



C o m m u n i t y E x p e r i e n c e D i s t i l l e d

Learning Apache Kafka

Second Edition

Start from scratch and learn how to administer Apache Kafka effectively for messaging

Nishant Garg

[PACKT] open source*
PUBLISHING community experience distilled

Learning Apache Kafka

Second Edition

Start from scratch and learn how to administer Apache Kafka effectively for messaging

Nishant Garg



BIRMINGHAM - MUMBAI

Learning Apache Kafka

Second Edition

Copyright © 2015 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: October 2013

Second edition: February 2015

Production reference: 1210215

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

ISBN 978-1-78439-309-0

www.packtpub.com

Credits

Author

Nishant Garg

Project Coordinator

Harshal Ved

Reviewers

Sandeep Khurana

Saurabh Minni

Supreet Sethi

Proofreaders

Stephen Copestake

Paul Hindle

Commissioning Editor

Usha Iyer

Indexer

Rekha Nair

Acquisition Editor

Meeta Rajani

Graphics

Sheetal Aute

Content Development Editor

Shubhangi Dhamgaye

Production Coordinator

Nilesh R. Mohite

Technical Editors

Manal Pednekar

Chinmay S. Puranik

Cover Work

Nilesh R. Mohite

Copy Editors

Merilyn Pereira

Aarti Saldanha

About the Author

Nishant Garg has over 14 years of software architecture and development experience in various technologies, such as Java Enterprise Edition, SOA, Spring, Hadoop, Hive, Flume, Sqoop, Oozie, Spark, Shark, YARN, Impala, Kafka, Storm, Solr/Lucene, NoSQL databases (such as HBase, Cassandra, and MongoDB), and MPP databases (such as GreenPlum).

He received his MS in software systems from the Birla Institute of Technology and Science, Pilani, India, and is currently working as a technical architect for the Big Data R&D Group with Impetus Infotech Pvt. Ltd. Previously, Nishant has enjoyed working with some of the most recognizable names in IT services and financial industries, employing full software life cycle methodologies such as Agile and SCRUM.

Nishant has also undertaken many speaking engagements on big data technologies and is also the author of *HBase Essentials*, Packt Publishing.

I would like to thank my parents (Mr. Vishnu Murti Garg and Mrs. Vimla Garg) for their continuous encouragement and motivation throughout my life. I would also like to thank my wife (Himani) and my kids (Nitigya and Darsh) for their never-ending support, which keeps me going.

Finally, I would like to thank Vineet Tyagi, CTO and Head of Innovation Labs, Impetus, and Dr. Vijay, Director of Technology, Innovation Labs, Impetus, for encouraging me to write.

About the Reviewers

Sandeep Khurana, an 18 years veteran, comes with an extensive experience in the Software and IT industry. Being an early entrant in the domain, he has worked in all aspects of Java- / JEE-based technologies and frameworks such as Spring, Hibernate, JPA, EJB, security, Struts, and so on. For the last few professional engagements in his career and also partly due to his personal interest in consumer-facing analytics, he has been treading in the big data realm and has extensive experience on big data technologies such as Hadoop, Pig, Hive, ZooKeeper, Flume, Oozie, HBase and so on.

He has designed, developed, and delivered multiple enterprise-level, highly scalable, distributed systems during the course of his career. In his long and fruitful professional life, he has been with some of the biggest names of the industry such as IBM, Oracle, Yahoo!, and Nokia.

Saurabh Minni is currently working as a technical architect at AdNear. He completed his BE in computer science at the Global Academy of Technology, Bangalore. He is passionate about programming and loves getting his hands wet with different technologies.

At AdNear, he deployed Kafka. This enabled smooth consumption of data to be processed by Storm and Hadoop clusters. Prior to AdNear, he worked with Adobe and Intuit, where he dabbled with C++, Delphi, Android, and Java while working on desktop and mobile products.

Supreet Sethi is a seasoned technology leader with an eye for detail. He has proven expertise in charting out growth strategies for technology platforms. He currently steers the platform team to create tools that drive the infrastructure at Jabong. He often reviews the code base from a performance point of view. These aspects also put him at the helm of backend systems, APIs that drive mobile apps, mobile web apps, and desktop sites.

The Jabong tech team has been extremely helpful during the review process. They provided a creative environment where Supreet was able to explore some of cutting-edge technologies like Apache Kafka.

I would like to thank my daughter, Seher, and my wife, Smriti, for being patient observers while I spent a few hours everyday reviewing this book.

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	1
Chapter 1: Introducing Kafka	5
Welcome to the world of Apache Kafka	5
Why do we need Kafka?	8
Kafka use cases	8
Installing Kafka	10
Installing prerequisites	10
Installing Java 1.7 or higher	10
Downloading Kafka	11
Building Kafka	12
Summary	12
Chapter 2: Setting Up a Kafka Cluster	13
A single node – a single broker cluster	14
Starting the ZooKeeper server	14
Starting the Kafka broker	15
Creating a Kafka topic	16
Starting a producer to send messages	17
Starting a consumer to consume messages	19
A single node – multiple broker clusters	20
Starting ZooKeeper	20
Starting the Kafka broker	20
Creating a Kafka topic using the command line	21
Starting a producer to send messages	21
Starting a consumer to consume messages	22
Multiple nodes – multiple broker clusters	22
The Kafka broker property list	23
Summary	24

Chapter 3: Kafka Design	25
Kafka design fundamentals	26
Log compaction	28
Message compression in Kafka	29
Replication in Kafka	31
Summary	33
Chapter 4: Writing Producers	35
The Java producer API	36
Simple Java producers	38
Importing classes	38
Defining properties	38
Building the message and sending it	39
Creating a Java producer with custom partitioning	42
Importing classes	42
Defining properties	43
Implementing the Partitioner class	43
Building the message and sending it	44
The Kafka producer property list	47
Summary	48
Chapter 5: Writing Consumers	49
Kafka consumer APIs	50
The high-level consumer API	50
The low-level consumer API	52
Simple Java consumers	54
Importing classes	54
Defining properties	54
Reading messages from a topic and printing them	55
Multithreaded Java consumers	59
Importing classes	59
Defining properties	59
Reading the message from threads and printing it	60
The Kafka consumer property list	64
Summary	65
Chapter 6: Kafka Integrations	67
Kafka integration with Storm	68
Introducing Storm	68
Integrating Storm	69
Kafka integration with Hadoop	73
Introducing Hadoop	73
Integrating Hadoop	76

Hadoop producers	76
Hadoop consumers	78
Summary	79
Chapter 7: Operationalizing Kafka	81
Kafka administration tools	81
Kafka cluster tools	81
Adding servers	83
Kafka topic tools	86
Kafka cluster mirroring	88
Integration with other tools	90
Summary	91
Index	93
