INDUSTRY DEVELOPMENTS AND MODELS

Data Lineage Management: Impact and Value

Stewart Bond

IDC OPINION

Data is at the core of digital transformation, and data without integrity won't be able to support digital transformation initiatives. A key component of data integrity includes being able to trust the data. A key component of data trust is lineage. If you don't know the lineage of your data, you don't know whether you can trust it. Data lineage has been important since before the 1st Platform, increased in importance on the 2nd Platform as data became more distributed, and is of significant importance on the 3rd Platform as the scale of data distribution and the variation of data sources are greater than ever before. Results from a survey of data integration software end users indicate that organizations that are tracking data lineage have more trustworthy data, are able to find data faster, and are able to better support security and privacy requirements compared with those organizations that are not tracking lineage. These survey results, combined with drivers for data with integrity in an era that introduces added challenges of schemaless and ever-changing big data persistence environments, are driving innovations in the metadata management segment of the data integration functional market tracked by IDC. IDC is also seeing metadata management and data lineage components becoming the cornerstones of emerging data intelligence solutions. Increasing numbers of regulatory requirements, regional diversity, data-driven decision making, complex security and privacy requirements, and the era of digital transformation are all driving new requirements for data lineage and expanding the definition to answer the five Ws of data:

- Who is using it?
- What does it mean?
- Where did it come from, and where is it?
- When was it captured, and when is it being used?
- Why is it being stored, and how is it being used?
- How has it changed, and how is it related?

April 2016, IDC #US40332715

IN THIS STUDY

This study provides an overview of the impact and value that data lineage solutions are making in organizations today and makes recommendations to technology solution providers and buyers on how best to meet the needs of data lineage in the era of digital transformation on the 3rd Platform.

SITUATION OVERVIEW

Data lineage is something that organizations have been trying to understand for some time. But why is there a resurgence in the need for better lineage? With so much data coming in from unmanaged external sources (such as sensors, data feeds, and the Internet of Things), the need to track that data as it is ingested, transformed, and moved around the data environment is critical to its effective use. Digital transformation on the 3rd Platform is driving a need for higher levels of data integrity as organizations need to have higher-quality data they can trust as their digital foundation.

Historically, data lineage has been traced in two dimensions or types: where and how. The need for better data intelligence is driving new requirements and expanding the definition and the metadata being captured, including what, when, why, and who. Notwithstanding these new demands, the basics of where and how lineage need to be mastered first for impact and value to be realized.

Where lineage traces the origin of data. How lineage traces how the data source was manipulated to produce the outcome. These two types of lineage can be further defined by levels of granularity: schema (coarse grained) and instance (fine grained). Schema-level where lineage traces the definition of an output data set backward at the point of consumption to understand what data sets were used to produce the output. Schema-level how lineage traces transformations applied to the source data set to produce the output data set. In contrast, fine-grained instance lineage looks at the data values within the schema, understanding where the values came from and how they have changed through transformation to produce the output.

Lineage types and granularities are best understood through an illustrative example. Consider an accounts payable report of the total amount of money paid to vendors over a given period. Schema-level where lineage would trace the output data set back to the source invoice and vendor tables in the accounts payable application. Schema-level how lineage would look at how the vendor and invoice tables were joined and the summarization functions that were performed on the invoice table to produce the total amount paid to each vendor. Instance-level where lineage on the output could trace the amount of money paid to a vendor back to the source invoices that were provided by the vendor. For full process tracing, instance lineage could also link back to the originating requisition, purchase order, and receipt transactions in addition to the approvals for payment.

Benefits of Using Data Lineage

There are many uses for lineage — delivering a wide range of impact and value for organizations that are managing it. Use cases include:

- **Governance:** Providing backward lineage of data to trace results to data owners and sources for quality assurance and access control, in addition to providing forward lineage that allows data owners to manage the use of their data (Combined with a business glossary, data lineage can enable data stewards to manage common definitions and understanding of data terms and fields.)

- **Compliance:** Providing evidence to regulatory bodies on where the data came from, who is using it, and how it has been changed
- **Change management:** Allowing users and developers to understand how a data element change will impact downstream systems and reports
- **Solution development:** Allowing better design, testing, and higher-quality deliverables by sharing lineage, glossary, and relationship metadata across distributed development teams
- **Storage optimization:** Providing insights into what data is being accessed and where, how often, and by whom it is being accessed as input to data archival and disposition decisions
- **Data quality:** Improving quality scores calculated through the application of business and standardization rules against the data, added to the metadata population for input to algorithms and decision making
- **Problem resolution:** Assisting with root-cause analysis in break-fix resolutions processes

A wider business-level benefit of lineage also exists — focusing on changed values of core master data entities that are shared among processes, departments, and applications. An example could be the marketing, sales, and/or service impact of a contact's change of title, department, address, or even employer. The U.S. Bureau of Labor Statistics reports that on average, people have approximately 11 different jobs throughout their career. Add to this the rate at which people in the United States move and change their names each year; the potential change in master data information could be significant depending on how much these statistics reflect the master data population within an organization. The ability to capture, validate, distribute, and trace these changes in a timely manner could lead to better protection of legacy revenue streams and the ability to capitalize on new revenue in B2C and B2B commercial relationships. Take as an example of cable companies offering moving assistance so that they can keep their records up to date and make offers for new services in new locations. Consider also organizations charting capabilities in sales, customer relationship management, and customer service management software.

Survey Results

In the fall of 2015, IDC ran a survey of 651 data integration software end users. Survey results indicated that data governance and metadata management software is making a difference and delivering value. Within the population of respondents that had implemented data governance and stewardship software, in combination with those that had also implemented metadata management software, 90% had also reported seeing tangible benefits. Almost all of the same population reported seeing the tangible benefits within the first year of implementation.

The survey data also provided insights into how the process of data dictionary (business glossary) and lineage management has impacted data quality and availability among respondents that are measuring such metrics.

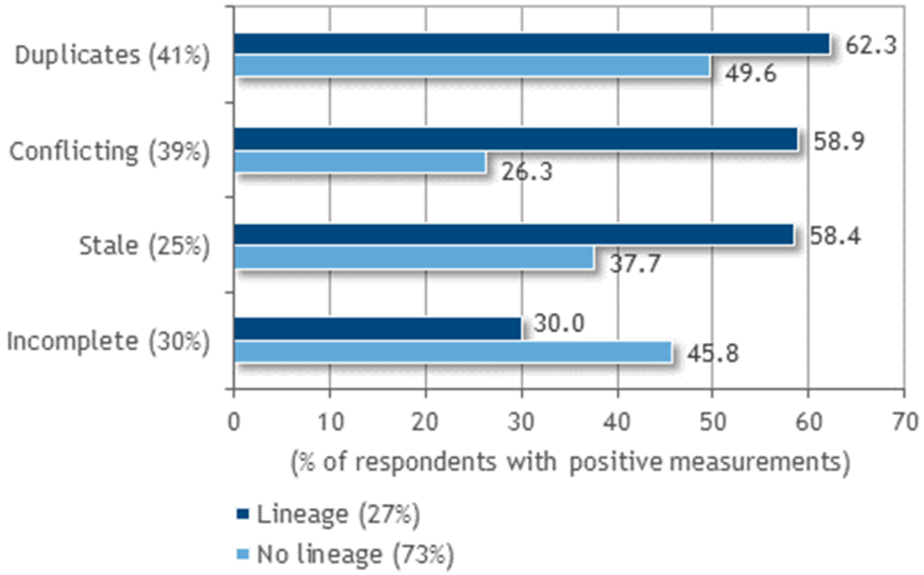
Figure 1 illustrates the difference between the number of respondents that have reported positive data quality measurements and have implemented the process of lineage management and the number of respondents that have reported positive measurements but are not doing lineage management.

FIGURE 1

Lineage Management Impact on Data Quality

Q. Which of the following data metrics have been tracked at your organization (duplicates, conflicting, stale/timeliness, incomplete)?

Q. Has your organization experienced a positive, negative, or neutral change in each metric?



n = 352

Base = respondents filtered to those that have implemented data integration software

Source: IDC's *Data Integration End User Survey,* 2015

Data lineage has a positive impact on reducing the amount of data duplication, reducing the number of data conflicts, and improving the timeliness of data (reducing staleness). Based on these results, we can only hypothesize that data lineage doesn't have positive impact on whether or not data sets are complete.

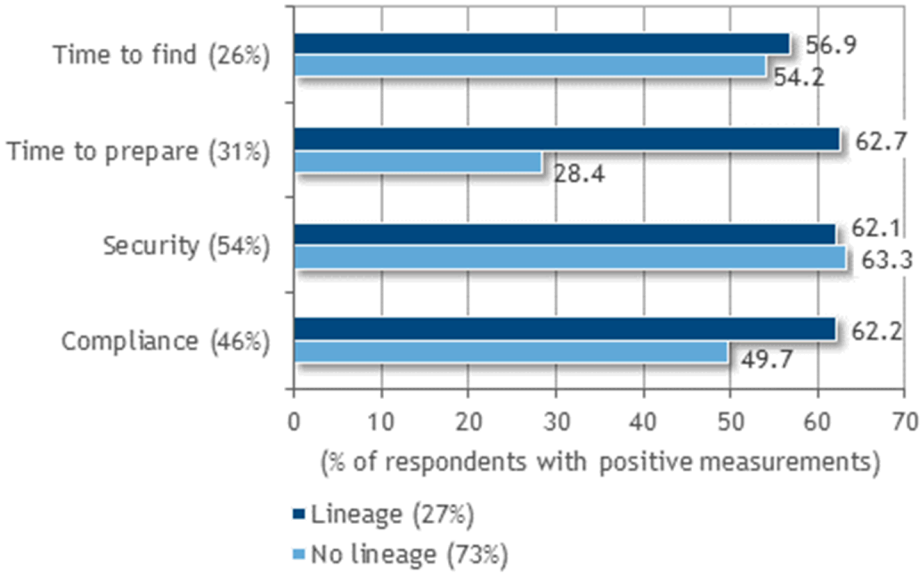
Figure 2 illustrates the difference between the number of respondents that have reported positive data availability measurements and have implemented the process of lineage management and the number of respondents that have reported positive measurements but are not doing lineage management.

FIGURE 2

Lineage Management Impact on Data Availability

Q. Which of the following data metrics have been tracked at your organization (time to prepare, time to find, security, compliance)?

Q. Has your organization experienced a positive, negative, or neutral change in each metric?



n = 352

Base = respondents filtered to those that have implemented data integration software

Source: IDC's *Data Integration End User Survey,* 2015

These results validate many of the qualitative and anecdotal evidence provided previously. There is some positive impact on reducing the time to find data but a significant impact on reducing the amount of time to prepare data for presentation, by a factor of more than twofold. Although one of the case studies referenced in the Case Studies section found lineage to be a benefit for security, it has not had an impact on security in the population of survey respondents. Last, lineage does appear to be having a positive impact on compliance.

Case Studies

Given the broad range of use cases and applications of data lineage, there isn't one return-on-investment (ROI) or value formula, but the following anecdotal evidences offer insight:

- **A major online payments solution provider (source: Primary Research Interview):**
  - The provider uses data lineage as input to solution design early in agile product development sprints. More than 80% of information about data elements used in solutions is available and consistent across distributed development teams, removing assumptions, improving the quality of solutions delivered in sprints, and reducing the length of time to value.

- Lineage also provides the teams with the ability to perform impact analysis of proposed changes and develop regression test cases.
- While the company hasn't quantified the value of lineage, it estimates that data lineage has saved at least one two-week sprint per project. Two weeks of on average 8–10 developers could be $20,000–40,000 per project depending on the average salary and the number of developers. This doesn't include additional time saved through better quality deliverables and fewer break-fix cycles after implementation.
- Audit cycle times have also been reduced as compliance metrics can be reported on through the use of the lineage metadata.

- A utility company (source: Primary Research Interview):
  - Prior to having data lineage available through an automated solution, the utility company had employed 15 data stewards across the organization, each responsible for data in different areas of the business. The company estimated the data stewards spent 30–50% of their time in data forensics: responding to business users' requests to know what the data in a report meant and where it came from.
  - After implementing an automated data lineage solution, the organization was able to deploy business user–friendly data lineage dashboards. As a result, the amount of time data stewards spent on forensics became negligible.
  - Automated data lineage discovery also resulted in uncovering security and compliance issues. Architectural documents and information had not been kept up to date throughout changes to systems, the data warehouse, data marts, and reports. As a result, resources couldn't fully comprehend backward (where did the data come from) and forward (who was using it) lineage. Data lineage helped bring the utility back into compliance with internal and external security policies.
- A U.S. state education system (source: SAS Case Study):
  - The education system has implemented a data lineage solution in its statewide longitudinal data system, covering 26 data sources, hundreds of reports, and thousands of definitions and fields. Lineage metadata is being used to track data changes across the state and identify the impact of changing an element prior to making the changes.
  - The education system claims to be saving on average 80 hours of coding effort on each proposed change through the utilization of backward and forward lineage: from the source data systems through to applications, OLAP cubes, and reports.
- An international bank (source: Teradata Case Study):
  - The bank operates in more than 15 different countries and is subject to more than 15 different sets of legislation and regulators. Data lineage helps the bank stay in business by being able to quickly satisfy audit reporting requirements across the growing number of regulations it is subject to.
  - Less time is spent reconciling data, making more time available for analysts to interpret data and discover insights that can lead to more business opportunities.
- A top 5 U.S. bank (source: ASG Case Study):
  - Initially, the bank required data lineage to assist with data forensic processes and meet federal regulatory audit requirements including TARP and Basel II. Since implementing an automated lineage discovery solution, the value of lineage is being applied in other areas of the business.

    - Lineage is also being used at multiple levels of change management and facilitation of application modernization projects and has been able to reduce operational risks. The bank has been able to decrease time-to-market windows of projects and increase efficiency and transparency.
    - The bank has been able to qualify the value of data lineage and has also been able to quantify the value of an automated lineage discovery solution. Tracking lineage in its complex systems environment was difficult and error prone with manual processes. Through the implementation of an automated lineage solution, the bank was able to reduce its efforts by 80-fold. A detailed analysis showed an approximate $1.1+ million savings on discovering the lineage of 10 key business elements across 100 applications.
- **An example of a case where lineage hasn't been tracked in a postsecondary education institute (source: Primary Research Interview):**
    - Once a year, the institute needs to run a report of professor teaching hours and report to the government for tax and funding purposes.
    - It takes the organization 20 minutes to run the report but three weeks to validate that the information in the report is correct. The organization has not implemented data lineage technology and, in this case, suffers from a lack of where and how lineage at schema and instance levels of granularity. Over time, the system that has been used to track lecture hours has been extended and changed, including changing field content without changing schema.
    - As a result, analysts need to go back through every data entry in every instance for each professor and manually validate the hours in the report.
    - Although it is only three weeks per year, it could represent a nontrivial expense depending on the salary and the number of analysts involved in the validation. The resources involved could be otherwise engaged in more value-added activities.

Data lineage is clearly having an impact on those organizations that are capturing, tracking, and managing it, delivering value in their efforts to be compliant, and improving the level of trust in data. Data integration vendors are responding by providing solutions for managing expanded metadata and lineage, including relationships among data entities driven by digital transformation on the 3rd Platform.

Data lineage solutions are no longer limited to data integration platforms collecting metadata about what happens inside the solution but increasingly are able to collect metadata from the outside looking in and across multiple system components and external data feeds for end-to-end lineage with relationships. The market is also seeing new software vendors emerging to solve the lineage and relationship problems without necessarily also providing data movement and transformation technology – scanning, sniffing, crawling, and parsing through application codes, ETL jobs, data models, SQL queries, report definitions, dashboards, social media data, and structured and unstructured data content to help answer the five Ws of data in the era of digital transformation on the 3rd Platform.

FUTURE OUTLOOK

The impact and the value of data lineage are clear, and the complexity of data lineage in the era of digital transformation on the 3rd Platform is driving innovation in solutions to capture and manage data lineage. Lineage is a key component in the ability to deliver higher-quality data to the business that is trustworthy, available, secure, and compliant.

Key Trends

- Zero-gap data lineage is becoming more important as organizations need to see the full picture of where the data came from and which system components it has traversed. Lineage information provided within the scope of a data integration or data warehouse platform alone doesn't provide the full picture. Data lineage solution vendors have addressed this in the past with manual interfaces but are increasingly introducing new technology for scanning application source codes, SQL queries and stored procedures, and custom-coded solutions to reduce lineage gaps.
- Data lineage is important and a part of emerging data intelligence solutions. In addition to focusing on the new insights that big data can provide, organizations and software vendors have also come to realize the importance and insights that intelligence about the data itself can provide.
- Data intelligence will be used to inform and improve data governance, improve data life-cycle management operations, assist in making data more secure and compliant, and deliver new insights to data owners and stewards.
- Data intelligence solutions will increase their focus on instance lineage. Where and how schema lineages have driven many of the solutions and innovations to date, but new requirements such as relationship tracking drive data lineage solutions down to the instance level. Instance-lineage trends will continue as more needs to be known about where the data for a specific instance of a product, customer, service, location, relationship, or other type of master data entity came from and how it has changed in its lifetime for the data itself to be trustworthy.

ESSENTIAL GUIDANCE

Advice for Technology Buyers

Organizations looking to improve their understanding of data lineage in their environment, whether data lineage capabilities already exist or need to be acquired, can benefit from the following best practices:

- **Scope lineage projects** to one business use case and one business problem at a time. As noted in this document, end-to-end lineage can be enterprisewide, but the full scope isn't realized overnight. Select a business use case and problem that lineage will have an impact on and deliver value for. It could be to meet a compliance need or improve data quality, availability, or security.
- **Gather requirements,** especially if looking for an automated metadata capture and lineage discovery solution. Understand the system, resource, and data constraints prior to shopping for a solution and at the beginning of every project. Inventory all the different technologies and system components that handle data within the scope of the business problem, and look for solutions that can interact with, and introspect each one for end-to-end lineage.
- **Evaluate technology alternatives** looking at vendor solutions that have unified metadata across their platform; are working to expand the type of metadata being captured to answer the five W's of data; and are able to capture lineage across the full data life cycle including what happens to data inside of SQL statements, stored procedures, business applications, and as it flows into or through relational, NoSQL, and big data repositories and ultimately how it is being consumed.

- **Develop and manage business glossaries and data catalogs** as part of the lineage solution. Business glossaries are key to understanding the family trees of data, and catalogs are critical to understanding where the roots and branches are within the organization.
- **Establish metrics, and measure** incremental improvements. Identify a metric that quantifies the business problem being solved. Take measurements before lineage is implemented, and subsequently take measurements throughout the implementation process to illustrate change and improvement.
- **Communicate success,** and socialize solutions with business users. Communication of the value lineage is bringing to the organization, demonstrated through improved metrics and quantifiable benefits, will contribute to project success and secure sponsorship for expanded lineage for solving additional business problems.
- **Iterate** through the use of an agile methodology or simply by continued expansion of lineage scope. Be prepared to continually iterate to bring more value to the business in more areas of the business.

LEARN MORE

Related Research

- *Worldwide Data Integration and Access Software 2016 Top 10 Predictions* (IDC #US40332615, January 2016)
- *Data Integration End User Survey: Deployment Report* (IDC #US40734015, December 2015)
- *Data Quality Market Demand* (IDC #257064, June 2015)
- *Worldwide Data Integration and Access Software Forecast, 2015-2019* (IDC #256768, June 2015)
- *Worldwide Data Integration and Access Software Market Shares, 2014: Year of Affirmation of Data Integration on the 3rd Platform* (IDC #256783, June 2015)
- *IDC's Worldwide Data Integration and Access Software Taxonomy, 2015* (IDC #255856, May 2015)

Synopsis

This IDC study provides an overview of the impact and value that data lineage solutions are making in organizations today and makes recommendations on how best to meet the needs of data lineage in the era of digital transformation on the 3rd Platform. Data lineage has been important since before the 1st Platform and is of significant importance on the 3rd Platform as the scale of data distribution and the variation of data sources are greater than ever before.

"Requirements in the era of digital transformation are expanding the definition of lineage to not only include where and how but also answer the rest of the five Ws of data," says Stewart Bond, director, Data Integration Software. "Data lineage is a core element in emerging data intelligence solutions, bringing more insight about the data itself and delivering even more impact and value for data-driven organizations."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com