COURSE CONTENT

MINING MASSIVE DATASETS								
I Semester: CSE								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
2515811	**Foundation**	3	0	0	3	40	60	100
Contact Classes: 45	**Tutorial Classes: Nil**	**Practical Classes: Nil**				**Total Classes: 45**		
Prerequisites: Students should be familiar with Data mining, algorithms, basic probability theory and Discrete math.								

Course Overview:

This course that focuses on techniques, algorithms, and systems for extracting meaningful patterns, trends, and knowledge from extremely large-scale datasets. With the exponential growth of data generated from social media, sensors, scientific computing, and online platforms, traditional data processing methods become inefficient.

Course Objectives:

1. To understand the fundamentals of large-scale data mining and distributed data processing techniques.
2. To develop skills in implementing MapReduce and parallel algorithms for big data analytics.
3. To apply similarity search, stream processing, and frequent itemset mining techniques on massive datasets.
4. To design recommendation systems, web advertising models, and ranking algorithms for online platforms.
5. To analyze social network graphs and clustering techniques for extracting meaningful insights from complex data.

Course Outcomes: After Completion of the Course, Students should be able to

1. Apply MapReduce and distributed file systems to process and analyze large-scale datasets for real-time analytics applications.
2. Implement similarity search and streaming data techniques to detect near-duplicate documents and monitor real-time data streams.
3. Analyze link structures, frequent itemset, and clustering algorithms to improve recommendation and ranking systems for e-commerce platforms.
4. Design web advertising and recommendation system algorithms, including collaborative filtering and dimensionality reduction, for personalized marketing.
5. Evaluate social network graph mining techniques, including graph clustering and partitioning, to extract insights for social media analytics.

UNIT - I: Data Mining-Introduction-Definition of Data Mining-Statistical Limits on Data Mining, **MapReduce and the New Software Stack-**Distributed File Systems, MapReduce, Algorithms Using MapReduce.

UNIT - II: Similarity Search: Finding Similar Items-Applications of Near-Neighbor Search, Shingling of Documents, Similarity-Preserving Summaries of Sets, Distance Measures.
Streaming Data: Mining Data Streams-The Stream Data Model, Sampling Data in a Stream, Filtering Streams.

UNIT - III: Link Analysis-PageRank, Efficient Computation of PageRank, Link Spam
Frequent Itemsets-Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream.
Clustering-The CURE Algorithm, Clustering in Non-Euclidean Spaces, Clustering for Streams and Parallelism.

UNIT - IV: **Advertising on the Web-**Issues in On-Line Advertising, On-Line Algorithms, The Matching Problem, The Adwords Problem, Adwords Implementation. **Recommendation Systems-**A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering, Dimensionality Reduction, The Netflix Challenge.

UNIT - V: Mining Social-Network Graphs-Social Networks as Graphs, Clustering of Social-Network Graphs, Partitioning of Graphs, Simrank, Counting Triangles.

TEXT BOOKS:

1. Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets, 3rd Edition.

REFERENCE BOOKS:

1. Jiawei Han & Micheline Kamber, Data Mining – Concepts and Techniques 3rd Edition Elsevier.
2. Margaret H Dunham, Data Mining Introductory and Advanced topics, PEA.
3. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann.

ELECTRONIC RESOURCES:

1. https://www.coursera.org/specializations/data-mining/
2. https://www.classcentral.com/course/mining-stanford-university-mining-massive-dataset-2406/
3. https://www.mygreatlearning.com/academy/learn-for-free/courses/data-mining1/
4. https://www.mygreatlearning.com/academy/learn-for-free/courses/mastering-big-data-analytics/

MATERIALS ONLINE:

1. Course template
2. Tutorial question bank
3. Tech talk and Concept Video 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. E-Learning Readiness Videos (ELRV)