



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

Apache Accumulo for Developers

Build and integrate Accumulo clusters with various cloud platforms

Guðmundur Jón Halldórsson

[PACKT] open source*
PUBLISHING community experience distilled

Apache Accumulo for Developers

Build and integrate Accumulo clusters with
various cloud platforms

Guðmundur Jón Halldórsson

[PACKT] open source 
PUBLISHING community experience distilled
BIRMINGHAM - MUMBAI

Apache Accumulo for Developers

Copyright © 2013 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

Production Reference: 1101013

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

ISBN 978-1-78328-599-0

www.packtpub.com

Cover Image by Gant Man (gantman@gmail.com)

Credits

Author

Guðmundur Jón Halldórsson

Reviewers

Einar Th. Einarsson

Andrea Mostosi

Pálmi Skowronski

Acquisition Editor

Joanne Fitzpatrick

Commissioning Editor

Sharvari Tawde

Technical Editors

Aparna Kumari

Krutika Parab

Pramod Kumavat

Hardik B. Soni

Project Coordinator

Akash Poojary

Copy Editors

Brandt D'Mello

Gladson Monterio

Alfida Paiva

Proofreader

Simran Bhogal

Indexer

Rekha Nair

Graphics

Abhinash Sahu

Ronak Dhruv

Production Coordinator

Manu Joseph

Cover Work

Manu Joseph

About the Author

Guðmundur Jón Halldórsson is a Software Engineer who enjoys the challenges of complex problems and pays close attention to detail. He is an annual speaker at the Icelandic Computer Society (SKY, <http://www.utmessan.is/>).

Guðmundur is a Software Engineer with extensive experience and management skills, and works for Five Degrees (www.fivedegrees.nl), a banking software company. The company develops and sells high-quality banking software. As a Senior Software Engineer, he is responsible for the development of a backend banking system produced by the company. Guðmundur has a B.Sc. in Computer Sciences from the Reykjavik University.

Guðmundur has a long period of work experience as a Software Engineer since 1996. He has worked for a large bank in Iceland, an insurance company, and a large gaming company where he was in the core EVE Online team.

Guðmundur is passionate about whatever he does. He loves to play online chess and Sudoku. And when he has time, he likes to read science fiction and history books.

He maintains a Facebook page to network with his friends and readers, and blogs about the wonders of programming and cloud computing at <http://www.gudmundurjon.net/>.

I would like to thank my two girls, Kolbrún and Bryndís, for their patience while I was writing this book, and researching in the area of cluster computing.

About the Reviewers

Einar Th. Einarsson has been hacking computers since childhood, and has worked both as a Programmer and a System Administrator for more than 15 years in diverse fields such as online gaming, anti-malware, biotech, and telecommunications, at companies such as CCP Games, FRISK Software, and deCODE Genetics. He is currently the CTO of a startup company focused on providing tools for the online poker world.

Andrea Mostosi is a passionate Software Developer. In 2003, while he was at high school, he started with a single-node LAMP stack and grew up by adding more languages, components, and nodes. He graduated in Milan and worked on several web-related projects. He is currently working with data, trying to discover information hidden behind huge datasets.

I would like to thank my girlfriend Khadija, who lovingly supports me in everything I do, and the people I collaborated with, for fun or for work, for everything they taught me. I would also like to thank Packt Publishing and its staff for this opportunity to contribute to this production.

Pálmi Skowronski holds a bachelor's and a master's degree in Computer Science from Reykjavík University, with a focus on machine-learning and heuristic searches.

Most recently, he has been working in the financial sector developing distributed enterprise solutions with Five Degrees as a Senior Developer, and is currently working on smart analysis of financial transactions with Meniga as a Software Specialist.

I would like to thank the author Mr. Halldórsson, a friend and colleague, for the many laughs and stimulating conversations we had during the writing of this book. May there be many more in the near future.

www.PacktPub.com

Support files, eBooks, discount offers and more

You might want to visit www.PacktPub.com for support files and downloads related to your book.

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.



<http://PacktLib.PacktPub.com>

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across 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 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 nine entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface	1
Chapter 1: Building an Accumulo Cluster from Scratch	5
Necessary requirements	6
Setting up Cygwin	7
Setting up Hadoop	8
SSH configuration	8
Creating a Hadoop user	9
Generating an SSH key for the Hadoop user	9
Installing Hadoop	10
Configuring Hadoop	11
core-site.xml	13
mapred-site.xml	14
hdfs-site.xml	14
hadoop-env.sh	15
Preparing the Hadoop filesystem	15
Starting the Hadoop cluster	16
Multi-node configurations	16
The NameNode website	18
The JobTracker website	19
The TaskTracker website	19
Setting up ZooKeeper	20
Installing ZooKeeper	20
Configuring ZooKeeper	21
Starting ZooKeeper	22
Setting up and configuring Accumulo	23
Installing Accumulo	23
Configuring Accumulo	24

Starting the Accumulo cluster	24
The Accumulo website	25
Connecting to the Accumulo cluster using Java	26
Summary	27
Chapter 2: Monitoring and Managing Accumulo	29
Monitoring	30
Setting up Ganglia	31
Configuring Ganglia	32
Setting up the Graylog2 server	33
Logging using Graylog2	33
Setting up Nagios	33
Hadoop	34
NameNode web interface	34
Finding the logfiles	35
How does Accumulo store files in Hadoop?	37
Live, dead, and decommissioning nodes	38
Accumulo	39
Monitoring a system's overview	41
Elasticity	41
Failover	42
Resource management	42
Summary	42
Chapter 3: Integrating Accumulo into Various Cloud Platforms	43
Amazon EC2	44
Prerequisites for Amazon EC2	44
Creating Amazon EC2 Hadoop and ZooKeeper cluster	44
Setting up Accumulo	48
Google Cloud Platform	49
Prerequisites for Google Cloud Platform	49
Creating the project	50
Installing the Google gcutil tool	50
Configuring credentials	50
Configuring the project	51
Creating the firewall rules	51
Creating the cluster	52
Hadoop	52
ZooKeeper	54
Accumulo	54
Deleting the cluster	55
Rackspace	57
Configuration	57

Network	57
Windows Azure	58
Prerequisites	58
Creating the cluster	58
Hadoop	59
ZooKeeper	60
Accumulo	61
Deleting the cluster	61
Summary	62
Chapter 4: Optimizing Accumulo Performance	63
Prerequisites	64
Hadoop performance	65
Baseline	65
Tuning	66
Tuning parameters for mapred-default.xml	66
HDFS	67
Tuning parameters for mapred-site.xml	68
Tuning parameters for hdfs-site.xml	69
ZooKeeper performance	69
ZooKeeper overview	70
Accumulo performance	70
Tuning parameters for accumulo-site.xml	71
Accumulo overview	71
Accumulo's performance summary	72
Tables	72
Comparing bulk ingest versus batch write	74
Accumulo examples	75
Summary	76
Chapter 5: Security	77
Visibility	79
Creating an Accumulo user	80
Creating tables in Accumulo	80
How does visibility work?	81
Security expression	85
Writing a Java client	85
Authorization	87
User authorizations	87
Handling secure authorization	88
Query Services Layer	88
Summary	88
