Name of the Programme: MSc Integrated
Course Code: IMC- 602
Title of the Course: Big Data Frameworks
Number of Credits: 6(4L-0T-2P)
Effective from AY: 2022-23

Prerequisites for the course	Probability and Statistics; Python Programming	
Objectives	• To understand the need of Big Data, challenges and different analytical architectures • Installation and understanding of Hadoop Architecture and its ecosystems • Processing of Big Data with Advanced architectures like Spark. • Describe graphs and streaming data in Spark	
Content Theory:	**Introduction to Big Data:** 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	9 hours
	Hadoop framework: 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 MapReduce jobs -	7 hours
	Hadoop Ecosystem : Introduction to Hadoop ecosystem technologies: Serialization: AVRO, Co-ordination: Zookeeper, Databases: HBase, Hive, Scripting language: Pig, Streaming: Flink, Storm	7 hours
	Spark framework: 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.	7 hours
	Data analysis with spark shell: Writing Spark Application - Spark Programming in Scala, Python, R, Java - Application Execution	6 hours
	Spark SQL and Graph X : SQL Context – Importing and Saving data – Data frames – using SQL – GraphX overview – Creating Graph – Graph Algorithms.	6 hours
	Spark Streaming: Overview – Errors and Recovery – Streaming Source – Streaming live data with spark	6 hours

Content Practical:	**Suggested Lab Assignments:** 1. Downloading and installing Hadoop; Understanding different Hadoop modes. Startup scripts, Configuration files. 2. Hadoop Implementation of file management tasks, such as Adding files and directories, Retrieving files and Deleting files 3. Implement of Matrix Multiplication with Hadoop Map Reduce 4. Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm. 5. Implementation of K-means clustering using Map Reduce 6. Installation of Hive along with practice examples. 7. Installation of HBase, Installing thrift along with Practice examples 8. Patrice importing and exporting data from various databases .	8 * 6 = 48 hours
Pedagogy	Assignment / Quiz / Project / Seminar	

References/ Readings	1. Mike Frampton, “Mastering Apache Spark”, Packt Publishing, 2015. 2. Tom White, “Hadoop: The Definitive Guide”, O’Reilly, 4thEdition, 2015. 3. Nick Pentreath, Machine Learning with Spark, Packt Publishing, 2015. 4. Mohammed Guller, Big Data Analytics with Spark, Apress, 2015. 5. Donald Miner, Adam Shook, “Map Reduce Design Pattern”, O’Reilly, 2012.
Course Outcomes	1. Understand big data fundamentals. 2. Learn big data technologies (e.g., Hadoop, Spark). 3. Analyze and process large datasets using distributed computing. 4. Apply big data analytics techniques for valuable insights.