BIG DATA ANALYTICS

Course Code : SWE2011 Slot : C_2

DIGITAL ASSIGNMENT - 1

Name : S. Deepan
Reg No : 19MIS0102

Explain the parallel implementation of apriori Algorithm based on the mapReduce with an Example and diagram.

The Apriori algorithm is one of the typical algorithm which is a seminal algorithm

MapReduce is a programming paradigm that runs in the background of Hadoop to provide Scalability and easy data-processing Solutions.

The Map task takes a set of data and converts it into another set of data,

- where individual elements are broken down into the tuples (key - value pairs).

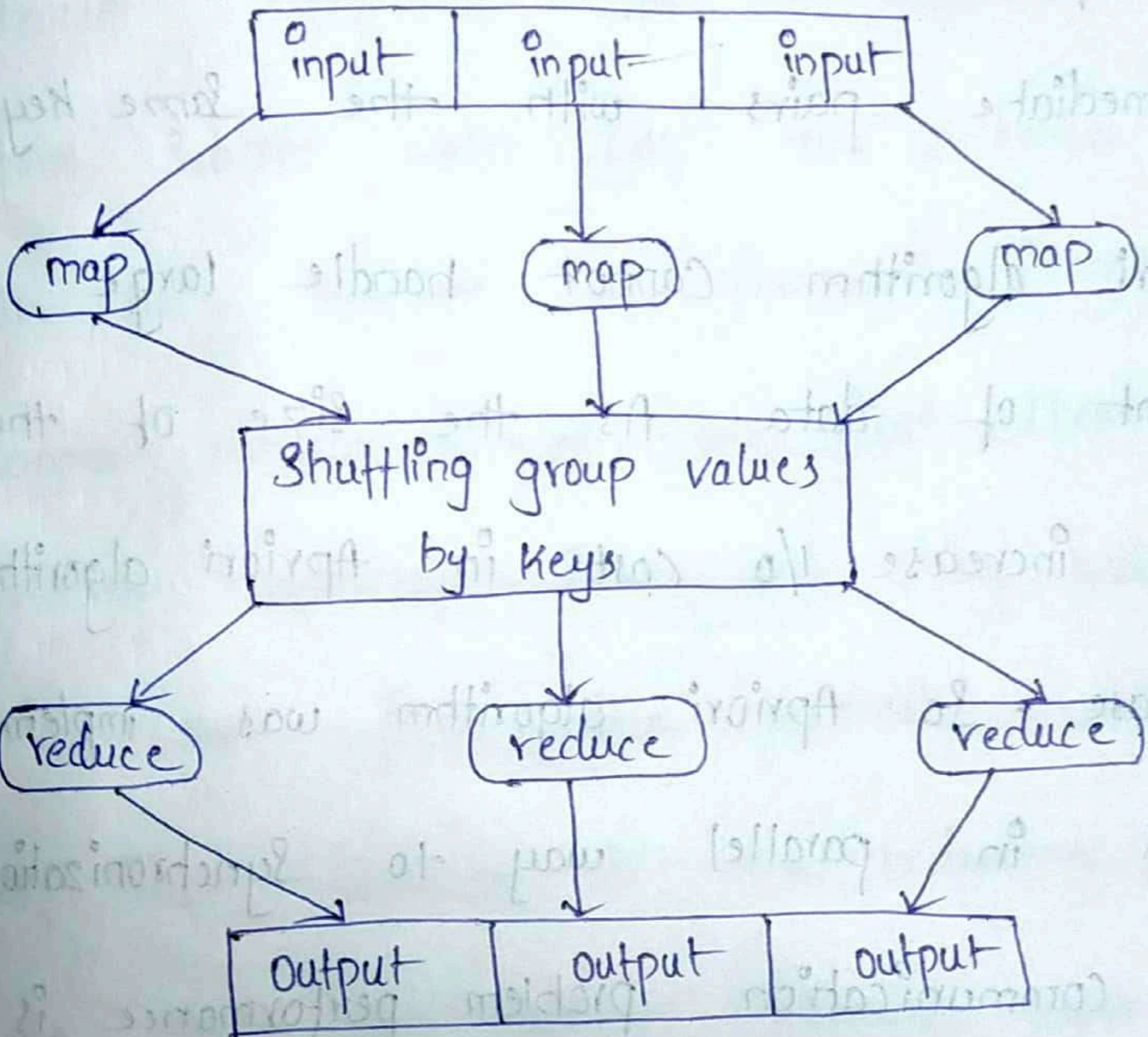
The Reduce task takes the output from the Map as an input and combines those data tuples (key - value pairs) into a smaller set of tuples.

- Some data preprocessing, clustering and classification algorithms have been implemented based on MapReduce

- The name of the Apriori Algoritm is based on the fact that the algorithm uses prior knowledge of frequent itemset property which is that all nonempty subsets of a frequent itemset must also be frequent.

MapReduce specifies the computation in

terms of a map and a reduce function, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks.



Map takes an input pair and produces a set of intermediate key/value pairs.

- The MapReduce library groups together all the intermediate values allocated with the same intermediate key and passes them to the reduce function. if could be normalize as map:: (key 1, value 1) list (key 2, value 2).

- MapReduce function can be Executed in the parallel on each set of the intermediate pairs with the same key.

- Apriori algorithm cannot handle large amount of data. As the size of the data increase. I/O cost in Apriori algorithm Increase. So Apriori algorithm was implemented in parallel way to Synchronization and communication problem performance is not good in case of large amount of data or when we use in big data

The Apriori algorithm is implemented in MapReduce framework. There are two steps in Apriori algorithm, one is candidate generation that finds the frequent itemsets and add them to the candidate sets. Second one is the count step, in this step all the candidate itemset compared with the minimum support. them subsets which fulfil the criteria can be selected as frequent itemsets.

- Mapper performed at the first step by dividing datasets into the key value pair, and finds the potential candidate set. Then reducer do the reducing part here which set quality minimum support. Such candidate will be selected as frequent item sets.

- In the implementation of the parallel Apriori algorithm in mapreduce is the most established algorithm for finding frequent item sets from a transactional dataset ; it needs to scan the dataset many times and to generate many candidate item sets. But when the dataset size is huge, both memory use and computational cost can still be very Expensive.

- The Apriori algorithm needs one kind of MapReduce. The map function performs the procedure of Counting each occurence of potential candidate of size K and the map stage realize the occurrences, counting for all the potential candidate in a parallel way. Then the reduce function

performs the procedure of summing the occurences counts.

The main aim of the Apriori algorithm in parallel using map reduce is to use the apriori algorithm which is a data mining algorithm along with the MapReduce

- This is mainly used to find the frequent item sets for a application which consists of various transaction. By using these algorithm we will take the inputs from the database sets present in the application and the output is given as frequent item sets.

Example :

The below table shows the transaction data of a store and to Explain the

parallel implementation of Apriori map reduce algorithm.

S.NO	Items	Date	ID
1	Coke, Milk	3-2	1
2	pizza	4-2	2
3	coke, pizza	5-2	3
4	Milk	6-2	4
5	Cracker, Milk	7-2	5

- let us assume three Map nodes, two transaction data are distributed to three map nodes

Map node	item, count
C_{11}	<cracke 1>, <coke, 1>
C_{12}	<coke, 2>, <Milk, 2>
C_{13}	<pizza 1>, <coke 1>

So from all the Map node reduce the nodes collect and compute C_1 that is the size and L_1 is size frequent item pairs that meets minimum support.

C_1 = { <cracker 1>, <coke 4>, <Milk, 2>,
<pizza 1> }

L_1 = { <coke 4>, <Milk 2>, <pizza 1> }

The item sets L_1 and L_2 can be used to produce association rule of the transaction by Apriori algorithm using map reduce

L_2 = {[<pizza, cracker>, 1],]

- The Apriori algorith uses a layer-by-layer iterative search method to count the support of each item sets by scanning the dataset. and also uses to find frequent itemsets

- So it is an iterative approach process and its two main components are candiclate itemsets generation and the frequent itemsets generation. The count distribution parallel version of apriori is best

to implement data distribution automatically

Example :

- let us consider in a departmental store there are list of items for example

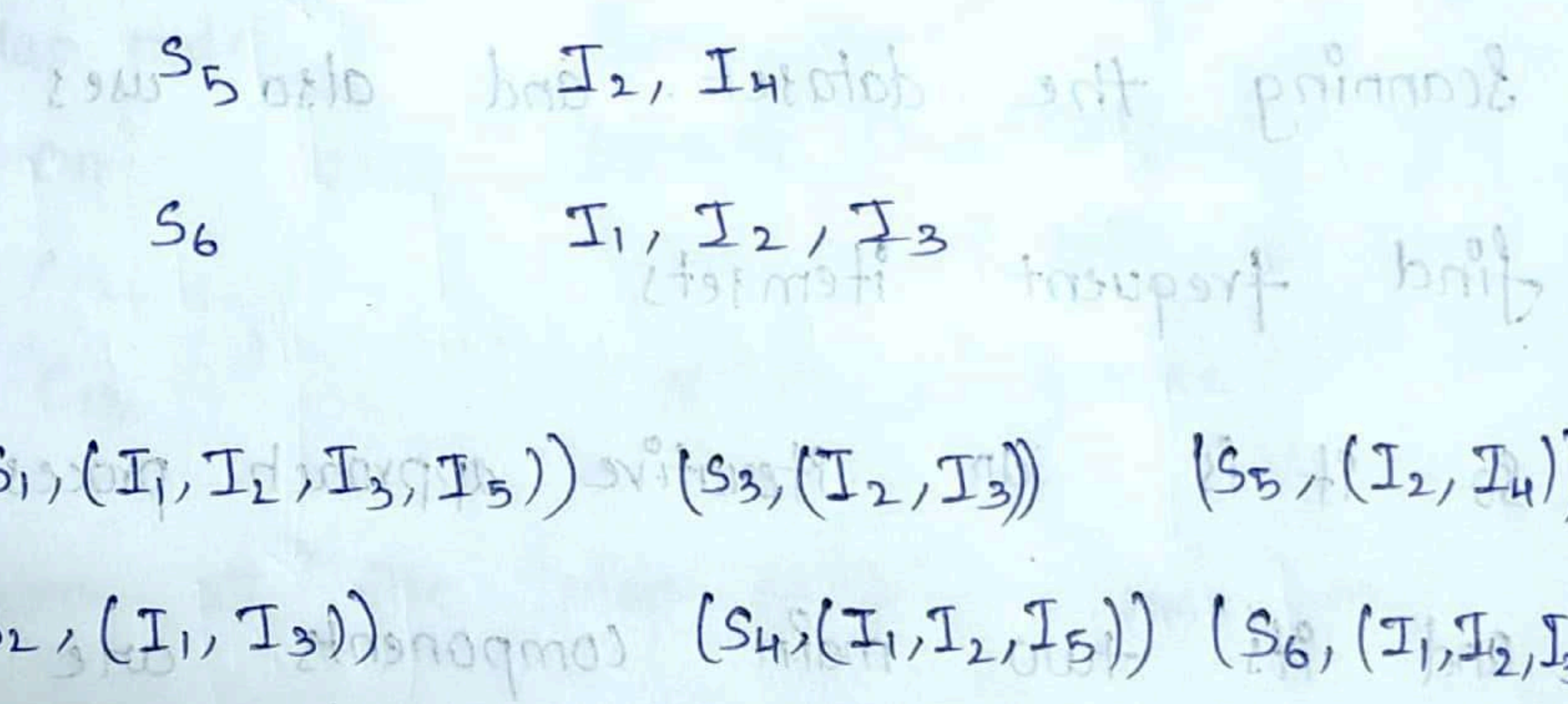
Dal → I_1, Rice → I_2, Sugar → I_3,

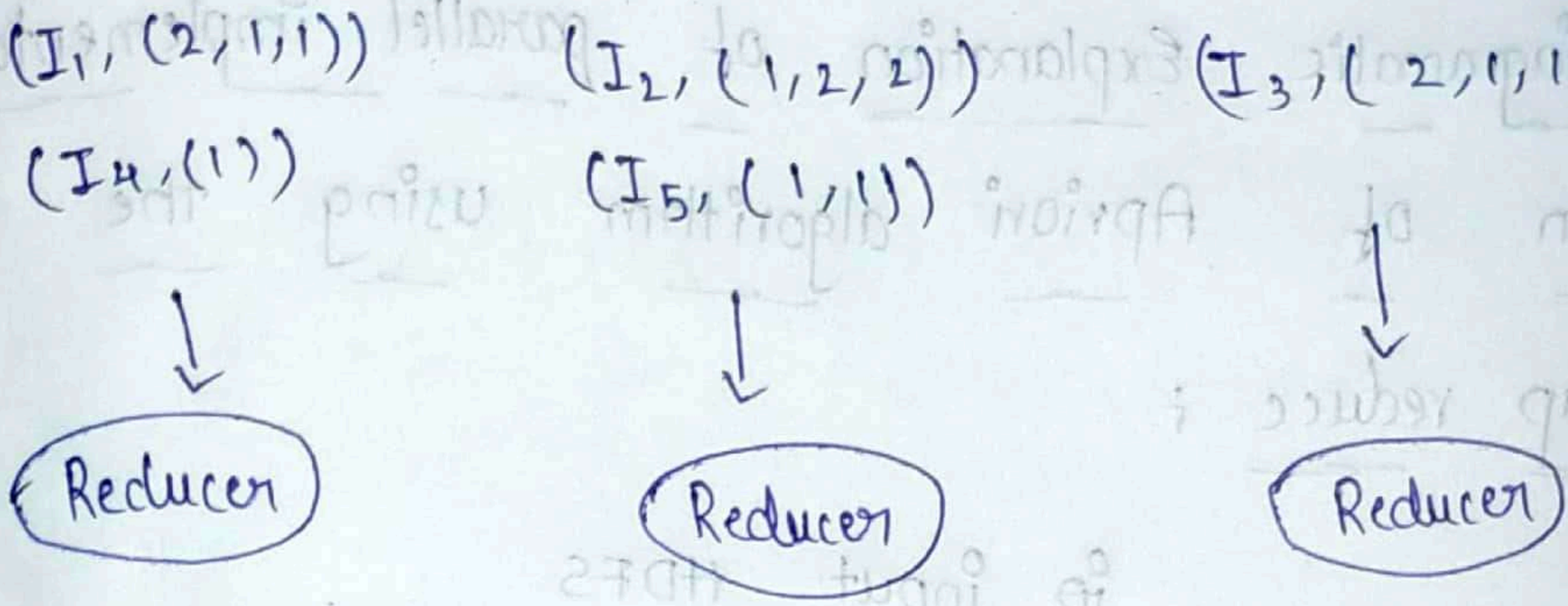
Salt → I_4, chilli powder → I_5

S.NO	Item list
S_1	I_1, I_2, I_3, I_5
S_2	I_1, I_3
S_3	I_2, I_3
S_4	I_1, I_2, I_5
S_5	I_2, I_4
S_6	I_1, I_2, I_3

$(S_1, (I_1, I_2, I_3, I_5))$ $(S_3, (I_2, I_3))$ $(S_5, (I_2, I_4))$

$(S_2, (I_1, I_3))$ $(S_4, (I_1, I_2, I_5))$ $(S_6, (I_1, I_2, I_3))$





So it basically collect data items the moves to intermediate values → Shuffle and output ← Exchange

- Examples for the Apriori algorithm based on parallel implementation using map-reduce.

- Crime detection and prediction - To analyse the crime in cities and urban areas

- Crowd mining : finding information from the social data to achieve better behavior of the residents

Diagramatic Explanation of parallel implementation of Apriori algorithm using the map reduce :