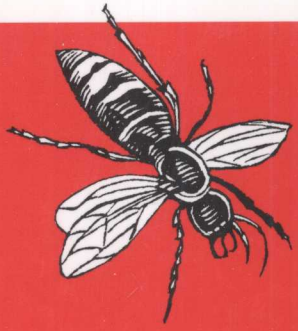
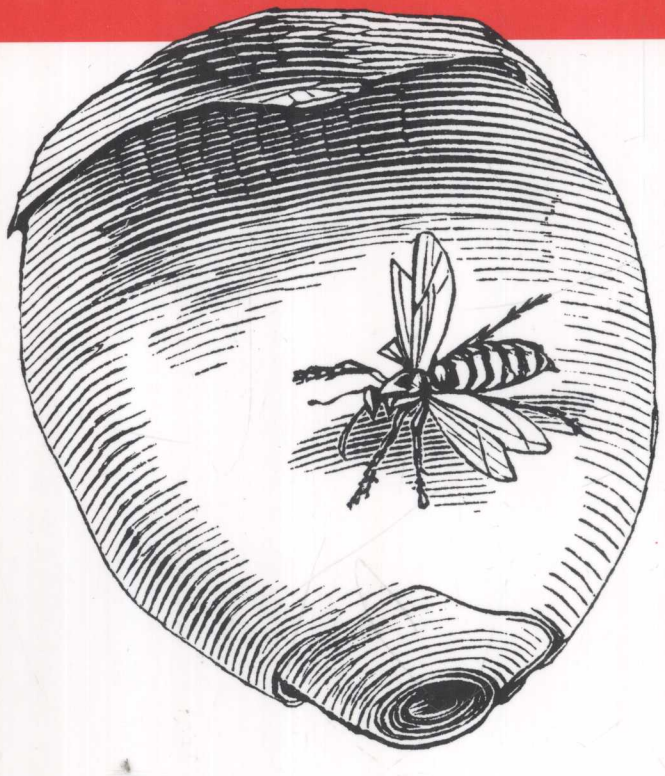
Hive编程（影印版）

Programming

Hive





Edward Capriolo,
Dean Wampler &
Jason Rutherglen 著

O'REILLY®
東南大學出版社

TP311.13
445

Hive编程（影印版）

Programming Hive

Edward Capriolo, Dean Wampler &
Jason Rutherglen 著

O'REILLY®

Beijing · Cambridge · Farnham · Köln · Sebastopol · Tokyo

O'Reilly Media, Inc.授权东南大学出版社出版

南京 东南大学出版社

图书在版编目（CIP）数据

Hive 编程：英文/(美)卡普里奥罗（Capriolo, E.），(美)万普勒（Wampler, D.），(美)卢森格林（Jason, R.)著. —影印本. —南京：东南大学出版社，2013.5

书名原文：Rrogramming Hive

ISBN 978-7-5641-4197-4

I. ① H… II. ①卡… ②万… ③卢… III. ①数据库系统－程序设计－英文 IV. ① TP311.13

中国版本图书馆 CIP 数据核字（2013）第 089649 号

江苏省版权局著作权合同登记

图字：10-2013-120 号

©2012 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Press, 2013. Authorized reprint of the original English edition, 2013 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版 2012。

英文影印版由东南大学出版社出版 2013。此影印版的出版和销售得到出版权和销售权的所有者 —— O'Reilly Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

Hive 编程（影印版）

出版发行：东南大学出版社
地　　址：南京四牌楼 2 号　　邮编：210096
出 版 人：江建中
网　　址：http://www.seupress.com
电子邮件：press@seupress.com
印　　刷：扬中市印刷有限公司
开　　本：787 毫米 × 980 毫米　16 开本
印　　张：22
字　　数：431 千字
版　　次：2013 年 5 月第 1 版
印　　次：2013 年 5 月第 1 次印刷
书　　号：ISBN 978-7-5641-4197-4
定　　价：54.00 元（册）

本社图书若有印装质量问题，请直接与营销部联系。电话（传真）：025-83791830

Preface

Programming Hive introduces *Hive*, an essential tool in the *Hadoop* ecosystem that provides an *SQL* (Structured Query Language) dialect for querying data stored in the *Hadoop Distributed Filesystem* (HDFS), other filesystems that integrate with Hadoop, such as *MapR-FS* and Amazon's *S3* and databases like *HBase* (the Hadoop database) and *Cassandra*.

Most data warehouse applications are implemented using relational databases that use SQL as the query language. Hive lowers the barrier for moving these applications to Hadoop. People who know SQL can learn Hive easily. Without Hive, these users must learn new languages and tools to become productive again. Similarly, Hive makes it easier for developers to port SQL-based applications to Hadoop, compared to other tool options. Without Hive, developers would face a daunting challenge when porting their SQL applications to Hadoop.

Still, there are aspects of Hive that are different from other SQL-based environments. Documentation for Hive users and Hadoop developers has been sparse. We decided to write this book to fill that gap. We provide a pragmatic, comprehensive introduction to Hive that is suitable for SQL experts, such as database designers and business analysts. We also cover the in-depth technical details that Hadoop developers require for tuning and customizing Hive.

You can learn more at the book's catalog page (*http://oreil.ly/Programming_Hive*).

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions. Definitions of most terms can be found in the Glossary.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

champion Hadoop as a solution internally. Even though I am now very familiar with Hadoop internals, Hive is still my primary method of working with Hadoop.

It is an honor to write a Hive book. Being a Hive Committer and a member of the Apache Software Foundation is my most valued accolade.

Dean Wampler

As a "big data" consultant at *Think Big Analytics*, I work with experienced "data people" who eat and breathe SQL. For them, Hive is a *necessary and sufficient condition* for Hadoop to be a viable tool to leverage their investment in SQL and open up new opportunities for data analytics.

Hive has lacked good documentation. I suggested to my previous editor at O'Reilly, Mike Loukides, that a Hive book was needed by the community. So, here we are...

Jason Rutherglen

I work at *Think Big Analytics* as a software architect. My career has involved an array of technologies including search, Hadoop, mobile, cryptography, and natural language processing. Hive is the ultimate way to build a data warehouse using open technologies on any amount of data. I use Hive regularly on a variety of projects.

Acknowledgments

Everyone involved with Hive. This includes committers, contributors, as well as end users.

Mark Grover wrote the chapter on Hive and Amazon Web Services. He is a contributor to the Apache Hive project and is active helping others on the Hive IRC channel.

David Ha and Rumit Patel, at M6D, contributed the case study and code on the Rank function. The ability to do Rank in Hive is a significant feature.

Ori Stitelman, at M6D, contributed the case study, Data Science using Hive and R, which demonstrates how Hive can be used to make first pass on large data sets and produce results to be used by a second R process.

David Funk contributed three use cases on in-site referrer identification, sessionization, and counting unique visitors. David's techniques show how rewriting and optimizing Hive queries can make large scale map reduce data analysis more efficient.

Ian Robertson read the entire first draft of the book and provided very helpful feedback on it. We're grateful to him for providing that feedback on short notice and a tight schedule.

此为试读，需要完整PDF请访问：www.ertongbook.com

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://oreil.ly/Programming_Hive*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

What Brought Us to Hive?

The three of us arrived here from different directions.

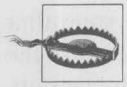
Edward Capriolo

When I first became involved with Hadoop, I saw the distributed filesystem and MapReduce as a great way to tackle computer-intensive problems. However, programming in the MapReduce model was a paradigm shift for me. Hive offered a fast and simple way to take advantage of MapReduce in an SQL-like world I was comfortable in. This approach also made it easy to prototype proof-of-concept applications and also to

`Constant width bold`
: Shows commands or other text that should be typed literally by the user.

`Constant width italic`
: Shows text that should be replaced with user-supplied values or by values determined by context.

This icon signifies a tip, suggestion, or general note.



This icon indicates a warning or caution.

Using Code Examples

This book is here to help you get your job done. In general, you may use the code in this book in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Programming Hive* by Edward Capriolo, Dean Wampler, and Jason Rutherglen (O'Reilly). Copyright 2012 Edward Capriolo, Aspect Research Associates, and Jason Rutherglen, 978-1-449-31933-5."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com*.

Safari® Books Online

Safari Books Online (*www.safaribooksonline.com*) is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

John Sichi provided technical review for the book. John was also instrumental in driving through some of the newer features in Hive like `StorageHandlers` and Indexing Support. He has been actively growing and supporting the Hive community.

Alan Gates, author of *Programming Pig* (*http://shop.oreilly.com/product/0636920018087.do*), contributed the HCatalog chapter. Nanda Vijaydev contributed the chapter on how Karmasphere offers productized enhancements for Hive. Eric Lubow contributed the SimpleReach case study. Chris A. Mattmann, Paul Zimdars, Cameron Goodale, Andrew F. Hart, Jinwon Kim, Duane Waliser, and Peter Lean contributed the NASA JPL case study.

Table of Contents

Preface xiii

1. Introduction 1
An Overview of Hadoop and MapReduce 3
Hive in the Hadoop Ecosystem 6
Pig 8
HBase 8
Cascading, Crunch, and Others 9
Java Versus Hive: The Word Count Algorithm 11
What's Next 13

2. Getting Started 15
Installing a Preconfigured Virtual Machine 15
Detailed Installation 16
Installing Java 16
Installing Hadoop 18
Local Mode, Pseudodistributed Mode, and Distributed Mode 19
Testing Hadoop 20
Installing Hive 21
What Is Inside Hive? 22
Starting Hive 23
Configuring Your Hadoop Environment 24
Local Mode Configuration 24
Distributed and Pseudodistributed Mode Configuration 26
Metastore Using JDBC 28
The Hive Command 29
Command Options 29
The Command-Line Interface 30
CLI Options 31
Variables and Properties 31
Hive "One Shot" Commands 34

Executing Hive Queries from Files 35
The .hiverc File 36
More on Using the Hive CLI 36
Command History 37
Shell Execution 37
Hadoop dfs Commands from Inside Hive 38
Comments in Hive Scripts 38
Query Column Headers 38

3. Data Types and File Formats 41
Primitive Data Types 41
Collection Data Types 43
Text File Encoding of Data Values 45
Schema on Read 48

4. HiveQL: Data Definition 49
Databases in Hive 49
Alter Database 52
Creating Tables 53
Managed Tables 56
External Tables 56
Partitioned, Managed Tables 58
External Partitioned Tables 61
Customizing Table Storage Formats 63
Dropping Tables 66
Alter Table 66
Renaming a Table 66
Adding, Modifying, and Dropping a Table Partition 66
Changing Columns 67
Adding Columns 68
Deleting or Replacing Columns 68
Alter Table Properties 68
Alter Storage Properties 68
Miscellaneous Alter Table Statements 69

5. HiveQL: Data Manipulation 71
Loading Data into Managed Tables 71
Inserting Data into Tables from Queries 73
Dynamic Partition Inserts 74
Creating Tables and Loading Them in One Query 75
Exporting Data 76

6. HiveQL: Queries 79
SELECT ... FROM Clauses 79
Specify Columns with Regular Expressions 81
Computing with Column Values 81
Arithmetic Operators 82
Using Functions 83
LIMIT Clause 91
Column Aliases 91
Nested SELECT Statements 91
CASE ... WHEN ... THEN Statements 91
When Hive Can Avoid MapReduce 92
WHERE Clauses 92
Predicate Operators 93
Gotchas with Floating-Point Comparisons 94
LIKE and RLIKE 96
GROUP BY Clauses 97
HAVING Clauses 97
JOIN Statements 98
Inner JOIN 98
Join Optimizations 100
LEFT OUTER JOIN 101
OUTER JOIN Gotcha 101
RIGHT OUTER JOIN 103
FULL OUTER JOIN 104
LEFT SEMI-JOIN 104
Cartesian Product JOINs 105
Map-side Joins 105
ORDER BY and SORT BY 107
DISTRIBUTE BY with SORT BY 107
CLUSTER BY 108
Casting 109
Casting BINARY Values 109
Queries that Sample Data 110
Block Sampling 111
Input Pruning for Bucket Tables 111
UNION ALL 112

7. HiveQL: Views 113
Views to Reduce Query Complexity 113
Views that Restrict Data Based on Conditions 114
Views and Map Type for Dynamic Tables 114
View Odds and Ends 115

此为试读，需要完整PDF请访问：www.ertongbook.com

8. HiveQL: Indexes 117
Creating an Index 117
Bitmap Indexes 118
Rebuilding the Index 118
Showing an Index 119
Dropping an Index 119
Implementing a Custom Index Handler 119

9. Schema Design 121
Table-by-Day 121
Over Partitioning 122
Unique Keys and Normalization 123
Making Multiple Passes over the Same Data 124
The Case for Partitioning Every Table 124
Bucketing Table Data Storage 125
Adding Columns to a Table 127
Using Columnar Tables 128
Repeated Data 128
Many Columns 128
(Almost) Always Use Compression! 128

10. Tuning 131
Using EXPLAIN 131
EXPLAIN EXTENDED 134
Limit Tuning 134
Optimized Joins 135
Local Mode 135
Parallel Execution 136
Strict Mode 137
Tuning the Number of Mappers and Reducers 138
JVM Reuse 139
Indexes 140
Dynamic Partition Tuning 140
Speculative Execution 141
Single MapReduce MultiGROUP BY 142
Virtual Columns 142

11. Other File Formats and Compression 145
Determining Installed Codecs 145
Choosing a Compression Codec 146
Enabling Intermediate Compression 147
Final Output Compression 148
Sequence Files 148

Compression in Action 149
Archive Partition 152
Compression: Wrapping Up 154

12. Developing 155
Changing Log4J Properties 155
Connecting a Java Debugger to Hive 156
Building Hive from Source 156
Running Hive Test Cases 156
Execution Hooks 158
Setting Up Hive and Eclipse 158
Hive in a Maven Project 158
Unit Testing in Hive with hive_test 159
The New Plugin Developer Kit 161

13. Functions 163
Discovering and Describing Functions 163
Calling Functions 164
Standard Functions 164
Aggregate Functions 164
Table Generating Functions 165
A UDF for Finding a Zodiac Sign from a Day 166
UDF Versus GenericUDF 169
Permanent Functions 171
User-Defined Aggregate Functions 172
Creating a COLLECT UDAF to Emulate GROUP_CONCAT 172
User-Defined Table Generating Functions 177
UDTFs that Produce Multiple Rows 177
UDTFs that Produce a Single Row with Multiple Columns 179
UDTFs that Simulate Complex Types 179
Accessing the Distributed Cache from a UDF 182
Annotations for Use with Functions 184
Deterministic 184
Stateful 184
DistinctLike 185
Macros 185

14. Streaming 187
Identity Transformation 188
Changing Types 188
Projecting Transformation 188
Manipulative Transformations 189
Using the Distributed Cache 189

Producing Multiple Rows from a Single Row 190
Calculating Aggregates with Streaming 191
CLUSTER BY, DISTRIBUTE BY, SORT BY 192
GenericMR Tools for Streaming to Java 194
Calculating Cogroups 196

15. Customizing Hive File and Record Formats 199
File Versus Record Formats 199
Demystifying CREATE TABLE Statements 199
File Formats 201
SequenceFile 201
RCFile 202
Example of a Custom Input Format: DualInputFormat 203
Record Formats: SerDes 205
CSV and TSV SerDes 206
ObjectInspector 206
Think Big Hive Reflection ObjectInspector 206
XML UDF 207
XPath-Related Functions 207
JSON SerDe 208
Avro Hive SerDe 209
Defining Avro Schema Using Table Properties 209
Defining a Schema from a URI 210
Evolving Schema 211
Binary Output 211

16. Hive Thrift Service 213
Starting the Thrift Server 213
Setting Up Groovy to Connect to HiveService 214
Connecting to HiveServer 214
Getting Cluster Status 215
Result Set Schema 215
Fetching Results 215
Retrieving Query Plan 216
Metastore Methods 216
Example Table Checker 216
Administrating HiveServer 217
Productionizing HiveService 217
Cleanup 218
Hive ThriftMetastore 219
ThriftMetastore Configuration 219
Client Configuration 219

17. Storage Handlers and NoSQL 221
Storage Handler Background 221
HiveStorageHandler 222
HBase 222
Cassandra 224
Static Column Mapping 224
Transposed Column Mapping for Dynamic Columns 224
Cassandra SerDe Properties 224
DynamoDB 225

18. Security 227
Integration with Hadoop Security 228
Authentication with Hive 228
Authorization in Hive 229
Users, Groups, and Roles 230
Privileges to Grant and Revoke 231
Partition-Level Privileges 233
Automatic Grants 233

19. Locking 235
Locking Support in Hive with Zookeeper 235
Explicit, Exclusive Locks 238

20. Hive Integration with Oozie 239
Oozie Actions 239
Hive Thrift Service Action 240
A Two-Query Workflow 240
Oozie Web Console 242
Variables in Workflows 242
Capturing Output 243
Capturing Output to Variables 243

21. Hive and Amazon Web Services (AWS) 245
Why Elastic MapReduce? 245
Instances 245
Before You Start 246
Managing Your EMR Hive Cluster 246
Thrift Server on EMR Hive 247
Instance Groups on EMR 247
Configuring Your EMR Cluster 248
Deploying hive-site.xml 248
Deploying a .hiverc Script 249

Setting Up a Memory-Intensive Configuration 249
Persistence and the Metastore on EMR 250
HDFS and S3 on EMR Cluster 251
Putting Resources, Configs, and Bootstrap Scripts on S3 252
Logs on S3 252
Spot Instances 252
Security Groups 253
EMR Versus EC2 and Apache Hive 254
Wrapping Up 254

22. HCatalog 255
Introduction 255
MapReduce 256
Reading Data 256
Writing Data 258
Command Line 261
Security Model 261
Architecture 262

23. Case Studies 265
m6d.com (Media6Degrees) 265
Data Science at M6D Using Hive and R 265
M6D UDF Pseudorank 270
M6D Managing Hive Data Across Multiple MapReduce Clusters 274
Outbrain 278
In-Site Referrer Identification 278
Counting Uniques 280
Sessionization 282
NASA's Jet Propulsion Laboratory 287
The Regional Climate Model Evaluation System 287
Our Experience: Why Hive? 290
Some Challenges and How We Overcame Them 291
Photobucket 292
Big Data at Photobucket 292
What Hardware Do We Use for Hive? 293
What's in Hive? 293
Who Does It Support? 293
SimpleReach 294
Experiences and Needs from the Customer Trenches 296
A Karmasphere Perspective 296
Introduction 296
Use Case Examples from the Customer Trenches 297

Glossary 305

Appendix: References 309

Index 313

此为试读，需要完整PDF请访问：www.ertongbook.com