Pulp - Story #3934

Story # 3202 (CLOSED - CURRENTRELEASE): As a user, I can sync RPM/SRPM/Erratum from a remote Yum/DNF repository

As a plugin writer, I can have a stage that removes duplicates

08/24/2018 02:02 PM - daviddavis

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	amacdona@redhat.com	**% Done:**	100%
Category:		**Estimated time:**	0:00 hour
Sprint/Milestone:	3.0.0		
Platform Release:		**Tags:**	
Groomed:	Yes	**Sprint:**	Sprint 46
Sprint Candidate:	Yes	**Quarter:**	

Description

Problem

Both rpm and docker have a situation where content units can mutate. The DeclarativeVersion will treat a mutated content unit as a new unit and add it to a RepositoryVersion with sync. This is in addition to the previous version (the unmutated one). This effectively adds the same unit to the RepositoryVersion twice.

What would be great is if the older one would be removed by DeclarativeVersion as part of the pipeline in some kind of configurable way.

Solution

Make a new stage called RemoveDuplicates that takes two parameters 'type' and 'field_list' (or tuple). 'type' is the content unit type that the stage should inspect. 'field_list' is the list of field names that needs to be unique within the RepositoryVersion. For example for RPM it will configure this stage with type=pulp_rpm.UpdateRecord and field_list=['id']. A Docker example would use the stage twice, first: `type=pulp_docker.Tag`, `field_list=['name', 'manifest']`; second: `type=pulp_docker.Tag`, field_list=`['name', 'manifest_list']`

This new stage will unassociate any units that are of type=type with the same field names as one of the units emitted in the DeclarativeContent stream. It will be a batching stage, handling batches of units at a time. (Note, batches might perform poorly here, since multiple types may be flowing through the stream.)

The stage can be used directly by plugin writers. This funcitonality will also be added as an option to DeclarativeVersion called remove_duplicates which will take the following form:

```
[{
    'type': 'pulp_rpm.UpdateRecord',
    'field_names': ['id']
}]
```

Notice how the stage takes only 1 duplicate type, but the DeclarativeVersion takes a list of them. The DeclarativeVersion will create one RemoveDuplicates stage for each item in the list, making the pipeline a variable length depending on the data passed into DeclarativeVersion.

These extra stages should be run before the AssociateContent stage.

Related issues:

Blocks RPM Support - Task #3954: Prevent duplicate Package content in repos	**CLOSED - CURRENTRELEASE**
Blocks File Support - Task #4028: Prevent duplicate files in repositories	**CLOSED - CURRENTRELEASE**
Blocks Container Support - Story #4172: Remove duplicate tags from repository...	**CLOSED - CURRENTRELEASE**

History

#1 - 08/24/2018 02:03 PM - daviddavis

- Description updated

#2 - 08/24/2018 02:04 PM - daviddavis

- Description updated

#3 - 08/24/2018 02:10 PM - daviddavis

- Subject changed from Remove duplicate UpdateRecords after performing sync to Remove duplicate UpdateRecords for repos after performing sync

#4 - 08/24/2018 02:10 PM - daviddavis

- Description updated

#5 - 08/24/2018 06:15 PM - bmbouter

- Tracker changed from Task to Story

- Project changed from RPM Support to Pulp

- Subject changed from Remove duplicate UpdateRecords for repos after performing sync to As a plugin writer, I can have a stage that removes duplicates

- Description updated

- Sprint/Milestone deleted (Pulp 3 RPM MVP)

Rewriting to be a generalized core stage.

#6 - 08/24/2018 10:58 PM - bmbouter

In order to work on this, it would be best if we could have a pulp-smash test committed that causes a mutated erratum associated to a repo version in addition to the original, unmutated erratum.

#7 - 08/27/2018 01:02 PM - daviddavis

bmbouter, agreed. Is there a pulp 2 smash test for this scenario that we could re-use?

#8 - 08/30/2018 04:39 PM - daviddavis

- Blocks Task #3954: Prevent duplicate Package content in repos added

#9 - 09/19/2018 09:24 PM - daviddavis

- Blocks Task #4028: Prevent duplicate files in repositories added

#10 - 10/05/2018 03:12 PM - daviddavis

Two comments on this:

- I think that this needs to accept a list of fields instead of a single field. Consider the case of duplicate rpms which are unique by nevra (5 fields) or docker tags (3 fields: name, manifest__pk, manifest_list__pk) as a docker repo could have two tags with the same name (one for a manifest and one for a manifest_list).
- Also, I wonder if this field_list should be defined on the content class kind of like how we define the natural key uniqueness on Content now.

#11 - 11/28/2018 04:27 PM - amacdona@redhat.com

- Related to Story #4172: Remove duplicate tags from repository during sync added

#12 - 11/28/2018 04:40 PM - amacdona@redhat.com

- Description updated

- Groomed changed from No to Yes

- Sprint Candidate changed from No to Yes

#13 - 11/28/2018 05:34 PM - jortel@redhat.com

- Sprint set to Sprint 46

#14 - 11/28/2018 05:44 PM - amacdona@redhat.com

- Related to deleted (Story #4172: Remove duplicate tags from repository during sync)

#15 - 11/28/2018 05:44 PM - amacdona@redhat.com

- Blocks Story #4172: Remove duplicate tags from repository during sync added

#16 - 12/04/2018 04:28 PM - amacdona@redhat.com

- Status changed from NEW to ASSIGNED

- Assignee set to amacdona@redhat.com

#17 - 12/05/2018 02:45 PM - amacdona@redhat.com

- Tags Pulp 3 RC Blocker added

#18 - 12/07/2018 02:07 PM - amacdona@redhat.com

- Status changed from ASSIGNED to POST

https://github.com/pulp/pulpcore-plugin/pull/7

#19 - 12/07/2018 10:19 PM - amacdona@redhat.com

- Status changed from POST to MODIFIED

- % Done changed from 0 to 100

Applied in changeset commit:pulpcore-plugin|c320a0d1bce8cd68ff5cc56d1f6fb023ff72ad64.

#20 - 04/25/2019 06:45 PM - daviddavis

- Sprint/Milestone set to 3.0.0

#21 - 04/26/2019 10:34 PM - bmbouter

- Tags deleted (Pulp 3, Pulp 3 RC Blocker)

#22 - 12/13/2019 06:10 PM - bmbouter

- Status changed from MODIFIED to CLOSED - CURRENTRELEASE