Cloudera Administrator Training for Apache Hadoop

Varighed: 3.00 Days   **Kursus Kode: GK3901**

Beskrivelse:

Learn to deploy, configure, and manage Cloudera's Apache Hadoop implementation and HDFS.
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:

- 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

Forudsætninger:

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

Test og certificering

- This course is part of the following programs or tracks:
- CCAH: Cloudera Certified Administrator for Apache Hadoop (CDH3)

Yderligere Kurser:
- Cloudera Training for Apache HBase
- Cloudera Training for Apache Hive and Pig

Indhold:

1. Hadoop and HDFS

- Why Hadoop?
- HDFS
- MapReduce
- Hive, Pig, HBase, and Other Ecosystem Projects

2. Planning Your Hadoop Cluster

- General Planning Considerations
- Choosing the Right Hardware
- Node Topologies
- Choosing the Right Software

3. Deploying Your Cluster

- Installing Hadoop
- Using SCM Express for Easy Installation
- Typical Configuration Parameters
- Configuring Rack Awareness
- Using Configuration Management Tools

4. Managing and Scheduling Jobs

- Starting and Stopping MapReduce Jobs
- FIFO Scheduler
- Fair Scheduler

5. Cluster Maintenance

- Checking HDFS with Fsck
- Copying Data with Distcp
- Rebalancing Cluster Nodes
- Adding and Removing Cluster Nodes
- Backup and Restore
- Upgrading and Migrating
- NameNode Metadata

6. 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

7. Populating HDFS from External Sources

- Using Sqoop
- Using Flume
- Best Practices for Data Ingestion

8. Installing and Managing Other Hadoop Projects

- Hive
- Pig
- HBase
- Metastore

Labs

- Lab 1: Install a Pseudo-Distributed Cluster
- Lab 2: Install a Hadoop Cluster
- Lab 3: Manage Jobs
- Lab 4: Use the FairScheduler
- Lab 5: Break the Cluster
- Lab 6: Verify the Cluster's Self-Healing Features
- Lab 7: Back Up and Restoring
- Lab 8: Configure the Hive Shared

Flere Informationer:

For yderligere informationer eller booking af kursus, kontakt os på tlf.nr.: 44 88 18 00

training@globalknowledge.dk

www.globalknowledge.dk

Global Knowledge, Stamholmen 110, 2650 Hvidovre