Hadoop Deep Dive

Tecnologias de Informação - Data & Analytics

- **Nível:**
- **Duração:** 40h

Sobre o curso

Apache Hadoop is one of the most popular frameworks for processing Big Data on clusters of servers. This Deep Dive course delves into data management in HDFS, advanced Pig, Hive, HBase, etc.

These advanced programming techniques will be beneficial to experienced Hadoop developers.

Destinatários

- This course targets professional developers who are looking for in-depth knowledge for Data Science with Apache Hadoop.

Objetivos

- Learn how to write HDFS/Mapreduce programs
- Learn how to write & utilize effectively Hive & Pig Scripts
- Understand basically how the administration part is even handled for a cluster setup
- Learn how to write & utilize effectively Flume & Zookeeper mechanisms
- Understand better on the internal architecture/design involved on all the Hadoop platforms
- Understand how to enhance your coding skills using Hbase & Sqoop tools
- To understand better on how a real time projects fit into the big data platform

---

Pré-requisitos

- Participants of this course need to have a good understanding of Java programming and SQL.

Programa

- Deep Dive into the Hadoop Distributed File System
- YARN Resource Management in Hadoop
- Internals of MapReduce
- SQL on Hadoop
- Real-Time Processing Engines
- Widely Used Hadoop Ecosystem Components
- Designing Applications in Hadoop
- Real-Time Stream Processing in Hadoop
- Machine Learning in Hadoop
- Hadoop in the Cloud
- Hadoop Cluster Profiling

Deep Dive into the Hadoop Distributed File System

- Defining HDFS
- Deep dive into the HDFS architecture
- NameNode internals
- DataNode internals
- Quorum Journal Manager (QJM)
- HDFS high availability in Hadoop 3.x
- Data management
- HDFS reads and writes
- Managing disk-skewed data in Hadoop 3.x
- Lazy persist writes in HDFS
- Erasure encoding in Hadoop 3.x
- HDFS common interfaces
- HDFS command reference

YARN Resource Management in Hadoop

- Architecture
- YARN job scheduling
- FIFO scheduler
- Capacity scheduler
- Fair scheduler
- Resource Manager high availability
- Node labels

- YARN Timeline server in Hadoop 3.x
- Opportunistic containers in Hadoop 3.x
- Docker containers in YARN
- YARN REST APIs
- YARN command reference

Internals of MapReduce

- Deep dive into the Hadoop MapReduce framework
- YARN and MapReduce
- MapReduce workflow in the Hadoop framework
- Common MapReduce patterns
- MapReduce use case
- Optimizing MapReduce

SQL on Hadoop

- Hive

Real-Time Processing Engines

- Spark
- Storm

Widely Used Hadoop Ecosystem Components

- Pig
- HBase
- Kafka
- Flume

Designing Applications in Hadoop

- File formats
- Data compression
- Serialization
- Data ingestion
- Data processing
- Common batch processing pattern
- Airflow for orchestration
- Data governance

Real-Time Stream Processing in Hadoop

- What are streaming datasets?
- Stream data ingestion
- Common stream data processing patterns
- Streaming design considerations
- Micro-batch processing case study
- Real-time processing case study

Machine Learning in Hadoop

- Machine learning steps
- Common machine learning challenges
- Spark machine learning
- Hadoop and R
- Mahout
- Machine learning case study in Spark

Hadoop in the Cloud

- Introduction to Cloudera
- Logical view of Hadoop in the cloud
- Network
- Managing resources
- Data pipelines
- High availability (HA)

Hadoop Cluster Profiling

- Benchmarking and profiling
- HDFS
- NameNode
- YARN
- Hive
- Mix-workloads