Data Product Management in the AI Age

Design and Manage Your Data Strategy to Get Ahead

Jessika Milhomem

Apress®

Data Product Management in the AI Age

Design and Manage Your Data Strategy to Get Ahead

Jessika Milhomem

Apress®

Data Product Management in the AI Age: Design and Manage Your Data Strategy to Get Ahead

Jessika Milhomem
Sao Paulo, Brazil

ISBN-13 (pbk): 979-8-8688-1314-6 ISBN-13 (electronic): 979-8-8688-1315-3
https://doi.org/10.1007/979-8-8688-1315-3

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shaul Elson
Development Editor: Laura Berendson
Coordinating Editor: Gryffin Winkler
Copy Editor: Kim Burton

Cover Image by Racool_studio on Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, New York, NY 10004. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a Delaware LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book can be found here: https://www.apress.com/gp/services/source-code.

If disposing of this product, please recycle the paper

Table of Contents

About the Author

Jessika Milhomem is a Data & AI leader with nearly two decades of experience helping companies turn data challenges into strategic growth. She specialized in bridging technical depth with business impact, performing her work by mixing her hands-on technical skills, cross-domain data strategy and product mindset. She has been leading high-performance teams in national and international projects across a wide range of industries, such as digital banking, digital fashion, telecommunication, education, automotive and consumer goods.

Her passion lies in integrating the triad of business contexts and needs, the power of data, and the team effort dynamic to deliver impactful results. That's why she works close to her teams and leads by empowering them to consume and produce data products and data as a product with strong strategy as a foundation to strengthen business opportunities and decisions in creating innovative customer solutions.

You can find more information about her work at `https://jessikamilhomem.com`.

About the Technical Reviewers

Kent Graziano, a.k.a. The Data Warrior, is an internationally recognized industry thought leader, award-winning speaker, author, and semi-retired Snowflake Data Cloud and Data Vault evangelist with too many decades in the industry to count! He is a Data Vault Master, Knight of the Oaktable, Oracle ACE Director – Alumni, and Taekwondo grandmaster.

Kent is a recognized expert in Snowflake Data Cloud, agile data warehousing, and agile data modeling. He has written numerous articles and blog posts, authored three Kindle books (available on Amazon.com), co-authored four other data-related books, and has done hundreds of presentations and podcasts nationally and internationally. Kent was a co-author of the first edition of *The Data Model Resource Book*, the first data vault book (with Dan Linstedt), and the technical editor for *Super Charge Your Data Warehouse* (the main technical book for DV 1.0).

Lucas Bittencourt Xavier is a data architect with extensive experience designing and building scalable data solutions for startups and large enterprises. He specializes in data architecture, modeling, and automation, creating efficient systems that transform raw data into valuable insights.

Passionate about data quality and performance, Lucas has worked across the full data lifecycle, from designing infrastructure to optimizing analytics. He enjoys solving complex data challenges and continuously improving systems to support business growth.

When he's not working with data, Lucas is constantly exploring new technologies and industry trends.

Acknowledgments

I dedicate this book to God, who always gave me physical and mental capabilities and opened all my opportunities. And to myself for believing in me and always fighting for my goals with love, obstinacy, resilience, and diligence!

I would like to wholeheartedly thank my beloved and loving husband, Ricardo, who is my partner in life from so many angles and perspectives. Of course, he read my book many times and gave me a lot of feedback! He was my first reader and editor!

I appreciate everything you do! I love you very much and the life we have been developing together during these long years. And there are so many that we'll come!

Thank you to my parents, Beth and João, who always did what they could and couldn't, not just for my existence but also for my intellectual development and values, as well as they did for my brothers! And you always did it with much love. I'm a warrior because I have wonderful teachers in my home! Thank you for the values you taught me! I love you!

Thank you to my brothers, Bruno Vinicius and Erika Araceli, for being my best friends, rooting for me, and motivating me whenever I needed it! You are very important to me, and I'm super proud about you both! I love you!

Enzo, Emma, Eric, my beloved nephews, I hope you see this book as an inspiration to dream big and work hard to achieve them! It's completely possible, although eventually hard (or super-ultra hard)! Everything we do with dedication, passion, diligence, and hard work is achievable! We need to do our part while God does them! I love you so much!

I would also like to thank my fifth-year Portuguese language teacher, Cristina Oliveira, who encouraged me to develop a greater love for reading.

Finally, I'd like to thank my editors, Shaul Elson, Gryffin Winkler, and Laura Berendson, who carefully guided me throughout this writing adventure! I do appreciate your work! You're amazing!

Also, I would like to thank my technical reviewers, Kent Graziano and Lucas Bittencourt, for all the detailed and fantastic collaborations carried out during this journey!

Introduction

The Motivation for This Book

You are in the all-hands meeting of the company you work for, and the C-level team is delivering a very inspirational pitch about the importance of innovating by using more and more artificial intelligence. They share some of the projects already in progress or provide a high-level overview of projects that will be implemented in the company. The company's professionals are inspired and very excited with the pitch, especially the senior leadership. After all, the company may achieve great results with all these innovations!

Now, you attend the business unit's leadership meetings to discuss the strategy planning for the following cycle. The teams have planned significant activities for their respective business settings to accomplish a common aim. Some are also considering novel approaches to applying data models and leveraging artificial intelligence for answers. Everything is mapped except the data quality initiatives you formed agreements or advised that they undertake. After drawing attention to it, the leaders state that what they have planned is sufficient for their key aims and it can accomplish the desired results alone through self-service data usage. Coincidentally, the individual contributors from their teams share the same mindset and will just need collaborative work with the data individual contributors.

At another moment, a leader of your business area asks for your help to investigate issues in the data of the teams you've just aligned with. They shared it with another business unit to enable the implementation of a new strategical business process. However, they have faced and are again impacting the business because the data results are mismatched from the expected behavior. It is unclear whether any process failed, the data is wrong due to a bug, or if there's another issue involved. Even worse, they don't know where to start the investigation.

Then, one day, a leader from another Business Unit needs to report consolidated company-level information to the C-level team or to a regulatory agency and comes to you asking for a source of true analytical data from your business domain to ensure the accuracy of the information. However, no such thing exists because each team works with their data using different approaches and rules, and none end-to-end data strategy was designed and considered in any roadmap Consequently, they either have partial information or none at all, as the rest of the areas, including you. However, even though, you and your team still need to work on that super innovative projects of AI that demands data, and are the expectation for the company's interests, that were already set up and communicated by the C-level for the company. And everything is important for the company. All the business needs are real.

Could you see yourself or someone in at least one of these scenarios? I imagine that they might not be uncommon to you. And yes, we could talk about many other "hypothetical scenarios"!

While there is no unique silver bullet to address all of them, concrete strategy and practical work designed to manage data products (from data to AI) is fundamental to ensure all the needs and expectations of the business. And it's critical that leaders have knowledge and experience to deal with it. For this, some skills and frameworks based on important fundamentals can and should be applied to address these challenges. It is the huge opportunity I saw to share the knowledge and techniques I developed throughout my professional career journey over almost two decades in different markets, also experiencing real-world cases with diverse businesses contexts, technical environments, and professionals (mid-senior managers/directors and individual contributors) from different markets.

We're in the era of artificial intelligence. The backbone of this technology is data by providing the content learned, while the algorithm is the brain that guides the muscles to understand the paths.

In other words, data is critical for any strategy nowadays. Thus, robust data product management is fundamental. If the data used is inaccurate or biased in some aspect, the performance of AI systems will be directly affected, as they may learn incorrect patterns. The potential impacts of this can be significant!

In 2021, Gartner stated: "Every year, poor data quality costs organizations an average of $12.9 million. Apart from the immediate impact on revenue, over the long term, poor quality data increases the complexity of data ecosystems and leads to poor decision making."[1]

To succeed, companies need robust data architectures that facilitate self-service analytics, enabling faster product tests and implementations in the short term. These initiatives may not necessarily require best-in-class governed and accurate data.

However, companies must also be capable of successfully scaling emergent solutions into consolidated ones within healthy platforms with foundation capabilities for the middle and long term.

Besides the obvious continuous discoveries for new business opportunities, any company has responsibilities for internal and external purposes that require curated data, fundamentally as the canonical data (a.k.a source of truth data).

These responsibilities entail the following internal and external needs.

- Supporting the design, development, delivery, and maintenance of solutions with scalability, including monitoring, policies, and AI/machine learning (ML) models

[1] Gartner.com (2021), *How to improve your data quality.* Available at https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality

- Providing efficient routine (business as usual) among professionals
- Upholding governance standards for internal and external purposes
- Feeding enterprise mechanisms that deliver information to customers
- Reporting ad hoc or recurrently reports to regulatory agencies

Due to these reasons, it is critical to have responsible professionals for the data strategy of businesses who know how to design and implement data products and data as a product.

Poor Data Quality Definition

The poor data lacks conformance of business rules, completeness of data, format compliance, consistency rules, accuracy verification, uniqueness verification, and timeliness validation (DAMA International, 2015).[2] This results in one and/or more of the following characteristics.

- Inaccurate data
- Duplicate data
- Incomplete data
- Non-compliant data

It can also be generated due to the lack of ownership and leadership of data.

[2] DAMA International, (2015). *Data DMBook*, Technics Publications p. 591 to 601

- Lack of awareness on the part of leadership and staff
- Lack of business governance
- Lack of leadership and management
- Difficulty in justification of improvements
- Inappropriate or ineffective instruments to measure value

Poor Data Quality Impacts Business

The poor data quality directly impacts the business, even if it is not always easy to demonstrate or if it is explicit.

These issues can lead to risks or impacts in at least four aspects.

- **Regulatory violations and fines** (e.g., reporting bad data) result in getting fined
- **Higher operational costs** from unfit processes using inefficient and non-standardized data architectures
- **Loss of (opportunities to generate) revenue** from not understanding the market and internal opportunities to generate revenues, reduce costs and losses
- **Negative reputational impact** from things like when companies announce they miscalculated something (e.g., underpaid compensation costs by Uber to its drivers[3])

[3] Business Insider (2017), *Uber is paying its NYC drivers 'tens of millions' because of an accounting error that underpaid them for years.* Available at https://www.businessinsider.com/uber-paying-nyc-drivers-tens-of-millions-accounting-error-2017-5

The Purpose of This Book

Accurate and very well-managed data is important, and it is not just for companies. It is fundamental for a lot of businesses.

It becomes even more critical now, in the initial era of AI, as data is the information that AI uses to learn and make decisions while using algorithms as rules to follow, process data, and learn from (David Pattel, 2023).[4]

With that in mind, this book aims to help you perform this important responsibility of managing data as a product by helping you to answer important questions.

- What is a data product, and why is it necessary?
- What is data as a product, and why is it necessary?
- What should I know to do this work?
- How should I design it?
- How should I get started?
- What and how should I do it?

Data Product Management in the AI Age is a book intended for several types of readers.

- Data executives who will lead the design and implementation of data products
- Product managers interested in becoming data product managers/owners
- Business people who will work with data teams, managing data products and their projects,

[4] David M Patel (2023). *Artificial Intelligence & Generative AI for Beginners: The Complete Guide.* Kindle Edition, p.23.

- Analytics engineers, data engineers, data scientists, and/or machine learning engineers who will implement data solutions
- Data-related professionals and developers who are aspiring data leadership and product management

The book does not presume a sophisticated agile and project management background. However, by its very nature, the material is grounded on its concepts.

The goal is to impart a significant understanding of practical data product management and, more importantly, give practical guidelines about how to perform it.

To address this, the book covers the following.

- **Introducing data product management fundamentals**: You'll learn about the concepts of the products, what are data products, data as a product and the origin of everything, the analytical architecture journey and its issues, and the necessary knowledge to manage it.
- **Putting data product management into practice**: You'll learn new techniques I designed, such as the Golden Data Platform, used to manage data as a product through a data product. And you'll also learn about the data product management framework and how to use the canvas I created. Finally, you'll learn how to put everything into practice, from the designing of the solution to the ongoing of it.

By the end of the book, you will be able to explain to your leaders, peers and team why data products and especially data as a product are important for the business. Not just explain it, but put it into practice and deliver value to the business in this new era!

Although it's not mandatory, I recommend you read the chapters sequentially for a deeper understanding of the topics approached and a gradual learning path.

PART I

Fundamentals for Data Product Management

CHAPTER 1

Introduction to Data Product Management

In this chapter, you'll learn the main concepts of data product management.

I'll explain what a product is and define data management, data governance, artificial intelligence, data products, data as a product, and how everything is related.

This chapter contains fundamental concepts that many experienced professionals still don't know and are fundamental for any seniority of professionals.

1.1. Definitions

Before talking about product definitions, it is fundamental to understand the definition of a problem. After all, products are created upon them. At least, they should be.

According to Collins Dictionary: "A problem is a situation that is unsatisfatory and causes difficulties for people."

Dictionary.com describes a problem as "1. any question or matter involving doubt, uncertainty, or difficulty. 2. a question proposed for solution or discussion."

J. Milhomem, *Data Product Management in the AI Age*,
https://doi.org/10.1007/979-8-8688-1315-3_1

As you can see, the problem definition is also an invitation to create a solution, which is the main core of the product's definition.

Now, let's use the same approach and evaluate the definition of a product through the dictionary!

Dictionary.com describes the product definition as follows: "1. a thing produced by labor. 2. a person or thing produced by or resulting from a process, as a natural, social, or historical one; result. 3. the totality of goods or services that a company makes available; output."

Scrum,[1] one of the most important agile[2] frameworks today, states, "A product is a vehicle to deliver value. It has a clear boundary, known stakeholders, and well-defined users or customers. A product could be a service, a physical product, or something more abstract."

This is the foundation concept for this book.

I enjoy the concept description the most because it represents the core of the product's expectations and what every work result should deliver: *value*—the foundation of everything I've been discussing since this book's beginning!

It means that a product is also measurable. It must enable us to measure the result of a problem solved by the product, even if it is a complex job to be done!

With that in mind and recapping the initial discussion, the main principle of a product is **to solve a problem for someone**. For this reason, it is important to be grounded in two main pillars when designing a product.

- The problem: Map the issue and its full context.
- The target audience: Map the interested public in the solution and its motivations.

---

[1] Simon Kneafsey (2014), "What is a product?" Scrum.org. Available at `https://www.scrum.org/resources/blog/what-product#:~:text=%E2%80%9CA%20product%20is%20a%20vehicle,,%20or%20something%20more%20abstract.%E2%80%9D&text=This%20is%20an%20intentionally%20broad,and%20applicable%20in%20multiple%20contexts`

[2] Agile is a set of methods and practices where solutions evolve through collaboration between self-organizing, cross-functional teams.

1.2. Data Products

Products exist because there are opportunities. In other words, there are issues/problems to be solved. It also applies to the data world.

In 2021, Gartner, one of the biggest consulting companies around the world that evaluates technologies solutions for decision-making and strategies of companies, shared that "Poor data quality costs organizations an average of $12.9 million. Apart from the immediate impact on revenue, over the long term, poor quality data increases the complexity of data ecosystems and leads to poor decision-making."[3]

In 2024, Fivetran and Vanson Bourne performed new research[4] with more than 500 companies from the United States, Germany, France, the United Kingdom, and Ireland that are operating in the private and public sectors. The results show that although 97% of the companies are investing in artificial intelligence (AI)/machine learning (ML) models in the next two years, 81% of them trust their AI/ML outputs, admitting they have data foundation inefficiencies, which are necessary to improve their business outcomes.

The data issues come from technical and non-technical aspects.

The following are technical aspects.

- Disorganized and siloed data
- IT infrastructure outdated
- Low quality of data
- Stale data
- Access to the data

[3] Manasi Sakpal (2021), "How to improve your data quality." Garter.com. Available at https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality

[4] Fivetran (2024), *Fivetran + Vanson Bourne report: AI in 2024.* Available at https://www.fivetran.com/resources/reports/fivetran-vanson-bourne-report-ai-in-2024

The following are non-technical aspects.

- Lack of buy-in and support from senior teams
- Lack of internal skills

Consequently, 40% of the companies admitted they had experienced data inaccuracies, hallucinations, and data biases in their AI outputs.

Moreover, in financial aspects, underperforming AI programs/models built on inaccurate or low-quality data cost 6% of global annual revenue, which represents approximately $336 million a year, based on data from 550 respondent organizations with more than $25 million annual revenue and an average of $5.6 billion.

Data quality becomes a key barrier to good business results. Especially as organizations reach the advanced stage, in a nutshell, for AI models to produce the best output and create impact for the business, generating the necessary value result, the data used must be of the foremost quality.

Data Management Fundamentals

Before talking about the types of data issues, it is critical to understand some data management fundamentals.

Unfortunately, during my almost two decades of experience, I'm used to seeing lack of knowledge as a huge deficiency for mid/senior data professionals (managers and individual contributors). This is one of the reasons for this book.

These concepts are fundamental not just for operational and tactical executions of any project, initiative, or product but for strategic thinking. After all, managing such important assets as the company's data requires a lot of strategic thinking and designing. At least, it should.

What Is Data Management?

In the 1950s, database theory and usability started with the data storage's existing solutions. The technologies kept evolving until the invention of relational databases in the 1970s, which were used as repositories until the present.

In the 1980s, databases started to be used for analytical purposes (traditionally nominated as decision support systems) as database marketing[5] designed to innovate and improve direct marketing[6] because it was previously done offline with strong manual effort.

Sequentially, the data warehouse concept emerged, engaging more business areas to leverage data organization and usability, and so on.

Due to the power of data, companies have started to recognize their data as a vital enterprise asset, enabling innovations and strategic achievements. Due to its importance, it became inefficient for the business to manage data in an ad hoc manner based on which kind of value it could derive. For this, it was necessary to change the way to manage data: with focus, intention, planning, leadership, and commitment.

According to DAMA International:[7] "Data management is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their life cycles."

Thus, data management is the processes and procedures required to manage data. Its main goal is to manage data as a valuable asset.

---

[5] Database marketing is a sort of direct marketing in which a company's client databases are utilized to create customized email lists for marketing campaigns, a practice known as customer segmentation.

[6] Direct marketing relies on one-on-one communication with a target audience. It includes tools like emails, phone calls, catalog marketing, and text messages.

[7] DAMA International, (2015), *Data DMBook*, Technics Publications.

With the data management necessity emerging, the DAMA-DMBoK (DAMA International's Guide to the Data Management Body of Knowledge) framework was designed in 2009 to enable structured data management, which received updates in 2015 to englobe new needs related to big data technologies.

Data Governance

Piethein Strengholt (2020)[8] says that data governance "consists of activities for implementing and enforcing authority and control over data management, including corresponding assets."

Data governance requires a larger engagement of the organization, always based on three main dimensions (people, technology, and processes) to put any aspect of data governance (knowledge areas) in place.

- **Roles and responsibilities (people)**: It's related to responsibilities regarding human aspects, such as legal definitions, ethical trade-offs, and social and economic considerations. It's also related to the roles of professionals focused on data management, which consists of the following.

- **Data owner** (also known as business owner): Accountable for the data of a specific process/domain. This role is responsible for data definitions, data quality, classifications, usability purposes, and so on.

    - **Data steward**: This professional ensures that the data policies and solutions comply with the definitions of data standards.

[8] Piethein Strengholt (2020), *Data Management at Scale.* O'Reilly Media, p. 185.