Kadi Sarva Vishwavidyalaya

Faculty of Engineering & Technology

Fourth Year Bachelor of Engineering (Computer)

(To be Proposed For: Academic Year 2020-21)

Subject Code: CE802-N	**Subject Title: Big Data Analytics**
Pre-requisite	

Teaching Scheme (Credits and Hours)

Teaching scheme				**Total Credit**	**Evaluation Scheme**					
L	**T**	**P**	**Total**		**Theory**		**Mid Sem Exam**	**CIA**	**Pract.**	**Total**
Hrs	**Hrs**	**Hrs**	**Hrs**		**Hrs**	**Marks**	**Marks**	**Marks**	**Marks**	**Marks**
03	00	02	05	04	03	70	30	20	30	150

Course Objectives:

- To provide an overview of an exciting growing field of big data analytics.
- To introduce the tools required to manage and analyze big data like Hadoop, NoSql Map-Reduce.
- To teach the fundamental techniques and principles in achieving big data analytics with scalability and streaming capability.
- To enable students to have skills that will help them to solve complex real-world problems in for decision support.

Outline of the Course:

Sr. No	**Title of the Unit**	**Minimum Hours**
1	Introduction to Big Data	06
2	Mining Data Streams	05
3	**Big Data Analytics and Big Data Analytics Techniques:**	06
4	Link Analysis	04
5	Frequent Item sets	07
6	Mining Social Network Graphs	05
7	NoSQL	04
8	Map Reduce and New Software Stack	05
9	Big Data Analytics Applications/Use cases and Visualization of Big Data	06

Total hours (Theory):48
Total hours (Lab):32
Total hours: 80

Kadi Sarva Vishwavidyalaya

Faculty of Engineering & Technology

Fourth Year Bachelor of Engineering (Computer)

(To be Proposed For: Academic Year 2020-21)

Detailed Syllabus

Sr. No	Topics	Lecture Hours	Weightage (%)
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. Structured Data, unstructured Data and semi Structured Data.	06	13
2	**Mining Data Streams:** **The Stream Data Model:** A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing. **Sampling Data in a Stream:** Obtaining a Representative Sample, The General Sampling Problem, Varying the Sample Size. **Filtering Streams:** The Bloom Filter, Analysis. **Counting Distinct Elements in a Stream** The Count-Distinct Problem, The Flajolet-Martin Algorithm, Combining Estimates, Space Requirements **Counting Ones in a Window:** The Cost of Exact Counts, The Datar-Gionis-Indyk- Motwani Algorithm, Query Answering in the DGIM	05	10
3	**Big Data Analytics and Big Data Analytics Techniques:** Big Data and its Importance, Drivers for Big data, Optimization techniques, Dimensionality Reduction techniques, Time series Forecasting, Social Media Mining and Social Network Analysis and its Application, Big Data analysis using Hadoop, Pig, Hive, Mongodb, Spark and Mahout, Data analysis techniques like Discriminant Analysis and Cluster Analysis,	06	13
4	**Link Analysis:** Page Rank Definition, Structure of the web, dead ends, Using Page rank in a search engine, Efficient computation of Page Rank: Page Rank Iteration Using Map Reduce, Use of Combiners to Consolidate the Result Vector. Topic sensitive Page Rank, link Spam, Hubs and Authorities.	04	08
5	**Frequent Itemsets:** **Handling Larger Datasets in Main Memory** Algorithm of Park, Chen, and Yu, The Multistage Algorithm, The Multihash Algorithm. **The SON Algorithm and Map Reduce Counting Frequent Items in a Stream** Sampling Methods for Streams, Frequent Itemsets in Decaying Windows.	07	15

6	**Mining Social-Network Graphs:** Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, SimRank, Counting triangles using Map-Reduce	05	10
7	**NoSQL:** 1. What is NoSQL? NoSQL business drivers; NoSQL case studies; 2. NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns; 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	04	08
8	**Map Reduce and New Software Stack:** **Distributed File Systems:** Physical Organization of Compute Nodes, 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 by MapReduce, Computing Natural Join by MapReduce, Grouping and Aggregation by MapReduce, Matrix Multiplication, Matrix Multiplication with One MapReduce Step.	05	10
9	**Big Data Analytics Applications/Use cases and Visualization of Big Data:** **Big Data Analytics Applications:** Big Data Analytics in E-Governance & Society, Applications in Science, Engineering, Healthcare, Visualization, Business etc. Case Study of Existing Big Data Analytics Systems. **Use cases and Visualization of Big Data:** Big Data visualization with the tools like D3, Kibana, and Grafana, Scala and Python for Data Visualization,	06	13
	Total	**48**	**100**

Instructional Method and Pedagogy:

At the start of course, the course delivery pattern, prerequisite of the subject will be discussed.

- Lectures will be conducted with the aid of multi-media projector, black board etc.
- Attendance is compulsory in lectures and laboratory which carries 5 marks weightage.

- One Internal exam will be conducted and same will be converted to equivalent of 15 Marks as a part of CIA.
- Assignments based on course will be given to the students at the end of each unit/topic and will be evaluated at regular interval.
- Surprise tests/Quizzes/Seminar will be conducted
- The Course includes a laboratory where students have an opportunity to build an appreciation for the concepts being taught in lectures.
- Experiments/Tutorials related to course content will be carried out in the laboratory.

Students Learning Outcome:

- Students will to build and maintain reliable, scalable, distributed systems with Apache Hadoop and Spark.
- Students will be able to write Map-Reduce based Applications
- Students will be able to design and build MongoDB based Big data Applications and learn MongoDB query Language
- Students will learn difference between conventional SQL query language and NoSQL and Graph processing and Visualization.

Students will learn tips and tricks for Big Data use cases and solutions.

E-Resources:

- http://www.bigdatauniversity.com

Reference Books:

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 andthe rest of us**, Manning Press.
4. Bart Baesens , Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, ,Wiley, 2014

References:

1. Bill Franks **, "Taming The Big Data Tidal Wave: Finding Opportunities In HugeData 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 EnterpriseClass Hadoop and Streaming Data',** McGraw Hill Education.
7. Boris Lublinsky, Kevin T. Smith, Alexey Yakubovich**, "Professional HadoopSolutions",** Wiley India.
8. Dirk Deroos et al., Hadoop for Dummies, Dreamtech Press, 2014.
9. Leskovec, Rajaraman, Ullman, Mining of Massive Datasets, Cambridge University Press.
10. I.H. Witten and E. Frank, Data Mining: Practical Machine learning tools and techniques.

List of experiments

No	Name of Experiment
1	Study of Hadoop ecosystem
2	Programming exercises on Hadoop
3	Programming exercises in No SQL
4	Programming exercises in MongoDB
5	Implementing simple algorithms in Map- Reduce (3) - Matrix multiplication, Aggregates, joins, sorting, searching etc.
6	Implementing any one Frequent Itemset algorithm using Map-Reduce
7	Implementing any one Clustering algorithm using Map-Reduce
8	Implementing any one data streaming algorithm using Map-Reduce
9	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.

Students will perform at least 8 programming exercises and implement one mini-project. The students can work in groups of 2 or 3.