INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

ALGORITHMIC TECHNIQUES FOR BIG DATA								
VIII Semester: CSE(DS)								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
ACDD34	Open Elective	3	0	0	3	40	60	100
Contact Classes: 48	**Tutorial Classes: Nil**	**Practical Classes: Nil**				**Total Classes: 48**		
Prerequisites: Basic knowledge of data structures, algorithms, probability theory, discrete mathematics, and programming in Python or Java.								

I. COURSE OVERVIEW:

This course introduces algorithmic paradigms and computational models tailored for processing and analyzing large-scale datasets efficiently. Students will learn sublinear and streaming algorithms, distributed computing frameworks like MapReduce, and probabilistic techniques used to handle massive data. Emphasis is placed on designing scalable algorithms with bounded resources (time and memory) and applying them to real-world big data problems in search, graphs, machine learning, and web analytics.

II. COURSE OBJECTIVES:

The students will try to learn:

I. The foundational models and challenges in algorithm design for massive datasets.
II. The core algorithmic techniques such as sampling, sketching, and streaming for sublinear computation.
III. The principles behind scalable distributed processing using MapReduce and Spark frameworks.
IV. The design and analysis of algorithms for large-scale applications in graph processing, data mining, and web search.

III. COURSE OUTCOMES:

At the end of the course students will be able to:

CO1 Explain the computational challenges and models used in big data algorithms.
CO2 Apply sampling, sketching, and streaming algorithms to approximate large-scale computations.
CO3 Implement distributed algorithms using MapReduce and Spark for parallel data processing.
CO4 Design efficient algorithms for searching, ranking, clustering, and classification at scale.
CO5 Analyze performance and resource trade-offs in memory-constrained and parallel environments.
CO6 Solve real-world big data problems by applying scalable algorithmic techniques.

IV. COURSE CONTENT:

Module I: Introduction to Big Data Algorithms (10)
Big data characteristics and computational bottlenecks, RAM vs. streaming model, External memory model, Communication complexity, Streaming model introduction, One-pass algorithms, Count-Min Sketch, Bloom filters, Hashing and fingerprinting, Basic probability review.

Module II: Sampling and Sketching Techniques (10)
Random sampling algorithms, Reservoir sampling, Priority sampling, Frequent item estimation, Heavy hitters problem, HyperLogLog for cardinality estimation, AMS Sketch for second moment, Histogram approximation, Quantile computation, Space vs. accuracy trade-offs.

Module III: Streaming and Sublinear Algorithms (10)
Sliding window algorithms, Approximate counting, Misra-Gries algorithm, Data streams vs batch processing, Dimensionality reduction basics.

Johnson-Lindenstrauss lemma, Locality-sensitive hashing (LSH), Matrix sketching and CUR decomposition, Sampling in high dimensions, Estimating norms and similarities.

Module IV: Distributed Models and MapReduce (10)
MapReduce programming model, Complexity analysis in MapReduce, Distributed sorting and searching, Matrix multiplication in MapReduce, PageRank algorithm, Join algorithms, Filtering and aggregation, MapReduce implementation with Hadoop, Introduction to Spark, Parallel algorithm design patterns.

Module V: Applications and Case Studies (8)
Graph algorithms for large graphs, Clustering and community detection, Recommendation systems, Mining data streams, Large-scale learning with stochastic gradient descent, Web crawling and indexing, Text and log data analysis, Case studies from industry (e.g., Google, Facebook, Netflix).

V. TEXTBOOKS:

1. S. Muthukrishnan, *Data Streams: Algorithms and Applications*, Now Publishers, 2005.
2. Rajeev Motwani and Jeffrey Ullman, *Mining of Massive Datasets*, Cambridge University Press, 3rd Edition, 2020.
3. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, *Mining of Massive Datasets*, Stanford University Textbook, Latest Edition.

VI. REFERENCE BOOKS:

1. S. Dasgupta, C.H. Papadimitriou, and U.V. Vazirani**, *Algorithms*, McGraw-Hill, 2008.**
2. Charu C. Aggarwal**, *Data Mining: The Textbook*, Springer, 2015.**
3. Andreas C. Müller & Sarah Guido**, *Introduction to Machine Learning with Python*, O'Reilly Media, 2016.**

VII. ELECTRONICS RESOURCES:

1. Stanford University – Mining of Massive Datasets Online Course
   http://www.mmds.org
2. NPTEL – Big Data Computing by Prof. Rajiv Misra (IIT Patna)
   https://nptel.ac.in/courses/106105175
3. MIT OpenCourseWare – Algorithms for Massive Data Sets
   https://ocw.mit.edu/courses/6-854j-advanced-algorithms-fall-2008/

4. Coursera – Algorithms for Big Data (University of California San Diego)
   https://www.coursera.org/learn/algorithms-for-big-data

VIII. MATERIALS ONLINE

1. Course template
2. Tutorial question bank
3. Tech talk topics
4. Open-ended experiments
5. Definitions and terminology
6. Assignments
7. Model question paper – I
8. Model question paper – II
9. Lecture notes
10. PowerPoint presentation
11. E-Learning Readiness Videos (ELRV)