WEB TRACKKER

Hadoop

Course Content

Hadoop Course Content

- ✓ Hadoop Overview, Architecture Considerations, Infrastructure, Platforms and Automation

Use case walkthrough

- ✓ ETL
- ✓ Log Analytics
- ✓ Real Time Analytics

Hbase for Developers :

NoSQL Introduction

- ✓ Traditional RDBMS approach
- ✓ NoSQL introduction
- ✓ Hadoop & Hbase positioning

Hbase Introduction

- ✓ What it is, what it is not, its history and common use-cases
- ✓ Hbase Client – Shell, exercise

Hbase Architecture

- ✓ Building Components
- ✓ Storage, B+ tree, Log Structured Merge Trees
- ✓ Region Lifecycle
- ✓ Read/Write Path

Hbase Schema Design

- ✓ Introduction to hbase schema
- ✓ Column Family, Rows, Cells, Cell timestamp
- ✓ Deletes
- ✓ Exercise - build a schema, load data, query data

Hbase Java API – Exercises

- ✓ Connection
- ✓ CRUD API
- ✓ Scan API
- ✓ Filters
- ✓ Counters
- ✓ Hbase MapReduce
- ✓ Hbase Bulk load

Hbase Operations, cluster management

- ✓ Performance Tuning
- ✓ Advanced Features
- ✓ Exercise
- ✓ Recap and Q&A

MapReduce for Developers

Introduction

- ✓ Traditional Systems / Why Big Data / Why Hadoop
- ✓ Hadoop Basic Concepts/Fundamentals

Hadoop in the Enterprise

- ✓ Where Hadoop Fits in the Enterprise
- ✓ Review Use Cases

Architecture

- ✓ Hadoop Architecture & Building Blocks
- ✓ HDFS and MapReduce

Hadoop CLI

- ✓ Walkthrough
- ✓ Exercise

MapReduce Programming

- ✓ Fundamentals
- ✓ Anatomy of MapReduce Job Run
- ✓ Job Monitoring, Scheduling
- ✓ Sample Code Walk Through
- ✓ Hadoop API Walk Through
- ✓ Exercise

MapReduce Formats

- ✓ Input Formats, Exercise
- ✓ Output Formats, Exercise

Hadoop File Formats

MapReduce Design Considerations

Hadoop File Formats

MapReduce Algorithms

- ✓ Walkthrough of 2-3 Algorithms

MapReduce Features

- ✓ Counters, Exercise
- ✓ Map Side Join, Exercise
- ✓ Reduce Side Join, Exercise
- ✓ Sorting, Exercise

Use Case A (Long Exercise)

- ✓ Input Formats, Exercise
- ✓ Output Formats, Exercise

MapReduce Testing

Hadoop Ecosystem

- ✓ Oozie
- ✓ Flume
- ✓ Sqoop
- ✓ Exercise 1 (Sqoop)
- ✓ Streaming API
- ✓ Exercise 2 (Streaming API)

- ✓ Hcatalog
- ✓ Zookeeper

HBase Introduction

- ✓ Introduction
- ✓ HBase Architecture

VIEW Types

- ✓ Default Views
- ✓ Overriden Views
- ✓ Normal Views

MapReduce Performance Tuning

Development Best Practice and Debugging

Apache Hadoop for Administrators Hadoop Fundamentals and Architecture

- ✓ Why Hadoop, Hadoop Basics and Hadoop Architecture
- ✓ HDFS and Map Reduce

Hadoop Ecosystems Overview

- ✓ Hive
- ✓ Hbase
- ✓ ZooKeeper
- ✓ Pig
- ✓ Mahout
- ✓ Flume
- ✓ Sqoop
- ✓ Oozie

Hardware and Software requirements

- ✓ Hardware, Operating System and Other Software
- ✓ Management Console

Deploy Hadoop ecosystem services

- ✓ Hive
- ✓ ZooKeeper
- ✓ HBase
- ✓ Administration
- ✓ Pig
- ✓ Mahout
- ✓ Mysql
- ✓ Setup Security

Enable Security – Configure Users, Groups, Secure HDFS, MapReduce, HBase and Hive

- ✓ Configuring User and Groups
- ✓ Configuring Secure HDFS
- ✓ Configuring Secure MapReduce
- ✓ Configuring Secure HBase and Hive

Manage and Monitor your cluster

Command Line Interface

Troubleshooting your cluster

Introduction to Big Data and Hadoop Hadoop Overview

- ✓ Why Hadoop
- ✓ Hadoop Basic Concepts
- ✓ Hadoop Ecosystem – MapReduce, Hadoop Streaming, Hive, Pig, Flume, Sqoop, Hbase, Oozie, Mahout
- ✓ Where Hadoop fits in the Enterprise
- ✓ Review use cases

Apache Hive & Pig for Developers Overview of Hadoop

- ✓ Why Hadoop
- ✓ Hadoop Basic Concepts
- ✓ Hadoop Ecosystem – MapReduce, Hadoop Streaming, Hive, Pig, Flume, Sqoop, Hbase, Oozie, Mahout

- ✓ Where Hadoop fits in the Enterprise

Overview of Hadoop

- ✓ Big Data and the Distributed File System
- ✓ MapReduce

Hive Introduction

- ✓ Why Hive?
- ✓ Compare vs SQL
- ✓ Use Cases

Hive Architecture – Building Blocks

- ✓ Hive CLI and Language (Exercise)
- ✓ HDFS Shell
- ✓ Hive CLI
- ✓ Data Types
- ✓ Hive Cheat-Sheet
- ✓ Data Definition Statements
- ✓ Data Manipulation Statements
- ✓ Select, Views, GroupBy, SortBy/DistributeBy/ClusterBy/OrderBy, Joins
- ✓ Built-in Functions
- ✓ Union, Sub Queries, Sampling, Explain

Hive Architecture – Building Blocks

- ✓ Hive CLI and Language (Exercise)
- ✓ HDFS Shell
- ✓ Hive CLI
- ✓ Data Types
- ✓ Hive Cheat-Sheet
- ✓ Data Definition Statements
- ✓ Data Manipulation Statements
- ✓ Select, Views, GroupBy, SortBy/DistributeBy/ClusterBy/OrderBy, Joins
- ✓ Built-in Functions
- ✓ Union, Sub Queries, Sampling, Explain

Hive Architecture – Building Blocks

- ✓ Hive CLI and Language (Exercise)
- ✓ HDFS Shell
- ✓ Hive CLI
- ✓ Data Types
- ✓ Hive Cheat-Sheet
- ✓ Data Definition Statements
- ✓ Data Manipulation Statements
- ✓ Select, Views, GroupBy, SortBy/DistributeBy/ClusterBy/OrderBy, Joins
- ✓ Built-in Functions
- ✓ Union, Sub Queries, Sampling, Explain

Hive Usecase implementation -(Exercise)

- ✓ Use Case 1
- ✓ Use Case 2
- ✓ Best Practices

Advance Features

- ✓ Transform and Map-Reduce Scripts
- ✓ Custom UDF
- ✓ UDTF
- ✓ SerDe
- ✓ Recap and Q&A

Pig Introduction

- ✓ Position Pig in Hadoop ecosystem
- ✓ Why Pig and not MapReduce
- ✓ Simple example (slides) comparing Pig and MapReduce
- ✓ Who is using Pig now and what are the main use cases
- ✓ Pig Architecture
- ✓ Discuss high level components of Pig
- ✓ Pig Grunt - How to Start and Use

Pig Latin Programming

- ✓ Data Types
- ✓ Cheat sheet
- ✓ Schema

- ✓ Expressions
- ✓ Commands and Exercise
- ✓ Load, Store, Dump, Relational Operations,Foreach, Filter, Group, Order By, Distinct, Join, Cogroup,Union, Cross, Limit, Sample, Parallel

Use Cases (working exercise)

- ✓ Use Case 1
- ✓ Use Case 2
- ✓ Use Case 3 (compare pig and hive)

Advanced Features, UDFs

Best Practices and common pitfalls

Mahout & Machine Learning

- ✓ Mahout Overview
- ✓ Mahout Installation
- ✓ Introduction to the Math Library
- ✓ Vector implementation and Operations (Hands-on exercise)
- ✓ Matrix Implementation and Operations (Hands-on exercise)
- ✓ Anatomy of a Machine Learning Application

Classification

- ✓ Introduction to Classification
- ✓ Classification Workflow
- ✓ Feature Extraction
- ✓ Classification Techniques (Hands-on exercise)

Evaluation (Hands-on exercise)

- ✓ Clustering
- ✓ Use Cases
- ✓ Clustering algorithms in Mahout
- ✓ K-means clustering (Hands-on exercise)
- ✓ Canopy clustering (Hands-on exercise)

Clustering

- ✓ Mixture Models
- ✓ Probabilistic Clustering – Dirichlet (Hands-on exercise)
- ✓ Latent Dirichlet Model (Hands-on exercise)

- ✓ Evaluating and Improving Clustering quality (Hands-on exercise)
- ✓ Distance Measures (Hands-on exercise)

Recommendation Systems

- ✓ Overview of Recommendation Systems
- ✓ Use cases
- ✓ Types of Recommendation Systems
- ✓ Collaborative Filtering (Hands-on exercise)
- ✓ Recommendation System Evaluation (Hands-on exercise)
- ✓ Similarity Measures
- ✓ Architecture of Recommendation Systems
- ✓ Wrap Up