Hadoop Administration

Tecnologias de Informação - Data & Analytics

- **Nível:**
- **Duração:** 28h

Sobre o curso

This course is designed for student that need to learn how to master Hadoop Admin activities like planning, installation, monitoring, configuration and performance tuning of large and complex Hadoop clusters.

You will learn to implement security using Kerberos and Hadoop YARN features using real-life use cases.

Destinatários

- This course targets IT professionals that want to work as effective Hadoop administrators.

Objetivos

- Learn how to install and configure a secure and stable Hadoop cluster.
- Understand the architecture, requirements and role of the individual components of core Hadoop.
- Be prepared to troubleshoot problems with Hadoop clusters and tune cluster performance.

Pré-requisitos

- Participants of this course should be comfortable with basic Linux system administration and basic scripting skills.
- Knowledge of Hadoop and Distributed Computing is not required.

Programa

- Hadoop Architecture and Deployment
- Maintaining Hadoop Cluster HDFS
- Maintaining Hadoop Cluster – YARN and MapReduce
- High Availability
- Schedulers

- Backup and Recovery
- Data Ingestion and Workflow
- Performance Tuning
- HBase Administration
- Cluster Planning
- Troubleshooting and Diagnostics
- Security

Hadoop Architecture and Deployment

- Overview of Hadoop Architecture
- Building and compiling Hadoop
- Installation methods
- Setting up host resolution
- Installing a single-node cluster – HDFS components
- Installing a single-node cluster – YARN components
- Installing a multi-node cluster
- Configuring Hadoop Gateway node
- Decommissioning nodes
- Adding nodes to the cluster

Maintaining Hadoop Cluster HDFS

- Configuring HDFS block size
- Setting up Namenode metadata location
- Loading data into HDFS
- Configuring HDFS replication
- HDFS balancer
- Quota configuration
- HDFS health and FSCK
- Configuring rack awareness
- Recycle or trash bin configuration
- Distcp usage
- Controlling block report storm
- Configuring Datanode heartbeat

Maintaining Hadoop Cluster – YARN and MapReduce

- Running a simple MapReduce program
- Hadoop streaming
- Configuring YARN history server
- Job history web interface and metrics
- Configuring ResourceManager components
- YARN containers resource allocations
- ResourceManager Web UI and JMX metrics
- Preserving ResourceManager states

High Availability

- Namenode HA using shared storage
- ZooKeeper configuration
- Namenode HA using Journal node
- Resourcemanager HA using ZooKeeper
- Rolling upgrade in HA
- Configuring shared cache manager
- Configuring HDFS cache
- HDFS snapshots
- Configuring storage-based policies
- Configuring HA for Edge nodes

Schedulers

- Configuring users and groups
- Fair Scheduler configuration
- Fair Scheduler pools
- Configuring job queues
- Job queue ACLs
- Configuring Capacity Scheduler
- Queuing mappings in Capacity Scheduler
- YARN and Mapred commands
- YARN label-based scheduling
- YARN SLS

Backup and Recovery

- Initiating Namenode saveNamespace
- Using HDFS Image Viewer
- Fetching parameters which are in-effect
- Configuring HDFS and YARN logs
- Backing up and recovering Namenode
- Configuring Secondary Namenode
- Promoting Secondary Namenode to Primary
- Namenode recovery
- Namenode roll edits – Online mode
- Namenode roll edits – Offline mode
- Datanode recovery – Disk full
- Configuring NFS gateway to serve HDFS
- Recovering deleted files

Data Ingestion and Workflow

- Hive server modes and setup
- Using MySQL for Hive metastore
- Operating Hive with ZooKeeper

- Loading data into Hive
- Partitioning and Bucketing in Hive
- Hive metastore database
- Designing Hive with credential store
- Configuring Flume
- Configure Oozie and workflows

Performance Tuning

- Tuning the operating system
- Tuning the disk
- Tuning the network
- Tuning HDFS
- Tuning Namenode
- Tuning Datanode
- Configuring YARN for performance
- Configuring MapReduce for performance
- Hive performance tuning
- Benchmarking Hadoop cluster

HBase Administration

- Setting up single node HBase cluster
- Setting up multi-node HBase cluster
- Inserting data into HBase
- Integration with Hive
- HBase administration commands
- HBase backup and restore
- Tuning HBase
- HBase upgrade
- Migrating data from MySQL to HBase using Sqoop

Cluster Planning

- Disk space calculations
- Nodes needed in the cluster
- Memory requirements
- Sizing the cluster as per SLA
- Network design
- Estimating the cost of the Hadoop cluster
- Hardware and software options

Troubleshooting and Diagnostics

- Namenode troubleshooting
- Datanode troubleshooting
- Resourcemanager troubleshooting

- Diagnose communication issues
- Parse logs for errors
- Hive troubleshooting
- HBase troubleshooting

Security

- Encrypting disk using LUKS
- Configuring Hadoop users
- HDFS encryption at Rest
- Configuring SSL in Hadoop
- In-transit encryption
- Enabling service level authorization
- Securing ZooKeeper
- Configuring auditing
- Configuring Kerberos server
- Configuring and enabling Kerberos for Hadoop