17CS4103 – BIG DATA AND HADOOP

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Category:** | Program Core | **Credits:** | 3 |
| **Course Type:** | Theory | **Lecture – Tutorial – Practical:** | 3-0-0 |
| **Prerequisite:** | Should have knowledge of one Programming Language (Java preferably), Practice of SQL (queries and sub queries) and exposure to Linux Environment. | **Sessional Evaluation:****Univ.Exam Evaluation:****Total Marks:** | 4060100 |
| **Objectives** | * To learn the concept and challenge of big data (3 V’s: volume, velocity, and variety).
* To apply skills and tools to manage and analyze the big data.
* Understand the Big Data Platform and provide an overview of Apache Hadoop
* Provide HDFS Concepts and understand Map Reduce Job.
* Apply analytics on Structured and Unstructured Data.
* Exposure to Data Analytics with R.
 |

|  |  |
| --- | --- |
| **Course Outcomes** | Upon successful completion of the course, the students will be able to: |
| CO1 | Understand the basics of Big Data analytics and study its life cycle |
| CO2 | Learn the fundamentals of R and Hadoop to develop simple data analysis applications |
| CO3 | Study various file access and process operations and their utilization in a distributed environment |
| CO4 | Perform I/O operations and learn the environment of MapReduce basics |
| CO5 | Explore the components of Hadoop and Hadoop Eco-System with supporting tools |
| CO6 | Develop Big Data Solutions using Hadoop Eco System tools  |
| **Course Content** | **UNIT - I****Introduction to Big Data Analytics**: Big Data Overview, State of the Practice in Analytics, Key Roles for the New Big Data Ecosystem, Examples of Big Data Analytics**Data Analytics Life Cycle**: Data Analytics Lifecycle Overview, Discovery, Data Preparation, Model Planning, Model Building, Communicate Results, Operationalize, Case Study: Global Innovation Network and Analysis (GINA).**UNIT – II****INTRODUCTION TO R AND HADOOP:** Introduction to R, Exploratory Data Analysis, History of Hadoop, Apache Hadoop, Analyzing Data with Unix tools, Analyzing Data with Hadoop, Hadoop Streaming, Hadoop Echo System. **UNIT – III**IBM Big Data Strategy, Introduction to Infosphere BigInsights and Big Sheets.**HDFS (Hadoop Distributed File System):** The Design of HDFS, HDFS Concepts, Command Line Interface, Hadoop file system interfaces, Data flow, Data Ingest with Flume and Scoop and Hadoop archives. **UNIT – IV****Hadoop I/O**: Compression, Serialization, Avro and File-Based Data structures. **MapReduce:** Anatomy of a MapReduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution, MapReduce Types and Formats, MapReduce Features.**UNIT – V****Hadoop Eco System - tools:****Pig:** Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases, Grunt, Pig Latin, User Defined Functions, Data Processing operators.**Hive:** Hive Shell, Hive Services, Hive Metastore, Comparison with Traditional Databases, HiveQL, Tables, Querying Data and User Defined Functions.**UNIT – VI****Hadoop Eco System - tools (Continued):****Hbase:** HBasics, Concepts, Clients, Example, Hbase Versus RDBMS.**Big SQL**: Introduction |
| **Text Books and References:** | **TEXT BOOKS**:1. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services**,** Published by John Wiley & Sons, Inc.
2. Tom White “Hadoop: The Definitive Guide” Third Edition, O’reily Media, 2012.

**REFERENCE BOOKS**:1. Michael Berthold, David J. Hand, "Intelligent Data Analysis”, Springer, 2007.
2. Jay Liebowitz, “Big Data and Business Analytics” Auerbach Publications, CRC press (2013).
3. Tom Plunkett, Mark Hornick, “Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R Enterprise and Oracle R Connector for Hadoop”, McGraw-Hill/Osborne Media (2013), Oracle press.
4. Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015
 |
| **E-Resources** | 1. <https://nptel.ac.in/courses>
2. <https://freevideolectures.com/university/iitm>
 |