Apache Hive Cookbook

Easy, hands-on recipes to help you understand Hive and its integration with frameworks that are used widely in today's big data world

Hanish Bansal
Saurabh Chauhan
Shrey Mehrotra

PACKT PUBLISHING
BIRMINGHAM - MUMBAI
Credits

Authors
Hanish Bansal
Saurabh Chauhan
Shrey Mehrotra

Reviewer
Aristides Villarreal Bravo

Commissioning Editor
Wilson D'souza

Acquisition Editor
Tushar Gupta

Content Development Editor
Anish Dhurat

Technical Editor
Vishal K. Mewada

Copy Editor
Dipti Mankame

Project Coordinator
Bijal Patel

Proofreader
Safis Editing

Indexer
Priya Sane

Graphics
Kirk D'Penha

Production Coordinator
Shantanu N. Zagade

Cover Work
Shantanu N. Zagade
About the Authors

Hanish Bansal is a software engineer with over 4 years of experience in developing big data applications. He loves to study emerging solutions and applications mainly related to big data processing, NoSQL, natural language processing, and neural networks. He has worked on various technologies such as Spring Framework, Hibernate, Hadoop, Hive, Flume, Kafka, Storm, and NoSQL databases, which include HBase, Cassandra, MongoDB, and search engines such as Elasticsearch.

In 2012, he completed his graduation in Information Technology stream from Jaipur Engineering College and Research Center, Jaipur, India. He was also the technical reviewer of the book *Apache Zookeeper Essentials*. In his spare time, he loves to travel and listen to music.

You can read his blog at [http://hanishblogger.blogspot.in/](http://hanishblogger.blogspot.in/) and follow him on Twitter at [https://twitter.com/hanishbansal786](https://twitter.com/hanishbansal786).

I would like to thank my parents for their love, support, encouragement and the amazing chances they've given me over the years.

Saurabh Chauhan is a module lead with close to 8 years of experience in data warehousing and big data applications. He has worked on multiple Extract, Transform and Load tools, such as Oracle Data Integrator and Informatica as well as on big data technologies such as Hadoop, Hive, Pig, Sqoop, and Flume.

He completed his bachelor of technology in 2007 from Vishveshwarya Institute of Engineering and Technology. In his spare time, he loves to travel and discover new places. He also has a keen interest in sports.

I would like to thank everyone who has supported me throughout my life.
Shrey Mehrotra has 6 years of IT experience and, since the past 4 years, in designing and architecting cloud and big data solutions for the governmental and financial domains.

Having worked with big data R&D Labs and Global Data and Analytical Capabilities, he has gained insights into Hadoop, focusing on HDFS, MapReduce, and YARN. His technical strengths also include Hive, Pig, Spark, Elasticsearch, Sqoop, Flume, Kafka, and Java.

He likes spending time performing R&D on different big data technologies. He is the co-author of the book *Learning YARN*, a certified Hadoop developer, and has also written various technical papers. In his free time, he listens to music, watches movies, and spending time with friends.

I would like to thank my mom and dad for giving me support to accomplish anything I wanted. Also, I would like to thank my friends, who bear with me while I am busy writing.
About the Reviewer

Aristides Villarreal Bravo is a Java developers, a member of the NetBeans Dream Team, and a Java User Groups leader.

He has organized and participated in various conferences and seminars related to Java, JavaEE, NetBeans, NetBeans Platform, free software, and mobile devices, nationally and internationally.

He has written tutorials and blogs about Java, NetBeans, and web development. He has participated in several interviews on sites such as NetBeans, NetBeans Dzone, and JavaHispano. He has developed plugins for NetBeans. He has been a technical reviewer for the book PrimeFaces Blueprints.

Aristides is the CEO of Javscaz Software Developers. He lives in Panamá.

To my mother, father, and all family and friends.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Chapter 1: Developing Hive</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Deploying Hive on a Hadoop cluster</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Deploying Hive Metastore</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Installing Hive</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Configuring HCatalog</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Understanding different components of Hive</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Compiling Hive from source</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Hive packages</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Debugging Hive</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Running Hive</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Changing configurations at runtime</td>
<td>18</td>
</tr>
<tr>
<td>Chapter 2: Services in Hive</td>
<td>Introducing HiveServer2</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Understanding HiveServer2 properties</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Configuring HiveServer2 high availability</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Using HiveServer2 Clients</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Introducing the Hive metastore service</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Configuring high availability of metastore service</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Introducing Hue</td>
<td>36</td>
</tr>
<tr>
<td>Chapter 3: Understanding the Hive Data Model</td>
<td>Introduction</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Using numeric data types</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Using string data types</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Using Date/Time data types</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Using miscellaneous data types</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Using complex data types</td>
<td>48</td>
</tr>
</tbody>
</table>
### Table of Contents

**Chapter 4: Hive Data Definition Language**
- Using operators 50
- Partitioning 57
- Partitioning a managed table 58
- Partitioning an external table 65
- Bucketing 65

**Chapter 5: Hive Data Manipulation Language**
- Introduction 89
- Loading files into tables 90
- Inserting data into Hive tables from queries 93
- Inserting data into dynamic partitions 96
- Writing data into files from queries 98
- Enabling transactions in Hive 99
- Inserting values into tables from SQL 101
- Updating data 104
- Deleting data 105