Pulp - Issue #4265

Repo Feed Url Change does not update lazy_content_catalog

12/13/2018 07:38 PM - crashdummymch

Status:	CLOSED - NOTABUG	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	ggainey	**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:			
Severity:	2. Medium	**Sprint Candidate:**	No
Version:		**Tags:**	Pulp 2
Platform Release:		**QA Contact:**	
Blocks Release:		**Complexity:**	
OS:	CentOS 7	**Smash Test:**	
Backwards Incompatible:	No	**Verified:**	No
Triaged:	Yes	**Verification Required:**	No
Groomed:	No	**Sprint:**	Sprint 56

Description

1. Description
   When updating a pulp rpm repo the lazy content catalog is not updated properly resulting in pulp_streamer continuing to utilize old feed url.

pulp-admin rpm repo update --repo-id myrepo --feed http://myfeedurl
lazy_content_catalog is not updated when a repo feed url is update.

1. syslog
   Dec 13 16:51:50 ip-10-222-253-104 pulp_streamer: pulp.streamer.server:INFO: Trying URL: http://myfeedurl/pub/centos/7/os/x86_64/Packages/tcl-8.5.13-8.el7.x86_64.rpm

1. Workaround

delete all offending entries in the mongodb lazy_content_catalog
Example:
db.lazy_content_catalog.find({url:{$regex:"http://myfeedurl.*"}}).pretty()
db.lazy_content_catalog.deleteMany({url:{$regex:"http://myfeedurl.*"}})

pulp-admin rpm repo sync --force-full on all repo's that had feed url changed
Example:
pulp-admin rpm repo sync run --repo-id myrepo --force-full
pulp-admin rpm repo publish run --repo-id myrepo

Once this is completed a lazy download will work properly without any restarts necessary

History

#1 - 12/20/2018 03:57 PM - rchan

- Project changed from Pulp CLI to Pulp

- Sprint/Milestone deleted (2.18.0)

I'm removing the target milestone of 2.18.0 - that field is reserved for where we plan to fix it. Is this where it was found? And changed the project from CLI to Pulp for triage to determine.

#2 - 12/20/2018 08:04 PM - CodeHeeler

- Triaged changed from No to Yes

#3 - 04/12/2019 09:56 PM - bmbouter

- Status changed from NEW to CLOSED - WONTFIX

#4 - 04/15/2019 10:04 PM - bmbouter

- Tags Pulp 2 added

#5 - 05/31/2019 03:41 PM - ggainey

- Status changed from CLOSED - WONTFIX to ASSIGNED

- Assignee set to ggainey

- Sprint set to Sprint 53

#6 - 06/04/2019 04:48 PM - amacdona@redhat.com

- Sprint changed from Sprint 53 to Sprint 54

#7 - 06/14/2019 10:43 PM - ggainey

LazyCatalogEntries (LCE) have a fully-qualified URL per content-unit.

Example:

{ "_id" : ObjectId("5d03ee1230f252067753d1ba"), "_ns" : "lazy_content_catalog", "path" : "/var/lib/pulp/content/units/rpm/f8/d7cba1691f3bc7e29a7c8966be06d8418de488f4ffa7f84ad472de6d604e9b/zebra-0.1-2.noarch.rpm", "importer_id" : "5d03ee0c30f2520188a220cc", "unit_id" : "d088a227-c260-4491-9866-b0e9fc84c6ef", "unit_type_id" : "rpm", **"url" : "https://repos.fedorapeople.org/pulp/pulp/fixtures/rpm-unsigned/zebra-0.1-2.noarch.rpm"**, "checksum" : "7aa66335d8ebc295d626abc0639135ff6dec6333d4e94e0da69ed720c5fdd5f0", "checksum_algorithm" : "sha256", "revision" : 1, "data" : { } }

When we change the feed, we could make a few assumptions and do surgery on the LCE.url field. The problem would be if the repository at the new feed-url doesn't match the same pattern as the one we're replacing - e.g., old-feed points at a Packages/<letter>/<NEVRA>.rpm repo, new-feed points at a Packages/<NEVRA>.rpm repository. There's no way for code to know how to build the url-path - only the new repository's metadata knows for sure.

The other option is to **delete** all the LCEs for the affected importer when changing a lazy-download-feed, requiring the user to force a complete resync if they change the feed, and let that rebuild the LCEs from scratch. This is essentially the path being used in the proposed workaround.

#8 - 06/17/2019 04:29 PM - dkliban@redhat.com

You have identified an interesting problem. You are correct that Pulp has no way of knowing what hte URL structure is at the new feed. Deleting the Lazy Catalog Entries and telling the user to resync is our best option. However, I am not sure how we are going to tell the user to re-sync.

#9 - 06/17/2019 06:02 PM - ttereshc

I think not all LCEs should be deleted for the repo or the old importer because this way users won't be able to combine on_demand repos.
Maybe it makes sense to update (or remove/create) only LCEs for the units which are provided by a new feed.
What do you think?

#10 - 06/18/2019 09:31 PM - ggainey

- Status changed from ASSIGNED to CLOSED - WONTFIX

@ttereshc:

Maybe it makes sense to update (or remove/create) only LCEs for the units which are provided by a new feed.What do you think

Well initially, my thought was that there's no way for the api-handler to do that, because it requires syncing the new feed's metadata in order to be able to know what the new RPMs are, and doing NEVRA/cksum compares on 'existing' LCEs. Except, it's worse than that :(

After a lot more investigation and a certain amount of tail-chasing, I think that a) your reservations are completely justified, b) the current behavior isn't a bug (mostly), and c) the use-case of "fix a broken on-demand url", while entirely reasonable, is not fixable in Pulp-2.

The root problem(s) here, is that a given feed can be in use in more than one repo, that a given feed can be in use as both on-demand and 'not' in different places, and that a given NEVRA can exist in more than one 'upstream' and more than one pulp-repo, but it is only ever on-disk **once** (and however it got there first, direct or on-demand, and from which upstream feed, 'wins') - all **in addition to** the problem that the directory-layout on a new feed, may not match the directory-layout on the feed an LCE was originally sync'd from, and there is therefore no way for code to 'fix up' the feed-urls in specific LCEs. Ay de mi.

At the moment where one decides "a URL I was using as a feed is broken, I want to fix it" (e.g., the problem from #4798 , which is the issue that sent me down this hole in the first place), LCEs for RPMs from that broken feed can be related to multiple repos. Finding all LCEs from a given feed them therefore requires passing in old-feed from the API (as opposed to 'find all LCEs related to 'this' repo-id). You would then need to change the feed-url **on all repos that are currently using old-feed LCEs**, delete those LCEs, and return to the caller a list of repos that now need to be force-resync'd to get the affected content back.

Except, in pulp-2, the 'current' feed in a repo may be perfectly fine - it's just that **at some point** the user set up to grab content from the now-broken on-demand feed, or even just pointed their repo at a feed that has RPMs with the **same NEVRAs** as the now-broken feed, and which is therefore linked to the LCE's-that-need-to-be-removed.

The human-assumption is an implicit 'I am getting a given NEVRA into one repository from one place'. The architectural reality is that NEVRA-to-feed-to-repo, are all many-to-many relationships, not one-to-one - and the current API doesn't let you manage them that way in Pulp-2, and in fact **cannot** let you manage them that way without some seriously low-level architectural refactoring.

Pulp-3's multiple-remotes is better, but will, I suspect, also be bitten by this gnarly problem; @dkliban has promised a pulp-3 issue to be opened.

The net of several days of digging here, is that I am going to re-close this as WONTFIX, and Pulp-3 discussions need to happen to figure out the best way to address.

If anyone can see a reasonable way to address, in Pulp-2, the use case of 'an on-demand feed URL is broken and I didn't get a chance to D/L all the RPMs first', in the multi-repo/multi-feed/same-NEVRA-from-more-than-one-place situation, I would be perfectly happy to reopen this issue and fixit.

#11 - 06/18/2019 10:28 PM - dkliban@redhat.com

An idea that came up during discussion on IRC with @ggainey and @jsherril

The use case would look like this:

1) User creates an on_demand repository/importer with a URL
2) User syncs the repository
3) User updates importer config with a new URL
4) User force-syncs the repository

The update importer task (3) would need to record that the importer's URL changed. The sync task (4) would see that the URL changed for the importer and update all the Lazy Catalog Entries associated with the RPMs discovered during the sync.

#12 - 06/18/2019 11:41 PM - dkliban@redhat.com

The list of RPMs that get downloaded or at least have LazyCatalogEntries created for them is generated here[0] and here[1]. If Pulp checked here[2] for a change in the importer URL, and only called

```
values.discard(unit.unit_key_as_named_tuple)
```

if the URL did not change, then it would be able to later use the list to update the Lazy Catalog Entries for all those units.

The opportunity to remove the old LazyCatalogEntry is here[3] and here[4]. In cases where URL has changed, the sync task would need to call

```
catalog.delete(unit)
```

before calling

```
catalog.add(unit, path)
```

.

[0] https://github.com/pulp/pulp_rpm/blob/2-master/plugins/pulp_rpm/plugins/importers/yum/sync.py#L689
[1] https://github.com/pulp/pulp_rpm/blob/2-master/plugins/pulp_rpm/plugins/importers/yum/sync.py#L733
[2] https://github.com/pulp/pulp_rpm/blob/2-master/plugins/pulp_rpm/plugins/importers/yum/existing.py#L153
[3] https://github.com/pulp/pulp_rpm/blob/2-master/plugins/pulp_rpm/plugins/importers/yum/sync.py#L892
[4] https://github.com/pulp/pulp_rpm/blob/2-master/plugins/pulp_rpm/plugins/importers/yum/sync.py#L836

#13 - 06/21/2019 03:20 PM - ggainey

- Status changed from CLOSED - WONTFIX to ASSIGNED

#14 - 06/21/2019 03:26 PM - ttereshc

- Sprint changed from Sprint 54 to Sprint 55

#15 - 07/08/2019 10:21 PM - ggainey

- Status changed from ASSIGNED to POST

2 PRs needed:

pulp: https://github.com/pulp/pulp/pull/3938
pulp_rpm: https://github.com/pulp/pulp_rpm/pull/1397

#16 - 07/12/2019 03:28 PM - dkliban@redhat.com

- Sprint changed from Sprint 55 to Sprint 56

#17 - 08/12/2019 09:58 PM - ggainey

- Status changed from POST to CLOSED - NOTABUG

OK, so after a lot of investigation/discussion/experimentation, I am withdrawing the proposed changesets and closing this as "NOTABUG". When you change an on-deman feed in Pulp2, you have to **sync the repo** once before Pulp has enough information to change the catalog-entries to start streaming content from the new feed-location. --update doesn't do that initial sync for you.

If anyone on this issue can find a reproducer that shows updating an on-demand feed, syncing, and **then** retrieving a NEVRA for the first time and seeing it come from the old-feed, please attach it here and re-open this issue.