FR. Conceicao Rodrigues College Of Engineering

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

Department of Computer Engineering

B.E. (Computer) (semester VIII) (2019-2020)
Course Outcomes & Assessment Plan

Subject: Big Data Analytics (BDA-CSDLO7032) **Credits-5**

Course Objectives:

1. To provide an overview of an exciting growing field of big data analytics.
2. To introduce programming skills to build simple solutions using big data technologies such as MapReduce and scripting for NoSQL, and the ability to write parallel algorithms for multiprocessor execution.
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.
5. To provide an indication of the current research approaches that is likely to provide a basis for tomorrow's solutions.

Syllabus:

Prerequisite: Some prior knowledge about Java programming, Basics of SQL, Data mining and machine learning methods would be beneficial.

1. **Introduction to Big Data and Hadoop**
   - 1.1 Introduction to Big Data,
   - 1.2 Big Data characteristics, types of Big Data,
   - 1.3 Traditional vs. Big Data business approach,
   - 1.4 Case Study of Big Data Solutions.
   - 1.5 Concept of Hadoop
   - 1.6 Core Hadoop Components; Hadoop Ecosystem.

2. **Hadoop HDFS and MapReduce**
   - 2.1 **Distributed File Systems**: Physical Organization of Compute Nodes, Large-Scale File-System Organization.
   - 2.2 **MapReduce**: The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures.
   - 2.3 **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
   - 2.4 Hadoop Limitations

3. **NoSQL**
   - 3.1 Introduction to NoSQL, NoSQL Business Drivers,
   - 3.2 **NoSQL Data Architecture Patterns:** Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns, NoSQL Case Study
   - 3.3 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:**
   - 4.1 **The Stream Data Model:** A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing.
   - 4.2 Sampling Data techniques in a Stream
   - 4.3 **Filtering Streams**: Bloom Filter with Analysis.
   - 4.4 Counting **Distinct Elements** in a Stream, Count-Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements
   - 4.5 Counting **Frequent Items in** a Stream, Sampling Methods for Streams, Frequent Itemsets in Decaying Windows.
   - 4.6 **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**
   - 5.1 **Distance Measures**: Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance.
   - 5.2 CURE Algorithm, Stream-Computing, A Stream-Clustering Algorithm, Initializing & Merging Buckets, Answering Queries.

6. **Real-Time Big Data Models**
   - 6.1 PageRank Overview, Efficient computation of PageRank: PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector.
   - 6.2 A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering.
   - 6.3 Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph.

Term Work: Assign a case study for group of 3/4 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 (Theory & Practical) ..... (05) Marks.

TOTAL: ..... (25) Marks.

Internal Assessment: Assessment consists of two class tests of 20 marks each. The first class test is to be conducted when approx. 40% syllabus is completed and second class test when additional 40% syllabus is completed. Duration of each test shall be one hour.

End Semester Theory Examination:

1. Question paper will comprise of 6 questions, each carrying 20 marks.
2. The students need to solve total 4 questions.
3. Question No.1 will be compulsory and based on entire syllabus.
4. Remaining questions (Q.2 to Q.6) will be selected from all the modules.

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

Teaching Scheme

Course Code	**Course Name**	**Teaching Scheme**			**Credits Assigned**			
		Theory	**Practical**	**Tutorial**	**Theory**	**Practical/Oral**	**Tut**	**Credits**
CSDLO 7032	Big Data Analytics	04	--	--	04	--	---	04
CSL704	Computational Lab-I	--	02	--	--	1	--	01

Examination Scheme

Course Code	Course Name	Examination Scheme							
		Theory Marks				Term Work	Practical	Oral	Total
		Internal Assessment			End Sem Exam				
		Test1	Test2	Avg					
CSDLO 7032	Big Data Analytics	20	20	20	80 (3hr)	25	---	25	100
CSL704	Computational Lab-I					25	--	25	50

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 2	**Introduction to Big Data:** Introduction to Big Data, Big Data characteristics, types of Big Data.	**02/07/2019** **05/07/2019**		PPT [1_BigData] – **Video1**, **[TB1]** /**UT1**
3	Traditional vs. Big Data business approach, Case Study of Big Data, Solutions.	**09/07/2019**		PPT[1_BigData]- **[TB1]** / **Group Discussion**
4	Big Data Case Study	**11/07/2019**		
5 6	**Introduction to Hadoop:** What is Hadoop? Core Hadoop Components;	**12/07/2019** **16/07/2019**		PPT[2_Hadoop]- **Video2,[TB1_4], Chart**/ **UT1**
7	Hadoop Ecosystem; Physical Architecture;.	**17/07/2019**		
8	Hadoop EcoSystem; Hadoop limitations.	**18/07/2019**		
9	**Hadoop HDFS and MapReduce: Distributed File Systems:** Physical Organization of Compute Nodes, Large-Scale File-System Organization.	**19/07/2019**		PPT[2_Hadoop]- **Video3,[TB1_4]** /**UT1**
10	**MapReduce:** The Map Tasks, Grouping by Key, The Reduce Tasks,	**23/07/2019**		PPT[2_Hadoop]- **[TB1_4]** /**PostLab**
11	Combiners, Details of MapReduce Execution, Coping With Node Failures. **Algorithms using MapReduce:** Word Count Problem	**24/07/2019**		
12	Matrix Vector Multiplication by MapReduce,	**25/07/2019**		ClassRoom Teaching - **[TB1_4]** / **Lab Expt, UT1**
13	Relational Algebra Operations. Computing Selections by MapReduce	**26/07/2019**		
14	Computing Projections by MapReduce, Union, Intersection and difference by	**30/07/2019**		

	MapReduce, Computing Natural join by MapReduce, Grouping and Aggregation by MapReduce			

15	Matrix Multiplication (One-step) Hadoop limitations.	**31/07/2019**		ClassRoom Teaching- **[TB1]/** **Lab Expt**
16 **17**	**NoSQL:** What is NoSQL? NoSQL business drivers; NoSQL case studies.	**01/07/2019** **02/08/2019**		PPT[3_NoSQL], Case Study- **[TB4]/UT1**
18	Variations of NoSQL architectural patterns: Key-value stores, Graph stores	**31/07/2019**		PPT[3_NoSQL], Case Study **[TB3_4]/ UT1**
19	Column family (Bigtable) stores, Document stores,	**06/08/2019**		
20	HBase NoSQL	**07/08/2019**		
21	BigTable NoSQL	**08/08/2019**		
22	MongoDB NoSQL	**09/08/2019**		
23	Neo4j NoSQL	**19/08/2019**		
24	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	**20/08/2019**		PPT[3_NoSQL]- **[TB4]/** **UT1**
25	**Mining Data Streams** **The Stream Data Model**: A Data-Stream-Management System	**21/08/2019**		
26	Stream Querie, Issues in Stream Processing. Examples of Stream Sources	**22/08/2019**		
27	**Sampling Data in a Stream** : Obtaining a Representative Sample	**23/08/2019**		
28	The General Sampling Problem, Varying the Sample Size.	**27/08/2019**		
29	**Filtering Streams**: The Bloom Filter, Analysis.	**28/08/2019**		
30	**Counting Distinct Elements in a Stream** The Count-Distinct Problem,	**29/08/2019**		
31	The Flajolet-Martin Algorithm,	**30/08/2019**		
32	Combining			

	Estimates, Space Requirements			
33	**Counting Ones in a Window**: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.	**11/09/2019**		
34	**Counting Frequent items in a Stream,** Sampling Methods for Streams, Frequent itemsets in a decaying Windows.	**12/09/2019**		
35	**Finding Similar Items and Clustering** Applications of Near-Neighbor Search,.	**13/09/2019**		
36	**Distance Measures:** Definition of a Distance Measure, Euclidean Distances, Cosine Distance,	**16/09/2019**		ClassRoom Teaching – **[TB1_4]/ Quiz1**
37	Edit Distance, Hamming Distance.	**17/09/2019**		
38	Jaccard Distance, Jaccard Similarity of Sets, Similarity of Documents, Collaborative Filtering as a Similar-Sets Problem	**18/09/2019**		
39	**Clustering** - CURE Algorithm,	**19/09/2019**		PPT- **[TB1_4]/**
40	Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets, Answering Queries	**20/09/2019**		ClassRoom Teaching- **[TB1_4] /UT2/ Lab_Expt**
41	**Real-Time Big Data Models** PageRank Definition, Structure of the web, dead ends, Using Page rank in a search engine	**24/09/2019**		
42	Efficient computation of Page Rank, PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector.	**25/09/2019**		PPT **[TB1_4]**
43	Topic sensitive Page Rank, link Spam	**26/09/2019**		
44	Hubs and Authorities.	**27/09/2019**		
45	**Recommendation Systems** A Model for Recommendation Systems, Content-Based Recommendations,	**01/10/2019** **03/10/2019**		Case Study-Seminar **[TB1_4] /UT2**
46	Collaborative Filtering.	**04/10/2019**		
47	**Mining Social-Network Graphs** Social Networks as Graphs, Clustering of Social-Network Graphs	**09/10/2019** **10/10/2019**		PPT **[TB4]**, HB **/UT2, Lab Expt**
48	Direct Discovery of Communities, SimRank, Counting triangles using Map-Reduce	**11/10/2019**		PPT **[TB4]**

Total Lectures : 48

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 Mcary and Ann Kelly —Making Sense of NoSQL – A guide for managers and the rest of us, Manning Press.

References 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
3. Jared Dean, —Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Wiley India Private Limited, 2014.
4. Jiawei Han and Micheline Kamber, —Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 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 in Analyzing Unstructured Data, Cambridge University Press, 2006.
7. Vojislav Kecman, —Learning and Soft Computing, MIT Press, 2010.

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:

CSDLO7032.1: Understand the key issues in big data management and its associated applications for business decisions and strategy. [B1: Understanding]

CSDLO7032.2: Develop problem solving and critical thinking skills in fundamental enabling Techniques like Hadoop and MapReduce in big data analytics. [B3: Application]

CSDLO7032.3: Collect, manage, store, query and analyze various forms of Big Data[B4: Analysis]

CSDLO7032.4: Interpret business models and scientific computing paradigms, and apply software tools for big data analytics. [B3: Application]

CSDLO7032.5: Adapt adequate perspectives of big data analytics in various applications like Recommender systems, Social Media applications etc. [B3: Application]

CSDLO7032.6: Solve Complex real world problems in various applications like Recommender systems, Social media applications, health and medical systems, etc. [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
CSDLO7032.1	3												3	
CSDLO7032.2	3	3	3	1	3				2				3	3
CSDLO7032.3	3	3	3	2	3				2				3	3
CSDLO7032.4	3	3	3	2	3				2				3	3
CSDLO7032.5	3	3	3	2	3	3			2			2	3	3
CSDLO7032.6	3	3	3	2	3	3			3	3	2	3	3	3
TOTAL	12	12	12	9	12	6			11	3	2	5	12	9
CO-PO MATRIX	**3**	**3**	**3**	**2**	**3**	**2**			**2**	**3**	**2**	**3**	**3**	**3**

Course Outcomes Target: [Target 2.5]

Upon completion of this course students will be able to

CSDLO7032.1: Understand the key issues in big data management and its associated applications for business decisions and strategy.

CSDLO7032.2: Develop problem solving and critical thinking skills in fundamental enabling Techniques like Hadoop and MapReduce in big data analytics. [B3: Application]

CSDLO7032.3: Collect, manage, store, query and analyze various forms of Big Data. [B4: Analysis]

CSDLO7032.4: Interpret business models and scientific computing paradigms, and apply software tools for big data analytics. [B3: Application]

CSDLO7032.5: Adapt adequate perspectives of big data analytics in various applications like Recommender systems, Social Media applications etc. [B3: Application]

CSDLO7032.6: Solve Complex real world problems in various applications like Recommender systems, Social media applications, health and medical systems, etc. [B3: Application]

CO ASSESSMENT TOOLS

	Direct Methods (80%)						**Indirect Methods (20%)**
CSDLO7032.1	Assignment 1 (60%)		UE –P (20%)	UE-O (20%)			(100%)
CSDLO7032.2	Test1 (30%)	Lab 1-2-4-6 (30%)	UE –P (20%)	UE-O (20%)			(100%)
CSDLO7032.3	Test1 (30%)	Lab 2-3-5 (30%)	UE –P (20%)	UE-O (20%)			(100%)
CSDLO7032.4	Test2 (30%)	Lab 7-8-9 (30%)	UE –P (20%)	UE-O (20%)			(100%)
CSDLO7032.5	Test2 (30%)	Lab 10-11 (10%)	UE –P (10%)	UE-O (10%)	MP (20%)	Assig2 (20%)	(100%)
CSDLO7032.6	MP (50%)	Lab 10-11-12 (30%)	UE –P (10%)	UE-O (10%)			(100%)

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 : VII **Subject Code :*CSDLO8032***

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)
Efforts(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 : VII **Subject Code :*CSDLO8032***

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 : VII **Subject Code :*CSDLO8032***

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 D Tues	Batch C Wed	Batch B Thurs	CO Mapping	Title/aim
01	16 July	17 July	18 July	CO2	Hadoop HDFS Practical
02	23 July	24 July	25 July	CO2, CO3	Use of Sqoop tool to transfer data between Hadoop and relational database servers.
03	30 July	31 July	1 Aug	CO3	To install and configure MongoDB/ Cassandra/ HBase/ Hypertable to execute NoSQL commands.
04	06 Aug	07 Aug	08 Aug	CO2	Write a program to implement word count program using MapReduce /PySpark.
05	06 Aug	07 Aug	08 Aug	CO3	Matrix – Vector Multiplication using Map-reduce, Aggregate, Join, Search, Sort. (Map-Reduce/PySpark)
06	20 Aug	21 Aug	22 Aug	CO2	Create HIVE Database and Descriptive analytics-basic statistics, visualization using Hive/PIG/R.
07	27Aug	28 Aug	29 Aug	CO4	Implementing DGIM algorithm using any Programming Language/ Implement Bloom Filter using any programming language.
08	17 Sept	11 Sept	12 Sept	CO4	Implementing any one Clustering algorithm (*K*-Means/CURE) using Map-Reduce/PySpark
09	24 Sept	18 Sept	18 Sept	CO4	Streaming data analysis – use flume for data capture, HIVE/PYSpark for analysis of twitter data, chat data, weblog analysis (nehal)
10	01 Oct	25 Sept	26 Sept	CO5	Implement basic PageRank algorithm using Map-reduce/PySpark.
11	05 Oct	09 Oct	10 Oct	CO5	Implement predictive Analytics techniques (regression / time series, etc.) using R/ Scilab/ Tableau/ Rapid miner.
12				**CO6**	**Mini Project:** One real life large data application using standard dataset (Group of 2/3).
	16 July				**Topic Submission**
	06 Aug				**Progress review**
	17 Sept				**Presentation and Demo**
	10 Oct				**Mini Project Report submission**

Assignments Plan

Assignments			
01	16 July 2019	CO4	**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
02	06 August	CO3-4	Topic of Study
03	30 August	CO4 /PO12	Study of Research Paper

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?