

Big Data Analytics

Course Code: UCST00016T

Course Title: Big Data Analytics

Semester: II

Credits: 04

Rationale

The rapid rise of the internet and digital transactions led to an exponential growth in demand for data storage and analytics. Big Data analytics is a data analysis methodology enabled by recent advances in technologies supports high-velocity data capture, storage and analysis. Besides many benefits, the Big Data also has raised few security and privacy concerns. This course introduces the students to the basic concepts of Big Data concerning their storage and analysis.

Course Outlines

Contents	No of Lectures
<u>Unit-I</u> Big Data: Introduction, Need of Big Data, Classification of Data, Characteristic Features, Big Data Types, Big Data Handling Techniques, Classification Methods for data and Big Data, Designing Data Architecture Design, Managing Data for Analysis, Data Sources, Data Quality, Data Preprocessing, Big Data vs Traditional Data, Big Data Applications, Different Big Data platforms, Phases in Big Data Analytics.	10
<u>Unit-II</u> Introduction to NoSQL, NoSQL Data Store, Big Data NoSQL, Features in NoSQL Transactions, Big Data No SQL Solutions, Significance of NoSQL, CAP theorem, NoSQL Data Architecture Patterns, Graph Databases, Variations of NOSQL Architectural Patterns, Using NoSQL to mange Big Data, Types of Big Data Problems, Comparison of NoSQL with other Databases, Shared Nothing Architecture for Big Data Tasks, Mongo DB Database, Features of Mongo DB, CASSANDRA Databases.	10
<u>Unit-III</u> Introduction to Hadoop, Hadoop Core Components, Features of Hadoop, Hadoop Ecosystem Components, Hadoop Streaming, Hadoop Pipes, Design of Hadoop Distributed File System, HDFS Concepts, HDFS Commands, Hadoop Yarn, Hadoop 2 Execution Model, Map reduce Framework, Mapreduce Programming Model, Hadoop Ecosystem, Ambari, HBase, Hive, Pig, Mahout.	10
<u>Unit-IV</u> Statistical Methods: Regression Analysis, Simple Linear Regression, Multiple	10

Annexure – III

Regressions, Regression Modelling, K-Nearest Neighbour Regression Analysis, Frequent Item sets and Association Rule Mining, Apriori Algorithm, Classification Concepts, Supervised Learning, Unsupervised Learning, K-Nearest Neighbour Classifier, Decision Tree Algorithm, Naive Bayes Classifier, Support Vector Machine Classifier, Adaboost and other Ensemble Classifiers. Clustering: Types of Data in Cluster Analysis, K- means clustering, k-medoids, Hierarchical Clustering, Recommendation System, Models for Recommendation Systems.	
<u>Unit-V</u> Introduction to R language, Advantages of R over other programming languages, Basic R programming commands, Data type in R, Loading and Handling Data in R, Expressions, Variables, Functions, Logical Values, Manipulating Text in Data using R, Different R functions for Understanding Data in Data Frames, Simple Analysis using R, Different Methods for Reading Data, Exploring Dataset in R, Missing Value Treatment in R, Predictive Analytics using R: Linear Regression, Logistic Regression analysis using case studies.	10

Course Outcomes

Upon successful completion of this course, candidates will be able to:

- Identify Big Data and its Business Implications.
- Understand the benefits that Big Data can offer to organizations.
- Handle the large datasets using Big Data Storage platforms like Hadoop and NoSQL.
- Analyze the different types of datasets using R programming language.
- Apply Machine Learning Techniques using R.

Text Books

1. Raj Kamal and Preeti Saxena, "Big Data Analytics: Introduction to Hadoop, Spark and Machine Learning", Mc Graw Hill Education (India) Private Limited 2019.
2. Seema Acharya, "Data Analytics Using R", Mc Graw Hill Education (India) Private Limited 2018.

Reference Books

1. Michael Minelli, Michelle Chambers, and AmbigaDhiraj, "Big Data, Big Analytics: Emerging Business 2013.
2. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, 2012.
3. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
4. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.
5. L. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.