

Course Content of Big Data Hadoop(Intermediate+ Advance)			
Pre-requisites: knowledge of Core Java/ Oracle: Basic of Unix			
S.no	Topics	Date	Status
1	Introduction to Big Data & Hadoop		
	Importance of Data& Data Analysis		
	What is Big Data		
	Big Data & its hype		
	Big Data Users & Scenerios		
	Structured Vs Unstruxtured Data		
	Challenges of Big Data		
	How to overcome the challenges		
	Divide & Conquer philosophy		
	Oveview of Hadoop		
	Hadoop and its file system- HDFS		
	History of Hadoop		
	Hadoop ecosystem		
	Hadoop Animal Planet		
	What is Hadoop		
	Key Distinction of Hadoop		
	Hadoop components		
	HDFS		
	Map Reduce		
	Why Distributed File System		
	The Design of HDFS		
	Hadoop Distributed File System		
	What is HDFS block		
	Why HDFS block is so large in HDFS		
	Name Node		
	Data Node		
Secondary Name Node			
A file in HDFS			

Hadoop Components/ Architecture		
Name node, Job Tracker, Data Node, Task Tracker & Secondary Name node		
Understanding storage components(Name node, Data Node & Secondary Name node)		
Understanding processing componets (Job Tracker & Task Tracker)		
Anatomy of file read		
Anatomy of file write		
Understanding Hadoop cluster		
Walkthrough of CDH VM setup		
Hadoop Cluster modes		
Standalone Mode		
Distributed Mode		
Hadoop Configuration files		
Core-site .xml		
hdfs-site.xml		
yarn-site.xml		
Understanding cluster configuration		
Map Reduce		
Meet Map Reduce		
Word count algorithm- Traditional approach		
Traditional approach on Distributed system & its's drawback		
Map reduce approach		
Input & Output forms of MR program		
Hadoop Data types		

	Map, shuffle& sort, Reduce Phases		
	Workflow & Transformaton of Data		
	Word Count code walkthrough		
	Input split & HDFS Block		
	Relation between split & block		
	MR Flow with single reduce task		
	MR flow with multipile reducers		
	Data locality Optimization		
	Speculative Execution		
	Combiner		
	Partitioner		
	Advance Map Reduce		
	Input Format & its hierarchy		
	Output format & its hierarchy		
	Using Compressin techniques		
	Side Data Distribution- Distibuted Cache		
	Joins		
	Map side join using Distributed Cache		
	Reduce side join		
	Secondary Sorting		
	MR Unit - An Unit testing framework		
2	Pig		
	What is Pig		
	Pig Vs SQL		
	Execution types or modes		
	Running Pig		
	Pig Data types		
	Pig Latin Diagnostic operators		
	Pig Latin Macro & UDF statements		
	Pig Latin commands		

	Pig Latin expressions		
	Schemas		
	Pig Functions		
	Pig Latin file loaders		
	Pig UDF & executing a Pig UDF		
	Pig Use Cases		
3	Hive		
	Introduction to Hive		
	Pig Vs Hive		
	Hive Limitation & Possibilities		
	Metastore		
	Hive QL		
	SQL Vs Hive QL		
	Hive Data Types		
	Data Storage		
	Manged & External Tables		
	Partitions & Buckets		
	Static Partitioning & Dynamic Partitioning		
	Storage Formats		
	File Formats- Sequence File & RC File		
	Using Compression in Hive		
	Built in Serdes		
	Importing Data(Using Load Data & Insert Into)		
	Alter & Drop Commands		
	Data Querying		
	Using MR Scripts		
Hive Joins			
Sub Queries			
Views			
4	Hbase		
	Introduction to NoSQL & Hbase		
	Hbase Use Cases		
	Row & Column oriented storage		

	Characteristics of huge DB		
	What is Hbase		
	Hbase Data Model		
	Hbase Logical model & physical storage		
	Hbase architecture		
	Hbase in operation (put, get, scan & delete)		
	Loading Data into Hbase		
	Hbase shell commands		
	Hbase operations through Java		
	Hbase operations through MR		
	Introduction to Zookeeper		
	Distributed Coordination		
	Zookeeper Data Model		
	Zookeeper Service		
	Introduction to Zookeeper		
	Distributed Coordination		
	Zookeeper Data Model		
	Zookeeper Service		
5	Sqoop		
	Introduction to Sqoop		
	Sqoop design		
	Sqoop basis commands		
	Sqoop table import flow of execution		
	Sqoop import commands- to HDFS, Hive& H Base tables		
	Sqoop Inremental Import		
	Incremental Append		
	Incremental Last Modified		
	Sqoop export flow of execution		
	Sqoop Export Command		
6	Flume		
	Flume Architecture		
	Flume Components		

	Streaming live Twitter data with Flume		
7	Hadoop 2.0 & YARN		
	Hadoop 1 Limitations		
	HDFS Federation		
	Name Node High Availability		
	Introduction to YARN		
	YARN applications		
	YARN Architecture		
	Anatomy of an YARN application		
8	Mongo DB		
	Spark Overview		
	Why Spark		
	Spark & Big Data		
	Spark Components		
	Resilient Distributed Data Sets		
	Data Operations on RDD		
	Spark Libraries		
9	Scala Object Oriented Programming		
	Introduction to Scala		
	Why Scala		
	Scala Vs Java		
	Installing of Scala		
	Installing of Sbt		
	Variable Declarations		
	Ranges		
	Partial Functions		
	Method Declerations		
	Literal's in Scala		
	Operators		
	Operator Overloading		
	Scala Control Statement		
	Call by Name and Call by value		
Pattern Macthings			

	Implicit Conversions		
	Traits		
	Abstraction		
	Inheritance		
	Collections(List, Tuple, Set, Arrays, Buffer, Map)		
	Scala Functional Programming		
	What is Functional Programming		
	Diff. between Functional & Imperative		
	Anonymous Functions		
	Closures		
	Currying		
	Functional Data Structures(Sequences, Map, Stes)		
	Traversal		
	Mapping		
	Flat Mapping		
	Filtering		
	Folding & Reducing		
10	Spark in Detailed		
	What is Spark		
	Why Spark		
	Diff. between Map Reduce & Spark		
	Diff. between Spark & Storm		
	Installation of Spark on HDFS		
	Installation of Spark on EC2		
	What is RDD's and creation of RDD's		
	Programming with RDD's		
	Working with Key/ Value pairs		
	Loading and saving your data		
	Advanced Spark Programming		
	Running on a Spark Cluster		

	Spark Streaming		
	Spark SQL		
	Spark MLIB		
	Spark Graphix		
	Tunning & Debugging Spark		
11	Kafka in Detailed		
	What is Kafka		
	Why Kafka		
	Installing Kafka on local mode		
	Installing Kafka on local mode with multiple servers		
	Creating custom producers		
	Creating custom consumers		
	Integrating Kafka to Storm		
	Integrating Kafka to Hadoop		
	Overview of Kafka Adminstration		