

Venkat Ankam

Big Data Analytics

A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters

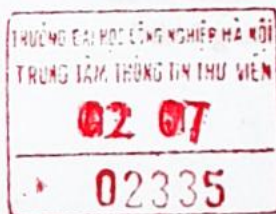


Packt >

Big Data Analytics

A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters

Venkat Ankam



Packt

BIRMINGHAM - MUMBAI

Big Data Analytics

Copyright © 2016 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: September 2016

Production reference: 12309016

Published by Packt Publishing Ltd.

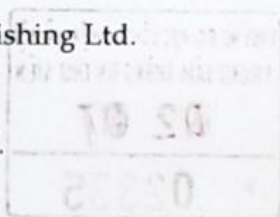
Livery Place

35 Livery Street

Birmingham B3 2PB, UK.

ISBN 978-1-78588-469-6

www.packtpub.com



Acknowledgement

I would like to thank Databricks for providing me with training in Spark in early 2014 and an opportunity to deepen my knowledge of Spark.

I would also like to thank Tyler Allbritton, principal architect, big data, cloud and analytics solutions at Tectonic, for providing me support in big data analytics projects and extending his support when writing this book.

Then, I would like to thank Mani Chhabra, CEO of Cloudwick, for encouraging me to write this book and providing the support I needed. Thanks to Arun Sirimalla, big data champion at Cloudwick, and Pranabh Kumar, big data architect at InsideView, who provided excellent support and inspiration to start meetups throughout India in 2011 to share knowledge of Hadoop and Spark.

Then I would like to thank Ashrith Mekala, solution architect at Cloudwick, for his technical consulting help.

This book started with a small discussion with Packt Publishing's acquisition editor Ruchita Bansali. I am really thankful to her for inspiring me to write this book. I am thankful to Kajal Thapar, content development editor at Packt Publishing, who then supported the entire journey of this book with great patience to refine it multiple times and get it to the finish line.

I would also like to thank Sumeet Sawant, Content Development Editor and Pranil Pathare, Technical Editor for their support in implementing Spark 2.0 changes.

I dedicate this book to my family and friends. Finally, this book would not have completed without the support from my wife, Srilatha, and my kids, Neha and Param, who cheered and encouraged me throughout the journey of this book.

Table of Contents

Preface	ix
Chapter 1: Big Data Analytics at a 10,000-Foot View	1
Big Data analytics and the role of Hadoop and Spark	3
A typical Big Data analytics project life cycle	5
Identifying the problem and outcomes	6
Identifying the necessary data	6
Data collection	6
Preprocessing data and ETL	6
Performing analytics	7
Visualizing data	7
The role of Hadoop and Spark	7
Big Data science and the role of Hadoop and Spark	8
A fundamental shift from data analytics to data science	8
Data scientists versus software engineers	9
Data scientists versus data analysts	10
Data scientists versus business analysts	10
A typical data science project life cycle	10
Hypothesis and modeling	11
Measuring the effectiveness	12
Making improvements	12
Communicating the results	12
The role of Hadoop and Spark	12
Tools and techniques	13
Real-life use cases	15
Summary	16
Chapter 2: Getting Started with Apache Hadoop and Apache Spark	17
Introducing Apache Hadoop	17
Hadoop Distributed File System	18
Features of HDFS	20

MapReduce	21
MapReduce features	21
MapReduce v1 versus MapReduce v2	22
MapReduce v1 challenges	23
YARN	24
Storage options on Hadoop	26
File formats	27
Compression formats	29
Introducing Apache Spark	30
Spark history	32
What is Apache Spark?	33
What Apache Spark is not	34
MapReduce issues	34
Spark's stack	37
Why Hadoop plus Spark?	40
Hadoop features	40
Spark features	41
Frequently asked questions about Spark	42
Installing Hadoop plus Spark clusters	43
Summary	47
Chapter 3: Deep Dive into Apache Spark	49
Starting Spark daemons	49
Working with CDH	50
Working with HDP, MapR, and Spark pre-built packages	50
Learning Spark core concepts	51
Ways to work with Spark	51
Spark Shell	51
Spark applications	53
Resilient Distributed Dataset	54
Method 1 – parallelizing a collection	54
Method 2 – reading from a file	55
Spark context	56
Transformations and actions	57
Parallelism in RDDs	59
Lazy evaluation	63
Lineage Graph	64
Serialization	65
Leveraging Hadoop file formats in Spark	66
Data locality	68
Shared variables	69
Pair RDDs	70

Lifecycle of Spark program	70
Pipelining	73
Spark execution summary	74
Spark applications	74
Spark Shell versus Spark applications	74
Creating a Spark context	75
SparkConf	75
SparkSubmit	76
Spark Conf precedence order	77
Important application configurations	77
Persistence and caching	78
Storage levels	79
What level to choose?	80
Spark resource managers – Standalone, YARN, and Mesos	80
Local versus cluster mode	80
Cluster resource managers	81
Standalone	81
YARN	82
Mesos	84
Which resource manager to use?	85
Summary	85
Chapter 4: Big Data Analytics with Spark SQL, DataFrames, and Datasets	87
History of Spark SQL	88
Architecture of Spark SQL	90
Introducing SQL, Datasources, DataFrame, and Dataset APIs	91
Evolution of DataFrames and Datasets	93
What's wrong with RDDs?	94
RDD Transformations versus Dataset and DataFrames Transformations	95
Why Datasets and DataFrames?	95
Optimization	96
Speed	96
Automatic Schema Discovery	97
Multiple sources, multiple languages	98
Interoperability between RDDs and others	98
Select and read necessary data only	98
When to use RDDs, Datasets, and DataFrames?	98
Analytics with DataFrames	99
Creating SparkSession	99
Creating DataFrames	100
Creating DataFrames from structured data files	100

Creating DataFrames from RDDs	100
Creating DataFrames from tables in Hive	103
Creating DataFrames from external databases	103
Converting DataFrames to RDDs	104
Common Dataset/DataFrame operations	104
Input and Output Operations	104
Basic Dataset/DataFrame functions	105
DSL functions	105
Built-in functions, aggregate functions, and window functions	106
Actions	106
RDD operations	106
Caching data	107
Performance optimizations	107
Analytics with the Dataset API	107
Creating Datasets	108
Converting a DataFrame to a Dataset	109
Converting a Dataset to a DataFrame	109
Accessing metadata using Catalog	109
Data Sources API	110
Read and write functions	110
Built-in sources	111
Working with text files	111
Working with JSON	111
Working with Parquet	112
Working with ORC	113
Working with JDBC	114
Working with CSV	116
External sources	116
Working with AVRO	117
Working with XML	117
Working with Pandas	118
DataFrame based Spark-on-HBase connector	119
Spark SQL as a distributed SQL engine	122
Spark SQL's Thrift server for JDBC/ODBC access	122
Querying data using beeline client	123
Querying data from Hive using spark-sql CLI	124
Integration with BI tools	125
Hive on Spark	125
Summary	125
Chapter 5: Real-Time Analytics with Spark	
Streaming and Structured Streaming	127
Introducing real-time processing	128
Pros and cons of Spark Streaming	129
History of Spark Streaming	130

Architecture of Spark Streaming	130
Spark Streaming application flow	132
Stateless and stateful stream processing	133
Spark Streaming transformations and actions	136
Union	136
Join	136
Transform operation	136
updateStateByKey	137
mapWithState	137
Window operations	137
Output operations	138
Input sources and output stores	139
Basic sources	140
Advanced sources	140
Custom sources	141
Receiver reliability	141
Output stores	141
Spark Streaming with Kafka and HBase	142
Receiver-based approach	142
Role of Zookeeper	144
Direct approach (no receivers)	145
Integration with HBase	146
Advanced concepts of Spark Streaming	147
Using DataFrames	147
MLlib operations	148
Caching/persistence	148
Fault-tolerance in Spark Streaming	148
Failure of executor	149
Failure of driver	149
Performance tuning of Spark Streaming applications	151
Monitoring applications	152
Introducing Structured Streaming	153
Structured Streaming application flow	154
When to use Structured Streaming?	156
Streaming Datasets and Streaming DataFrames	156
Input sources and output sinks	157
Operations on Streaming Datasets and Streaming DataFrames	157
Summary	161

Chapter 6: Notebooks and Dataflows with Spark and Hadoop	163
Introducing web-based notebooks	163
Introducing Jupyter	164
Installing Jupyter	165
Analytics with Jupyter	167
Introducing Apache Zeppelin	169
Jupyter versus Zeppelin	171
Installing Apache Zeppelin	171
Ambari service	172
The manual method	172
Analytics with Zeppelin	173
The Livy REST job server and Hue Notebooks	176
Installing and configuring the Livy server and Hue	177
Using the Livy server	178
An interactive session	178
A batch session	180
Sharing SparkContexts and RDDs	181
Using Livy with Hue Notebook	181
Using Livy with Zeppelin	184
Introducing Apache NiFi for dataflows	185
Installing Apache NiFi	185
Dataflows and analytics with NiFi	186
Summary	189
Chapter 7: Machine Learning with Spark and Hadoop	191
Introducing machine learning	192
Machine learning on Spark and Hadoop	193
Machine learning algorithms	194
Supervised learning	195
Unsupervised learning	195
Recommender systems	196
Feature extraction and transformation	197
Optimization	198
Spark MLlib data types	198
An example of machine learning algorithms	200
Logistic regression for spam detection	200
Building machine learning pipelines	203
An example of a pipeline workflow	204
Building an ML pipeline	205
Saving and loading models	208
Machine learning with H2O and Spark	208
Why Sparkling Water?	208

An application flow on YARN	208
Getting started with Sparkling Water	210
Introducing Hivemall	211
Introducing Hivemall for Spark	211
Summary	212
Chapter 8: Building Recommendation Systems with Spark and Mahout	213
Building recommendation systems	214
Content-based filtering	214
Collaborative filtering	214
User-based collaborative filtering	215
Item-based collaborative filtering	215
Limitations of a recommendation system	216
A recommendation system with MLlib	216
Preparing the environment	217
Creating RDDs	218
Exploring the data with DataFrames	219
Creating training and testing datasets	222
Creating a model	222
Making predictions	223
Evaluating the model with testing data	223
Checking the accuracy of the model	224
Explicit versus implicit feedback	225
The Mahout and Spark integration	225
Installing Mahout	225
Exploring the Mahout shell	226
Building a universal recommendation system with Mahout and search tool	230
Summary	234
Chapter 9: Graph Analytics with GraphX	235
Introducing graph processing	235
What is a graph?	236
Graph databases versus graph processing systems	237
Introducing GraphX	237
Graph algorithms	238
Getting started with GraphX	238
Basic operations of GraphX	238
Creating a graph	239
Counting	242
Filtering	242
inDegrees, outDegrees, and degrees	243

Triplets	244
Transforming graphs	244
Transforming attributes	245
Modifying graphs	245
Joining graphs	246
VertexRDD and EdgeRDD operations	247
GraphX algorithms	248
Triangle counting	250
Connected components	250
Analyzing flight data using GraphX	252
Pregel API	254
Introducing GraphFrames	256
Motif finding	259
Loading and saving GraphFrames	260
Summary	261
Chapter 10: Interactive Analytics with SparkR	263
Introducing R and SparkR	263
What is R?	264
Introducing SparkR	265
Architecture of SparkR	266
Getting started with SparkR	267
Installing and configuring R	267
Using SparkR shell	268
Local mode	268
Standalone mode	269
Yarn mode	269
Creating a local DataFrame	270
Creating a DataFrame from a DataSources API	271
Creating a DataFrame from Hive	271
Using SparkR scripts	273
Using DataFrames with SparkR	275
Using SparkR with RStudio	280
Machine learning with SparkR	282
Using the Naive Bayes model	282
Using the k-means model	284
Using SparkR with Zeppelin	285
Summary	287
Index	289

Preface

Big Data Analytics aims at providing the fundamentals of Apache Spark and Hadoop, and how they are integrated together with most commonly used tools and techniques in an easy way. All Spark components (Spark Core, Spark SQL, DataFrames, Datasets, Conventional Streaming, Structured Streaming, MLLib, GraphX, and Hadoop core components), HDFS, MapReduce, and Yarn are explored in great depth with implementation examples on Spark + Hadoop clusters.

The Big Data Analytics industry is moving away from MapReduce to Spark. So, the advantages of Spark over MapReduce are explained in great depth to reap the benefits of in-memory speeds. The DataFrames API, the Data Sources API, and the new Dataset API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help in building streaming applications. New structured streaming concept is explained with an Internet of Things (IOT) use case. Machine learning techniques are covered using MLLib, ML Pipelines and SparkR; Graph Analytics are covered with GraphX and GraphFrames components of Spark.

This book also introduces web based notebooks such as Jupyter, Apache Zeppelin, and data flow tool Apache NiFi to analyze and visualize data, offering Spark as a Service using Livy Server.

What this book covers

Chapter 1, Big Data Analytics at a 10,000-Foot View, provides an approach to Big Data analytics from a broader perspective and introduces tools and techniques used on Apache Hadoop and Apache Spark platforms, with some of most common use cases.

Chapter 2, Getting Started with Apache Hadoop and Apache Spark, lays the foundation for Hadoop and Spark platforms with an introduction. This chapter also explains how Spark is different from MapReduce and how Spark on the Hadoop platform is beneficial. Then it helps you get started with the installation of clusters and setting up tools needed for analytics.

Chapter 3, Deep Dive into Apache Spark, covers deeper concepts of Spark such as Spark Core internals, how to use pair RDDs, the life cycle of a Spark program, how to build Spark applications, how to persist and cache RDDs, and how to use Spark Resource Managers (Standalone, Yarn, and Mesos).

Chapter 4, Big Data Analytics with Spark SQL, DataFrames, and Datasets, covers the Data Sources API, the DataFrames API, and the new Dataset API. There is a special focus on why DataFrame API is useful and analytics of DataFrame API with built-in sources (Csv, Json, Parquet, ORC, JDBC, and Hive) and external sources (such as Avro, Xml, and Pandas). Spark-on-HBase connector explains how to analyze HBase data in Spark using DataFrames. It also covers how to use Spark SQL as a distributed SQL engine.

Chapter 5, Real-Time Analytics with Spark Streaming and Structured Streaming, provides the meaning of real-time analytics and how Spark Streaming is different from other real-time engines such as Storm, trident, Flink, and Samza. It describes the architecture of Spark Streaming with input sources and output stores. It covers stateless and stateful stream processing and using receiver-based and direct approach with Kafka as a source and HBase as a store. Fault tolerance concepts of Spark streaming is covered when application is failed at driver or executors. Structured Streaming concepts are explained with an Internet of Things (IOT) use case.

Chapter 6, Notebooks and Dataflows with Spark and Hadoop, introduces web-based notebooks with tools such as Jupyter, Zeppelin, and Hue. It introduces the Livy REST server for building Spark as a service and for sharing Spark RDDs between multiple users. It also introduces Apache NiFi for building data flows using Spark and Hadoop.

Chapter 7, Machine Learning with Spark and Hadoop, aims at teaching more about the machine learning techniques used in data science using Spark and Hadoop. This chapter introduces machine learning algorithms used with Spark. It covers spam detection, implementation, and the method of building machine learning pipelines. It also covers machine learning implementation with H2O and Hivemall.

Chapter 8, Building Recommendation Systems with Spark and Mahout, covers collaborative filtering in detail and explains how to build real-time recommendation engines with Spark and Mahout.

Chapter 9, Graph Analytics with GraphX, introduces graph processing, how GraphX is different from Giraph, and various graph operations of GraphX such as creating graph, counting, filtering, degrees, triplets, modifying, joining, transforming attributes, Vertex RDD, and EdgeRDD operations. It also covers GraphX algorithms such as triangle counting and connected components with a flight analytics use case. New GraphFrames component based on DataFrames is introduced and explained some concepts such as motif finding.

Chapter 10, Interactive Analytics with SparkR, covers the differences between R and SparkR and gets you started with SparkR using shell scripts in local, standalone, and Yarn modes. This chapter also explains how to use SparkR with RStudio, DataFrames, machine learning with SparkR, and Apache Zeppelin.

What you need for this book

Practical exercises in this book are demonstrated on virtual machines (VM) from Cloudera, Hortonworks, MapR, or prebuilt Spark for Hadoop for getting started easily. The same exercises can be run on a bigger cluster as well.

Prerequisites for using virtual machines on your laptop:

- RAM: 8 GB and above
- CPU: At least two virtual CPUs
- The latest VMWare player or Oracle VirtualBox must be installed for Windows or Linux OS
- Latest Oracle VirtualBox, or VMWare Fusion for Mac
- Virtualization enabled in BIOS
- Browser: Chrome 25+, IE 9+, Safari 6+, or Firefox 18+ recommended (HDP Sandbox will not run on IE 10)
- Putty
- WinScP

The Python and Scala programming languages are used in chapters, with more focus on Python. It is assumed that readers have a basic programming background in Java, Scala, Python, SQL, or R, with basic Linux experience. Working experience within Big Data environments on Hadoop platforms would provide a quick jump start for building Spark applications.