# Big Data Analytics - II

**Credit Distribution, Eligibility and Pre-Requisites of the Course**

| Course title & Code | Credits | Credit distribution of the course | | | Eligibility criteria | Pre-requisite of the course (if any) |
|---|---|---|---|---|---|---|
| | | Lecture | Tutorial | Practical/ Practice | | |
| **Big Data Analytics-II** | **2** | **1** | **0** | **1** | **Class XII** | **Big Data Analytics-I** |

## Learning Objectives

The Learning Objectives of this course are as follows:

- Provide hands-on Hadoop Eco System.
- Provide an overview of Apache Spark.
- To understand Machine Learning with Big Data.
- Provide an overview of GPU Computing.

## Learning Outcomes

The Learning Outcomes of this course are as follows:

- After studying this course, students will be able to develop Big Data Solutions using Hadoop Eco System.
- After studying this course, students will be able to use Spark for the effective analysis of the Big Data.
- After studying this course, students will be able to use Machine Learning Techniques with Big Data.
- After studying this course, students will be able to use GPU Computing for parallel computations.

## Syllabus:

**Unit-1: Big Data Systems- The Advances (3 hours)**

Data flow, Data Ingest with Flume and Scoop and Hadoop archives, Hadoop I/O: Compression, Serialization, Avro and File-Based Data structures.

**Unit-2: Introduction to Apache Spark (3 hours)**

Introduction, Architecture of Spark, Resilient Distributed Datasets, Spark Transformations, Writing Spark Application - Spark Programming in Scala, Python, R, Java - Application Execution

**Unit-3: Machine Learning with Big Data (3 hours)**

Introduction to machine learning, Supervised vs Unsupervised learning, Cluster analysis, understanding k means clustering, Implementation of k means clustering with Map Reduce.

**Unit-4: Introduction to GPU Computing (6 hours)**

Introduction to GPU Computing, CUDA Programming Model, CUDA API, Simple Matrix, Multiplication in CUDA, CUDA Memory Model, Shared Memory Matrix Multiplication, Additional CUDA API Features.

**Practical Exercises (30 hours)**

- Implementation of Matrix Multiplication with Hadoop Map Reduce.
- Implementation of K-means clustering using Map Reduce.
- To study and implement basic functions and commands in R/Python programming.
- To build Word cloud, a text mining method using R/Python for easy to understand and visualization than a table data.
- To implement clustering program using R/Python programming

**Essential/recommended readings**


- Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.
- Mike Frampton, "Mastering Apache Spark", Packt Publishing, 2015.
- Tom White, "Hadoop: The Defective Guide", 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.


**Examination scheme and mode:**

Evaluation scheme and mode will be as per the guidelines notified by the University of Delhi