ITE6013 Big Data Analytics

LTPJC
3 0 0 4 4

Pre-requisite: Nil

Objectives:

- Use the Hadoop Distributed File System for storing large datasets and run distributed computations over those datasets with MapReduce.
- Become familiar with Hadoop's data and I/O building blocks for compression, data integrity, serialization and persistence.
- Discover common pitfalls and advanced features for writing real-world MapReduce programs.

Expected Outcome :

On completion of this course, student should be able to

- Design, build and administer a dedicated Hadoop cluster, or run Hadoop in the cloud.
- Apply best practices to extend data warehousing with Hadoop and other big data technologies across business operations and industries to enable big data analytics.
- Implement algorithms for analyzing and mining data streams and social network graphs.

Module	Topics	L Hrs	SLO
1	**Overview of Big Data and Data Analytics:** Overview of Big Data: Characteristics of Big Data-Big Data Sources- Challenges in Big Data processing-Scalability issues; Business Intelligence v/s Data Analytics-Need of Data Analytics-Data Analytics in Industries-Role of the Data Scientist.	6	2
2	**Hadoop and HDFS:** The Design of HDFS- HDFS Concepts- Blocks - Namenodes and Datanodes; The Command-Line Interface: Basic File system Operations; Hadoop File systems: Interfaces-The Java Interface-Data Flow; Hadoop I/O: Data Integrity-Compression-Serialization-File-based data structures.	6	2
3	**MapReduce:** Analyzing the Data with Unix Tools- Analyzing the Data with Hadoop- Map and Reduce- Java MapReduce; Data Flow-Combiner Functions- Running a Distributed MapReduce Job; Hadoop Streaming; Hadoop Pipes.	6	2
4	**Application development using MapReduce framework:** The Configuration API- Configuring the Development Environment- Writing a Unit Test- Running Locally on Test Data- Running on a Cluster- Tuning a Job- MapReduce Workflows.	6	7
5	**Working of MapReduce:** Anatomy of a MapReduce Job Run-Failures-Job Scheduling-Shuffle and Sort-Task Execution; MapReduce Types and Formats- Input Formats- Output formats-MapReduce Features- Counters- Sorting-Joins.	6	2
6	**Analytics for Big Data in motion:** Mining Data Streams: The Stream Data Model- Sampling data in a stream- Filtering Streams- The Bloom filter; Counting distinct elements in a stream- The Flajolet-Martin Algorithm. How stream works-Streams Processing Language; Apache Spark - Introduction- Features of Apache Spark- Components of Spark- Resilient Distributed Datasets- Data Sharing using Spark RDD-Spark Streaming.	6	14

7	**Analysis of Social Network Data-**Mining Social Network Graphs: Clustering of Social Network Graphs- Direct Discovery of Communities- Partitioning of Graphs- Finding overlapping communities- Simrank; Sentiment analysis- Document sentiment classification- Rules of Sentiment Composition- Sentiment analysis using Twitter data.	6	14
8	**Applications of Big Data Analytics in Industry**	3	17
# Mode: Flipped Class Room, [Lecture to be videotaped], Use of physical and computer models to lecture, Visit to Industry, Min of 2 lectures by industry experts		**Total Lecture Hours** **45**	
Text Books 1. Tom White, "Hadoop: The definitive guide",3rd Edition, O'Reilly Media, Inc., 2012. **Reference Books** 1. Jure Leskovec, Anand Rajaraman, Jeff Ullman, "Mining of Massive Datasets", 2nd Edition, Cambridge University Press, UK, 2011. 2. Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill, 2012. 3. Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies,Cambridge University Press, 2015. 4. Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, " Learning Spark:Lightning-Fast Big Data Analysis", O'Reilly Media, 2015. 5. David Loshin, Morgan, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL and Graph", Kaufman Publishers, 2013.			
Compiled by : Prof. Nancy Victor			
Date of approval by the Academic Council : 18.03.16			