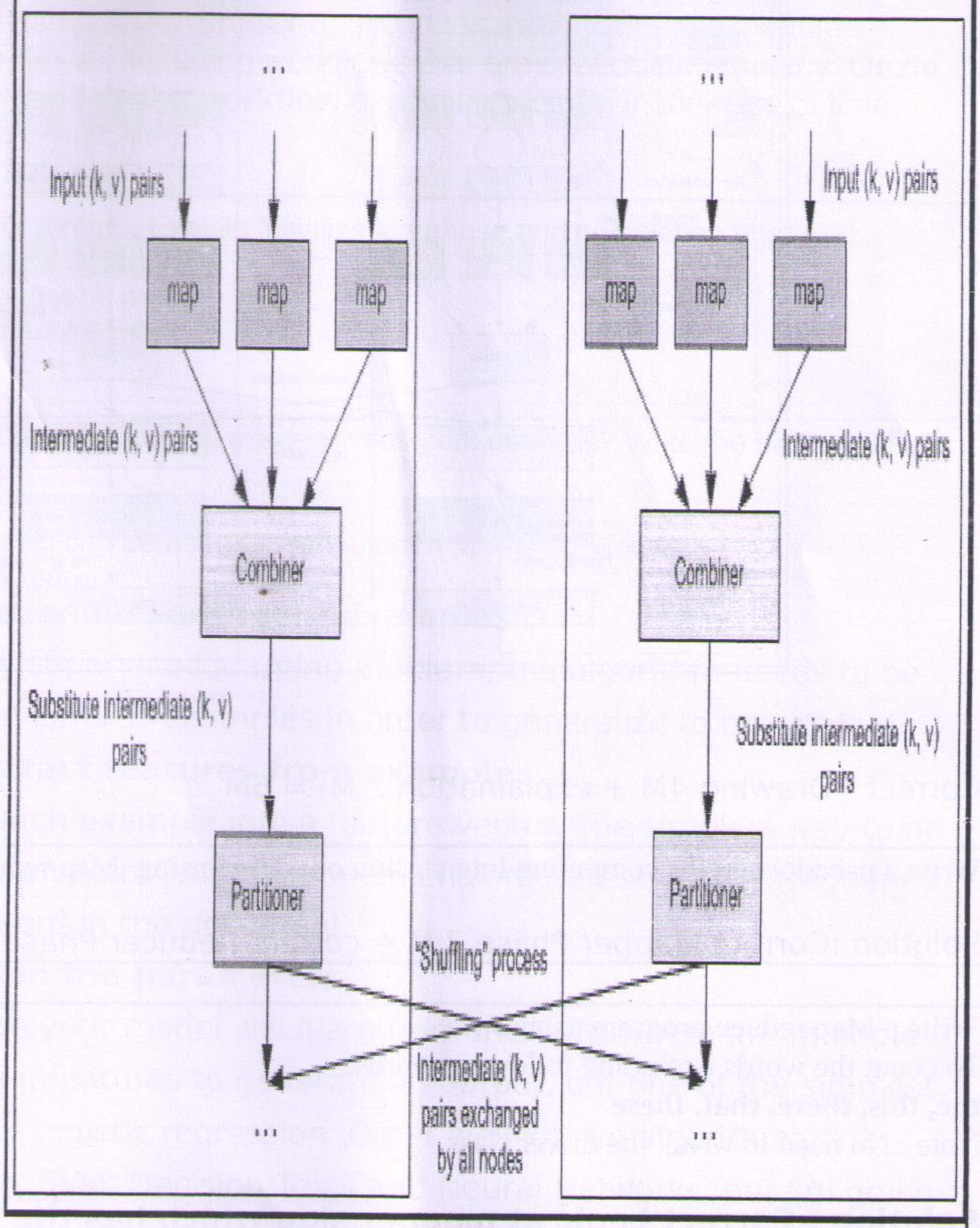
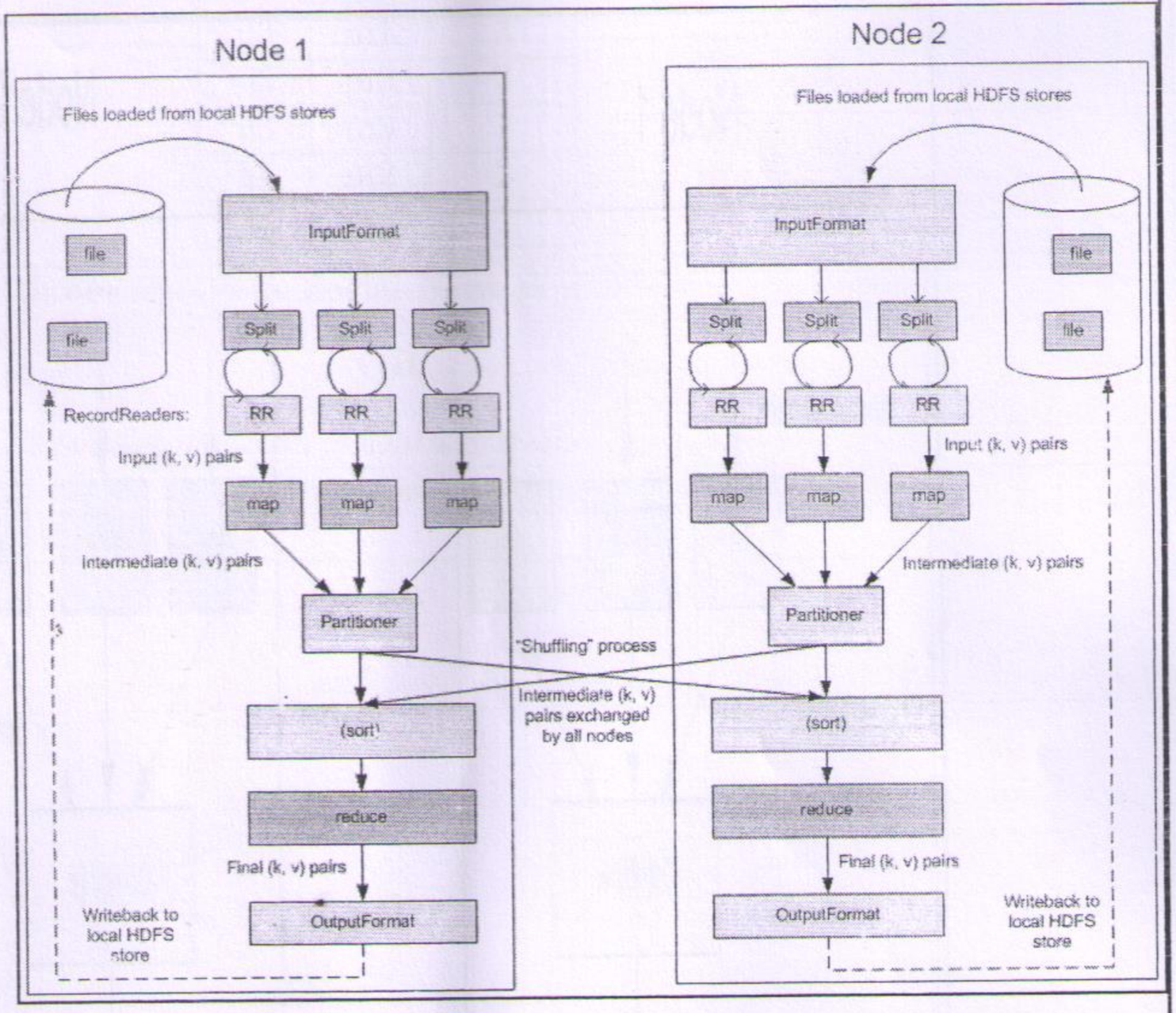
M. Tech - CMPN - I

Code : CE912
Name : Big Data Analytics and Management
Course Code : CE912 Exam Date 28/11/2019 10 – 1 pm

Sr No / Q. No.	Question	Marks
Q1 (a)	Jusify - Combiner can be viewed as mini-reducers in the map phase. Draw a scenario to explain combiner in Map and Reduce phase Solution :	6



Correct - Drawing 4M + explaination 2 M = 6M

	OR What is partitioning phase – Draw and explain need of partitioners in map reduce you can take any relevant scenarion and explain . Solution :  Correct - Drawing 4M + explaination 2 M = 6M	
Q1(b)	Write a pseudocode for computing Intersection operation using Map reduce Solution :Correct Mapper Phase 3 M + correct Reducer Phase 3M = 6M	6
Q 2 (a)	Write a Map reduce program using Java - To count the words excluding following words :- **the, this, there, that, these** Note : No need to write the driver class **Solution : Correct Logic at mapper side which has the condition of filtering given words** **If this logic is NOT written on mapper side then 2 marks will be deducted**	6

	Write a pseudocode for computing Difference operation using Map reduce Solution : Solution :Correct Mapper Phase 3 M + correct Reducer Phase 3M = 6M	6
Q 3 (a)	What is Reducer side join ? Write a pseudocode for the data set having following schema : Sch_Customer_sale(Tran_ID, Date, CustID, Amount, Equipment, city, state, mode) Solution : Correct explaination of Reducer side join = 2 Marks And Correct pseudocode = 4 Marks Total = 6 marks	6
Q 3(b)	What is the use of Oozie coordinator. Oozie Coordinator jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability. **Coordinator** applications allow users to schedule complex workflows, including workflows that are scheduled regularly. **Oozie Coordinator** models the **workflow** execution triggers in the form of time, data or event predicates. correct use: 6 marks	6
Q4a	Write CURE Algorithm. Explain initialization phase and completion phase. CURE Algorithm: 2mark initialization phase: 2 mark completion phase: 2 mark OR How amazon reviews can be used for sentiment analysis? Write the basic steps for it. amazon reviews can be used for sentiment analysis: 2mark basic steps: 4 mark **Step 1: Get some sentiment examples** As for every supervised learning problem, the algorithm needs to be trained from labeled examples in order to generalize to new data. **Step 2: Extract features from examples** Transform each example into a feature vector. The simplest way to do it is to have a vector where each dimension represents the frequency of a given word in the document. **Step 3: Train the parameters** This is where your model will learn from the data. There are multiple ways of using features to generate an output, but one of the simplest algorithms is logistic regression. Other well-known algorithms are Naive Bayes, SVM, Decision Trees and Neural Networks, but I'm going to use logistic regression as an example here. In the simplest form, each feature will be associated with a weight. Let's say the word "love" has a weight equal to +4, "hate" is -10, "the" is 0 ... For a given example, the weights corresponding to the	6

features will be summed, and it will be considered "positive" if the total is > 0, "negative" otherwise. model will then try to find the optimal set of weights to maximize the number of examples in our data that are predicted correctly.

If you have more than 2 output classes, for example if you want to classify between "positive", "neutral" and "negative", each feature will have as many weights as there are classes, and the class with the highest weighted feature sum wins.

Step 4: Test the model

After we have trained the parameters to fit the training data, we have to make sure our model generalizes to new data, because it's really easy to overfit. The general way of regularizing the model is to prevent parameters from having extreme values.

Q4b

Discuss how item based collaborative filtering can be applied for the example given. — 8

ID	user	item	rating
241	u1	m1	2
222	u1	m3	3
276	u2	m1	5
273	u2	m2	2
200	u3	m1	3
229	u3	m2	3
231	u3	m3	1
239	u4	m2	2
286	u4	m3	2

Step1. finding the matrix 1 mark
Step2.forming the vectors 1 mark
Step3. find the distance measures and similarity 4 mark

	m1	m2	m3
m1	1	0.97	0.79
m2	0.97	1	0.90
m3	0.79	0.90	1

Step4: Recommendation rank 2 mark

Q5a) — 3

Write any three varieties in Social Network graph.

Collaboration graph, Information Linkage Graph, Heterogeneous Social network etc- 1 mark each

	OR Write any three applications of Data Streams. Each application -1 mark	
5b)	Why are dead-ends and spider traps a problem and why do teleports solve the problem? dead-ends and spider traps a problem: 4 marks teleports solve the problem: 3 marks	7