BC304 BIG DATA ANALYTICS

Course description and Objectives: The main objectives of this course is to enable the students with basic data analytic skills like regression analysis, classification techniques, clustering techniques, association rule mining. Further, this course also enables the students how to scale the above algorithms with different data environments like massive amounts of data, streaming data, distributed data and provide hands on experience on real world problems using the above theoretical background.

Course Outcomes: · Necessary theory background to process analytics. · Processing analytics on small scale data. · Mining from massive datasets. · Mining from distributed datasets.

UNIT - I Introduction To Big Data : Introduction to Big Data Platform – Traits of Big data - Challenges of Conventional Systems - Web Data – Evolution Of Analytic Scalability - Analytic Processes and Tools - Analysis vs Reporting - Modern Data Analytic Tools - Statistical Concepts: Sampling Distributions - ReSampling - Statistical Inference - Prediction Error.

UNIT - II Data Analysis : 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 : 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 - Real time Analytics Platform(RTAP) Applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions.

UNIT - IV Frequent Item sets and Clustering: 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 NonEuclidean Space – Clustering for Streams and Parallelism.

UNIT - V Frameworks And Visualization : MapReduce – Hadoop, Hive, MapR – Sharding – NoSQL Databases - S3 - Hadoop Distributed File Systems – Visualizations - Visual Data Analysis Techniques - Interaction Techniques; Systems and Analytics Applications - Analytics using Statistical packagesApproaches to modeling in Analytics – correlation, regression, decision trees, classification, association-Intelligence from unstructured information-Text analytics-Understanding of emerging trends and technologies-Industry challenges and application of Analytics

TEXT BOOKS: 1. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.

2. AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2012.

3. Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley & sons, 2012.

REFERENCE BOOKS: 1. Glenn J. Myatt, “Making Sense of Data”, John Wiley & Sons, 2007 2. Pete Warden, “Big Data Glossary”, O’Reilly, 2011. 3. Jiawei Han, MichelineKamber “Data Mining Concepts and Techniques”, Second Edition, Elsevier, Reprinted 2008