HADOOP / BIGDATA SYLLABUS

- **BigData Introduction and Hadoop** Fundamentals
  - Data Storage and Analysis
  - Comparison with RDBMS

- **Hadoop – A Brief History**

- **MapReduce – Part1**
  - Map and Reduce
  - Sample Program
  - Combiner
  - Practitioners and Custom Partitioned

- **Hadoop Streaming & Pipes**

- **HDFS**
  - Blocks
  - NN & DN
  - HDFS Federation & High Availability

- **HDFSClients**
  - HDFS Command Line
  - HDFS CLI – File System Operations Lab
  - HDFS Web UI
  - HDFS Java Client
  - HDFS Java Client – File System Operations Lab
  - CRUD Operations using Java Client
  - Anatomy of File Read and File Write
  - DistCp
  - Cluster balancing

- **YARN – Cluster Management (Hadoop 2.x)**
  - How Yarn Applications run?
  - YARN vs MapReduce
  - YARN Scheduling
    - Capacity Scheduler
    - Fair Scheduler
    - FIFO Scheduler

- **Map Reduce – Part2**
  - Env Setup
  - Tool and ToolRunner
  - Mapper
  - Reducer
  - Driver program
  - How to package the job
  - MapReduce WebUI

- How MapReduce Job run?
- Shuffle & Sort
- Speculative Execution

- **InputFormats**
  - Input Splits and Record Reader
  - Default Input Formats
  - Implement Custom Input Format

- **OutputFormats**
  - Default Output formats
  - Output Record Reader

- **Compression**
  - Map Output
  - Final Output
  - Splittable vs Non Splittable
  - Compression Codecs

- **Serialization**
  - Data types –default
  - Writable vs Writable Comparable
  - Custom Data types – Custom Writable/Comparable

- **File Based Data structures**
  - Sequence file
  - Reading and Writing into Sequence file
  - Map File

- **Tuning MapReduce Jobs**

- **Advanced MapReduce**
  - Counters
    - Built-In Counters Classification
    - User Defined Counters
  - Sorting
    - Partial Sort
    - Total Sort
    - Secondary Sort
  - Joins
    - Map-side joins
    - Reduce-side joins
    - Distributed Cache

- **Hive**
  - Comparison with RDBMS
  - HQL
  - Data types
  - Tables
  - Importing and Exporting
  - Partitioning and Bucketing – Advanced.

- Joins and Join Optimization.
- Functions- Built in & user defined
- Advanced Optimization of HQL
- Storage File Formats – Advanced
- Loading and Storing Data
- SerDes– Advanced

- **Sqoop**
  - Important basics
  - Import – Deep dive
  - Export – Deep dive
  - Sqoop Optimization – Incremental Load
  - Many more

- **PIG**
  - Important basics
  - Pig Latin
  - Data types
  - Functions – Built-in, User Defined
  - Loading and Storing Data

- **Flume**
  - Configure Flume and Import data
  - Architecture and LAB

- **Oozie**
  - Different workflow jobs
  - Ooze scheduler.
  - LAB – covers advanced topics
  - CAP theorem
  - HBaseArchitecture
  - HBase Clients – Java Client
  - Loadling Data
  - UDF,UDAF,UDTFs

SYLLABUS CONTENTS CAN BE MODIFIED BASED ON STUDENTS REQUIREMENTS