Developing Solutions Using Apache Hadoop

Course description

Training course is designed for developers who want to better understand how to create Apache Hadoop solutions. This 30 Hours provides Java programmers the necessary training for creating enterprise solutions using Apache Hadoop. It consists of a prudent combination of interactive lecture and extensive hand-on lab exercises.

Student Take away

- Study Material
- Learning stuff
- Sample project for practice

Developing Solutions Using Apache Hadoop online training curriculum

Introduction

- What is Big Data?
- Introduction to Analytics and the need for big data analytics
- Hadoop Solutions - Big Picture
- Hadoop distributions
- Apache Hadoop
- Cloud era Hadoop
- Horton Works and Other Hadoop distributions
- Comparing Hadoop Vs. Traditional systems
- Data Retrieval - Radom Access Vs. Sequential Access

Hadoop: Basic Concepts

- What is Hadoop?
- The Hadoop Distributed File System
- Hadoop Map Reduce Works
- Anatomy of a Hadoop Cluster

➢ **Hadoop demons**

- **Master Daemons**
- **Slave Daemons**

Master Daemons

- Name node
- Job Tracker
- Secondary name node

Slave Daemons

- Job tracker
- Task tracker

HDFS (Hadoop Distributed File System)

Blocks and Splits

- Input Splits
- HDFS Splits

- **Data Replication**
- **Data high availability**
- **Data Integrity**
- **Cluster architecture and block placement**

Accessing HDFS

- JAVA Approach
- CLI Approach

Practices & Programming Performance Tuning

Developing Map Reduce Programs in Local Mode

- Running without HDFS and Map reduce

Developing Map Reduce Programs in Pseudo-distributed Mode

- Running all daemons in a single node

Developing Map Reduce Programs in Fully distributed mode

- Running daemons on dedicated nodes

- Map Reduce
- Map Reduce Overview
- Word Count Problem
- Word Count Flow and Solution
- Map Reduce Flow
- Algorithms for simple problems
- Algorithms for complex problems

Developing the Map Reduce Application

- Data Types
- File Formats
- Explain the Driver, Mapper and Reducer code
- Running locally
- Running on Cluster
- Hands on exercises

How Map-Reduce Works?

- Anatomy of Map Reduce Job runs
- Job Submission
- Job Initialization
- Task Assignment
- Job Completion
- Job Scheduling
- Job Failures
- Shuffle and sort
- Hands on Exercises

Hadoop Ecosystem

Hive

- Hive concepts
- Hive architecture
- Hive meta-store
- Install and configure hive on cluster
- Create database, access it from java client
- Partitions
- Buckets
- Hive JDBC
- Joins in hive
- Inner joins
- Outer Joins
- Hive UDF
- Hive UDAF
- Hive UDTF
- Develop and run sample applications in Java to access hive
- Load New York Stock Exchange data into Hive and process it using Hive

PIG

- Pig basics
- Install and configure PIG on a cluster
- PIG Vs Map Reduce and SQL
- Pig Vs Hive
- Write sample Pig Latin scripts
- Modes of running PIG
- Local mode
- Map Reduce mode
- PIG UDFs

PIG Macros

Hbase

- HBase and Zookeeper concepts
- HBase architecture
- Region server architecture
- File storage architecture
- HBase basics
- Column access
- Scans
- HBase use cases
- Install and configure HBase on a multi node cluster
- Create database, Develop and run sample applications
- Access data stored in HBase using clients like Java, Python and Pearl Map Reduce client to access the HBase data

Sqoop

- Install and configure Sqoop on cluster
- Connecting to RDBMS
- Installing MySql
- Import data from Oracle/MySql to hive
- Export data to Oracle/MySql
- Internal mechanism of import/export
- Import millions of records into HDFS from RDBMS using Sqoop

YARN and MR2

- YARN
- MR2 Demons (Resource Manager, Application Manager, App Master, Node Mgr)
- Execution of an MR2 Job
- Name Node High Availability
- Name Node federation

Hadoop Administrative Tasks

Setup Hadoop cluster

- Install and configure Hadoop
- Make a fully distributed Hadoop cluster on a single laptop/desktop
- Monitoring the cluster
- Performance Tuning

Hadoop POC/real-life experience