Data Mesh

Principles, patterns, architecture, and strategies for data-driven decision making

Pradeep Menon

www.bpbonline.com

First Edition 2024

Copyright © BPB Publications, India

ISBN: 978-93-55519-962

All Rights Reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means or stored in a database or retrieval system, without the prior written permission of the publisher with the exception to the program listings which may be entered, stored and executed in a computer system, but they can not be reproduced by the means of publication, photocopy, recording, or by any electronic and mechanical means.

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

To View Complete
BPB Publications Catalogue
Scan the QR Code:

www.bpbonline.com

Kup książkę

Dedicated to

My beloved wife ***Archana****,*
My charming daughter ***Anaisha****, and*
My handsome pet buddy ***Pablo***

Kup książkę

About the Author

Pradeep Menon is an accomplished technology professional with over 20 years of extensive expertise in Data, AI, Analytics, and Cloud Computing. Currently serving as the CTO for Digital Natives in ASEAN at Microsoft, Pradeep is pivotal in spearheading the adoption and strategic implementation of Generative AI across the region. His career highlights a robust background with roles at Microsoft and Alibaba Cloud, where he successfully led major initiatives in data and AI, greatly enhancing business strategies and operational efficiency across Asia.

Pradeep's approach seamlessly integrates high-level strategic discussions with C-suite executives and detailed technical implementations, making him a key figure in driving digital transformation. His technical and strategic acumen has resulted in significant revenue growth and enhanced competitive positioning for numerous enterprises.

A thought leader and visionary, Pradeep's contributions extend beyond corporate borders. He is the acclaimed author of "Data Lakehouse in Action" and a revered voice on the international speaking circuit, illuminating pathways in technology with his insights. His academic credentials— an MS in Business Analytics from NYU Stern and an MBA from Strathclyde—marry technical prowess with strategic insight, underscoring his holistic approach to innovation and leadership in the digital age.

Kup książkę

About the Reviewer

Rajesh Ghosh is a solutions engineer and data enthusiast whose extraordinary journey has transformed him into a thought leader in his field. With a knack for innovative problem-solving and a passion for empowering data-driven decisions, Rajesh has spearheaded transformative initiatives that have modernized critical information technology systems and advanced data engineering and analytics capabilities across organizations. His expertise in data engineering and architecture has earned him widespread recognition and respect within the industry.

Kup książkę

Acknowledgement

I am deeply grateful to my family and friends for their unwavering support throughout the creation of this book. Special thanks to my wife, Archana, and my daughter, Anaisha, whose love and patience have been my anchor, and to Pablo, my loyal pet, whose companionship has brightened many writing sessions.

I extend heartfelt thanks to BPB Publications for their expertise and guidance in bringing this project to fruition. Their dedication throughout the journey of revising and perfecting this book has been invaluable.

I also owe a tremendous debt of gratitude to my colleagues and customers at Microsoft. Their willingness to collaborate, share insights, and apply these principles in the field of data architecture has profoundly shaped this work.

To all the readers and supporters of this book, your encouragement and engagement mean the world to me. Thank you for your enthusiasm and for believing in the value of this work.

Each of you has contributed to making this book not just a collection of pages but a vibrant, living dialogue on data architecture. I am sincerely thankful for every contribution, conversation, and word of encouragement that has turned this vision into reality.

Kup książkę

Preface

In the rapidly evolving world of data management, the shift from traditional centralized architectures like data lakes and warehouses to a decentralized, domain-oriented approach marks a revolutionary change. This book dives deep into this transformative concept known as Data Mesh, which redefines how data is handled across organizations. This book is crafted for data professionals eager to understand and implement a structure that promotes agility, scalability, and resilience within their data ecosystems.

Data Mesh represents a paradigm shift, focusing on treating data as a product and emphasizing decentralized governance. This approach aligns closely with the needs of modern businesses that require rapid access to diverse, distributed data sources. By breaking down the traditional silos, Data Mesh enables a more collaborative and flexible data management environment. This book is designed not only to introduce the concept but also to provide a detailed guide on implementing Data Mesh effectively.

This book embarks on a comprehensive exploration of Data Mesh, guiding readers through the transformative shift from traditional centralized data architectures to a decentralized, domain-oriented framework. The journey begins by establishing a contextual foundation for Data Mesh, followed by a historical overview of data architecture evolution, highlighting the necessity for such an innovative approach. As the chapters progress, readers delve into the core principles and patterns of Data Mesh, gaining insights into how it fosters agility, scalability, and resilience in data management. The book then navigates through the practical aspects of implementing Data Mesh, covering data governance, cataloging, sharing, and security, each treated with depth and precision to facilitate understanding and application. Finally, the book culminates with practical examples and real-world applications, illustrating how to operationalize Data Mesh effectively within various organizational contexts. This structured journey equips data professionals with the knowledge to not only understand but also implement Data Mesh to enhance their data management practices and stay ahead in the rapidly evolving data landscape.

By the conclusion of this book, readers will not only grasp the theoretical underpinnings of Data Mesh but will also be equipped with practical knowledge and strategies to implement these concepts in their day-to-day operations. Whether you are a seasoned data architect, a Chief Data Officer, or a curious analyst, This book offers valuable insights and guidelines that will help you stay at the forefront of data management technology. This book is your

Kup książkę

comprehensive guide to navigating the complexities of modern data architectures and leveraging the full potential of Data Mesh to drive business value.

Chapter 1: Establishing the Data Mesh Context – This chapter introduces the Data Mesh concept by delineating its need within modern data management paradigms. It sets the stage by describing the shift from centralized systems to a more fluid, decentralized architecture, explaining how this approach aligns with the demands of big data and agile enterprises.

Chapter 2: Evolution of Data Architectures – This chapter traces the development of data architectures from traditional databases and data warehouses to modern data lakes and beyond. It highlights the limitations of earlier systems and sets the rationale for the adoption of Data Mesh, presenting a historical perspective that underscores the evolution toward decentralized data domains.

Chapter 3: Principles of Data Mesh Architecture - This chapter delves into the core principles that define the Data Mesh framework. It explains each principle in detail, providing the theoretical foundation necessary for understanding and implementing Data Mesh.

Chapter 4: The Patterns of Data Mesh Architecture – This chapter explores various architectural patterns within Data Mesh, including decentralized topologies and hybrid models. It offers guidelines on how to select and implement these patterns based on specific organizational needs and data strategies.

Chapter 5: Data Governance in a Data Mesh - This chapter discusses the unique challenges and solutions for governing data in a decentralized context. It covers strategies for maintaining data quality, managing metadata, ensuring compliance, and aligning data governance with organizational goals within the Data Mesh framework.

Chapter 6: Data Cataloging in a Data Mesh - This chapter focuses on effective data cataloging practices that enhance the discoverability and usability of data across decentralized domains. It details the strategy, processes, and tools for building a comprehensive data catalog that supports the Data Mesh's collaborative and agile nature.

Chapter 7: Data Sharing in a Data Mesh - This chapter examines the topologies for secure and efficient data sharing across different domains within a Data Mesh. It provides insights into designing data-sharing strategies that balance autonomy with oversight, which is crucial for fostering an integrated yet flexible data environment.

Chapter 8: Data Security in a Data Mesh - This chapter addresses the critical aspects of securing a decentralized data architecture. It lays out the detailed framework for data

Kup książkę

security Data Mesh environments that covers the organization, inter-domain, and intra-domain security.

Chapter 9: Data Mesh in Practice - This chapter culminates the learnings from all previous chapters, synthesizing the principles, patterns, governance, cataloging, sharing, and security strategies into a cohesive framework for implementing Data Mesh in practice. It lays out step-by-step guidelines for operationalizing Data Mesh within various organizational contexts, providing a comprehensive roadmap that translates theoretical concepts into actionable strategies.

Kup książkę

Coloured Images

Please follow the link to download the
Coloured Images of the book:

https://rebrand.ly/e8b279

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At **www.bpbonline.com**, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

Kup książkę

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Kup książkę

Table of Contents

1. Establishing the Data Mesh Context 1

Introduction 1

Structure 1

Objectives 2

Navigating the modern data landscape 2

Data Warehouses 3

Data Lakes 4

Data Lakehouse 5

Need for a macro data architecture pattern 6

Conclusion 7

Key takeaways 7

2. Evolution of Data Architectures 9

Introduction 9

Structure 10

Objectives 10

Era of monolithic data architecture 11

Birth of Relational Database Management System 11

Monolithic systems: Rise and challenges 11

Era of Data Warehouses 14

Decoupling analytics and online transaction processing 14

Inmon vs. Kimball: Divergent approaches 15

Challenges in the Enterprise Data Warehouse paradigm 17

The perfect storm 18

The exponential growth of data 19

The increase in computing power 20

The decrease in storage cost 20

The rise of cloud computing 20

Kup książkę

The advancements in Artificial Intelligence ... 21
Paving the way to the era of Data Lakes ... 21
The era of Data Lakes ... 22
Origins of the Hadoop ecosystem ... 22
Key components of the Hadoop ecosystem ... 23
The Data Lake architecture pattern ... 24
Benefits of Data Lake over the traditional EDW pattern ... 26
Challenges of the Data Lake pattern ... 27
From Data Lake to Data Swamp ... 27
The evolution of the Data Lakehouse architecture pattern ... 27
The era of Data Lakehouses ... 28
Symbiotic rise of cloud computing and the Data Lakehouse ... 28
Data Lakehouse pattern ... 29
Adoption of Data Lakehouse ... 30
Challenges with the Data Lakehouse architecture ... 31
Introduction to Data Mesh ... 32
Conclusion ... 33
Key takeaways ... 33

3. Principles of Data Mesh Architecture ... 35

Introduction ... 35
Structure ... 35
Objectives ... 36
Understanding domains and nodes ... 36
Domain ... 36
Central unit ... 37
Subunits ... 37
Node ... 39
The interplay between domains and nodes ... 40
Foundations of the principles ... 42
The overarching goal: The balance between governance and flexibility ... 42
The architectural principles ... 44

Kup książkę

Methodology for examining the principles 45
Principle 1: Domain-oriented ownership 46
Aspects of the principle of domain-oriented ownership 47
Complete lifecycle ownership 47
Context preservation in data management 48
Decentralized governance to enhance data quality 48
Business alignment and domain autonomy 49
Seamless cross-domain interoperability 50
Rationale for the principle of domain-oriented ownership 51
Overcoming organizational silos with domain-oriented ownership 51
Cultivating responsibility through domain-oriented ownership 51
Augmenting agile responses with domain-oriented ownership 52
Enriching data insights and intelligence through domain diversity 53
Facilitating organizational learning 53
Implications of the principle of domain-oriented ownership 54
Realigning roles and responsibilities for data ownership 54
Creating a resilient operational framework through data decentralization 55
Enhancing data intelligence and value creation across the organization 55
Revising data governance policies for domain diversity 56
Decentralizing budget allocation for data ownership 57
Principle 2: Reimagining data as a product 57
Aspects of the principle of reimagining data as a product 58
Redefining data products as first-class citizens 59
Aligning data products with business domains and use cases 59
Ensuring discoverability, accessibility, and compliance of data products 60
Ensuring reliability, consistency, and interoperability of data products 60
Continuous feedback and iterative improvement 60
The rationale for the principle of reimagining data as a product 61
Empowering domain teams to manage their own data products 61
Breaking down silos and enhancing quality 62

Kup książkę

Managing data products with ownership and lifecycle 62
Enhancing data consumer experience with data products 63
Strategically leveraging data assets 63
The implication of the principle of reimagining data as a product 64
Redefining data roles with data products 64
Transforming data management with agile, lean, and DevOps practices 65
Facilitating technological innovation for data products 65
Empowering data consumers with data products 65
Facilitating enriched insights through cross-domain collaboration 66
Principle 3: Empowering with self-serve data infrastructure 66
The aspects of the principle of empowering with self-serve data infrastructure 68
Fostering a decentralized data infrastructure 68
Leveraging platform thinking 69
Adopting self-service tools 69
Pivoting toward a domain-driven design 70
Creating an agile self-serve data infrastructure using DataOps 71
Rationale for the principle of empowering with self-serve data infrastructure 72
Accelerating data value with decentralized empowerment 72
Enhancing business agility with rapid data product development 72
Achieving data scalability and resilience with distributed architecture 73
Promoting resource efficiency and cost-effectiveness 74
Enhancing cross-functional collaboration 74
Implication of the principle of empowering with self-serve data infrastructure 75
Seamless integration of tools and platforms 75
Empowering teams through training and skill development 76
Ensuring data security and compliance 76
Building a resilient data architecture 77
Enhanced data discovery and accessibility 77
Conclusion 78
Key takeaways 79

Kup książkę

4. The Patterns of Data Mesh Architecture 81

Introduction 81

Structure 82

Objectives 82

Data mesh component model 82

Domain 83

Domain node 84

Data Catalog 85

Data Share 87

Bringing it all together as a domain unit 88

Fully governed data mesh architecture 89

Fully federated data mesh architecture 92

Hybrid data mesh architecture 95

Domain placement methodology 97

Functional context 98

People and skills 99

Regulations 99

Operations 99

Technical capabilities 100

Methodology in action 100

Parameter weightage 101

Parameter score 101

Conclusion 102

Key takeaways 102

5. Data Governance in a Data Mesh 105

Introduction 105

Structure 106

Objectives 106

Importance of data governance 106

Vitality of data governance in a data mesh paradigm 106

Consequences of Lax in governance in a data mesh 107

Kup książkę

Traditional data governance: A centralized approach 108
Data governance in other architectural patterns 108
Challenges of traditional governance in the data mesh framework 110
Data mesh governance framework 111
The governance goals 113
The seven objectives 113
The three governance components 114
Organizational bodies and roles 115
Key roles and interactions 117
Data governance processes 118
Data product definition 119
Data product cataloging 120
Data product quality assurance 121
Data product security 122
Data sharing 123
Data governance policies 124
Data product policies 125
Data cataloging policies 126
Data sharing policies 127
Conclusion 129
Key takeaways 130

6. Data Cataloging in a Data Mesh 131

Introduction 131
Structure 132
Objectives 132
The role of data cataloging 132
Data cataloging as a means of data utility 134
Data cataloging as a means of data governance 135
Principles of data cataloging 136
Developing a data cataloging strategy 138
Step 1: Defining the scope and objectives 138
Step 2: Assessing the current state and gaps 139

Kup książkę

Step 3: Designing the desired state and roadmap *141*
Implementing the data cataloging strategy 142
Understanding the domain *143*
Establishing the domain structure *145*
Identifying cataloging elements *146*
Cataloging the domain *149*
Monitoring catalog usage and effectiveness *150*
Conclusion 151
Key takeaways 152

7. Data Sharing in a Data Mesh 155

Introduction 155
Structure 156
Objectives 156
Role of data sharing 156
Information dissemination *157*
Data value creation *158*
Principles for data sharing *159*
Domain data autonomy *159*
Data interoperability *160*
Contextual data sharing *161*
Quality-first approach *161*
Collaborative data stewardship *162*
Patterns for data sharing *163*
Publish-subscribe pattern *163*
Request-response *166*
Push-pull *168*
Implementing the data-sharing strategy 171
Step 1: Identifying appropriate data sharing pattern *172*
Step 2: Establishing the data sharing protocol *174*
Step 3: Creating secure infrastructure and access control interfaces *175*
Step 4: Monitoring and performance optimization *177*
Conclusion 177
Key takeaway 178

Kup książkę

8. Data Security in a Data Mesh 179

Introduction 179
Structure 180
Objectives 180
Security challenges in a decentralized system 181
Challenge 1: Data privacy across domains 182
Challenge 2: Unauthorized data access 182
Challenge 3: Data integrity and consistency 183
Challenge 4: Network security in a distributed environment 184
Challenge 5: Scalability of security measures 184
SECURE: Principles of data mesh security 185
S: Scalable Security Protocols 186
E: Encryption and Secure Data Transfer 187
C: Consistent Data Integrity Checks 188
U: Unified Access Control 189
R: Robust Privacy Standards 190
E: End-to-End Data Protection 191
Data Mesh Security Strategy: The three-circle approach 191
Circle 1: Organizational security 193
Circle 2: Inter-Domain Security 196
Circle 3: Intra-Domain Security 200
Components of Data Mesh Security 202
Data security component 203
Data encryption 203
Data masking 204
Data backup 205
Data classification 206
Network security 208
Firewall 208
Virtual Private Network 209
Intrusion detection system 209
Transport Layer Security 210
Public Key Infrastructure 211

Kup książkę

Access management *212*
Authentication *212*
Authorization *213*
Key management *214*
Access audit and compliance *216*
Conclusion 217
Key takeaways 218

9. Data Mesh in Practice 219
Introduction 219
Structure 220
Objective 220
Domain-Architecture-Operations overview 220
Domain: The foundation 222
Step 1: Define the Domain *223*
Step 2: Domain placement *225*
Step 3: Define the Domain Node *227*
Architecture: Building the blueprint 228
Step 1: Create the domain data cataloging strategy *229*
Step 2: Define the domain data sharing pattern *231*
Step 3: Define the Data Mesh security strategy *233*
Operations: From blueprint to action 236
Step 1: Establish the governance structure *237*
Step 2: Data Mesh technology selection *239*
Step 3: Operationalizing the Data Mesh *241*
A measured approach: Tracking progress and impact *241*
The feedback loop: Continuous improvement through learning *242*
Conclusion 243
Key takeaways 244

Appendix: Key terms 245

Index 255-261

Kup książkę

Chapter 1
Establishing the Data Mesh Context

Introduction

Decades ago, *Clive Humbly*, a respected mathematician and data science pioneer, stated, *Data is the new oil*. Today, his words hold even greater significance as we are in a data-driven era where effective data management has become a critical aspect of transformation.

In the digital age, data has emerged as one of the most valuable assets for organizations worldwide. In this chapter, we embark on a journey through the intricate maze of the modern data landscape. We begin by navigating the contemporary data ecosystem and understanding its complexities and challenges. From the structured realms of Data Warehouses to the vast expanses of Data Lakes and the hybrid environment of the Data Lakehouse, we explore each architecture's nuances, strengths, and limitations. As we progress, we recognize the growing need for a more encompassing solution – a macro data architecture pattern. This pattern seeks to address the unique challenges extensive and multifaceted organizations face in today's data-driven world. Join us as we unravel the intricacies of these architectures and pave the way for a more holistic approach to data management.

Structure

In this chapter, we will introduce the following:

- Navigating the modern data landscape.
- Need for a macro data architecture pattern.

Kup książkę

Objectives

The primary objective of this chapter is to provide readers with a foundational understanding of the contemporary data landscape. We aim to demystify the core architectures that dominate today's data management practices, from the structured world of Data Warehouses to the expansive domains of Data Lakes and the integrative approach of Data Lakehouses in subsequent chapters. By exploring these architectures, we highlight their merits and challenges. Furthermore, we underscore the emerging need for a macro data architecture pattern, emphasizing its significance in addressing the complexities of large-scale data management.

Lastly, this chapter serves as a precursor to the deeper discussions in the subsequent chapters, offering a brief overview of the topics and insights. Through this chapter, we aspire to equip readers with a holistic perspective on modern data architectures and set the stage for the following comprehensive exploration.

Navigating the modern data landscape

Data management has become increasingly complex in today's digital world, with different patterns and structures emerging in analytics. This highlights the growing importance and intricacy of managing data. Among these patterns, four architectures have emerged as the most prevalent: **Data Warehouses**, **Data Lakes**, and the hybrid model known as the **Data Lakehouse**.

Different architectures have specific capabilities and purposes for managing and analyzing data. These architectures guide us through the vast and sometimes challenging world of data.

The following table summarizes the advantages and disadvantages of each of these architectural patterns:

Pattern	Key features	Advantages	Disadvantages
Data Warehouse	Subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.	• Integrated data • Improved data quality and consistency • Better decision-making	• Complexity and cost • Data latency • Limited flexibility

Kup książkę

Pattern	Key features	Advantages	Disadvantages
Data Lake	A centralized repository that allows you to store all your structured and unstructured data at any scale.	• Has better flexibility • Has better scalability • Relatively cost-effective	• Prone to becoming Data swamps. • Challenging Security implementations • Complexity to process unstructured data. • Requires greater governance.
Data Lakehouse	A unified platform for various data workloads, such as descriptive, predictive, and prescriptive analytics	• Provides greater flexibility as compared to other patterns. • Provides better performance as compared to other patterns. • Supports all types of analytics due to its unified approach.	• Prone to becoming Data swamps if governance is not in place. • Requires more organizational maturity. • More complexity due to scale and scope.

***Table 1.1**: Advantages and disadvantages of various architectural patterns*

In the upcoming sections, we will examine these architectures in greater detail, discussing their advantages, disadvantages, and relevance in today's data environment.

Data Warehouses

The concept of a Data Warehouse has been introduced previously. *Bill Inmon* first introduced it in the 1970s. He defined it as a **subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.** The idea was to create a central repository where data from various sources could be stored and analyzed. Over time, data warehousing has evolved with technological advancements, but the core concept remains the same.

A Data Warehouse is a centralized repository where data from various sources is consolidated, transformed, and stored. This data is typically structured and processed, making it suitable for analysis and reporting. Data Warehouses are used by organizations to support business intelligence activities, including data analytics, reporting, and decision-making. They provide a historical data view, enabling trend analysis and strategic planning.

Kup książkę