

D4Science Infrastructure - Task #1215

Install Dataminer in production environment

Oct 22, 2015 03:24 PM - Gianpaolo Coro

| | | | |
|--|---|------------------------|---------------------|
| Status: | Closed | Start date: | Oct 22, 2015 |
| Priority: | High | Due date: | |
| Assignee: | Andrea Dell'Amico | % Done: | 100% |
| Category: | High-Throughput-Computing | Estimated time: | 0.00 hour |
| Target version: | D4Science Infrastructure Upgrade to gCube 3.9.0 | | |
| Infrastructure: | Production | | |
| Description An installation of Dataminer in the production environment is required. The installation should follow the guidelines at this wiki page: https://gcube.wiki.gcube-system.org/gcube/DataMiner_Installation and the instructions in the tickets related to this ticket. Furthermore, apart from the machine for the single instance, another machine is required, with an Apache service installed, to act as proxy for DataMiner. This will be used to balance the requests to several instances that will be running in D4Science. The name of this balancer should be "dataminer.d4science.org". The Apache service should answer on the 80 port, but also each Dataminer instance should be available on that port. | | | |
| Related issues: | | | |
| Related to D4Science Infrastructure - Task #422: Report libraries and configur... | | Closed | Jul 23, 2015 |
| Related to D4Science Infrastructure - Task #421: Create an automatic installa... | | Closed | Jul 23, 2015 |
| Related to D4Science Infrastructure - Task #1502: Automate the dataminer inst... | | Closed | Nov 23, 2015 |
| Related to D4Science Infrastructure - Support #1837: Add SmartGenericWorkers ... | | Closed | Dec 17, 2015 |

History

#1 - Oct 22, 2015 03:25 PM - Gianpaolo Coro

- Related to Task #422: Report libraries and configuration components for DataMiner added

#2 - Oct 22, 2015 03:25 PM - Gianpaolo Coro

- Related to Task #421: Create an automatic installation package for DataMiner added

#3 - Nov 23, 2015 04:08 PM - Andrea Dell'Amico

- Related to Task #1502: Automate the dataminer installation as a loadbalanced service added

#4 - Dec 11, 2015 05:13 PM - Andrea Dell'Amico

- % Done changed from 0 to 20

Following the dev provisioning, I've created the ansible variable groups needed to install a dataminer cluster in production.

#5 - Dec 15, 2015 11:53 AM - Gianpaolo Coro

@andrea.dellamico@isti.cnr.it @tommaso.piccioli@isti.cnr.it Who is responsible for creating the machines for the production environment and run the installation scripts?

#6 - Dec 15, 2015 07:02 PM - Andrea Dell'Amico

One load balancer and two dataminer servers also in production. Other dataminer servers can be added at a later time

Required specs for the load balancer:

- Hostname: dataminer-lb1-p-d4s.d4science.org (with dataminer.d4science.org as alias)
- Ubuntu 14.04 LTS
- 1 GB of RAM

- 2 virtual CPUs
- 10 GB disk

Required specs for the dataminer VMs, from the wiki page https://wiki.gcube-system.org/index.php/DataMiner_Installation:

- Hostname: dataminer[1:2]-p-d4s.d4science.org
- Ubuntu 12.04.5 LTS
- 6 GB of RAM
- 10 virtual CPUs
- 10 GB of HD space

#7 - Dec 16, 2015 11:47 AM - Andrea Dell'Amico

- Status changed from New to In Progress
- Assignee changed from Tommaso Piccioli to Andrea Dell'Amico

#8 - Dec 16, 2015 12:17 PM - Andrea Dell'Amico

```
dataminer1-p-d4s.d4science.org IP: 146.48.122.251
dataminer2-p-d4s.d4science.org IP: 146.48.123.64
dataminer-lb1-p-d4s.d4science.org IP: 146.48.123.71
```

#9 - Dec 16, 2015 01:17 PM - Andrea Dell'Amico

- % Done changed from 20 to 60

The VMs are up and the provisioning of the three hosts is running.

#10 - Dec 16, 2015 03:50 PM - Andrea Dell'Amico

- Status changed from In Progress to Feedback
- % Done changed from 60 to 100

The dataminer production cluster is ready to be tested. The main URL is <http://dataminer.d4science.org/wps/>

#11 - Dec 16, 2015 05:56 PM - Gianpaolo Coro

- Status changed from Feedback to In Progress
- % Done changed from 100 to 90

I added Dataminer to the BiodiversityLab, ScalableDataMining and BiOnym VREs. It works very well with "local" algorithms. However no SmartGenericWorker is present in these VREs and should be added. I'm going to open another ticket and put it as dependency of this ticket.

#12 - Dec 16, 2015 05:59 PM - Gianpaolo Coro

- Related to Support #1837: Add SmartGenericWorkers to the VREs added

#13 - Dec 17, 2015 07:03 PM - Gianpaolo Coro

Test on all the algorithms were successful using 4 workers behind the scenes. I communicated the main dataminer link (load balancer) to FishBase people.

I have two questions on the installation @andrea.dellamico@isti.cnr.it :

- 1 - is the Dataminer configuration downloaded on-the fly during the installation? In other words: will it always be updated?
- 2 - is it possible to copy the encryption keys (<https://goo.gl/cSXOBq>) also in the PARALLEL_PROCESSING folder?

#14 - Dec 17, 2015 10:45 PM - Andrea Dell'Amico

Gianpaolo Coro wrote:

Test on all the algorithms were successful using 4 workers behind the scenes. I communicated the main dataminer link (load balancer) to FishBase people.

I have two questions on the installation @andrea.dellamico@isti.cnr.it :

- 1 - is the Dataminer configuration downloaded on-the fly during the installation? In other words: will it always be updated?

Nope. It was not requested, so the download of the svn files is done once. If the matter is to run a svn update at regular intervals it can be done easily. Otherwise, please give details.

- 2 - is it possible to copy the encryption keys (<https://goo.gl/cSXOBq>) also in the PARALLEL_PROCESSING folder?

Yes, I'll do.

#15 - Dec 17, 2015 11:19 PM - Gianpaolo Coro

Thank you Andrea,
the cfg and the PARALLEL_PROCESSING folders would be better downloaded from this SVN location in the next installations:

<https://svn.d4science.research-infrastructures.eu/gcube/trunk/data-analysis/DataMinerConfiguration>

#16 - Dec 17, 2015 11:27 PM - Andrea Dell'Amico

Gianpaolo Coro wrote:

Thank you Andrea,
the cfg and the PARALLEL_PROCESSING folders would be better downloaded from this SVN location in the next installations:

Do we need to perform a svn update regularly or not?

<https://svn.d4science.research-infrastructures.eu/gcube/trunk/data-analysis/DataMinerConfiguration>

I was using this (public path, to avoid authentication and problems with the untrusted TLS certificate):

http://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/DataMinerConfiguration/{cfg,PARALLEL_PROCESSING}

I don't think there's a difference between the private and the public link.

#17 - Dec 17, 2015 11:35 PM - Gianpaolo Coro

So I misunderstood. We do not need an automatic periodic update. Those folders need to be installed only at the beginning and should be updated only when a new release is available.
Furthermore, your link is good, it is the free-access read-only version of my link. If your script downloads that folder at each installation there is no problem.

#18 - Dec 17, 2015 11:44 PM - Andrea Dell'Amico

Gianpaolo Coro wrote:

If your script downloads that folder at each installation there is no problem.

Yes. I only changed the svn link and you changed the wiki accordingly, so that it now points to the public access svn repository.

#19 - Dec 18, 2015 02:30 PM - Andrea Dell'Amico

- Status changed from In Progress to Feedback

Andrea Dell'Amico wrote:

2 - is it possible to copy the encryption keys (<https://goo.gl/cSXOBq>) also in the PARALLEL_PROCESSING folder?

Yes, I'll do.

The playbook is ready and I tried it on the dev instances. I see that you installed the production keys there. Do we need both the dev and the production keys, or under 'PARALLEL_PROCESSING' the production keys are needed even if it's the dev environment?

#20 - Dec 18, 2015 03:51 PM - Gianpaolo Coro

Perhaps the best would be having the production keys installed on the prod. servers and the development key on the dev. servers. From a technical point of view there is no issue if you copy all the keys on the services, since the scopes are assigned at infrastructure level and the belonging infrastructure depends on the GHN configuration.

#21 - Dec 18, 2015 04:13 PM - Andrea Dell'Amico

Gianpaolo Coro wrote:

Perhaps the best would be having the production keys installed on the prod. servers and the development key on the dev. servers.

It's what I'm (the playbook is) doing.

#22 - Dec 18, 2015 05:29 PM - Andrea Dell'Amico

- Status changed from Feedback to Closed
- % Done changed from 90 to 100