Subject Code:	Big Data Frameworks	L,T,P,J,C 2,0,2,4,4
Objective	The course objective is to impart an understanding of the challenges in storing and processing big data and how to use different big data frameworks effectively to store and process big data.	
Expected Outcomes	After successfully completing the course the student should be able to a) Discuss the challenges in Big Data. b) Describe the need of different big data frameworks. c) Write MapReduce programming in both Hadoop and Spark Framework. d) Write programs in Spark Streaming, SPARK SQL and GraphX	
SLO's	2,5,7,9,14,17	

Module	Topics	LHrs	SLO
1	INTRODUCTION TO BIG DATA3 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		2
2	Hadoop Framework Hadoop – Requirement of Hadoop Framework - Design principle of Hadoop –Comparison with other system - Hadoop Components – 6 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		5,7,9
3	Hadoop Ecosystem Introduction to Hadoop ecosystem technologies: Serialization: AVRO, Co-ordination: Zookeeper, Databases: HBase, Hive, Scripting language: Pig, Streaming: Flink, Storm	3	5,7,9
4	 Framework Overview of Spark – Hadoop vs Spark – Cluster Design – Cluster Management – performance,Application Programming interface (API): Spark Context, Resilient Distributed Datasets, Creating RDD, RDD Operations, Saving RDD - Lazy Operation – Spark Jobs	5Spark	5,7,9

5	4Interactive Data Analysis with Spark Shell Writing Spark Application - Spark Programming in Scala, Python, R, Java - Application Execution		5,7,9
6	Spark SQL and GraphX SQL Context – Importing and Saving data – Data frames – using SQL – GraphX overview – Creating Graph – Graph Algorithms	5	5,7,9
7	Spark Streaming Overview – Errors and Recovery – Streaming Source – Streaming live data with spark	3	5,7,9
8	Recent Trends in Big Data Analytics Framework	1	2
	Lab(IndicativeListofExperiments(inthe areasof) 1. HDFS Commends 2. MapReduce Program to show the need of Combiner 3. MapReduce I/O Formats –Text, key- value 4. MapReduce I/O Formats - NLine, Multiline 5. Sequence file Input / Output Formats 6. Secondary sorting 7. Distributed Cache & Map Side Join, Reduce side Join 8. Building and Running a Spark Application 9. Wordcount in Hadoop and Spark 10. Manipulating RDD 11. Inverted Indexing in Spark 12. Sequence alignment problem in Spark 13. Implementation of Matrix algorithms in Spark 14. Spark Sql programming 15. Building Spark Streaming application	30	14

Project# Generally ateamproject[5to10 members] # Concepts studied inXXXX should have beenused # Down to earthapplication and innovative idea should have beenattempted # ReportinDigital formatwithall drawings usingsoftware package tobesubmitted. # Assessment ona continuous basis with a min of3 reviews. Projects may be given as group projects The following is the sample project that can be given to students to be implemented: 1. Predicting forest cover 2. Anomaly detection 3. Text Analytics 4. Co-occurrence of terms in social networks using GraphX 5. HITS algorithm 6. Geospatial and Temporal data analytics	60[Non Contact hrs]	17

ReferenceBooks

1. Mike Frampton, “Mastering Apache Spark”, Packt Publishing, 2015.
2. Tom White, “Hadoop: The Definitive Guide”, O’Reilly, 4th Edition, 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, “MapReduce Design Pattern”, O’Reilly, 2012

Big Data Frameworks

Knowledge Areas that contain topics and learning outcomes covered in the course

Knowledge Area	Total Hours of Coverage
CS: AL(Algorithms and Complexity) / CE: CAO	3
CS: PL(Programming Languages) / CE: CAO	24
CS: DS / CE: DSC	3

Body of Knowledge coverage
[List the Knowledge Units covered in whole or in part in the course. If in part, please indicate which topics and/or learning outcomes are covered. For those not covered, you might want to indicate whether they are covered in another course or not covered in your curriculum at all. This section will likely be the most time-consuming to complete, but is the most valuable for educators planning to adopt the CS2013 guidelines.]

KA	Knowledge Unit	Topics Covered	Hours
CE: AR	Memory System Organization and Architecture	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	3
CE: PD CS: PL / CE: PRFLanguage	Parallel algorithms, Analysis and Programming Pragmatics	Hadoop – Requirement of Hadoop Framework - 12 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 SQL Context – Importing and Saving data – Data frames Writing Spark Application - Spark Programming in Scala, Python, R, Java - Application Execution	
CS: PL / CE: PRFProgramming	Advanced Constructs	Map side join, Reduce Side Join, Secondary sorting, Pipelining MapReduce jobs Overview – Errors and Recovery – Streaming Source – Streaming live data with spark Introduction to Hadoop ecosystem technologies: Serialization: AVRO, Co-ordination: Zookeeper, Databases: HBase, Hive, Scripting language: Pig, Streaming: Flink, Storm	12

CE/NC CS: DS / CE: DSC	Social networking Graphs and Trees	using SQL – GraphX overview – Creating Graph – 3 Graph Algorithms	
		Total Hours	30

Where does the course fit in the curriculum?
[In what year do students commonly take the course? Is it compulsory? Does it have pre-requisites, required following courses? How many students take it?]

This course is a

- Core subject
- Suitable from 1st semester onwards.
- Knowledge of Java programming language is essential.

What is covered in the course?
[A short description, and/or a concise list of topics - possibly from your course syllabus.(This is likely to be your longest answer)]

Part 1: Introduction to Big Data
This part of the course gives introduction to the basics of bigdata, characteristics of big data, challenges involved and the need for bigdata frameworks

Part II: Hadoop Framework
Describes the Hadoop Architecture and compares it with legacy distributed computing. This part of the course also introduces data storage in Hadoop and writing MapReduce code. The essential ecosystems of Hadoop are introduced in this part.
Part III: Spark and Streaming
This part of the course, introduces Spark tool, Graph algorithms and streaming. Spark will lead to interactive data analysis and supports streaming.

What is the format of the course?
[Is it face to face, online or blended? How many contact hours? Does it have lectures, lab sessions, discussion classes?]

This Course is designed with 150 minutes of in-classroom sessions per week, as well as 200 minutes of non-contact time spent on implementing course related project. Generally this course has the combination of lectures, in-class discussion, case studies, guest-lectures, mandatory off-class reading material, quizzes.

How are students assessed?
[What type, and number, of assignments are students are expected to do? (papers, problem sets, programming projects, etc.). How long do you expect students to spend on completing assessed work?]

- Students are assessed on a combination group activities, classroom discussion, projects, & continuous and final assessment tests.
- Additional weightage will be given based on their rank in crowd sourced projects.
- Students can earn additional weightage based on certificate of completion of a related MOOC course or any online course completion.

Session wise plan

S.No	Topic Covered	Class Hour	Levels of mastery	Reference Book
1	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	3	Familiarity	R2
2	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	6	Usage	R2, R5
3	Introduction to Hadoop ecosystem technologies: Serialization: AVRO, Co-ordination: Zookeeper, Databases: HBase, Hive, Scripting language: Pig, Streaming: Flink, Storm	3	Familiarity	R2, R3
4	Overview of Spark – Hadoop vs Spark – Cluster Design – Cluster Management – performance,Application Programming interface (API): Spark Context, Resilient Distributed Datasets, Creating RDD, RDD Operations, Saving RDD - Lazy Operation – Spark Jobs Writing Spark Application - Spark Programming in Scala,	5	Familiarity	R1, R3, R4
5	Python, R, Java - Application Execution SQL Context – Importing and Saving data – Data frames –	4	Usage	R1, R3, R4
6		5	Usage	R1, R3, R4

	using SQL – GraphX overview – Creating Graph – Graph Algorithms			
7	Overview – Errors and Recovery – Streaming Source – Streaming live data with spark	3	Usage	R1, R3, R4
8	Recent Trends	1		
Total hours		30		