

CLOUD COMPUTING

Revised 3/13/2012

/training/etc

The Art of Knowledge.

This Page Intentionally Left Blank

Table of Contents

Hadoop for Administrators.....	1
Hadoop for Developers.....	2
Introduction to Cloud Computing.....	3
Introduction to Cloud Computing/Hadoop for Administrators Combo.....	4

This Page Intentionally Left Blank

Course Description:

This course provides administrators with the fundamentals required to successfully implement and maintain Hadoop clusters. After an overview of Hadoop and its capabilities, you will examine best practices for deploying Hadoop clusters, determining hardware needs, and monitoring Hadoop clusters. You will see how to handle failures of Hadoop components and how to add and remove those components from your Hadoop cluster. In addition to exploring how to install Hadoop you will learn to install other related technologies such as Hive, Pig, and Accumulo.

Who Should Attend:

This course is intended for Administrators who are interested in learning how to deploy and manage a Hadoop cluster. Students have previous experience with UNIX or Linux.

Prerequisites:

Students should have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the student appreciate the more advanced data processing examples.

Benefits of Attendance:

Upon completion of this course, students will be able to:

- Set up Hadoop in a cluster and write data analytic programs
- Present design patterns and practices of programming MapReduce
- Grasp all the knobs and levers for running Hadoop
- Write meaningful programs in a MapReduce framework
- Understand basic concepts of MapReduce applications developed using Hadoop, including framework components
- Use Hadoop for a variety of data analysis tasks

Course Outline:**Hadoop Overview**

Why Hadoop?
HDFS Concepts
Blocks
Namenodes and Datanodes
MapReduce
Interfaces
Hive, Pig, HBase and other ecosystem projects

Planning a Hadoop Cluster

General Planning
Choosing Hardware
Node Topologies
Choosing the Software

Setting Up a Hadoop Cluster

Cluster Setup and Installation
Installing Java
Creating a Hadoop User
Installing Hadoop
Testing the Installation
SSH Configuration
Hadoop Configuration
Configuration Management
Environment Settings
Important Hadoop Daemon Properties
Hadoop Daemon Addresses and Ports
Other Hadoop Properties
Post Install
Benchmarking a Hadoop Cluster
Hadoop Benchmarks
User Jobs
Hadoop in the Cloud
Hadoop on Amazon EC2

Administering Hadoop

HDFS
Persistent Data Structures
Safe Mode
Audit Logging
Tools
Monitoring
Logging
Metrics
Java Management Extensions
Maintenance
Routine Administration Procedures
Commissioning and Decommissioning Nodes
Upgrades

Managing and Scheduling Jobs

Starting and stopping MapReduce jobs
Hands-On Exercise: Managing jobs
The FIFScheduler

The Fair Scheduler
Hands-On Exercise: Using the FairScheduler

Cluster Maintenance

Checking HDFS with fsck
Hands-On Exercise: Breaking the Cluster
Copying data with distcp
Rebalancing cluster nodes
Adding and removing cluster nodes
Backup And Restore
Upgrading and Migrating
The NameNode Metadata

Cluster Monitoring, Troubleshooting and Optimizing

Hadoop Log Files
Using the NameNode and JobTracker Web UIs
Interpreting Job Logs
Monitoring with Ganglia
Other monitoring tools
General Optimization Tips
Benchmarking Your Cluster

Populating HDFS from External Sources

Using Sqoop
Using Flume
Best Practices for Data Ingestion

Installing and Managing Other Hadoop Projects

Hive
Pig
HBase
Zookeeper
Accumulo

Case Studies

Course Description:

You will learn how to use Apache Hadoop and write MapReduce programs. You will begin with a quick overview of installing Hadoop, setting it up in a cluster, and then proceed to writing data analytic programs. The course will present the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. The course will further examine related technologies such as Hive, Pig, and Apache Accumulo. Apache Accumulo is a highly scalable structured store based on Google's BigTable, written in Java and operates over the Hadoop Distributed File System (HDFS). Hive is data warehouse software for querying and managing large datasets. Pig is a platform to take advantage of parallelization when running data analysis. Finally, you will observe how Hadoop works in and supports cloud computing and explore examples with Amazon Web Services and case studies.

Who Should Attend:

The course is intended for programmers, architects, and project managers who have to process large amounts of data offline.

Prerequisites:

Students should have a basic familiarity with Linux administration and Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) is helpful to appreciate the more advanced data processing examples.

Benefits of Attendance:

Upon completion of this course, students will be able to:

- Understand basic concepts of MapReduce applications developed using Hadoop
- Understand how Hadoop works in and supports cloud computing and explore examples with Amazon Web Services and case studies
- Use Apache Hadoop and write MapReduce programs

Course Outline:**What is Hadoop?**

Understanding distributed systems and Hadoop
Comparing SQL databases and Hadoop
Understanding MapReduce
Counting words with Hadoop—running your first program
History of Hadoop

Starting Hadoop

The building blocks of Hadoop
Setting up SSH for a Hadoop cluster
Running Hadoop
Web-based cluster UI

Components of Hadoop

Working with files in HDFS
Anatomy of a MapReduce program
Reading and writing

Writing basic MapReduce programs

Constructing the basic template of a MapReduce program
Counting things
Adapting for Hadoop's API changes
Streaming in Hadoop
Improving performance with combiners

Advanced MapReduce

Chaining MapReduce jobs
Joining data from different sources
Creating a Bloom filter

Programming Practices

Developing MapReduce programs
Monitoring and debugging on a production cluster
Tuning for performance

Cookbook

Passing job-specific parameters to your tasks
Probing for task-specific information
Partitioning into multiple output files
Inputting from and outputting to a database
Keeping all output in sorted order

Managing Hadoop

Setting up parameter values for practical use
Checking system's health
Setting permissions
Managing quotas
Enabling trash
Removing DataNodes
Adding DataNodes
Managing NameNode and Secondary NameNode

Recovering from a failed NameNode
Designing network layout and rack awareness
Scheduling jobs from multiple users

Running Hadoop in the cloud

Introducing Amazon Web Services
Setting up AWS
Setting up Hadoop on EC2
Running MapReduce programs on EC2
Cleaning up and shutting down your EC2 instances
Amazon Elastic MapReduce and other AWS services

Programming with Pig

Installing Pig
Running Pig
Learning Pig Latin through Grunt
Speaking Pig Latin
Working with user-defined functions
Working with scripts
Seeing Pig in action—example of computing similar patents

Hadoop Related Technologies

Hive
Apache Accumulo
Other Hadoop-related stuff

Case studies

Converting 11 million image documents from the New York Times archive
Mining data at China Mobile
Recommending the best websites at StumbleUpon
Building analytics for enterprise search—IBM's Project ES2

Course Description:

This course will cover the benefits and challenges of cloud computing. Students will define cloud computing and learn methods to assess the appropriateness of in-house or hosted solutions. We will examine the suitability of several cloud technologies and walk through the steps to choose a solution, calculate costs, and develop deployment and training plans.

Who Should Attend:

This class is for Network Administrators interested in learning about Cloud Computing and becoming Cloud Administrators.

Prerequisites:

Students should have Network Plus experience.

Benefits of Attendance:

Upon completion of this course, students will be able to:

- Describe the benefits and challenges of cloud computing
- Define cloud computing
- Assess the appropriateness of in-house or hosted solutions
- Examine the suitability of cloud technologies and choose the appropriate solution
- Calculate costs
- Develop deployment and training plans

Course Outline:**Define Cloud Computing**

Common Definitions
Architecture
Infrastructure as a Service
Platform as a Service
Software as a Service

Assess Available Technologies

Weigh Benefits and Challenges
Strategic Analysis
Risk Impact
Financial Impact

Design Cloud Solution

Requirements Analysis
Draft an Architecture
Preliminary Design

Select Cloud Technology

Take Application Inventory
Find Stakeholders and Business Criteria
Find Technical Criteria
Service Delivery Model
User Profiles and Configuration
Identity Management

Integrate with Current Technology

Technical Design
Connectivity
Physical Infrastructure
Availability
Business Continuity / Disaster Recovery
Security

Implementation

Transitioning to the Cloud
Migrate Code
Migrate Data
Employee Training

Operation

Service Strategy and Design
Administration
Service Request and Change Management
Monitoring
End-user and IT Support

Controlling Your Cloud

Ensuring Compliance
Ensuring Privacy
Responding to Incidents
Governance

Adapting Your Cloud

Improvement Processes
Working with New Technology

Future Trends**Vendors Overview**

Course Description:

This course combines the following courses into a single course:

Hadoop for Administrators
Introduction to Cloud Computing

The description of each course is listed below:

This course provides administrators with the fundamentals required to successfully implement and maintain Hadoop clusters. After an overview of Hadoop and its capabilities, you will examine best practices for deploying Hadoop clusters, determining hardware needs, and monitoring Hadoop clusters. You will see how to handle failures of Hadoop components and how to add and remove those components from your Hadoop cluster. In addition to exploring how to install Hadoop you will learn to install other related technologies such as Hive, Pig, and Accumulo.

This course will cover the benefits and challenges of cloud computing. Students will define cloud computing and learn methods to assess the appropriateness of in-house or hosted solutions. We will examine the suitability of several cloud technologies and walk through the steps to choose a solution, calculate costs, and develop deployment and training plans.

Who Should Attend:

This course combines several courses, the descriptions for who should attend are listed below: This course is intended for Administrators who are interested in learning how to deploy and manage a Hadoop cluster. Students have previous experience with UNIX or Linux. This class is for Network Administrators interested in learning about Cloud Computing and becoming Cloud Administrators.

Prerequisites:

Students should have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the student appreciate the more advanced data processing examples. Students should have Network Plus experience.

Benefits of Attendance:

Upon completion of this course, students will be able to:

- Set up Hadoop in a cluster and write data analytic programs
- Present design patterns and practices of programming MapReduce
- Grasp all the knobs and levers for running Hadoop
- Write meaningful programs in a MapReduce framework
- Understand basic concepts of MapReduce applications developed using Hadoop, including framework components
- Use Hadoop for a variety of data analysis tasks
- Describe the benefits and challenges of cloud computing
- Define cloud computing
- Assess the appropriateness of in-house or hosted solutions
- Examine the suitability of cloud technologies and choose the appropriate solution
- Calculate costs
- Develop deployment and training plans

Course Outline:

<p>Hadoop Overview</p> <ul style="list-style-type: none"> Why Hadoop? HDFS Concepts Blocks Namenodes and Datanodes MapReduce Interfaces Hive, Pig, HBase and other ecosystem projects <p>Planning a Hadoop Cluster</p> <ul style="list-style-type: none"> General Planning Choosing Hardware Node Topologies Choosing the Software <p>Setting Up a Hadoop Cluster</p> <ul style="list-style-type: none"> Cluster Setup and Installation Installing Java Creating a Hadoop User Installing Hadoop Testing the Installation SSH Configuration Hadoop Configuration Configuration Management Environment Settings Important Hadoop Daemon Properties Hadoop Daemon Addresses and Ports 	<ul style="list-style-type: none"> Other Hadoop Properties Post Install Benchmarking a Hadoop Cluster Hadoop Benchmarks User Jobs Hadoop in the Cloud Hadoop on Amazon EC2 <p>Administering Hadoop</p> <ul style="list-style-type: none"> HDFS Persistent Data Structures Safe Mode Audit Logging Tools Monitoring Logging Metrics Java Management Extensions Maintenance Routine Administration Procedures Commissioning and Decommissioning Nodes Upgrades <p>Managing and Scheduling Jobs</p> <ul style="list-style-type: none"> Starting and stopping MapReduce jobs Hands-On Exercise: 	<ul style="list-style-type: none"> Managing jobs The FIFScheduler The Fair Scheduler Hands-On Exercise: Using the FairScheduler <p>Cluster Maintenance</p> <ul style="list-style-type: none"> Checking HDFS with fsck Hands-On Exercise: Breaking the Cluster Copying data with distcp Rebalancing cluster nodes Adding and removing cluster nodes Backup And Restore Upgrading and Migrating The NameNode Metadata <p>Cluster Monitoring, Troubleshooting and Optimizing</p> <ul style="list-style-type: none"> Hadoop Log Files Using the NameNode and JobTracker Web UIs Interpreting Job Logs Monitoring with Ganglia Other monitoring tools General Optimization Tips Benchmarking Your Cluster 	<p>Populating HDFS from External Sources</p> <ul style="list-style-type: none"> Using Sqoop Using Flume Best Practices for Data Ingestion <p>Installing and Managing Other Hadoop Projects</p> <ul style="list-style-type: none"> Hive Pig HBase Zookeeper Accumulo <p>Case Studies</p> <p>Define Cloud Computing</p> <ul style="list-style-type: none"> Common Definitions Architecture Infrastructure as a Service Platform as a Service Software as a Service <p>Assess Available Technologies</p> <ul style="list-style-type: none"> Weigh Benefits and Challenges Strategic Analysis Risk Impact 	<ul style="list-style-type: none"> Financial Impact <p>Design Cloud Solution</p> <ul style="list-style-type: none"> Requirements Analysis Draft an Architecture Preliminary Design <p>Select Cloud Technology</p> <ul style="list-style-type: none"> Take Application Inventory Find Stakeholders and Business Criteria Find Technical Criteria Service Delivery Model User Profiles and Configuration Identity Management <p>Integrate with Current Technology</p> <ul style="list-style-type: none"> Technical Design Connectivity Physical Infrastructure Availability Business Continuity / Disaster Recovery Security <p>Implementation</p> <ul style="list-style-type: none"> Transitioning to the Cloud Migrate Code Migrate Data 	<ul style="list-style-type: none"> Employee Training <p>Operation</p> <ul style="list-style-type: none"> Service Strategy and Design Administration Service Request and Change Management Monitoring End-user and IT Support <p>Controlling Your Cloud</p> <ul style="list-style-type: none"> Ensuring Compliance Ensuring Privacy Responding to Incidents Governance <p>Adapting Your Cloud</p> <ul style="list-style-type: none"> Improvement Processes Working with New Technology <p>Future Trends</p> <p>Vendors Overview</p>
--	---	---	--	---	---