Prasad V. Potluri Siddhartha Institute of Technology:: Vijayawada.
Department of Computer Science and Engineering

I/II M.Tech. (CSE) - (Second Semester)

17CSCS2T2 BIG DATA ANALYTICS Credits: 4

Lecture: 4 Periods/week **Internal Assessment: 40 Marks**
Semester end examination: 60 Marks

Course Description

This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including MapReduce and Hadoop and its ecosystem.

Course Outcomes:

At the end of the course, students should be able to:

CO1: Learn tips and tricks for Big data use cases and solutions

CO2: Learn about build and maintain reliable, scalable, distributed systems in big data using Apache Hadoop

CO3: Apply MapReduce concepts in Distributed environemnt

CO4: Able to apply Hadoop ecosystem components

Unit-1

Introduction to Big data and Hadoop: Introduction – Distributed file system, Big data and its importance, Six V's, Drivers for Big Data, Big data Analytics, Applications of Big data, algorithms using MapReduce.

Introduction to Hadoop: Big Data – Apache Hadoop & Hadoop EcoSystem – Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce - Data Serialization.

Unit-2

Hadoop Architecture, Hadoop Storage: HDFS, Common Hadoop Shell commands , Anatomy of File Write and Read., NameNode, Secondary NameNode, and DataNode, Hadoop MapReduce paradigm, Map and Reduce tasks, Job, Task trackers - Cluster Setup – SSH & Hadoop Configuration – HDFS Administering –Monitoring & Maintenance.

Unit-3

MAP REDUCE: Introduction – distributed file system – algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce – Hadoop - Understanding the Map Reduce architecture - Writing Hadoop MapReduce Programs - Loading data into HDFS - Executing the Map phase - Shuffling and sorting - Reducing phase execution.

Unit-4

HADOOP ECOSYSTEM AND YARN: Hadoop ecosystem components - Schedulers - Fair and Capacity, Hadoop 2.0 New Features- NameNode High Availability, HDFS Federation, MRv2, YARN, Running MRv1 in YARN. Introduction to Hive, HBASE, HiveQL, Zookeeper.

Text Books:

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, ISBN: 9788126551071, 2015.
2. Chris Eaton, Dirk deroos et al. , "Understanding Big data ", McGraw Hill, 2012.
3. Tom White, "HADOOP: The definitive Guide" , O Reilly 2012.

Reference Books:

1. Vignesh Prajapati, "Big Data Analytics with R and Haoop", Packet Publishing 2013.
2. Tom Plunkett, Brian Macdonald et al, "Oracle Big Data Handbook", Oracle Press, 2014. http://www.bigdatauniversity.com/
3. Jy Liebowitz, "Big Data and Business analytics",CRC press, 2013.
4. Big Data and Analytics, Seema Acharya, Subhashini Chellappan, Wiley Publications, 2015.