

From Origin to Destination, Tracing Your Data Ancestry

Getting the most out of your data requires knowing what data you have, where it is, and where it came from—plus understanding its quality and value to the organization. But you can't understand your data in a business context, much less track its physical existence and lineage or maximize its security, quality and value, if it's scattered across different silos in numerous applications.

Data lineage enables data tracking from origin to destination across its lifespan and all the processes it's involved in. It also plays a vital role in data governance. Beyond the ability to know where data came from and whether or not it can be trusted, there's an element of statutory reporting and compliance often requiring knowledge of how that same data (known or unknown, governed or not) has changed over time. A platform that provides insights like data lineage, impact analysis and full-history capture serves as a central hub from which everything can be discovered about the data—whether in a data lake, data vault or traditional data warehouse.

In a traditional data management organization, spreadsheets are used to manage the incoming data design, what's known as the "pre-ETL" mapping documentation, but this doesn't provide visibility or auditability. In fact, each unit of work represented in these 'mapping documents' becomes an independent variable in the overall

around "data at rest" combined with "data in motion," as well as difficulties with legacy architectures, means organizations spend more time finding the data they need rather than using it to produce meaningful business outcomes.

Organizations need to create and sustain an enterprise-wide view of, and easy access to, underlying metadata, but that's a tall order given numerous data types and sources never designed to work together, with infrastructures cobbled together over time with disparate technologies, poor documentation and little thought for downstream integration. So the applications and initiatives that depend on a solid data infrastructure may be compromised, resulting in faulty analyses.

These issues can be addressed with a strong data management strategy underpinned by technology that enables the data quality the business requires, which encompasses data cataloging (integration of data sets from various

management architecture; construct business glossaries; assess what data aligns with specific business rules and policies; and inform how that data is transformed, integrated and federated throughout business processes—complete with full documentation.

Centralized design, immediate lineage and impact analysis, and change-activity logging means you will always have answers readily available, or just a few clicks away. Subsets of data can be identified and generated via predefined templates, generic designs generated from standard mapping documents, and pushed via ETL process for faster processing via automation templates.

With automation, data quality is systematically assured and the data pipeline is seamlessly governed and operationalized to the benefit of all stakeholders. Without such automation, business transformation will be stymied. Large companies with thousands of systems, files and processes will be particularly challenged by a manual approach. And outsourcing data management efforts to professional services firms only increases costs and schedule delays.

erwin Mapping Manager automates enterprise data mapping and code generation for faster time-to-value and greater accuracy for data movement projects, as well as synchronizes data in motion with data management and governance efforts. Map data elements to their sources within a single repository to determine data lineage, deploy data warehouses and other Big Data solutions, and harmonize data integration across platforms. There's no need for specialized resources with knowledge of ETL and database procedural code, making it easy for business analysts, data architects, ETL developers, testers and project managers to collaborate for faster decision-making.

This article was contributed by Danny Sandwell of erwin, Inc., as part of a content sponsorship. It originally appeared in the November 2018 issue of SD Times.



system development lifecycle, and therefore nearly impossible to learn from, much less standardize. The key to accuracy and integrity in any exercise is eliminating human error—that doesn't mean eliminating humans from the process, but incorporating the right tools to reduce the likelihood of error as humans apply thought to the work.

Knowing what data you have, where it lives and where it came from is complicated. The lack of visibility and control

sources), mapping, versioning, business rules and glossaries maintenance and metadata management (associations and lineage).

An automated, metadata-driven framework for cataloging data assets and their flows across the business provides an efficient, agile and dynamic way to generate data lineage from operational source systems (databases, data models, file-based systems, unstructured files and more) across the information