RPM Support - Story #4527

Improve performance of rpm duplicate nevra check

03/11/2019 05:09 AM - rmcgover

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	rmcgover	**% Done:**	100%
Category:		**Estimated time:**	0:00 hour
Sprint/Milestone:	2.19.0		
Platform Release:	2.19.0	**Tags:**	Pulp 2
Groomed:	Yes	**Sprint:**	Sprint 50
Sprint Candidate:	No	**Quarter:**	

Description

In current versions of Pulp 2.x, uploading an RPM to a repo will remove other RPMs with the same NEVRA.

Currently, we are upgrading from an old version of Pulp 2.7, and I've found that performance of import_uploaded_unit tasks for RPMs has regressed significantly. In Pulp 2.7, imports would usually take around 0.5s. In Pulp 2-master, imports to the same repos have taken from 8 to 130 seconds, depending on the size of the repo.

By debugging I've found most of the time is spent in this duplicate check (remove_unit_duplicate_nevra).

This issue is for improving the performance of remove_unit_duplicate_nevra to reduce the severity of the performance regression.

Related issues:

Copied to RPM Support - Test #4566: Improve performance of rpm duplicate nevr...	**CLOSED - COMPLETE**

Associated revisions

Revision 3bfdbd84 - 03/11/2019 05:11 AM - rmcgover

Improve performance of remove_unit_duplicate_nevra

This function, which is used whenever a new RPM is uploaded, was slower than necessary.

The old implementation would first use find_repo_content_units, which queries for all unit IDs of the required type in the repo and then performs a unit query combining the relevant NEVRA with the IDs.

In fact finding all those unit IDs is measurably slow for a large repo, and it's much faster (while still correct) to simply search for the RPMs to remove directly. It's harmless if this finds some RPMs with same NEVRA which are already not in the repo.

On our installation, for a repo with ~24000 RPMs, this reduced the runtime of this method from ~8 seconds to <0.1 seconds.

fixes #4527 https://pulp.plan.io/issues/4527

History

#1 - 03/11/2019 05:17 AM - rmcgover

Pull request: https://github.com/pulp/pulp_rpm/pull/1297

#2 - 03/11/2019 05:18 AM - rmcgover

- Status changed from ASSIGNED to POST

#3 - 03/13/2019 03:22 PM - ttereshc

- Groomed changed from No to Yes

- Sprint set to Sprint 50

#4 - 03/13/2019 03:24 PM - rmcgover

- Status changed from POST to MODIFIED

- % Done changed from 0 to 100

Applied in changeset 3bfdbd849aaad46cbe98bb9ba7d1dd16d63b5726.

#5 - 03/14/2019 05:16 PM - ttereshc

- Platform Release set to 2.19.0

#6 - 03/14/2019 05:23 PM - ttereshc

- Sprint/Milestone set to 2.19.0

#8 - 03/20/2019 10:03 AM - ttereshc

- Status changed from MODIFIED to 5

#9 - 03/25/2019 01:53 PM - bherring

- Copied to Test #4566: Improve performance of rpm duplicate nevra check added

#10 - 04/02/2019 10:52 PM - ttereshc

- Status changed from 5 to CLOSED - CURRENTRELEASE

#11 - 04/15/2019 10:00 PM - bmbouter

- Tags Pulp 2 added