Hadoop, Part 2 of 4: ETL and MapReduce

page 1

Meet the expert: Kevin McCarty is a computer professional with over 30 years of experience in the industry as a programmer, project manager, database administrator, architect, and data scientist. He is a Microsoft Certified Trainer with over 25 individual certifications in programming and database technologies and serves as the chapter leader of the Boise SQL Server Users Group. A former Army officer and Eagle Scout, he holds a doctorate in Computer Science and a lifelong love of learning.

Prerequisites: This course assumes that students have some programming background and some familiarity with a Unix-based operating system. No specific experience with Java programming language or Hadoop is required. As with any such course, the more experience you bring to the course, the more you'll get out of it. This course moves quickly through a broad range of topics, but it does not require any prior experience with Hadoop.
The course does assume that you are well familiarized with how to use the version of Windows that you are running. For example, the course might say simply "Open PuTTY" without explaining how to do that. You should also be able to navigate the folder hierarchy using Windows Explorer.

Runtime: 01:40:19

Course description: In this course, Hadoop expert Kevin McCarty takes a closer look at some of the major components underpinning Hadoop – services such as Mahout, Oozie, and ZooKeeper, and languages such as Pig and Hive. He will examine the Hadoop architecture and look at some ETL tools Hadoop provides for moving data between a Hadoop cluster and external servers. Finally, McCarty will demonstrate a simple application in Java and follow that up with a deep dive into MapReduce including a look at automation using the Linux Chron Utility

Course outline:

Big Data Sources And ETL
- Introduction
- Where Do You Find Big Data?
- Big Data Sources - Volume
- Big Data Sources - Variety
- Structured Data
- Semi-Structured
- Unstructured Data
- Problems with Big Data
- Data Integrity
- Data Completeness
- Data Format
- Data Timeliness
- How Do We Process Big Data?
- What Is ETL? - Extraction
- What Is ETL? - Transform
- What Is ETL? - Load
- Summary

ETL Demonstration
- Introduction
- In This Exercise...
- Demo: Sqoop
- Demo: Working with Tables
- Demo: ETL
- Summary

Understanding MapReduce
- Introduction
- What Is MapReduce?
- History of MapReduce
- MapReduce - Benefits
- MapReduce - Limitations
- Demo: MapReduce
- Demo: Create a Jar File
- Summary

MapReduce Demonstration
- Introduction
- Demo: MapReduce Setup
- Demo: Word Count Program
- Summary

Developing MapReduce
- Introduction
- Language Support
- How Streaming Works
- Creating a MapReduce Application
- MapReduce - Execution
- MapReduce - Main
- MapReduce - The Mapper
- MapReduce - The Reducer
- Demo: Create Java File
- Demo: MapReduce
- Demo: Map Method
- Demo: Reduce Function
- Summary

Schedule MapReduce
- Introduction
- Ad-Hoc vs. Scheduling
- Cron Jobs
- Cron Tables
- Creating a Cron Job
- Example Cron Job Text
- Demo: Cron Scheduling
- Summary