Course code	**LARGE SCALE DATA PROCESSING**	**L**	**T**	**P**	**J**	**C**
CSE3025		**2**	**0**	**2**	**4**	**4**
Pre-requisite	**-**					**Syllabus version**
						v. xx.xx

Course Objectives:

Objectives of this course is to

- Understand different characteristics of big data
- Understand the requirement of big data frameworks
- Learn the concepts of distributed file system
- Provide MapReduce programming environment
- Understand need of inverted indexing and graph data analytics

Expected Course Outcome:

After successfully completing the course the student should be able to

(1) Define the characteristics of big data and explain the data science life cycle.
(2) Differentiate between conventional and contemporary distributed framework.
(3) Characterize storage and processing of large data.
(4) Implement and demonstrate the use of the hadoop eco- system.
(5) Compare scalable frameworks for large data.
(6) Identify independent tasks in a program that may be parallelized.
(7) Decompose a problem into map and reduce operations for implementation.
(8) Recognize different input output formats for map reduce programs.
(9) Design programs to analyze large scale text data.
(10) Identify problems suitable for use of graph mining in large data processing.

Student Learning Outcomes (SLO):	**2,11,17**

Module:1	**INTRODUCTION TO BIG DATA AND ANALYTICS**	**4 hours**	**SLO:2**

Big Data Overview – Characteristics of Big Data – Business Intelligence vs Data Analytics

Module:2	**NEED OF DATA ANALYTICS**	**4 hours**	**SLO: 11**

Data Analytics Life Cycle – Data Analytics in Industries Exploring Big data – Challenges in handling Big Data

Module:3	**Big Data Tools**	**4 hours**	**SLO: 17**

Need of Big data tools - understanding distributed systems - Overview of Hadoop – comparing SQL databases and Hadoop – Hadoop Eco System - Distributed File System: HDFS, – Design of HDFS – writing files to HDFS – Reading files from HDFS

Module:4	**Hadoop Architecture**	**6 hours**	**SLO: 11**

Hadoop Daemons - Hadoop Cluster Architecture– YARN–Advantages of YARN

Module:5	**Introduction to MapReduce**	**6 hours**	**SLO: 11**

Developing MapReduce Program – Anatomy of Map Reduce Code - Simple Map Reduce Program - counting things – Map Phase – shuffle and sort - Reduce Phase – Master slave architecture –Job Processing in hadoop – Map Reduce Pipelining

Module:6	**Map Reduce Programming Concepts**	**3 hours**	**SLO: 17**

Use of Combiner - Block vs Split Size - working with Input and output format – Key,Text, Sequence, NLine file format, XML file format.			
Module:7	**Inverted Indexing and Graph Analytics**	**3 hours**	**SLO: 17**
Web crawling – inverted index – Baseline and revised implementation - Graph Representation – Parallel Breadth first search – page rank – issues with graph processing.			
	Total Lecture hours:	**30 hours**	

Text Book(s)	
1.	Tom White, Hadoop The Definitive Guide, O'Reilly, 4th Edition, 2015
Reference Books	
1. 2. 3.	Alex Holmes, Hadoop in Practice, Manning Shelter Island, 2012 Chuck Lam, Hadoop in Action. Manning Shelter Island, 2011 Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with Map Reduce, 2010
Mode of Evaluation:	

List of Challenging Experiments (Indicative)		**SLO: 14,17**
1.	Setting up Hadoop in Single node / Multinode environment	X hours
2.	Command line interface with HDFS	X hours
3.	Counting things using MapReduce	X hours
4.	Map Reduce Program to show the need of Combiner	X hours
5.	Map Reduce I/O Formats – key- value, Text	X hours
6.	Map Reduce I/O Formats – N line	
7.	Multiline I/O	
8.	Parallel Breadth First Search	
9.	Sequence file Input / Output Formats	
10.	Baseline Inverted Indexing using Map Reduce	
11.	Revised Inverted Indexing using Map Reduce	
12	Matrix Factorization using Map Reduce	
13	Video Processing using Map Reduce	
14	BioInformatics (Protein/Gene Sequence etc) processing with MapReduce	
	Total Laboratory Hours	X hours

Project:

# Generally a team project [5 to 10 members]
# Concepts studied in XXXX should have been used
# Down to earth application and innovative idea should have been attempted
#Report in Digital format with all drawings using software package to be submitted. [Ex. 1. Design of a traffic light system using sequential circuits OR 2. Design of digital clock]
#Assessment on a continuous basis with a min of 3 reviews.

Projects may be given as group projects

The following is the sample project that can be given to students to be implemented in the Hadoop environment using appropriate tools.

1. Implementing association rule mining
2. Implementing closed item set mining
3. Implementing maximal item set mining

4. Solving sequence alignment problem – (Bio informatics) 5. Solving Data Science problems from Kaggle website			
Mode of evaluation:			
Recommended by Board of Studies	DD-MM-YYYY		
Approved by Academic Council	No. xx	Date	DD-MM-YYYY

<u>CO-PO MAPPING:</u>

	PO 2	PO 3	PO 8	PO 9	PO 10	PO 11	PO 12	PO 13	PO 15	PO 16	PO 18	PO 20
CO1												
CO2												
CO3												
CO4												
CO5												
CO6												
CO7												

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

[Kindly refer ACM Computer Science Recommendation (CS 2013)- and ACM Computer Engineering Recommendation CE 2004.]

Knowledge Area	**Total Hours of Coverage [Theory+Practical]**
CS: IM(Information Management)	4+ 8
CS: PD(Parallel and Distributed Computing)	20 +4
CS: SF(System Fundamental)	6 + 18
Total	60 Hours [30 + 30]

2.1 Body of Knowledge coverage

[List the Knowledge Units covered in whole in the course. This section will likely be the most time-consuming to complete, but is the most valuable for educators planning to adopt the CS2013/ CE 2004 guidelines.]

KA	**Knowledge Unit**	**Topics Covered**	**Hours**
CS: IM	IM/Indexing	Web crawling inverted index Baseline and revised implementation page rank	4
CS: PD	Parallel Algorithms, Analysis, and Programming,	Move the computation , Parallel Graphs, MapReduce	4
CS: PD	Parallel Decomposition,	Distributed File System Cluster Architecture Independence and partitioning, Data and task decomposition,	2
CS: SF	SF / Parallelism,	Task parallelism, MapReduce,	4

3. 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

- Elective Course.
- Suitable from 5th semester onwards.
- Knowledge of any one programming language is essential.

4. 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)]

4.1 Part 1: Introduction to Big Data

It introduces what is big data and its life cycle, challenges in big data analytics and handling large scale data.

4.2 Part II: Big data tools and architecture

This section covers the need of big data tools, hand-on exposure to store and process the data using Hadoop Distributed File System and MapReduce programming respectively. Hadoop daemons and YARN architecture is discussed.

4.3 Part III: Inverted Indexing and Graph Analytics

This section deals with storing and processing text data, introduces graph algorithms for analytics, discusses pagerank algorithms as a case study.

5. 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 100 minutes of in-classroom sessions per week, 60 minutes of video/reading instructional material per week, 100 minutes of lab hours per week, as well as 200 minutes of non-contact time spent on implementing course related project. Generally this course have the combination of lectures, in-class discussion, case studies, guest-lectures, mandatory off-class reading material, quizzes.

6. 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, and continuous, final assessment tests.

- Additional weightage will be given based on their rank in crowd sourced projects/ Kaggle like competitions.

- Students can earn additional weightage based on certificate of completion of a related MOOC course.

7. Session wise plan

Student Outcomes Covered: 2, 11, 17

Sl. No	Topic Covered	Class Hour	Lab Hour	levels of mastery	Reference Book	Remarks
1	Big Data Overview – Characteristics of Big Data –	2		Usage	1	
2	Business Intelligence vs Data Analytics	2		Usage	1,	
3	Need of Data Analytics – Data Analytics Life Cycle – Data Analytics in Industries –Exploring Big data –	2		Usage	1	
4	Challenges in handling Big Data	2		Usage	1	
5	Need of Big data tools - understanding distributed systems –	2		Familiarity	1	
6	Overview of Hadoop – comparing SQL databases and Hadoop – Hadoop Eco System	2		Usage	1	
7	Distributed File System:		4	Familiarity	1	LAB

	HDFS, – Design of HDFS – writing files to HDFS – Reading files from HDFS					Component
8	Hadoop Daemons - Hadoop Cluster Architecture – YARN – Advantages of YARN –	3		Usage	1	
9	Developing MapReduce Program – Anatomy of MR Code - Simple Map Reduce Program - counting things	3	4	Usage	1	LAB Component
10	Map Phase – shuffle and sort - Reduce Phase – Master slave architecture – Job Processing in hadoop – Map Reduce Pipelining	3	2	Usage	1	LAB Component
11	MapReduce Programming Concepts– Use of Combiner - Block vs Split Size	2	4	Usage	1	LAB Component
12	working with Input and output format – Key,Text,		4		1,2,3	LAB Component

13	Sequence, NLine file format, XML file format		4	Assessment	1,2,3	LAB Component
14	Web crawling – inverted index – Baseline and revised implementation	2	4	Usage	4	LAB Component
15	- Graph Representation – Parallel Breadth first search –	3		Usage	4	
16	page rank – issues with graph processing	2	4	Usage	4	LAB Component
Total hours covered		30 Hours (2 Credit hours /week ✍ 15 Weeks schedule)	30 Hours (2 Credit hours / week)			