BIG DATA AND HADOOP
(Elective-1)

Course Code: 13CS2108

L	P	C
4	0	3

Course Outcomes:

At the end of the course, a student will be able to:

CO1: Explain Big data and Apache Hadoop Eco system; and install Hadoop software.

CO2: List out and explain the Design concepts of HDFS(Hadoop Distributed File system); and describe the anatomy of reading a file and writing a file in HDFS and Coherency Model

CO3: Distinguish between analyzing data using Unix tools and using Java MapReduce API in Hadoop framework

CO4: Apply the Concepts of Hadoop, create programs and execute them in Hadoop environment and analyze the results

CO5: Use different tools in Hadoop Framework and Compare between services provided in traditional databases and database services provided in Hadoop.

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.