Course Code	Course/Subject Name	Credits
CSC702	**Big Data Analytics**	**3**

Prerequisite: Some prior knowledge about Java programming, Basics of SQL, Data mining and machine learning methods would be beneficial.	
Course Objectives:	
1	To provide an overview of an exciting growing field of big data analytics.
2	To introduce programming skills to build simple solutions using big data technologies such as MapReduce and scripting for NoSQL, and the ability to write parallel algorithms for multiprocessor execution
3	To teach the fundamental techniques and principles in achieving big data analytics with scalability and streaming capability.
4	To enable students to have skills that will help them to solve complex real-world problems in decision support.
5	To provide an indication of the current research approaches that is likely to provide a basis for tomorrow's solutions.
Course Outcomes:	
1	Understand the key issues in big data management and its associated applications for business decisions and strategy.
2	Develop problem solving and critical thinking skills in fundamental enabling techniques like Hadoop, Map reduce and NoSQL in big data analytics.
3	Collect, manage, store, query and analyze various forms of Big Data.
4	Interpret business models and scientific computing paradigms,and apply software tools for big data analytics.
5	Adapt adequate perspectives of big data analytics in various applications like recommender systems, social media applications etc.
6	Solve Complex real world problems in various applications like recommender systems, social media applications, health and medical systems, etc.

Module	Detailed Contents	Hours
01	**Introduction to Big Data & Hadoop** 1.1 Introduction to Big Data, 1.2 Big Data characteristics, types of Big Data, 1.3 Traditional vs. Big Data business approach, 1.4 Case Study of Big Data Solutions. 1.5 Concept of Hadoop 1.6 Core Hadoop Components; Hadoop Ecosystem	**06**
02	**Hadoop HDFS and Map Reduce** 2.1 Distributed File Systems: Physical Organization of Compute Nodes, Large-Scale File-System Organization. 2.2 MapReduce: The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures. 2.3 Algorithms Using MapReduce: Matrix-Vector Multiplication by MapReduce, Relational-Algebra Operations, Computing Selections by MapReduce, Computing Projections by MapReduce, Union, Intersection, and Difference by MapReduce 2.4 Hadoop Limitations s.	**10**
03	**NoSQL** 3.1 Introduction to NoSQL, NoSQL Business Drivers, 3.2 NoSQL Data Architecture Patterns: Key-value stores, Graph stores, Column family (Bigtable)stores, Document stores, Variations of NoSQL architectural patterns, NoSQL Case Study 3.3 NoSQL solution for big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer; NoSQL systems to handle big data problems. peer-to-peer; Four ways that NoSQL systems handle big data problems	**06**
04	**Mining Data Streams** 4.1 The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing. 4.2 Sampling Data techniques in a Stream 4.3 Filtering Streams: Bloom Filter with Analysis. 4.4 Counting Distinct Elements in a Stream, Count-Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements 4.5 Counting Frequent Items in a Stream, Sampling Methods for Streams, Frequent Itemsets in Decaying Windows. 4.6 Counting Ones in a Window: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.	**12**
05	**Finding Similar Items and Clustering** 5.1 Distance Measures: Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance. 5.2 CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets, Answering Queries.	**08**
06	**Real-Time Big Data Models** 6.1 PageRank Overview, Efficient computation of PageRank: PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector. 6.2 A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering. 6.3 Social	**10**

	Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph.	

Textbooks:	
1	Anand Rajaraman and Jeff Ullman —Mining of Massive Datasets‖, Cambridge University Press,
2	Alex Holmes —Hadoop in Practice‖, Manning Press, Dreamtech Press.
3	Dan Mcary and Ann Kelly —Making Sense of NoSQL‖ – A guide for managers and the rest of us, Manning Press.
References:	
1	Bill Franks , —Taming The Big Data Tidal Wave: Finding Opportunities In Huge Data Streams With Advanced Analytics‖, Wiley
2	Chuck Lam, —Hadoop in Action‖, Dreamtech Press
3	Jared Dean, —Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners‖, Wiley India Private Limited, 2014.
4	Jiawei Han and Micheline Kamber, —Data Mining: Concepts and Techniques‖, Morgan Kaufmann Publishers, 3rd ed, 2010.
5	Lior Rokach and Oded Maimon, —Data Mining and Knowledge Discovery Handbook‖, Springer, 2nd edition, 2010.
6	Ronen Feldman and James Sanger, —The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data‖, Cambridge University Press, 2006.
7	Vojislav Kecman, —Learning and Soft Computing‖, MIT Press, 2010

Assessment:	
Internal Assessment:	
The assessment consists of two class tests of 20 marks each. The first class test is to be conducted when approx. 40% syllabus is completed and second class test when additional 40% syllabus is completed. Duration of each test shall be one hour.	
End Semester Theory Examination:	
1	Question paper will comprise a total of six questions.
2	All questions carry equal marks.

3	Question 1 and question 6 will have questions from all modules. Remaining 4 questions will be based on the remaining 4 modules.
4	Only four questions need to be solved.
5	In question paper weightage of each module will be proportional to the number of respective lecture hours as mentioned in the syllabus.