D4Science Infrastructure - Task #10279

Large file upload limit on the workspace

Nov 09, 2017 02:08 PM - Claudio Atzori

Status:	Closed	**Start date:**	Nov 09, 2017
Priority:	Urgent	**Due date:**	
Assignee:	Costantino Perciante	**% Done:**	100%
Category:	Other	**Estimated time:**	0.00 hour
Target version:	UnSprintable		
Infrastructure:	Production		

Description

I'm trying to upload large files (up to 20Gb each) on the OpenAIRE datathon workspace using the curl command described in the wiki:

https://gcube.wiki.gcube-system.org/gcube/Home_Library_REST_API

Unfortunately the transfer fails with error "413 Request Entity Too Large".

Could you please adjust the Nginx configuration to allow large files in? 50Gb should be enough.

Related issues:

Related to D4Science Infrastructure - Incident #10306: mongo3-p-d4s.d4science...	**Closed**	**Nov 10, 2017**

History

#1 - Nov 09, 2017 02:40 PM - Andrea Dell'Amico

- Status changed from New to In Progress

- % Done changed from 0 to 100

I increased it to 50GB in the workspace nginx configuration.

#2 - Nov 09, 2017 02:42 PM - Andrea Dell'Amico

- Status changed from In Progress to Feedback

#3 - Nov 10, 2017 09:34 AM - Marek Horst

I managed to upload successfully 8GB file. It took 49 mins.

When I tried to upload 12GB file (arxiv-text-20171017.gz in /Workspace/VRE Folders/1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4) it finished in 73 mins, client obtained valid response but when I took a look at the workspace I found 253MB file size. File seems to be truncated.

Execution time seems to be proportional so I guess whole file was uploaded.

How is this possible?

#4 - Nov 10, 2017 09:46 AM - Andrea Dell'Amico

@costantino.perciante@isti.cnr.it @roberto.cirillo@isti.cnr.it do you find any error on the worskpace?

#5 - Nov 10, 2017 12:21 PM - Pasquale Pagano

@marek.horst@gmail.com, have you tried to refresh the workspace to see if the dimension is correct? The file dimension could be invalid at first moment since it is initially guessed. Have you tried to download the file to see if it is valid?
You could perform these simple verification while the team verify the logs of the workspace.

#6 - Nov 10, 2017 12:38 PM - Marek Horst

Pasquale Pagano wrote:

> @marek.horst@gmail.com, have you tried to refresh the workspace to see if the dimension is correct? The file dimension could be invalid at first moment since it is initially guessed. Have you tried to download the file to see if it is valid?
> You could perform these simple verification while the team verify the logs of the workspace.

Yeap, I have. I've signed out and in (I even had to do this otherwise I wasn't able to download - I guess I could create dedicated ticket for that case), started downloading the file to my local machine with 247MB file size indicated by chrome. The thing is as soon as I got to:

```
247 MB of 247 MB
```

the download speed plummeted to 0 B/s and stuck on this value. Downloaded temporary file size didn't increase since then.

#7 - Nov 10, 2017 12:44 PM - Marek Horst

Marek Horst wrote:

> The thing is as soon as I got to:
>
> ```
> 247 MB of 247 MB
> ```
>
> the download speed plummeted to 0 B/s and stuck on this value. Downloaded temporary file size didn't increase since then.

For the record: downloading process has just finished. File was truncated at mentioned 247MB:

```
$ gunzip arxiv-text-20171017.gz
```

```
gzip: arxiv-text-20171017.gz: unexpected end of file
```

#8 - Nov 10, 2017 12:51 PM - Marek Horst

Btw, right now I am uploading epmc-meta-20171017.gz file (8GB) to:

```
/Workspace/VRE Folders/1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4
```

so you can monitor the process.

#9 - Nov 10, 2017 12:54 PM - Marek Horst

Marek Horst wrote:

> Btw, right now I am uploading epmc-meta-20171017.gz file (8GB) to:
>
> ```
> /Workspace/VRE Folders/1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4
> ```
>
> so you can monitor the process.

The very same thing happened: I got valid response:

```
<string>/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4/epmc-meta-20171017.gz</string>
```

but the file got truncated at 159MB.

#10 - Nov 10, 2017 01:14 PM - Marek Horst

The command I am using at my machine is the following:

```
curl --header "Transfer-Encoding: chunked" --header "gcube-token: XXX-my-token-XXX" --request POST -T "$fileName" --header "Content-Type: application/javascript" 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload?name='$fileName'&description=myDataset&parentPath=/Home/marek.horst/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon%20datasets/Dataset%20%234'
```

I've added Transfer-Encoding: chunked header when finding the ways to make my script working but I guess it shouldn't cause the problem since I managed to upload some of the files successfully with this command.

#11 - Nov 10, 2017 02:05 PM - Pasquale Pagano

- Priority changed from Normal to Urgent

Dear @marek.horst@gmail.com, today there is a strike in Italy and more than half of the CNR personnel is on strike. I am sorry but it is hard to support you today.
I uploaded 200 GB without any problem but my archive is composed of several 2 GB files on average. On Monday it will be our first priority. Can you postpone your activity until Monday?

#12 - Nov 10, 2017 02:13 PM - Marek Horst

Pasquale Pagano wrote:

> Dear @marek.horst@gmail.com, today there is a strike in Italy and more than half of the CNR personnel is on strike. I am sorry but it is hard to support you today.
> I uploaded 200 GB without any problem but my archive is composed of several 2 GB files on average. On Monday it will be our first priority. Can you postpone your activity until Monday?

Sure, no worries, yesterday Alessia mentioned you are going to be on a strike this Friday so I was prepared ;)

Please try to test with larger files, because I managed to upload 2GB file without problems (once I was also lucky with 8GB).

#13 - Nov 10, 2017 06:43 PM - Andrea Dell'Amico

- Related to Incident #10306: mongo3-p-d4s.d4science.org went out of memory more than once added

#14 - Nov 13, 2017 05:22 PM - Marek Horst

It seems only the "large enough" files are truncated. Here is the summary of already uploaded files:

file name	uploaded file size	stored file size
arxiv-text-20171017.gz	13GB	247MB
epmc-meta-20171017.gz	8.2GB	159MB
arxiv-meta-20171017.gz	2GB	2GB
repec-meta-20171017.gz	644M	644M

All the files up to 2GB file size are valid and full content is available. Each time each "large enough" file is truncated repeatedly at the very same length: 8.2GB at 159MB, 13GB at 247MB.

I wasn't sure at which point file becomes "large enough" so I have conducted several tests and it turned out we're OK up to 4GB (4294967296). Uploading 5GB (5368709120) resulted EXACTLY in 1GB (1 048 576 kB) file and as you probably guess - uploading 4294967297 (4GB+1B) file results in 1B file (probably 1B: workspace shows 1kB because the kB is the smallest size unit).

We are clearly bashing the wall at 4294967296 bytes, this explains why we got e.g. 253142kB file for arxiv-text-20171017.gz:

13144119378 % 4294967296 = 259217490 (253142kB)

So it is almost all clear... All but the fact I managed to successfully upload 8GB file last Thursday (9th of November). Maybe the file went different path, e.g. the transfer was handled by different machine?

#15 - Nov 13, 2017 06:08 PM - Pasquale Pagano

I added few more people from the CNR team to this issue since they are working on the scene to understand the issue.

The architecture of the service you are using is quite complex and distributed clearly. The backend is a distributed storage system. We uploaded a single file up to 50GB without any problem by pushing the file with a storage client. So the issue should not be in the backend technology.
Rather, it should be on one of the following components: the HL client you are using, the HL service backend, the service we use to store metadata and ACL about the files (JackRabbit), the reverse proxy.
The CNR team already verified the configuration of the single components of the architecture and there should not be any limit to 4 GB as you spotted.

I invite CNR team to report any additional information given your recent analysis.

#16 - Nov 13, 2017 06:29 PM - Marek Horst

Thank you Pasquale.

Tomorrow I will test curl upload against other server to eliminate client side problem. This is unlikelly but possible.

#17 - Nov 13, 2017 06:37 PM - Roberto Cirillo

Marek Horst wrote:

> I managed to upload successfully 8GB file. It took 49 mins.
>
> When I tried to upload 12GB file (arxiv-text-20171017.gz in /Workspace/VRE Folders/1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4) it finished in 73 mins, client obtained valid response but when I took a look at the workspace I found 253MB file size. File seems to be truncated.
>
> Execution time seems to be proportional so I guess whole file was uploaded.
>
> How is this possible?

I confirm that the file arxiv-text-20171017.gz is present on our storage. The file size is the following: 13.144.119.378 bytes. Please @mhorst@icm.edu.pl Could you confirm the file size?
Maybe the workspace interface is referring to another corrupted file. I need further analysis on that.

#18 - Nov 13, 2017 06:46 PM - Marek Horst

Roberto Cirillo wrote:

> I confirm that the file arxiv-text-20171017.gz is present on our storage. The file size is the following: 13.144.119.378 bytes. Please @mhorst@icm.edu.pl Could you confirm the file size?

Yes, I can confirm the file size is valid. So it seems it was properly uploaded in the end. But when downloading I receive truncated file.

#19 - Nov 14, 2017 09:27 AM - Roberto Cirillo

- Status changed from Feedback to In Progress

- Assignee changed from _InfraScience Systems Engineer to Roberto Cirillo

#20 - Nov 14, 2017 09:58 AM - Marek Horst

Marek Horst wrote:

> Roberto Cirillo wrote:
>
> > I confirm that the file arxiv-text-20171017.gz is present on our storage. The file size is the following: 13.144.119.378 bytes. Please @mhorst@icm.edu.pl Could you confirm the file size?
>
> Yes, I can confirm the file size is valid. So it seems it was properly uploaded in the end. But when downloading I receive truncated file.

This may also explain:

```
29 items, 69.19 GB
```

present in the left bottom corner of workspace panel:

https://services.d4science.org/group/d4science-services-gateway/workspace

It seems like every file I have uploaded so far was taken into account because in the workspace I can sum all the files up to ~11GB.

#21 - Nov 14, 2017 10:40 AM - Costantino Perciante

We managed to fix the largest of them (i.e. arxiv-text-20171017.gz); you can now download the entire file.

I kindly ask @roberto.cirillo@isti.cnr.it to report the real sizes of the files in the /Datathon datasets/Dataset #4/ shared folder, so that we can check them all

#22 - Nov 14, 2017 11:35 AM - Roberto Cirillo

- Assignee changed from Roberto Cirillo to Costantino Perciante

The file epmc-meta-20171017.gz is present on our storage with the following size: 8753058643 bytes.
However this file is no more present on the "Dataset #4" folder but only in the Trash.
We should restore the file from trash and set the correct metadata value.

#23 - Nov 14, 2017 11:45 AM - Marek Horst

Roberto Cirillo wrote:

> The file epmc-meta-20171017.gz is present on our storage with the following size: 8753058643 bytes.
> However this file is no more present on the "Dataset #4" folder but only in the Trash.
> We should restore the file from trash and set the correct metadata value.

I guess I have removed this file after I saw the incorrect size.

You don't need to restore it because I've started re-upload of this file few minutes ago (before I've seen your comment).

#24 - Nov 14, 2017 12:36 PM - Marek Horst

Marek Horst wrote:

> Roberto Cirillo wrote:
>
> > The file epmc-meta-20171017.gz is present on our storage with the following size: 8753058643 bytes.
> > However this file is no more present on the "Dataset #4" folder but only in the Trash.
> > We should restore the file from trash and set the correct metadata value.
>
> I guess I have removed this file after I saw the incorrect size.
>
> You don't need to restore it because I've started re-upload of this file few minutes ago (before I've seen your comment).

Done. I am about to upload 3 more files:

- epmc-text-20171017.gz (19GB, ongoing)
- other-meta-20171017.gz
- other-text-20171017.gz

#26 - Nov 14, 2017 03:04 PM - Costantino Perciante

Marek Horst wrote:

> Done. I am about to upload 3 more files:
>
> - epmc-text-20171017.gz (19GB, ongoing)
> - other-meta-20171017.gz
> - other-text-20171017.gz

Please let us know as soon as you finish with them

#27 - Nov 15, 2017 11:59 AM - Roberto Cirillo

- Status changed from In Progress to Feedback

- Assignee changed from Costantino Perciante to Marek Horst

#28 - Nov 15, 2017 12:59 PM - Marek Horst

Costantino Perciante wrote:

> Please let us know as soon as you finish with them

The last one (and the largest one: 32GB) is ongoing. I will give you a note when it's done.

#29 - Nov 15, 2017 03:55 PM - Marek Horst

Marek Horst wrote:

> Costantino Perciante wrote:
>
> > Please let us know as soon as you finish with them
>
> The last one (and the largest one: 32GB) is ongoing. I will give you a note when it's done.

Hmm, I'm not sure other-text-20171017.gz was properly uploaded. Even though it appeared in workspace I didn't receive valid response from server. After 224 minutes of uploading I got:

```
<html><body><h1>504 Gateway Time-out</h1>
The server didn't respond in time.
</body></html>
```

Can you check the real size in storage layer? It should be 34022668402.

#30 - Nov 15, 2017 03:57 PM - Marek Horst

- Assignee changed from Marek Horst to Costantino Perciante

#31 - Nov 15, 2017 04:02 PM - Costantino Perciante

- Assignee changed from Costantino Perciante to Roberto Cirillo

#32 - Nov 15, 2017 06:43 PM - Roberto Cirillo

Unfortunately the file named other-text-20171017.gz in the storage layer is 28635607040 so I think it is not complete.

#33 - Nov 15, 2017 06:46 PM - Roberto Cirillo

I've tried to check the integrity with gzip -t and I confirm, the file is not complete. @andrea.dellamico@isti.cnr.it any suggestion?

#34 - Nov 16, 2017 09:28 AM - Marek Horst

I tried to reupload other-text-20171017.gz file, this time I got:

```
<!DOCTYPE html>
<html>
<head>
<title>Error</title>
<style>
    body {
        width: 35em;
        margin: 0 auto;
        font-family: Tahoma, Verdana, Arial, sans-serif;
    }
</style>
</head>
<body>
<h1>An error occurred.</h1>
<p>Sorry, the page you are looking for is currently unavailable.<br/>
Please try again later.</p>
<p>If you are the system administrator of this resource then you should check
the <a href="http://nginx.org/r/error_log">error log</a> for details.</p>
<p><em>Faithfully yours, nginx.</em></p>
</body>
</html>
```

after 106m. I've just triggered upload processes once again.

#35 - Nov 16, 2017 09:40 AM - Roberto Cirillo

Before retry could you try to do a double compression of the file like "tar.gz"?

#36 - Nov 16, 2017 09:45 AM - Costantino Perciante

It would be also better to perform a multipart-like post request. Like this:

curl --header "gcube-token:**********" --request POST -v -F name=file_name -F parentPath=/Home/test.user/Workspace/ -F data=@/path/to/file -F description=test 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload'

#37 - Nov 16, 2017 11:44 AM - Marek Horst

Roberto Cirillo wrote:

> Before retry could you try to do a double compression of the file like "tar.gz"?

Unfortunately I don't have enough diskspace left on a gateway machine where the gz file was created.

But would that help anyway? Do you suggest creating tar.gz on already compressed gz file or making tar.gz out of the source txt file?

#38 - Nov 16, 2017 11:48 AM - Marek Horst

Costantino Perciante wrote:

> It would be also better to perform a multipart-like post request. Like this:
>
> curl --header "gcube-token:**********" --request POST -v -F name=file_name -F parentPath=/Home/test.user/Workspace/ -F data=@/path/to/file -F description=test 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload'

Multipart upload is ongoing.

I hope -F data=@/path/to/file works differently than --data-binary @/path/to/file (which I had to replace with -T /path/to/file) because whole file was stored in memory before uploading.

#39 - Nov 16, 2017 12:22 PM - Roberto Cirillo

Marek Horst wrote:

> Roberto Cirillo wrote:
>
> > Before retry could you try to do a double compression of the file like "tar.gz"?
>
> Unfortunately I don't have enough diskspace left on a gateway machine where the gz file was created.
>
> But would that help anyway? Do you suggest creating tar.gz on already compressed gz file or making tar.gz out of the source txt file?

I think it's not possible in one step to convert from gz to tar.gz. If you have the .txt file without compression (and enough diskspace left), you could try to convert it to "tar.gz" otherwise we should find another solution.

#40 - Nov 16, 2017 12:39 PM - Marek Horst

Roberto Cirillo wrote:

> Marek Horst wrote:
>
> > Roberto Cirillo wrote:
> >
> > > Before retry could you try to do a double compression of the file like "tar.gz"?
> >
> > Unfortunately I don't have enough diskspace left on a gateway machine where the gz file was created.
> >
> > But would that help anyway? Do you suggest creating tar.gz on already compressed gz file or making tar.gz out of the source txt file?
>
> I think it's not possible in one step to convert from gz to tar.gz. If you have the .txt file without compression (and enough diskspace left), you could try to convert it to "tar.gz" otherwise we should find another solution.

I wouldn't expect major gain when packaging single txt file with tar.gz instead of gz (in fact the source for gzipping is not file but the stream generated by hadoop fs -cat piped command). AFAIK tar by itself does not introduce any compression. This could be useful when dealing with thousands of source files - then we could benefit from creating single gzipped tar archive.

#41 - Nov 16, 2017 01:48 PM - Costantino Perciante

Marek Horst wrote:

> Costantino Perciante wrote:
>
> > It would be also better to perform a multipart-like post request. Like this:
> >
> > curl --header "gcube-token:**********" --request POST -v -F name=file_name -F parentPath=/Home/test.user/Workspace/ -F data=@/path/to/file -F description=test 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload'
>
> Multipart upload is ongoing.
>
> I hope -F data=@/path/to/file works differently than --data-binary @/path/to/file (which I had to replace with -T /path/to/file) because whole file was stored in memory before uploading.

I uploaded a 20 GB file on our development infrastructure, with the same command, from my pc without problems. It doesn't load the whole file in memory

#42 - Nov 16, 2017 04:55 PM - Marek Horst

This time I got:

```
<!DOCTYPE html>
<html>
<head>
<title>Error</title>
<style>
    body {
        width: 35em;
        margin: 0 auto;
        font-family: Tahoma, Verdana, Arial, sans-serif;
    }
</style>
</head>
<body>
```

```
<h1>An error occurred.</h1>
<p>Sorry, the page you are looking for is currently unavailable.<br/>
Please try again later.</p>
<p>If you are the system administrator of this resource then you should check
the <a href="http://nginx.org/r/error_log">error log</a> for details.</p>
<p><em>Faithfully yours, nginx.</em></p>
</body>
</html>
```

but the file seems to be properly uploaded and the file size is correct. At least rounded to kB, not sure if it was uploaded to the very last byte.

Could you please verify its integrity with gzip -t command?

#43 - Nov 16, 2017 05:39 PM - Costantino Perciante

Marek Horst wrote:

> This time I got:
>
> ```
> <!DOCTYPE html>
> <html>
> <head>
> <title>Error</title>
> <style>
>     body {
>         width: 35em;
>         margin: 0 auto;
>         font-family: Tahoma, Verdana, Arial, sans-serif;
>     }
> </style>
> </head>
> <body>
> <h1>An error occurred.</h1>
> Sorry, the page you are looking for is currently unavailable.<br/>
> Please try again later.
> If you are the system administrator of this resource then you should check
> the <a href="http://nginx.org/r/error_log">error log</a> for details.
> Faithfully yours, nginx.
> </body>
> </html>
> ```
>
> but the file seems to be properly uploaded and the file size is correct. At least rounded to kB, not sure if it was uploaded to the very last byte.
>
> Could you please verify its integrity with gzip -t command?

I guess nginx closes the connection after a while. @andrea.dellamico@isti.cnr.it could confirm its connection timeout. I experienced the same behaviour in my yesterday's test

#44 - Nov 16, 2017 07:00 PM - Roberto Cirillo

Marek Horst wrote:

> This time I got:
>
> ```
> <!DOCTYPE html>
> <html>
> <head>
> <title>Error</title>
> <style>
>     body {
>         width: 35em;
>         margin: 0 auto;
>         font-family: Tahoma, Verdana, Arial, sans-serif;
>     }
> </style>
> </head>
> <body>
> <h1>An error occurred.</h1>
> Sorry, the page you are looking for is currently unavailable.<br/>
> Please try again later.
> If you are the system administrator of this resource then you should check
> the <a href="http://nginx.org/r/error_log">error log</a> for details.
> Faithfully yours, nginx.
> </body>
> ```

```
</html>
```

> but the file seems to be properly uploaded and the file size is correct. At least rounded to kB, not sure if it was uploaded to the very last byte.
>
> Could you please verify its integrity with gzip -t command?

I've verified, the file has been properly uploaded to our storage layer.

#45 - Nov 16, 2017 09:43 PM - Andrea Dell'Amico

Costantino Perciante wrote:

> I guess nginx closes the connection after a while. @andrea.dellamico@isti.cnr.it could confirm its connection timeout. I experienced the same behaviour in my yesterday's test

Yes it does. It's configurable, but raising it opens the server to the risk of very simple DDOS.

#46 - Nov 17, 2017 10:05 AM - Marek Horst

All done.

There is only one file left we need to fix its size: epmc-meta-20171017.gz. All the other were already checked I guess.

#47 - Nov 17, 2017 12:29 PM - Costantino Perciante

- Assignee changed from Roberto Cirillo to Marek Horst

Marek Horst wrote:

> All done.
>
> There is only one file left we need to fix its size: epmc-meta-20171017.gz. All the other were already checked I guess.

Fixed.. please close this ticket if everything is ok now

#48 - Nov 17, 2017 12:31 PM - Marek Horst

- Status changed from Feedback to Closed

Guys, thank you for your help. I really appreciate it :)

#49 - Nov 29, 2017 01:16 PM - Marek Horst

- Status changed from Closed to In Progress

I need to reopen this ticket because we have to reupload packages.

Currently the problem is I am unable to delete README.txt file via the workspace. File is located in:

/Home/marek.horst/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4

The window with the following message is shown:

Sorry, an error has occurred on the server when deleting item. LockException: Node locked. Impossible to remove itemID d6bfe4cf-18bf-48a4-8fde-cb8e0b58effd

The other thing (probably related to the deletion issue) is I am unable to upload new version of this file. I am receiving the following, pretty enigmatic error message:

java.lang.ClassCastException: java.lang.String cannot be cast to org.gcube.common.homelibary.model.items.ItemDelegate

#50 - Nov 29, 2017 01:17 PM - Marek Horst

- Assignee changed from Marek Horst to Costantino Perciante

#51 - Nov 29, 2017 01:39 PM - Costantino Perciante

There is a scheduled maintainance downtime of the infrastructure today. Some of the errors may be related to this. It should end at 16:00. Could you wait untill it finishes? Then we can better check the problem

#52 - Nov 29, 2017 05:50 PM - Marek Horst

Costantino Perciante wrote:

> There is a scheduled maintainance downtime of the infrastructure today. Some of the errors may be related to this. It should end at 16:00. Could you wait untill it finishes? Then we can better check the problem

Is it finished already?

I am constantly receiving **504 Gateway Time-out** but as you already explained before: this a feature, not a bug ;)

One problem is after upload finishes the file do not appear in the workspace instantly, it takes several minutes to show up.

Second problem is I have uploaded very last 10GB epmc-text-20171017.gz file but it does not appear in workspace even though over 1h has passed since upload finished. I have reuploaded it once again - still no sight.

Generally speaking whole uploading process is pretty painful... OpenAIRE datathon starts tomorrow so it will be nice to have all files there before then.

#53 - Nov 29, 2017 06:21 PM - Costantino Perciante

Marek Horst wrote:

> Costantino Perciante wrote:
>
> > There is a scheduled maintainance downtime of the infrastructure today. Some of the errors may be related to this. It should end at 16:00. Could you wait untill it finishes? Then we can better check the problem
>
> Is it finished already?
>
> I am constantly receiving **504 Gateway Time-out** but as you already explained before: this a feature, not a bug ;)
>
> One problem is after upload finishes the file do not appear in the workspace instantly, it takes several minutes to show up.
>
> Second problem is I have uploaded very last 10GB epmc-text-20171017.gz file but it does not appear in workspace even though over 1h has passed since upload finished. I have reuploaded it once again - still no sight.
>
> Generally speaking whole uploading process is pretty painful... OpenAIRE datathon starts tomorrow so it will be nice to have all files there before then.

Everything should work now, so if something fails we must check the reason why it did. We are not able to find the epmc-text-20171017.gz file on the workspace/storage area. Could you retry to upload it? Moreover, did you face a "gateway timeout" during the upload of this file?

#54 - Nov 29, 2017 06:50 PM - Marek Horst

Costantino Perciante wrote:

> Everything should work now, so if something fails we must check the reason why it did.

OK.

> We are not able to find the epmc-text-20171017.gz file on the workspace/storage area. Could you retry to upload it?

Just triggered another upload with the following script:

```
curl --header "gcube-token: XXXXXXXXXXXXXXXXXXXXXXXX" --request POST -v -F name=$fileName -F parentPath='/Home/marek.horst/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4' -F data=@"$fileName" -F description=test 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload'
```

It should finish in ~38 minutes as this was the time of previous uploads of this file.

> Moreover, did you face a "gateway timeout" during the upload of this file?

Every file upload resulted in gateway timeout returned to client.

#55 - Nov 29, 2017 08:20 PM - Marek Horst

Another attempt of epmc-text-20171017.gz upload has just finished with the same result.

Time taken suggest whole file should be uploaded although I cannot find it in workfspace.

#56 - Nov 30, 2017 04:26 PM - Costantino Perciante

Marek Horst wrote:

> Another attempt of epmc-text-20171017.gz upload has just finished with the same result.
>
> Time taken suggest whole file should be uploaded although I cannot find it in workfspace.

Marek, we incremented some nginx parameters to avoid timeouts. Could you retry once again?

#57 - Nov 30, 2017 04:31 PM - Marek Horst

Costantino Perciante wrote:

> Marek, we incremented some nginx parameters to avoid timeouts. Could you retry once again?

I just triggered upload. It should finish in ~40 minutes.

#58 - Nov 30, 2017 05:14 PM - Marek Horst

Marek Horst wrote:

> Costantino Perciante wrote:
>
> > Marek, we incremented some nginx parameters to avoid timeouts. Could you retry once again?
>
> I just triggered upload. It should finish in ~40 minutes.

This is what I got:

```
< HTTP/1.1 504 Gateway Time-out
< Server: nginx
< Date: Thu, 30 Nov 2017 16:02:57 GMT
< Content-Type: text/html
< Content-Length: 537
< ETag: "5315bd25-219"
< Strict-Transport-Security: max-age=15768000
* HTTP error before end of send, stop sending
<
<!DOCTYPE html>
<html>
<head>
<title>Error</title>
<style>
    body {
        width: 35em;
        margin: 0 auto;
        font-family: Tahoma, Verdana, Arial, sans-serif;
    }
</style>
</head>
<body>
<h1>An error occurred.</h1>
<p>Sorry, the page you are looking for is currently unavailable.<br/>
Please try again later.</p>
<p>If you are the system administrator of this resource then you should check
the <a href="http://nginx.org/r/error_log">error log</a> for details.</p>
<p><em>Faithfully yours, nginx.</em></p>
</body>
</html>
```

after 33 minutes of uploading. It always finishes after the same amount of time (33-34 minutes).

Current total workspace size is 149.74GB and I think it was smaller before (somewhere around 139GB). But this is more a guess/hint, I didn't remember it well.

#59 - Nov 30, 2017 06:10 PM - Costantino Perciante

Marek Horst wrote:

> Marek Horst wrote:
>
> > Costantino Perciante wrote:
> >
> > > Marek, we incremented some nginx parameters to avoid timeouts. Could you retry once again?
> >
> > I just triggered upload. It should finish in ~40 minutes.
>
> This is what I got:

```
< HTTP/1.1 504 Gateway Time-out
< Server: nginx
< Date: Thu, 30 Nov 2017 16:02:57 GMT
< Content-Type: text/html
< Content-Length: 537
< ETag: "5315bd25-219"
< Strict-Transport-Security: max-age=15768000
* HTTP error before end of send, stop sending
<
<!DOCTYPE html>
<html>
<head>
<title>Error</title>
<style>
    body {
        width: 35em;
        margin: 0 auto;
        font-family: Tahoma, Verdana, Arial, sans-serif;
    }
</style>
</head>
<body>
<h1>An error occurred.</h1>
<p>Sorry, the page you are looking for is currently unavailable.<br/>
Please try again later.</p>
<p>If you are the system administrator of this resource then you should check
the <a href="http://nginx.org/r/error_log">error log</a> for details.</p>
<p><em>Faithfully yours, nginx.</em></p>
</body>
</html>
```

> after 33 minutes of uploading. It always finishes after the same amount of time (33-34 minutes).
>
> Current total workspace size is 149.74GB and I think it was smaller before (somewhere around 139GB). But this is more a guess/hint, I didn't remember it well.

Dear Marek, the problem could be related to the network (or something else). I've just uploaded 10 Gbytes file without any problem at all here at CNR.

In order to understand the amount of available bandwidth (hoping for an almost symmetric upload/download connection), could you try to download this file http://ftp.d4science.org/knime/knime-full_3.3.2.linux.gtk.x86_64.tar.gz and tell us how long the operation takes?

It would be very helpful.

Thanks

#60 - Dec 01, 2017 11:59 AM - Marek Horst

As a last yesterday's resort I have used the old command I'd been using to upload packages before (but which at some point failed to upload large files and was replaced by recommended multipart version):

```
curl --header "gcube-token: XXX" --request POST -T "$fileName" --header "Content-Type: application/javascript"
 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload?name='$fileName'&description=myDa
taset&parentPath=/Home/marek.horst/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1
st_OpenAIRE_Datathon/Datathon%20datasets/Dataset%20%234'
```

which finally succeeded, in contrary to multipart-like command:

```
curl --header "gcube-token: XXX" --request POST -v -F name=$fileName -F parentPath='/Home/marek.horst/Workspac
e/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon datasets/Data
set #4' -F data=@"$fileName" -F description=test 'https://workspace-repository.d4science.org/home-library-weba
pp/rest/Upload'
```

which repeatedly failed with 10GB file (succeeded for smaller files).

I hope this will help you in finding the real cause of the problem. You could download this file and try to reupload it again using the second command...

#61 - Dec 01, 2017 01:06 PM - Costantino Perciante

Marek Horst wrote:

> As a last yesterday's resort I have used the old command I'd been using to upload packages before (but which at some point failed to upload large files and was replaced by recommended multipart version):
>
> ```
> curl --header "gcube-token: XXX" --request POST -T "$fileName" --header "Content-Type: application/javascript" 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload?name='$fileName'&description=myDataset&parentPath=/Home/marek.horst/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon%20datasets/Dataset%20%234'
> ```
>
> which finally succeeded, in contrary to multipart-like command:
>
> ```
> curl --header "gcube-token: XXX" --request POST -v -F name=$fileName -F parentPath='/Home/marek.horst/Workspace/MySpecialFolders/d4science.research-infrastructures.eu-OpenAIRE-1st_OpenAIRE_Datathon/Datathon datasets/Dataset #4' -F data=@"$fileName" -F description=test 'https://workspace-repository.d4science.org/home-library-webapp/rest/Upload'
> ```
>
> which repeatedly failed with 10GB file (succeeded for smaller files).
>
> I hope this will help you in finding the real cause of the problem. You could download this file and try to reupload it again using the second command...

I'm glad you eventually succeeded, even if this makes the issue more problematic to be solved. We need to investigate it a bit further

#62 - Feb 21, 2018 04:07 PM - Costantino Perciante

- Status changed from In Progress to Closed