Apache™ Hadoop® with Spark™ in One Day

This one-day workshop covers the essential introductory aspects of the Apache Hadoop and Spark ecosystems. After completing the workshop attendees will gain an understanding Hadoop's technical value proposition and acquire hands-on experience with some basic Hadoop tools including Spark (with Zeppelin GUI). Approximately 30% of the workshop time is devoted to assisted hands-on exercises.

Unique Features:

- Instructional and hands-on exercises are performed on a local Hadoop cluster
- All attendees receive a copy of the book "*Hadoop 2 Quick Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem*" (written by the presenter)
- A high-touch small group atmosphere with full access to example source code and a post workshop Q&A page
- On-site lunch and breaks are included with cost of the workshop
- Discounts on desk-side Hadoop systems are available to all participants

> *"An excellent introduction to Hadoop concepts and tools. The hands-on lessons were a great launching pad for further exploration".*
>
> *– workshop particpant*

Intended Audience:

Though no prior database experience is needed, those that work in the traditional database and data warehousing sectors should find the workshop useful. Devops and application programmers should find the material helps them understand the Hadoop processing models and ecosystem. Finally, those involved in data science or statistics will learn about how Hadoop and Spark can be used as an analytics tool.

What You Will Learn:

Attendees will learn why Hadoop is different from more traditional approaches to data analysis. The Hadoop core components will be presented and related to the various Hadoop capabilities. In addition, attendees will gain hands-on experience with the Hadoop Distributed File Systems (HDFS), the Hadoop resource manager (YARN), and several high level Hadoop tools including Spark. After completing the workshop attendees will be able to use and navigate a production Hadoop cluster and develop their own projects by building on the workshop examples.

About the Presenter:

Douglas Eadline, PhD, is a consultant and writer in the Big Data (Hadoop) and High Performance Computing (HPC) industries. Doug has written hundreds of articles, white papers, and instructional documents covering many aspects of HPC and Hadoop computing. Prior to starting and editing the popular *ClusterMonkey.net* website in 2005, he served as editor-in-chief for *ClusterWorld Magazine*, and was senior HPC editor for *Linux Magazine*. He has authored *Hadoop Fundamentals LiveLessons, Second Edition* (2015), and *Apache Hadoop YARN LiveLessons* (2014), and is coauthor of *Apache Hadoop YARN* (2014) and *Hadoop 2 Quick-Start Guide* (2016), all from Addison-Wesley.

Prerequisites:

- Familiarity with the Linux command line and text editing is helpful
- A wifi capable laptop with an up-to-date web browser and an `ssh` client (For Windows users, we highly recommend MobaXterm, http://mobaxterm.mobatek.net)
- To simplify connection issues, attendees can rent a preconfigured Chromebook that meets these requirements

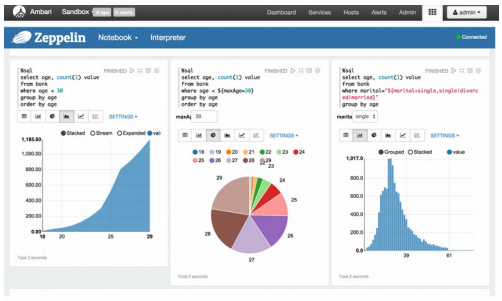
Workshop Outline

WELCOME 8:30AM

Why is Hadoop Such a Big Deal?
- A Brief History of Apache Hadoop
- What is Big Data?
- Hadoop as a Data Lake
- Apache Hadoop V2 is a Platform
- The Apache Hadoop Project Ecosystem
- Apache Spark

Hadoop Distributed File System (HDFS) Basics
- HDFS Basics
- HDFS User Commands

Hadoop MapReduce Framework
- The MapReduce Model
- MapReduce Data Flow



The Zeppelin Web GUI used for Spark Examples

BREAK 10:00-10:30 (Cluster Help Available)

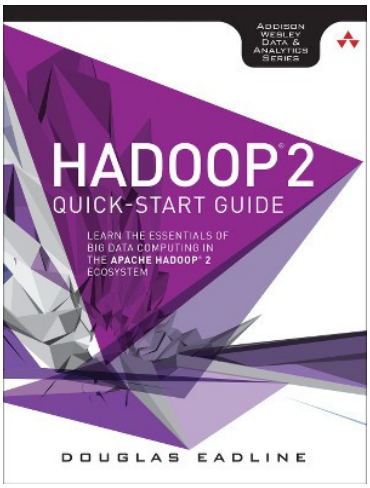
Running Example MapReduce Programs and Benchmarks
- Apache Hadoop Example Programs
- Apache Hadoop Benchmarks
- Viewing Application Progress

Hands-on Exercises covering HDFS and Hadoop Examples

LUNCH 12:15-1PM

Examples of Essential Hadoop Tools
- Using Apache Pig
- Using Apache Hive
- Using Apache HBase
- Using Apache Spark with Zeppelin

Tools and Topics Worth Knowing About
- Hadoop Distributions and Installation
- Manage Hadoop Workflows with Apache Oozie
- Apache YARN and Application Frameworks
- Finding more information about Hadoop



All Participants Receive a Copy of the Hadoop 2 Quick-Start Guide

BREAK 2:30-2:45PM

Hands-on Exercises covering Pig, Hive, HBase, Spark

WRAP-UP and Q&A 4:15-4:30PM

About the Unique Hands-On Hadoop Cluster Experience

Unlike other Hadoop and Spark trainings, this workshop employs a live (on-site) four-node Hadoop cluster. The local nature of the cluster allows responsive and robust examples to be run as part of the workshop. This unique approach avoids slow or flakey Internet cloud connections and/or notebook pseudo clusters that impede the learning experience. The cluster is running the latest version of **Hortonworks Hadoop** software (HDP with Ambari) and is accessed via attendees laptop over a private wireless network.