FR. Conceicao Rodrigues College Of Engineering

Father Agnel Ashram, Bandstand, Bandra-west, Mumbai-50

Department of Information Technology

B.E. (IT) (semester VIII) (2018-2019)

Lesson Plan:

Subject: Big Data Analytics (ITC802) **Credits-5**

Course Code	Course Name	Teaching Scheme Hrs./Week			Credits Assigned			
		Theory	Practical	Tutorial	Theory	Practical/Oral	Tutorial	Total
ITC802	Big Data Analytics	04	02	---	04	01	---	05

Course Code	Course Name	Examination Scheme							
		Theory Marks				Term Work	Practical	Oral	Total
		Internal assessment			End Sem. Exam				
		Test 1	Test 2	Avg. of 2 Tests					
ITC802	Big Data	20	20	20	80	25	---	25	150

Course Objectives:

1. To provide an overview of an exciting growing field of big data analytics.

2. To introduce the tools required to manage and analyze big data like Hadoop, NoSql Map-Reduce.

3. To teach the fundamental techniques and principles in achieving big data analytics with scalability and streaming capability.

4. To enable students to have skills that will help them to solve complex real-world problems in for decision support.

Course Outcomes: At the end of this course a student will be able to:

2. Understand the key issues in big data management and its associated applications in intelligent business and scientific computing.

3. Acquire fundamental enabling techniques and scalable algorithms like Hadoop, Map Reduce and NO SQL in big data analytics.

4. Interpret business models and scientific computing paradigms, and apply software tools for big data analytics.

5. Achieve adequate perspectives of big data analytics in various applications like recommender systems, social media applications etc.

DETAILED SYLLABUS:

DETAILED SYLLABUS:

Sr. No.	Module	Detailed Content	Book	Hours
1	Introduction to Big Data	Introduction to Big Data, Big Data characteristics, types of Big Data, Traditional vs. Big Data business approach, Case Study of Big Data Solutions.	**From Ref. Books**	**03**
2	Introduction to Hadoop	What is Hadoop? Core Hadoop Components; Hadoop Ecosystem; Physical Architecture; Hadoop limitations.	Hadoop in Practise Chapter 1	**02**
3	NoSQL	1. What is NoSQL? NoSQL business drivers; NoSQL case studies; 2. NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns; 3. Using NoSQL to manage big data: What is a big data NoSQL solution? Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution	**No-SQL book**	**04**

		models: master-slave versus peer-to-peer; Four ways that NoSQL systems handle big data problems		
4	MapReduce and the New Software Stack	**Distributed File Systems :** Physical Organization of Compute Nodes, Large-Scale File-System Organization. **MapReduce:** The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures. **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, Computing Natural Join by MapReduce, Grouping and Aggregation by MapReduce, Matrix Multiplication, Matrix Multiplication with One MapReduce Step.	**Text Book 1**	**06**

5	Finding Similar Items	Applications of Near-Neighbor Search, Jaccard Similarity of Sets, Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem . **Distance Measures:** Definition of a Distance Measure , Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance.	**Text Book 1**	**03**
6	Mining Data Streams	**The Stream Data Model**: A Data-Stream-Management System, Examples of Stream Sources, Stream Querie, Issues in Stream Processing. **Sampling Data in a Stream** : Obtaining a Representative Sample , The General Sampling Problem, Varying the Sample Size. **Filtering Streams**: The Bloom Filter, Analysis. **Counting Distinct Elements in a Stream** The Count-Distinct Problem, The Flajolet-Martin Algorithm, Combining Estimates, Space Requirements . **Counting Ones in a Window**: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.	**Text Book 1**	**06**
7	Link Analysis	PageRank Definition, Structure of the web, dead ends,	**Text**	**05**

		Using Page rank in a search engine, Efficient computation of Page Rank: PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector. Topic sensitive Page Rank, link Spam, Hubs and Authorities.	**Book 1**	
8	Frequent Itemsets	**Handling Larger Datasets in Main Memory** Algorithm of Park, Chen, and Yu, The Multistage Algorithm, The Multihash Algorithm. **The SON Algorithm and MapReduce** **Counting Frequent Items in a Stream** Sampling Methods for Streams, Frequent Itemsets in Decaying Windows	**Text Book 1**	**05**
9	Clustering	CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets,	**Text**	**05**

[redacted]

		Answering Queries	**Book 1**	
10	Recommendation Systems	A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering.	**Text Book 1**	**04**
11	Mining Social-Network Graphs	Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, SimRank, Counting triangles using Map-Reduce	**Text Book 1**	**05**

Text Books:

1. Anand Rajaraman and Jeff Ullman "**Mining of Massive Datasets**", Cambridge University Press,

2. Alex Holmes "Hadoop in Practice", Manning Press, Dreamtech Press.

3. Dan McCreary and Ann Kelly "**Making Sense of NoSQL" – A guide for managers and the rest of us**, Manning Press.

References:

- Bill Franks **, "Taming The Big Data Tidal Wave: Finding Opportunities In Huge Data Streams With Advanced Analytics**", Wiley

- Chuck Lam, "**Hadoop in Action**", Dreamtech Press

- Judith Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman, "**Big Data for Dummies",** Wiley India

- Michael Minelli, Michele Chambers, Ambiga Dhiraj, “**Big Data Big Analytics: Emerging Business Intelligence And Analytic Trends For Today's Businesses”,** Wiley India

- Phil Simon**, “Too Big To Ignore: The Business Case For Big Data”,** Wiley India

- Paul Zikopoulos, Chris Eaton**, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data’,** McGraw Hill Education.

- Boris Lublinsky, Kevin T. Smith, Alexey Yakubovich**, “Professional Hadoop Solutions”,** Wiley India.

Oral Exam:

An oral exam will be held based on the above syllabus.

Term work:

Assign a case study for group of 2/3 students and each group to perform the following experiments on their case-study; Each group should perform the exercises on a large dataset created by them.

Term work: (15 marks for programming exercises + 10 marks for mini-project)

Suggested Practical List: Students will perform at least 8 programming exercises and implement one mini-project. The students can work in groups of 2/3.

1. Study of Hadoop ecosystem
2. 2 programming exercises on Hadoop
3. 2 programming exercises in No SQL
4. Implementing simple algorithms in Map- Reduce (3) - Matrix multiplication, Aggregates, joins, sorting, searching etc.
5. Implementing any one Frequent Itemset algorithm using Map-Reduce
6. Implementing any one Clustering algorithm using Map-Reduce
7. Implementing any one data streaming algorithm using Map-Reduce
8. Mini Project: One real life large data application to be implemented (Use standard Datasets available on the web)
   a) Twitter data analysis
   b) Fraud Detection
   c) Text Mining etc.

Theory Examination:

- Question paper will comprise of 6 questions, each carrying 20 marks.
- Total 4 questions need to be solved.

- Q.1 will be compulsory, based on entire syllabus where in sub questions of 2 to 3 marks will be asked.
- Remaining question will be randomly selected from all the modules.

Weight age of marks should be proportional to number of hours assigned to each module.

2. Course Outcome Statement

Sr.No.	Course Outcome Statement
ITC802.1	Explain characteristics of and trends in big data.
ITC802.2	Use tools like hadoop and NoSQL to solve big data related problems.
ITC802.3	Apply appropriate algorithms for extracting knowledge from given dataset.
ITC802.4	Apply Big data analytics in real life applications.

3.CO-PO and CO-PSO Mapping

Course Name	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO1		3		2									2	3
CO2		2	2		3								2	3
CO3		2	3	2	2				1	1			2	3
CO4		2			1								2	3

4. CO Assessment Tools

	Direct Methods						Indirect Methods
							Course Exit Survey
ITC802.1	UT1(30%)	A1(40%)	Oral(10%)	UE(20%)			100%
ITC802.2	UT1(20%)	P1-P5(50%)	Oral(10%)	UE(20%)			100%
ITC802.3	UT2(20%)	P6-P7(30%)	Presentation(20%)	Oral(10%)	UE(20%)		100%
ITC802.4	UT2(30%)	P8-P9(20%)	A2(20%)	Oral(10%)	UE(20%)		100%

5. Course Outcomes Target:

Upon Completion of this course, students will be able to :

ITC802.1: Explain characteristics of and trends in big data.[B2:Comprehension]

Target level: 2.0

ITC802.2: Use tools like hadoop and NoSQL to solve big data related problems.[B3:Application]

Target level: 2.0

ITC802.3: Apply appropriate algorithms for extracting knowledge from given dataset.[B3:Application]

Target level: 2.0

ITC802.4:Apply Big data analytics in real life applications. [B3:Application]

Target level: 2.0

6.Content Beyond Curriculum

1. Research papers to be presented in a group of 4 students.

7.Lesson Plan

No of classes available:	47	1. No of Classes taken: 2.Total Remedial Lectures	47	
Sr. No.	Topic Planned with CO	Planned Date	Actual Date	Delivery Mechanisms
	Don’t forget to include CO dissemination			
1.	Introduction to Big Data(ITC802.1)	04-01-2019	04-01-2019	Blackboard, ppt, notes
2.	Introduction to Hadoop(ITC802.2)	08-01-2019	08-01-2019	Blackboard, ppt, notes, videos
3.	NoSQL(ITC802.2)	12-02-2019	12-02-2019	Blackboard, ppt, notes, videos
4.	MapReduce and New Software Stack(ITC802.2)	22-02-2019	22-02-2019	Blackboard, ppt
5.	Finding Similar Items(ITC802.2)	29-01-2019	29-01-2019	Blackboard, notes, videos
6.	Mining Data Streams(ITC802.3)	05-03-2019	05-03-2019	Blackboard, notes, videos
7.	Link Analysis(ITC802.3)	19-03-2019	19-03-2019	Blackboard, notes
8.	Frequent Itemsets(ITC802.3)	27-03-2019	27-03-2019	Blackboard, notes
9.	Clustering(ITC802.3)	28-03-2019	28-03-2019	Blackboard, notes
10.	Recommendation Systems(ITC802.4)	03-04-2019	03-04-2019	Blackboard, notes
11.	Mining Social Network Graphs(ITC802.4)	05-04-2019	05-04-2019	Blackboard, notes

Date wise lecture plan

Date	**Topic Taught**	**Date**	**Topic Taught**
01-01-2019	Introduction to the course and course outcomes	01-01-2019	Big data and its characteristics
02-01-2019	Types of big data	03-01-2019	Traditional Vs. big data business approach
04-01-2019	Big Data Case study	08-01-2019	What is Hadoop? Core Components of Hadoop
08-01-2019	Hadoop Ecosystem, Physical architecture, Hadoop Limitations	09-01-2019	DFS, Physical organization of compute nodes, Large scale file system organization
10-01-2019	MapReduce: Map tasks, grouping by key and reduce tasks	11-01-2019	Combiners, details of mapredue execution, Coping with node failures
15-01-2019	Matrix-vector multiplication by MapReduce	16-01-2019	Relational algebra operations: selection, projection, set operators
17-01-2019	Natural Join, Grouping and aggregation	18-01-2019	Matrix Multiplication in two phase mapreduce
22-01-2019	Matrix multiplication using one phase map reduce	23-01-2019	Applications of Near neighbor search, jaccard similarity of sets, Similarity of documents
24-01-2019	CF as similar sets problem, Definition of distance measure, Euclidean distances	25-01-2019	Jaccard distance, cosine distance
29-01-2019	Edit and hamming distance	30-01-2019	What is NoSQL? NOSQL business drivers, NOSQL case studies
01-02-2019	Key value stores, Graph stores	07-02-2019	Column family stores, Document stores
10-02-2019	Variations of NoSQL patterns, Big data NoSQL solution, types of big data problems	12-02-2019	Analysis of big data using shared nothing architecture, distribution models, 4 ways No SQL systems handle big data problems
21-02-2019	Data stream model, examples of stream sources, stream query, issues in stream processing	22-02-2019	Sampling in data streams
26-02-2019	Bloom Filter and analysis	27-02-2019	Counting distinct elements in stream
28-02-2019	FM algorithm, combining estimates and space requirements	01-03-2019	Counting ones in a window, DGIM algorithm
01-03-2019	Query answering in DGIM	05-03-2019	Decaying window
06-03-2019	PR definition, structure of web, dead ends	06-03-2019	Using PR in search engines
13-03-2019	Efficient computation of Page rank, use of Mapreduce in PR calculation	14-03-2019	Use of combiners, Topic sensitive PR
19-03-2019	Link spam, Hubs and authorities	20-03-2019	PCY algorithm

22-03-2019	Multistage and multihash algorithm	26-03-2019	SON algorithm using Mapreduce
27-03-2019	Sampling methods for stream, Frequentitemsets in decaying windows	28-03-2019	CURE algorithm, Stream computing and clustering algorithm, Initialization and merging of buckets, answering queries
29-03-2019	A model of recommendation systems	02-04-2019	Content based recommendations
03-04-2019	Collaborative Filtering	04-04-2019	Social networks as graph, Clustering of social network graphs, Direct discovery of communities
05-04-2019	SimRank, Counting triangles using mapreduce		

8.Lab Plan

		Batch	Planned Dates	Actual Dates	Relevant CO
1	Installation and Configuration of Hadoop	A	18/1/19	18/1/19	ITC802.2
		B	15/1/19	15/1/19	ITC802.2
		C	16/1/19	16/1/18	ITC802.2
		D	17/1/19	17/1/19	ITC802.2
2	Counting number of words in a file using Map Reduce.	A	25/1/19	25/1/19	ITC802.2
		B	22/1/19	22/1/19	ITC802.2
		C	23/1/19	23/1/19	ITC802.2
		D	24/1/19	24/1/19	ITC802.2
3	Finding Maximum Temperature using Map Reduce	A	1/2/19	1/2/19	ITC802.2
		B	29/1/19	29/1/19	ITC802.2
		C	30/1/19	30/1/19	ITC802.2
		D	7/2/19	7/2/19	ITC802.2
4	Matrix Multiplication using Map Reduce	A	8/2/19	8/2/19	ITC802.2
		B	5/2/19	5/2/19	ITC802.2
		C	30/1/19	30/1/19	ITC802.2
		D	21/2/19	21/2/19	ITC802.2
5	CRUD operations in MongoDB	A	22/2/19	22/2/19	ITC802.2
		B	26/2/19	26/2/19	ITC802.2
		C	22/2/18	22/2/19	ITC802.2
		D	28/2/19	28/2/19	ITC802.2
6	Implementation of Bloom filter in python	A	1/3/19	1/3/19	ITC802.3

		B	5/3/19	5/3/19	ITC802.3
		C	27/2/19	27/2/19	ITC802.3
		D	6/3/19	6/3/19	ITC802.3
7	Implementation of K-means using Map Reduce	A	8/3/19	8/3/19	ITC802.3
		B	12/3/19	12/3/19	ITC802.3
		C	6/3/19	6/3/19	ITC802.3
		D	7/3/19	7/3/19	ITC802.3
8	Implementation of Recommendation System in R	A	22/3/19	22/3/19	ITC802.4
		B	19/3/19	19/3/19	ITC802.4
		C	14/3/19	13/3/19	ITC802.4
		D	14/3/19	14/3/19	ITC802.4
9	Social Network Analysis using Map Reduce	A	29/3/19	29/3/19	ITC802.4
		B	26/3/19	26/3/19	ITC802.4
		C	20/3/19	20/3/19	ITC802.4
		D	28/3/19	28/3/19	ITC802.4
10	Presentation of a case study/mini project	A	5/4/19	5/4/19	ITC802.3
		B	2/4/19	2/4/19	ITC802.3
		C	27/3/19	27/3/19	ITC802.3
		D	4/9/19	4/9/19	ITC802.3

9.Assignment Plan

Assignment No.	Date	Topics with CO
1	14-03-2019	Introduction to big data.(ITC802.1)
2	05-04-2019	Recommendation systems and Social Network Analysis (ITC802.4)