D4Science Infrastructure - Incident #6217

Tagme application: High load

Dec 14, 2016 06:09 PM - Roberto Cirillo

Status:	Closed	**Start date:**	Dec 14, 2016
Priority:	Urgent	**Due date:**	
Assignee:	Marco Cornolti	**% Done:**	100%
Category:	Application	**Estimated time:**	0.00 hour
Target version:	UnSprintable		
Infrastructure:	Production		

Description

I've restarted the tagme application because the load has been increased over 200.
A lot of requests are coming from the following user:

```
sum1kawa.pregc@160.16.92.217
```

from Tokyo (Japan)

If the number of requests will increase we should alert or block the user. Otherwise we should think a different solution.

History

#1 - Dec 14, 2016 06:20 PM - Pasquale Pagano

- Priority changed from Normal to Urgent

Ciao @cornolti@di.unipi.it, you have not the right but we can block the activity of a single user by invalidating temporarily his token. You could also send a private message to him to notify that we are going to block him temporarily or permanently according to his behaviour.
In the future, it is better to clarify this in the documentation of TagMe by reporting some information about:

- maximum number of concurrent requests (if the load was up to 200, it means that he was sending hundreds parallel requests considering the TagMe performance)
- maximum number of request per day

What do you think?

#2 - Dec 14, 2016 06:50 PM - Marco Cornolti

Hi guys, thanks for pointing this out.

I'd say blocking the user is too radical, and I'd keep it as a measure of last resort. I will contact him in private and ask him to limit the number of concurrent queries.

There is a global maxConnections parameter in Tomcat, that defaults to 500. I have lowered it to 16 (2 * the number of cores, do you think this is fine?). But it would be more appropriate if there was any way to configure the Gcube authorization module to limit the number of per-user parallel requests, keeping a per-user queue. Is there any way to do that?

#3 - Dec 14, 2016 06:58 PM - Marco Cornolti

Okay I have asked the user to limit the number of parallel queries to eight. I also asked him to tell me what he did, so we have more information on what caused the crash. If he keeps on submitting a high number of queries, please cut him out.

Thanks!

#4 - Dec 15, 2016 10:51 AM - Roberto Cirillo

Marco Cornolti wrote:

> Hi guys, thanks for pointing this out.
>
> I'd say blocking the user is too radical, and I'd keep it as a measure of last resort. I will contact him in private and ask him to limit the number of concurrent queries.
>
> There is a global maxConnections parameter in Tomcat, that defaults to 500. I have lowered it to 16 (2 * the number of cores, do you think this is fine?). But it would be more appropriate if there was any way to configure the Gcube authorization module to limit the number of per-user parallel requests, keeping a per-user queue. Is there any way to do that?

I think it's better to limit the number of parallel queries to the single user. If we restrict only the global connections to 16, a single user could use all the available connections cutting out other users

#5 - Dec 15, 2016 10:53 AM - Roberto Cirillo

- Status changed from New to Closed

- % Done changed from 0 to 100

#6 - Jan 08, 2017 04:45 PM - Andrea Dell'Amico

- Status changed from Closed to In Progress

- % Done changed from 100 to 70

It seems that the same user is abusing the service again.

#7 - Jan 09, 2017 11:03 AM - Roberto Cirillo

I've restarted the service right now

#8 - Jan 09, 2017 11:26 AM - Marco Cornolti

Andrea Dell'Amico wrote:

> It seems that the same user is abusing the service again.

Can you be more specific about that? Was the user activity making the service unreachable, or the response time too high? I just need to check if the limits on concurrency that I have set are fine or should be stricter.

I guess the solution would be to serve requests with a round-robin policy, with a request queue for each user. Do you have any idea on how to do that? Should it be done on the server side or rather by the gcube authentication?

#9 - Jan 10, 2017 11:26 AM - Roberto Cirillo

- Status changed from In Progress to Closed

- % Done changed from 70 to 100

Marco Cornolti wrote:

> Andrea Dell'Amico wrote:
>
> > It seems that the same user is abusing the service again.
>
> Can you be more specific about that? Was the user activity making the service unreachable, or the response time too high? I just need to check if the limits on concurrency that I have set are fine or should be stricter.

Unfortunately I cannot see the logs now: there is a black hole in the log file. I assume the problem was the same: a lot of requests from the ip above. However this time the load was to 20, the last time the load was over 200. I think is better to restrict again the number of concurrency from a single user or alternatively you could wait the next high load from better analyze the situation.

> I guess the solution would be to serve requests with a round-robin policy, with a request queue for each user. Do you have any idea on how to do that? Should it be done on the server side or rather by the gcube authentication?

Unfortunately we don't have at the moment this kind of mechanism but of course we should think another solution.
I'm going to open another thread for discussing about this topic.I'm going to close this one because the load is in the range now.