareholders.

Banks with distinctive transaction-testing capabilities shine in three areas. First, they have well-defined operating models that conduct transaction testing as an ongoing exercise (rather than a one-off effort), with clearly assigned roles, procedures, and governance oversight. The findings from transaction tests are funneled into existing data-governance processes that assess the impact of identified issues and remediate them.

Second, they strategically automate and expedite transaction testing, utilizing modern technology and tools. While no tools exist that span the

end-to-end process, leading banks are using a combination of best-in-class solutions for critical capabilities (such as document management and retrieval), while building wraparound workflows for integration.

Finally, they apply a risk-based approach to define their transaction-testing methodology. For example, leading banks often select the population for testing by combining data criticality and materiality with other considerations. These could include the persistence or resolution of issues identified in previous tests. Similarly, the size and selection of samples from that population will be related to the population's risk characteristics. While most leading banks opt for a minimum sample size and random sampling, some also use data profiling to inform their sampling, pulling in more samples from potentially

problematic accounts. The review or testing of these samples is often done at an account level (rather than a report level) to allow for cross-report integrity checks, which examine the consistency of data across similar report disclosures.

Although banks have in general made fair progress with data programs, their approaches to building data-management capabilities vary greatly in cost, risk, and value delivered. In the absence of more coordinated guidance from regulators, it is incumbent upon the banking industry to pursue a broader and more harmonized data-control framework based on the risks that need to be managed and the pace of automation to ensure data efforts are sustainable.

Tony Ho is an associate partner in McKinsey's New York office, where **Jorge Machado** and **Kayvaun Rowshankish** are partners; **Satyajit Parekh** is a knowledge expert in the Waltham office, and **John Walsh** is a senior adviser in the Washington, DC, office.

Copyright © 2020 McKinsey & Company. All rights reserved.