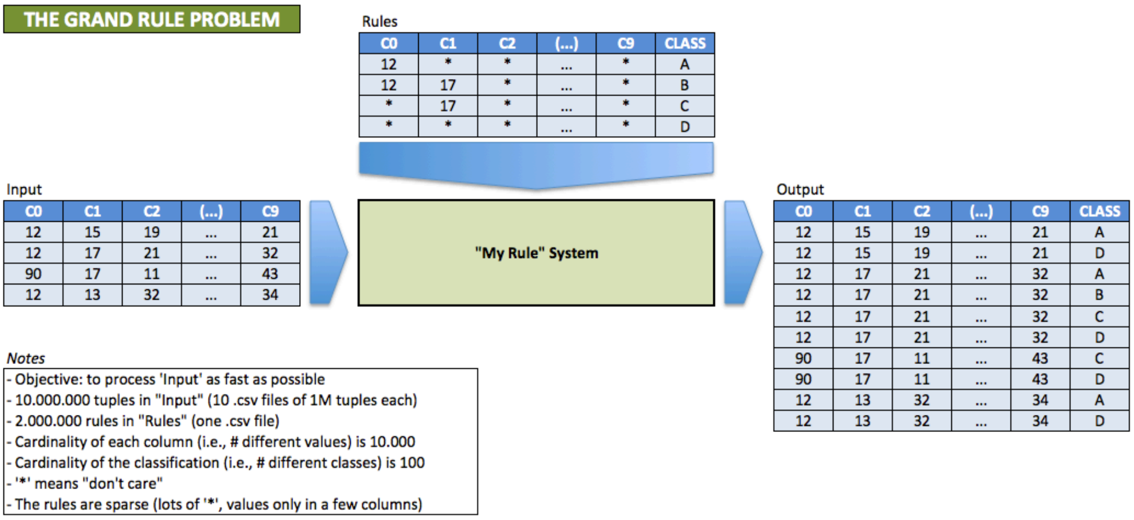
# The Grand Rule Problem of 2011 - GPGPU

Recently you have been approach by an industrial partner (a bank) that wants to classify the expenses of its clients according to a number of rules. In practice, for each bank transaction of an input file, they want to see if it matches a number of predefined rules. A bank transaction is a tuple with 10 attributes – $(a_0, a_1, a_2, a_3, \ldots, a_9)$, where each attribute is an integer (32 bit). Each rule is also a tuple with 10 attributes plus a classification "c". I.e., $(r_0, r_1, r_2, r_3, \ldots, r_9, c)$. Each rule attribute can either be a number (32 bit), or a "*", which means "don't care". For example, the transaction (1,2,3,4,5,6,7,8,9,0) matches the rule (1,2,3,4,5,6,7,8,9,0,111) and the rule (1,*,*,*,*,*,*,*,0,222), but it does not match either the rules (2,2,3,4,5,6,7,8,9,0,333) or (2,*,*,*,*,*,*,*,0,444).

Your challenge is to implement a system that allows efficiently solving this problem for large amounts of data. The next image illustrates the problem.



Rules

| C0 | C1 | C2 | (...) | C9 | CLASS |
|---|---|---|---|---|---|
| 12 | * | * | ... | * | A |
| 12 | 17 | * | ... | * | B |
| * | 17 | * | ... | * | C |
| * | * | * | ... | * | D |

Input

| C0 | C1 | C2 | (...) | C9 |
|---|---|---|---|---|
| 12 | 15 | 19 | ... | 21 |
| 12 | 17 | 21 | ... | 32 |
| 90 | 17 | 11 | ... | 43 |
| 12 | 13 | 32 | ... | 34 |

Output

| C0 | C1 | C2 | (...) | C9 | CLASS |
|---|---|---|---|---|---|
| 12 | 15 | 19 | ... | 21 | A |
| 12 | 15 | 19 | ... | 21 | D |
| 12 | 17 | 21 | ... | 32 | A |
| 12 | 17 | 21 | ... | 32 | B |
| 12 | 17 | 21 | ... | 32 | C |
| 12 | 17 | 21 | ... | 32 | D |
| 90 | 17 | 11 | ... | 43 | C |
| 90 | 17 | 11 | ... | 43 | D |
| 12 | 13 | 32 | ... | 34 | A |
| 12 | 13 | 32 | ... | 34 | D |

*Notes*

- Objective: to process 'Input' as fast as possible
- 10.000.000 tuples in "Input" (10 .csv files of 1M tuples each)
- 2.000.000 rules in "Rules" (one .csv file)
- Cardinality of each column (i.e., # different values) is 10.000
- Cardinality of the classification (i.e., # different classes) is 100
- '*' means "don't care"
- The rules are sparse (lots of '*', values only in a few columns)

In particular, the objective is to process 10 input files containing the transactions of the last 10 days. Each input file has 1M tuples. The rule table is fixed and has 2M rules. Most of the rules are actually empty (i.e., with "*"), having numbers in only a few columns of each rule. The number of unique values in each column is 10.000 and the number of different classifications is 100. Your solution should process the data as efficiently as possible, generating an output file for each input one.

You will be given access to a test data set and the complete dataset for the project. One piece of advice: try to implement a good algorithm before making it parallel. As an indication, you should be able to process, at least, 6.000 transactions/sec on a Core2-Duo laptop, running at 2.5GHz with 4Gb of RAM (without using any GPU, just the CPU). Doing twice that is readily attainable.

At http://sse.dei.uc.pt/grand_rule/dataset.tgz you will find the test dataset. In particular, there will be three directories there, which you can use for tests:

- TINY_INPUT → Contains a very small dataset (input transactions, rules, expected output)
- SMALL_INPUT → A small dataset (input transactions, rules, expected output)
- THE_PROBLEM → The actual problem to solve (10 files with input transactions; a rule file with 2M rules; the expected output for each file – 10 files, sorted for easy comparison with your output)

Good Luck!

PS:
This file is available at http://sse.dei.uc.pt/grand_rule/grand_rule.pdf