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

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

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 H20 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.