|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Disciplines | N | D | K | M | File Size | single CPU | CPU cluster | Time GPU |
| Flow Cytometry | 10^6 | 24 | 100s | 100 | 146MB | 281 sec with 12 cores |  | 9.4 sec using 1 GPU |
| Forest Cover Type | 581012 | 54 | 7 | 100 | 191MB | 30.4 sec with 12 cores |  | 1.1sec using 1 GPU |
| Census-Income Data | 299285 | 40 | 10s | 100 | 79 MB | 15.3 sec with 12 cores |  | 5.6 sec using 1 GPU |
| YahooEig | 1.4 billion | 6 | 100s | Unknow | 0.2TB | Very long due to memory swapping | 8 minutes with 128 cores with  MapReduce | Cannot fit into GPU memory (6GB) |
| Quantum Physics Dataset | 100,000 | 78 | 2 | 100 | 47.5M | 1.93 sec with 12 cores |  | 0.16 sec using 1 GPU |

Typical Disciplines using C-means Classification Algorithms

Dataset References:

1. Flow Cytometry Data Set: <http://flowrepository.org/>
2. Forest Cover Type Data Set: <http://archive.ics.uci.edu/ml/datasets/Covertype>
3. Census Income Data Set: <http://archive.ics.uci.edu/ml/datasets/Census+Income>
4. Top 6 eigenvector of adjacency matrix of web graph crawled Yahoo: <http://www.yahoo.com>
5. Quantum Physics Dataset: <http://osmot.cs.cornell.edu/kddcup/>

Paper References:

1. Scalable Data Clustering using GPU Clusters
2. Clustering Billions of Data Points Using GPUs
3. Speedup of Fuzzy clustering through stream processing on graphics processing units.
4. A Data-Clustering Algorithm On Distributed Memory Multiprocessors
5. Speedup of Fuzzy and Possibilistic Kernel and c-Means for Large-Scale Clustering
6. Parallel Fuzzy c-Means Clustering for Large Data Sets