Pulp - Refactor #5701

Performance improvement in remote duplicates

11/13/2019 04:15 AM - dalley

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	dalley	**% Done:**	100%
Category:		**Estimated time:**	0:00 hour
Sprint/Milestone:	3.0.0		
Platform Release:		**Tags:**	
Groomed:	No	**Sprint:**	Sprint 63
Sprint Candidate:	No	**Quarter:**	

Description

The current implementation of the "Remove Duplicates" functionality is probably lacking in efficiency. It looks like this:

```
query_for_repo_duplicates_by_type = defaultdict(lambda: Q())
for item in repository_version.added():
    detail_item = item.cast()

    if detail_item.repo_key_fields == ():
        continue
    unit_q_dict = {
        field: getattr(detail_item, field) for field in detail_item.repo_key_fields
    }
    item_query = Q(**unit_q_dict) & ~Q(pk=detail_item.pk)
    query_for_repo_duplicates_by_type[detail_item._meta.model] |= item_query

for model in query_for_repo_duplicates_by_type:
    _logger.debug(_("Removing duplicates for type: {}".format(model)))
    qs = model.objects.filter(query_for_repo_duplicates_by_type[model])
    repository_version.remove_content(qs)
```

While I haven't measured the exact impact, the individual item.cast() for each item is probably quite expensive. What would likely improve the situation is one of the following:

Proposal 1:

1. Sort these into groups based on their pulp_type which is present on the master Content model.
2. Look up the detail content models that represent the pulp_type strings
3. Query the detail content models directly, in bulk, provided a list of PKs, instead of cast() individually
4. Then within each type group check for duplicates

Proposal 2:

Alternately, each repository can list all of the content types it supports, which would allow us to skip item 2 above (maybe item 1 also) and would allow us to provide an extra layer of protection around making sure you can't have e.g. file content in an RPM repository which we can't easily or centrally guarantee otherwise.

Related issues:

Related to RPM Support - Issue #5688: Large memory consumption when syncing R...	**CLOSED - CURRENTRELEASE**

Associated revisions

Revision fb3cda3c - 12/06/2019 08:00 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision bfc50d61 - 12/06/2019 08:03 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision 8acbb8a0 - 12/06/2019 08:03 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision 3caded8c - 12/06/2019 08:45 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision 3caded8c - 12/06/2019 08:45 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision 1af38824 - 12/06/2019 09:05 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision 140d8495 - 12/06/2019 09:17 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701

Revision 0604a7b4 - 12/07/2019 08:12 PM - dalley

Add additional content type verification to RepositoryVersion

re: #5701 https://pulp.plan.io/issues/5701

Revision 1861257b - 12/08/2019 05:11 PM - dalley

Improve performance of removing duplicates

closes: #5701 https://pulp.plan.io/issues/5701

Revision f4f190c2 - 12/10/2019 03:21 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701 (cherry picked from commit 140d849544fa2467ba30db16e6b5163274a4b3d0)

Revision cc2fb46a - 12/11/2019 03:26 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701 (cherry picked from commit 3caded8c2e6a5d3be5be218713741e7e4ae515ed)

Revision cc2fb46a - 12/11/2019 03:26 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701 (cherry picked from commit 3caded8c2e6a5d3be5be218713741e7e4ae515ed)

Revision 6b2af564 - 12/11/2019 03:52 PM - dalley

Add additional content type verification to RepositoryVersion

re: #5701 https://pulp.plan.io/issues/5701 (cherry picked from commit 0604a7b41cc5c394b9769d13f31fd0744d6d0ca7)

Revision 66dc8749 - 12/11/2019 03:53 PM - dalley

Improve performance of removing duplicates

closes: #5701 https://pulp.plan.io/issues/5701 (cherry picked from commit 1861257bee238b91488101c1ae257484fa48ab87)

Revision 32c1d08f - 12/11/2019 04:07 PM - dalley

Add CONTENT_TYPES to repo definition

Required PR: https://github.com/pulp/pulpcore/pull/441

re: #5701 https://pulp.plan.io/issues/5701 (cherry picked from commit fb3cda3c7314d0d7c5bb3e0c1e440808959be191)

History

#1 - 11/13/2019 04:16 AM - dalley

- Description updated

#2 - 12/03/2019 04:36 PM - dalley

- Tracker changed from Refactor to Issue

- Severity set to 2. Medium

- Triaged set to No

#3 - 12/03/2019 04:46 PM - fao89

- Tracker changed from Issue to Refactor

- % Done set to 0

#4 - 12/03/2019 04:47 PM - fao89

- Related to Issue #5688: Large memory consumption when syncing RHEL 7 os x86_64 added

#5 - 12/03/2019 06:26 PM - dalley

- Status changed from NEW to ASSIGNED

- Assignee set to dalley

#6 - 12/06/2019 03:33 PM - dalley

- Status changed from ASSIGNED to POST

- Sprint set to Sprint 62

https://github.com/pulp/pulpcore/pull/441
https://github.com/pulp/pulp_file/pull/329
https://github.com/pulp/pulp_rpm/pull/1549
https://github.com/pulp/pulp_python/pull/264
https://github.com/pulp/pulp_maven/pull/27
https://github.com/pulp/pulp_container/pull/25
https://github.com/pulp/pulp_ansible/pull/264

#7 - 12/06/2019 03:43 PM - rchan

- Sprint changed from Sprint 62 to Sprint 63

#8 - 12/09/2019 09:19 PM - dalley

- Status changed from POST to MODIFIED

- % Done changed from 0 to 100

Applied in changeset pulpcore|1861257bee238b91488101c1ae257484fa48ab87.

#9 - 12/13/2019 05:00 PM - bmbouter

- Sprint/Milestone set to 3.0.0

#10 - 12/13/2019 06:31 PM - bmbouter

- Status changed from MODIFIED to CLOSED - CURRENTRELEASE