

OBJECTIVES:

The Student should be made to:

- Be exposed to big data
- Learn the different ways of Data Analysis
- Be familiar with data streams
- Learn the mining and clustering
- Be familiar with the visualization

UNIT I INTRODUCTION TO BIG DATA**8**

Introduction to Big Data Platform – Challenges of conventional systems - Web data – Evolution of Analytic scalability, analytic processes and tools, Analysis vs reporting - Modern data analytic tools, Stastical concepts: Sampling distributions, resampling, statistical inference, prediction error.

UNIT II DATA ANALYSIS**12**

Regression modeling, Multivariate analysis, Bayesian modeling, inference and Bayesian networks, Support vector and kernel methods, Analysis of time series: linear systems analysis, nonlinear dynamics - Rule induction - Neural networks: learning and generalization, competitive learning, principal component analysis and neural networks; Fuzzy logic: extracting fuzzy models from data, fuzzy decision trees, Stochastic search methods.

UNIT III MINING DATA STREAMS**8**

Introduction to Streams Concepts – Stream data model and architecture - Stream Computing, Sampling data in a stream – Filtering streams – Counting distinct elements in a stream – Estimating moments – Counting oneness in a window – Decaying window - Realtime Analytics Platform(RTAP) applications - case studies - real time sentiment analysis, stock market predictions.

UNIT IV FREQUENT ITEMSETS AND CLUSTERING**9**

Mining Frequent itemsets - Market based model – Apriori Algorithm – Handling large data sets in Main memory – Limited Pass algorithm – Counting frequent itemsets in a stream – Clustering Techniques – Hierarchical – K- Means – Clustering high dimensional data – CLIQUE and PROCLUS – Frequent pattern based clustering methods – Clustering in non-euclidean space – Clustering for streams and Parallelism.

UNIT V FRAMEWORKS AND VISUALIZATION**8**

MapReduce – Hadoop, Hive, MapR – Sharding – NoSQL Databases - S3 - Hadoop Distributed file systems – Visualizations - Visual data analysis techniques, interaction techniques; Systems and applications:

TOTAL: 45 PERIODS**OUTCOMES:**

The student should be made to:

- Apply the statistical analysis methods.
- Compare and contrast various soft computing frameworks.
- Design distributed file systems.
- Apply Stream data model.
- Use Visualisation techniques

TEXT BOOKS:

1. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer, 2007.
2. Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, Cambridge University Press, 2012.

REFERENCES:

1. Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with advanced analytics, John Wiley & sons, 2012.
2. Glenn J. Myatt, Making Sense of Data, John Wiley & Sons, 2007 Pete Warden, Big Data Glossary, O'Reilly, 2011.
3. Jiawei Han, Micheline Kamber "Data Mining Concepts and Techniques", Second Edition, Elsevier, Reprinted 2008.