| Course Code | Course Name | Teaching Scheme Hrs./Week | | | Credits Assigned | | | |
|---|---|---|---|---|---|---|---|---|
| | | Theory | Practical | Tutorial | Theory | Practical/Oral | Tutorial | Total |
| BEITC802 | Big Data Analytics | 04 | 02 | --- | 04 | 01 | --- | 05 |

| Course Code | Course Name | Examination Scheme | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Theory Marks | | | | Term Work | Practical | Oral | Total |
| | | Internal assessment | | | End Sem. Exam | | | | |
| | | Test 1 | Test 2 | Avg. of 2 Tests | | | | | |
| BEITC802 | Big Data Analytics | 20 | 20 | 20 | 80 | 25 | --- | 25 | 150 |

**Course Objectives:**

1. To provide an overview of an exciting growing field of big data analytics.
2. To introduce the tools required to manage and analyze big data like Hadoop, NoSql Map-Reduce.
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 for decision support.

**Course Outcomes: At the end of this course a student will be able to:**

1. Understand the key issues in big data management and its associated applications in intelligent business and scientific computing.
2. Acquire fundamental enabling techniques and scalable algorithms like Hadoop, Map Reduce and NO SQL in big data analytics.
3. Interpret business models and scientific computing paradigms, and apply software tools for big data analytics.
4. Achieve adequate perspectives of big data analytics in various applications like recommender systems, social media applications etc.

**DETAILED SYLLABUS:**

| Sr. No. | Module | Detailed Content | Book | Hours |
|---|---|---|---|---|
| 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 Data Solutions. | **From Ref. Books** | **03** |
| 2 | Introduction to Hadoop | What is Hadoop? Core Hadoop Components; Hadoop Ecosystem; Physical Architecture; Hadoop limitations. | Hadoop in Practise<br>Chapter 1 | **02** |
| 3 | NoSQL | 1. What is NoSQL? NoSQL business drivers; NoSQL case studies;<br>2. NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns;<br>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 | **No-SQL book** | **04** |
| 4 | MapReduce and the New Software Stack | **Distributed File Systems :** Physical Organization of Compute Nodes, Large-Scale File-System Organization.<br>**MapReduce:** The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures.<br>**Algorithms Using MapReduce**:<br>Matrix-Vector Multiplication by MapReduce , Relational-Algebra Operations, Computing Selections by MapReduce,<br>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. | **Text Book 1** | **06** |

| | | | | |
|---|---|---|---|---|
| 5 | Finding Similar Items | Applications of Near-Neighbor Search, Jaccard Similarity of Sets, Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem .<br>**Distance Measures:** Definition of a Distance Measure , Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance. | **Text Book 1** | **03** |
| 6 | Mining Data Streams | **The Stream Data Model**: A Data-Stream-Management System, Examples of Stream Sources, Stream Querie, Issues in Stream Processing.<br>**Sampling Data in a Stream** : Obtaining a Representative Sample , The General Sampling Problem, Varying the Sample Size.<br>**Filtering Streams**:<br>The Bloom Filter, Analysis.<br>**Counting Distinct Elements in a Stream**<br>The Count-Distinct Problem, The Flajolet-Martin Algorithm, Combining Estimates, Space Requirements .<br>**Counting Ones in a Window**:<br>The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows. | **Text Book 1** | **06** |
| 7 | Link Analysis | 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.<br>Topic sensitive Page Rank, link Spam, Hubs and Authorities. | **Text Book 1** | **05** |
| 8 | Frequent Itemsets | **Handling Larger Datasets in Main Memory**<br>Algorithm of Park, Chen, and Yu, The Multistage Algorithm, The Multihash Algorithm.<br>**The SON Algorithm and MapReduce**<br>**Counting Frequent Items in a Stream**<br>Sampling Methods for Streams, Frequent Itemsets in Decaying Windows | **Text Book 1** | **05** |
| 9 | Clustering | CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets, | **Text** | **05** |

| | | | | |
|---|---|---|---|---|
| | | Answering Queries | **Book 1** | |
| 10 | Recommendation Systems | A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering. | **Text Book 1** | **04** |
| 11 | Mining Social-Network Graphs | Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, SimRank, Counting triangles using Map-Reduce | **Text Book 1** | **05** |

**Text Books:**


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 McCreary 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. Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman, "**Big Data for Dummies",** Wiley India
4. Michael Minelli, Michele Chambers, Ambiga Dhiraj, "**Big Data Big Analytics: Emerging Business Intelligence And Analytic Trends For Today's Businesses",** Wiley India
5. Phil Simon**, "Too Big To Ignore: The Business Case For Big Data",** Wiley India
6. Paul Zikopoulos, Chris Eaton**, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data',** McGraw Hill Education.
7. Boris Lublinsky, Kevin T. Smith, Alexey Yakubovich**, "Professional Hadoop Solutions",** Wiley India.

**Oral Exam:**

An oral exam will be held based on the above syllabus.

**Term work:**

Assign a case study for group of 2/3 students and each group to perform the following experiments on their case-study; Each group should perform the exercises on a large dataset created by them.

**Term work: (15 marks for programming exercises + 10 marks for mini-project)**

**Suggested Practical List:** Students will perform at least 8 programming exercises and implement one mini-project. The students can work in groups of 2/3.

1. Study of Hadoop ecosystem
2. 2 programming exercises on Hadoop
3. 2 programming exercises in No SQL
4. Implementing simple algorithms in Map- Reduce (3) - Matrix multiplication, Aggregates, joins, sorting, searching etc.
5. Implementing any one Frequent Itemset algorithm using Map-Reduce
6. Implementing any one Clustering algorithm using Map-Reduce
7. Implementing any one data streaming algorithm using Map-Reduce
8. Mini Project: One 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 etc.

**Theory Examination:**

- Question paper will comprise of 6 questions, each carrying 20 marks.
- Total 4 questions need to be solved.
- Q.1 will be compulsory, based on entire syllabus where in sub questions of 2 to 3 marks will be asked.
- Remaining question will be randomly selected from all the modules.

Weight age of marks should be proportional to number of hours assigned to each module.