Cloudera Administrator Training for Apache Hadoop

Varighet: 3.00 Days **Kurskode: GK3901**

Beskrivelse:

In this hands-on course, you will be introduced to the basics of Hadoop, Hadoop Distributed File System (HDFS), MapReduce, Hive, Pig, and HBase. You will cover core administration skills, such as cluster deployment, job management, and ongoing Hadoop maintenance and monitoring, as you gain the expertise to support your environments in day-to-day activities.
This course covers concepts addressed on the Cloudera Certified Administrator for Apache Hadoop (CCAH) exam and includes a CCAH exam voucher you'll receive at the end of class.

Målgruppe:

System administrators looking to understand all of the steps necessary to operate and manage Apache Hadoop clusters

Agenda:

- In this course you will learn:
- HDFS and MapReduce
- Optimal hardware configurations for Hadoop clusters
- Network considerations to take into account when building out your cluster
- Configure Hadoop options for best cluster performance
- Configure the FairScheduler to provide service-level agreements for multiple users of a cluster
- Maintain and monitor your cluster
- Load data into the cluster from dynamically generated files using Flume and from relational database management systems using Sqoop
- System administration issues with other Hadoop projects such as Hive, Pig, and HBase

Forkunnskaper:

- Basic level of Linux system administration experience
- Prior knowledge of Apache Hadoop is not required

Test og sertifisering

This course is part of the following programs or tracks:

- CCAH: Cloudera Certified Administrator for Apache Hadoop (CDH3)

Påfølgende kurs:

- Cloudera Training for Apache HBase
- Cloudera Training for Apache Hive and Pig

Innhold:

Hadoop and HDFS

- Why Hadoop?
- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Managing and Scheduling Jobs

- Starting and Stopping MapReduce Jobs
- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore
- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia

- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Planning Your Hadoop Cluster

- General Planning Considerations

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore

- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Cluster Maintenance

- Checking HDFS with Fsck

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating

- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster

- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes

- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes

- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Populating HDFS from External Sources

- Using Sqoop

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker

- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Deploying Your Cluster

- Installing Hadoop

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters

- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp

- Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Installing and Managing Other Hadoop Projects

- Hive

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler

- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software

- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Cluster Monitoring, Troubleshooting, and Optimizing

- Hadoop Log Files

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software

- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Labs

- Install a Pseudo-Distributed Cluster
- Install a Hadoop Cluster
- Manage Jobs
- Use the FairScheduler
- Break the Cluster
- Verify the Cluster's Self-Healing Features
- Back Up and Restoring
- Configure the Hive Shared

- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools
- FIFO Scheduler
- Fair Scheduler
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata
- Using the NameNode and JobTracker Web UIs
- Interpreting Job Logs
- Monitoring with Ganglia
- Other Monitoring Tools
- General Optimization Tips
- Benchmarking Your Cluster
- Using Flume
- Best Practices for Data Ingestion
- Pig
- HBase
- Metastore

Ytterligere informasjon:

For mer informasjon eller kursbooking, vennligst ring oss 22 95 66 00

info@globalknowledge.no

www.globalknowledge.no

Grenseveien 97, 0663 Oslo, PO Box 6256 Etterstad, 0606 Oslo, Norway