Lesson Plan

B.E. (CE-A) (Semester VII)

Subject: Big Data Analytics (BDA-CSC702)

Subject code: BDA-CSC702

Teacher-in-charge: Prof. Ankita Amburle

Academic Term: July – October 2022

Module		Content	Hrs
1		**Introduction to Big Data and Hadoop**	**2**
	1.1	Introduction to Big Data - Big Data characteristics and Types of Big Data Traditional vs. Big Data business approach	
	1.2	Case Study of Big Data Solutions Concept of Hadoop, Core Hadoop Components; Hadoop Ecosystem.	
2		**Hadoop HDFS and MapReduce**	**8**
	2.1	Distributed File Systems: Physical Organization of Compute Nodes, Large-Scale File-SystemOrganization. MapReduce: The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details ofMapReduce Execution, Coping With Node Failures.	
	2.2	Algorithms Using MapReduce: Matrix-Vector Multiplication by MapReduce, Relational- Algebra Operations, Computing Selections by MapReduce, Computing Projections by MapReduce, Union, Intersection, and Difference by MapReduce Hadoop Limitations	
3		**NOSQL**	**10**
	3.1	Introduction to NoSQL, NoSQL Business Drivers, NoSQL Data Architecture Patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns, NoSQL Case Study	
	3.2	NoSQL solution for big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer- to-peer; NoSQL systems to handle big data problems.	

4		**Mining Data Streams:**	11
	4.1	The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing. Sampling Data techniques in a Stream Filtering Streams: Bloom Filter with Analysis.	
	4.2	Counting Distinct Elements in a Stream, Count-Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements Counting Frequent Items in a Stream, Sampling Methods for Streams, Frequent Itemsets inDecaying Windows. Counting Ones in a Window: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.	
5		**Finding Similar Items and Clustering**	**4**
	5.1	A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering Case Study: Product Recommendation	
	5.2	Social Networks as Graphs, Clustering of Social- Network Graphs, Direct Discovery ofCommunities in a social graph	
6		**Real-Time Big Data Models**	**4**
	6.1	Exploring Basic features of R, Exploring RGUI, Exploring RStudio, Handling Basic Expressions in R, Variables in R, working with Vectors, Storing and Calculating Values in R, Creating and using Objects, interacting with users, Handling data in R workspace, Executing Scripts, Creating Plots, Accessing help and documentation in R.	
	6.2	Reading datasets and Exporting data from R, Manipulating and Processing Data in R, Using functions instead of script, built-in functions in R.Data Visualization: Types, Applications.	

Course Objectives:

1. To provide an overview of the big data platforms, its use cases and Hadoop ecosystem.
2. To introduce programming skills to build simple solutions using big data technologies such asMapReduce, Scripting for No SQL and R.
3. To learn the fundamental techniques and principles in achieving big data analytics withscalability and streaming capability.
4. To enable students to have skills that will help them to solve complex real-world problems for decisionsupport.

Course Outcomes:

Upon completion of this course students will be able to:

CSC702.1: Understand the building blocks of Big Data Analytics.
CSC702.2: Apply fundamental enabling techniques like Hadoop and
                 MapReduce insolving real world problems
CSC702.3: Understand different NoSQL systems and how it handles big data.
CSC702.4: Apply advanced techniques for emerging applications like
                 stream analytics.
CSC702.5: Achieve adequate perspectives of big data analytics in various
                 applications likerecommender systems, social media
                 applications, etc.
CSC702.6: Apply statistical computing techniques and graphics for
            analyzing big data.

CO-PO-PSO Mapping:

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CSC702.1	3												3	
CSC702.2	3	3	3	1	3				2				3	3
CSC702.3	3	3	3	2	3				2				3	3
CSC702.4	3	3	3	2	3				2				3	3
CSC702.5	3	3	3	2	3	3			2			2	3	3
CSC702.6	3	3	3	2	3	3			3	3	2	3	3	3

CO Assessment Tools:

Course Outcomes	Indirect Method (20%)							
	Unit Tests		Assignments		Quizzes		End Sem Exam	Course exit survey
	1	2	1	2	1	2		
CSDC7022.1	20%	--	20%	--	10%	--	50%	100%
CSDC7022.2	20%	--	20%	--	10%	--	50%	100%
CSDC7022.3	--	25%	--	25%	10%	--	50%	100%
CSDC7022.4	--	20%	--	20%	--	10%	50%	100%
CSDC7022.5	--	20%	--	20%	--	10%	50%	100%
CSDC7022.6	--	20%	--	20%	--	10%	50%	100%

CO calculation= (0.8 *Direct method +

0.2*Indirect method) Rubrics for

assessing Course Outcome with each

assessment tool:

Assignment:

Indicator				
Timeline (2)	More than two dayslate (0)	Two days late (1)	One day late (2)	On time (3)
Correctness (4)	All questions correct (4)	One point deducted for each incorrect answer		
Completion (4)	All questions answered (4)	One point will be deducted for each incomplete or un-attempted question		

Curriculum Gap identified: (with action plan)

1. **Nil**

Content beyond syllabus:

1. Link Analysis (Extra Session)

Sr.No.	Content Beyond Syllabus	Action Plan	PO Mapping
1	Link analysis	Planned one lecture.	PO2, PSO2

Modes of content delivery

Modes of Delivery	Brief description of content delivered
Class room lecture	1. Introduction to Big Data and Hadoop 2. Hadoop HDFS and MapReduce 3. NOSQL 4. Mining Data Streams 5. Finding Similar Items and Clustering 6. Real-Time Big Data Models
Assignments	1. **Assignment 1: based on** 1. Introduction to Big Data and Hadoop 2. NOSQL Assignment 2. **based on remaining modules**
Quizzes	Quiz 1: on 1. Introduction to Big Data and Hadoop 2. Hadoop HDFS and MapReduce 3. NOSQL Quiz 2: on 4. Mining Data Streams: 5.Finding Similar Items and Clustering 6.Real-Time Big Data Models

Text Books:
1. Cre Anand Rajaraman and Jeff Ullman ―Mining of Massive Datasets‖, Cambridge UniversityPress
2. Alex Holmes ―Hadoop in Practice‖, Manning Press, Dreamtech Press.
3. Dan Mcary and Ann Kelly ―Making Sense of NoSQL‖ – A guide for managers and the rest of us,Manning Press.
4. DT Editorial Services, ―Big Data Black Book‖, Dreamtech Press
5. EMC Education Services,‖Data Science and Big Data Analytics‖,Wiley

References books:
1. Bill Franks , ―Taming The Big Data Tidal Wave: Finding Opportunities In Huge Data StreamsWith Advanced Analytics, Wiley
2. Chuck Lam, ―Hadoop in Action‖, Dreamtech Press
3. Jared Dean, ―Big Data, Data Mining, and Machine Learning: Value Creation for

Business Leadersand Practitioners, Wiley India Private Limited, 2014.

4. Jiawei Han and Micheline Kamber, —Data Mining: Concepts and Techniques, Morgan KaufmannPublishers, 3rd ed, 2010.

5. Lior Rokach and Oded Maimon, —Data Mining and Knowledge Discovery Handbook, Springer,2nd edition, 2010.

6. Ronen Feldman and James Sanger, —The Text Mining Handbook: Advanced Approaches inAnalyzing Unstructured Data, Cambridge University Press, 2006.

7. Vojislav Kecman, —Learning and Soft Computing, MIT Press, 2010.

CLASS	BE Computer Engineering (A), Semester VII
Academic Term	July- October 2022
Subject	**Big Data Analytics (BDA-CSC702)**

Periods (Hours) per week	*Lecture*	**3**	
	Practical		
	Tutorial		

Evaluation System		*Hours*	*Marks*
	Theory examination	3	80
	Internal Assessment	--	20
	Practical Examination	--	--
	Oral Examination	--	--
	Term work	--	--
	Total	--	100

Time Table	*Day*	*Time*
	Tuesday	9:45-10:45 AM
	Wednesday	12-1 PM
	Friday	9:45-10:45 AM

Course Content and Lesson plan

Week	Lecture No.	Date		Topic	Remarks
		Planned	**Actual**		
				Module 1: **Introduction to Big Data**	
1	1	19-07-22	19-07-22	Introduction to Big Data,	
	2	21-07-22	21-07-22	Big Data characteristics, types of Big Data.	
	3	22-07-22	22-07-22	Types of Big Data.	
	4	27-07-22	27-07-22	Traditional vs. Big Data business approach,Case Study of Big Data, Solutions.	
	5	28-07-22	28-07-22	Big Data Case Study	
	6	19-07-22	19-07-22	What is Hadoop? Core Hadoop Components;	
	7	21-07-22	21-07-22	Hadoop Ecosystem; Physical Architecture;.	
	8	22-07-22	22-07-22	Hadoop EcoSystem; Hadoop limitations.	
				Module 2: **Hadoop HDFS and MapReduce:**	

2	9	5-08-22	5-08-22	Physical Organization of Compute Nodes, Large-Scale File-System Organization.	
	10	10-08-22	10-08-22 12-08-22	MapReduce: The Map Tasks, Grouping byKey, The Reduce Tasks,	
	11	12-08-22	18-08-22	Combiners, Details of MapReduce Execution, Coping With Node Failures.**Algorithms using MapReduce:** Word Count Problem	Assignment 1 on Module 1&2
	12	18-08-22	23-8-22	Matrix Vector Multiplication by MapReduce,	
	13	23-08-22(2)	23-8-22(Extra Lec)	Relational Algebra Operations. Computing Selections by MapReduce MapReduce, Computing Natural join by MapReduce, Grouping and Aggregation by MapReduce	
	14	24-08-22	24-08-22	Matrix Multiplication (One-step)Hadoop limitations.	
Module 3: **NoSQL**					
3	15	6-09-22	6-9-22	What is NoSQL? NoSQL business drivers; NoSQL case studies.	Assignment on module 2
	16	7-09-22	7-09-22	Variations of NoSQL architectural patterns:Key-value stores, Graph stores	Holidays from 31/08 to 04/09due to Ganesh Festival
	17	13-09-22	13-09-22	Column family (Bigtable) stores, Document stores	
	18	20-09-22	20-09-22	HBase NoSQL, BigTable NoSQL	
	19	21-09-22	21-09-22	MongoDB NoSQL, Neo4j NoSQL	
	20	25-09-22	25-09-22	Using NoSQL to manage big data: What is abig data NoSQL solution? Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer;Four ways that NoSQL systems handle big data Problem	
Module 4: **Mining Data Streams**					

	21	26-09-22	26-09-22	A Data-Stream- Management System, Stream Queries, Issuesin Stream Processing. Examples of Stream Sources	
4	22	27-09-22	27-09-22	**Sampling Data in a Stream:** Obtaining a Representative Sample, The General Sampling Problem, Varying the Sample Size.	Discussion on module 3
	23	28-09-22	28-09-22	**Filtering Streams**: The Bloom Filter, Analysis, **CountingDistinct Elements in a Stream** The Count-Distinct Problem, The Flajolet- Martin Algorithm **Counting Frequent items in a Stream,** Sampling Methods for Streams, Frequentitemsets in a decaying Windows.	

			Module 5: **Finding Similar Items and Clustering**		
5	24	8-10-22	8-10-22	Applications of Near-Neighbor Search **Distance Measures:** Definition of a Distance Measure, Euclidean Distances, Cosine Distance,	
	25	09-10-22	09-10-22	Edit Distance, Hamming Distance, Jaccard Distance, Jaccard Similarity of Sets,Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem	
	26	10-10-22	10-10-22	**Clustering** - CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets,Answering Queries	

			Module 6: **Real-Time Big Data Models**		
6	27	12-10-22	12-10-22	PageRank Definition, Structure of the web,dead ends, Using Page rank in a search engine	
	28	21-10-22	21-10-22	Efficient computation of Page Rank, PageRank Iteration Using MapReduce, Useof Combiners to Consolidate the Result Vector.	
	29	21-10-22	21-10-22	Topic sensitive Page Rank, link Spam Hubs and Authorities.	

	30	21-10-22	21-10-22	A Model for Recommendation Systems, Content-Based Recommendations,	
	31	22-10-22	21-10-22	Collaborative Filtering. Social Networks as Graphs, Clustering of Social-Network Graphs	
	32	23-10-22	22-10-22	Direct Discovery of Communities, SimRank,Counting triangles using Map-Reduce	

Submitted By	Approved By	
Prof. Ankita Amburle	ii) Dr. Sujata Deshmukh	Sign:
Sign:	ii) Dr. B. S. Daga	Sign:
	iii) Prof. Merly Thomas	Sign:
	iv) Prof. Roshni Padate	Sign:
	v) Prof. Kalpana Deorukhkar	Sign:
Date of Submission:	Date of Approval:	
Remarks by DQAC (if any)		