Enterprise Data Management - Data Lineage
Data Lineage has been a hot topic in data management for a while now. This is especially true in the financial sector. Regulatory and supervisory authorities no longer rely on reports that banks and insurers provide on their financial and risk positions. They also want to be able to have a look ‘under the bonnet’: they want to be assured that the data used to generate these reports is “as complete, appropriate, and accurate as possible”, in the language of the Solvency II standards for data quality. Data lineage offers this look under the bonnet.

Data lineage was the topic for the sixth in a series of roundtable discussions on themes in enterprise data management. KPMG together with ASG Technologies hosted a meeting with participants from both the financial and a range of other sectors. The guest speakers were Werner Zuurbier, Manager IT Strategy at Univé, and Frank Berkvens and Rudolf Kunkel from ASG.
Before we look at some of the reasons for the growing interest in data lineage, it is helpful to provide a definition. Many can be found, almost all are variations on this one:

Data lineage traces the flow of data, starting from the lowest level of detail, all the way from its origin to its ultimate destination. It shows how data moves through processes and systems, and how it is transformed and aggregated along the way.

**Drivers for data lineage**

Of all the drivers for data lineage, changes in the regulatory environment are arguably the most important. The financial services sector moves strongly towards data-driven reporting. Solvency II and IFRS 17 for insurers and BCBS 239 - ‘Principles for effective risk data aggregation and risk reporting’ - for banks are examples of regulation with a major impact on financial institutions’ data management. Other regulated sectors like healthcare experience similar developments, and the EU’s General Data Protection Regulation that goes into effect in May 2018 will also affect many sectors.

Data lineage also helps to deal with changes in the technology environment. As companies move from traditional mainframes and data warehouse systems to cloud solutions and data lakes, it becomes more complex to ‘track & trace’ data in their landscape, a development that data lineage can counter. A related driver is the increasing complexity of data within organisations: as companies start using data sources of increasing variety and scale, the ability to ‘track & trace data’ becomes more important. Finally, data is no longer ‘just’ a supportive asset, but is increasingly recognised as an asset with value in itself. This may justify additional investments, e.g. in data lineage, that enable companies to create value from data.

**The case for data lineage**

In light of the developments outlined above, the case for data lineage becomes easier to make. The benefits offered by data lineage can be found in the following inter-related areas:

**Data governance**

Data lineage not only has technical aspects, it is also about people. Data lineage can provide real insights into who works with the data, how data is shared between groups of users and departments, and who owns the data and the systems it resides in. Data lineage can help you to identify the kind of data policies, governance, roles and responsibilities that you have in place for specific data.

**Data quality**

Data lineage is vital for data quality measurement. We often see that the root cause of a data quality issue is located ‘upstream’, near the point of origin, data lineage helps you find the exact location more easily.

**Impact analysis**

Whereas data lineage helps you find your way back to the root cause of data quality issues, it also works in the other direction. Data lineage can be used to study how changes in your IT system or in business processes in which data is used affect specific products or reports ‘downstream’.

**Traceability and control**

Data lineage enables you to identify the key risks in your data cycle, and to check if the proper controls are already in place or need to be improved.

**Regulatory and compliance**

In many cases data lineage is simply a must for compliance, if not explicitly so, then certainly from a practical point of view. In addition, compliance requires that certain standards be met in other areas that benefit from data lineage, e.g. data governance and data quality.
There are essentially two ‘extreme’ approaches to the implementation of data lineage. On one end of the spectrum, there is the fully manual approach in which all mappings throughout the IT landscape - from source to target - are identified, visualised and maintained manually, using generic tools like Microsoft Excel and PowerPoint. On the other end of the spectrum, we find fully automated solutions for data lineage. Tools are available that use scanners to read all the coding and all the data elements and flows in your landscape and create, maintain and update data lineage automatically.

In practice, many data lineage tools are available that facilitate a hybrid approach. Such tools can for example visualise data lineage automatically, but still require manual entry and maintenance of mappings and of source and target systems.

Regardless of which approach you choose, one overriding concern should be to start timely. Unfortunately, experience tells us that too often companies postpone working on data lineage when it’s already ‘two minutes to midnight’. By that time, they have really boxed themselves into a corner: there is no time left to select and implement tooling, and pressure from supervisory authorities is mounting. The only option left at that point is to just start: forget about efficiency, only effectiveness counts. Deploy as many resources as needed to get rid of all the ‘red flags’ and meet the minimum standards.

In the remainder of this paper we will describe both ‘extreme’ approaches to data lineage. We will first lay out a step-by-step framework for a manual implementation, and see how such an approach worked out in the case of Univé. After that we will briefly show how a fully automatic solution offered by ASG works.

The manual way can be considered the long road to data lineage. It requires many hours of work and includes many routine, iterative steps. When a short-term solution is needed or when companies want to start with a pilot project in data lineage, it can however be a valuable approach.

Prepare: Preparation is the crucial initial stage in any data lineage approach, particularly when you need to keep the workload manageable. A key question to answer upfront regards the scope of the project: do you want to have data lineage for one flow or for your entire IT landscape? Another key question: what is the purpose? Should your data lineage project provide insight into business processes, or achieve compliance with Solvency II? And which stakeholders should be involved in this project?

Start at the highest level in your IT landscape and divide your systems and databases into buckets for the main categories, for example: source systems, middleware where aggregation or calculations are done, and reporting tools at the end.

The advice here is to start at the end and work backwards into flow. Take a report or other data element at end of flow, identify what the source of that field is, in which system and table it is located, and also what the technical and business names of that element are. Of course, this requires many iterations, and this is the stage where the most hours have to be put in.

Connect each source and target with each other in an illustration, made for example in PowerPoint or with the support of tooling.

This can be a particularly challenging stage, because a manual approach provides only a snapshot, a view of data lineage at one moment in time. It is not interactive: once something changes in a process or the IT environment, data lineage will not be updated automatically and requires renewed effort to find out how this change affects your data lineage.
Werner Zuurbier is Manager IT Strategy at Univé, a cooperative insurer with over 1.5 million customers who together hold over four million insurance policies. Mr. Zuurbier heads the department of IT Strategy within the IT organisation with a total of around 100 employees. He is accountable, among other things, for the architecture, security and quality of Univé’s IT landscape, as well as for data management and quality. He also calls himself a “first-line compliance officer.”

First things first: data governance
Solvency II forced Univé to think about data management in a way it never had to do before. “We really had to take data management to the next level; regulators, supervisors, IT auditors and external accountants more or less forced us to.” The first step Univé took was creating a data governance structure. “We started with governance because somebody has to be responsible, and somebody has to do the work. Our assumption was that a data management organisation separate from operations was not going to work. We decided therefore that data management should be part of people’s daily jobs.”

Univé has built a virtual data management organisation, with roles at three levels - strategic, tactical and operational - embedded in regular functions. The data governance board consists of finance, operations, actuarial and other managers who also act as data owners, and who are responsible for bringing data management to the next level. Around 15 data stewards work to resolve data issues. Recently a ‘tactical team’ was created that takes up data issues that do not get resolved via standard procedures.

A considerable challenge, and quite an achievement
In the next step, a detailed inventory of the data landscape was made; all the places where data relevant to Solvency II resided were identified, and the flows of this data were mapped. This exercise involved around 74 essential data fields and was all done by hand, in Excel and PowerPoint. And although a few years before Univé had rationalised its IT landscape and got rid of much of its legacy systems, this was still a considerable challenge. In 2016 though, Univé was compliant with Solvency II standards for data quality. “Right now, we have a data management organisation, we have good insights into our data assets and our data flows, and we know what the issues and gaps are that still exist in data quality and accuracy. That is quite an achievement, but as I said, we have to go to the next level.”

The next steps
This step to the next level actually involves several steps, in different directions. One is extending the scope of data lineage beyond Solvency II, to the domain of GDPR (data privacy) and possibly to customer data. Progress also requires finding an alternative to the manually drawn data flows. “These have been very helpful so far in pointing out to people where exactly data issues arise and who should take action to resolve them. But of course, these flows are very difficult to keep up-to-date, and sooner rather than later we will have to start using tooling to maintain and visualise our data flows.”

Getting to the next level also means Mr Zuurbier has to keep pushing to improve data quality and to emphasise how that creates value. “Last year data quality was around 97%, which is quite good considering we’re dealing with millions and millions of records. Even a small issue related to Solvency II, however, may have big consequences. That means we have to do impact analyses and act to resolve these quality issues. In addition, we need to dispel the notion that working on data quality is something we do for others. When data stewards analyse data issues together we discover how an issue that arises in one department causes a lot of rework in another. In this sense data lineage generates insights into the potential business value of improvements in data quality.”

A final step to reach the next level will have to be made in data governance. “For now, data management is a role within a regular job description, but with increases in scope and ambition, the demands on people become too high. I expect that within one or two years we will have added dedicated data management specialists.”

The goal
Where will all these steps lead? The goal for 2021 is to have real-time insight into and corrections of data quality issues, and an automated data impact assessment of changes in data flows. “When we change something in data flows or systems, I want to have real-time insight with one press of a button in how it affects our compliance with Solvency II or privacy regulations.”
One of the challenges companies face is to make their lineage documentation sustainable to keep their lineage documentation current. Many, if not all, of them are looking for automated solutions and ASG introduces a solution that facilitates this. It supports automated capturing and maintenance of data lineage. During the roundtable, ASG Technologies presented their Enterprise Data Intelligence solution. Rudolf Kunkel and Frans Berkvens of ASG explained the functionalities of an automated lineage approach. The Enterprise Data Intelligence solution delivers a tool-agnostic solution that includes discovery of mainframe, distributed and other ETL code, analysing functionalities to ensure there are no gaps in the end-to-end lineage. They explained that this discovery is based on automated scanners which ‘crawl’ through mainframes, platforms, interfaces and systems, capturing metadata of data attributes. This metadata is used to determine logical connections between data attributes. Visualising all these connections delivers the actual data lineage.

Mr Kunkel explained that “IT and data landscapes within organisations are usually complex. So even though ASG can deliver over 220 predefined scanners, in most cases some additional scanning solutions will need to be built by ASG”. So, what if data is connected through manual processes? Mr Kunkel showed how manual activities can be ‘stitched’ into the data lineage. This stitching is an inventory of data connections in office applications. MS Excel of MS Access are hard to eliminate within business processes. Mr Kunkel therefore expects manual stitching to be part of lineage for some time. “In practice, we see that end-to-end data lineages are partially made up of automated retrieval and partially of manual stitching.” However, companies more and more have plans in place to work towards mitigating the amount of manual processes. So, in the longer term, ASG expects that automated lineage will increasingly gain ground within organisations.

The Enterprise Data Intelligence solution also offers the possibility to show the data quality for each attribute within the data lineage. This is done through interfacing with data quality tooling solutions.

Having an up-to-date data lineage with data quality insights available delivers fast insights into every data attribute within every organisational department, which is specifically relevant to data stewards, data architects and data owners.
Contact

Ronald Jonker
Service Leader
Enterprise Data management
KPMG Advisory N.V.

tel.:  +31 10 453 4113
mobile: +31 6 512 05 427
e-mail: jonker.ronald@kpmg.nl

KPMG
Laan van Langerhuize 1
Amstelveen

P.O. Box 74500
1070 DB Amsterdam

KPMG is a leading Data Management Consultancy firm

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.