D4Science Infrastructure - Task #117

Execute FAO MSY on the complete FAO Dataset

May 18, 2015 04:11 PM - Gianpaolo Coro

Status:	Closed	**Start date:**	Jun 11, 2015
Priority:	Normal	**Due date:**	
Assignee:	Gianpaolo Coro	**% Done:**	100%
Category:	High-Throughput-Computing	**Estimated time:**	0.00 hour
Target version:	CommunitySupport		
Infrastructure:	Production		

Description

FAO MSY needs to be executed on 5500 species. This requires empowering the production Cloud processing environment.

History

#1 - May 20, 2015 11:13 AM - Luca Frosini

- Target version changed from 29 to zz - UnSprintable

#2 - May 20, 2015 11:13 AM - Luca Frosini

- Target version changed from zz - UnSprintable to 29

#3 - May 26, 2015 12:51 PM - Pasquale Pagano

- Tracker changed from Bug to Task

- Project changed from 2 to D4Science Infrastructure

- Category set to High-Throughput-Computing

- Target version changed from 29 to CommunitySupport

- Start date changed from May 18, 2015 to Jun 11, 2015

- Infrastructure Production added

#4 - May 27, 2015 02:18 PM - Gianpaolo Coro

- Status changed from New to In Progress

- % Done changed from 0 to 10

Tests have started for this huge computation.
I expect the computation time to be exponential descendant.
A linear estimate of the computation time is 4 days.

#5 - May 29, 2015 03:15 PM - Gianpaolo Coro

- % Done changed from 10 to 70

As it happened also in other cases (e.g. the Length-Weight algorithm), the effect of parallelising an R script is to reduce the computational time more than linearly.
With the latest input provided by FAO, the computation time of the sequential run is around **30 days**.
Using 60 nodes, instead, the lower usage of memory and disk has the effect to reduce the computation time to **15h and 20 minutes**.

Thus, the time reduction with respect to the sequential case is 97.8%

I have run the computation two times to double-check the time and the output.

The execution produces the following output files:
Main output: http://goo.gl/bJ1ZRx
Auxiliary output: http://goo.gl/slPIwA

In the list of the 5565 input species, there are 49 species on which the script crashes. This requires further investigation by FAO, in order to produce a patch or to discard these species.
The list of the 49 species records is here: http://goo.gl/kAPPaA

I will update this ticket as soon as FAO will have answered on how to proceed for the species without output.

#6 - Jun 10, 2015 11:26 AM - Gianpaolo Coro

- Status changed from In Progress to Closed

- % Done changed from 70 to 100

The results have been sent to Yimin Ye of FAO.