FR. Conceicao Rodrigues College Of Engineering

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

Department of Computer Engineering

B.E. (Computer) (semester VIII) (2018-2019)
Course Outcomes & Assessment Plan

Subject: Big Data Analytics (BDA-CPE8035) **Credits-5**

Syllabus:

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.

2. **Introduction to Hadoop:** What is Hadoop? Core Hadoop Components; Hadoop Ecosystem; Physical Architecture; Hadoop limitations.

3. **NoSQL:** What is NoSQL? NoSQL business drivers; NoSQL case studies; NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns; 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 models: master-slave versus peer-to-peer; Four ways that NoSQL systems handle big data Problem

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.

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.

6. **Mining 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: 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, TheDatar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.

7. **Link Analysis**: PageRank Definition, tructure of the web, dead ends, 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.

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

9. **Clustering:** CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets, Answering queries

10. **Recommendation Systems:** A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering.

11. **Mining Social-Network Graphs:** Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, SimRank, Counting triangles using Map-Reduce

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

The distribution of marks for term work shall be as follows:
Programming Exercises: …….. (10) Marks.
Mini project: …….. (10) Marks.
Attendance …….. (05) Marks
TOTAL: …….. (25) Marks.

Internal Assessment:
Internal Assessment consists of two tests. Test 1, an Institution level central test, is for 20 marks and is to be based on a minimum of 40% of the syllabus. Test 2 is also for 20 marks and is to be based on the remaining syllabus. Test 2 may be either a class test or assignment on live problems or course project.

Practical/Oral examination:
An oral exam will be held based on the above syllabus.

Course Objectives (optional):

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.

FR. Conceicao Rodrigues College Of Engineering

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

Department of Computer Engineering

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

Lecture Plan:

Subject: Big Data Analytics (BDA-CPE8035) **Credits-5**

Time Table (2 week):

	Prof. Swati Ringe										**With Effect from 09th January 2018**
	8.45-9.30	**9.30-10.15**	**10.15-11.00**	**11.00-11.15**	**11.15-12.00**	**12.00-12.45**	**12.45-13.15**				
Mon				**BREAK**			**LUNCH**				
Tues		**BDA BEC**									
Wed						**BDA BEC**					
Thurs		**BDA BEC**	**BDA BEC**								
Fri						**BDA BEC**					

Time Table (Regular):

Prof. Swati Ringe						**With Effect from 14th January 2019**				
	8.45-9.45	**9.45-10.45**	**10.45-11.00**	**11.00-12.00**	**12.00-01.00**	**13.00-13.30**	**13.30-14.30**	**14.30-15.30**	**15.30-16.30**	**16.30-17.30**
Mon			**BREAK**			**LUNCH**	**OSL (SEC-D)**		**OSL (SEC-C)**	
Tues		**BDA BEC**						**OSL (SEC-D)**		
Wed		**BDA BEC**					**OSL (SEC-C)**			
Thurs		**BDA BEC**		**BDA (BEC-A)**			**OSL (SEC-A)**			
Fri		**BDA BEC**					**OSL (SEC-A)**			

Total Load: 4T + 14P = 18 + MENTOR

Lecture Plan : SEM VIII-BDA-CPE8035

Modes of Content Delivery:

i	Class Room Teaching	v	Self Learning Online Resources	ix	Industry Visit
ii	Tutorial	vi	Slides	x	Group Discussion
iii	Remedial Coaching	vii	Simulations/Demonstrations	xi	Seminar
iv	Lab Experiment	viii	Expert Lecture	xii	Case Study

No.	Portion to be covered	Planned date	Actual date	Content Delivery - Reference /Assessment Method
1.	**Introduction to Big Data:** Introduction to Big Data, Big Data characteristics, types of Big Data.	01/01/2019	01/01/2019	PPT [1_BigData] – Video1, [TB1] /UT1
2	Traditional vs. Big Data business approach, Case Study of Big Data, Solutions.	02/01/2019	02/01/2019	PPT[1_BigData]- [TB1] / Group Discussion
3	**Introduction to Hadoop:** What is Hadoop? Core Hadoop Components;	03/01/2019 (2 lectures)	03/01/2019	PPT[2_Hadoop]- Video2,[TB1_4], Chart/ UT1
4	Hadoop Ecosystem; Physical Architecture; Hadoop limitations.	04/01/2019	04/01/2019	
5	**MapReduce and the New Software Stack: Distributed File Systems:** Physical Organization of Compute Nodes, Large-Scale File-System Organization.	08/01/2019	08/01/2019	PPT[2_Hadoop]- Video3,[TB1_4] /UT1
6	**MapReduce:** The Map Tasks, Grouping by Key, The Reduce Tasks,	09/01/2019	09/01/2019	PPT[2_Hadoop]- [TB1_4] /PostLab
7	Combiners, Details of MapReduce Execution, Coping With Node Failures.	10/01/2019	11/01/2019	
10	**Algorithms using MapReduce:** Matrix Vector Multiplication by MapReduce, Relational Algebra Operations. Computing Selections by MapReduce	11/01/2019 15/01/2019	10/01/2019 10/01/2019	ClassRoom Teaching - [TB1_4] / Lab Expt, UT1
11	Computing Projections by MapReduce, Union, Intersection and difference by MapReduce, Computing Natural join by MapReduce, Grouping and Aggregation by MapReduce	16/01/2019	16/01/2019	

12	Matrix Multiplication (One-step)	**17/01/2019**	**15/01/2019**	ClassRoom Teaching- **[TB1]/ Lab Expt**
13	**NoSQL:** What is NoSQL? NoSQL business drivers; NoSQL case studies.	**18/01/2019**	**17/01/2019**	PPT[3_NoSQL], Case Study- **[TB4]/UT1**
14	Variations of NoSQL architectural patterns: Key-value stores, Graph stores	**22/01/2019**	**18/01/2019**	PPT[3_NoSQL], Case Study **[TB3_4]/ UT1**
15	Column family (Bigtable) stores, Document stores,	**23/01/2019**	**22/01/2019**	
16	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 models: master-slave versus peer-to-peer; Four ways that NoSQL systems handle big data Problem	**24/01/2019** **25/01/2019**	**23/01/2019** **24/01/2019**	PPT[3_NoSQL]- **[TB4]/ UT1**
17	**Finding Similar Items** Applications of Near-Neighbor Search, Jaccard Distance, Jaccard Similarity of Sets, Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem .	**29/01/2019**	**25/02/2019**	ClassRoom Teaching – **[TB1_4]/ Quiz1**
18	**Distance Measures:** Definition of a Distance Measure, Euclidean Distances, Cosine Distance,	**30/01/2019**	**29/02/2019**	
19	Edit Distance, Hamming Distance.	**06/02/2019**		
20	**Link Analysis** PageRank Definition, Structure of the web, dead ends, Using Page rank in a search engine	**07/02/2019**	**30/02/2019** **06/02/2019**	PPT- **[TB1_4]/**
21	Efficient computation of Page Rank	**20/02/2019**	**07/02/2019**	ClassRoom Teaching- **[TB1_4] /UT2/ Lab_Expt**
22	PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector.	**21/02/2019**	**20/02/2019**	
23	Topic sensitive Page Rank, link Spam	**22/02/2019**	**21/02/2019**	
24	Hubs and Authorities.	**26/02/2019**	**22/02/2019**	
25	**Mining Data Streams** **The Stream Data Model**: A Data-Stream-Management System	**27/02/2019**	**26/02/2019**	PPT **[TB1_4]**
26	Stream Querie, Issues in Stream Processing.Examples of Stream Sources	**28/02/2019**	**27/02/2019**	

27	**Sampling Data in a Stream** : Obtaining a Representative Sample	**01/03/2019**	**28/02/2019**	
28	The General Sampling Problem, Varying the Sample Size.	**05/03/2019**	**01/03/2019**	
29	**Filtering Streams**: The Bloom Filter, Analysis.	**06/03/2019**	**05/03/2019**	Expert Lecture **[TB1_4]/** **Quiz, Guest Lect**
30	**Counting Distinct Elements in a Stream**The Count-Distinct Problem, The Flajolet-Martin Algorithm, Combining Estimates, Space Requirements	**07/03/2019**	**06/03/2019**	Classroom Teaching, **[TB1_4]/** **UT2**
31	**Counting Ones in a Window**: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm,Query Answering in the DGIM Algorithm, Decaying Windows.	**08/03/2019**	**07/03/2019**	Classroom Teaching, **[TB1_4]/** **UT2**
32	**Frequent Itemsets**-Handling Larger Datasets in Main Memory	**12/03/2019**	**08/03/2019**	Classroom Teaching **[TB1_4]/UT2**
33	Algorithm of Park, Chen, and Yu	**13/03/2019**	**12/02/2019**	
34	The Multistage Algorithm, The Multihash Algorithm.	**14/03/2019**	**13/02/2019**	Seminar **[TB1_4]**
35	The SON Algorithm and MapReduce	**19/03/2019**	**14/02/2019**	
36	Counting Frequent Items in a Stream Sampling Methods for Streams, Frequent Itemsets in Decaying Windows	**20/03/2019**	**19/03/2019**	
37	**Clustering** - CURE Algorithm,	**22/03/2019**	**20/03/2019**	PPT- **[TB1_4]** **/Lab Expt**
38	Stream-Computing	**26/03/2019**	**22/03/2019**	
39	A Stream-Clustering Algorithm,	**27/03/2019**	**26/03/2019**	
40	Initializing & Merging Buckets, Answering Queries	**28/03/2019**	**27/03/2019**	
41	**Recommendation Systems** A Model for Recommendation Systems, Content-BasedRecommendations,	**29/03/2019**	**03/04/2019**	Case Study-Seminar **[TB1_4]** **/UT2**
42	Collaborative Filtering.	**02/04/2019**	**04/04/2019**	
43	**Mining Social-Network Graphs** Social Networks as Graphs, Clustering of Social-Network Graphs	**03/04/2019**	**28/03/2019**	PPT **[TB4]**, HB **/UT2, Lab Expt**
44	Direct Discovery of Communities	**04/04/2019**	**29/03/2019**	PPT **[TB4]**
45	SimRank, Counting triangles using Map-Reduce	**11/04/2019**	**02/04/2019**	

Total Lectures : 46

Course Code	Course Name	Teaching Scheme			Credits Assigned			
		Theory	Practical	Tutorial	Theory	Practical/Oral	Tut	Credits
CPE 8035	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				
		Test1	Test2	Avg					
CPE 8035	Big Data Analytics	20	20	20	80	25	---	25	150

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.

The distribution of marks for term work shall be as follows:
· Programming Exercises: …….. (10) Marks.
· Mini project: …… (10) Marks.
· Attendance ……..…….. (05) Marks
TOTAL: …….. (25) Marks.

Internal Assessment:

Internal Assessment consists of two tests. Test 1, an Institution level central test, is for 20 marks and is to be based on a minimum of 40% of the syllabus. Test 2 is also for 20 marks and is to be based on the remaining syllabus. Test 2 may be either a class test or assignment on live problems or course project.

Practical/Oral examination:

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

Text Books/ Reference Books:

Te x t B o o k s :

[TB1]- Anand Rajaraman and Jeff Ullman "Mining of Massive Datasets", Cambridge University Press,
[TB2]-Alex Holmes "Hadoop in Practice", Manning Press, Dreamtech Press.
[TB3]-Dan McCreary and Ann Kelly "Making Sense of NoSQL" – A guide for managers and the rest of us, Manning Press.
[TB4]- VijayaLaxmi, Radha Shankarmani, "Big Data Analytics", Wiley.

Reference Books:

1. Bill Franks, "Taming The Big Data Tidal Wave: Finding Opportunities In Huge Data Streams With Advanced Analytics", Wiley
2. Chuck Lam, "Hadoop in Action", Dreamtech Press

Reference

[HB1]-Handbook with sample real life problems solution

Slides

Reference Web Resources:

1. Stanford University Lecture series on Mining Massive Data Sets.
2. BigDataUniversity web site.

Course Outcomes:

Upon completion of this course students will be able to:

CPE8035.1: Explain characteristics of and trends in big data.[**B2:Comprehension**]

CPE8035.2: Solve big data related problems using the tools like Hadoop and NoSQL.**[B3:Application]**

CPE8035.3: Apply appropriate algorithms for extracting knowledge from given BigDataSet. **[B3:Application]**

CPE8035.4: Simulate real life applications of big data analytics. **[B3:Application]**

Mapping of CO and PO/PSO

Relationship of course outcomes with program outcomes: Indicate 1 (low importance), 2 (Moderate Importance) or 3 (High Importance) in respective mapping cell.

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CSC8035.1	3			1					2				3	
CSC8035.2	3	3	2		3				2				3	3
CSC8035.3	3	3	3		2				2	1		3	3	3
CSC8035.4	3	3	3		2				2	2	2		3	3
TOTAL	9	6	8	1	7				8	3	2	3	12	9
CO-PO MATRIX	**3**	**3**	**2.66**	**1**	**2.33**				**2**	**1.5**	**2**	**3**	**3**	**3**

Justification

PO2: COs are mapped to this PO1 and 2 because the students understand and analyze the Big Data generated..

PO3: CO8035.2 , CO8035.3 and CO8035.4 are mapped to this PO3 because the students design and develop the mini software system using Big Data technologies.

PO4: CO8035.1 and CO8035.3 are mapped to PO4 because the students investigate the case studies.

PO5: CO8035.2 , CO8035.3 and CO8035.4 are mapped to this PO5 because the students use the tools like hadoop, mapreduce, MongoDB, Mahout tools to do the mini analytics system

PO9: CO8035.1-4 are mapped to this PO9 because the students work in a team and individually to develop the mini software system. (2 because not multidisplinary)

PO10: CO302.3 and CO302.4 are mapped to this PO10 because the students have do presentation and submit written report of the mini software system.

PO11: CO302.4 are mapped to PO11 because They do presentation of their mini project and research paper.

PO12: CO302.3 is mapped to PO12 because the students study research papers on a particular topic and summarizes them.

PSO1: All COs are mapped to PSO1 because the graduates will be able to apply fundamental knowledge of Big Data Analytics to provide computer base solution to real world problems.

PSO2: All COs are mapped to this PSO2 because the students design and implement the mini software system with consideration of Analytics

CO Assessment Tools:

CSC8035.1: **Direct Methods(80%):** Test1(Q1) Quiz1 UniExamThUniExamOral

CO1dm = 0.3T1(Q1-5M) + 0.3Quiz1+ 0.2UTh + 0.2UO.

InDirectMethods(20%): Course exit survey

CO1idm

CSC8035.1 = 0.8*CO1dm + 0.2* CO1idm

CSC8035.2: **Direct Methods(80%):** Test 1(Q2) Labs1-5 Assign1 UniExamThUniExamOral

CO2dm = 0.3T1(Q2-15M) + 0.3Lab1-5 +0.1ASSIGN1 + 0.2UTh + 0.1UO.

InDirectMethods(20%): Course exit survey

CO2idm

CSC8035.2 = 0.8*CO2dm + 0.2* CO2idm

CSC8035.3: **Direct Methods(80%):** Test2 (Q1) Labs6-8 MiniProject UniExamThUniExamOral

CO3dm = 0.4UT2 (Q1-15M)+ 0.1MP +0.2Lab6-8+ 0.2UTh + 0.1UPO.

InDirectMethods(20%): Course exit survey

CO3idm

CSC8035.3 = 0.8*CO3dm + 0.2* CO3idm

CSC8035.4: **Direct Methods(80%):** Test2(Q2) MiniProject UniExamThUniExamOral

CO4dm = 0.2UT2 (Q 2-5M) + 0.4MiniProject + 0.2UTh + 0.2UO.

InDirectMethods(20%): Course exit survey

CO4idm

CSC8035.4 = 0.8*CO4dm + 0.2* CO4idm

Course Outcomes Target:

Upon completion of this course students will be able to:

CPE8035.1: Explain characteristics of and trends in big data.[**B2:Comprehension**]
Target level: 2.5

CPE8035.2: Solve big data related problems using the tools like Hadoop and NoSQL. **[B3:Application]**
Target level: 2.5

CPE8035.3: Apply appropriate algorithms for extracting knowledge from given BigDataSet. **[B3:Application]**
Target level: 2.5

CPE8035.4: Simulate real life applications of big data analytics. **[B3:Application]**
Target level: 2.5

Content Beyond Syllabus:

1. Blooms Filter (Guest Lecture)
2. Research Paper study individually.

Curriculum Gap:

The students need to know basics of Data Mining Algorithms.

In order to achieve the course objectives, there are some topics listed below are not given much importance.

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

Department of Computer Engineering
Academic Term: Jan-April 2019

Rubrics for Lab Experiments

Class : B.*E. Computer* **Subject Name :*BDA***
Semester : VIII **Subject Code :*CPE8035***

Practical No:	
Title:	
Date of Performance:	
Roll No:	
Name of the Student:	

Evaluation:

Indicator	**Very Poor**	**Poor**	**Average**	**Good**	**Excellent**
Timeline (2)	More than three sessions late (0)	More than two sessions late (0.5)	Two sessions late (1)	One session late (1.5)	Early or on time (2)
Completeness(3)	N/A	N/A	Not Completed (1)	Partially Completed (2)	Completed(3)
Legibility(3)	N/A	N/A	Poor(1)	Good(2)	Very Good(3)
PostLab(2)	N/A	N/A	N/A	Partially Correct(1)	All Correct(2)

Total Marks :
Signature of the Teacher :

Department of Computer Engineering
Academic Term : Jan-April 2019

Rubrics for Assignments

Class : B.*E. Computer* **Subject Name :*BDA***
Semester : VIII **Subject Code :*CPE8035***

Assignment No:	
Title:	
Date of Performance:	
Roll No:	
Name of the Student:	

Rubrics for Assignment Grading:

Indicator	**Very Poor**	**Poor**	**Average**	**Good**	**Excellent**
Timeline (2)	More than three sessions late (0)	More than two sessions late (0.5)	Two sessions late (1)	One session late (1.5)	Early or on time (2)
Organization (3)	N/A	Very poor readability and not structured (0.5)	Poor readability and somewhat structured (1)	Readable with one or two mistakes and structured (2)	Very well written and structured without any mistakes (3)
Level of content (3)	N/A	Major points are omitted or addressed minimally (0.5)	All major topics are covered, the information is accurate.(1)	Most major and some minor criteria are included. Information is Accurate (2)	All major and minor criteria are covered and are accurate. (3)
Depth of Knowledge(2)	N/A	One answer correct(0.5)	Two answers correct(1)	Three answers correct(1.5)	Four answers correct(2)

Total Marks :
Signature of the Teacher :

Department of Computer Engineering
Academic Term: Jan-April 2019
Rubrics for Mini Project

Class : B.*E. Computer* **Subject Name :*BDA***
Semester : VIII **Subject Code :*CPE8035***

Practical No:	
Title:	
Date of Performance:	
Roll No:	
Name of the Student:	

Rubric for Mini Project

Indicator	**Very Poor**	**Poor**	**Average**	**Good**	**Excellent**
Timeline: **Maintains project deadline (2)**	Project not done (0)	More than two session late (0.5)	Two sessions late (1)	One session late (1.5)	Early or on time (2)
Completeness: **Complete all parts of project (2)**	N/A	< 40% complete (0.5)	~ 60% complete (1)	~ 80% complete(1.5)	100% complete(2)
Application design: **(4)**	Design aspects are not used (0)	Poorly designed (1)	Project with limited functionalities (2)	Working project with good design (3)	Working project with good design and advanced techniques are used (4)
Presentation(2)	Not submitted report (0)	Poorly written and poorly kept report(0.5)	Report with major mistakes(1)	Report with less than 3-4 mistakes (1.5)	Well written accurate report(2)

Total marks: **Signature of Teacher:**

List of Experiments/Mini Project Plan

Expt No.	Batch A Fri	CO Mapping	Title/aim
01	17/01/19	CO2	Study and Installation of Hadoop Ecosystem
02	24/01/19	CO2	Count the frequency of word using Map Reduce.
03	31/01/19	CO2	Perform CRUD operations in MongoDB.
04	07/2/19	CO2	Matrix – Vector Multiplication using Map-reduce.
05	21/02/19	CO2	Matrix – Matrix Multiplication using Map-reduce.
06	28/02/19	CO3	Implement PM algorithm using Map-reduce.
07	07/03/19	CO3	Implement basic PageRank algorithm using Map-reduce.
08	14/03/19	CO3	Implement PCY algorithm for frequent itemsetmining.
09		**CO4**	**Mini Project:** One real life large data application using standard dataset (Group of 2/3).
	3/2/19		**Topic Submission**
	10/3/19		**Progress review**
	24/3/19		**Presentation and Demo**
	13/4/19		**Mini Project Report submission**

Assignments Plan

Assignments			
01	14/02/2019	CO2-3	Topic of Study
02	31/03/2019	CO4 /PO12	Study of Research Paper
03	31/03/2019	CO1	**Provide the trends and solution using Big data Analytics. (use diagrams)** **Traffic Analysis** 1. An organization wants to create a real-time traffic analysis and prediction application that can be used to control traffic congestion and streamline traffic flow. The application must be targeted to provide cost optimization in commuting and help reduce waiting time and pollution levels. Data has to be captured from existing government provided datasets that include sources such as traffic-camera, traffic sensor, GPS and weather prediction systems. The government data needs to be coupled with social media to assist in predicting traffic speed and volume on reads. The analysis scenarios include the following: Analysis of historical data to gain insights and understand patterns of behavior of traffic and road incidents, Prediction of traffic speed and volume well ahead of time, Based on analysis of real-time and historical traffic data prediction of alternate cost-effective commute paths by analyzing situational traffic conditions across the entire transportation network. The application needs to provide a catalog of services based on social media, governmental data and different data options.
			Telecom Industry **2.** A telecommunication organization needs a solution for analyzing customer behaviour and viewing patterns in advance of rollout of video-over-IP (VOIP) offereings. The logs have to be compared to region specific, feature specific existing system data spread across multiple applications. Because the volume of data is already huge and the VOIP logs data will add many terabutes, the organization is looking for a robust solution to apply across all devices and systems.
			3. Health Care Sector

Assignment on Recommendation systems and mining social network graphs

Q.1

	M1	M2	M3	M4	M5	M6	M7	M8
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Treating the utility matrix representing the ratings on a 1-5 star scale of eight movies provided by users A, B, C.
Compute the following from the data of the matrix.

1. Treat the utility matrix as Boolean, compute the jaccard distance between each pair of users.
2. Repeat Part(1), but use Cosine distance.
3. Treat ratings of 3,4 and 5 as 1,2 and blank as 0. Compute the Jaccard distance between each pair of users.
4. Repeat part (3), but use the cosine distance.
5. Normalize the matrix by subtracting from each non blank entry the average value for its user.
6. Using the normalized matrix from part(5), compute the cosine distance between each pair of users.

Q. 2

Write an algorithm for finding triangles in social network graphs. How to use the algorithm using Map Reduce?