RPM Support - Issue #1843

Pulp publishes invalid PULP_DISTRIBUTION.xml metadata

04/13/2016 09:59 PM - jcline@redhat.com

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	High	**Due date:**	
Assignee:	jcline@redhat.com	**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:			
Severity:	3. High	**Groomed:**	No
Version:	2.8.0	**Sprint Candidate:**	No
Platform Release:	2.8.3	**Tags:**	Pulp 2
OS:		**Sprint:**	Sprint 1
Triaged:	Yes	**Quarter:**	

Description

If a repository contains a PULP_DISTRIBUTION.xml metadata file, it is possible for Pulp to re-publish it with invalid data. This causes a second Pulp server syncing from the first to fail. Specifically, files are referenced in the PULP_DISTRIBUTION.xml file that do no exist in the version published by Pulp[0] (but do exist upstream).

For example, the RHEL6[2] kickstart repository contains a PULP_DISTRIBUTION.xml file that references `repodata/productid`. During sync this is downloaded along with the XML file, but when the repository is published, it is explicitly skipped.

Ultimately, this occurs because Pulp blindly syncs and publishes this PULP_DISTRIBUTION.xml file[1] while filtering content retrieved using it.

To fix this, we should be generating/altering the PULP_DISTRIBUTION.xml file we publish to ensure we don't create invalid metadata. However, a bigger question is whether or not filtering content[0] is even appropriate. I suspect it is not. This issue is not meant to address that problem, though.

[0] https://github.com/pulp/pulp_rpm/blob/pulp-rpm-2.8.2-1/plugins/pulp_rpm/plugins/distributors/yum/publish.py#L796-L797
[1] https://github.com/pulp/pulp_rpm/blob/pulp-rpm-2.8.2-1/plugins/pulp_rpm/plugins/importers/yum/parse/treeinfo.py#L437-L441
[2] https://cdn.redhat.com/content/dist/rhel/server/6/6Server/x86_64/kickstart/

Associated revisions

Revision 9f97669b - 04/19/2016 03:45 PM - Jeremy Cline

Regenerate PULP_DISTRIBUTION.xml on publish if necessary

The PULP_DISTRIBUTION.xml file used to be saved from an upstream repository and republished without modification. This is problematic because files referenced by that file are filtered out during a publish. This commit is a short-term work-around to that problematic workflow. Without it, Pulp (or anything else using PULP_DISTRIBUTION.xml) will attempt to download files that don't exist in the published repository.

fixes #1843

History

#1 - 04/14/2016 07:05 AM - mmccune@redhat.com

- Severity changed from 2. Medium to 3. High

- Version set to 2.8.0

this is fairly severe in that it breaks a good porting of RHEL provisioning. moved to High severity

#2 - 04/14/2016 03:34 PM - jcline@redhat.com

- Description updated

#3 - 04/14/2016 05:20 PM - jcline@redhat.com

- Subject changed from Pulp-to-pulp distribution syncing is almost certainly broken in some cases to Pulp publishes invalid PULP_DISTRIBUTION.xml metadata

- Description updated

- Status changed from NEW to ASSIGNED

- Assignee set to jcline@redhat.com

I've re-written the issue to narrow the focus, since the original was very broad. There are already several known issues with distributions (issue #1768 which was only a very short-term fix and doesn't address the incorrect modeling and #1769 which describes content we fail to mirror).

I intend to ensure Pulp doesn't publish metadata that references files that doesn't exist. However, it may be that it won't reference files that *need* to exist. I don't know what is using (or not using) `repodata/productid` and I find it troubling that we don't mirror upstream, but I don't think I should to tackle all the problems we have as part of this issue.

#4 - 04/14/2016 05:43 PM - mhrivnak

A simple work-around that would improve, but not fix the situation, would be to do the same filtering during sync that we do during publish. Then at least pulp deployments with that change would happily ignore the same files that publish ignores.

As you point out, a better option is to modify the XML at publish time to filter out any files that don't actually get published. This would be more effort, but is still very doable.

And of course the best option would require figuring out why exactly pulp ignores those files, document that somewhere (at least in the code if not elsewhere), and determine if skipping those files is in fact appropriate.

To unblock katello, perhaps a combination of the first two would be valuable. You could probably make a PR for the first work-around very quickly, and then follow with the second option shortly thereafter. That would buy us time to further investigate why pulp is doing this at all. What do you think of that?

#5 - 04/14/2016 10:40 PM - mhrivnak

- Priority changed from Normal to High

- Sprint/Milestone set to 19

- Platform Release set to 2.8.3

#6 - 04/15/2016 05:27 PM - mhrivnak

- Triaged changed from No to Yes

#7 - 04/15/2016 09:47 PM - jcline@redhat.com

- Status changed from ASSIGNED to POST

https://github.com/pulp/pulp_rpm/pull/846

Note that the first suggested work-around in note 4 isn't possible because it would break lazy syncs.

#8 - 04/19/2016 04:01 PM - Anonymous

- Status changed from POST to MODIFIED

- % Done changed from 0 to 100

Applied in changeset 9f97669b4227a948fb5235ebf05eef478caf7a6c.

#9 - 04/27/2016 12:39 AM - semyers

- Status changed from MODIFIED to 5

#10 - 05/06/2016 04:52 PM - pthomas@redhat.com

- Status changed from 5 to 6

verified

```
[root@ibm-x3250m4-03 ~]# pulp-admin rpm repo sync run --repo-id rhel6
+----------------------------------------------------------------------+
                    Synchronizing Repository [rhel6]
+----------------------------------------------------------------------+

This command may be exited via ctrl+c without affecting the request.

Downloading metadata...
[|]
... completed

Downloading repository content...
[-]
[==================================================] 100%
RPMs:       0/0 items
Delta RPMs: 0/0 items

... completed

Downloading distribution files...
[==================================================] 100%
Distributions: 0/0 items
... completed

Importing errata...
[-]
... completed

Importing package groups/categories...
[-]
... completed

Cleaning duplicate packages...
[-]
... completed

Task Succeeded

Copying files
[-]
... completed

Initializing repo metadata
[-]
... completed

Publishing Distribution files
[|]
... completed

Publishing RPMs
[/]
... completed

Publishing Delta RPMs
```

```
... skipped

Publishing Errata
[-]
... completed

Publishing Comps file
[==================================================] 100%
212 of 212 items
... completed

Publishing Metadata.
[-]
... completed

Closing repo metadata
[-]
... completed

Generating sqlite files
... skipped

Publishing files to web
[\]
... completed

Writing Listings File
[-]
... completed

Writing Listings File
[-]
... completed

Task Succeeded
```

#11 - 05/17/2016 09:32 PM - semyers

- Status changed from 6 to CLOSED - CURRENTRELEASE

#13 - 03/08/2018 07:21 PM - bmbouter

- Sprint set to Sprint 1

#14 - 03/08/2018 07:21 PM - bmbouter

- Sprint/Milestone deleted (19)

#15 - 04/15/2019 10:31 PM - bmbouter

- Tags Pulp 2 added