4/4 B.Tech - SECOND SEMESTER

IT8T2D **BIG DATA ANALYTICS** **Credits:3**

Lecture: 3 Periods/week **Internal assessment: 30 marks**

Practice/Interaction: 1Period/week **Semester end examination: 70 marks**

Objectives:

- Cloud and lays a strong foundation of Apache Hadoop (Big data framework).
- The HDFS file system, MapReduce frameworks
- Hadoop tools like Hive, and Hbase
- Analyzing data with UNIX tools
- Sorting. Map side and Reduce side joins.

Outcomes:

Students will be able to

- Understand the fundamentals of Big cloud and data architectures.
- Learn the concepts of HDFS file systems and interfaces and able to keep HDFS cluster balanced
- Familiarize with map reduce classes, combiner functions and can run map reduce job.
- Aware of classic map reduce and able to apply shuffle and sort on map reducer side.
- Understand The Hive Shell.

Prerequisites:

File Structures, Databases, Java, UNIX

UNIT-I

Introduction to Big Data. Importance of Big Data. Map Reduce and example psuedocodes for some problems. A brief history of Hadoop. Apache hadoop and the Hadoop EcoSystem. VMWare Installation of Hadoop.

UNIT-II

The design of HDFS. HDFS concepts. Command line interface to HDFS. Hadoop File systems. Interfaces. Java Interface to Hadoop. Anatomy of a file read. Anatomy of a file write. Replica placement and Coherency Model. Parallel copying with distcp, Keeping an HDFS cluster balanced.

UNIT-III

Introduction. Map reduce: introduction, Analyzing data with unix tools. Analyzing data with hadoop. Java MapReduce classes (new API). Data flow, combiner functions, Running a distributed MapReduce Job. Configuration API. Setting up the development environment. Managing configuration. Writing a unit test with MRUnit. Running a job in local job runner. Running on a cluster.Launching a job. The MapReduce WebUI.

UNIT-IV

Classic Mapreduce. Job submission. Job Initialization. Task Assignment. Task execution. Progress and status updates. Job Completion. Shuffle and sort on Map and reducer side. Configuration tuning. Map Reduce Types. Input formats. Output formats ,Sorting. Map side and Reduce side joins.

UNIT-V

The Hive Shell. Hive services. Hive clients. The meta store. Comparison with traditional databases. Hive Ql. Hbasics. Concepts. Implementation. Java and Mapreduce clients. Loading data, web queries.

Text Books:

1. Tom White, Hadoop, “The Definitive Guide”, 3rd Edition, O’Reilly Publications, 2012.
2. Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch, “Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data”, 1st Edition, TMH, 2012.

Reference:

1. Frank J. Ohlhorst, “Big Data Analytics: Turning Big Data Into Big Money”, 2nd Edition, TMH,2012.

e-Learning Resources:

1. http://www.cloudera.com/content/cloudera-content/cloudera-docs/HadoopTutorial/CDH5/Hadoop-Tutorial.html