D4Science Infrastructure - Task #9127

Remove no longer files stored in Jackrabbit before migration to postgres

Jul 03, 2017 09:50 AM - Costantino Perciante

Status:	Closed	**Start date:**	Jul 03, 2017
Priority:	High	**Due date:**	
Assignee:	Costantino Perciante	**% Done:**	100%
Category:	Other	**Estimated time:**	0.00 hour
Target version:	Production Jackrabbit Migration from Derby to PostgreSQL		
Infrastructure:	Production		

Description

Before MongoDB (the actual storage used behind the workspace), files were stored into jackrabbit itself (through the DataStore facility). Some GBs of no longer used files need to be freed up. In order to do so:

- a script must be executed to replace the payload (if any) of file nodes with an empty payload (a background job that can be executed with workspace up and running);
- the DataStore Garbage Collector must be executed (it needs the workspace down, so we need to schedule its execution properly and extimate how long it takes to finish)

We tested once this operations onto a snapshot of the current content of jackrabbit in production and we were able to free up 25GB of files. However, we need to better extimate the time needed for the whole task to finish.

It is important because will let us migrate the datastore into postgres (no need of shared file system if we plan to have replicated jcr instances) and speedup the migration phase

History

#2 - Jul 03, 2017 10:40 AM - Costantino Perciante

- Description updated

#3 - Jul 05, 2017 12:42 PM - Costantino Perciante

- Status changed from New to In Progress

- % Done changed from 0 to 30

We managed to free up more or less the same space in a couple of hours on a snapshotted workspace. Valentina's script can run with jcr up and running and then the datastore garbage collector will take 2 hours more or less to actually remove data.

We can schedule its execution during the downtime needed for gcube 4.6 upgrade, can't we?

Moreover, a backup of production workspace is needed just before we proceed. Last but not least, the datastore will be migrated on postgres as well.

#4 - Jul 05, 2017 02:37 PM - Pasquale Pagano

Costantino Perciante wrote:

> We managed to free up more or less the same space in a couple of hours on a snapshotted workspace. Valentina's script can run with jcr up and running and then the datastore garbage collector will take 2 hours more or less to actually remove data.

If we can run it on the production workspace, why we have to wait fro gcube 4.6 rollout?
Can it e used to clean up the current version of the workspace?

#5 - Jul 05, 2017 02:43 PM - Costantino Perciante

Pasquale Pagano wrote:

> If we can run it on the production workspace, why we have to wait fro gcube 4.6 rollout?
> Can it e used to clean up the current version of the workspace?

Unfortunately the datastore garbage collector needs jackrabbit down (tomcat/home library must be down during its execution). This means that Valentina's script can run when we want and with jackrabbit up, but then we need to shut it down for the collector.

#6 - Jul 10, 2017 10:21 AM - Costantino Perciante

- Status changed from In Progress to Closed

- Assignee changed from Valentina Marioli to Costantino Perciante

- % Done changed from 30 to 100

This activity has been successfully concluded on friday 7/07/17. 28GB of useless data has been removed.