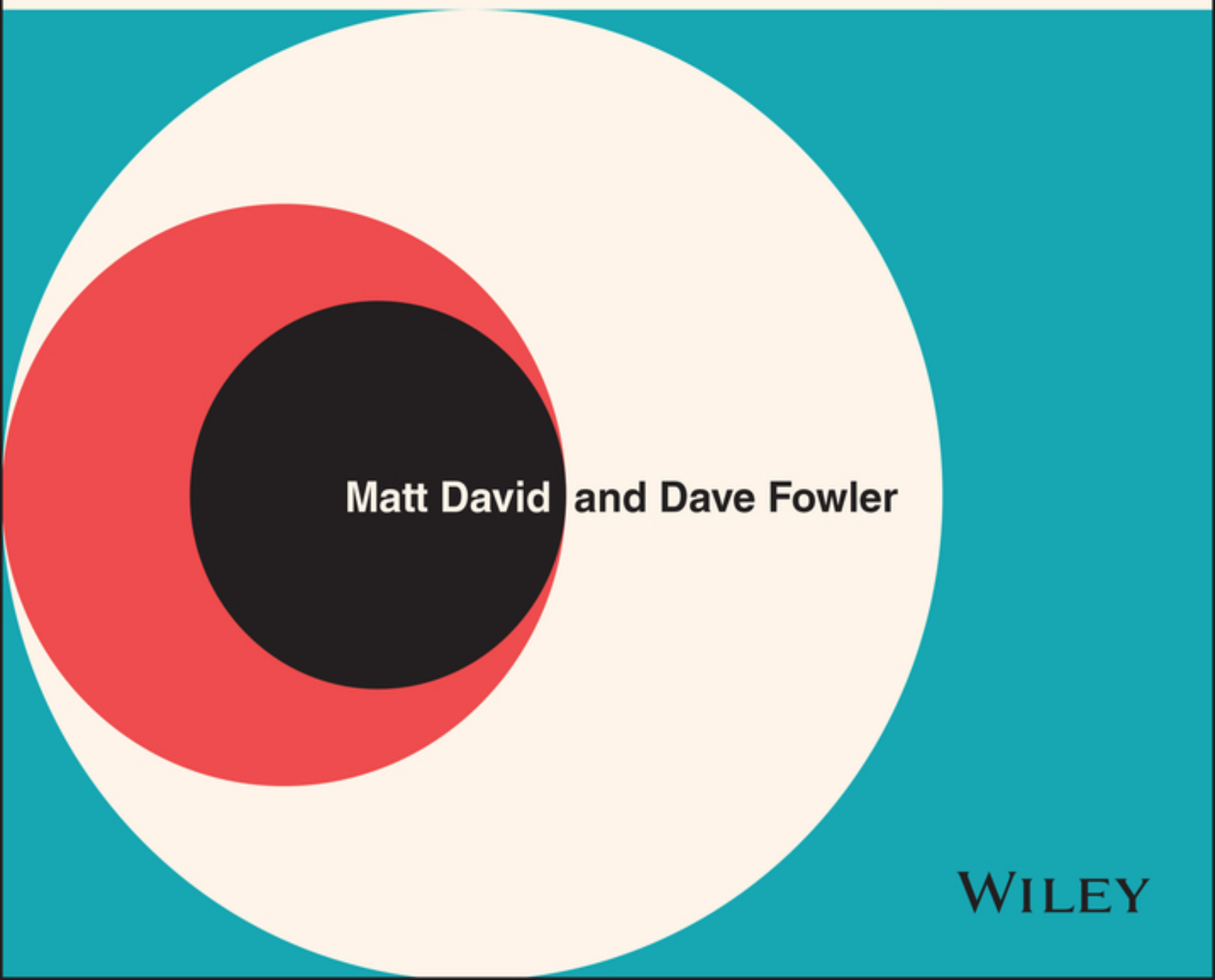
The Informed Company

How to build modern agile data stacks that drive winning insights



Matt David and Dave Fowler

WILEY

Table of Contents

List of Tables

List of Illustrations

The Informed Company

How to Build Modern Agile Data Stacks that Drive Winning Insights

Dave Fowler
Matt David

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data

Names: Fowler, Dave (Computer scientist), author. | Matt David, author.

Title: The informed company : how to build modern agile data stacks that drive winning insights / Dave Fowler, Matt David.

Description: Hoboken, New Jersey : Wiley, [2022] | Includes index.

Identifiers: LCCN 2021028324 (print) | LCCN 2021028325 (ebook) | ISBN 9781119748007 (paperback) | ISBN 9781119748021 (adobe pdf) | ISBN 9781119748014 (epub)

Subjects: LCSH: Data structures (Computer science) | Big data. | Cloud computing.

Classification: LCC QA76.9.D35 F69 2022 (print) | LCC QA76.9.D35 (ebook) | DDC 005.7/3—dc23

LC record available at https://lccn.loc.gov/2021028324

LC ebook record available at https://lccn.loc.gov/2021028325

Cover image: © Neo Geometric/Shutterstock

Cover design: Wiley

To my mother who continues to be my most supportive and patient teacher. As a software engineer you taught me to code for my sixth-grade science project. Today as a Data Analyst you helped a 38-year-old me in discussions and edits of this book. Thank you for always supporting and encouraging my curiosities and for all your love.

— Dave Fowler

I dedicate this book to my Mom, an educator who is fueled by helping others learn. Thank you for always believing in me and being an example of how much you can affect other people's lives.

— Matt David

About This Book

Why Write This Book

Most comprehensive books on analytics architecture that we've found are over a decade old, most of them pre-cloud. Because there really isn't a modern equivalent to Kimball's seminal *The Data Warehouse Toolkit*, today's data teams have to reinvent the principles of building a data stack. Too often, they do this without guidance. To solve this problem, we have created a best-practices guide for bootstrapping and nurturing a technologically current data warehouse.

Who This Book Is For

We wrote this book for whoever values data and believes that informed companies are competitive. It's a book for the working professional who is creating a practical, modern data stack. It's for the lone analyst or the professional embedded in a team. It's for anyone interested in what design practices underlie robust data architecture, the kind that equips entire companies with business intelligence insights. At its heart, this book is written with collaboration in mind (Figure A.1).

Figure A.1 Data management is a collaborative process.

Who This Book Is Not For

This book is not written for “big data” professionals. To be clear, even large corporations like Doordash, Discord, and the owners of *The Financial Times* and *The New York Times* (all previous customers of ours) do not qualify as big data companies. As a rule of thumb, the big data label applies to data architectures with raw input that exceeds 100 GB per day.

No doubt, many elements of this text map onto the big data workflow, especially since warehouses support all sorts of tables, not just, say, event streams. However, our aim is to focus on the central pillars of a modern data stack, so that the widest set of readers can readily benefit from the information herein. In this spirit, we forgo recommendations for mega-scale architectures.

This book is not for AI-enabled teams and does not cover AI workflows, machine learning models, or real-time operational use cases. Instead, its goal is to provide best practices for building and maintaining a robust data analytics stack (i.e. the analytics foundation on which an AI workflow can be built).

If you are a small business that can run everything with Quickbooks and Excel, that ability is great. Data is important for all companies, but if these tools are already serving you well, the book may not offer helpful guidance. If you start exceeding the data capacity of Excel or bring in a data source that needs to be in a database to be analyzed, then keep reading.

Who Wrote the Book

This book was written by Dave Fowler and Matt David.

Dave Fowler has worked in BI for over a decade, and has always looked for ways to `JOIN teams ON data`. He wants to enable any working professional (not just data analysts) to explore and understand their data. As the founder and CEO of Chartio, Dave has spent the last 11 years leading the development of a self-service BI product that aims to do just that. Chartio's suite of tools make it easy for anyone at a data-driven business to browse their schemas, merge various data sources, and produce beautiful dashboards. In March 2021, Atlassian acquired Chartio and is integrating it into their platform.

Matt David has worked in product management and education for eight years. As data becomes a necessary skill for more and more jobs, he passionately advocates for data literacy among the workforce. As the current head of The Data School, he oversees the production of free, online resources focused on leveraging data within companies. Recent book topics include SQL optimization, data governance, and common analysis biases. Dave started The Data School, and together he and Matt have grown it into an important free resource for the data community. He previously worked at Udacity and General Assembly teaching analytics.

Dave and Matt decided to co-write this book after seeing how many people struggle when constructing data stacks and then trying to use them. This book was created with the support of many employees at Chartio. They graciously provided insights into how customers model their data and collected frequently asked data-infrastructure questions. Their contributions guided the production of this text.

Who Edited the Book

This book was reviewed and edited by Emilie Schario, Mila Page, and David Yerrington. Emilie is the head of data at

Netlify and previously helped build Gitlab's entire data organization. She regularly writes and speaks on all things related to modern data. Mila is a developer relations advocate at dbt Labs, the makers of dbt (data build tool). She helps data professionals learn and apply modern analytics-engineering practices, and is an organizer for Coalesce, the dbt Community's annual conference. David is a Data Science Consultant and was the Global Lead Data Science Instructor at General Assembly. He helps people around the world better leverage their data. Emilie, Mila, and David have shaped the narrative and content of this book. Their (sometimes) line-by-line feedback has ensured that we can proudly stand behind our recommendations.

Influences

We've drawn on several sources of information and opinion when writing this text. While at Chartio, we worked with hundreds of modern cloud-based customers. We've collected, implemented, and refined these practices ourselves, and through writing this book, vetted them further with partners and customers. We've also learned from the data community through dataschool.com, blogs like Tristan Handy's, and data-focused slack communities.

And lastly, it's worth noting and thanking some classic books that informed the previous generation of warehousing toolkits. We honor them by echoing their terminology and best practices wherever possible:

- *Agile Data Warehouse Design* by Lawrence Corr
- *The Data Warehouse Toolkit* by Ralph Kimball
- *Information Dashboard Design* by Stephen Few (my review here)

How This Book Was Written

This book originates in part from a project within The Data School (Figure A.2), a collection of free online books and interactive tutorials on managing and leveraging data (see dataschool.com). These resources are always expanding, much like the articles of Wikipedia: each round of updates sees our ebooks cover additional topics, go deeper on established ideas, share more real-world examples, and better deliver that content. Our goal is to maintain and improve these resources and keep them modern.

Figure A.2

Source: The Data School

Few are complete “experts” in all of the areas of modern data governance, and the landscape is changing all of the time. If you have a story to share, or a chapter you think is missing, or a new idea, email us. Even if you don't know what specifically to share, but you don't mind sharing your story, please reach out as we are particularly interested in adding real-world experiences and insights.

There is already too much jargon in the data world, often created by talented vendor marketing teams. We try to stick with the most common and straightforward words

that are already in use. For any jargon we do find necessary, we include a definition.

There are many books for the old ways of working with data. We're highlighting current best practices here, so we ignore outdated terminology and techniques. In a few cases where it is beneficial to talk about industry evolution—like the change from ETL to ELT—we teach ELT and discuss the choice in a separate chapter.

Almost every part of this book could be contentious to someone, in some use case or to some vendor. In writing this book, it is tempting to bring up the caveats everywhere and write what would ultimately be a very defensive and overly explained book. We believe this type of book is way less useful for people seeking straightforward advice. Where we have a strong opinion, we don't argue it; we just go with it. Where we think the user has a legitimate choice to make, we pose those options.

This book aims to provide a broad overview and general guidelines on how to set up a data stack. We intentionally gloss over the details of launching a Redshift instance, writing SQL, or using various BI products. That would clutter the text, repeat what's already on the internet, and make the read quite stale.

How to Read This Book

The book starts with a quick overview and decision charts about what the stages are and what stage is appropriate for you. This book is structured with a section for each of the four stages, and if you'd like, you can jump ahead to the stage you're at.

Not every company needs the entirety of this book. As a growing company's data needs expand, more and more of the book becomes valuable. Note, though, many best

practices presented at each stage appear when they start to be relevant. These practices assume they are useful from the point they appear in the book, onward, to avoid redundancy. So it may benefit you to at least skim those earlier stages even if you and your company are further ahead.

At the end of the book we have a section where we describe what has changed in the data world that makes this new architecture relevant and performant. We avoid explaining how our recommendations differ from previous practices like Kimball Dimensional modeling so as not to clutter the experience. Such discussions are necessary, however, and we've put them in this last section of the book.

Lastly, throughout the book you will see the following icons:

Definitions

They are related to a term found on the same page. For example, on this page, the term “data lake” is mentioned. A data lake is a staging area for several data sources.

Protips

Protips expand on an idea or provide additional information about a topic related to what you read within a given chapter.

Foreword

In 2015, I used a product called Amazon Redshift. At the time, I had spent the prior 15 years of my career in a variety of roles all centered around their use of data, from analytics to marketing to operations. And while I considered my data competency my biggest professional differentiator, I had also become deeply frustrated. For all of the supposed progress in the data ecosystem, it was still slow, hard, and expensive to get insights out of data.

But my first experience with Redshift is where that all changed for me. I have such a visceral memory of the first hour I spent with the product: queries I ran returned so fast that it seemed like absolute magic. I had spent years and years of my career writing queries and waiting for the MacOS “spinner” icon to stop spinning. Now, all the sudden, these same queries weren’t 20% faster…they were 10 to 100 to 1000x faster. I felt like I had superpowers.

I'll let Dave and Matt actually explain how the modern data warehouse can achieve these types of performance results, but for now, just trust me that it can and does. Given that, the fascinating question is actually: *what does this mean for people like you and me?*

What kind of “people” do I mean? You know—people who are involved in making decisions at companies and want to use data to help us. People who likely over the years have acquired a variety of data skills, whether that's Excel VLOOKUPs, Google Analytics dashboards, SQL, or any of a thousand other options. People who have always felt like *it shouldn’t be so hard to do basic stuff* when it comes to data (but of course, for some reason, it always has been).

What I've come to realize in the years since my initial experience with Redshift is that the modern data stack has exactly two very important impacts for us:

1. With this far better tooling, we can grow our impact on the organizations we work for *dramatically*. I'm going to say something really stupid and obvious, but: **if your queries return 100x faster, you can know 100x more stuff.** And that means that you'll just be tremendously more valuable in the insights you're able to provide and the decisions you can make.
2. As a result of #1, our career options are just far, far greater. When I started my career, “data analyst” was a junior position that you attempted to graduate out of as quickly as possible to move onto other things. Now data analysts are high-leverage, strategic employees with earning potential that mirrors that of software engineers.

There really has been a change in what a single analyst is able to accomplish with some fairly simple tooling and some very accessible knowledge and skills. That single individual can now construct an entire sophisticated data infrastructure by plugging together some inexpensive off-the-shelf tools, and then get to work getting insights. Or, that individual can operate as an integral part of a large and sophisticated data team, forming a part of the nervous system for the modern enterprise. The same technology, same skill sets, same knowledge is required regardless how big or small the organization.

This is the modern data stack. To harness its power, you need to know SQL (learnable a day!) and you need to know some basic best practices that folks like the authors of this book have been deep in the weeds developing and recording over the past half-decade. It's surprisingly

achievable—one of the magical things about these tools is just how little advanced technical knowledge is required.

This book provides the foundational knowledge you'll need to navigate the modern data stack. From there, you'll be able to dive in yourself, get your hands dirty, and start asking questions of your fellow practitioners. And when you do, make sure to head over to join me in the dbt Community—my Slack handle there is @tristan and I'd love to hear from you!