DATA LINEAGE
from a business perspective

IRINA STEENBEEK
I first heard the term “data lineage” years ago when our team had implemented a data warehouse solution. One of the consultants came with an Excel sheet and suggested documenting data lineage. The IT team’s reaction was simple and direct: “No way.” They argued that nobody needed this information, and in the worst-case scenario, they could check the system. Sometime later, the topic of data lineage arose again during the implementation of a compliance-related project. One colleague of mine had unsuccess fully tried to gather requirements for data lineage. At some point, he desperately said: “Everybody needs data lineage. Nobody can explain what they mean by that.” So I took over his task. Since then, data lineage has become my area of expertise and my hobby. Over the years since, I have witnessed significant changes in data lineage and observed several new trends.

**DATA LINEAGE TRENDS**

The situation with data lineage demonstrates several trends spanning over the course of years. The three most notable trends are:

1. The increasing pressure of regulatory and business requirements demands documentation of data lineage.
   Years ago, data lineage was a luxury. Nowadays, it has become a daily demand. Recently, different regulators have issued a substantial number of legislative documents with particular requirements for data management. To meet these requirements, companies in diverse industries have to implement data lineage. A rapidly and unpredictably changing economic environment demands corresponding developments in a business environment. Any business change touches upon data. Examples of business changes are projects, digital transformation, big data, advanced data analytics initiatives, and the implementation of cloud solutions. The successful implementation of these changes requires knowledge of the location of data and the transformation that data undergoes along data chains. Data lineage serves as the source of such knowledge.

2. Both technical and business professionals have demonstrated needs and interest in data lineage.
   A while ago, only some technical professionals knew about data lineage. But even fewer had experience with it. Nowadays, data lineage has become a frequently used term among business professionals. For many of them, this concept remains abstract. It still doesn’t plead with the fact that data lineage has become one of the most in-demand business needs.

3. Many different data lineage software solutions have appeared on the market.
   Until recently, the documentation of data lineage has commonly taken place in Microsoft Excel and Word. Recently, several advanced data lineage solutions have been offered on the market. Companies of different sizes and from various industries may find a solution that fits their needs and resources.

Along with these trends, I have recognized several challenges with data lineage.

**CHALLENGES WITH DATA LINEAGE IMPLEMENTATION**

The implementation of data lineage experiences a lot of challenges. The most critical ones are the following:

1. Data lineage remains an abstract concept for many users.
   Data lineage is a complex concept. The data management community doesn’t have an aligned definition of it. Therefore, each company should start a data lineage initiative with the development of a data lineage metamodel.

2. The implementation is complex and time-consuming.
   In any case, the implementation of data lineage requires much effort and many resources. The proper identification of requirements and scope is one of the key success factors.

3. Even if implemented, data management and business professionals don’t use it as expected.
   At the beginning of the data lineage initiative, many stakeholders are not familiar with the concept. Their initial expectations often do not match the real outcomes. Furthermore, the use of data lineage requires some technical skills and knowledge. All of these factors lead to the situation when the results of data lineage implementation remain unclaimed.

The knowledge of trends and experience to overcome the challenges described above are what inspired me to write this book.

**KEY GOALS AND TARGET AUDIENCE**

This book uncovers different aspects of data lineage to data management and business professionals. This book aims to:

- Provide the definition and model of data lineage
  Data lineage is a complex concept, and every company may define the key components of data lineage differently and in a way that best meets the company’s needs.
- Demonstrate best practices in data lineage implementation
  The implementation of data lineage is time and resource-consuming. To make it successful, every company should define the appropriate scope, methods, and solutions.
- Discuss key business areas of data lineage usage
  Proper usage of data lineage should pay back the investments spent on the data lineage initiative. Different business functions may enjoy the outcomes of data lineage.

Several groups of professionals can use this book in different ways:

- Data management and business professionals can develop ideas about data lineage and its application areas.
  There are few resources about the data lineage concept. Articles on the Internet and the sites of data lineage providers are the main sources. The data lineage concept has a lack of aligned terminology. The situation causes challenges for newcomers to become familiar with the subject. This book offers a deep analysis of data lineage and also proposes a metamodel and corresponding terminology. It eases the communication regarding data lineage between different stakeholders.
- Professionals with a technical background may gain a better understanding of business needs and requirements for data lineage.
Different stakeholders have a significantly various understanding, needs, and requirements to data lineage. Technical professionals mainly focus on the implementation of metadata lineage at a physical level. Such a term says nothing to business professionals. This book DOES NOT cover the technical aspects of different data lineage solutions. Instead, it assists technical professionals in building a bridge between their viewpoints and those held by businesspeople. Project management professionals can become familiar with the best practices of data lineage implementation. A proper scope and appropriate methods of implementation are the key success factors of any project. Many factors influence the choice of scope, methods, and approaches. Project management professionals can receive practical recommendations and become familiar with techniques to develop data lineage business cases. The book also provides an overview of some data lineage software solutions.

“All theory, dear friend, is gray, but the golden tree of life springs ever green.”
- Johann Wolfgang von Goethe
The famous quote by Goethe aptly expresses the core idea of this book. We need the theory of data lineage to define our goals, needs, and requirements, as the implementation of data lineage gives us insight into data behavior. However, only the active usage of data lineage allows companies to achieve their goals and satisfy their needs. The three pillars of a data lineage business case are shown in Figure 1:

- Theory
- Implementation
- Usage

Figure 1: Three pillars of a data lineage business case.

The structure of this book follows this three-pillar idea.

**THE STRUCTURE AND CONTENT OF THIS BOOK**

The book consists of three parts that correspond to each pillar: theory, implementation, and usage. Each part includes several chapters that disclose the content of each pillar. A case study in Part 4 demonstrates the practical example of setting up a business case for data lineage. Below, you will find a brief overview of parts and chapters.

**Part 1: Clarifying the concept of data lineage**

Data lineage is a complex concept. The complexity causes ambiguous interpretations of its meaning. Part 1 aims to resolve these challenges.

Chapter 1 provides an overview of different visions regarding data lineage. It also includes an analysis of several concepts that overlap the concept of data lineage.

Chapter 2 discusses key business drivers to document data lineage, such as legislative requirements, business changes, and data management initiatives.

Chapter 3 defines the term “metamodel” and outlines the approach to design the metamodel of data lineage.

Chapter 4 describes in-depth the key components of the data lineage metamodel. Business processes, IT systems, and data models at different levels are examples of data lineage components.

Chapter 5 discusses different types of data lineage. Various business stakeholders have different expectations and requirements for data lineage. The common understanding of data lineage remains the same. At the same time, the way to document data lineage varies depending on the stakeholders’ needs.

By the end of Part 1, readers will gain comprehensive knowledge about the data lineage concept. It forms the basis for the next step: the implementation of data lineage. Different aspects of this will be discussed in Part 2.

**Part 2: Implementing data lineage**

Many factors influence and determine the success of data lineage implementation.

Chapter 6 describes the nine-step methodology to building a data lineage case. The subsequent chapters of Part 2 examine these steps in detail.

Chapter 7 focuses on the key parameters to scope a data lineage initiative. The correct scoping ensures the feasibility of such an initiative.

Chapter 8 provides an overview of key data lineage stakeholders. It also describes in detail the roles involved in the data lineage initiative and their accountabilities. The analysis of different factors that define the roles’ design is also a part of this chapter.

Chapter 9 provides a methodology and template for defining and documenting data lineage requirements. The distinction between metadata and data value lineage requirements is explained. Clear and feasible requirements are one of the keys to implementation success.

Chapter 10 discusses different approaches to implementing data lineage. Several factors influence the various approaches. A method of data lineage documentation, scope parameters, the direction of the documentation, and project management style are examples.

Chapter 11 offers a high-level overview of data lineage solutions that exist in the market. The documentation of data lineage requires suitable software. The most important thing about the solution is to find such that meet your current requirements while leaving space for future developments.

Chapter 12 pays attention to the practical aspects of data lineage documentation. Descriptive and automated data lineage are two different methods of metadata lineage documentation. The application of each of these methods follows diverse steps and has its characteristics.

Chapter 13 shares the practical experience about the success factors of data lineage implementation. By the end of Part 2, readers will have learned several practical insights and recommendations for the implementation of data lineage.

Active involvement of data lineage stakeholders in its usage is discussed in Part 3.
Part 3: Using data lineage

Data lineage outcomes can be used for different business purposes. Data management initiatives, critical data elements, data quality checks and controls are examples of such usage.

Chapter 14 highlights the application of data lineage for the definition of critical data. The concept of critical data is used in different contexts. Critical data assists in the scoping of different data lineage initiatives. The knowledge of data lineage is the key to discovering critical data along data chains.

Chapter 15 demonstrates the importance of data lineage for data quality. Gathering data quality requirements, designing, and building data quality checks and controls can hardly be performed without data lineage.

Chapter 16 discusses the usage of data lineage for impact and root-cause analysis. Different data management initiatives demand such analysis. Data lineage is the only means to perform it.

Chapter 17 shows a possibility to use data lineage for financial planning and analysis tasks. Driver-based modeling techniques and data lineage have a lot in common.

Chapter 18 discusses relationships between data lineage documentation and the implementation of data management frameworks. My experience brought me to the following conclusion. The setup of a data management framework follows the logic of the documentation of data lineage.

By the end of Part 3, readers will have gained a clear view of the three pillars of a data lineage business case: theory, implementation, and usage.

Part 4 consolidates the information received.

Part 4: Case study: “Build a data lineage business case”

Here we will have set the theory of data lineage aside and instead have some fun. In this section of this book, you will read a short story about a fictitious company and its journey in documenting data lineage.

Additional materials

Additional materials add value to this book and assist data lineage newcomers in proceeding with the data lineage initiative. I have added the following:

Template 1: Data lineage requirements

This template assists in gathering business requirements to scope the data lineage initiative.

The document contains general requirements and requirements to key data lineage components that a company would need to capture. The relationships between them are also a part of this template.

Template 2: The scope and progress of data lineage initiative

This template assists in communicating the scope of data lineage initiatives. It is also a good tool to demonstrate the progress of implementation.
Data lineage has become a daily demand. However, data lineage remains an abstract/unknown concept for many users. The implementation is complex and resource-consuming. Even if implemented, it is not used as expected. This book uncovers different aspects of data lineage for data management and business professionals. It provides the definition and metamodel of data lineage, demonstrates best practices in data lineage implementation, and discusses the key areas of data lineage usage.

Several groups of professionals can use this book in different ways:

• Data management and business professionals can develop ideas about data lineage and its application areas.
• Professionals with a technical background may gain a better understanding of business needs and requirements for data lineage.
• Project management professionals can become familiar with the best practices of data lineage implementation.

ABOUT THE AUTHOR

Dr. Irina Steenbeek is a data management practitioner with more than 11 years of experience. The key areas of her professional expertise are data management maturity assessment, implementation of data management frameworks, and data lineage. Irina has practical experience in software implementation such as ERP and DWH/BI, management consultation, financial and business controls, and data science.

Throughout the years, she has worked for global institutions as well as large- and medium-sized organizations in different sectors, including financial institutions, professional services, and IT companies.

In 2016, she has founded Data Crossroads - a training and coaching services enterprise in data management. Data Crossroads focuses on assisting companies in improving their decision-making by setting up an effective data management framework that fits their business goals and resources.

Irina is a strong believer that the success of data management initiatives is based on the combination of a pragmatic approach and clear and transparent methodology. She has shared her approach and implementation experience in her previous books: The “Orange” Model of Data Management, The Data Management Toolkit, and The Data Management Cookbook. She is also the author of various whitepapers and articles on the topic of data management.