Introduction to Apache Hadoop

Overview

Everywhere you look today, enterprises are embracing big data-driven customer relationships and building innovative solutions based on the insights gained from various data sources. According to IBM, every day we create 2.5 quintillion bytes of data — so much that **90% of the data in the world today has been created in the last two years**. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, just to name a few. This data is big data.

The demand for storing this unprecedented amount of information is enough of a challenge, but when you add the need for analytics, the technology requirements truly start to strain even the state-of-the-art IT platforms. Fortunately, the Open Source community has stepped up to this challenge and collaboratively developed a storage and processing layer under the name of Apache Hadoop. Add the dozens of other projects integrating with Apache Hadoop and you have the whole Hadoop ecosystem.

The Hadoop ecosystem, along with the data management platform it enables, is growing at an unprecedented rate, with 73% of Hadoop cluster deployments now in production - a number which continues to rise.

The demand for individuals who have experience managing this platform is also accelerating. According to the IT Skills and Certifications Pay Index research from Foote Partners, “the need for big data skills also continues to lead to pay increases -- about 8% over the last year”. Now is exactly the right time to build an exciting and rewarding career managing big data with Apache Hadoop.

This introductory course is taught by Hadoop experts from The Linux Foundation’s ODPi collaborative project. As host to some of the world's leading open source projects, The Linux Foundation provides training and networking opportunities to help you advance your career.

This course is perfect for IT professionals seeking a high-level overview of Hadoop and some of its quintessential ecosystem projects, and who want to find out if a Hadoop-driven big data strategy is the right solution to meet the data retention and analytics needs of their IT organization. This course also helps anyone who wants to set up a small scale Hadoop test environment to gain experience working with this exciting open source technology.

Prerequisites

- Experience with Linux
- Basic familiarity with Java applications

Course Outline

Chapter 1. Welcome & Introduction

Chapter 2. Enterprise Data Management

Chapter 3. Core Hadoop Architecture

- HDFS Architecture
- YARN Architecture

Chapter 4. Parallel Processing

- Basic Principles of Parallel Processing
- MapReduce Overview
- Apache Spark Overview
- Apache Hive Overview

Chapter 5. Deploying Hadoop

Chapter 6. Hadoop Security

Chapter 7. The Road Ahead

Final Exam