Community Experience Distilled

HBase Design Patterns

Design and implement successful patterns to develop scalable applications with HBase

Mark Kerzner Sujee Maniyam

[PACKT] PUBLISHING open source* community experience distilled

HBase Design Patterns

Design and implement successful patterns to develop scalable applications with HBase

Mark Kerzner

Sujee Maniyam

[PACKT] PUBLISHING

open source community experience distilled

BIRMINGHAM - MUMBAI

HBase Design Patterns

First published: December 2014

Production reference: 1171214

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78398-104-5

www.packtpub.com

Cover image by Asher W (wishkerman@hotmail.com)

Credits

Authors

Mark Kerzner

Sujee Maniyam

Reviewers

Ricky Ho

Raghu Sakleshpur

Sergey Tatarenko

Commissioning Editor

Amarabha Banerjee

Acquisition Editor

Nikhil Karkal

Content Development Editor

Anand Singh

Technical Editors

Tanvi Bhatt

Rikita Poojari

Copy Editors

Dipti Kapadia

Shambhavi Pai

Neha Vyas

Project Coordinator

Akash Poojary

Proofreaders

Simran Bhogal

Ameesha Green

Paul Hindle

Indexer

Monica Ajmera Mehta

Graphics

Abhinash Sahu

Production Coordinator

Conidon Miranda

Cover Work

Conidon Miranda

About the Authors

Mark Kerzner holds degrees in law, math, and computer science. He has been designing software for many years and Hadoop-based systems since 2008. He is a cofounder of Elephant Scale LLC, a big data training and consulting firm, as well as the co-author of the open source book *Hadoop Illuminated*. He has authored other books and patents as well. He knows about 10 languages and is a Mensa member.

I would like to acknowledge the help of my colleagues, in particular Sujee Maniyam, and last but not least, of my multitalented family.

Sujee Maniyam has been developing software for 15 years. He is a hands-on expert of Hadoop, NoSQL, and cloud technologies. He is a founder and the Principal at Elephant Scale (`http://elephantscale.com/`), where he consults and teaches big data technologies. He has authored a few open source projects and has contributed to the Hadoop project. He is an author of the open source book *Hadoop Illuminated* (`http://hadoopilluminated.com/`).

He is the founder of the Big Data Gurus meetup in San Jose, CA. He has presented at various meetups and conferences.

You can find him on LinkedIn at `http://www.linkedin.com/in/sujeemaniyam` or read more about him at `http://sujee.net`.

I would like to acknowledge all the kind souls who have helped me along the way. Special thanks go to my co-author, Mark Kerzner, who is also a cofounder and a dear friend and, not the least, thanks goes to my family—my understanding wife and two young kids—who put up with my crazy schedule.

About the Reviewers

Ricky Ho is a data scientist and programmer, providing advisory and development services in big data analytics, machine learning, and distributed system design projects. He has a wide range of technical interests but is especially passionate about the intersection of machine learning and big data. He has served as the Principal Architect of Microsoft advertising, implementing scalable prediction systems to optimize advertising revenue within a large, web scale deployment. Prior to this, he was a researcher at Adobe's lab where he processed web log data for predictive analytics related research. Before that, he was a distinguished architect in PayPal's risk management team, where he developed a fraud detection system using machine learning and anomaly detection algorithms. Ricky holds 10 patents in distributed computing and cloud resource optimization. He is also an active technical blogger and shares what he learns on his blog at `http://horicky.blogspot.com`.

Raghu Sakleshpur is a technologist at heart who works in the field of big data, developing and designing solutions specifically in the Hadoop ecosystem. He started off his career in distributed (clustered) systems and transitioned to developing Enterprise Java application (middleware) space, only to return to his true passion of handling big data in both scaled up and scaled out architectures. He is currently working with Intel in the field of big data and spends a good portion of his time working with customers and partners alike to define optimal architectures for specific big data needs.

I would like to thank and acknowledge the patience of my family for putting up with my weird work hours and a special thanks to Mark Kerzner for providing me with the privilege to review his work.

Sergey Tatarenko is a senior software developer in a major legal e-discovery company in Austin, TX. He received his MSc in Computer Science from Ben-Gurion University of the Negev in Israel and has worked as a software developer since 1999. He started his professional career at Clockwork Solutions, Israel, and worked on a product that was used to build discrete event simulation models. Later, he lead a team of software developers in HyperRoll, but staying farther away from actual software development was not so much fun. In 2008, Sergey agreed to relocate to USA and help his previous employer to finish building their product. In April 2013, he decided to get himself more exposed to big data and started working for a leading legal e-discovery company in Austin, TX. In addition to being a software developer, Sergey is a proud father of three beautiful kids—Ilia, Antony, and Emilia—and a happy husband to his beautiful wife, Ilona. He is also a very active member in the Russian-speaking community of Austin, an enthusiastic builder of Arduino projects at home, and an occasional fisherman.

I would like to thank Mark Kerzner for his wonderful lectures on Hadoop; my dear wife, Ilona, for just being around; and my sweet kids for leaving me some spare time to work on this book. This is my first review and I sincerely hope, my comments will help to make this book better.

---

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.PacktPub.com`.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.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 Packt books and eBooks.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

Software plays a paramount role in today's world, and NoSQL databases are an important part of the modern stack. They are found wherever a subsecond response access to vast amounts of information is needed. However, there is a huge gap between the first "Hello World" example in a NoSQL database and creating practical, scalable, and stable applications. The aim of this book is to fill this gap and to give you practical guidelines for building NoSQL software.

The book is specifically formulated in terms of HBase, and there are a few areas of design where HBase might be different from Cassandra or MongoDB, for example, but most of the design patterns discussed here can be transferred to other NoSQL databases. You are expected to invest efforts in learning, which will lead to rewarding skills in the end.

What this book covers

Chapter 1, *Starting Out with HBase*, covers what HBase is and the various ways in which you can install it on your computer or cluster of computers, with practical advice on the development environment.

Chapter 2, *Reading, Writing, and Using SQL*, covers the HBase shell and gives the first example of Java code to read and write data in HBase. It also covers using the Phoenix driver for higher-level access, which gives back SQL, justifying the "Not-only-SQL" meaning of NoSQL.

Chapter 3, *Using HBase Tables for Single Entities*, covers the simplest HBase tables to deal with single entities, such as the table of users. Design patterns in this chapter emphasize on scalability, performance, and planning for special cases, such as restoring forgotten passwords.

Chapter 4, *Dealing with Large Files*, covers how to store large files in HBase systems. It also covers the alternative ways of storing them and the best practices extracted from solutions for large environments, such as Facebook, Amazon, and Twitter.

Chapter 5, *Time Series Data*, shows that stock market, human health monitoring, and system monitoring data are all classified as time series data. The design patterns for this organize time-based measurements in groups, resulting in balanced, high-performing HBase tables. Many lessons are learned from OpenTSDB.

Chapter 6, *Denormalization Use Cases*, discusses one of the most common design patterns for NoSQL denormalization, where the data is duplicated in more than one table, resulting in huge performance benefits. It also shows when to unlearn one's SQL normalization rules and how to apply denormalization wisely.

Chapter 7, *Advanced Patterns for Data Modeling*, shows you how to implement a many-to-many relationship in HBase that deals with transactions using compound keys.

Chapter 8, *Performance Optimization*, covers bulk loading for the initial data load into HBase, profiling HBase applications, benchmarking, and load testing.

What you need for this book

All the software used in this book is open source and free. You need Linux and Internet access. The book teaches you how to download and install the rest.

Who this book is for

If you deal with implementing practical big data solutions, involving quick access to massive amounts of data, this is the book for you. Primarily intended for software developers and architects, it can also be used by project managers, investors, and entrepreneurs who plan software implementations.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "The key is what you save when EC2 created the key pair for you, and `<cm-url>` is the URL of the server where you run the Cloudera Manager."