Syllabus

IT1110	**DATA SCIENCE AND BIG DATA ANALYTICS**	L	T	P	C
	Total contact hours – 60	2	0	2	3
	Prerequisite				
	Knowledge of Statistics and Probability, Java and XML is preferred				

PURPOSE

Today's world is data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. This course provides grounding in basic and advanced analytic methods and an introduction to big data analytics technology and tools, including MapReduce and Hadoop.

INSTRUCTIONAL OBJECTIVES

1	Learn about the basics of data Science and to understand the various supervised and unsupervised learning techniques.
2	Bringing together several key technologies used in manipulating, storing, and analyzing big data from different perspectives.
3	Understanding the Hadoop architecture and implementation of MapReduce Application.

UNIT I - INTRODUCTION TO DATA SCIENCE (6 hours)

Introduction of Data Science – Basic Data Analytics using R – R Graphical User Interfaces – Data Import and Export – Attribute and Data Types – Descriptive Statistics – Exploratory Data Analysis – Visualization Before Analysis – Dirty Data – Visualizing a Single Variable – Examining Multiple Variables – Data Exploration Versus Presentation.

UNIT II- ADVANCED ANALYTICAL THEORY AND METHODS (6 hours)

Overview of Clustering – K-means – Use Cases – Overview of the Method – Perform a K-means Analysis using R – Classification – Decision Trees – Overview of a Decision Tree – Decision Tree Algorithms – Evaluating a Decision Tree – Decision Tree in R – Bayes' Theorem – Naïve Bayes Classifier – Smoothing – Naïve Bayes in R.

UNIT III-BIG DATA FROM DIFFERENT PERSPECTIVES (6 hours)

Big data from business Perspective: Introduction of big data-Characteristics of big data-Data in the warehouse and data in Hadoop- Importance of Big data- Big data Use cases: Patterns for Big data deployment. Big data from Technology Perspective: History of Hadoop-Components of Hadoop-Application Development in Hadoop-Getting your data in Hadoop-other Hadoop Component.

UNIT IV- HADOOP DISTRIBUTED FILE SYSTEM ARCHITECTURE (6 hours)

HDFS Architecture – HDFS Concepts – Blocks – NameNode – Secondary NameNode – DataNode – HDFS Federation – Basic File System Operations – Data Flow – Anatomy of File Read – Anatomy of File Write.

UNIT V- PROCESSING YOUR DATA WITH MAPREDUCE (6 hours)

Getting to know MapReduce – MapReduce Execution Pipeline – Runtime Coordination and Task Management – MapReduce Application – Hadoop Word Count Implementation.

TEXT BOOKS

1. David Dietrich, Barry Heller and Beibei Yang, “Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data”, Wiley, ISBN 13:9788126556533, 2015.
2. Paul Zikopoulos, Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, “Understanding Big Data: Analytics for Enterprise Class Hadoop and streaming Data”, The McGraw-Hill Companies, ISBN : 978-0-07-179054-3, 2012.
3. Tom White, "Hadoop: The Definitive Guide", 4th Edition, O'Reilly, ISBN: 9789352130672, 2015.
4. Biris Lublinsky, Kevin T. Smith and Alexey Yakubovich, “Professional Hadoop Solutions”, Wiley, ISBN 13:9788126551071, 2015.

LIST OF EXPERIMENTS **(30 hours)**

1. Basic Data Analytic Methods using R
2. Preparing and training data based on K-means clustering analysis using R
3. Preparing and training data based on Decision Tree Classification analysis using R
4. Preparing and training data based on Naïve Bayes Classification analysis using R
5. Hadoop Distributed File System Commands
6. Hadoop Word Count Implementation using MapReduce
7. Implementation of Matrix Multiplication using MapReduce