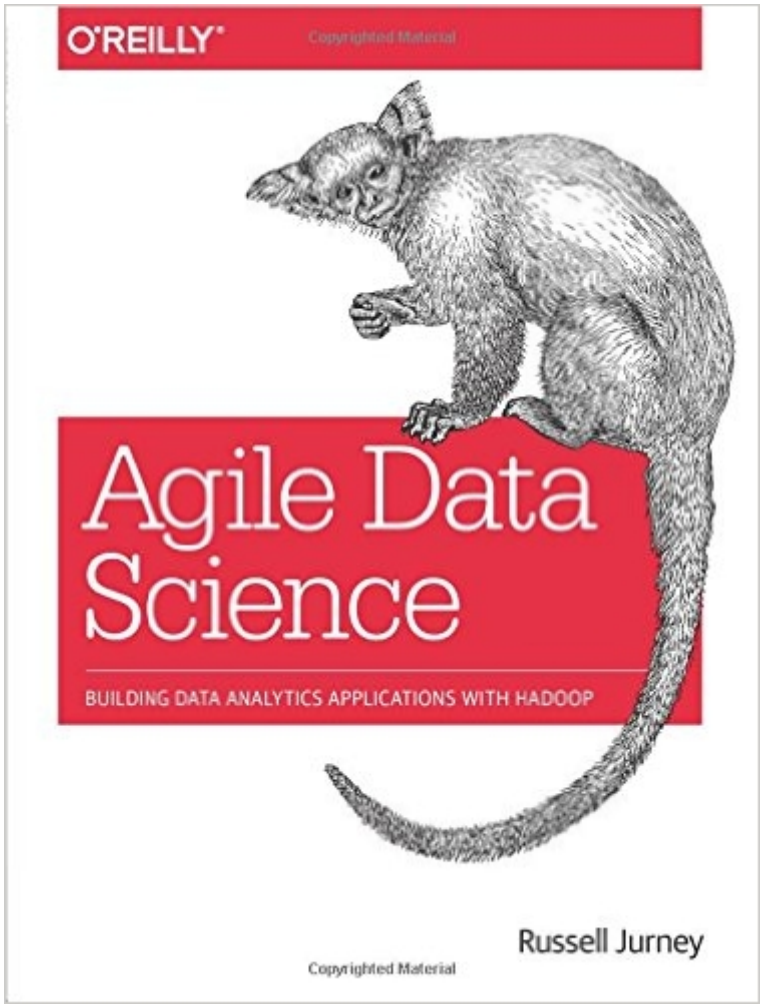
Agile Data Science: Building Data Analytics Applications With Hadoop



Synopsis

Mining big data requires a deep investment in people and time. How can you be sure youâ ™re building the right models? With this hands-on book, youâ ™ll learn a flexible toolset and methodology for building effective analytics applications with Hadoop.Using lightweight tools such as Python, Apache Pig, and the D3.js library, your team will create an agile environment for exploring data, starting with an example application to mine your own email inboxes. Youâ ™ll learn an iterative approach that enables you to quickly change the kind of analysis youâ ™re doing, depending on what the data is telling you. All example code in this book is available as working Heroku apps.Create analytics applications by using the agile big data development methodologyBuild value from your data in a series of agile sprints, using the data-value stackGain insight by using several data structures to extract multiple features from a single datasetVisualize data with charts, and expose different aspects through interactive reportsUse historical data to predict the future, and translate predictions into actionGet feedback from users after each sprint to keep your project on track

Book Information

Paperback: 178 pages

Publisher: O'Reilly Media; 1 edition (October 28, 2013)

Language: English

ISBN-10: 1449326269

ISBN-13: 978-1449326265

Product Dimensions: 7 x 0.4 x 9.2 inches

Shipping Weight: 1.1 pounds (View shipping rates and policies)

Average Customer Review: 3.6 out of 5 starsÂ Â See all reviewsÂ (8 customer reviews)

Best Sellers Rank: #428,112 in Books (See Top 100 in Books) #77 inÂ Books > Computers & Technology > Programming > Languages & Tools > Ruby #229 inÂ Books > Computers & Technology > Databases & Big Data > Data Modeling & Design #270 inÂ Books > Computers & Technology > Databases & Big Data > Data Mining

Customer Reviews

The story is nice, but the code that forms the basis of the entire project behind the book DOESN'T COMPILE. The author has - as of today (June 9, 2014) - removed all of the github references to the project.I"m half way through the book, have been practicing Agile development techniques for several years, and I am not quite sure what in particular makes this book about Data Science 'Agile'

based.One thing that he does nicely is explain the Pig code he uses, but I can’t use those programs because the Python programs that gather the data that feed Pig will not compile, even after I de-bugged his code for several hours. (Example: the author made reference to an RFC inline in the Python code that would have NEVER compiled. NEVER. Line 11 gmail.py from call to email utilitiies)

I was once told by a chief data scientist that they would rather teach a mathematician programming than a programmer math (to be a data scientist). After being a data scientist for some time now I would have to respectfully disagree. 85% of data science is plumbing and I wouldn’t hire a physicist to be a plumber. Indeed modern data scientists really do need to be full-stack developers trapped in an academic’s body.Jurney nails it! He offers tools and methodologies adapted to common data science workflows and their associated pitfalls wherein we spend 85% of our time plumbing and 15% of our time integrating some off-the-shelf algorithm to find deep insight.So, for new data scientists or 3rd-4th year grad students who have balanced their Twitter API hack with NSF grant deadlines, this is ABSOLUTELY REQUIRED READING.

Book review - Agile Data Science by Russell Jurney, O’Reilly MediaThe subtitle "Building Data Analytics Applications with Hadoop" of this book says more about the book than the actual title "Agile Data Science". However the subtitle will probably fool most people. Before reading this book I believed that Hadoop with the the distributed file-system HDFS. If you are looking for a book about building applications on the of HDFS then this book IS NOT for you. It turns out that Hadoop is much more than just HDFS.Do not buy this book for learning about agile software development methodologies. There are some rather strange comments about personal and private space requirement for creative workers as well as mentioning of "Easy access to large-format printing is a requirement for the agile environment." The discussion about agile methods for working with data science is interesting. The basic question is if it is possible to bridge agile methods and data science since science in it’s nature does not consists of a predefined set of tasks. It seems to me that the tools and software used in chapter 3 are called agile an hence is the process agile. In part II of the book the application build is chapter 3 is refined in a number of steps that the author calls iterative. But again, that does not make the process agile. I am not saying that the author is wrong but the point about the agile method and how process and tools interact to make the development agile is not entirely clear to me.This is NOT a book about the inner workings of Hadoop. Please refer to "Hadoop: The Definitive Guide" by Tom White for O’Reilly Media for a thorough introduction to

Hadoop. Instead the book takes a very practical approach and show us how to build agile applications using various Hadoop components like Pig, MapReduce, and the Avro serialization framework. In addition you will see how to move data into the popular noSQL database MongoDB and how to use ElasticSearch to search the data. Finally, all the collected data is accesses through a lightweight web application build with Python and Flask with visual enhancement made in Bootstrap and D3.Agile Data Science covers a lot of material and uses lots of different software and tools. If you want to run the examples in the book you have two options 1) a user-contributed Linux Vagrant image is available with most of the required software or 2) you can follow along the instructions given in the book and the accompanied Github project and install the software yourself. In either case you have to pay close attention to software versions. All of the examples work but it does require some effort the get them running and if you feel uncomfortable using a terminal and command line you might have a hard time playing with the examples.Being able to work in an agile way with data science is quite important but I do not feel that the attempt made by the author convinced me that the suggested framework will work in a practical setting.The main value of this book is definitely chapter 3 where Jurney show us how to go from zero to a working data science application. The application is literally build from ground up starting with data collection over storing data to build a web front-end. This chapter is alone worth the price of the entire book.Part II of the books contains interesting material about data visualizations and prediction models. For many readers some prior knowledge about Naive Bayes and the Natural Language Toolkit would most likely be useful to fully understand the implications of the predictions made around what makes an email likely to receive a response.I review for the O`Reilly Reader Review Program and I want to be transparent about my reviews so you should know that I received a free copy of this ebook in exchange of my review.

One of the problems with data science is that any description of what is encountered takes on the appearance of a mythical unicorn, noone person could possibly have all of the skills required. And it gets worse when you add to the standard set of statistics, domain knowledge, and programming the ability to deploy the application into a high speed environment. This book is not going to make a data scientist an expert in running a data center, but it is useful to give someone who has the rest of the skills an understanding of the environment their work will be deployed into.One of the conflicts between the data scientist/analyst and information technology groups is that while the data scientist gives the data owned by the organization its value, IT is charged with storing the data and providing the access. And in a high velocity, high volume environment of big data, not understanding how the

architecture works can lead to the data scientist creating valid solutions that cannot be applied in the actual day to day working environment. That is where this book comes in. The book has associated virtual machines in software repository so that the data scientist who does not know anything about infrastructure and the software stack that the data and the analysis rides on can see how everything fits together.The book title is misleading. This is not a book about data analytics. This is a book for data analysts so they know how their analytical application is deployed and applied to day-to-day use in enterprise environments. For that reason it is useful.Disclaimer: I received a free electronic copy of this book as part of the Oreilly Press Blogger program.

Introduction to Learning the Basics of Agile Project Management (Agile Project Management, Agile Software Development, Scrum) Agile Product Management: (Box Set): Agile Estimating & Planning Your Sprint with Scrum & User Stories 21 Tips (scrum, scrum master, agile development, agile software development) Agile Project Management: The Agile PMO: Leading the Effective, Value Driven and Agile Project Management Office (Agile Business Leadership Book 1) Agile Project Management: Agile Revolution, Beyond Software Limits: A Practical Guide to Implementing Agile Outside Software Development (Agile Business Leadership, Book 4) Agile Project Management: An Inclusive Walkthrough of Agile Project Management (Agile Project Management, Agile Software Developement, Scrum, Project Management) Analytics: Data Science, Data Analysis and Predictive Analytics for Business MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems