

Introduction to Big Data

- What is Big Data
- Big Data Facts
- The Three V's and Big Data

Understanding Hadoop

- What is Hadoop?
- Why learn Hadoop?
- Relational Databases Vs. Hadoop
- Motivation for Hadoop
- 6 Key Hadoop Data Types

The Hadoop Distributed File system (HDFS)

- What is HDFS?
- HDFS components
- Understanding Block storage
- The Name Node
- The Data Nodes
- Data Node Failures
- HDFS Commands
- HDFS File Permissions

The MapReduce Framework

- Overview of MapReduce
- Understanding MapReduce
- The Map Phase
- The Reduce Phase
- WordCount in MapReduce
- Running MapReduce Job

Planning Your Hadoop Cluster

- Single Node Cluster Configuration
- Multi-Node Cluster Configuration

Cluster Maintenance

- Checking HDFS Status
- Breaking the cluster
- Copying Data Between Clusters
- Adding and Removing Cluster Nodes
- Rebalancing the cluster
- Name Node Metadata Backup
- Cluster Upgrading

Installing and Managing Hadoop Ecosystem Projects

- Sqoop
- Flume
- Hive
- Pig
- HBase
- Oozie

Managing and Scheduling Jobs

- Managing Jobs
- The FIFO Scheduler
- The Fair Schedule
- How to stop and start jobs running on the cluster

Cluster Monitoring, Troubleshooting, and Optimizing

- General System conditions to Monitor
- Name Node and Job Tracker Web Uis
- View and Manage Hadoop's Log files
- Ganglia Monitoring Tool
- Common cluster issues and their resolutions
- Benchmark your cluster's performance

Populating HDFS from External Sources

How to use Sqoop to import data from RDBMSs to HDFS
How to gather logs from multiple systems using Flume
Features of Hive, Hbase and Pig
How to populate HDFS from external Sources