17CS41P2 – BIG DATA AND HADOOP LABORATORY

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Category:** | Program Core | **Credits** | 2 |
| **Course Type:** | Laboratory | **Practical:** | 0-0-3 |
| **Prerequisite:** | Require the basics of Database Management Systems and awareness on social networks | **Sessional Evaluation:****Univ. Exam Evaluation:****Total Marks:** | 4060100 |
| **Course Objectives** | * Optimize business decisions and create competitive advantage with Big Data analytics, Derive business benefit from unstructured data.
* Imparting the architectural concepts of Hadoop and introducing mapreduce paradigm.
* Study and practice various tools in Hadoop echo system.
 |

|  |  |
| --- | --- |
| **Course Outcomes** | Upon the successful completion of the course, the students will be able to: |
| CO1 | Prepare data summarization, apply data modeling techniques to large data sets, create applications for Big Data analytics and build a complete business data analytic solutions |
| **Course Content** | 1. Perform setting up and Installing Hadoop in its two operating modes such as
2. Pseudo distributed and Fully distributed.
3. Use web based tools to monitor your Hadoop setup.
4. Implement the following file management tasks in Hadoop:
	* Adding files and directories
	* Retrieving files
	* Deleting files
5. Benchmark and stress test an Apache Hadoop cluster.
6. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.
7. Find the number of occurrence of each word appearing in the input file(s)
8. Performing a MapReduce Job for word search count (look for specific keywords in a file).
9. Write a MapReduce program that mines weather data. Weather sensors collecting data every hour at many locations across the globe gather large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record-oriented. (Data available at: https:// github.com/ tomwhite / hadoopbook / tree/ master/ input/ ncdc / all )
10. Find average, max and min temperature for each year in NCDC data set?
11. Filter the readings of a set based on value of the measurement, Output the line of input files associated with a temperature value greater than 30.0 and store it in a separate file.
12. Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data.
13. Write a Pig Latin scripts for finding TF-IDF value for book dataset (A corpus of eBooks available at: Project Gutenberg).
14. Install and Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes.
 |
| **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>
 |