



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 Oozie Essentials

Unleash the power of Apache Oozie to create and manage your Big Data and machine learning pipelines in one go

Jagat Jasjit Singh

[PACKT] open source*
PUBLISHING community experience distilled

Apache Oozie Essentials

Unleash the power of Apache Oozie to create and manage your Big Data and machine learning pipelines in one go

Jagat Jasjit Singh



BIRMINGHAM - MUMBAI

Apache Oozie Essentials

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: December 2015

Production reference: 1011215

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

ISBN 978-1-78588-038-4

www.packtpub.com

Credits

Author

Jagat Jasjit Singh

Project Coordinator

Shweta H Birwatkar

Reviewers

Siva Prakash

Rahul Tekchandani

Proofreader

Safis Editing

Commissioning Editor

Dipika Gaonkar

Indexer

Priya Sane

Acquisition Editor

Tushar Gupta

Production Coordinator

Melwyn Dsa

Content Development Editor

Preeti Singh

Cover Work

Melwyn Dsa

Technical Editor

Dhiraj Chandanshive

Copy Editor

Roshni Banerjee

About the Author

Jagat Jasjit Singh works for one of the largest telecom companies in Melbourne, Australia, as a big data architect. He has a total experience of over 10 years and has been working with the Hadoop ecosystem for more than 5 years. He is skilled in Hadoop, Spark, Oozie, Hive, Pig, Scala, machine learning, HBase, Falcon, Kafka, GraphX, Flume, Knox, Sqoop, Mesos, Marathon, Chronos, Openstack, and Java. He has experience of a variety of Australian and European customer implementations. He actively writes on Big Data and IoT technologies on his personal blog (<http://jugnu.life>). Jugnu (a Punjabi word) is a firefly that glows at night and illuminates the world with its tiny light. Jagat believes in this same philosophy of sharing knowledge to make the world a better place. You can connect with him on LinkedIn at <https://au.linkedin.com/in/jagatsingh>.

All the (author side) earnings of this book will go towards charity. Please consider donating, if you have not purchased this book directly, at <http://www.pingalwara.net/donations.html>. You can donate with your PayPal account or credit card.

This book is dedicated to Almighty God, who gave me everything, my parents, and the wonderful people from the Omnia project at Commonwealth Bank of Australia (<https://github.com/CommBank>). I would like to acknowledge the help of Tushar Gupta, Dhiraj Chandanshive, Roshni Banerjee, and Preeti Singh from Packt Publishing in writing this book.

About the Reviewers

Siva Prakash has been working in the field of software development for the last 7 years. Currently, he is working with CISCO, Bangalore. He has an extensive development experience in desktop-, mobile-, and web-based applications in ERP, telecom, and the digital media industry. He has passion for learning new technologies and sharing knowledge thus gained with others. He has worked on big data technologies for the digital media industry. He loves trekking, travelling, music, reading books, and blogging.

He is available on LinkedIn at <https://www.linkedin.com/in/techsivam>.

Rahul Tekchandani is a Hadoop software developer who specializes in building and developing Hadoop data platforms for big financial institutions. With experience in software design, development, and support, he has engineered strong, data-driven applications using the Cloudera's Hadoop Distribution. Rahul has also worked as an information architect to support data sanitization and data governance.

Prior to his career in software development, he completed his masters in Management of Information Systems at University of Arizona and worked on academic projects for top tech and banking companies.

He currently lives in Charlotte, North Carolina. Visit his developer's blog at www.rahultekchandani.com to see what he is currently exploring, and to learn more about him.

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	v
Chapter 1: Setting up Oozie	1
Configuring Oozie in Hortonworks distribution	1
Installing Oozie using tar ball	6
Creating a test virtual machine	7
Building Oozie source code	10
Summary of the build script	10
Codehaus Maven move	11
Download dependency jars	11
Preparing to create a WAR file	12
Create a WAR file	13
Configure Oozie MySQL database	14
Configure the shared library	16
Start server testing and verification	16
Summary	17
Chapter 2: My First Oozie Job	19
Installing and configuring Hue	19
Oozie concepts	23
Workflows	23
Coordinator	24
Bundles	24
Book case study	24
Running our first Oozie job	25
Types of nodes	29
Control flow nodes	29
Action nodes	30
Oozie web console	30
The Oozie command line	32
Summary	33

Chapter 3: Oozie Fundamentals	35
Chapter case study	35
The Decision node	41
The Email action	41
Expression Language functions	42
Basic EL constants	42
Basic EL functions	43
Workflow EL functions	43
Hadoop EL constants	43
HDFS EL functions	44
Email action configuration	45
Job property file	46
Submission from the command line	49
Workflow states	50
Summary	51
Chapter 4: Running MapReduce Jobs	53
Chapter case study	53
Running MapReduce jobs from Oozie	54
The job.properties file	56
Running the job	56
Running Oozie MapReduce job	56
Coordinators	57
Datasets	58
Frequency and time	61
Cron syntax for frequency	62
Timezone	64
The <done-flag> tag	65
Initial instance	65
My first Coordinator	65
Coordinator v1 definition	65
job.properties v1 definition	66
Coordinator v2 definition	67
job.properties v2 definition	69
Checking the job log	70
Running a MapReduce streaming job	70
Summary	71
Chapter 5: Running Pig Jobs	73
Chapter case study	73
The Pig command line	74
The config-default.xml file	76
Pig action	77

Pig Coordinator job v2	80
Parameters in the Dataset's input and output events	84
current(int n)	84
hoursInDay(int n)	85
daysInMonth(int n)	85
latest(int n)	85
Coordinator controls	86
Pig Coordinator job v3	88
Summary	90
Chapter 6: Running Hive Jobs	91
Chapter case study	91
Running a Hive job from the command line	92
Hive action	93
Validating Oozie Workflow	96
Hive 2 action	97
Parameterization of Coordinator jobs	99
dateOffset(String baseDate, int instance, String timeUnit)	99
dateTzOffset(String baseDate, String timezone)	100
formatTime(String timeStamp, String format)	100
Summary	101
Chapter 7: Running Sqoop Jobs	103
Chapter case study	103
Running Sqoop command line	104
Sqoop action	106
HCatalog	108
HCatalog datasets	109
HCatalog EL functions	110
HCatalog Coordinator functions	110
Pig script	112
The job.properties file	112
The Sqoop action Coordinator	114
Running the job	115
Checking data in the Hive table	116
Summary	117
Chapter 8: Running Spark Jobs	119
Spark action	120
Bundles	124
Data pipelines	128
Summary	129

Chapter 9: Running Oozie in Production	131
Packaging and continuous delivery	131
Oozie in secured cluster	137
Rerun	140
Rerun Workflow	140
Rerun Coordinator	141
Rerun Bundle	141
Summary	142
Index	143

Preface

With the increasing popularity of Big Data in enterprise, every day more and more workloads are being shifted to Hadoop.

To run those regular processing jobs on Hadoop, we need a scheduler that can act as cron for all data pipelines. Oozie plays this role in the Big Data world.

This book introduces you to the world of Oozie using a step-by-step case study-based approach.

What this book covers

Chapter 1, Setting up Oozie, covers how to install and configure Oozie in Hadoop cluster. We will also learn how to install Oozie from the source code.

Chapter 2, My First Oozie Job, covers running a "Hello World" equivalent first Oozie job. It also introduces the concept of Workflow, Coordinator, and Bundles.

Chapter 3, Oozie Fundamentals, introduces the fundamental concepts of control nodes, expression language, web console, and running Oozie jobs from Hue.

Chapter 4, Running MapReduce Jobs, teaches how to run MapReduce jobs from Oozie and explores the concepts of Coordinators, Datasets, and cron-based frequency schedules.

Chapter 5, Running Pig Jobs, teaches how to run Pig jobs from Oozie. We will also cover the concept of parameterization of Datasets and Coordinator controls.

Chapter 6, Running Hive Jobs, introduces how to run Hive jobs and discusses the concepts of parameterization of Coordinator actions.

Chapter 7, Running Sqoop Jobs, shows how to run Sqoop jobs from Oozie and introduces the concept of HCatalog Datasets and EL functions.

Chapter 8, Running Spark Jobs, shows how to run Spark jobs. It also introduces the concept of Bundles and how they are used to group a set of Coordinator jobs.

Chapter 9, Running Oozie in Production, covers how to package the code for production deployments and how to rerun the jobs that have failed.

What you need for this book

To follow the tutorial and code examples in this book, you need to have access to Hadoop cluster or you can configure a single node virtual machine-based cluster. You should have a good laptop/desktop, preferably with a Linux operating system or Windows with VirtualBox installed.

Who this book is for

This book is for anyone who is familiar with basics of Hadoop and Hive, and now wants to automate the data and machine learning pipelines using Apache Oozie.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "Now, edit the `torrc` file placed at the `/etc/tor/` directory."

Most of the code in the book is XML. A block of code is set as follows:



```
<workflow-app name="My_first_Workflow"
xmlns="uri:oozie:workflow:0.5">
  <start to="fs-2178"/>
  <kill name="Kill">
    <message>Action failed </message>
  </kill>
  <action name="fs-2178">
    <fs>
      <delete path='${nameNode}/user/hue' />
```



```
</fs>
<ok to="End"/>
<error to="Kill"/>
</action>
<end name="End"/>
</workflow-app>
```

Any command-line input or output is written as follows:

```
# $ hadoop fs -ls /user/hue/learn_oozie
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "**Go to Settings | Networking | Port Forwarding** , Click on Add new port forwarding."

 Warnings or important notes appear in a box like this. 

 Tips and tricks appear like this. 

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book – what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail feedback@packtpub.com, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at www.packtpub.com/authors.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at <http://www.packtpub.com> for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit <http://www.packtpub.com/support> and register to have the files e-mailed directly to you.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books – maybe a mistake in the text or the code – we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting <http://www.packtpub.com/submit-errata>, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to <https://www.packtpub.com/books/content/support> and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at copyright@packtpub.com with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at questions@packtpub.com, and we will do our best to address the problem.

1

Setting up Oozie

Oozie is a workflow scheduler system to run Apache Hadoop jobs. Oozie Workflow jobs are **Directed Acyclic Graphs (DAGs)** of actions. More information on DAG can be found at https://en.wikipedia.org/wiki/Directed_acyclic_graph. Actions tell *what* to do in the job. Oozie supports running jobs of various types such as Java, Map-reduce, Pig, Hive, Sqoop, Spark, and Distcp. The output of one action can be consumed by the next action to create a chain sequence.

Oozie has client-server architecture, in which we install the server for storing the jobs and using client we submit our jobs to the server.

In this chapter, we will learn how to install Oozie for learning purpose and in production. For learning purposes, we will build Oozie from the source code, and for production we will use Hadoop distribution by Hortonworks. Throughout the book, we will use Hortonworks single node virtual machine. If you are using a different Hadoop distribution, you should not worry at all. All distribution packages are the same for Oozie software, which is made by the Apache community (<http://oozie.apache.org>).

After reading this chapter, we will be able to:

- Configure Oozie in Hortonworks distribution using Ambari
- Install Oozie using the source code provided as tar ball by the Apache Oozie website


Configuring Oozie in Hortonworks distribution

In this section, we will learn how to configure Oozie inside Hortonworks Hadoop distribution using Ambari. We will configure the Oozie server to use a MySQL database instead of the default Derby database to store all job information.

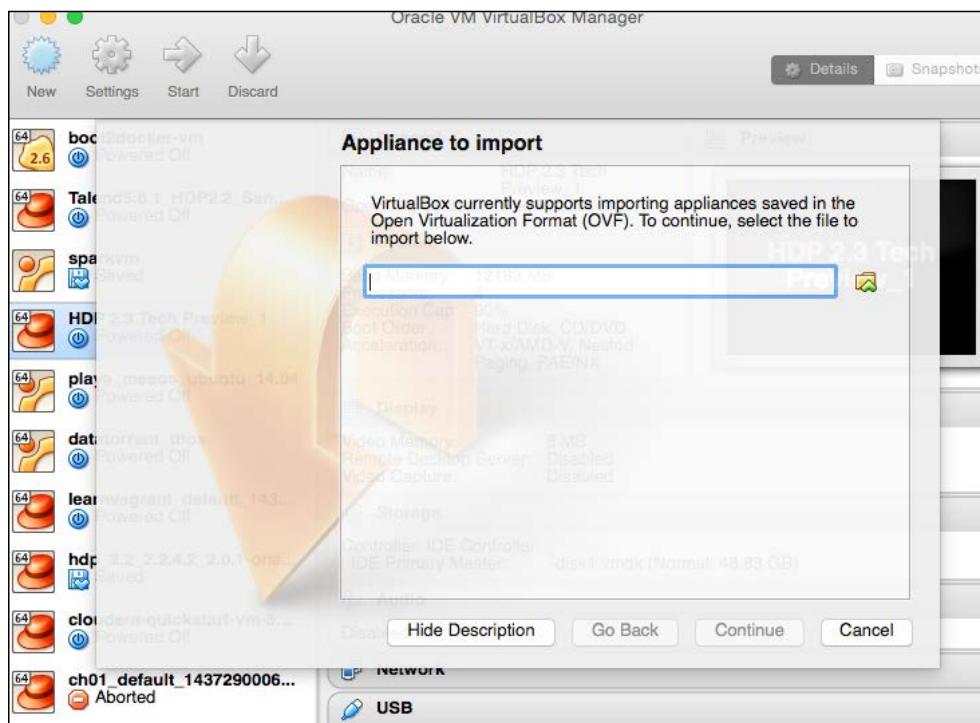
We will use a virtual machine to learn how to configure Oozie in Hortonworks Hadoop distribution. Most of other distributions, such as Cloudera, Pivotal, and so on, have similar steps.

Let's start with the following steps:

1. If you don't have VirtualBox on your machine, then download and install VirtualBox from <https://www.virtualbox.org/wiki/Downloads>.
2. Download the Hortonworks single node virtual machine from <http://hortonworks.com/hdp/downloads/>. It will take 1-2 hours depending upon your Internet connection speed.

[ It is always good to store the virtual machine images in a common folder. For example, I have folder in my machine such as ~/dev/vm/. It makes virtual machine image management easier.]

3. After the download is complete, open the VirtualBox and click on **File | Import Appliance:**



Import appliance