



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

Talend for Big Data

Access, transform, and integrate data using Talend's
open source, extensible tools

Bahaaldine Azarmi

[PACKT] open source*
PUBLISHING community experience distilled

Talend for Big Data

Access, transform, and integrate data using Talend's open source, extensible tools

Bahaaldine Azarmi



BIRMINGHAM - MUMBAI

Talend for Big Data

Copyright © 2014 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: February 2014

Production Reference: 2170214

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

ISBN 978-1-78216-949-9

www.packtpub.com

Cover Image by Abhishek Pandey (abhishek.pandey1210@gmail.com)

Credits

Author

Bahaaldine Azarmi

Project Coordinator

Ankita Goenka

Reviewers

Simone Bianchi

Vikram Takkar

Proofreader

Mario Cecere

Acquisition Editors

Mary Nadar

Llewellyn Rozario

Indexers

Hemangini Bari

Tejal Soni

Content Development Editor

Manasi Pandire

Production Coordinator

Komal Ramchandani

Technical Editors

Krishnaveni Haridas

Anand Singh

Cover Work

Komal Ramchandani

Copy Editor

Alfida Paiva

About the Author

Bahaaldine Azarmi is the cofounder of *reach5.co*. With his past experience of working at Oracle and Talend, he has specialized in real-time architecture using service-oriented architecture products, Big Data projects, and web technologies.

I like to thank my wife, Aurelia, for her support and patience throughout this project.

About the Reviewers

Simone Bianchi has a degree in Electronic Engineering from Italy, where he is living today, working as a programmer to develop web applications using technologies such as Java, JSP, jQuery, and Oracle. After having a brief experience with the Oracle Warehouse Builder tool, and as soon as the Talend solution came out, he started to extensively use this new tool in all his data migration/integration tasks as well as develop ETL layers in data warehouse projects. He also developed several Talend custom components such as tLogGrid, tDBFInput/Output, which you can download from the TalendForge site, and the ones to access/store data on the Web via SOAP/REST API.

I'd like to thank Packt Publishing to have chosen me to review this book, as well as the very kind people who work there, to have helped me to accomplish my first review at my best.

A special dedication to my father Americo, my mother Giuliana, my sisters Barbara and Monica, for all their support over the years, and finally to my little sweet nephew and niece, Leonardo and Elena, you are my constant source of inspiration.

Vikram Takkar is a freelance Business Intelligence and Data Integration professional with nine years of rich hands-on experience in multiple BI and ETL tools. He has a strong expertise in technologies such as Talend, Jaspersoft, Pentaho, Big Data-MongoDB, Oracle, and MySQL. He has managed and successfully executed multiple projects in data warehousing and data migration developed for both Unix and Windows environments. He has also worked as a Talend Data Integration trainer and facilitated training for various corporate clients in India, Europe, and the United States. He is an impressive communicator with strong leadership, analytical, and problem-solving skills. He is comfortable interacting with people across hierarchical levels for ensuring smooth project execution as per the client's specifications. Apart from this, he is a blogger and publishes articles and videos on open source BI and ETL tools along with supporting technologies on his YouTube channel at www.youtube.com/vtakkar. You can follow him on Twitter @VikTakkar and you can visit his blog at www.vikramtakkar.com.

I would like to thank the Packt Publishing team for again giving me the opportunity to review their book. Earlier, I reviewed their *Pentaho and Big Data Analytics* book.

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: Getting Started with Talend Big Data	5
Talend Unified Platform presentation	5
Knowing about the Hadoop ecosystem	7
Prerequisites for running examples	8
Downloading Talend Open Studio for Big Data	9
Installing TOSBD	9
Running TOSBD for the first time	10
Summary	12
Chapter 2: Building Our First Big Data Job	13
TOSBD – the development environment	13
A simple HDFS writer job	16
Checking the result in HDFS	25
Summary	25
Chapter 3: Formatting Data	27
Twitter Sentiment Analysis	27
Writing the tweets in HDFS	28
Setting our Apache Hive tables	31
Formatting tweets with Apache Hive	35
Summary	38
Chapter 4: Processing Tweets with Apache Hive	39
Extracting hashtags	39
Extracting emoticons	44
Joining the dots	46
Summary	48