

Subject Code: 2CS2010306	Subject Title: Big Data Analytics (Elective-III)
Prerequisite:	Working knowledge of programming language and database concepts

Course Objective:

The objectives of the course are to:

- Understand basics of Big Data and Big Data Tools (Hadoop, MapReduce)
- Understand fundamental techniques used for Big data analytics

Teaching Scheme (Hours per week)				Evaluation Scheme (Marks)				
Lecture	Tutorial	Practical	Credit	Theory		Practical		Total
				University Assessment	Continuous Assessment	University Assessment	Continuous Assessment	
3	-	3	6	60	40	30	20	150

Subject Contents			
Sr. No	Topic	Total Hours	Weight (%)
1	Overview – Big Data Tools: Big Data: Introduction to Big Data, Big Data characteristics, Types of Big Data, Traditional vs. Big Data approach Hadoop: What is Hadoop? Core Hadoop Components, Hadoop Ecosystem (Hbase, Hive, Hcatalog, Pig, Sqoop, Oozie, Mahout, ZooKeeper), Physical Architecture, Hadoop limitations. NoSQL: What is NoSQL? NoSQL business drivers, NoSQL case studies (Amazon Dynamo DB, Google's Big Table, MongoDB, Neo4j), NoSQL data architecture patterns (Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores), Variations of NoSQL architectural patterns; Using NoSQL to manage big data Map Reduce: MapReduce and New Software stack (Distributed File Systems, Physical Organization of Compute Nodes), The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures	9	25
2	Finding Similar items: Nearest neighbor Search, Applications of Near-Neighbor Search, Similarity of documents (Plagiarism detection, Document clustering, News aggregation), Collaborative Filtering as a Similar-Sets Problem, Recommendations based on user ratings, Distance Measures (Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance)	9	25
3	Data Stream Mining: The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing. Sampling Data in a Stream, Filtering Streams (The Bloom Filter), Counting Distinct Elements in a Stream, Counting Ones in a Window, Decaying Windows.	6	15

4	Link Analysis: PageRank, Efficient Computation of Page Rank (PageRank implementation using MapReduce, Use of Combiners to Consolidate the Result Vector), Topic sensitive Page Rank, link Spam, Hubs and Authorities.	3	10
5	Recommendation Systems and Social Network Graphs: Recommendation Systems: A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering. Social Network Graphs: Application of social network mining, Social network as a graph, Types of social networks, Clustering of social Graphs, Direct discovery of communities in a social Graph, SimRank	9	25

Course Outcome:

At the end of this course, the student would be able

- To understand the importance of Big Data and Big Data Tools for solving real world problems.
- To understand the mechanism of fundamental techniques used for the Big Data Analytics.

List of References:

1. Radha Shankarmani, M Vijayalakshmi, Big Data Analytics, 2nd Edition, Wiley
2. Jure Leskovec, AnandRajaraman, Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, Second Edition, 2014.
3. Seema Acharya, Subhashini Chhellaappan, BIG Data and Analytics , , Willey
4. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques", Morgan Kaufman Publications, Third Edition, 2011.
5. VigneshPrajapati, "Big Data Analytics with R and Hadoop", Packet Publishing 2013
6. Big Data Black Book, DreamTech

Web Resources:

1. <http://www.bigdatauniversity.com>
2. <http://www.mongodb.com>
3. <http://hadoop.apache.org>

List of Experiments:

Note: The experiment list provided beneath is for reference only. The course teacher may Change/formulate it as per his/her methodology and requirement.

Sr. No	Practical Experiments
1.	HADOOP: How to run word count program on Hadoop and HDFS
2.	2.1 MongoDB Installation Steps 2.2 Run various commands like help (), auth (), createUser (), cloneDatabase() etc. 2.3 Run various database related command 2.4 Try create Collection () and collection. Drop() methods
3.	3.1 Create a file into the local file system and export it to the HDFS systems 3.2 To work with Pig in both Hadoop local mode and HDFS / MapReduce mode 3.3 Define schema using Load command and validating it. 3.4 Try various Relational operators
4	Algorithms Using MapReduce: Matrix-Vector Multiplication by MapReduce, Relational operators (Selection, projection, Union, Intersection and difference), Computing Natural Join by MapReduce, Grouping and Aggregation by Map Reduce, Matrix Multiplication, Matrix Multiplication with One MapReduce Step.