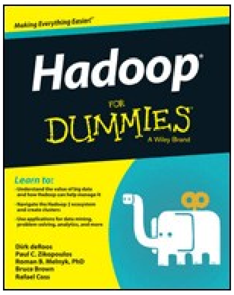


Hadoop for Dummies

by Dirk deRoos et al.

John Wiley & Sons (US). (c) 2014. Copying Prohibited.

Chapter 18: Ten Hadoop Resources Worthy of a Bookmark

In This Chapter

- Learning Hadoop — for free
- Finding the Hadoop information you need — fast
- Setting up a lifelong learning plan for Hadoop

From its roots in the early 2000s as an Internet search engine indexer, Hadoop has evolved to become a large-scale, general-purpose computing platform. Indeed, competence in Hadoop is one of the hottest skills you can list on a résumé in today's IT job market. If we can tell you one thing from our collective century-plus years of IT experience across multiple jobs and technology domains, it's that you should *never* reach the finish line of your learning roadmap.

Hadoop continues to evolve in a fascinating manner — especially when you consider all the Apache subprojects (and associated projects) that work within the Hadoop ecosystem. You're off to a great start with this book, though this fast-paced environment will continue to change. For example, many new processing frameworks for YARN are being developed that will introduce a wide variety of data processing options to Hadoop. We believe that the Hive project will explode with innovation, especially when you add YARN (see Chapter 7) and Tez (see Chapter 7 again) to the mix. The point? If you want to stay on top of Hadoop, you have to invest in it with a lifelong learning plan.

We highlight the free areas of Hadoop training in keeping with its open source spirit. You'll see that most vendors have found that they can make money delivering top-notch Hadoop training, so they often have both options: for-fee and for-free.

In this chapter, we describe what we think are ten terrific Hadoop resources that are worthy of creating a bookmark in your browser. These resources not only pick up from where we leave off in this book but also help you create a lifelong learning plan for Hadoop. From virtual universities, to 'zines and websites and more, you can continue learning in order to stay on the leading edge of the Hadoop curve, or simply to ensure that you have a solid understanding of the technology.

Central Nervous System: Apache.org

The Apache Software Foundation (ASF) is the central community for open source software projects. (***Note***: The group's charter stipulates that Apache software must be used for public good — so we're assuming that you'll use Hadoop for tasks other than finding better ways to increase the cost of gas.) Not just any project can be an Apache project — many consensus-driven processes convert a piece of software from its initial designs and beta code (its *incubator* status) to full-fledged, generally available software. You can find more about ASF at http://apache.org.

The ASF isn't just where projects like Hadoop are managed — it's where they "live and breathe." Today, there are hundreds of Apache projects. With this in mind, you should bookmark the Apache Hadoop page (http://projects.apache.org/projects/hadoop.html) as one of your mainstay learning resources. This site is important because you can access the source code there. You can also open or view Hadoop-related defects or bugs; view the license; access mailing lists for the community; download a versioned Hadoop feature, component, or branch (not just those marked stable); and more.

At this point, you've entered the Valhalla of Hadoop links — the Apache.org site is über-Hadoop Land, and is the home of Hadoop's development. You can think of the site as Hadoop's central nervous system. We make no guarantee that everything there is easy to consume, but its information is generally valuable — and straight from the source of Hadoop's developers.

Tweet This

Twitter isn't the place to learn Hadoop per se — after all, you can't easily master MapReduce programming in lessons that span only 140 characters. Be that as it may, quite a number of big data gurus are on Twitter, and they express opinions and point to resources that can make you a smarter Hadoop user.

A number of top-influencer lists in the Twitter landscape cover Hadoop and big data, and that's the best way to find these Hadoop personalities and add them to your Twitter lists. Here are a couple notable lists where you can find the most distinguished personalities covering Hadoop and big data on Twitter — including some of the authors of this book:

- **#BigData100 (Big Data Republic)**: tinyurl.com/ouk6lb8
- **Top 200 Big Data Influencers (Onalytica)**: tinyurl.com/oq6677s

Hortonworks University

Hortonworks University (hortonworks.com/hadoop-training) provides Hadoop training and certifications. The site offers Hadoop courses built for either administrator or developer practitioners with the option of a rigorous certification program. Hortonworks employs some of the deepest and most noted Hadoop experts in the world, so you're assured of quality expertise behind the courseware.

Hortonworks University has for-free and for-fee training. We focus on the free stuff in this chapter, so we think that the place you'll head to is the Hortonworks Sandbox (hortonworks.com/products/Hortonworks-sandbox) and its Resources page (hortonworks.com/resources). If you're looking for fee-based training, you can find it there as well.

The Hortonworks Sandbox gives you a portable Hadoop environment with an accompanying set of tutorials that cover a wide arrange of features from the latest HDP distribution. (This distribution is also used extensively in for-fee training).

The aforementioned Resources pages provide a wide array of document-based tutorials, videos, presentations, demos, and more. It also provides a decent roadmap to get started, aptly named "Getting Started with Hadoop" (hortonworks.com/get-started).

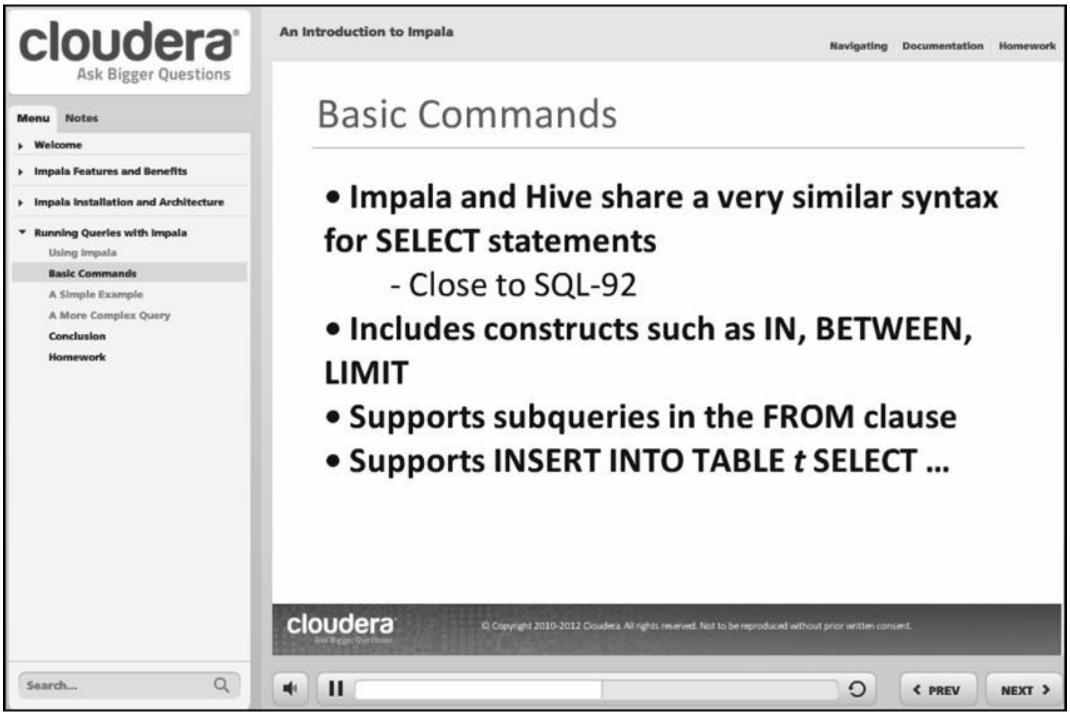
Cloudera University

Cloudera University (university.cloudera.com) is similar in its business model and charter to Hortonworks University, providing a number of learning avenues that run the gamut from traditional text to video. Cloudera is a prominent fixture in the Hadoop world. (Doug Cutting, the "father" of Hadoop is its chief architect.) The site offers an extensive set of courses, and more, which are based on the Cloudera Distribution for Hadoop (CDH).

Some courses are offered for a fee with in-classroom instruction, but one option lets you take certain courses for free in an online video series — for example, Cloudera Essentials for Apache Hadoop, at university.cloudera.com/onlineresources.html. When we took the course, we would have liked to have seen more-engaging materials in the courseware, but the instructors are engaging, considering that you're watching a recorded video (plus, nobody gets mad at you for chewing gum while in class).

We think that the Introduction to Data Science class at Cloudera University is pretty cool. Next to the *big data* label, *data science* is likely one of the most overused, or most misunderstood, labels — but people in that profession are commanding even higher salaries than Hadoop experts are. Cloudera even has a certification program (the Cloudera Certified Professional: Data Scientist), which we found to be a unique and terrific idea.

Cloudera University includes a number of modules in its e-learning catalog (university.cloudera.com/onlineresources/elearning.html). Because Cloudera is focused on Hadoop as well as on its own set of Hadoop add-ons, the site offers training in the full spectrum of features that Cloudera brings to the table. For example, in Figure 18-1 you can see an example of the course An Introduction to Impala. (Impala, if you're curious, is the Cloudera alternative to Hive; we cover Impala in Chapter 15.)



Basic Commands

- Impala and Hive share a very similar syntax for SELECT statements
  - Close to SQL-92
- Includes constructs such as IN, BETWEEN, LIMIT
- Supports subqueries in the FROM clause
- Supports INSERT INTO TABLE *t* SELECT ...

cloudera

© Copyright 2010-2012 Cloudera. All rights reserved. Not to be reproduced without prior written consent.

Figure 18-1: The Cloudera University e-learning course on Cloudera Impala

To get started with CDH and create an environment to complement your knowledge of Hadoop (or any Cloudera technology, for that matter), you can find packaged code and virtual images at cloudera.com/content/support/en/downloads.html.

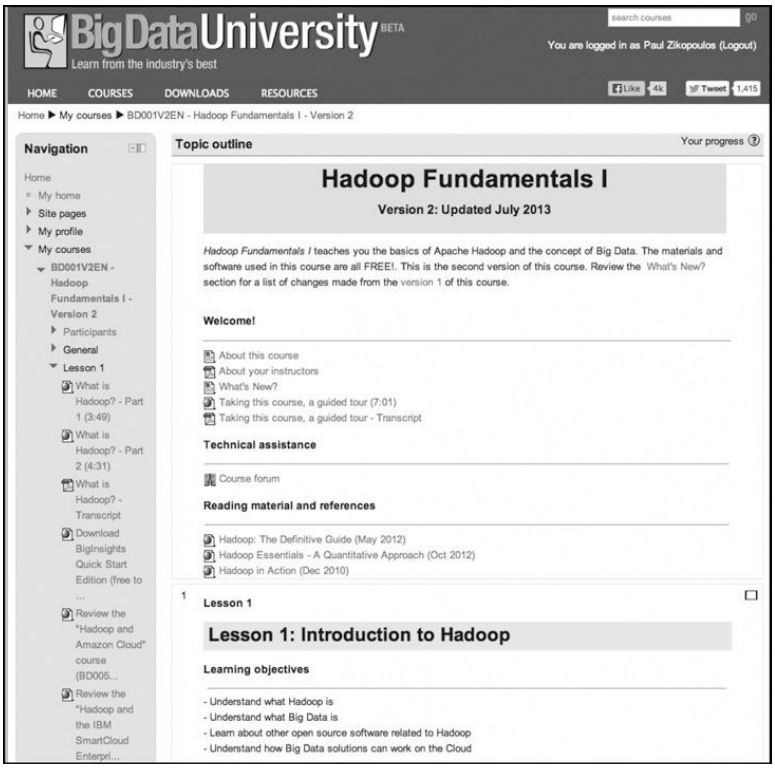
BigDataUniversity.com

BigDataUniversity.com (the case doesn't matter when you enter the URL in your browser) is a fantastic resource for learning about — you guessed it — big data. Of course, big data isn't just Hadoop, so you'll find more than Hadoop resources at this site. This university has over 100,000 students enrolled and learning about Hadoop and big data every day.

You'll notice right off the bat that this isn't a typical IBM resource. For example, you won't fill out dozens of fields and answer all kinds of questions that make you feel like you're getting set up for a cold call. We like to think of BigDataUniversity.com as free, in-your-place and at-your-pace Hadoop training. The word *free* is the key here: Unlike the other two universities we detail in this chapter, there isn't a fee-based component anywhere on the site.

The university moniker for this site isn't an accident — it has quite an extensive list of courseware that expands beyond Hadoop (bigdatauniversity.com/courses). From a Hadoop perspective, you won't just find courses on "Hadoop Fundamentals," but also "Hadoop and the Amazon Cloud," "Hadoop Reporting and Analytics" and more — including some database stuff. That's why we really like this resource – it gives off a Swiss Army knife vibe that gives you a place to expand your Hadoop knowledge even further into the big data domain.

Courses at BigDataUniversity.com are composed of traditional reading materials, mixed with multimedia, and code examples. An example of the Hadoop Fundamentals I course is shown in Figure 18-2.


BigDataUniversity BETA
Learn from the industry's best
search courses
go
You are logged in as Paul Zikopoulos (Logout)
HOME
COURSES
DOWNLOADS
RESOURCES
Like 4k
Tweet 1,415
Home ▶ My courses ▶ BD001V2EN - Hadoop Fundamentals I - Version 2
Navigation
Home
My home
Site pages
My profile
My courses
BD001V2EN - Hadoop Fundamentals I - Version 2
Participants
General
Lesson 1
What is Hadoop? - Part 1 (3:49)
What is Hadoop? - Part 2 (4:31)
What is Hadoop? - Transcript
Download BigInsights Quick Start Edition (free to ...
Review the "Hadoop and Amazon Cloud" course (BD005...
Review the "Hadoop and the IBM SmartCloud Enterpri...
Topic outline
Your progress ?

Hadoop Fundamentals I

Version 2: Updated July 2013

Hadoop Fundamentals I teaches you the basics of Apache Hadoop and the concept of Big Data. The materials and software used in this course are all FREE!. This is the second version of this course. Review the What's New? section for a list of changes made from the version 1 of this course.

Welcome!

About this course
About your instructors
What's New?
Taking this course, a guided tour (7:01)
Taking this course, a guided tour - Transcript

Technical assistance

Course forum

Reading material and references

Hadoop: The Definitive Guide (May 2012)
Hadoop Essentials - A Quantitative Approach (Oct 2012)
Hadoop in Action (Dec 2010)

1 Lesson 1 □

Lesson 1: Introduction to Hadoop

Learning objectives

- Understand what Hadoop is
- Understand what Big Data is
- Learn about other open source software related to Hadoop
- Understand how Big Data solutions can work on the Cloud

Figure 18-2: The Hadoop Fundamentals I course on the Big Data University site

From the navigation panel on the left, you can see multiple lessons and even a teaching assistant that can provide technical assistance if you get stuck. When you're done with a course, you can take a test. If you pass — you get a certificate!

Another nice feature of this site is that you can leverage the IBM Smart Cloud and create your own Hadoop cluster for free.

If you'd rather host your Hadoop platform locally, you can use IBM InfoSphere BigInsights, IBM's own Hadoop distribution. A Quick Start Edition (available at www.ibm.com/developerworks/downloads/im/biginsightsquick) comes with its own set of tutorials, which showcase not only Hadoop but also certain IBM enhancements. (The BigInsights Quick Start Edition includes the Text Analytics Toolkit, for example, which includes an Eclipse-based text analytics development environment with an accompanying SQL-like declarative language that runs on Hadoop, and other platforms.) You can use any Hadoop distribution for the courseware on BigDataUniversity.com.

Planet Big Data Blog Aggregator

We love it when the name of a site tells you exactly what it does — like planet Big Data Blog Aggregator (www.planetbigdata.com): It's an aggregator of blogs about big data, Hadoop, and other related topics on the planet (well, on Planet Earth anyway).

Both big names and no-names show up on the site, but that's helpful: Though there's undoubtedly commitment to Hadoop by Cloudera, Hortonworks, IBM, and others, it's often refreshing and valuable to get exposure to the thoughts and opinions of grass roots, non-affiliated practitioners by communities not tied to a specific vendor in your learning roadmap.

TIP Are you a big data blogger? Get your blog included in the planet Big Data Blog Aggregator list by e-mailing planetbigdata@gmail.com.

Quora's Apache Hadoop Forum

The Quora Apache Hadoop forum (www.quora.com/Apache-Hadoop) is the cornerstone for anyone looking to find out more about Hadoop, or about big data in general, for that matter.

As in any forum, the range of questions and answers you can find at this site is dizzying, but they all lead you to what you're looking for: knowledge. The site has linkages to Hadoop and to its individual components — for example, it has specific forums for MapReduce, HDFS, Pig, HBase, and more. The site also has associated Hadoop forums; for example, Cloudera and Hortonworks have specific discussion groups for their distributions — a testament to how popular this forum is.

Of course, as you transform yourself into a Hadoop demigod, you can answer questions that are posted to the forum and develop your Hadoop influence. (A lot of the active participants in this forum are on the Twitter lists we identify earlier in this chapter.)

The IBM Big Data Hub

The IBM Big Data Hub (www.ibmbigdatahub.com) is an excellent place to learn about Hadoop and its ecosystem. Despite being owned and operated by IBM, this site's content isn't always linked with IBM products.

The IBM Big Data Hub provides any visitor with enough knowledge to quench anyone's thirst for big data. You'll find all sorts of blogs, videos, analysts' articles, use cases, infographics, presentations, and more. It's truly a treasure trove of big data resources. This site also aggregates videos from the IBM Big Data and Analytics page at YouTube (youtube.com/user/ibmbigdata), which leads you into even more top-notch resources. For example, it has videos such as "What Is Big Data?" and "What Is Hadoop?" that feature some of the authors of this book.

Conferences Not to Be Missed

There are many Hadoop conferences, and even more big data conferences. We're recommending the Hadoop Summit (hadoopsummit.org) and Strata Hadoop World (strataconf.com) as the quintessential conferences not to be missed. Typically, a distribution vendor co-sponsors these conferences. For example, Yahoo! and Hortonworks sponsor the Hadoop Summit, and Cloudera is the co-sponsor of Strata Hadoop World.

Both Strata Hadoop World and the Hadoop Summit are *the* gathering places of the brightest Hadoop minds in the business; these conferences attract a wide array of Hadoop-interested professionals, including decision makers, architects, developers, analysts, and more.

The Strata Hadoop World name didn't come by accident; two formerly separate and independent conferences (Strata and Hadoop World) have now joined forces to become one of the world's largest gatherings of the Apache Hadoop community. A look at the curriculum makes obvious its focus on all aspects of Hadoop — from sessions devoted to hands-on practitioners to sessions devoted to business use cases.

The Hadoop Summit can be considered a competitor to Strata Hadoop World (though if you're lucky, your bosses will pay for you to go to both). The summit features the same themes and, likely, a lot of the same presenters. One aspect that we find appealing is that the conference tracks are chosen by the community at large as opposed to a conference committee. In the June 2013 Hadoop Summit that took place in San Jose, over 6,000 community members cast over 15,000 votes to create the seven tracks that became the pillars of the conference.

TIP If your appetite leans more in the direction of big data, we think that the yearly IBM Insight Conference (www.ibm.com/software/data/2013-conference) is a must-attend event. It not only features deep, hands-on Hadoop labs and sessions but also runs the gamut of big data topics, including stream computing, governance, the interaction of Hadoop and relational databases, and more.

The Google Papers That Started it All

What is now known as Hadoop has its genesis in a number of papers written by Google employees who were focused on the problem of indexing the Web. While the Apache Nutch project (an open source technology for crawling the Web) was turning its focus on scaling outward in order to index higher volumes of web data, Google published a paper, "The Google File System" (October 2003: research.google.com/archive/gfs.html), which greatly influenced Doug Cutting and his Nutch co-founder, Mike Cafarella. Shortly after, Google released its paper "MapReduce: Simplified Data Processing on Large Clusters" (December 2004: research.google.com/archive/mapreduce.html).

Together, the concept of a distributed file system and a large-scale parallel processing framework were taken by Cutting and Cafarella to develop Apache Hadoop. Of course, Cutting commercialized this work while at Yahoo!, and the rest, as they say, is history.

TIP Here's a great question for a game of Trivial Pursuit for IT geeks: Whatever happened to Mike Cafarella, who cofounded Hadoop with Doug Cutting?” The answer? He's an associate professor at the University of Michigan, and he's working on the Hadoop-complementary project RecordBreaker. Some call him the "Pete Best of big data." (Pete Best was the original drummer for The Beatles.)

A host of other Google papers have influenced the Hadoop ecosystem as well. For example, Google's paper “Bigtable: A Distributed Storage System for Structured Data (November 2006: research.google.com/archive/bigtable.html) is the inspiration behind HBase, among other NoSQL technologies.

Though these papers represent the original ideas behind Hadoop, and parts of its ecosystem, as a tribute to where it all began, we've included Google Research (research.google.com) and its collection of groundbreaking research papers in our list. Even today, reading these papers gives you a strong appreciation of where Hadoop came from and, potentially, some ideas of where it might evolve.

The Bonus Resource: What Did We Ever Do B.G.?

Considering the impact that Google has had on Hadoop, we thought it prudent to toss in one more related resource to keep in mind if you're on the hunt for Hadoop information: Google. (It's fair to lump YouTube into Google because not only does Google own it, but it has also become one of the top three Internet search sites.) From watching how to bake a pie to solving a problem on your computer to learning about Hadoop, after you type what you're looking for, there's a great chance that you'll find it. All this, of course, makes us wonder: What did we ever do B.G. (before Google)?