Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous Institute Affiliated to University of Mumbai)

Course Code	Course Name	Teaching Scheme (Hrs/week)			Credits Assigned			
		L	T	P	L	T	P	Total
CPE8035	Elective-III Big Data Analytics	4	-	--	4	-	--	4
		Examination Scheme						
		ISE		MSE	ESE			
		10		30	100 (60% Weightage)			

Pre-requisite Course Codes		-
At end of successful completion of this course, student will be able to		
Course Outcomes	CO1	Identify challenges in big data management and inadequacy of existing technology to analyze big data.
	CO2	Apply scalable algorithms based on Hadoop and Map Reduce to perform Big Data Analytics..
	CO3	Apply NoSQL tools to solve big data problems.
	CO4	Use stream data model to provide real time analysis of big data.
	CO5	Discover information from social network graphs.

Module No.	Topics	Ref.	Hrs.
1	**Introduction to Big Data** Introduction to Big Data, Big Data characteristics, types of Big Data,Traditional vs. Big Data business approach, Case Study of Big DataSolutions.	1-5	**03**
2	**Introduction to Hadoop** What is Hadoop? Core Hadoop Components; Hadoop Ecosystem;Physical Architecture; Hadoop limitations.	1-5	**03**
3	**NoSQL** What is NoSQL? NoSQL business drivers; NoSQL case studies NoSQL data architecture patterns: Key-value stores, Graph stores,Column family (Bigtable) stores, Document stores, Variations of NoSQLarchitectural patterns; Using NoSQL to manage big data: What is a big data NoSQL solution?Understanding the types of big data problems; Analyzing big data with ashared-nothing architecture; Choosing distribution models: master-slaveversus peer-to-peer; Four ways that NoSQL systems handle big dataproblems	1-5	**04**
4	**MapReduce and the New Software Stack** Distributed File Systems: Physical Organization of Compute Nodes,	1-5	**06**

	Large- Scale File-System Organization. MapReduce: The Map Tasks, Grouping by Key, The Reduce Tasks Combiners, Details of MapReduce Execution, Coping With Node Failures, Algorithms Using MapReduce: Matrix-Vector Multiplication by MapReduce,Relational-Algebra Operations, Computing Selections by MapReduce,Computing Projections by MapReduce, Union, Intersection, and Difference byMapReduce, Computing Natural Join by MapReduce, Grouping andAggregation by MapReduce, One MapReduce Step.		
5	**Finding Similar Items** Applications of Near-Neighbor Search, Jaccard Similarity of Sets,Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem, Distance Measures: Definition of a Distance Measure, Euclidean, Distances, Jaccard Distance, Cosine Distance, Edit Distance, HammingDistance.	1-5	**03**
6	**Mining Data Streams** The Stream Data Model: A Data-Stream-Management System,Examples of Stream Sources, Stream Querie, Issues in Stream Processing, Sampling Data in a Stream: Obtaining a Representative Sample, TheGeneral Sampling Problem, Varying the Sample Size. Filtering Streams: The Bloom Filter, Analysis. 6.4 Counting Distinct Elements in a StreamThe Count-Distinct Problem, The Flajolet-Martin Algorithm, CombiningEstimates, Space Requirements. Counting Ones in a Window:The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm,Query Answering in the DGIM Algorithm, Decaying Windows.	1-5	**06**
7	**Link Analysis** PageRank Definition, Structure of the web, dead ends, Using Page ranking a search engine, Efficient computation of Page Rank: PageRankIteration Using MapReduce, Use of Combiners to Consolidate the Result Vector, Topic sensitive Page Rank, link Spam, Hubs and Authorities.	1-5	**05**
8	**Frequent Itemsets** Handling Larger Datasets in Main MemoryAlgorithm of Park, Chen, and Yu, The Multistage Algorithm, The Multihash Algorithm, The SON Algorithm and MapReduce, Counting Frequent Items in a StreamSampling Methods for Streams, Frequent Itemsets in Decaying Windows	1-5	**05**
9	**Clustering** CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm,Initializing & Merging Buckets, Answering Queries	1-5	**05**
10	**Recommendation Systems** A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering.	1-5	**04**

11	**Mining Social-Network Graphs** Social Networks as Graphs, Clustering of Social-Network Graphs, DirectDiscovery of Communities, SimRank, Counting triangles using Map-Reduce	1-5	**04**
		Total	**48**

References:

[1] AnandRajaraman and Jeff Ullman "Mining of Massive Datasets", Cambridge University Press,

[2] Alex Holmes "Hadoop in Practice", Manning Press, Dreamtech Press.

[3] Dan McCreary and Ann Kelly "Making Sense of NoSQL" – A guide for managers and the rest of us, Manning Press.

[4] Bill Franks , "Taming The Big Data Tidal Wave: Finding Opportunities In Huge Data Streams With Advanced Analytics", Wiley

[5] Chuck Lam, "Hadoop in Action", Dreamtech Press