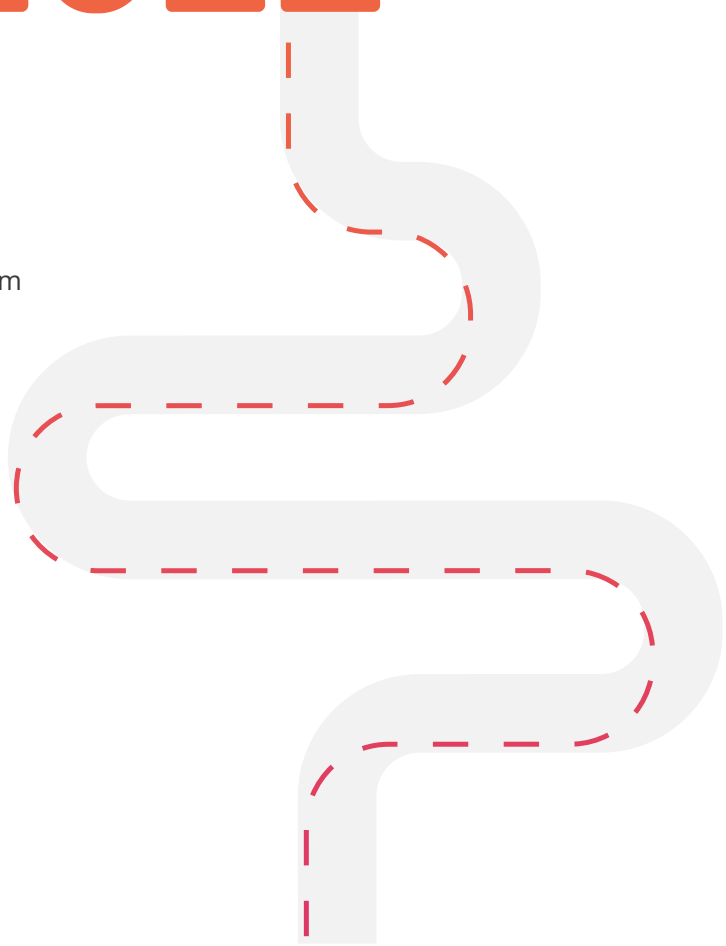
OCTOPAI eBook

The Essential Guide to DATA LINEAGE IN 2022

eBook

Malcolm Chisholm, Ph.D.

President, Data Millennium



What Is Data Lineage?

The core idea behind data lineage is the ability to fully understand how data flows from one place to another within the infrastructure that was built to house and process it. It seems like it should not be a difficult problem, but it is. In fact, it is a huge issue for organizations as they face 2022 and beyond. If organizations do not know where their data comes from or goes, they have uncontrolled environments that have risk on many different levels. Having uncontrolled data environments means that it is also very difficult to extract value from data, and data is the new oil or new gold. Organizations that cannot extract value from data stand a good chance of being outcompeted and replaced by organizations that can.

Our modern technological civilization is full of examples that would make data lineage seem easy by comparison. An oil refinery is a gigantic piece of infrastructure, built to a precise specification, where the operators know exactly what is happening with the liquid products flowing through it – admittedly with a few very rare, and sometimes tragic, exceptions. Modern telephone networks are another exquisite example of complex engineering, being fully controlled from Network Operating Centers (NOCs) so that calls go through from origin to destination without a hitch.

We can agree that data lineage is a hard problem to solve, but just what is it? To answer this, think of an item of data being captured for the first time within the firewalls of an organization, perhaps via data entry. The days when data stayed put in the same silo where it was first captured are long gone. Inevitably, the item of data will be sent to other data stores (databases or files), and from these to yet more, until finally, our item of data ends up as a piece of information in one or more data consumption platforms such as reports, operational systems or even customer facing applications. As the data item travels it may be replicated, or transformed to standardize it, or used in calculations to generate other data elements that enrich the overall data environment. All of this – where the data is stored, the pathways it travels, the changes that happen to it along the way, how it becomes a constituent of other data, and where it appears in the various data consumption platforms – make up data lineage.

Understanding Horizontal and Vertical Data Lineage

Yet there is another twist to data lineage we need to appreciate. It can exist at different levels, each of which has its own particular characteristics and value.

Data lineage was first documented manually as overall flows between systems. This high level, known as **horizontal data lineage**, is usually at the dataset level. It has the advantage of providing a big picture, showing, say, how customer data flows among the organization's systems. This is of enormous help to architects and business users – as far as it goes. But in many cases, it does not go far enough and there are many times when we need to pick a point in the horizontal data lineage and dig deeper. This is where **vertical data lineage** comes in. With vertical data lineage, we go through successive layers of detail until we get to the ultimate level, which is column-to-column (or column-to-report element) plus the transformations that happen at each of these units of data movement. Vertical data lineage answers questions such as where a particular data value in a report came from, or how a data value is transformed as it flows between two columns. It is useful to people like BI Analysts trying to solve a report discrepancy, or data analysts trying to understand the true scope of what exists in an environment that is to be migrated to a new platform.

Data Lineage and Business Processes

So far we have described data lineage and how it can be horizontal and vertical, but this has been done in the context of systems and data objects. But there is another frequently overlooked, but extremely important, aspect of data lineage. Data lineage represents a good deal of the business processes that occur in an organization. Long ago, all processes in an organization were manual, and information went from one person, or department, to another where individual people processed it. Today, all that has changed. We can think of systems having replaced the people who did the processing, and data lineage having replaced the ways in which information was sent between people and departments. An organization must ask itself if the business value chain represented by data lineage is the most effective and efficient way to implement business processes that is possible. All too often, organic growth in IT architecture and point-to-point data transfers accumulate over time to distort how business value chains are actually implemented. This is where data lineage becomes a strategic concern for enterprises, closely linked to their overall business model.

Truth be told, most organizations are still a long way off from using data lineage in business strategy. Nevertheless, the more technical use cases for data lineage are still extremely valuable, and we will look at a range of them now to gain a better appreciation of what data lineage is and the value it provides.

Use Case 1
Assurance of Data Integrity in Reports

A compelling argument for the need for data lineage is quickly resolving end-user doubts about the reliability of the data they are seeing in their reports.

Many BI developers, and report developers in the business, live in constant terror of being asked by a business user to confirm the accuracy of some strange "blip" of data that the user is seeing in a report. The most important thing to understand here is that it is not whether there is an error in the report or not that matters. What matters is whether the developer can give the business user a convincing explanation of what is going on within a reasonable time or not – even if it is an error. Failure to deliver a timely explanation inevitably makes the business user suspect that they are only seeing the tip of the iceberg, and they have no reason to trust the whole suite of reports they are working with.

With data lineage, a BI developer can trace back the lineage of the offending data element and inspect each node in the data lineage chain to determine what is happening. Clarity of the situation – whether good or bad – is achieved.

Our first couple of use cases have focused on more technical aspects. However, data lineage has a wider application, as we will now see.

Use Case 2
Impact Analysis

Changes involving data objects are frequent in an organization. A major headache is the determination of who may be affected if the change is implemented. One way of approaching this is to describe the change and ask a wide range of staff if they think that something in their area may be affected. Exactly which staff should be asked is difficult to know, because the impact is not known. Continuously asking a very broad range of individuals in an organization if they might be affected, when most of them inevitably will not, risks disruption that will lead to complaints about those managing the proposed change. All too often a team managing a change only makes a perfunctory effort to assess impact, and simply makes the change and waits to see if anyone will complain right away.

Data lineage is a huge advantage in impact analysis. The data objects downstream of where the change will be implemented are identified, along with the business users who interact with them. This is important because not every impact is a technical system or data impact, and business process changes may be required.

Use Case 3
Tracking Personal Information (PI)

Data Privacy has exploded in the past few years, with widespread concerns about the misuse of personal information (PI) by governments, corporations, and criminals. New laws have been enacted in several jurisdictions, and the European Union's GDPR explicitly calls out the need to know where PI is located in the data landscape.

The traditional approach to the need to know where PI is located has been data profiling. This involves examining each column in each relational database table (or other non-relational data objects) to try to infer if it is PI or not. However, data lineage provides a better solution. If all the data flows are known, then if a column at any node in a data flow pathway can be identified as PI, then every node in that pathway is logically the same piece of PI. This makes it scalable to have analysts identify PI, since this only has to be done once in any particular pathway. Furthermore, data lineage extends into reporting layers as well as databases, unlike traditional data profiling.

This means that if we can identify a data element in a report as PI, which a report user should be easily able to do, then we can find PI columns across all the lineage pathways involving this data element. An added bonus is that knowing what reports contain PI makes it easier to govern their dissemination – both inside and outside the enterprise.

Using a data lineage tool to bring a data environment under control from the perspective of PI is a concern of Data Governance and any Legal or Privacy function in an organization. This demonstrates how valuable data lineage can be outside of IT.

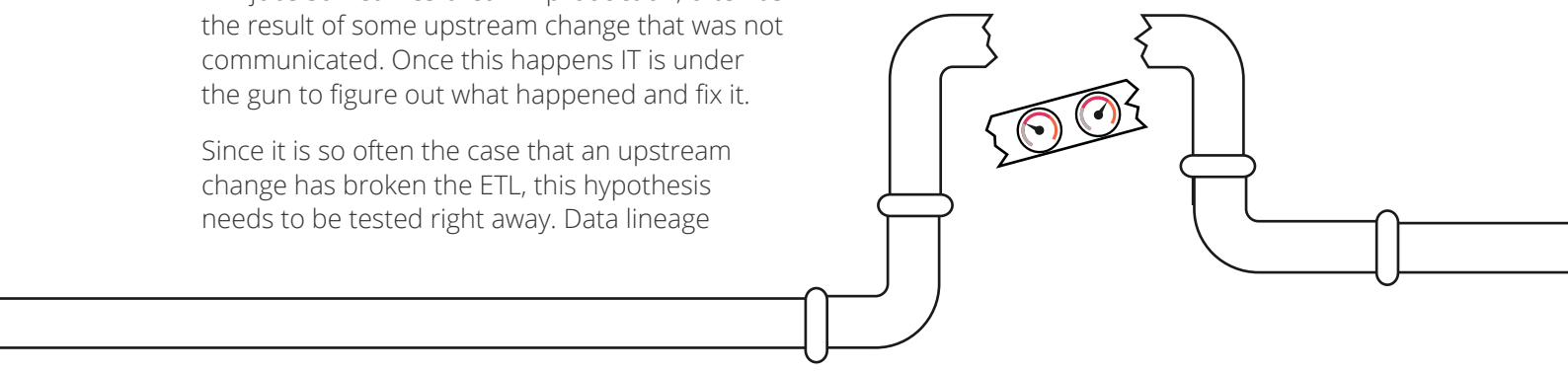
Use Case 4
Broken ETL

This use case is often a consequence of the lack of implementation of the use case we have just discussed. Extract-Transform-and-Load (ETL) tasks move data around the organization, often reshaping, enriching, and integrating the data as they do so. Closely related is ELT (Extract-Load-and-Transform). In ELT, data is taken from a source, loaded into target database, and then transformed. With ETL, there is a more traditional approach of doing the transformation before loading the data. We shall use "ETL" to cover both approaches.

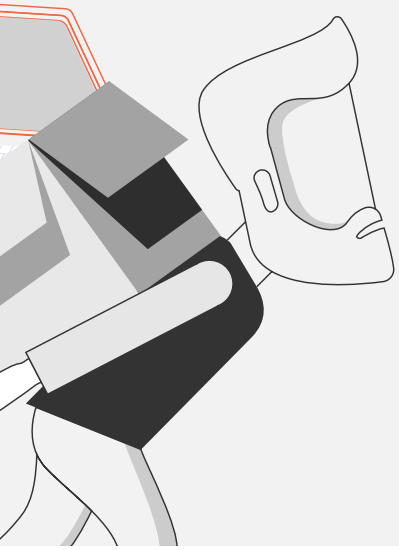
ETL jobs sometimes break in production, often as the result of some upstream change that was not communicated. Once this happens IT is under the gun to figure out what happened and fix it.

Since it is so often the case that an upstream change has broken the ETL, this hypothesis needs to be tested right away. Data lineage allows IT staff to trace back the pathway to the ETL job. With this, it is a comparatively simple job to investigate what if anything has changed in this pathway and fix it.

Most importantly, data lineage can pinpoint what is broken. This means that the root cause of the problem can be detected and analyzed. All too often, this is not done, and downstream workarounds are implemented that further distort the overall architecture. The role of data lineage in root cause analysis and error correction cannot be overstated.



Use Case 5
Migration of Applications and Reports



Migration of applications and reports is needed for many reasons. Support can run out for hardware or software. A new software product can become available with very attractive features. This is often the case with reporting software which seems to turn over on a very frequent basis. Today, however, there is a generational shift away from on-premise data centers to the Cloud, and this is the example of migration we will focus on.

When a migration away from on-premise occurs, there is not simply a need to replicate the data structures, processing logic, and reports in the Cloud. The flow of data also has to be replicated. This means understanding the existing data lineage.

There is an overwhelming temptation to adopt a "lift and shift" approach to migration, meaning that the legacy environment, with all its blemishes, is replicated as closely as possible in the Cloud environment. This is incredibly appealing to project managers and sponsors who want to drive down risk and deliver the project on time. Yet it is a very grave risk to the long-term viability of the mature enterprises that take this approach. A long chain of periodic migrations in an enterprise that has been around for decades heap idiosyncrasy upon idiosyncrasy and workaround upon workaround with each migration.

The point here is that mature enterprises should use data lineage during a migration to understand their business processes and re-engineer them. Data lineage is not just a map of how data flows – it reveals an understanding of how the ultimate business processes have been implemented. Ideally, a mature enterprise will abstract back from the data lineage to what these processes should be today and redesign them.

That said, there are some additional quick wins that can be gained from data lineage during a migration. In particular, identifying data and report objects where data "dead-ends" and which are not used is extremely helpful. There is no point in migrating something that is not used. Thus, these "dead-end" objects and the data lineage pathways to them can be discarded in the migration.

Use Case 6
Assurance of Data Integrity in Reports

In the 1990's Data Administration became very popular, only to largely disappear in the recessions of the early 2000's. It never became a serious component of the Data Governance revolution that began in 2005, probably because of the intensely manual aspect of Data Administration. Yet, data lineage now offers a solution to this problem. Examples of the concerns of Data Administration that data lineage can address include:

- Continuously monitoring for tables and ETL processes that are not used in reporting, and go nowhere. This is not just within the context of the migration projects we mentioned earlier but is a continuous activity of Data Administration.
- Discovering and remediating datatype discrepancies that corrupt data as it flows. A target column may be shorter than a source column, and truncate data. Or a target column may be numeric and remove meaningful leading zeroes from a source column that is character data, but contains only numbers.
- Discovering suspicious data extracts, such as "private" files that might be used for ungoverned data processing, or may even possibly be fraudulent.

There are actually a wide range of use cases that exist within Data Administration in addition to these examples. Without data lineage it is difficult to see how these could be meaningfully addressed at the scale of an entire enterprise data ecosystem.

The Role of Automated Data Lineage

At this point, we have described what data lineage is and illustrated it with a set of use cases where it is particularly valuable. However, it is natural to ask how data lineage actually gets done.

Traditionally, it has been documented during IT development efforts – at least sometimes. Even where there are attempts to document data lineage, they are nearly always wasted effort. The environment evolves, but the documentation is usually not updated. And if there is even the slightest suspicion of the accuracy of the documentation, then all of it is suspect, and nobody trusts it.

Traditionally, this has left us with manual efforts to trace data lineage whenever the need arises – which is pretty frequent. These manual efforts are costly, error-prone, and frustrating to all involved. But today, there is an alternative – automated data lineage discovery.

The tools that enable automated data lineage address the use cases described above very effectively, for the following reasons:

Automated data lineage tools scale well. Very often there are not enough technical IT staff to do manual data lineage work.

Automated data lineage tools are accurate. Manual effort is error-prone, and different analysts may document data lineage differently leading to interpretation problems.

Automated data lineage tools deal well with complexity. Even moderately sized enterprises have a huge number of columns and ETL processes in their data landscapes.

Automated data lineage tools are very fast. They can scan huge environments and produce data lineage diagrams in just minutes. It would take humans between several days to several months to do the same work. As we have seen, there are use cases where answers about data lineage are needed immediately.

IT managers and executive sponsors sometimes make the mistake of adopting a viewpoint that the need for data lineage is very intermittent – like migrations projects - so why should they invest in an automated data lineage tool that will eat up recurrent annual license fees? Sometimes they think it is better to hire consultants to document the data lineage manually when it is needed – a one-time cost. The use cases discussed earlier show that this is a short-sighted view and an automated data lineage tool needs to be on hand for the use cases that are permanent in nature, and others where a rapid response is needed. That really is the situation for most enterprises as they face 2022 and beyond.

Conclusion

In this eBook we have described data lineage in detail with illustrations from several core use cases. We have seen how useful it can be from an IT perspective and a Data Governance perspective. In fact, the importance and value of data lineage goes well beyond what we have described as it is needed to successfully address Data Quality (e.g. source-target reconciliation), Master Data Management (e.g. flows into integration processes), and other aspects of Data Governance (e.g. selecting the best source of data). We have also seen barriers to adoption, including:

- The perception that automated data lineage is needed on an infrequent basis
- Inertia in IT based on data lineage being impractical in the past, and so never discussed
- A lack of understanding of the use cases due to the problems they solve simply being ignored

Yet we have also clearly demonstrated the value of automated data lineage. Going back to our oil refinery metaphor, no refinery could operate without the instrumentation to understand what is happening in it at any time. Why should we expect a complex data environment to function efficiently and without risk if we do not even have a map of how it is laid out? From a perspective of business strategy, operational efficiency, and risk mitigation this map is needed and it is precisely what an automated data lineage tool provides. Perhaps of all the use cases, we have laid out, the most strategic is the first one wherein migrations the opportunity exists to reengineer business processes to match business objectives. The most tactically useful is perhaps the last one, with the ability to quickly diagnose what broke an ETL process. Yet the combination of all of the use cases we have described provides an overwhelming justification for the acquisition of an automated data lineage tool. So overwhelming in fact that we can expect widespread adoption of these tools in 2022.