Electronics & ICT Academy
(Under Ministry of Electronics & Information Technology)

Indian Institute of Technology Guwahati, Guwahati, Assam, Pin 781039

Phone: +91-361-2582503, 2582536 Email: eictacad@iitg.ernet.in

DETAILED POINT-WISE COURSE SYLLABUS

DAY 1

1. Understanding BigData and Hadoop

Understand Big Data, the limitations of the existing solutions for Big Data problem, how Hadoop solves the Big Data problem, the common Hadoop ecosystem components, Hadoop Architecture, HDFS, Anatomy of File Write and Read, how MapReduce Framework works.

Theory:

Big Data, Limitations and Solutions of existing Data Analytics Architecture, Hadoop, Hadoop Features, Hadoop Ecosystem, Hadoop 2.x core components, Hadoop Storage: HDFS, Hadoop Processing: MapReduce Framework, Hadoop Different Distributions.

2. Hadoop Architecture and HDFS

Learn the Hadoop Cluster Architecture, Important Configuration files in a Hadoop Cluster, Data Loading Techniques.

Theory:

Hadoop 2.x Cluster Architecture - Federation and High Availability, A Typical Production Hadoop Cluster, Hadoop Cluster Modes, Common Hadoop Shell Commands, Hadoop 2.x Configuration Files, log files,lib files.

Practical/Demo:

Hadoop local repository set ups on all Linux systems.

Demo of various shell commands, relevance of each configuration files, log file, lib files will be done after Hadoop cluster set up.

Day2

1. Hadoop Cluster set up

How to set up multinode cluster. Basic Hadoop Administration Commands

Theory:

Single node cluster and Multi node cluster set up Hadoop Administration. Run Various hadoop commands.

Practical/Demo:

Assuming each student to have access to one Linux machine. Group of 2 to make multi node cluster with all best practices. Configuration changes, tuning cluster.

Big data High Availability Configuration

Day3

1. Hadoop Map reduce Framework and HBASE(NoSqDB)

Hadoop MapReduce framework and the working of MapReduce on data stored in HDFS. You will understand concepts like Input Splits in MapReduce, Combiner & Partitioner and Demos on MapReduce using different data sets.

Theory:

MapReduce Use Cases, Traditional way Vs MapReduce way, Why MapReduce, Hadoop 2.x MapReduce Architecture, Hadoop 2.x MapReduce Components, YARN MR Application Execution Flow, YARN Workflow, Anatomy of MapReduce Program, Demo on MapReduce. Input Splits, Relation between Input Splits and HDFS Blocks, MapReduce: Combiner & Partitione..Hbase set up, monitoring HBASE,Region servers Common issues. Troubleshooting approach.

Practical/Demo:

▪ Sample Map reduce jobs on test data.

2. Run MapReduce Jobs and monitor

Run Mapreduce jobs, monitor and troubleshoot from admin UI's/Hue

Theory:

Teragen ,Terasort,Hadoop jars,Mapreduce wordcount programs,Resource Manager UI,logs,Hadoop Logs

Practical/Demo:

Troubleshooting from AdminUI's, Hue.

Benchmarking via teragen,terasort hadoop jars etc.

Day4

1. Learn NoSQL Data Management

Awareness and use cases of various NoSql DB's

Theory:

Comparison which NoSql and associated benefits.

Practical/Demo:

No Demo.

2. PIG

learn Pig, types of use case we can use Pig, tight coupling between Pig and MapReduce, and Pig Latin scripting, PIG running modes, PIG UDF.

Theory:

About Pig, MapReduce Vs Pig, Pig Use Cases, Programming Structure in Pig, Pig Data Types, Shell and Utility Commands, Pig Latin : Relational Operators, File Loaders, Group Operator, COGROUP Operator, Built In Functions (Eval Function, Load and Store Functions, Math function, String Function, Date Function.

Practical:

▪ scripting via pig

3. Hive

This module will help you in understanding Hive concepts, Hive Data types, Loading and Querying Data in Hive, running hive scripts and Hive UDF.

Theory:

Hive Background, Hive Use Case,Creating and quering Hive tables.. Hive Vs Pig, Hive Architecture and Components, Metastore in Hive, Limitations of Hive, Comparison with Traditional Database, Hive Data Types and Data Models, Partitions and Buckets, Hive Tables(Managed Tables and External Tables).

Practical:

HiveQl commands to create and manage tables.

Day5

1. Sqoop

Sqoop concepts and use cases

Theory:

Install Mysql, load data and move data to from mysql database to Hadoop and vice versa

Practical/Demo:

Sqoop demo for ETL

2. Advanced technologies with Hadoop

In this module you will learn Spark ecosystem and its components, how scala is used in Spark, SparkContext. You will learn how to work in RDD in Spark. Demo will be there on running application on Spark Cluster,Introduction to Kafka.

Theory:

What is Apache Spark, Spark Ecosystem, Spark Components, History of Spark and Spark Versions/Releases, What is Scala?, Why Scala?, SparkContext, RDD.Benifits of in memory processing

Practical:

▪ No demo.

Electronics & ICT Academy
(Under Ministry of Electronics & Information Technology)

Indian Institute of Technology Guwahati, Guwahati, Assam, Pin 781039

Phone: +91-361-2582503, 2582536 Email: eictacad@iitg.ernet.in

Day6

1. Administration and Maintenance

Namenode/Datanode directory structures and files, File system image and Edit log, The Checkpoint Procedure, Namenode failure and recovery procedure, Safe Mode, Metadata and Data backup, Potential problems and solutions / what to look for, Adding and removing nodes, Lab: MapReduce File system Recovery.

Theory:

Namenode/Datanode directory structures and files, File system image and Edit log, The Checkpoint Procedure, Namenode failure and recovery procedure, Safe Mode, Metadata and Data backup, Potential problems and solutions / what to look for, Adding and removing nodes, Lab: MapReduce File system Recovery

Practical:

- Capacity Expansion and planning of Hadoop cluster. Adding new nodes.

2. Job Scheduler: Map reduce job submission flow

How to schedule Jobs on the same cluster, FIFO Schedule, Fair Scheduler and its configuration

Theory:

How to schedule Jobs on the same cluster, Capacity scheduler and other schedulers and its configuration

Practical/Demo:

Job Scheduling.