White Paper

Find and Eliminate Data Errors with Automated Discovery and Data Lineage

David Loshin,
President of Knowledge Integrity, Inc.

OCTOPAI

Introduction

Organizations have long struggled to identify and take advantage of opportunities for improving data quality and trust. And in recent years, a combination of growing corporate reliance on reporting, analytics, and data science applications, an expanding body of regulations mandating protection of personal data, and increasing concerns about unauthorized or improper data use have motivated the establishment of enterprise data governance programs and corresponding institutionalization of data stewardship practices intended to encompass data quality management processes.

Participants on data governance councils work with data stewards to solicit and formalize business user data quality requirements, specify data quality metrics, and devise methods for measuring compliance with data validity expectations. However, there is often a disconnect between the intended objectives of defining and approving data quality policies and the methods for implementing compliance with those policies. Worse yet, data stewards are often tasked with critical data quality management responsibilities without the proper training, tools, or familiarity of how to find and fix critical data quality challenges.

Data stewards need awareness of **data lineage** – mappings that document the origins of data, the processing paths through which data flows, and the descriptions of the transformations applied to the data along those different paths. Without data lineage, data stewards will be unable to perform the root cause analysis necessary to identify and consequently remedy data quality issues. This allows data flaws to continue to propagate into the business intelligence environments, leading to inconsistent reporting and analyses that influence individuals to make bad decisions.

But as much as data lineage is a critical corporate capability, many companies struggle to establish a data lineage catalog, let alone have on-demand access to accurate and reliable data lineage mappings. That is because in most environments, the complexity of how data flows from the various sources to the business intelligence environments demands manual review, inspection, and documentation. In other words, not only do organizations need to introduce techniques, procedures, and tools that will operationalize the different aspects of data quality measurement, reporting, analysis, and remediation, they also need modern technologies that leverage new algorithms and machine learning to automate how data lineage maps are built. This paper suggests that data lineage can inspire a fundamental behavior change to transition from being an immature to a more mature organization when it comes to data quality and stewardship: transitioning from correcting data to fixing processes. We examine two operational practices for data stewardship: root cause analysis and instituting data controls, and show how data lineage is critical to operationalizing data stewardship in two ways:

1. Providing the information necessary for tracing how data instances flow through an end-to-end process to find the point where data flaws are introduced,

2. Laying out the end-to-end information flows that enable the data steward to determine the appropriate locations to integrate data controls to proactively monitor for quality and automatically generate alerts.

Data Correction vs. Process Correction

Within any end-to-end business process, data in various formats (such as relational database records, JSON objects, rows in a CSV file, or XML objects) move from a variety of acquisition or origination points, through a sequence of processing stages, to a final target destination such as a data warehouse or materialized within some analysis or report. Downstream data consumers, such as business analysts, rely on these reports and analyses to inform business decisions. Yet not all data instance comply with business expectations; sometimes data flaws are introduced that can cause a halt in the processing, or potentially worse, delay or negatively influence important decision-making. Identifying and addressing data errors is a critical data stewardship task, and when a data flaw is found, a data steward must be notified to evaluate the situation and remediate the issue.

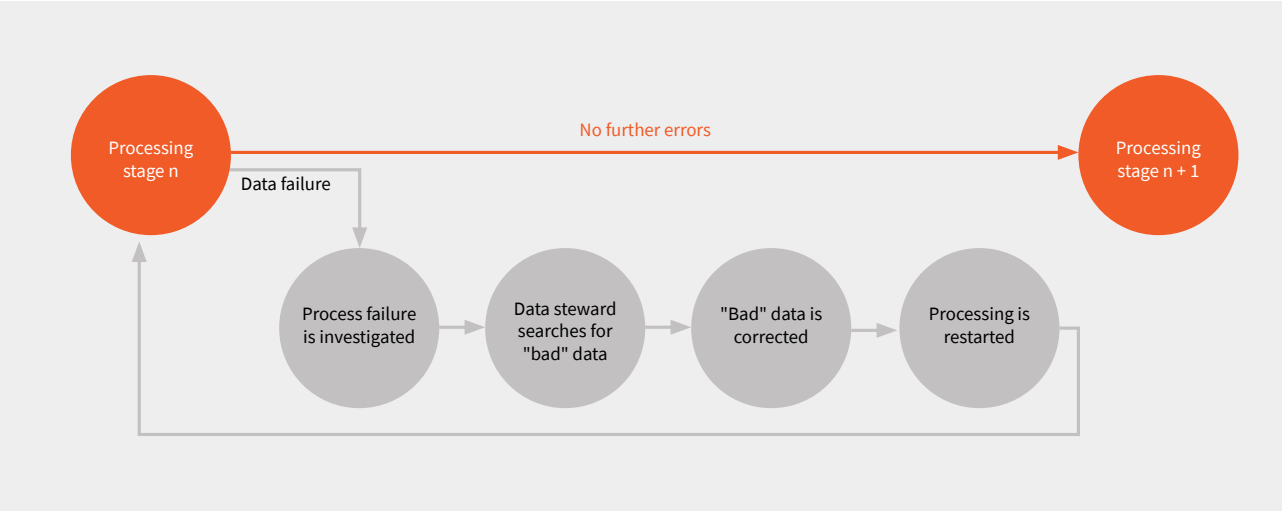

Figure 1: An immature data error remediation process. Note that even though the errors are corrected, they may recur during a subsequent execution.

In immature organizations, a data steward will examine the break in the processing or the flawed report and try to determine the erred data instance that caused the failure. Once the offending record is identified, the data steward may try to ascertain the nature of the error, figure out the values that were wrong, *correct those values*, and restart the processing. Even if it were easy to find the bad data, the problem with this approach is that while it enables today's processing to continue, it does not treat the root cause; no steps have been taken to prevent that data problem from recurring in the future. That being said, the problem is made much more complex when the data stewards do not have visibility into the data lineage and they are unable to track and find the location of the data failure.

In more mature environments, the goal is to go beyond just correcting data errors and instead eliminate their root cause. The data steward's job is to track the flow of data from their origination points across the processing flows, determine how the faulty data instances were introduced into the process flow in the first place, figure out what caused the data to "go bad," **correct the process**, and then restart the processing. Instead of just placing a temporary bandage on the process with selective data correction, this approach, which takes advantage of data lineage, aims to prevent the recurrence of the data failures and avert the possibilities of impacted decision-making.

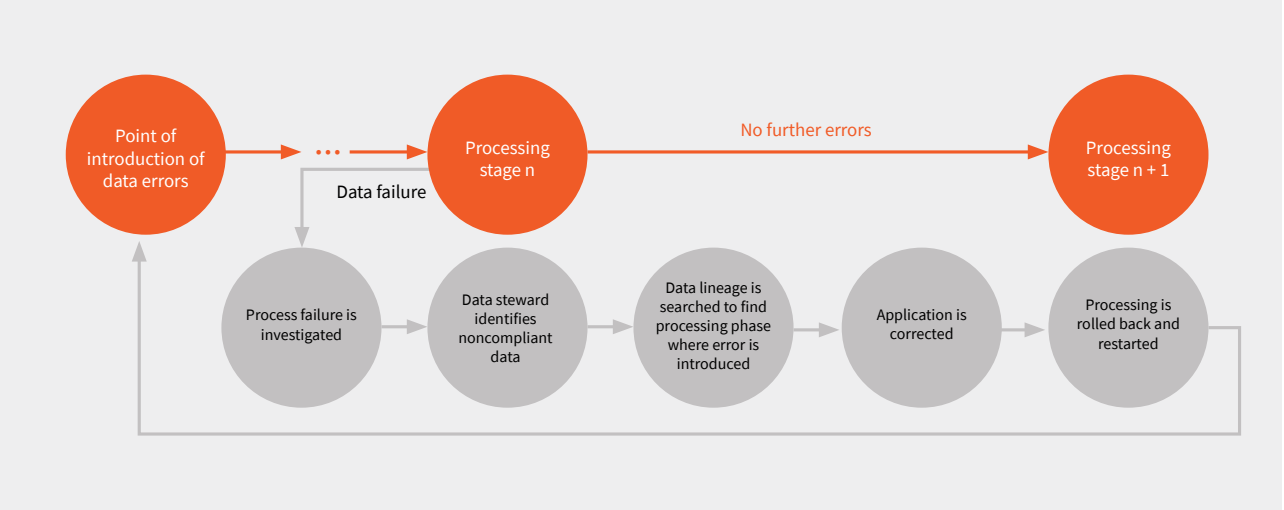

Figure 2: A process reflecting mature data stewardship. Note that since the process is corrected, the risk of introducing the same error is greatly reduced.

Root Cause Analysis with Data Lineage: Reverse-Tracing the Data Flow

Data lineage simplifies the root cause analysis process by providing visibility into the sequence of processing stages through which the data flow. The quality of the data can be examined at each point of the processing flow, enabling a process for finding the point of introduction of data errors.

The process begins at the point where the error was discovered, and proceeds *backward through the lineage* to find where the error was introduced, following this procedure:

1. Identify the phase in the data lineage where the data issue was discovered. This becomes the *point of investigation*.
2. Review the data lineage to identify the processing phase(s) that precede the point of investigation. These are called the *preceding phases.*
3. For each of the preceding phases:
    a. Determine if the data error was present when at the start of the preceding phase. If the data error was present, then the current phase becomes the point of investigation; Jump back to step 2 (to investigate the phases preceding this current point of investigation).
    b. If the data error was not present at the start of the preceding phase, then the data steward has identified that this preceding phase is the one in which the error was introduced. This phase is called the *point of introduction.*

Once the point of introduction has been found, the data steward can work with the application owners to develop small tests, review the code, and determine how the data became noncompliant. Once the root cause has been found, the application can be corrected to prevent that specific error from being introduced in the future.

In other words, as data stewards apply a systematic reverse-tracking of the erred data from the final destination back to the point where the error is introduced, the organization gains the ability to reduce and potentially eliminate recurring errors that can impact critical business decisions. **Recognize, though, that this root cause analysis procedure is difficult, if not impossible to perform without a robust /complete data lineage map.**

Data Quality Rules and Data Controls

Tracing the data flows to find the origination point for data errors is a reactive process. However, the root cause analysis process reveals how data lineage can be used for proactive data quality. The data steward's role is not only to react to acute data issues when they are discovered, but also to solicit data consumer data quality expectations and to transform these expectations into data quality rules. These data quality rules can be used to institute data controls that measure and quantify the levels of data quality at different points of the information flows to report that end-user quality expectations are met.

Consider a simple example of a report that aggregates sales amounts from a variety of eCommerce, telesales, and brick and mortar Point-of-Sale applications on a daily basis. To make it a little more challenging, there are interim aggregations performed for each line of business and for each sales channel. There is an expectation that all sales transactions refer to line items, and that each of those line items has a product code, a price, and the quantity ordered.

This expectation defines a set of data quality completeness rules, namely that in the sets of sales transaction line item records extracted from the various sales systems, the product code, price, and quantity ordered fields may not be empty. A data validation control can be configured to validate that this assertion is true, and the data lineage maps can be consulted to determine the most appropriate locations in the data processing flows to insert those controls to alert a data steward to a noncompliance as early as possible in the process.

When a record is found to violate a data quality rule, a data steward is automatically notified and assigned to investigate and remediate the issue. That way, the source of a data error can be corrected before flawed data can flow downstream and be incorporated into a report without even being detected!

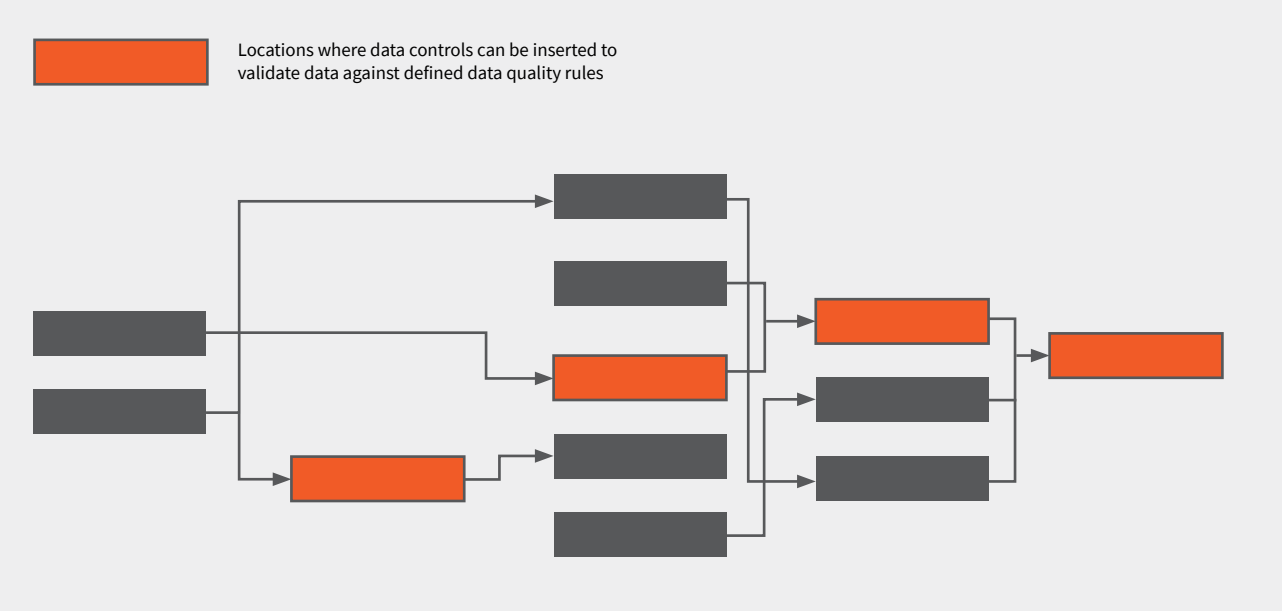

Figure 3: These dotted arrows point to locations in the data lineage where data controls can be inserted.

Using data controls that implement sets of data quality rules to proactively monitor for data inconsistencies and errors represents greater level of maturity in operationalized data stewardship. Automated monitoring and notification not only alerts the data steward to the point in the process flow where the data flaw was introduced, it also characterizes the type of error based on the specific rule (or rules) that were violated. In our example, should the data control indicate a noncompliance, the breadth of the problem can be narrowed to an issue with ensuring that the sales line item records have been properly populated. This significantly optimizes the root cause analysis procedure, and allows the data steward to more rapidly evaluate alternatives for corrective action!

As with the root cause analysis process, knowledge of the data lineage is critical. **Without properly documented data lineage, the data stewards would not be able to determine the processing phases to integrate the data validation controls.**

Automated Data Lineage Mapping

Some metadata management tools provide a "single-point" view of corporate metadata, listing data elements and describing their technical characteristics (such as data type and length). Conventional data lineage adds a horizontal dimension by listing the stages through which the data flow, allowing data professionals to better understand the sequence of events leading to the population of data warehouses and the production of analyses and reports. However, the most innovative offerings not only capture the processing sequences, they *drill down into the data integration, ETL, and report-generation applications and capture the inner/vertical lineage that cannot be captured manually*.

Data lineage:

- Represents the flow of information;
- Documents the end-to-end path through which the data were conveyed from point of origin to the site of persistence;
- Documents the transformations applied to the data at each stage of the path; and
- Provides visibility and traceability across the organization's information architecture.

Collecting the metadata and documenting the process flows and data lineage manually requires a significant amount time, resources, and is prone to error, especially in larger organizations with many reports and analyses. Therefore, it is important to look for tools that can automatically map data lineage across the enterprise. When considering alternatives, look for products that can:

- Scan a wide variety of data sources to assemble a summary of source data metadata. These sources may include static data sets (such as CSV files or spreadsheets), databases, and more complex reporting tools from different vendors used to configure data accessibility and materialization of reports.
- Collect the metadata into a centralized repository that is visible by different roles in the organization to help validate and augment the metadata descriptions.
- Interpret the metadata, infer data types, and link reference data and data elements from across different sources.
- Capture the full horizontal data lineage as well as drill down into ETL and application code to capture the vertical data lineage to provide a multidimensional view into the data production flow.
- Provide a visual presentation of how data flow across the organization.
- Index the data elements and provide a search capability to rapidly trace the flow of data from origination point to all downstream targets.

Summary

As more and more organizations are recognizing the importance of data governance, the need to support operational data stewardship demands that tools for automated data lineage be available. Whether your organization is seeking to establish best practices for root cause analysis or for proactive data validation, having rapidly-developed, searchable data lineage mapping technologies will help to optimize the data steward's day-to-day activities.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader, TDWI instructor, and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices, as the author of numerous books and papers on data management. Knowledge Integrity provides expert consulting guiding clients in developing and launching successful data governance, data quality, and analytics programs. David can be reached at Loshin@knowledge-integrity.com