राष्ट्रीय प्रौद्योगिकी संस्थान पटना / NATIONAL INSTITUE OF TECHNOLOGY PATNA
संगणक विज्ञान एंव अभियांत्रिकी विभाग / DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
अशोक राजपथ, पटना-८००००५, बिहार / ASHOK RAJPATH, PATNA-800005, BIHAR
Phone No.: 0612-2372715, 2370419, 2370843, 2371929 Ext- 200, 202 Fax-0612-2670631 Website: www.nitp.ac.in

No:- Date:

CSXX2821:***Big Data Analytics***
L-T-P-Cr: 2-0-2-3

Course Objectives:

- To understand the competitive advantages of big data analytics
- To understand the big data frameworks
- To learn data analysis methods
- To learn stream computing
- To gain knowledge on Hadoop related tools such as HBase, Cassandra, Pig, and Hive for big data analytic

Course Outcomes:

At the end of this course, the students will be able to:

1. Understand how to leverage the insights from big data analytics
2. Analyze data by utilizing various statistical and data mining approaches
3. Perform analytics on real-time streaming data
4. Understand the various NoSql alternative database models

Sl. No	Course Outcome (CO)	Mapping to PO
1	Understand how to leverage the insights from big data analytics	PO1, PO2
2	Analyze data by utilizing various statistical and data mining approaches	PO1, PO2, PO3, PO6, PO8
3	Perform analytics on real-time streaming data	PO1, PO3, PO6, PO7, PO8
4	Understand the various NoSql alternative database models	PO1, PO2, PO3, PO6, PO8

UNIT I INTRODUCTION TO BIG DATA

Big Data – Definition, Characteristic Features – Big Data Applications - Big Data vs Traditional Data - Risks of Big Data - Structure of Big Data - Challenges of Conventional Systems - Web Data – Evolution of Analytic Scalability - Evolution of Analytic Processes, Tools and methods - Analysis vs Reporting - Modern Data Analytic Tools.

UNIT II BIG DATA FRAMEWORK

Distributed File Systems - Large-Scale FileSystem Organization – HDFS concepts - MapReduce Execution, Algorithms using MapReduce, Matrix-Vector Multiplication – Hadoop YARN, Spark.

UNIT III DATA ANALYSIS

Statistical Methods:Regression modeling, Multivariate Analysis - Classification: SVM & Kernel Methods - Rule Mining - Cluster Analysis, Types of Data in Cluster Analysis, Partitioning Methods,Hierarchical Methods, Density Based Methods, Grid Based Methods, Model Based Clustering Methods, Clustering High Dimensional Data - Predictive Analytics.

UNIT IV MINING DATA STREAMS

Streams: Concepts – Stream Data Model and Architecture - Sampling data in a stream - Mining Data Streams and Mining Time-series data - Real Time Analytics Platform (RTAP) Applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions.

UNIT V BIG DATA FRAMEWORKS

Introduction to NoSQL – Aggregate Data Models – Hbase: Data Model and Implementations – Hbase Clients – Examples – .Cassandra: Data Model – Examples – Cassandra Clients – Hadoop Integration. Pig – Grunt – Pig Data Model – Pig Latin – developing and testing Pig Latin scripts. Hive – Data Types and File Formats – HiveQL Data Definition – HiveQL Data Manipulation – HiveQL Queries.

REFERENCES:

1. Bill Franks, ―Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics‖, Wiley and SAS Business Series, 2012.
2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", 2013.
3. Michael Berthold, David J. Hand, ―Intelligent Data Analysis‖, Springer, Second Edition, 2007.

4. Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
5. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.
6. Richard Cotton, "Learning R – A Step-by-step Function Guide to Data Analysis, , O'Reilly Media, 2013.