Introduction to Hadoop Administration (TTDS6503)

Overview

Apache Hadoop is an open source framework for creating reliable and distributable compute clusters. Hadoop provides an excellent platform (with other related frameworks) to process large unstructured or semi-structured data sets from multiple sources to dissect, classify, learn from and make suggestions for business analytics, decision support, and other advanced forms of machine intelligence. This is an introductory-level, hands-on lab-intensive course geared for the administrator (new to Hadoop) who is charged with maintaining a Hadoop cluster and its related components. You will learn how to install, maintain, monitor, troubleshoot, optimize, and secure Hadoop.

Prerequisite Comments

This is an introductory-level course designed to teach experienced systems administrators how to install, maintain, monitor, troubleshoot, optimize, and secure Hadoop. Previous Hadoop experience is not required.
Next-Steps / Follow on Training: We offer a wide variety of Big Data, Analytics, Data Science, Hadoop, AI / Machine Learning, Python / R / Scala / Java Programming and other related courses that can advance your skills to the next level after this course. Please see our Big Data Training Suite course list for details, or please inquire for next step recommendations for follow on courses or Learning Plan options based on your role and goal.

Target Audience

This is an introductory-level course designed to teach experienced systems administrators how to install, maintain, monitor, troubleshoot, optimize, and secure Hadoop. Previous Hadoop experience is not required.

Course Objectives

Working within in an engaging, hands-on learning environment, guided by our expert team, attendees will learn to:
Understand the benefits of distributed computing
Understand the Hadoop architecture (including HDFS and MapReduce)
Define administrator participation in Big Data projects
Plan, implement, and maintain Hadoop clusters
Deploy and maintain additional Big Data tools (Pig, Hive, Flume, etc.)
Plan, deploy and maintain HBase on a Hadoop cluster
Monitor and maintain hundreds of servers
Pinpoint performance bottlenecks and fix them

Course Outline

1 - Introduction

Hadoop history and concepts
Ecosystem
Distributions
High level architecture
Hadoop myths
Hadoop challenges (hardware / software)

2 - Planning and installation

Selecting software and Hadoop distributions
Sizing the cluster and planning for growth
Selecting hardware and network
Rack topology
Installation
Multi-tenancy
Directory structure and logs
Benchmarking

3 - HDFS operations

Concepts (horizontal scaling, replication, data locality, rack awareness)
Nodes and daemons (NameNode, Secondary NameNode, HA Standby NameNode, DataNode)
Health monitoring
Command-line and browser-based administration
Adding storage and replacing defective drives

4 - MapReduce operations

Parallel computing before MapReduce: compare HPC versus Hadoop administration
MapReduce cluster loads
Nodes and Daemons (JobTracker, TaskTracker)
MapReduce UI walk through
MapReduce configuration
Job config
Job schedulers
Administrator view of MapReduce best practices
Optimizing MapReduce
Fool proofing MR: what to tell your programmers
YARN: architecture and use

New Horizons Estonia
+372 555 11 819
ADDRESS1
Tallinn, 10132

Contact Us
info@newhorizons.ee
www.newhorizons.com

5 - Advanced topics

Hardware monitoring
System software monitoring
Hadoop cluster monitoring
Adding and removing servers and upgrading Hadoop
Backup, recovery, and business continuity planning
Cluster configuration tweaks
Hardware maintenance schedule
Oozie scheduling for administrators
Securing your cluster with Kerberos
The future of Hadoop

New Horizons Estonia
+372 555 11 819
ADDRESS1
Tallinn, 10132

Contact Us
info@newhorizons.ee
www.newhorizons.com