COURSE CONTENT

MINING MASSIVE DATASETS								
I Semester: CSE								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
2215818	**Foundation**	**L**	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
		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 fundamental concepts and challenges involved in mining and processing massive datasets.
2. To learn parallel and distributed computing techniques such as MapReduce for handling large-scale data efficiently.
3. To study algorithms for similarity search, streaming data analysis, clustering, and frequent itemset mining.
4. To explore recommendation systems, web advertising models, and social network graph mining techniques.
5. To develop the ability to design and implement scalable data mining algorithms for realworld big data applications.

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

1. Analyze different perspectives and frameworks of software quality (including ISO-9126) and their role in identifying, measuring, and improving correctness and defect properties in software systems.
2. Evaluate techniques for defect prevention, reduction, and containment within software quality assurance to improve fault tolerance, safety, and risk management.
3. Formulate quality planning strategies, assess software processes, and propose improvements through quality engineering practices.
4. Develop and manage effective test strategies by planning, executing, measuring, and automating test activities to ensure software quality and reliability.
5. Apply coverage and usage-based testing techniques, including operational profile construction, to evaluate and improve the reliability of software systems through case studies.

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)