ITE2013	Big Data Analytics	L	T	P	J	C
		3	0	0	4	4
Pre-requisite	**ITE1003**	**Syllabus version**				
		1.0				

Course Objectives:

- To introduce Big Data and Data analytics lifecycle to address business challenges that leverage big data.
- To understand the importance of mining data streams and social network graphs.
- To introduce big data analytics technology and tools including MapReduce and Hadoop.

Expected Course Outcome:

1) Reframe a business challenge as an analytics challenge.
2) Create models and identify insights that can lead to actionable results.
3) Design of big data analytics projects.
4) Use tools such as MapReduce / Hadoop.
5) Implement suitable analytics for big data clustering for resolving challenges in real-time business problems
6) Develop suitable social network analysis models, appraise the quality of the inputs, gain understanding from the outcomes.
7) Implement Multiple and huge scaling analytics tools for resolving contemporary big data challenges

Student Learning Outcomes (SLO):	**7, 14**
[7]	Having computational thinking
[14]	An ability to design and conduct experiments, as well as to analyze and interpret data

Module:1	**Big Data Concepts and Environment**	**6 hours**

Big Data Overview-Big Data Challenges and Opportunities- Data analytics lifecycle overview – Phases of Data Analytics: Discovery, Data preparation, Model planning, Model building, Communicate results, Operationalize – Case Study.

Module:2	**Overview of Hadoop and HDFS**	**6 hours**

Introduction to Hadoop - The Distributed File System: HDFS, GPFS – The Design of HDFS – HDFS-Concepts-Blocks, Name Nodes and Data Nodes; Components of Hadoop- Hadoop Cluster Architecture-Batch Processing- Serialization - Hadoop ecosystem of tools-NoSQL .

Module:3	**Map Reduce**	**6 hours**

MapReduce Basics - Functional Programming Roots - Mappers and Reducers - The Execution Framework -MapReduce Algorithm Design –Shuffling, Grouping, Sorting- Custom Partitioners and Combiners- MapReduce Formats and Features.

Module:4	**Algorithms for Handling Big Data**	**6 hours**
Random Forest Algorithm, Unstructured Data Analytics, Randomized Matrix Algorithms in Parallel and Distributed Environments, Mahout: Probabilistic Hashing for Efficient Search and Learning on Massive Data, Dirichlet process clustering, Latent Dirichlet Allocation, Singular value decomposition, Parallel Frequent Pattern mining, Complementary Naive Bayes classifier, Random forest decision tree based classifier.		
Module:5	**Lambda Architecture**	**6 hours**
Different layers of Lambda Architecture, Data storage on the batch layer. Serving Layer- Requirements for a serving layer database, Indexing strategies. Speed Layer- Storing and Computing Real time views, Queuing and Streaming – Illustration using Cassandra data model.		
Module:6	**Big Data Clustering**	**6 hours**
K-means Algorithms - K-Means Basics - Initializing Clusters for K-Means -Picking the Right Value of k - The Algorithm of Bradley, Fayyad, and Reina - Processing Data in the BFR Algorithm.		
Module:7	**Mining Social Network Graphs**	**6 hours**
Link Analysis: Page Rank- Efficient computation of Page Rank- Topic Sensitive Page Rank- Link Spam- Hubs and Authorities. Mining Social Network Graphs: Web Advertising: Online and Offline Algorithms; Social Network Graphs: Clustering of Social Network Graphs- Direct Discovery of Communities- Partitioning of Graphs- Finding overlapping communities- Simrank- Counting Triangles- Neighborhood properties of Graphs.		
Module:8	**Contemporary issues:**	**3 hours**
	Total Lecture hours:	**45 hours**

Text Book(s)	
1.	Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill, 2015.
Reference Books	
1.	Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Jimmy, Morgan & Claypool Synthesis, 2010.
2.	Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, Cambridge University Press, 2014.
3.	Tom White, Hadoop, the Definitive guide, O'Reilly Media, 2015.
4.	Noreen Burlingame, Little Book of Big Data, Ed. 2016.

Recommended by Board of Studies	05-03-2016		
Approved by Academic Council	No. 40	Date	18-03-2016