BU.330.740. Large Scale Computing with Hadoop

Assignment 1: Frequent Itemset Mining Using MapReduce

Learning Goal: using MapReduce framework to implement frequent doubleton itemsets

Input Data:

The original data is stored in transaction.dat. Each line is a transaction containing multiple items separated by space (item1 item2 item3 · · · itemn)

Output results:

Set *support* threshold *s* = 2, which means the output should be all doubletons that appear at least two times in the transactions.

Implementation:

The mapper class
Your map function would take this original file and generate an intermediate output. The input key would be line number in input file. The input value would be the content in each line. The output key would be the doubleton itemsets. The output value is 1. You can use *tab* to separate the key and the value.

For example:

1 2 3 ⇒

(1,2)   1

(1,3)   1

(2,3)   1

2 3 ⇒

(2,3)   1

The reducer class
Your reduce function would aggregate all values for each key. The output key would be itemsets. The output value is the number of occurrence of each corresponding key. In the example case, it will generate the following outputs:

(2,3)   2

Test Locally (optional):

You can test your scripts locally before deploying on AWS using the technics introduced in Lab3.

Submission:

You need to submit one zipped file, containing all .py scripts and output results (using AWS Hadoop cluster), and submit it through the blackboard. Please try to comment your codes for critical sections and make your codes as readable as possible.

Extension Materials:

There are frequent itemset mining dataset repository on http://www.cs.rpi.edu/~zaki/Workshops/FIMI/data/, and some extra resources on the website for you to explore after you finish with assignment 1.