Pulp - Story #7252

Story # 6134 (CLOSED - CURRENTRELEASE): [EPIC] Pulp import/export

As a plugin writer, I have a way to map Content to Repositories in Pulp exports

07/31/2020 03:20 PM - daviddavis

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	daviddavis	**% Done:**	100%
Category:		**Estimated time:**	0:00 hour
Sprint/Milestone:	3.6.0		
Platform Release:		**Tags:**	
Groomed:	Yes	**Sprint:**	Sprint 79
Sprint Candidate:	No	**Quarter:**	

Description

Problem

Currently, the Pulp import/export code divides up Content (and its related information) into RepositoryVersion folders when exporting. At import time, a new task is launched for each folder to import this Content and automatically create a new version of the Repository with all the Content that gets imported[0].

The problem is that DistributionTree Repositories have Content that is not directly tied to the new RepositoryVersion. Instead, this Content belongs to subrepos.

As pulpcore handles the association of Content to Repositories, pulpcore needs to provide a way for plugin writers to be able to map Content to particular Repositories.

Solutions

The naïve solution would be to export RepositoryContent. But this is somewhat problematic as RepositoryContent has no helpful information: it has a repository id (instead of the repo natural key), a content id (another non-natural key), and version information (which probably won't match the downstream).

Option 1

Create a json dump that matches repository name to a list of content upstream_ids. After the content is imported, the upstream_ids should match the upstream pulp instance and the import code could simply go through and create a new repo version for each repository name. If there are no changes, Pulp will automatically not create a new repo version (which is the current behavior).

Option 2

Add repository name to each exported Content unit. This seems like a simple solution but its implementation is a bit more complex. There are a few obstacles like figuring out how to handle Content units that might belong to multiple Repositories (do we export the Content multiple times?). Also we'd have to tell the Content model resource to populate the field but ignore it during runtime and then extract this mapping of Repository to Content to build our new repo versions.

[0]
https://github.com/pulp/pulpcore/blob/7157e4452bb3e1c480217a037829c34dcbf5bc75/pulpcore/app/tasks/importer.py#L135-L150

Associated revisions

Revision 82c0f147 - 08/10/2020 08:32 PM - daviddavis

Allow import/export of content not directly tied to repos

fixes #7252

History

#1 - 07/31/2020 06:34 PM - daviddavis

Talking this over with ttereshc some and this will probably cause the size of the exports to increase. We should try to make this change optional and only for repos that have subrepos.

#2 - 07/31/2020 08:33 PM - ggainey

- Status changed from NEW to ASSIGNED

- Assignee set to daviddavis

- Groomed changed from No to Yes

- Sprint set to Sprint 78

David and I have spent a number of sessions talking this over. The takeaway for me is, Option-1 above is the most reliable/straightforward way to solve this problem.

We already have the upstream_id available at Content-level, it exists specifically to handle the relinking case where we don't have a 'natural key' we can rely on. If we take into the design the observation made by ttereshc that we only want to do this when we **know** there are subrepos involved (as opposed to 'for every repo', when for most we don't need this) we can be more efficient with our disk-usage and in our execution-time at import.

We've struggled to find a cleaner way to rebuild DistributionTrees on import. The real-world of kickstarts is moderately ugly; it is, I suppose, only appropriate that dealing with them be a little less than elegant as well.

#3 - 08/03/2020 09:55 PM - pulpbot

- Status changed from ASSIGNED to POST

PR: https://github.com/pulp/pulpcore/pull/819

#4 - 08/07/2020 04:35 PM - rchan

- Sprint changed from Sprint 78 to Sprint 79

#5 - 08/10/2020 04:52 PM - daviddavis

- Parent task set to #6134

#6 - 08/10/2020 08:32 PM - daviddavis

- Status changed from POST to MODIFIED

- % Done changed from 0 to 100

Applied in changeset pulpcore|82c0f14730ddd87fe7d6de63e30e7c32ed7ceab0.

#7 - 08/13/2020 10:20 PM - dkliban@redhat.com

- Sprint/Milestone set to 3.6.0

#8 - 08/13/2020 11:39 PM - pulpbot

- Status changed from MODIFIED to CLOSED - CURRENTRELEASE