European Week of Astronomy and Space Science

Enhanced SOM distributed processing for the classification of large spectroscopic data in the Gaia mission

Marco Antonio Álvarez González
marco.antonio.agonzalez@udc.es

Computer Science Faculty, A Coruña, Spain

The idea

The Bigdata phenomenon makes the neccesity of computational power a reality.

In our institutions we dispose of a lot of machines available and idle most part of the time. During night or weekends, for example.

Configure a flexible cluster of Spark, being capable of manage a fully dynamic number of nodes, promoving the massive resource utilization.

The system

Over the last months we have been developing an automated system, called **SparkFlex**, which allows us to add machines to our cluster in execution time, contributing with the running tasks at that moment.

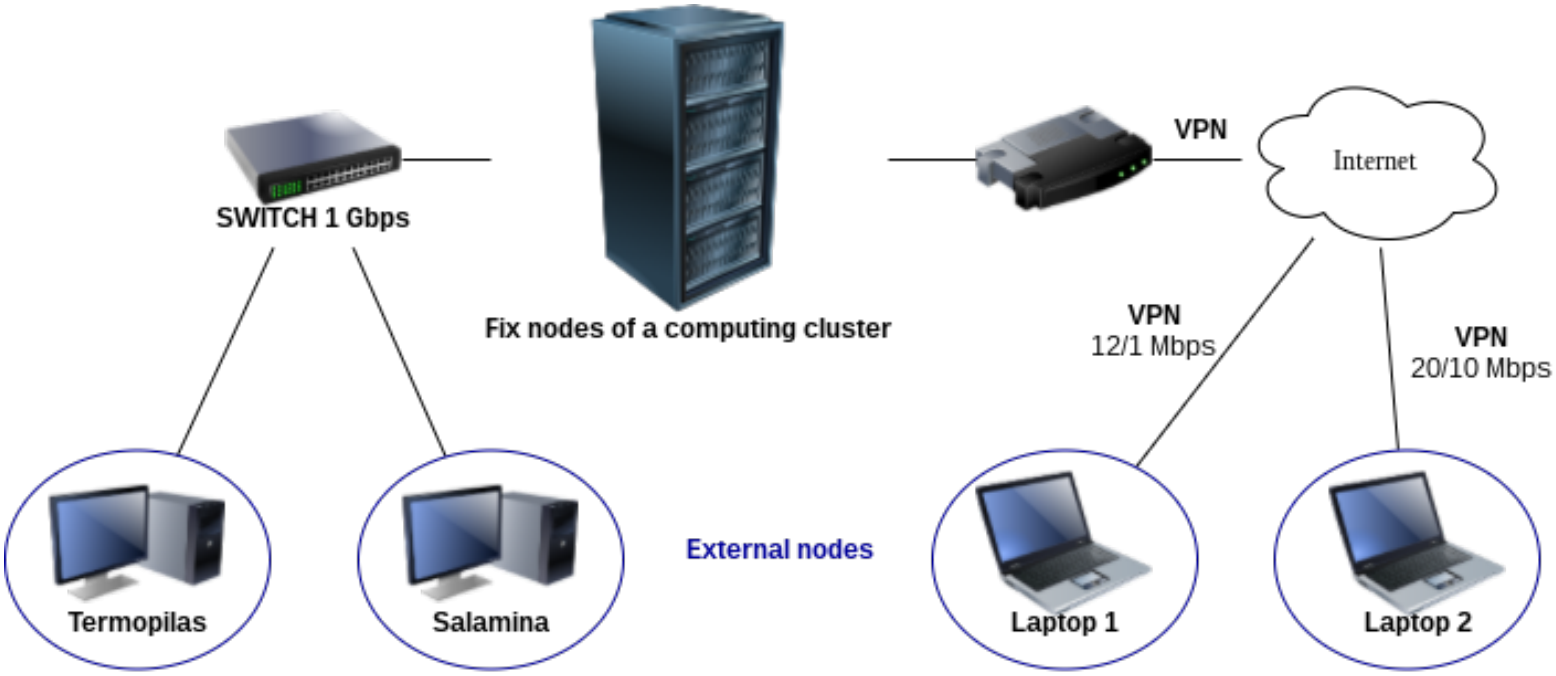
We carried out several tests over the system, using different configurations and different file distributions (HDFS, NFS, Local...), and finally we build a Virtual Machine, which is prepared to be used by any one.

How it works

Any user could contribute to a single task with these simple steps:

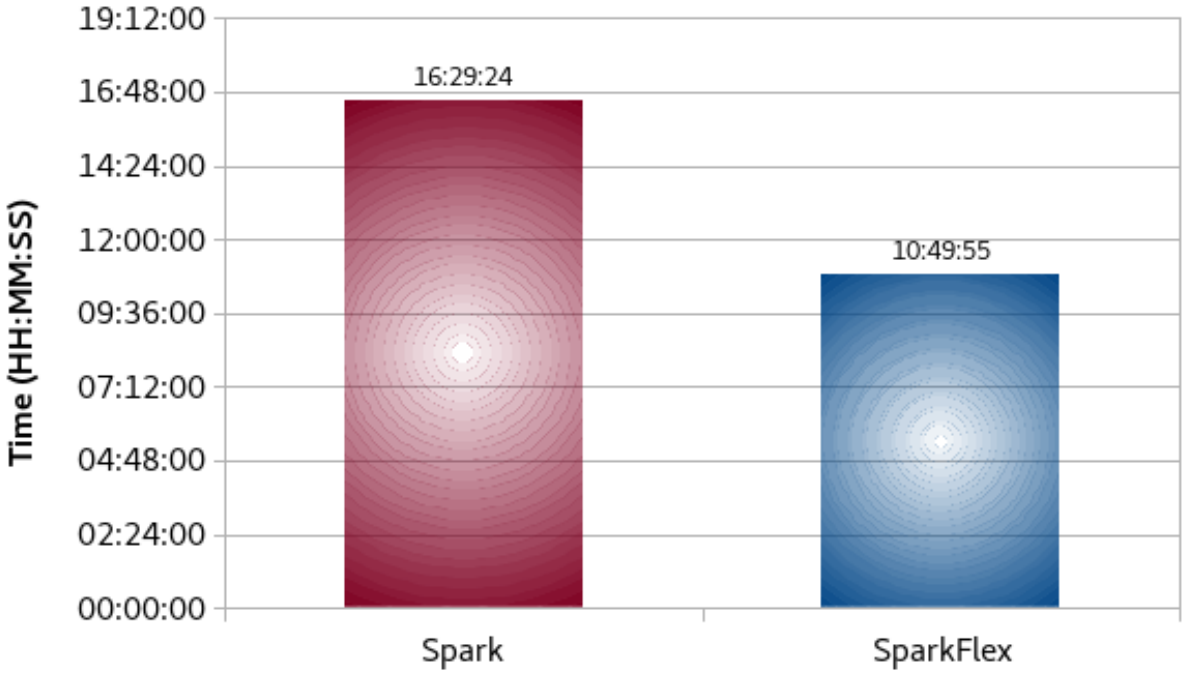
1. Download a Virtual Machine with SparkFlex installed
2. Assign the resources available in the computer to the Virtual Machine, as many as wanted
3. Start the Virtual Machine, it automatically connects and joins to the cluster

Environment



Test environment

Results



SOM training with 1 Million elements

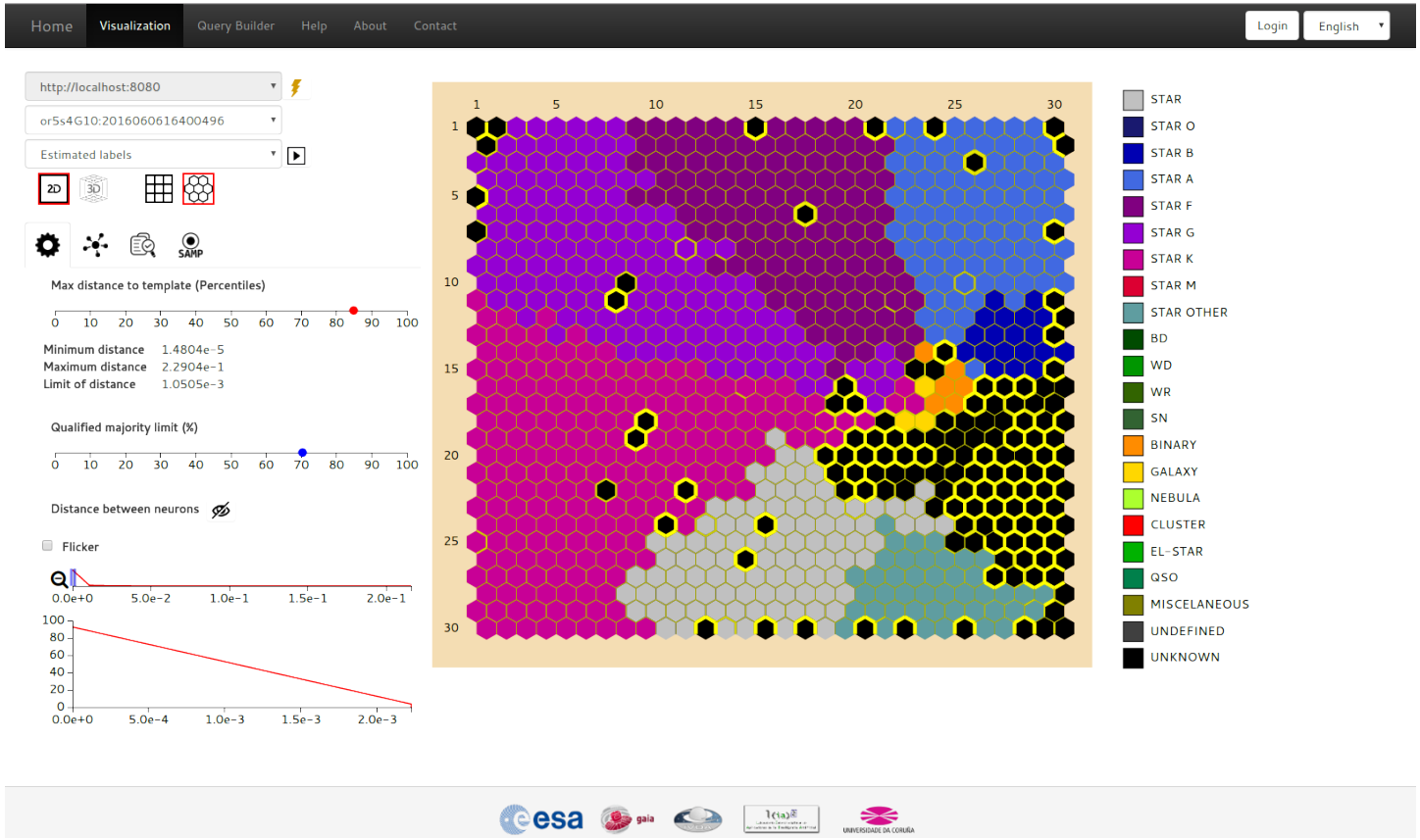
Spark: 2 nodes with 56 cores and 92 GiB of RAM in total
SparkFlex: Spark + 2 external nodes (15 GiB and 6 cores)

Issues

- There are minimum requirements to execute the Virtual Machine in a dynamic node. In our case, 2.5 GiB of RAM and 1 core
- The bandwidth is an important factor... slow connections result in timeouts with a high traffic of data in HDFS

GUASOM

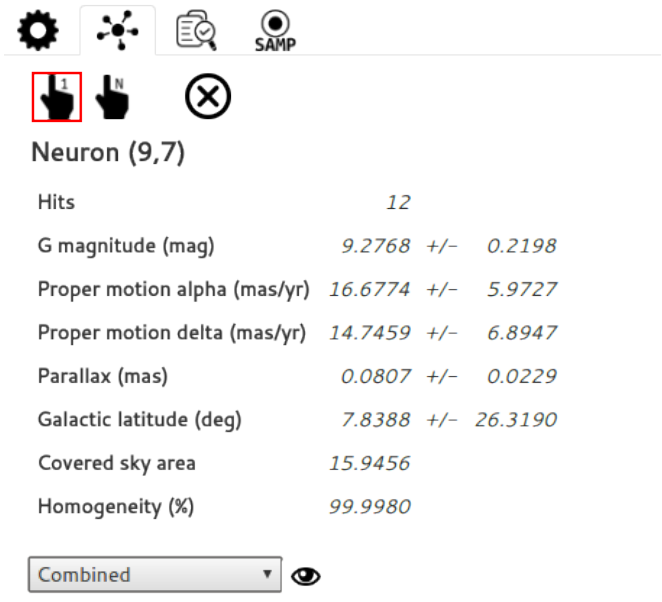
Web visualization tool aimed to analyze Self Organizing Maps.



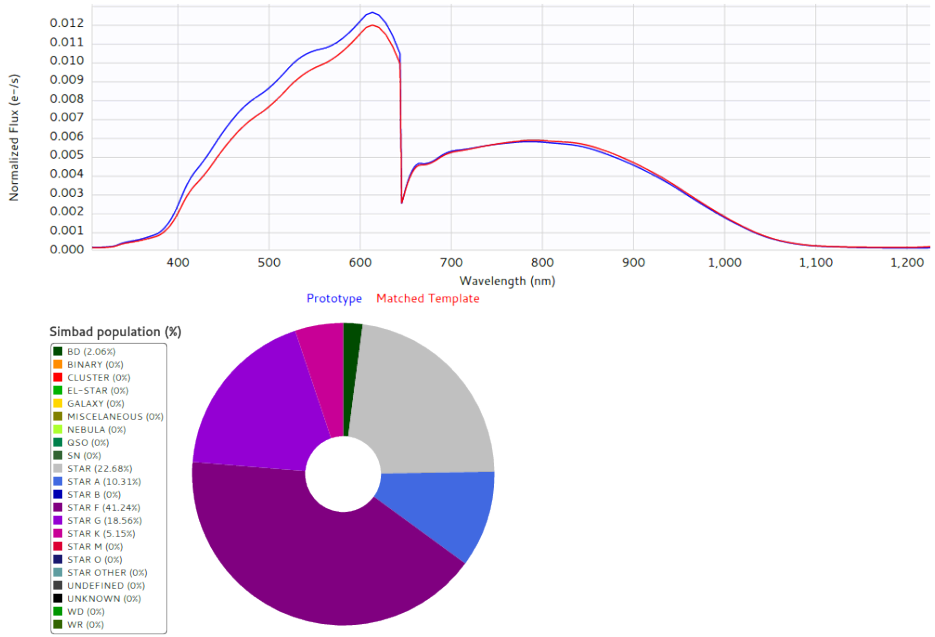
Representative label for each neuron

GUASOM

Features like neuron information, SAMP communication or crossmatch with external catalogues are available.



(d) Astrometry information



(e) Spectra and population

Neuron Information

Conclusions

- The dynamic resource assignment has demonstrated its utility in extreme intensive computing
- We can make the most of our idle computers, combining them in the cluster in an easy way
- We develop a practical demonstration that it is possible to configure a Spark cluster to hot-plug machines on it (even in running jobs)
- A preconfigured Virtual Machine is a portable and platform independent technique to achieve the purpose
- Additionally we have develop a web visualization tool for our use case (SOM)

References

Marco Antonio Álvarez, Carlos Dafonte, Daniel Garabato, and Minia Manteiga.

Analysis and Knowledge Discovery by Means of Self-Organizing Maps for Gaia Data Releases, pages 137–144.

Springer International Publishing, 2016.

Diego Fustes, Carlos Dafonte, Bernardino Arcay, Minia Manteiga, Kester Smith, Antonella Vallenari, and Xavier Luri.

SOM ensemble for unsupervised outlier analysis. Application to outlier identification in the Gaia astronomical survey.

Expert Systems with Applications, 40(5):1530–1541, 2013.

Daniel Garabato, Carlos Dafonte, Minia Manteiga, Diego Fustes, Marco A. Alvarez, and Bernardino Arcay.

A distributed learning algorithm for self-organizing maps intended for outlier analysis in the GAIA – ESA mission.

Atlantis Press, 2015.