**B.E. (Sem. VII)**

Course / Subject Name: **BigData Analytics**

Course Code: CPE8035

Academic Term: Jan 2018 – June 2018

Syllabus:

|  |  |  |
| --- | --- | --- |
| **Module** | **Contents** | **Hours** |
| **1** | **Introduction to Big Data**  1.1 Introduction to Big Data, Big Data characteristics, types of Big Data, Traditional vs. Big Data business approach, Case Study of Big Data Solutions. | **03** |
| **2** | **Introduction to Hadoop**  2.1 What is Hadoop? Core Hadoop Components; Hadoop Ecosystem; Physical Architecture; Hadoop limitations. | **03** |
| **3** | **NoSQL**  3.1 What is NoSQL? NoSQL business drivers; NoSQL case studies;  3.2 NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns;  3.3 Using NoSQL to manage big data: What is a big data NoSQL solution? Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer; Four ways that NoSQL systems handle big data problems |  |
| **4** | **MapReduce and the New Software Stack**  4.1 Distributed File Systems: Physical Organization of Compute Nodes, LargeScale File-System Organization.  4.2 MapReduce: The Map Tasks, Grouping by Key, The Reduce Tasks, 06 Combiners, Details of MapReduce Execution, Coping With Node Failures.  4.3 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, Computing Natural Join by MapReduce, Grouping and Aggregation by MapReduce, Matrix Multiplication, Matrix Multiplication with One MapReduce Step. | **04** |
| **5** | **Finding Similar Items**  5.1 Applications of Near-Neighbor Search, Jaccard Similarity of Sets, Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem  5.2 Distance Measures: Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance | **06** |
| **6** | **Mining Data Streams**  6.1 The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing.  6.2 Sampling Data in a Stream: Obtaining a Representative Sample, The General Sampling Problem, Varying the Sample Size.  6.3 Filtering Streams: The Bloom Filter, Analysis.  6.4 Counting Distinct Elements in a Stream The Count-Distinct Problem, The Flajolet-Martin Algorithm, Combining Estimates, Space Requirements.  6.5 Counting Ones in a Window: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows. | **03** |
| **7** | **Link Analysis**  7.1 PageRank Definition, Structure of the web, dead ends, Using Page rank in a search engine, Efficient computation of Page Rank: PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector.  7.2 Topic sensitive Page Rank, link Spam, Hubs and Authorities. | **06** |
| **8** | Frequent Itemsets  8.1 Handling Larger Datasets in Main Memory Algorithm of Park, Chen, and Yu, The Multistage Algorithm, The Multihash Algorithm.  8.2 The SON Algorithm and MapReduce  8.3 Counting Frequent Items in a Stream Sampling Methods for Streams, Frequent Itemsets in Decaying Windows | **05** |
| **9** | **Clustering**  9.1 CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets, Answering Queries | **05** |
| **10** | **Recommendation Systems**  10.1 A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering | **04** |
| **11** | **Mining Social-Network Graphs**  11.1 Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, SimRank, Counting triangles using MapReduce | **04** |