BIG DATA AND HADOOP
(Elective-1)

Course Code: 13CS2108 **L P C**
4 0 3

Pre requisites: Data base Management Systems, Object Oriented Programming through Java.

Course Educational Objectives:

- This course introduces the fundamental concepts of cloud and lays a strong foundation of Apache Hadoop (Big data framework).
- The HDFS file system, MapReduce frameworks are studied in detail. Hadoop tools like Hive, and Hbase, which provide interface to relational databases, are also covered as part of this course work.

Course Outcomes:

- The student will be able to understand the fundamentals of Big cloud and data architectures.
- The student will be able to understand HDFS file structre and Mapreduce frameworks, and use them to solve complex problems, which require massive computation power.
- The student will be able to use relational data in a Hadoop environment, using Hive and Hbase tools of the Hadoop Ecosystem.

UNIT-I

Introduction to Big Data. What is Big Data. Why Big Data is Important. Meet Hadoop. Data. Data Storage and Analysis. Comparison with other systems. Grid Computing. A brief history of Hadoop. Apache hadoop and the Hadoop EcoSystem. Linux refresher; 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. 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. MapReduce 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.

References:

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