cloudera®

Spark Guide

Important Notice
© 2010-2017 Cloudera, Inc. All rights reserved.

Cloudera, the Cloudera logo, and any other product or service names or slogans contained in this document are trademarks of Cloudera and its suppliers or licensors, and may not be copied, imitated or used, in whole or in part, without the prior written permission of Cloudera or the applicable trademark holder.

Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation. All other trademarks, registered trademarks, product names and company names or logos mentioned in this document are the property of their respective owners. Reference to any products, services, processes or other information, by trade name, trademark, manufacturer, supplier or otherwise does not constitute or imply endorsement, sponsorship or recommendation thereof by us.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Cloudera.

Cloudera may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Cloudera, the furnishing of this document does not give you any license to these patents, trademarks copyrights, or other intellectual property. For information about patents covering Cloudera products, see http://tiny.cloudera.com/patents.

The information in this document is subject to change without notice. Cloudera shall not be liable for any damages resulting from technical errors or omissions which may be present in this document, or from use of this document.

Cloudera, Inc.
1001 Page Mill Road, Bldg 3
Palo Alto, CA 94304
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

Release Information

Version: CDH 5.6.x
Date: August 24, 2017

Table of Contents

Apache Spark Overview.....5

Running Your First Spark Application.....6

Spark Application Overview.....8

Spark Application Model.....8

Spark Execution Model.....8

Developing Spark Applications.....9

Developing and Running a Spark WordCount Application.....9

Using Spark Streaming.....13

Spark Streaming Example.....13

Enabling Fault-Tolerant Processing in Spark Streaming.....15

Configuring Authentication for Long-Running Spark Streaming Jobs.....16

Using Spark SQL.....16

Spark SQL Example.....16

Ensuring HiveContext Enforces Secure Access.....17

Interaction with Hive Views.....17

Using the spark-avro Library to Access Avro Data Sources.....17

Using Spark MLlib.....21

Running a Spark MLlib Example.....21

Enabling Native Acceleration For MLlib.....21

Accessing Data Stored in Amazon S3 through Spark.....22

Building Spark Applications.....24

Building Applications.....24

Building Reusable Modules.....25

Configuring Spark Applications.....26

Configuring Spark Application Properties in spark-defaults.conf.....27

Configuring Spark Application Logging Properties.....28

Running Spark Applications.....29

Submitting Spark Applications.....29

spark-submit Options.....30

Cluster Execution Overview.....31