

**Pre-requisite:** The course is ideal for experienced software developers who have experience in traditional data management technologies and want to upgrade their skills.

#### Module 1

#### 1. Introduction

This section provides an introduction to class participants, schedule, agenda, knowledge areas and classroom environment.

- Class/Instructor/Student Introduction
- Agenda coverage What will be covered/our agenda, exercise format
- Pre-requisites what knowledge is expected for student's to have before course start

# 2. Class Project

In this section we will discuss the student class project to be worked on at the end. This is a project, designed and implemented by the student that can use his or her own data sets or one of the pre-defined data sets available.

### 3. Big Data

What data challenges do we face that have given rise to products like Hadoop as well as other NoSQL products.

- Big Data And modern data challenges
  - o Scale Out vs. Scale Up
  - o Internet Scale
  - o Offline Batch vs. Online Transactions
  - Big Data and NoSQL
- Hadoop Introduction
- Hadoop Distributions vendor landscape
- Comparison to Traditional Equivalent Products
- Why Use Hadoop and Common Use Cases

## 4. Hadoop Overview

Hadoop overview, its architecture, changes in architecture between major versions and the discussion of the vast Hadoop Eco-system.

- Architecture
  - Hadoop 1.0 Architecture
  - o Hadoop 2.0 Architecture
- Execution modes
  - Single
  - Pseudo-distributed
  - Distributed
- Hadoop Eco-System
  - o Ambari
  - o Mesos
  - HCatalog
  - o HBASE
  - o Hive
  - o Chukwa
  - o Pig
  - o Paraquet



- o Oozie
- Zookeeper
- Sqoop
- o Flume
- o Realtime Interfaces, e.g. Impala, Drill
- Exercise install VM, run through various exercises to demonstrate Hadoop and the various components.

### Module 2

## 1. Hadoop Components

This module takes a systematic approach through the main high level Hadoop architecture components. Covered is how they interface and depend on each other. These will be covered in more details in later modules.

- Setup setup Hadoop VM, install / check required development tools
- HDFS
  - NameNode
  - DateNode
  - Shell & Basic Commands
  - o HDFS Monitoring UI
  - Exercise adding files to, manipulating and accessing files in HDFS
- Map/Reduce Intro
  - o S mod
  - o Reduce
  - Job Tracker / Task Tracker
  - Exercise Simple Example of count word frequency
- YARN
  - M/R and Non M/R Jobs
  - o Resource Manager
  - Application Master
  - Resource Model

### Module 3

## 1. HDFS In-Depth

This module continues the HDFS section from the last module and takes an in-depth look at HDFS, architecture and use,

- Structure and Architecture
- File System Commands
- Importing / Exporting Data
- Java API
- Common Java Classes
- Example Usage
- Exercise file manipulation using Java.

# **Module 4**

#### 1. MapReduce In-Depth

This module continues the MapReduce section from the earlier introduction and takes an indepth look at M/R, processing steps, and use,

- Theory and application
- M/R Model



- M/R Framework
- M/R and YARN
  - o Daemons
  - o MapReduce1 vs. YARN
- V1 vs. New (V2) API Topics
- Exercise Map Side Join
- Exercise Reduce Side Join

## 2. Writing MapReduce Jobs

This module takes an in-depth look at MapReduce workflow, configuration and running dependent jobs.

- Fundamentals of a Map/Reduce API
- Basics of Map/Reduce Job Run
- The sort/shuffle/merge aggregation phases
- Job scheduling in Hadoop
- M/R Flow of Data
- Sample Map/Reduce Job walk-through
- Hadoop API walk-through
- Reducer Class
- Mapper Class
- Combiner
- Structuring complex jobs
- Composing MapReduce jobs into workflows
- Exercise implement complex workflow implemented in Java.

## **Module 5**

# 1. Map Reduce Formats

Within this module the student will learn about data formats supported within Hadoop and common implementation techniques for handling these. This will be critical to create efficient and robust MapReduce applications.

- Key and Value Types
- InputSplit
- InputFormat
- Partitioner
  - o Hash
  - o Custom
- Input /Output Formats
  - Working with text, XML, and JSON
  - Understanding SequenceFiles, Avro, and Protocol Buffers
  - Working with custom data formats, writing formatters
- Using the Hadoop Data Types
- Creating and using custom Data Types
- Exercise Sorting
- Exercise Sampling



# 2. MapReduce Development Topics

This sections looks at advanced topics connected with development of M/R jobs including testing, debugging and patterns for code organization.

This sections continues the look at

- Tool and Toolrunner Interface
- Configuration and Properties
- Exception Handling
- Logging
- Packaging M/R Applications
- Common Issues
  - o Dependency Versions & ClassPath
- Instrumentation
  - Predefined Counters
  - User Defined
  - Counters/Exercise
- Unit Testing M/R Jobs
  - o Using MRUnit
- Performance Considerations for M/R
- MapReduce Design Patterns and anti-patterns
- Exercise testing and packaging a M/R job
- Exercise debugging M/R job using Eclipse

### Module 6

### 1. Hadoop Streaming

This module looks at streaming and its use for developing quick MapReduce jobs.

- Streaming Overview
- Use Cases and Overview
- Streaming and M/R
- Exercise developing M/R job using Hadoop Streaming API

### 2. MapReduce and YARN

At this point has a solid understanding of concepts, architecture and development of MapReduce v1 jobs, and we will now look at MapReduce v2 and YARN.

- YARN In Depth
- YARN and M/R 2.0 Daemons
- Components
  - o Client
  - Resource Manager
  - Node Manager
  - M/R Application Master
  - HDFS
  - o Memory Model and Configuration for components'
- Handling Failures
- M/R and YARN command line tools

## **Module 7**

## 1. M/R Abstraction Frameworks



Alternative methods of running MapReduce jobs are examined in this module. As a developer it is important to understand these different frameworks and their use.

- Important Communication Protocols
  - o ProtoBuf
  - o Thrift
  - o Avro
- Paraquet
  - Paraquet M/R
- Hive
  - Overview of Hive
  - o HiveQL
  - Integrating M/R Into Hive Queries
  - o Exercise Hive Basic Commands and HiveQL
- Pig
- Overview of Pig
- o Pig Latin
- o Exercise Pig Join, Sort, Filter
- Cascading
  - Exercise Implement MapReduce job using Cascading

## Module 8

### 1. HBase

HBase is a column family type database built on top of Hadoop. This module looks at HBase, its data model and using the Java API for performing CRUD and more complex functions.

- Introduction
- Architecture
- Data Model
  - Tables
  - o Rows
  - Families
  - o Cell
  - Regions
- Data Types and Data Modeling
- Java API
  - Implementations/Libraries
  - CRUD Operations
- Exercise Simple Java app for performing CRUD operations
- Integration With M/R
  - o Scan API
  - Scan Caching
  - Scan Batching
  - Filters
- Exercise complex App for querying large amounts of data

### Module 9

# 1. Class Project

- Final Design
- Implementation





- Presentation to class
- 2. Questions and Wrap-Up