Pulp - Issue #8305

Deleting a remote used as source for live content corrupts ContentArtifact records

02/25/2021 04:42 AM - dalley

Status:	NEW	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:		**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:			
Severity:	3. High	**Groomed:**	No
Version:		**Sprint Candidate:**	No
Platform Release:		**Tags:**	
OS:		**Sprint:**	Sprint 110
Triaged:	Yes	**Quarter:**	

Description

- Create a repository and on_demand remote, and sync them.
- Delete the remote

The deletion of the Remote deletes the RemoteArtifacts, leaving behind ContentArtifact attached to neither Artifacts nor Remotes, making them effectively corrupted and unpublishable.

```
# create repository, remote
remote = remote_api.create(gen_file_remote(policy='on_demand'))
repo = repo_api.create(gen_repo())

# sync the repository
repository_sync_data = RepositorySyncURL(remote=remote.pulp_href)
sync_response = repo_api.sync(repo.pulp_href, repository_sync_data)
task = monitor_task(sync_response.task)

# delete the remote
monitor_task(remote_api.delete(remote.pulp_href).task)

# ^---- problem occurs here, now RemoteArtifacts deleted, now ContentArtifact is broken

publish_response = publications_api.create({"repository_version": task.created_resources[0]})
monitor_task(publish_response.task)  # boom publish failure
```

This is more pernicious because content units can move throughout repositories, and if the remote is ever deleted, every repo can be broken at once with no safeguards.

Related issues:

Related to Pulp - Issue #9101: Content_artifact is not updated	**CLOSED - CURRENTRELEASE**
Has duplicate Ansible Plugin - Issue #7924: Sync doesn't create RemoteArtifacts	**CLOSED - DUPLICATE**

History

#1 - 02/25/2021 05:07 AM - dalley

We should probably introduce an actual constraint that enforces this, so it blows up immediately if it ever occurs.

#2 - 02/25/2021 09:03 PM - dalley

- File reproduce_publish_error.py added

- File deleted (reproduce_publish_error.py)

- Subject changed from Some sequences of events can result in invalid ContentArtifact records to Deleting a remote used as source for live content corrupts ContentArtifact records

- Description updated

#3 - 02/25/2021 09:08 PM - dalley

- Priority changed from Normal to High

#4 - 02/25/2021 09:39 PM - dalley

- Description updated

#5 - 03/02/2021 06:16 PM - dalley

Recreating the remote and re-syncing does fix the content artifacts, but if you don't notice the problem immediately, or if you copied the content around between repos, it would be practically impossible to know how to resolve the issue.

There's some discussion on solutions here: https://hackmd.io/_3gsUVdyQwy50Nc5pMXN-g

#6 - 03/05/2021 04:54 PM - fao89

- Triaged changed from No to Yes

#7 - 03/09/2021 03:18 PM - mdellweg

Idea to solve the problem: Add a force flag to the DELETE call. If force is not specified, and the remote is referenced by either a repository or a remote artifact, the call will fail.

#8 - 03/09/2021 03:19 PM - dalley

- Priority changed from High to Normal

#9 - 05/17/2021 08:12 PM - dalley

- Related to Issue #7924: Sync doesn't create RemoteArtifacts added

#10 - 05/17/2021 09:18 PM - bmbouter

- Related to deleted (Issue #7924: Sync doesn't create RemoteArtifacts)

#11 - 05/17/2021 09:19 PM - bmbouter

- Has duplicate Issue #7924: Sync doesn't create RemoteArtifacts added

#12 - 07/13/2021 05:57 PM - dalley

- Sprint set to Sprint 100

#13 - 07/13/2021 05:57 PM - ggainey

I'd go for 1c) from the associated hackmd. I wouldn't even give the option of --force - the state your pulp-instance gets left in is pretty horrific, even if you meant to do it. Having some way to list the content/repo-versions/artifacts, or a list of "do an immediate sync on the following repos before attempting" would be great.

#14 - 07/15/2021 11:08 PM - rchan

- Sprint changed from Sprint 100 to Sprint 101

#15 - 07/21/2021 05:37 PM - ipanova@redhat.com

during review of a PR there has been raised an idea about what to do with the content that does not have RA not Artifact

```
Do you think it would be reasonable to prevent content with no Artifact and no RemoteArtifact from being added
 to new repository versions via one of the validation hooks? We want to keep the historical records around, bu
t the content is no longer useful or functional, so it doesn't make sense to let it spread into new repository
 versions.
```

EDIT: add a --force flag which can be specified when new repo-version is being created to allow such content

#16 - 07/21/2021 09:38 PM - dalley

More discussion from the PR

I have not manually tested rpm plugin yet, but skimmed through pulpcore code - plugins that use directly pulp's content app handler, should be able to gracefully handle this. Uploaded content will have artifact set to none as well it won't have any remoteartifacts so a 404 will just be raised https://github.com/pulp/pulpcore/blob/master/pulpcore/content/handler.py#L681. Since pulp-container has a subclassed version of handler, i needed to modify couple of lines https://gist.github.com/ipanova/bd5821b55a1e01245fe7556dc3791ddd which led to this output:

```
$ podman pull localhost:24817/test/repo --tls-verify=false
```

```
Trying to pull localhost:24817/test/repo:latest...
Error: Error initializing source docker://localhost:24817/test/repo:latest: Error reading manifest latest in l
ocalhost:24817/test/repo: StatusCode: 404, 404: Not Found
(pulp) [vagrant@pulp3-source-fedora34 ~]$
```

tldr, i think we're fine just need to audit plugins that subclass the Handler.

EDIT: some plugins contain content that is expected to always have artifact. We should not touch those content types during reclaim disk space. For example: rpm modules and defaults, container tags, manifests and config blobs. For the rest of the content types for which disk space was supposed to correctly reclaim the artifact, the code needs to be adjusted so it takes into account situation when ca.artifact is None content._artifacts.get() returns ObjectDoesNotExist

```
[ipanova@fluffy pulp_rpm]$ git grep '\._artifacts'
pulp_rpm/app/migrations/0003_DATA_incorrect_json.py:            modulemd_index.update_from_string(module._arti
facts.first().file.read().decode(), True)
pulp_rpm/app/tasks/publishing.py:            mod_yml.write(modulemd._artifacts.get().file.read())
pulp_rpm/app/tasks/publishing.py:            mod_yml.write(default._artifacts.get().file.read())
[ipanova@fluffy pulp_rpm]$
[ipanova@fluffy pulp_rpm]$
[ipanova@fluffy pulp_rpm]$ cd ..
[ipanova@fluffy pulp3]$ cd pulp_container/
[ipanova@fluffy pulp_container]$ git grep '\._artifacts'
pulp_container/app/migrations/0007_clear_tags_artifacts_refs.py:        tag._artifacts.clear()
pulp_container/app/migrations/0007_clear_tags_artifacts_refs.py:            tag._artifacts.add(tag.tagged_mani
fest._artifacts.get())
pulp_container/app/redirects.py:            artifact = manifest._artifacts.get()
pulp_container/app/redirects.py:            artifact = blob._artifacts.get()
pulp_container/app/registry.py:            artifact = tag.tagged_manifest._artifacts.get()
pulp_container/app/registry_api.py:        artifact = manifest._artifacts.get()
pulp_container/app/registry_api.py:        artifact = blob._artifacts.get()
pulp_container/app/schema_convert.py:        config_artifact = manifest.config_blob._artifacts.get()
pulp_container/app/schema_convert.py:    manifest_artifact = manifest._artifacts.get()
pulp_container/app/tasks/sync_stages.py:            with man._artifacts.get().file.open() as content_file:
pulp_container/app/tasks/tag.py:    artifact = manifest._artifacts.all()[0]
[ipanova@fluffy pulp_container]$
```

#17 - 08/02/2021 07:44 PM - ipanova@redhat.com

- Sprint changed from Sprint 101 to Sprint 102

#18 - 08/06/2021 07:14 PM - dalley

- Related to Issue #9101: Content_artifact is not updated added

#19 - 08/12/2021 05:23 PM - rchan

- Sprint changed from Sprint 102 to Sprint 103

#20 - 08/27/2021 05:08 PM - rchan

- Sprint changed from Sprint 103 to Sprint 104

#21 - 09/10/2021 12:28 AM - rchan

- Sprint changed from Sprint 104 to Sprint 105

#22 - 09/23/2021 11:54 PM - rchan

- Sprint changed from Sprint 105 to Sprint 106

#23 - 10/08/2021 03:16 PM - rchan

- Sprint changed from Sprint 106 to Sprint 107

#24 - 10/21/2021 06:35 PM - rchan

- Sprint changed from Sprint 107 to Sprint 108

#25 - 11/04/2021 10:21 PM - rchan

- Sprint changed from Sprint 108 to Sprint 109

#26 - 11/19/2021 09:39 PM - rchan

- Sprint changed from Sprint 109 to Sprint 110

Files

reproduce_publish_error.py	1.28 KB	02/25/2021	dalley