FR. Conceicao Rodrigues College Of Engineering

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

Department of Information Technology

B.E. (IT) (semester VIII) (2019-2020)

Lesson Plan:

Subject- Big Data Analytics(ITC801)

Course Code	Course Name	Theory	Practical	Tutorial	Theory	Practical /Oral	Tutorial	Total
ITC801	Big Data Analytics	04	--	--	04	--	--	04

Course Code	Course Name	Examination Scheme							
		Theory Marks				Term Work	Practical & Oral	Oral	Total
		Internal assessment			End Sem. Exam				
		Test1	Test2	Avg. of two Tests					
ITC801	Big Data Analytics	20	20	20	80	--	--	--	100

Course Objectives: Students will try:

1. To provide an overview of an exciting growing field of Big Data analytics.
2. To discuss the challenges traditional data mining algorithms face when analyzing Big Data.
3. To introduce the tools required to manage and analyze big data like Hadoop, NoSql Map-Reduce.
4. To teach the fundamental techniques and principles in achieving big data analytics with scalability and streaming capability.
5. To introduce to the students several types of big data like social media, web graphs and data streams.
6. To enable students to have skills that will help them to solve complex real-world problems in for decision support.

Course Outcomes: Student will be able to:

1. Explain the motivation for big data systems and identify the main sources of Big Data in the real world.
2. Demonstrate an ability to use frameworks like Hadoop, NOSQL to efficiently store retrieve and process Big Data for Analytics.
3. Implement several Data Intensive tasks using the Map Reduce Paradigm
4. Apply several newer algorithms for Clustering Classifying and finding associations in Big Data
5. Design algorithms to analyze Big data like streams, Web Graphs and Social Media data.

6. Design and implement successful Recommendation engines for enterprises.

Prerequisites: Database Management System.

Detailed syllabus:

Sr. No.	Module	Detailed Content	Hours	CO Mapping
0	**Prerequisites**	Data Mining, database Systems, Algorithms	02	**--**
I	Introduction to Big Data	Introduction to Big Data, Big Data characteristics, types of Big Data, Traditional vs. Big Data business approach, Big Data Challenges, Examples of Big Data in Real Life, Big Data Applications	03	CO 1
II	Introduction to Big Data Frameworks: Hadoop, NOSQL	What is Hadoop? Core Hadoop Components; Hadoop Ecosystem; **Overview of** : Apache Spark, Pig, Hive, Hbase, Sqoop **What is NoSQL?** NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, **Mongo DB**	10	CO 2
III	MapReduce Paradigm	**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 . **Illustrating** use of MapReduce with use of real life databases and applications.	09	CO 3
IV	Mining Big Data Streams	**The Stream Data Model:** A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing. **Sampling Data in a Stream :** Sampling Techniques. **Filtering Streams:** The Bloom Filter	07	CO 5

		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.		
V	Big Data Mining Algorithms	**Frequent Pattern Mining** : Handling Larger Datasets in Main Memory Basic Algorithm of Park, Chen, and Yu. The SON Algorithm and MapReduce. **Clustering Algorithms:** CURE Algorithm. Canopy Clustering, Clustering with MapReduce **Classification Algorithms:** Parallel Decision trees, Overview SVM classifiers, Parallel SVM, K-Nearest Neighbor classifications for Big Data, One Nearest Neighbour.	10	CO 4
VI	Big Data Analytics Applications	**Link Analysis** : PageRank Definition, Structure of the web, dead ends, Using Page rank in a search engine, Efficient computation of Page Rank: PageRank Iteration Using MapReduce, Topic sensitive Page Rank, link Spam, Hubs and Authorities, HITS Algorithm. **Mining Social- Network Graphs** : Social Networks as Graphs, Types , Clustering of Social Network Graphs, Direct Discovery of Communities, Counting triangles using Map-Reduce. **Recommendation Engine**s: A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering.	11	CO 4 CO 6

Text Books:

1. Radha Shankarmani, M Vijayalakshmi, "Big Data Analytics", Wiley Publications,
2. Anand Rajaraman and Jeff Ullman "Mining of Massive Datasets", Cambridge University Press.
3. Alex Holmes "Hadoop in Practice", Manning Press, Dreamtech Press.
4. Professional NoSQL Paperback, by Shashank Tiwari, Dreamtech Press
5. MongoDB: The Definitive Guide Paperback, Kristina Chodorow (Author), Michael Dirolf, O'Reilly Publications

References:

1. Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, Bart Baesens , WILEY Big Data Series.
2. Big Data Analytics with R and Hadoop by Vignesh Prajapati Paperback, Packt Publishing Limited
3. Hadoop: The Definitive Guide by Tom White, O'ReillyPublications

Assessment:

Internal Assessment for 20 marks:

Consisting of **Two Compulsory Class Tests**

Approximately 40% to 50% of syllabus content must be covered in First test and remaining 40% to 50% of syllabus contents must be covered in second test.

End Semester Examination: Some guidelines for setting the question papers are as:

- Weightage of each module in end semester examination is expected to be/will be proportional to number of respective lecture hours mentioned in the syllabus.
- Question paper will comprise of total **six questions**, **each carrying 20 marks.**
- **Q.1** will be **compulsory** and should **cover maximum contents of the syllabus**.
- **Remaining question will be mixed in nature** (for example if Q.2 has part (a) from module 3 then part (b) will be from any other module. (Randomly selected from all the modules.)
- Total **four questions** need to be solved.

2. Course Outcome Statement

Sr.No.	Course Outcome Statement
ITC801.1	Explain characteristics of and trends in big data.
ITC801.2	Use tools like hadoop and NoSQL to solve big data related problems.
ITC801.3	Apply appropriate algorithms for extracting knowledge from given dataset.
ITC801.4	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
ITC801.1		3		2									2	3
ITC801.2		2	2		3								2	3
ITC801.3		2	3	2	2				1	1			2	3
ITC801.4		2			1								2	3

4. CO Assessment Tools

	Direct Methods						**Indirect Methods**
							Course Exit Survey
ITC801.1	UT1(30%)	A1(40%)	Oral(10%)	UE(20%)			100%
ITC801.2	UT1(20%)	P1-P5(50%)	Oral(10%)	UE(20%)			100%
ITC801.3	UT2(20%)	P6-P7(30%)	Presentation(20%)	Oral(10%)	UE(20%)		100%
ITC801.4	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 :

ITC801.1: Explain characteristics of and trends in big data.[B2:Comprehension]

Target level: 2.0

ITC801.2: Use tools like hadoop and NoSQL to solve big data related problems.[B3:Application]

Target level: 2.0

ITC801.3: Apply appropriate algorithms for extracting knowledge from given dataset.[B3:Application]

Target level: 2.0

ITC801.4:Apply Big data analytics in real life applications. [B3:Application]

Target level: 2.0

6.Content Beyond Curriculum

1. Mini Project to be implemented in Python/R/Java.

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
1.	Introduction to Big Data(ITC801.1)	09-01-2020	09-01-2020	Blackboard, ppt, notes
2.	Introduction to Big Data Frameworks; Hadoop, NoSQL(ITC801.2)	24-01-2020	24-01-2020	Blackboard, ppt, notes, videos
3.	MapReduce Paradigm(ITC801.2)	28-01-2020	14-02-2020	Blackboard, ppt
4.	Mining Big Data Streams(ITC801.3)	11-03-2020		Blackboard, notes, videos
7.	Big Data Mining Algorithms(ITC801.3)	24-03-2020		Blackboard, notes
8.	Big Data Analytics Applications(ITC801.4)	10-04-2020		Blackboard, notes

Date wise lecture plan

Date	**Topic Taught**	**Date**	**Topic Taught**
07-01-2020	Introduction to the course and course outcomes	08-01-2020	Big data and its characteristics
09-01-2020	Types of big data	10-01-2020	Traditional Vs. big data business approach

14-01-2020	Big Data Case study	15-01-2020	What is Hadoop? Core Components of Hadoop
16-01-2020	Hadoop Ecosystem, Physical architecture, Hadoop Limitations	17-01-2020	DFS, Physical organization of compute nodes, Large scale file system organization
21-01-2020	Apache Spark	22-01-2020	NoSQL Data stores
23-01-2020	Key-value stores, Graph Stores	24-01-2020	Column family stores, Document Stores
28-01-2020	MapReduce: Map tasks, grouping by key and reduce tasks	28-01-2020	Combiners, details of mapredue execution, Coping with node failures
29-01-2020	Matrix-vector multiplication by MapReduce	30-01-2020	Relational algebra operations: selection, projection, set operators
04-02-2020	Natural Join, Grouping and aggregation	05-02-2020	Matrix Multiplication in two phase mapreduce
06-02-2020	Matrix multiplication using one phase map reduce	07-02-2020	Applications of Near neighbor search, jaccard similarity of sets, Similarity of documents
11-02-2020	CF as similar sets problem, Definition of distance measure, Euclidean distances	12-02-2020	Jaccard distance, cosine distance
13-02-2020	Edit and hamming distance	14-02-2020	What is NoSQL? NOSQL business drivers, NOSQL case studies

8.Lab Plan

		Batch	Planned Dates	Actual Dates	Relevant CO
1	Installation and Configuration of Hadoop	A	21/1/20	21/1/20	ITC801.2
		B	23/1/20	23/1/20	ITC801.2
		C	24/1/20	24/1/20	ITC801.2
		D	22/1/20	22/1/20	ITC801.2
2	Counting number of words in a file using Map Reduce.	A	28/1/20	28/1/20	ITC801.2
		B	30/1/20	30/1/20	ITC801.2
		C	31/1/20	31/1/20	ITC801.2
		D	29/1/20	29/1/20	ITC801.2
3	Finding Maximum Temperature using Map Reduce	A	4/2/20	4/2/20	ITC801.2
		B	6/2/20	6/2/20	ITC801.2
		C	7/2/20	7/2/20	ITC801.2
		D	5/2/20	5/2/20	ITC801.2
4	Matrix Multiplication using Map Reduce	A	11/2/20	11/2/20	ITC801.2
		B	13/2/20	13/2/20	ITC801.2
		C	14/2/20	14/2/20	ITC801.2

		D	12/2/20	12/2/20	ITC801.2
5	CRUD operations in MongoDB	A	3/3/20		ITC801.2
		B	5/3/20		ITC801.2
		C	6/3/20		ITC801.2
		D	4/3/20		ITC801.2
6	Implementation of Bloom filter in python	A	17/3/20		ITC801.3
		B	19/3/20		ITC801.3
		C	20/3/20		ITC801.3
		D	18/3/20		ITC801.3
7	Implementation of K-means using Map Reduce	A	17/3/20		ITC801.3
		B	19/3/20		ITC801.3
		C	20/3/20		ITC801.3
		D	18/3/20		ITC801.3
8	Implementation of Recommendation System in R	A	24/3/20		ITC801.4
		B	26/3/20		ITC801.4
		C	27/3/20		ITC801.4
		D	1/4/20		ITC801.4
9	Social Network Analysis using Map Reduce	A	31/3/20		ITC801.4
		B	2/4/20		ITC801.4
		C	3/4/20		ITC801.4
		D	1/4/20		ITC801.4
10	Presentation of a case study/mini project	A	7/4/20		ITC801.3
		B	9/4/20		ITC801.3
		C	10/4/20		ITC801.3
		D	8/4/20		ITC801.3

9.Assignment Plan

Assignment No.	Date	Topics with CO
1	5-03-2020	Introduction to big data.(ITC801.1)
2	07-04-2020	Recommendation systems and Social Network Analysis (ITC801.4)