MSDP 201: Big Data Analytics   **Total Hours: 30**

Learning Objectives: The Learning Objectives of this course are as follows:

- To Understand the Big Data Platform and its Uses
- Provide an overview of Apache Hadoop
- Provide HDFS Concepts and Interfacing with HDFS.
- Provide hands on Hadoop Eco System
- To understand spark framework

Learning outcomes: The Learning Outcomes of this course are as follows:

- After studying this course, students will be able to identify Big Data and its Business Implications.
- After studying this course, students will be able to list the components of Hadoop and Hadoop Eco-System.
- After studying this course, students will be able to access and process data on distributed file system.
- After studying this course, students will be able to manage job execution in Hadoop environment.
- After studying this course, students will be able to develop Big Data Solutions using Hadoop Eco System.

SYLLABUS

Unit I: Fundamentals of Big Data Analysis: **7 Hours**

Data Storage and Analysis, Characteristics of Big Data, Big Data Analytics, Typical Analytical Architecture, Requirement for new analytical architecture, Challenges in Big Data Analytics - Need of big data frameworks.

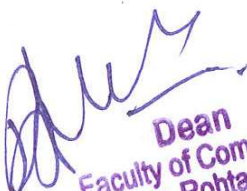**Unit II: Hadoop Framework:**   **7 Hours**

Hadoop, Requirement of Hadoop Framework, Design principle of Hadoop -Comparison with other system, Hadoop Components Hadoop 1 vs Hadoop 2, Hadoop Daemon's – HDFS Commands, Map Reduce Programming: I/O formats, Map side join, Reduce Side Join, Secondary sorting, Pipelining Map Reduce jobs

Unit III: HDFS (Hadoop Distributed File System):   **8 Hours**

The Design of HDFS, HDFS Concepts, Command Line Interface, Hadoop file system interfaces, Data flow, Data Ingest with Flume and Scoop and Hadoop archives, Hadoop I/O: Compression, Serialization, Avro and File-Based Data structures.

Unit IV: Spark Framework and Data Analysis with Spark Shell: **8 Hours**

Introduction to GPU Computing, CUDA Programming Model, CUDA APL, Simple Matrix, Multiplication in CUDA, CUDA Memory Model, Shared Memory Matrix Multiplication, Additional CUDA API Features. Writing Spark Application - Spark Programming in Scala, Python, R, Java - Application Execution.

Text & References:

- Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.
- Mike Frampton, "Mastering Apache Spark", Packt Publishing, 2015.
- Tom White, "Hadoop: The DefinitiveGuide", O'Reilly, 4th Edition, 2015.
- Nick Pentreath, Machine Learning with Spark, Packt Publishing, 2015.
- Mohammed Guller, Big Data Analytics with Spark, Apress,2015.
- Donald Miner, Adam Shook, "Map Reduce Design Pattern", O'Reilly, 2012

Dean
Faculty of Commerce
Jamuhar, Rohtas (Bihar)

Coordinator
IQAC (NAAC)
Gopal Narayan Singh University
Jamuhar, Sasaram, Rohtas (Bihar)