GUJARAT TECHNOLOGICAL UNIVERSITY
Integrated- MCA
Year – V (Semester – IX) (W.E.F. JUNE 2017)

Subject Name: Big Data Analytics (BDA)
Subject Code: 4490605

1. Objectives:

1. To understand basics of Big Data and Big Data Tools (Hadoop, MapReduce)
2. To understand fundamental techniques used for Big data analytics

2. Prerequisites: Working knowledge of Programming Language and Database Concepts

3. Course Contents:

Sr. No.	Course Content	No. of sessions
1	**Unit 1: Introduction** **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	**07**
2	**Unit 2: MapReduce** **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. **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 MapReduce, Matrix Multiplication, Matrix Multiplication with One MapReduce Step.	**06**
3	**Unit 3: 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)	**03**

4	**Unit 4: 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.	**04**
5	**Unit 5: 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.	**04**
6	**Unit 6: Frequent Item Set Mining** Market-Basket Model, Algorithm for finding Frequent Itemsets, Handlig Larger Datasets in Main Memory, Algorithm of Park, Chen, and Yu, The Multistage Algorithm, The Multihash Algorithm, The SON Algorithm and MapReduce, Counting Frequent Items in a Stream, Sampling Methods for Streams, Frequent Itemsets in Decaying Windows	**05**
7	**Unit 7 : Clustering** Overview of clustering techniques, Hierarchical Clustering, Partitioning Methods, CURE algorithm, Clustering stream	**04**
8	**Unit 8: 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	**07**

4. Text Book(s):

1. Radha Shankarmani, M Vijayalakshmi, "Big Data Analytics", 2nd Edition, Wiley

5. Other Reference Books:

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

Web Resources

a. http://www.bigdatauniversity.com
b. http://www.mongodb.com
c. http://hadoop.apache.org/

6. Unit wise coverage from Text book(s):

Unit 1	Topics
I	Chapter 1.1 to 1.4,2,3
II	Chapter 4
III	Chapter 5
IV	Chapter 6
V	Chapter 7
VI	Chapter 8
VII	Chapter 9
VIII	Chapter 10,11

7. Accomplishment

Students can design algorithms by employing Map Reduce technique for solving Big Data problems

8. Laboratory Exercises

1. Basics (not to be considered for Practical Exam)
    a. Study of Hadoop ecosystem
    b. Programming exercises on Hadoop e.g.-Word count program
    c. Programming exercises in NoSQL/MongoDB
    d. Implementing simple algorithms in Map- Reduce - Matrix multiplication, Aggregates, joins, sorting, searching etc.
2. Implementing Frequent Itemset algorithm using Map-Reduce
3. Implementing Clustering algorithm using Map-Reduce
4. Implementing any one data streaming algorithm using Map-Reduce
5. Consider very large vectors indicsating n-dimensional data points, Write a program using MapReduce for computing distances (Jaccard Distance and Cosine Distance)
6. Real life large data application to be implemented (Use standard Datasets available on the web)
    a. Twitter data analysis
    b. Fraud Detection
    c. Text Mining
    d. Plagiarism Analysis
    e. Basket Analysis
7. Chapter End programming assignments from text Book