Pulp - Issue #8295

Disc Usage during Repository Sync

02/23/2021 11:07 AM - wibbit

Status:	NEW	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:		**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:			
Severity:	2. Medium	**Groomed:**	No
Version:		**Sprint Candidate:**	No
Platform Release:		**Tags:**	
OS:		**Sprint:**	
Triaged:	Yes	**Quarter:**	Q2-2021

Description

While performing a repository sync recently I noticed that the WORKING_DIRECTORY was of a much much larger size than I expected.

It is currently assigned 150GB, yet failed to sync 3 repositories (I believe RHEL7 Source, RHEL7 Debug, and possibly RHEL7 Server) due to running out of disk space.

I was of the understanding that the working directory was not intended to hold the entire repository till completion, but only the individual unit's being transferred, on a successful transfer, that would then be ingested and end up as an artefact and stored in the artefact's path.

I believe an additional side affect of this, is that though units have been downloaded cleanly, and I believe could have been converted to an artefact, this has not been done, and a complete re-download is required as evidenced by the fact that on an additional sync the same behaviour occurs.

Maybe this would be the case any way, unless the artefact is converted to content, I'm unsure of those specifics.

Related issues:

Related to Pulp - Issue #8296: Pulp worker directories not cleaned up	**NEW**
Related to Pulp - Issue #7316: Files are not being deleted from storage when ...	**CLOSED - CURRENTRELEASE**

History

#1 - 02/25/2021 05:24 PM - dalley

- Project changed from RPM Support to Pulp

#2 - 02/25/2021 05:24 PM - dalley

- Related to Issue #8296: Pulp worker directories not cleaned up added

#3 - 02/25/2021 05:35 PM - dalley

- Related to Issue #7316: Files are not being deleted from storage when calling the method delete() added

#4 - 03/02/2021 04:55 PM - daviddavis

- Triaged changed from No to Yes

#5 - 04/08/2021 05:21 PM - ttereshc

Just for some context. It might be a pulpcore issue but it also can be pulp_rpm specific because working directory is recreated a lot in pulp_rpm.

#6 - 04/08/2021 05:44 PM - ipanova@redhat.com

- Sprint set to Sprint 94

#7 - 04/09/2021 10:50 AM - lmjachky

- Status changed from NEW to ASSIGNED

- Assignee set to lmjachky

#8 - 04/16/2021 11:13 PM - rchan

- Sprint changed from Sprint 94 to Sprint 95

#9 - 04/30/2021 06:14 PM - rchan

- Sprint changed from Sprint 95 to Sprint 96

#10 - 05/02/2021 02:00 PM - lmjachky

- File sync_disk_usage_change.patch added

I made a couple of experiments and noticed that our pipeline is not really creating artifacts from temporary files along the way during the sync process. This is caused by the parameter minsize (https://github.com/pulp/pulpcore/blob/354383883032277e7a1f7dc7ddf2dc0a5bc40fad/pulpcore/plugin/stages/api.py#L84) that retains the pipeline from saving artifacts unless the queue is filled up (https://github.com/pulp/pulpcore/blob/eda2c890214a76e6f6ffa18cc939d04273b1fa13/pulpcore/plugin/stages/artifact_stages.py#L229).

Furthermore, all the temporary files are being deleted only when the whole sync process ends. Because of the parameter delete in NamedTemporaryFile (https://github.com/pulp/pulpcore/blob/354383883032277e7a1f7dc7ddf2dc0a5bc40fad/pulpcore/download/base.py#L122), we should handle the removal manually by ourselves. The working directory (temporary files stored under a specific task) is cleared as a whole only when the sync process reasonably finishes.

Also, another problem is that in some cases pulp_rpm do not handle the removal of temporary repodata files at all. When I synced from the http://mirrors.sonic.net/epel/playground/8/Everything/x86_64/os/, the downloaded repodata (3.9GB) were not removed even when the sync process successfuly ended. It seemed to me that I never reached this return statement: https://github.com/pulp/pulp_rpm/blob/f14be05d58f06e794676a5222b443cd9d082f031/pulp_rpm/app/tasks/synchronizing.py#L428, weird. Yet, the sync proceeded without failures.

So, I propose to examine some of the default parameters we changed in https://github.com/pulp/pulpcore/pull/440/files and call os.unlink right after we create artifacts. I am attaching some of the advised changes in the description and unassigning myself.

#11 - 05/02/2021 02:00 PM - lmjachky

- Status changed from ASSIGNED to NEW

- Assignee deleted (lmjachky)

#12 - 05/14/2021 04:46 PM - rchan

- Sprint deleted (Sprint 96)

- Quarter set to Q2-2021

We are putting this one on hold for now since this Sprint was busy, but do plan to work on it this quarter.

Files

sync_disk_usage_change.patch	2.15 KB	05/02/2021	lmjachky