

MANTA and the Data Lineage Imperative

By Robin Bloor, PHD



MANTA and the Data Lineage Imperative

By Robin Bloor, PHD

The Sins of Our Fathers

At the time of the Renaissance, Fra Luca Pacioli was a Franciscan friar, a mathematician, and a friend of Leonardo da Vinci. He is called “The Father of Accounting” not because he invented double-entry bookkeeping, but because he popularized it. The accounting systems of that era can be thought of as the first true data processing systems, not because they recorded data—that had been done for millennia—but because they embodied an audit trail that demonstrated the lineage of the data and kept it under control.

Given that the virtue of accounting systems had been proved over centuries, it is odd that when modern commercial computing was born in the 1950s and 1960s, nobody insisted on having an audit trail for data. No doubt the computer pioneers of the time would excuse themselves by claiming that in those ancient times data storage was awkwardly expensive, but the truth is that it was never tried.

” The Absence of Data Lineage is a technology mistake.“

 **The Bloor Group**

It's water under the bridge, of course, but let's not pretend that there was never a time when a full record of data lineage could not have been introduced. It's far too late now; a gazillion business applications are currently now dropping trillions of transactions into billions of databases. The imp escaped from the bottle a long time ago, and it has been playing impish tricks ever since.

The Data Landscape

Corporate IT environments are, in general, a tangled labyrinth of data files and databases. They are awkward to navigate, there are no coherent data maps, and misunderstandings and missteps are frequent. While some organizations may be less disorganized than others, the common truth is that the data landscape is chaotic and its organization is inadequate.

We learned this the hard way—by living with it and experiencing it. And now a major intervention may be needed, but if we'd been smarter when we populated this landscape, it wouldn't be so. It all began a long long time ago with the fundamental error of not providing an audit trail for data.

We could label this as a problem of data governance, but in our view, that's ducking the issue. And although data lineage will help in many areas of data governance, as it will in other aspects of IT, like Data Ops,

“Data Lineage touches many things.”

let’s call this out for what it is. The absence of data lineage is a technology mistake.

So if you’re involved in data governance initiatives then tracking data lineage is likely to be something you need to

investigate, and if you’re concerned about the strategic or tactical organization of data: data migration, data quality, data integration, data catalogs and so on, then data lineage ought to attract your attention anyway. Data lineage touches a lot of things.

MANTA and Data Lineage

MANTA FLOW delivers capabilities that allow you to analyze, map, visualize and enhance data lineage in a corporate systems environment.

If you’re a pragmatist at heart, you will like it.

It is technology agnostic, happily integrating with a wide array of other data technologies and governance tools, including those from the big names like Oracle, Informatica, and Callibra, as well as from lesser lights. It refers to itself a Unified Lineage Platform, which means in practice that it focuses unapologetically on data lineage.



You will get a clearer picture of MANTA if we briefly describe how it works. You choose a system or databases or collection of files and register them in the MANTA repository. MANTA then proceeds to analyze what you have pointed it at. It crawls its way through the data definitions and code (including ETL code and SQL scripts) and builds a map. This process is fully automatic. It has been designed to decipher and unravel record definitions, database scripts, ETL workflows and BI requests. It recognizes the intricacies of specific products and makes use of them. It will even include flat files in the maps that it creates.

“If you’re a pragmatist at heart, you will like it.”

When you come to use it, and for example, query a database table,

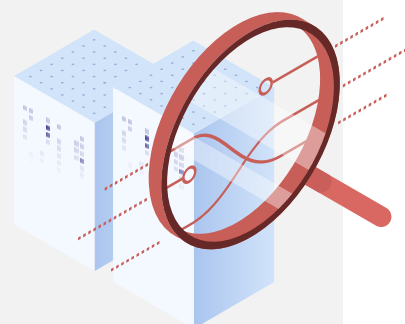
it presents you with a screen similar to the one shown above (which in this instance is chasing down a date). This is a well-defined connected structure you can navigate. You can specify the level of granularity for the map to provide, database level, table level or column level, then click on a single data item and follow the paths it takes. You get to see the origin and the destiny of data at whatever level you choose. It even points out indirect flows, highlighting, for example, where one particular data filter or database join influences the flow of data.

Where Rubber Meets The Road

Customers of MANTA will tell you that it has multiple areas of application and, given the map it creates of the data landscape, that should be no surprise. It is an excellent documentation tool for example. However, there are specific areas where it is frequently employed to great advantage. They are:

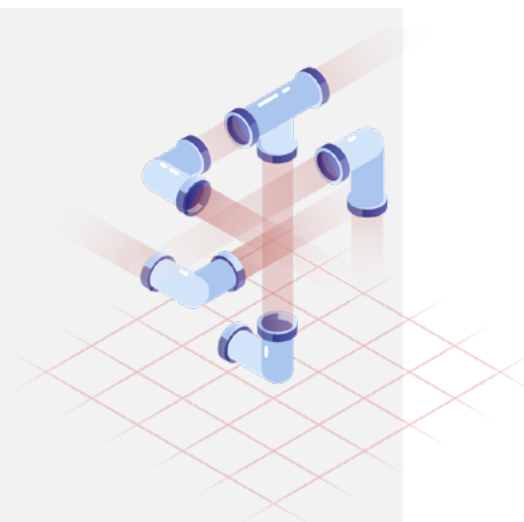
Regulatory Governance

Data lineage is vitally important in complying with regulations like the EU's GDPR, which place onerous demands on businesses that store and use personal data. You only need to read an article or two about it to realize that business will get into trouble if it cannot provide an audit trail for every single item of personal data that it gathers uses and stores. The same applies with Basel II and other such regulatory imperatives. Audit trails are always a feature of the requirements.



Data Pipeline Management

In recent years we have begun to think of data as flowing in a pipeline or even streaming from one destination to another. MANTA is particularly useful for IT shops that wish to manage data flows in that way or are already doing so. It can be used to design data pipelines at any level of granularity down to the data item and it will keep audit trail of where data came from, what it was used for and where it went.



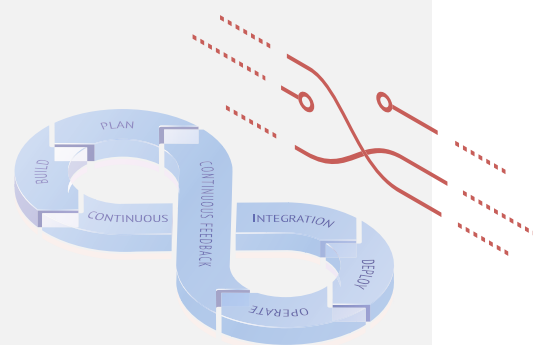
Data Migration

Data migrations are often difficult. That is particularly the case when migrating from one database to another in order, say, to leap into the cloud—perhaps from Oracle to a cloud-based Snowflake database. The company will likely save a packet of money, but only if the migration goes smoothly. MANTA will deliver a detailed validation of such a migration. If the data lineage snapshot of the system prior to migration matches the migrated system, the migration is clean.



Data Ops

There are two distinct areas of Data Ops activity where MANTA can prove useful. It will be useful to developers to carry out impact analysis, whether they are building a whole new application or simply making a small change. It can be useful in fault diagnosis, particularly in situations where an ETL job fails. It can also be useful in creating a comprehensive map of all the BI applications to carry out audits of who is using what and why.



Versatility

Every organization that makes extensive use of data is faced with the problem of trying to remain effective while undergoing constant digital transformation. In our view, having a data lineage capability reduces that malaise.

Data scientists need data lineage to prove the validity of the algorithms they create. Business analysts need it to source reliable data for the BI capabilities. Data stewards need it to help manage data catalogs and keep the metadata valid. It can even help in areas such as data quality and data security. It is, in our view, a core technology.

Talk to Ernie Ostic, SVP of products at MANTA, and he'll tell you that

“ The data lineage problem is exactly the same now as it was then.”

while the tools and environments have changed in many ways since 1982, with windows and fancy GUIs, the data lineage problem is exactly the same now as it was then.

He's right.

Only now, there are a lot more databases and data files, by an order of magnitude.

It's time we fixed it.

The Data Lineage Spectrum

There are several products that do data lineage. MANTA distinguishes itself from this group in several ways. First is its focus and integration capabilities. The company took the decision long ago to devote itself to data lineage and to integrate with other software rather than to leak into adjacent product areas. Thus, you will find it implemented in sites where it lives symbiotically with data catalog companies that have their own data lineage capabilities.

The point of course, is that MANTA does it better, at a finer level of granularity, in particular able to do lineage right down at the column level and offering better support for and integration with other products.

Unlike most other data lineage products, it captures the lineage history, using an underlying graph database that is indexed by time-frame. You could say it keeps an audit trail of the data lineage.

Clearly this is a necessity for database migration projects but can also prove invaluable for fault diagnosis unexpected happens.

Arguably, **MANTA is best in class.**

The Bottom Line

We have not told the full story of MANTA here. In a sense we've only scratched the surface. Suffice to say that we are deeply impressed with this capability, not just because the software is well designed and well-engineered, but also because it attacks a problem that deserves to be attacked and defeated.

We advise those businesses who are investigating data lineage technology to take a good long look—and if you think it's suitable then step forward to a proof of concept.

” This is a necessity for database migration projects but can also prove invaluable for fault diagnosis unexpected happens.”
