HADOOP Course Content   -   In Depth HadoopTraining
Syed

I. Introduction to Big Data and Hadoop
* What is Big Data?
* What are the challenges for processing big data?
* What technologies support big data?
*3V's of BigData and Growing.
* What is Hadoop?
* Why Hadoop and its Use cases
* History of Hadoop
* Different Ecosystems of Hadoop.
* Advantages and Disadvantages of Hadoop
* Real Life Use Cases

II. HDFS (Hadoop Distributed File System)
* HDFS architecture
* Features of HDFS
* Where does it fit and Where doesn't fit?
* HDFS daemons and its functionalities
* Name Node and its functionality
* Data Node and its functionality
* Secondary Name Node and its functionality
* Data Storage in HDFS
* Introduction about Blocks
* Data replication
*Accessing HDFS
* CLI(Command Line Interface) and admin commands
* Java Based Approach
*Hadoop Administration
*Hadoop Configuration Files
*Configuring Hadoop Domains
*Precedence of Hadoop Configuration
*Diving into Hadoop Configuration
*Scheduler
*RackAwareness
*Cluster Administration Utilities
*Rebalancing HDFS DATA
*Copy Large amount of data from HDFS
*FSImage and Edit.log file theoretically and practically.

III. MAPREDUCE
Map Reduce architecture
* JobTracker , TaskTracker and its functionality
* Job execution flow
* Configuring development environment using Eclipse
* Map Reduce Programming Model
* How to write a basic Map Reduce jobs
* Running the Map Reduce jobs in local mode and distributed mode
* Different Data types in Map Reduce

* How to use Input Formatters and Output Formatters in Map Reduce Jobs
* Input formatters and its associated Record Readers with examples
* Text Input Formatter
* Key Value Text Input Formatter
* Sequence File Input Formatter
* How to write custom Input Formatters and its Record Readers
* Output formatters and its associated Record Writers with examples
* Text Output Formatter
* Sequence File Output Formatter
* How to write custom Output Formatters and its Record Writers
* How to write Combiners, Partitioners and use of these
* Importance of Distributed Cache
* Importance Counters and how to use Counters

Advance MapReduce Programming
Joins - Map Side and Reduce Side
* Use of Secondary Sorting
* Importance of Writable and Writable Comparable Api's
* How to write Map Reduce Keys and Values
* Use of Compression techniques
* Snappy, LZO and Zip
* How to debug Map Reduce Jobs in Local and Pseudo Mode.
* Introduction to Map Reduce Streaming and Pipes with examples
*Job Submission
*Job Initialization
*Task Assignment
*Task Execution
*Progress and status bar
*Job Completion
*Failures
*Task Failure
*Tasktracker failure
*JobTracker failure
*Job Scheduling
*Shuffle & Sort in depth
* Diving into Shuffle and Sort
* Dive into Input Splits
* Dive into Buffer Concepts
*Dive into Configuration Tuning
*Dive into Task Execution
*The Task assignment Environment
*Speculative Execution
*Output Committers
*Task JVM Reuse
*Multiple Inputs & Multiple Outputs
*Build In Counters
* Dive into Counters – Job Counters & User Defined Counters
* Sql operations using Java MapReduce
* **Introduction to YARN (Next Generation Map Reduce)**

IV. Apache HIVE

* Hive Introduction
* Hive architecture
* Driver
* Compiler
* Semantic Analyzer
* Hive Integration with Hadoop
* Hive Query Language(Hive QL)
* SQL VS Hive QL
* Hive Installation and Configuration
* Hive, Map-Reduce and Local-Mode
* Hive DLL and DML Operations
* Hive Services
* CLI
*Schema Design
*Views
*Indexes
* Hiveserver

Metastore
* embedded metastore configuration
* external metastore configuration
* Transformations in Hive
* UDFs in Hive
* How to write a simple hive queries
* Usage
*Tuning
* Hive with HBASE Integration
* Need to add some more R&D done by myself

V. Apache PIG

Introduction to Apache Pig
Map Reduce Vs Apache Pig
* SQL Vs Apache Pig
* Different data types in Pig
* Modes Of Execution in Pig
* Local Mode
* Map Reduce Mode
* Execution Mechanism
* Grunt Shell
* Script
* Embedded
* Transformations in Pig
* How to write a simple pig script
* UDFs in Pig
* Pig with HBASE Integration
* Need to add some more R&D done by myself

VI. Apache SQOOP

* Introduction to Sqoop
* MySQL client and Server Installation
* How to connect to Relational Database using Sqoop
* Sqoop Commands and Examples on Import and Export commands.
*Transferring an Entire Table
*Specifying a Target Directory
*Importing only a Subset of data
*Protecting your password
*Using a file format other than CSV
*Compressing Imported Data
*Speeding up Transfers
*Overriding Type Mapping
*Controlling Parallelism
*Encoding Null Values
*Importing all your tables
*Incremental Import
*Importing only new data
*Incrementing Importing Mutable data
*Preserving the last imported value
*Storing Password in the Metastore
*Overriding arguments to a saved job
*Sharing the MetaStore between sqoop client
*Importing data from two tables
*Using Custom Boundary Queries
*Renaming Sqoop Job instances
*Importing Queries with duplicate columns
*Transferring data from Hadoop
*Inserting Data in Batches
*Exporting with All or Nothing Semantics
*Updating an Existing Data Set
*Updating or Inserting at the same time
*Using Stored Procedures
*Exporting into a subset of columns
*Encoding the Null Value
*Encoding the Null Value Differently
*Exporting Corrupted Data

VII. Apache FLUME

* Introduction to flume
* Flume agent usage

VIII Apache Hbase

* Hbase introduction

* Hbase basics
* Column families
* Scans
* Hbase installation
* Hbase Architecture
* Storage
* WriteAhead Log
* Log Structured MergeTrees
* Mapreduce integration
* Mapreduce over Hbase
* Hbase Usage
* Key design
* Bloom Filters
* Versioning
* Filters
* Hbase Clients
* REST
* Thrift
* Hive
* Web Based UI
* Hbase Admin
* Schema definition
* Basic CRUD operations

IX. Apache OOZIE
* Introduction to Oozie
* Executing workflow jobs

X. Hadoop Installation on Linux, All other ecosystems installations on Linux.

XI. Cluster setup (200 Nodes cluster) knowledge sharing with setup document.

XII. Cloudera & Hortonworks