DataManagement Insight | Handbook

Data Lineage Handbook 2019

@DataMgmtInsight

Search Data Management Insight

www.datamanagementinsight.com

A-Team Insight

JUST PUBLISHED

A-Team Insight | Special Report

RegTech
DataManagement
TradingTech

FRTB
Special Report

sponsored by REFINITIV

@ATeamInsight
Search: A-Team Insight
www.a-teaminsight.com

Detailed insight into the data management requirements for FRTB

This detailed guide will enable you to:

- Learn more about the key requirements of FRTB and how they will affect your business;
- Understand the revised risk models and what you need to do to comply;
- Identify the additional data you will need to source in order to pass the rigorous new testing procedures;
- Explore the data management challenges you will need to be aware of and what you can do about them;
- Investigate best practice implementation and what you need to be doing in order to meet the upcoming deadline in an efficient and cost-effective manner.

Download this guide at
www.datamanagementinsight.com

A-Team Insight

Contents

Introduction 5

Foreword 6

Overview 8

Challenges and opportunities 16

Technology solutions 26

Approaches to implementation 30

Roundtable discussion 33

Outlook 38

A-TEAMGROUP

Editor
Sarah Underwood
sarah.underwood@a-teamgroup.com

A-Team Group
Chief Executive Officer
Angela Wilbraham
angela@a-teamgroup.com

President & Chief Content Officer
Andrew P. Delaney
andrew@a-teamgroup.com

Editorial
Sarah Underwood
sarah.underwood@a-teamgroup.com

RegTech Editor
Lauren McAughtry
lauren@a-teamgroup.com

Sales - Data Management Insight and RegTech Insight (mid & back office)
Jo Webb
jo@a-teamgroup.com

Sales TradingTech Insight and RegTech Insight (front office)
Phil Ansley
phil@a-teamgroup.com

Marketing Operations Manager
Leigh Hill
leigh@a-teamgroup.com

Director of Event Operations
Jeri-Anne McKeon
jeri-anne@a-teamgroup.com

Events Content Manager
Lorna Van Zyl
lorna@a-teamgroup.com

Group Marketing Manager
Claire Snelling
claire@a-teamgroup.com

Social Media Manager
Jamie Icenogle
jamie@a-teamgroup.com

Client Services Manager
Ron Wilbraham
ron@a-teamgroup.com

Production Manager
Sharon Wilbraham
sharon@a-teamgroup.com

Design
Victoria Wren
victoria@wr3n.com

Postal Address
Church Farmhouse
Old Salisbury Road
Stapleford, Salisbury
Wiltshire, SP3 4LN
+44-(0)20 8090 2055
info@a-teamgroup.com
www.a-teamgroup.com
www.a-teaminsight.com
www.datamanagementinsight.com

Introduction

An extensive guide to the challenges and opportunities of data lineage

Welcome to our latest handbook on data lineage, a response to growing interest in the business outcomes firms are seeing as a result of successful implementations, and also of the operational gains that can be made in terms of reduced costs, increased efficiency and improved risk management.

This handbook covers the complete scope of data lineage, with a view to helping you gain management buy-in and budget, decide whether to build or buy a solution, and to take a best practice approach to deployment. It also considers suitable technology tools to build lineage tracking solutions, and the potential of innovative technologies such as cloud, machine learning and artificial intelligence to increase automation, support sustainable lineage and identify previously unseen and potentially rich data applications.

A roundtable discussion covered in the handbook, hosted by A-Team Group and joined by the handbook's sponsors, MarkLogic and Bloomberg, highlights the importance of data lineage for firms in capital markets, as well as the challenges and opportunities of implementation. The sponsors also provide recommendations on how to get data lineage right.

The benefits of data lineage described in the handbook span operations, business and regulatory compliance, and touch on the real possibility of using lineage to support data-driven business strategy and uncover new business possibilities.

Of course, no data management handbook dedicated to capital markets would be complete without reference to regulation, which is covered in a special section outlining regulatory requirements for data lineage and how they can be fulfilled.

We will continue to update you on data lineage developments with blogs on Data Management Insight, one of three technology channels for capital markets participants available on our brand new A-TeamInsight website – www.a-teaminsight.com. You will also be able to find out more about data lineage by registering for A-Team webinars and events, and signing up for our weekly newsletters, all of which you can do via the new website.

Finally, thank you to MarkLogic and Bloomberg for sponsoring this handbook. We hope it will be a useful guide to delivering a successful data lineage programme at your organisation.

Angela Wilbraham
CEO
A-Team Group

Foreword

Lineage by Default

by Giles Nelson, Chief Technology Officer, MarkLogic

If you are betting on a data-driven future, you need to be able to understand the flow of data. That is data lineage – the story of what happens to data over time.

Organisations have come to recognise that data is one of their key assets. It's not just a component of planning and tactics, but of business value in itself. Understanding the value of data leads to new ways of interacting with customers and partners, new products, and new ways of doing business.

Data lineage is powering some businesses today and will be powering all data-driven businesses of the future. With data lineage, you can understand where your data comes from, how it is being used, how it is being changed, and where it's going. You can't evolve your business, nor the way it uses data, without understanding the story of that data.

For example, consider a request from a regulator on trades and why those trades were made. Understanding the full history of each trade as well as the history of the information that led to the trade is vital to efficient and effective compliance that can scale. Data lineage gives this history and the causation.

Choices in technology are important. A traditional relational database is great at storing the connections between the current versions of records, especially if they follow a strict schema, but the history of a record is lost every time there's an update. Records aren't just connected to each other. Many times, an individual record can be the result of merging multiple data sources, which means that you can't assume that an individual piece of data has a single, easily tracked audit trail that proves its provenance and tells its story.

MarkLogic®

Because of this complication, data lineage can't just be tacked on to an existing database. Data lineage isn't just a record of changes, but a deep history that records the origins of records, reasons behind changes, and information about the trustworthiness and completeness of a record's sources.

In the future, it is paramount that data lineage is built into applications from the start. For any application that uses data in a meaningful way, data lineage needs to be a part of the project from requirements to design and through to deployment. It is the same lesson we learned about application security 10 years ago. Data lineage requires a better approach to data integration and new patterns to record changes and usage.

It is clear that regulations, whether to protect personal data or make trading more transparent, are requiring greater levels of traceability and auditability. The regulatory landscape will only continue to become more complex and demanding. Governance of customer information, financial trade records and data usage threatens to become a crippling burden unless technology can keep pace. Data lineage features heavily in meeting these regulatory requirements and, in fact, can be part of promising paths towards making compliance easier (believe it or not).

The discussion is just starting around digital regulation and the automation of governance and oversight. If governance could adapt to new regulations automatically, compliance would not take as much effort and risk as it frequently does today.

We are keeping an eye on digital regulation. It could be the answer to the challenge of compliance in our evolving world, and pervasive data lineage would be the key.

Overview

Data lineage traces data from source to destination, noting every move the data makes and taking into account any changes to the data during its journey. Over recent years, it has become a critical concern and challenge for data managers working in capital markets as it is instrumental in regulatory compliance, provides operational transparency to reduce risk and cost, and is important to managing data for analytics. It also supports the increasingly favoured 'data first' approach to managing data for the benefit of the business.

Initially implemented without specific regulatory requirements to track data across individual data management projects, data lineage rose to prominence following the implementation of BCBS 239 in January 2016, a Basel Committee on Banking Supervision (BCBS) rule designed to improve data aggregation and reporting across financial markets, as well as accountability for data. This required improvements in data governance and data lineage that have since been reinforced by other regulations and financial institutions' recognition of the importance of accurate, complete and sustainable data lineage.

What is data lineage?

Data lineage covers the lifecycle of data, from its origins, through what happens to the data when it is processed by different systems, and where it moves from and to over time. It can be applied to most types of data and systems, and is particularly valuable in complex, high volume data environments. It is also a key element of data governance, providing an understanding of where data comes from, how systems process the data, how it is used and by whom. It also plays well into improving data quality.

Data lineage is sometimes referred to as technical lineage, which represents the flow of physical data through underlying applications, services and data stores, or business lineage, which requires the same underlying technicalities but is perceived as a driver of business intelligence and better business decisions.

By building a picture of how data flows through an organisation and is transformed from source to destination, it is possible to create complete audit trails of data points, an aspect of lineage that has become increasingly necessary to meeting regulatory requirements and ensuring data integrity for the business.

While data lineage helps to track data and identify different processes involved in the data flow and their dependencies, metadata management – the management of data that describes data – is key to capturing enterprise data flow and presenting data lineage. Data lineage solutions based on metadata collect and integrate consistent end-to-end metadata throughout an organisation, and create a metadata repository that is accessible and makes complete data lineage information available to different user groups.

Data lineage is usually represented visually to show the movement of data from source to destination, changes to the data and how it is transformed by processes or users as it moves from one system to another across an enterprise, and how it splits or converges after each move. Visualisation can demonstrate data lineage at different levels of granularity, perhaps at a high level providing data lineage that shows what systems data interacts with before it reaches destination. As the granularity increases, it is possible to provide detail around the particular data such as its attributes and the quality of the data at specific points in the data lineage.

> Scope of data lineage implementation is often determined by regulatory requirements, enterprise data management strategy, data impact and critical data elements

The scope of lineage implementation is often determined by regulatory requirements, enterprise data management strategy, data impact and critical data elements of an organisation. It is not necessary to boil the ocean, but instead identify regulatory requirements for data lineage and business areas to which its application is beneficial.

In many financial firms, users of data lineage include business managers and analysts, compliance professionals, strategy developers, data governance teams, data modellers, and IT management, development and support.

Importance of data lineage

Data lineage is critical to both regulatory compliance and business opportunity.

From a business perspective, and at a base level, data lineage helps firms stay on the right side of regulators and avoid the penalties of non-compliance

From a regulatory perspective, compliance has been tightened up considerably since the 2008 financial crisis with subsequent regulations been designed to avoid a repeat of similar circumstances. Rather than merely producing reports for compliance, these regulations – including BCBS 239, Markets in Financial Instruments Directive II (MiFID II), General Data Protection Regulation (GDPR), Fundamental Review of the Trading Book (FRTB) and the Comprehensive Capital Analysis and Review (CCAR) – now require firms to implement data lineage to demonstrate exactly how they came to the results published in reports. Using data lineage, firms can not only prove the accuracy of results, but also take a proactive approach to identifying and fixing any gaps in required data.

Complete data lineage can also reduce the burden of regulation by providing operational transparency and reducing risk and costs. Its metadata can help firms consolidate regulatory reporting by identifying data that is used across numerous regulations and move towards processing the data once for multiple purposes. Similarly, metadata for data lineage can ease the burden and cost of implementing new regulations.

From a business perspective, and at a base level, data lineage helps firms stay on the right side of regulators and avoid the penalties of non-compliance. Equally important, it helps firms gain an understanding of their data and the impact on data of any changes to strategy, systems and processes. With an understanding of data, firms can gain the benefits of data lineage beyond compliance, including the ability to spot new business opportunities, make better decisions, increase efficiency and reduce costs.

Regulatory drivers

Regulations driving financial institution to implement data lineage include those noted above and detailed here. The use of data lineage in each case is slightly different depending on the specific regulatory data requirement, but the overall theme is the same, to be able to demonstrate where data originated, trace its journey through an organisation, and prove how it has been changed along the way.

BCBS 239

Basel Committee on Banking Supervision Rule 239 (BCBS 239) came into force on January 1, 2016 and is designed to improve risk data aggregation and reporting. It is based on 14 principles that underpin accurate risk aggregation and reporting in normal times and times of crisis. To achieve compliance, banks must capture risk data across the organisation, establish consistent data taxonomies, and store data in a way that makes it easily accessible and straightforward to understand.

> The use of data lineage for regulatory compliance is slightly different depending on the specific regulatory data requirement, but the overall theme is the same

Data lineage requirement: Data lineage must be implemented to support risk aggregation, data accuracy and reporting. Also, and conversely, to ensure risk data can be traced back to its origin and risk reports can be defended.

MiFID II

Markets in Financial Instruments Directive II (MiFID II) is a principles based directive issued by the EU. It took effect on January 3, 2018, and aims to increase transparency across Europe's financial markets and ensure investor protection. The demand for reference and market data for both pre- and post-trade transparency, including trade reporting and transaction reporting, is unprecedented, leading to data management challenges including sourcing required data, reporting in near real-time, and uploading reference and market data to MiFID II mechanisms including Approved Publication Arrangements (APAs) and Approved Reporting Mechanisms (ARMs).

Firms subject to GDPR are dependent on data lineage to track data and provide transparency about where it is and how it used, and allow data subjects to exercise their rights

Data lineage requirement: MiFID II operations can benefit from data lineage in a number of ways. Lineage can be used to identify any gaps in trade reporting data, and any similarities across numerous regulatory reporting obligations. It can also be used to map MiFID II reporting data from source systems to APAs and ARMs and vice versa.

GDPR

General Data Protection Regulation (GDPR) is an EU data privacy regulation that came into force on May 25, 2018. It is designed to harmonise data privacy laws across Europe and protect EU citizens' data privacy. The requirements of GDPR include gaining explicit consent to process personal data, giving data subjects access to their personal data, ensuring data portability, notifying authorities and individuals of data breaches, and giving individuals the right to be forgotten.

Data lineage requirement: Firms subject to GDPR are dependent on data lineage to track data and provide transparency about where it is and how it used. Data lineage provides the ability to demonstrate compliance with the regulation and, from a data subject's perspective, supports access to personal data and the execution of other rights such as the right to be forgotten.

FRTB

Fundamental Review of the Trading Book (FRTB) regulation will take effect in 2022. It is a response to the 2008 financial crisis, which exposed fundamental weaknesses in the design of the trading book regime, and focuses on a revised internal model approach to market risk and capital requirements, a revised standardised approach, a shift from value at risk to an expected shortfall measure of risk, incorporation of the risk of market illiquidity, and reduced scope for arbitrage between banking and trading books. Its data management challenges include data sourcing, facilitating capital calculations, and gathering historical data as well as real-price observations for executed trades, or committed quotes, to meet requirements around non-modellable risk factors (NMRFs) and the linked risk factor eligibility test.

Data lineage requirement: To satisfy the demands of FRTB, data lineage may be needed to track historical data and trade data aggregation required for the risk factor eligibility test of NMRFs, essentially the provision of at least 24 real price observations of the value of the risk factor over the previous 12 months.

CCAR

The Comprehensive Capital Analysis and Review (CCAR) is an annual exercise carried out by the Federal Reserve to assess whether the largest bank holding companies (BHCs) operating in the US have sufficient capital to continue operations throughout times of economic and financial stress, and have robust, forward-looking capital planning processes that account for their unique risks. From a data management perspective, CCAR requires data sourcing, analytics and risk data aggregation for stress tests designed to assess the capital adequacy of BHCs and for regulatory reporting purposes.

> To satisfy the demands of FRTB, data lineage may be needed to track historical data and trade data aggregation required for the risk factor eligibility test of NMRFs

> Few firms can claim complete and entirely successful data lineage, but most have developed a regulatory response that is beginning to yield operational and business benefits

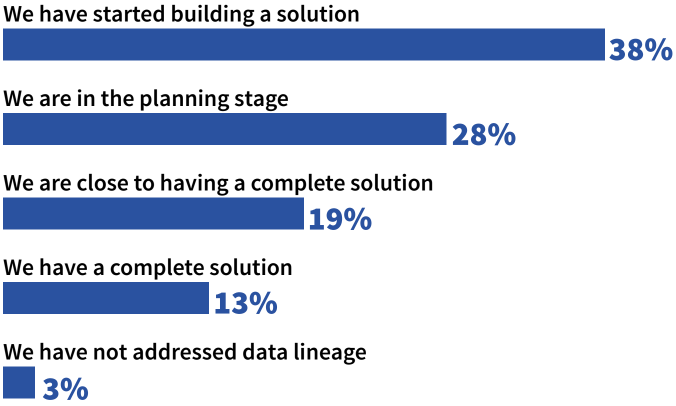
Data lineage requirement: CCAR requires attribute level data lineage to track data from source to destination and ensure the validity and veracity of capital plans. Data lineage can also be used to identify any data gaps in reporting and highlight any data quality issues.

Supply and demand

Over the past few years, a number of established data management vendors have brought data lineage solutions to market, as have start-ups and young companies dedicated to lineage. Some take a technical approach, others a business approach, but their common challenge is to meet growing market demand for automated data lineage that can cross complex data environments and ensure regulatory compliance and deliver business benefit.

On the demand side, recognition and adoption of data lineage has tracked increasing regulation since the financial crisis. Few firms can claim complete and entirely successful systems, but most have developed a regulatory response that is beginning to yield operational and business benefits.

How much progress has your organisation made on data lineage?



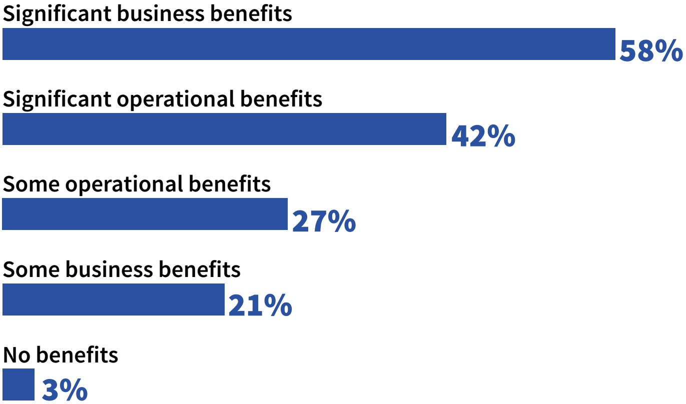
Source: A-Team Group data lineage webinar, 2018

Unsurprisingly, most progress has been made at Tier 1 banks and other large organisations subject to extensive regulation and with the resources to implement data lineage, although all firms that want to stay in the game are likely to need data lineage across some aspects of their business going forward. The results of a poll run during an A-Team Group webinar in 2018 that discussed 'How to Get Data Lineage Right', showed 13% of respondents with a complete solution, 19% close to having a complete solution, 38% starting to build and 28% in the planning stage. The remaining 3% had not yet addressed data lineage.

Most progress has been made at Tier 1 banks and other large organisations subject to extensive regulation and with resources to implement lineage

Looking at the extent of business and operational benefits organisations are gaining, or expect to gain, from data lineage, another poll noted 58% of respondents gaining or expecting to gain significant business benefits, 42% significant operational benefits, 27% some operational benefits, and 21% some business benefits. Just 3% suggested they would gain no benefits.

What extent of business and operational benefits is your organisation gaining, or expecting to gain, from data lineage?



Source: A-Team Group data lineage webinar, 2018

Challenges and opportunities

Overview

Like most data management programmes, data lineage includes inherent challenges and potential opportunities. The challenges range from winning management buy-in for initial projects to understanding and tracking huge volumes of data with complex links across a big data environment. The opportunities range from improved data quality to better decision making and identifying business opportunities.

Challenges

The challenges of data lineage tend to fall into three buckets – operations, technology and data management – and while many are ongoing pain points for data managers across all sorts of programmes, some are specific to data lineage.

Operational challenges

The operational challenges of data lineage start with winning management buy-in and funding for a solution that can be expensive, requires significant human input, and offers only a modicum of advantage in early implementation. Poor understanding of data lineage and its potential benefits by senior executives can stymie approval, while the prospect of lengthy and complex projects could be enough to bring the shutters down.

Questions to consider at the outset of a data lineage project include:

- Where are we now, why do we need data lineage?
- What extent of lineage would be optimal?
- How can we win management buy-in?
- Do we need a champion for data lineage?
- How much will it cost now and going forward?
- How much can we do with allocated budget?
- Do we have required skills internally?
- What are the internal cultural issues of data lineage?

You can begin to answer these questions by ensuring senior management understands the importance of data and benefits of data lineage, and starting small. Decide whether a pilot project is going to provide insight into business processes or achieve an element of regulatory compliance, prioritise the most important and relevant data, scope the project carefully, and identify stakeholders that should be involved.

In the first instance, it may be useful to assess where required data comes from manually and create baseline data lineage before considering automation. It is also important to make sure a pilot project is scalable and could include additional data or other areas of the organisation before making a business case.

Proving the concept of data lineage and demonstrating quick wins to the business should, at least in some cases, be enough to start the journey towards a larger data lineage programme spanning part or all of the organisation.

While a good start to any data management project means it should gain momentum, the success of data lineage is particularly dependent on people and their approaches. It takes a range of data and metadata management skills to develop and maintain data lineage, but if data producers and consumers don't see its value, they are unlikely to fall in with the cause and follow carefully created data lineage processes. These producers and consumers need to look beyond their own environment and understand how the organisation can benefit from data lineage.

That is not to say any data lineage. As data lineage can be expensive to build and manage, it is important to understand what level of data lineage users require. Depending on resources, it may or may not be possible to match extensive requirements, so the initial aim must be to build a data lineage solution that delivers value and is right-sized for consumers, with later iterations providing more detail around data and data flows.

Data ownership and accountability is an ongoing challenge that many organisations with huge amounts of data, myriad systems and applications, and little appetite among employees to take responsibility for data have failed to resolve. Data lineage isn't a silver bullet, but by tracking data and showing how it is used and by whom, it does add some clarity to data and allows responsibility for specific areas of data to be allocated to their rightful owners.

Technology challenges

The technology challenges of data lineage reflect growing numbers of regulations with overlapping lineage requirements and smarter auditors and regulators asking for responses to questions on demand. Advances in technology add to the challenge, with cloud-based applications and services, and big data systems – not to mention emerging machine learning, artificial intelligence and natural language processing technologies – creating a complex data infrastructure. Data can be managed in new and interesting ways, but keeping track of it and ensuring it can be trusted is increasingly difficult.

At the heart of addressing these challenges, and a challenge in itself, is the selection of a solution, or solutions, to support an organisation's data lineage. Early implementations of data lineage were often built in-house as few vendor solutions were available, more recently many firms have moved to hybrid in-house and vendor solutions, or migrated entirely to vendor solutions as data lineage has advanced towards becoming a commodity.

Whether you plan to build or buy, these questions are worth considering before final decisions are made:

- How much lineage is already in place?
- To what extent will manual lineage continue to be necessary?
- How will lineage be documented?
- How will it need to be scaled?
- How will impact assessment be managed?
- What is the long-term aim for automation?
- Which areas of the organisation will be covered and at what level in terms of technical and business lineage?
- How will data lineage be sustained?
- What skills will be required?
- How much will it cost?

There are no catch-all answers to these questions and few organisations that will find answers to all the questions in one solution, leading many to implement a combination of in-house developed and vendor deployed solutions.

Whatever the selected solution, however, it will not provide value in isolation. It is important to consider how data lineage and its metadata will integrate with the rest of an organisation's business metadata as this will provide rich data and the ability to slice and dice the data. Lineage also needs to run alongside an organisation's systems development lifecycle plan to ensure it is maintained as technologies are changed. And, of course, scalable and flexible technology is essential, not only to master growing volumes of existing data types, but also to embrace additional datasets, alternative data, data resulting from mergers and acquisitions, and data that we have yet to discover.

Data management challenges

Implementing data lineage is a complex data management task that could include huge volumes of data, the creation of metadata, multiple legacy systems, mountains of spreadsheets, disparate systems, siloed data, uncharted data flows and mixed data formats.

The potential impact of regulatory change must also be assessed, data quality considered, and manual processes brought into the lineage framework.

Big data, data lakes and repositories raise issues around how data is stored, tagged and linked to other data and systems, while outsourced data and automated data feeds need to be mined and brought into the data lineage scheme.

Data management questions that need to be considered before data lineage is implemented include:

- Is all the data valuable?
- Is the data duplicated?
- Is some of the data redundant?
- Is the data internal?
- Is the data external and correctly licensed?
- What tools are required to find answers to these questions?

Reflecting these questions, an early inventory of an organisation's data can start the process of identifying which data is important to the business and should be part of a data lineage programme, which data can be left as is, and which data can be scrapped. Data in legacy systems and black boxes will difficult, if not impossible, to capture, as will data that changes continually but not consistently.

Considering the scope and scale of these data management challenges, particularly in large organisations, data lineage utopia is not in sight, but there are tools and solutions that can break the backbone of implementation and provide a sturdy platform on which to build and maintain data lineage that can provide useful and timely information to the business.

Opportunities

The opportunities of data lineage can also be divided into buckets, in this case two buckets – business and operations.

Business opportunities

The business opportunities of successful and sustainable data lineage should more than justify the cost of implementation, although that can be frustratingly difficult to prove in the planning and proof of concept stage. From the ground up, they include:

A better understanding of data: This may sound simple, but understanding data that is used and stored across an organisation can be difficult when it includes masses of internal data, several sources of external data, data silos and data in different formats. By applying data lineage, it is possible for both technologists and business analysts to gain a greater understanding of the data a company holds, where it is, what it is used for, its current value and potential. With a good understanding of data, it is also possible to assign responsibility for data ownership to individuals, departments or lines of business.

Data discovery capabilities: Data lineage supports the ability to decide what data is important to the business and enables the data to be accessed quickly. This is crucial to business decisions and can help firms remain competitive and identify new business opportunities.

A focus on data quality and accuracy: Data quality is often an objective of data lineage programmes and can be a natural benefit of tracking data, eliminating duplicate or redundant data, and improving data reliability. Alternatively, data quality can be built into data lineage by monitoring, checking and improving the accuracy and consistency of specific datasets that must be of high quality. The audit element of lineage helps data managers trace errors back to their source, fix any problems and improve data accuracy on an ongoing basis.

Improved data reliability: By tracking data from its origin to its destination, data lineage can identify any gaps in the data that need to be filled, which is often done manually, and reduce the data remediation cycle. Lineage can also point to and eliminate data duplication. First phase implementation begins the journey towards improved data reliability, which will continue as data lineage matures.

More accurate analytics: More reliable and better quality data that is understood and easily accessible supports improved analytics and the knock-on effect of better business decisions.

Better business decisions: By supporting a better understanding of an organisation's data and providing access to trusted data quickly and efficiently, data lineage allows the business to make smarter, faster and better informed decisions. Decisions can be made more proactively where there is data lineage and defended on the basis of being able to determine the exact data underlying any decision.

Identifying business opportunities: A better understanding of data and the visualisation of data and processes, allows organisations to identify new business opportunities, such as the potential to create new products by combining certain data and processes, or the possibility of finding an external partner to upscale and commercialise specific datasets.

Business intelligence and change management: The ability of data lineage to expose an organisation's data lends itself well to business intelligence and change management. What-if analyses can be made using existing data and processes, starter projects can be undertaken to predict outcomes

The only global open data standard enabling effective data management.

Financial instrument global identifier*

Only through an open, shared framework will the financial industry be able to finally address core data quality and lineage issues across legacy codes and standards.

90% of firms globally using more than one existing standard to identify instruments.

48% of firms pointing to incorrect or incomplete instrument identification as cause for a growing percentage of operational errors.

86% of firms agree that an open, shared framework that can establish relationships between different existing legacy instrument identifiers is needed.

Streamline your trading workflow and reduce operational risk. For inquiries regarding FIGI integration:

✉ support@OpenFIGI.com

🌐 OpenFIGI.com

🖥 FIGI **<GO>**

Source: TABB Group Research, April 2018

*FIGI is provided under the MIT Open Source license, with Governance by the Financial Domain Task

OpenFIGI

Bloomberg

of change, and favourable projects can be developed quickly using existing and new resources. Rather than calling on IT to build new systems from scratch, the business can discover how new commercial concepts could work before investing in systems.

Gains in operations

The operational gains of data lineage are as significant as the business opportunities. Some are purely operational, others offer both operational and business benefits. Among them are:

Improved data governance: Data lineage is an important component of data governance and when implemented successfully can support the role of governance in managing the availability, quality and security of data across an organisation.

More responsive regulatory compliance: By creating complete and trusted datasets with transparency and a full audit trail, data lineage eases the burden of gathering the right data for regulatory reporting and helps firms defend decisions when challenged by regulators.

Data lineage can also support data harmonisation across multiple regulations by discovering common data that can be generated once for use by several regulations. On this basis, and by reducing duplication of effort, when a new regulation is introduced, it is also possible to establish what part of the required data is already being governed and documented by another regulation.

Increased efficiency: By eliminating duplicated data and redundant data and systems, and providing a clear view of data and how it changes and moves around an organisation, data lineage can provide increased operational efficiency that can support both cost reduction and business needs for fast access to trusted data.

Impact assessment: Data lineage can be used to study how changes to an organisation's systems infrastructure, business processes and/or data could affect specific products or reports downstream.

Reduced risk: By collecting large amounts of data, organisations expose themselves to regulatory and business liabilities around data breaches and the disclosure of sensitive data. Data lineage can reduce the amount of data held by a company, improve data management and understanding of the data, and provide an audit trail that should help firms avoid the penalties associated with data breaches and wrongful disclosure. Visualisation of data lineage allows organisations to identify key risks in the data cycle and check if controls are in place or need to be improved.

Cost reduction: While often expensive to implement, data lineage offers a number of ways to reduce costs. The need to review data across an organisation as a first step towards successful data lineage allows firms to identify and delete duplicated data, focus on data silos and decide their fate, and discover unused data that can be eradicated as well as redundant systems that can be switched off. This will optimise a firm's data footprint and reduce the challenges and costs of data management.

Data lineage processes also provide an opportunity to review licensed data, which may be licensed more than once in any single organisation or not used to any great extent. It can help firms avoid the penalties of using unlicensed data and renew licenses with data vendors to make external data provision more efficient and cost effective.

The costs and length of transformation programmes and smaller business change projects could be reduced by data lineage and data discovery that identify data and processes that can be reused. Similarly, data lineage can support and reduce the costs of data modernisation and migration programmes.

Data ownership: With a hodgepodge of data and numerous data silos, firms find it difficult, if not impossible, to assign data ownership and accountability to individuals, departments or lines of business and make it stick. Data lineage goes some way to solving the problem by clarifying where data is, who uses it and what for. This allow data ownership to be handed over to the relevant individual, department or line of business that can best exploit the data for financial or operational gain.

Technology solutions

Overview

While most data lineage projects in financial firms start as in-house manual developments responding to regulatory requirements, times are changing. An increasingly regulated environment, growing volumes of data, complex data infrastructures and the need to react quickly to changes and provide fast access to business data are driving firms towards a mix of in-house and vendor, or purely vendor, technology solutions. These come in varying types from enterprise solutions to cloud based services, their commonality being in bringing automation and increased timeliness, flexibility and accuracy to previously manual data lineage processes.

Build or buy?

The answer to the question of whether to build or buy a data lineage solution depends on the particularities of an organisation, whether it is large or small, has already built some automated lineage in house or is working with predominantly manual systems, and what target outcomes it is aiming to achieve.

Early data lineage developments often took an in-house, manual approach that maps data across the IT landscape from source to destination and uses generic tools such as Microsoft Excel and PowerPoint to maintain and visualise lineage. These tools have limitations when it comes to scaling, automating and ensuring the accuracy of data lineage, although they are still in use.

At the other end of the spectrum is fully automated, zero gap data lineage supported by vendor solutions, although in practice, many firms still need hybrid solutions covering both manual and automated lineage, and including both in-house and vendor elements.

A poll question asked during a recent A-Team Group webinar on data lineage showed 42% of respondents using in-house and vendor solutions, 29% building in house, 26% using vendor enterprise solutions, 26% using consultancy support, and 10% using vendor managed solutions.

Underlying technologies

Today, most data lineage solutions are coupled to traditional relational databases and data warehouses, and include automation tools, data management, visualisation and, increasingly, cloud technologies. The combination of data lineage and relational databases can work well, but it can be limiting in large and complex environments with huge amounts of data that must be traced, but are constantly changing. An alternative solution here is a multi-model NoSQL database that has a flexible schema, efficient processing and storage of both structured and unstructured data, and the ability to support high performance queries in a large environment.

Metadata is another underlying component of data lineage. It can be collected from enterprise data flow and integrated in a metadata repository. Data lineage is presented through a metadata abstraction layer that allows data lineage information to be consumed by different user groups, perhaps business users, IT managers, business analysts and the data governance team.

Innovative technology options supporting data lineage and making their way into the market include blockchain, machine learning, artificial intelligence (AI) and graph databases. Blockchain technology is a potential candidate for data lineage as it provides both consensus on the most recent version of golden copy data and an immutable historical record of data.

Machine learning has a part to play by learning and repeating required actions within data lineage, such as automatically tagging private or sensitive data elements and visualising these. AI will go further, perhaps replicating data lineage processes and identifying new business propositions without human intervention.

Graph database technology is also a good match for lineage as it is relatively easy to model data flows in a graph, data relationships can be queried in real time, and a graph schema can evolve to accommodate new data and relationships. The database's query language can be used to understand what data is used by whom, and which systems and reports would be impacted by a change in a particular process. Graph visualisation can help technology and business users investigate data lineage.

Automated data lineage solutions offer many benefits, including the ability to trace data errors, identify discrepancies and missing data, and control access to information

Automation

A typical automation solution for data lineage includes functionality that captures and documents data flows, such as flows of financial instruments, from the data source to their final destination, perhaps a regulatory or internal report. Drilldown functionality allows particular points in the lineage to be inspected more closely, while traceability and audit ensure it is possible to track the journey of a piece of data though across an organisation and verify its accuracy. Filtering capabilities allow users to filter for different data categories, such as reference data or trade data, and understand the data's corresponding attributes.

Data capture often includes the capture of business logic and/or metadata that can be stored in a repository and used to create source to target data lineage, eliminate duplicated or redundant data, and provide business and technical users with the ability to locate, understand, and manage information that supports business operations.

In terms of data quality, an organisation's critical data elements can be identified in data lineage and data quality checks can be established across the organisation as part of a data governance strategy.

These types of automated solutions offer many benefits, including the ability to trace data errors, identify discrepancies, control access to information and model what would happen if a new process or department were added to the business. They can also reduce time spent on validating data accuracy and put trusted information in the hands of decision makers. As yet, however, they do not provide 100% automation, but can reach levels of around 70-80%.

Visualisation

Another important facet of data lineage is visualisation technology, which can provide a real-time view of data moving through an organisation's processes and systems, improve the understanding of data, highlight any defects in data flows, and visualise the impact of any changes to regulatory or business data or systems on upstream and downstream systems. Ultimately, it can help firms determine operational improvements and identify potential business opportunities. Automated data lineage and visualisation tools should be made available to all stakeholders in a lineage programme.

Vendor approaches

Vendor solutions cover similar data lineage functionality in terms of capturing data and creating lineage that makes trusted data accessible to business users and can accommodate changes on the fly. There may be slight differences in underlying technologies, scope and potential for automation, but the key difference, for the moment at least, is delivery, with some vendors providing cloud-based solutions or managed services that can be up and running relatively quickly, and others offering enterprise software solutions that need to be implemented and maintained in-house. Going forward, however, data lineage is likely to follow the steady flow of data, applications and analytics into the cloud.

Approaches to implementation

Overview

Best practice approaches to data lineage implementation initially driven by regulatory compliance requirements are evolving as more firms adopt the discipline for not only regulatory, but also business reasons, as technology tools support increasing automation, and as successful programmes result in significant benefits.

> Start with a small pilot project with a well defined scope that will have a relatively large impact on the organisation. The project needs to demonstrate scalability and true business value

Not all best practice approaches fit all financial firms, but here are some guidelines that should be considered ahead of any data lineage project.

Planning and preparation: Early planning and preparation are key to data lineage to ensure projects stay within scope and budget, and are delivered to an expected specification. Consider the scope of the effort, whether the drivers are about regulation or business, or both, and whether the data lineage should cover one data flow, many flows or the organisation's entire IT landscape. It is also important to determine whether a project will cover lineage for data elements, information assets or databases, although these can be combined.

Early planning must also detail how metadata for data lineage will be included in the organisation's metadata repository and integrated with business metadata to deliver maximum benefits.

Structured and unstructured data: A decision must also be taken on whether to include structured and unstructured data in a lineage project. While structured data lineage is relatively straightforward, perhaps using a handful of attributes that uniquely identify each piece of data, unstructured data is not linear and may require critical data elements to be identified and included in lineage.

Start small: Start with a small pilot project with a well-defined scope that will have a relatively large impact on the organisation.

This project may take a manual approach as a starter to demonstrating the potential of data lineage automation. It also needs to show scalability, efficiency and true business value to win management buy-in and funding.

Understand required skills: Assess whether your organisation has sufficient resources to implement and maintain a data lineage project or whether it must look externally for help. A project is likely to need a champion and an implementation team, as well as business involvement, metadata experts, data custodians and stewards.

Manual data lineage: Consider the scope of the project, its objectives, and the elements in the dataflows that will be covered, such as data sources, systems that aggregate or calculate data, and reporting tools. Make an inventory of all data, circle data specific to the area to which lineage is being applied and discover how the data originates and moves between people, processes, services and products. The level of data granularity that will be included in lineage must be decided and it is then possible to map data flows from source to end point.

Ensure the data lineage is documented accurately and completely, and that the documentation can be updated easily when changes are made to data, systems and flows. This will sustain lineage, but can be difficult in a large or dynamic environment, suggesting development of at least some automated lineage.

Make an inventory of all data, circle data specific to the area to which lineage is being applied and discover how the data originates and moves between people, processes, services and products

Automated data lineage: Early planning and preparation for automated data lineage follow similar steps to manual lineage. With scope, flows and data granularity established, an automation tool can be implemented to gather required data and/or metadata and demonstrate data flows. These tools should be able to react quickly to any changes to data or systems, verify the origins of data, and help firms identify any missing data, data quality issues that need to be addressed, and any other weaknesses in the lineage.

Early planning and preparation for automated data lineage requires scope, flows and data granularity to be established, and an automation tool to be selected to gather data

As mentioned previously, automation is unlikely to provide a complete solution to data lineage, with experts suggesting 70% to 80% is closer to the mark. The remaining portion will need manual intervention, which is also important to monitoring the automated process to ensure it is working correctly, managing exceptions and responding to alerts raised by the solution.

Software development lifecycle (SDLC): To ensure data lineage remains accurate, efficient, accessible and sustainable, it is important to consider how it is going to be included in an organisation's SDLC. Lineage must also be documented to support further systems and software development.

Roundtable discussion

To discover more about vendor views on data lineage in capital markets, how best lineage can be implemented, technology options, and the benefits of successful deployment, we hosted a roundtable discussion with the key sponsors of this handbook, MarkLogic and Bloomberg. We also asked them to make some recommendations to help you get data lineage right.

A-Team: What is your company's interest in data lineage?

MarkLogic: As a leading operational and transactional enterprise NoSQL database provider, MarkLogic has built data lineage into its platform from the ground up. Original data is imported 'as is' and can then be progressively transformed using 'envelope' patterns to enable and maintain data lineage. The original data is stored, and transformations are kept as metadata. The entire lifecycle of every piece of data can therefore be replayed and examined. Our customers rely on our technology and built-in data lineage across many use cases.

Bloomberg: Data is core to everything Bloomberg does, from our open data in OpenFIGI and our Legal Entity Identifier (LEI) Local Operating Unit (LOU) to the data Bloomberg customers rely on daily to provide the most reliable information in the industry so they can trade, execute and manage their risk globally.

A-Team: How important is data lineage to capital markets participants?

Bloomberg: Lineage is one of the pillars supporting cleanliness, usability and accuracy of data. More and more, users are understanding that context is critical to understanding if a piece or set of data is appropriate for a particular use – lineage has a key role in determining that. Where data has originated from, why it was originally created, and what changes have been made to it over time, and by who, all play into the question of whether the data, regardless of the label applied to it, actually fits the context and definition you need.

MarkLogic: Data lineage is critically important. Regulators are requesting more and more information from capital market firms related to trading activities. Markets in Financial Instruments Directive II (MiFID II) is a good example of this, where the Financial Conduct Authority (FCA) is requiring capital market firms to provide greater transparency into trade execution and reconstruct specific trades. This is very difficult to do without data lineage.

A-Team: What are the key challenges of implementation?

MarkLogic: Implementation challenges are usually a result of poor technology choices and data governance approaches, or both. Enterprises that use relational or single-model database systems struggle with data lineage due to the disconnected nature of the resulting data architecture. Only with the ability to handle multiple data models, and keep all original and transformed data as metadata, can enterprises improve data governance and efficiently handle changing business requirements.

Bloomberg: Put simply, implementation is difficult. You can be successful just using a technological solution or approach, but to really get it right, you need to marry in disciplines that many in financial services are less familiar with, and where there is a lack of expertise. Notably, ontologies and semantic views have gained traction over the past few years, but we are still missing a basic understanding of linguistic issues that need to underpin those approaches. I'm referring to context – without an understanding of how different communities in financial services define similar terms, semantic implementations may not take into account inputs that have fundamentally changed the data over time. Lineage is not a simple audit trail, and managing metadata about lineage can become very complex in a short amount of time.

A-Team: What types of technology can enhance data lineage?

Bloomberg: It is less about the technology and more about the concepts and methodology. Understanding that data is not static – both from a perspective that data changes over time, as well as how or why you are looking at a piece of data can change its meaning or intent –

is key to changing the way we approach technical solutions in the first place. Yes, ontologies such as RDF/OWL and the like are useful, but if you view the data as existing only in a single context, you will possibly ignore alternate representations, misrepresent or simplify concepts, or create a massive model that is unwieldy and ultimately not useful.

MarkLogic: From our point of view, data lineage should be a core function of the database. If it is not, or if you have multiple disconnected databases, then your ability to trace and audit data across your enterprise will be a major challenge, and your applications will suffer the consequences.

A-Team: What are the benefits of successful data lineage?

MarkLogic: It is clear that regulations, whether to protect personal data or make trading more transparent, are requiring greater levels of traceability and auditability. The regulatory landscape will only continue to become more complex and more demanding. Governance of customer information, financial trade records and data usage threatens to become a crippling burden unless technology can keep pace. Data lineage features heavily in meeting these regulatory requirements, and in fact, can be a part of promising paths towards making compliance easier.

Bloomberg: Again, the simple answer is better information and, therefore, better decisions. For many general, broad-based macro factors, unclean data is usually mitigated by the volume of cleaner data. But as you get to more specific decisions, looking at the trees as opposed to the forest, so to speak, details become important. Differences between the trees in the forest – their makeup and lineage – become vastly more important if you are picking one over the other to build with.

A-Team: Finally, what recommendations would you make to help firms get data lineage right?

Bloomberg: Be very clear about the community the data is for, and if that changes throughout the data lifecycle.

Know it is OK to have the same data mean different things, as long as context is included and translations are available from one meaning to the other.

Solutions are not static, lineage can branch and evolve – especially when you are not looking.

MarkLogic: Your choice of technology is important. A multi-model operational and transactional NoSQL database system, like MarkLogic, provides maximum flexibility in handling most complex data integration requirements, with built-in ability to trace data lineage across all data sources.

Evaluate data provenance and lineage as part of your enterprise data governance model and practices. You will not only be in a better position to address current requirements, such as those related to regulatory reporting, but also changes required for new regulatory compliance mandates.

Build data lineage into your applications from the start. For any application that uses data in a meaningful way, data lineage needs to be part of the project from requirements to design and through to deployment. It is the same lesson we learned about building in application security 10 years ago.

Overcoming these limitations requires a better approach to data integration and new patterns to record changes and usage.

DataManagement Summit 2019

From A-Team Insight

NEW YORK

SEPTEMBER 19th

DataManagement Insight

From A-Team Insight

@datamgmtinsight

Search: Data Management Insight

www.datamanagementinsight.com

Outlook

From a standing start in 2016, data lineage took off when regulation BCBS 239 came into force with the aim of averting further financial crises on the scale of the 2008 debacle by improving data aggregation and reporting across capital markets.

It has since grown into a critical function for many financial institutions as it delivers not only regulatory compliance, but also benefits such as fast access to reliable information, smarter analytics, better decision making, improved impact assessment of any changes to an organisation's IT landscape, and ultimately, the potential to identify new business opportunities. From an operational perspective, data lineage can help firms reduce costs and risk by eliminating redundant systems and data.

In coming years, data lineage will continue to be driven by regulatory requirements, which will intensify in their demands for increasingly granular data, more detailed audit trails and faster time to data discovery. Data lineage will also be pushed up the agenda by escalating business interest in its significant beneficial outcomes. The build or buy balance is likely to tip in favour of vendors as financial institutions look for more functionality that can be implemented quickly and updated frequently.

Development of data lineage for regulatory, operational and business purposes will be coupled to technology advances including cloud, machine learning and artificial intelligence. These technologies should go some way to increasing automation, supporting sustainable lineage and identifying previously unseen and potentially rich applications of data.

Evolving data standards, metadata, semantics and ontologies will also feature more prominently, easing the burden of the regulated and the regulator, integrating data lineage into operational infrastructure and weaving it into every data dependent business strategy.

While a blue sky scenario depicts zero-gap, 100% automated data lineage, and these are the results we should pursue, the end game beyond regulatory compliance will, at least for the foreseeable future, remain the same – to move from defensive to offensive data intelligence that drives better decisions and unlocks new business opportunities.