Pulp - Issue #6347

File content list does not return all unique content units

03/16/2020 07:23 PM - sajha

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	fao89	**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:	3.3.0		
Severity:	2. Medium	**Groomed:**	No
Version:		**Sprint Candidate:**	No
Platform Release:		**Tags:**	Katello
OS:		**Sprint:**	Sprint 70
Triaged:	Yes	**Quarter:**	

Description

https://pulp-file.readthedocs.io/en/latest/restapi.html#operation/content_file_files_list does not return all unique content units when used with a page size of ~2000 and there are multiple pages of results. To reproduce:

1. Create and sync a large file repo (http://quartet.usersys.redhat.com/pub/fake-repos/large_file/)
2. Retrieve content units for the repo using page_size 2000
3. Out of 70,000 content units, roughly 69k unique records are returned.

Several units are repeated across pages leading to this discrepancy since we only query till offset 68000 with the 2000 page_size and 70k results don't consist of 70k unique units.

Associated revisions

Revision 5ef9ee83 - 04/08/2020 05:38 PM - Fabricio Aguiar

Fixed non unique content units on content list

https://pulp.plan.io/issues/6347 closes #6347

Revision 937fbddf - 04/08/2020 11:17 PM - Fabricio Aguiar

Fixed non unique content units on content list

https://pulp.plan.io/issues/6347 closes #6347

(cherry picked from commit 5ef9ee839bb23d15f0754e44e53358789ce57eb0)

Revision 8f6fb1bd - 04/14/2020 10:17 PM - Fabricio Aguiar

StableOrderingFilter only for NamedModelViewSet

https://pulp.plan.io/issues/6347 fixes #6347

History

#1 - 03/17/2020 03:32 PM - fao89

- Triaged changed from No to Yes

- Sprint set to Sprint 68

#2 - 03/20/2020 04:21 PM - rchan

- Sprint changed from Sprint 68 to Sprint 69

#3 - 03/25/2020 05:53 PM - jsherril@redhat.com

- Tags Katello-P1 added

- Tags deleted (Katello-P2)

#4 - 03/25/2020 05:58 PM - dkliban@redhat.com

This may be a problem for other plugins such as pulp_container. We should see if we can add the sort to all content by default.

#5 - 04/02/2020 10:11 PM - dkliban@redhat.com

All Content models should be sorted by the 'pulp_created' field.

#6 - 04/03/2020 01:54 PM - lmjachky

- Status changed from NEW to ASSIGNED

- Assignee set to lmjachky

#7 - 04/03/2020 04:41 PM - dkliban@redhat.com

- Project changed from File Support to Pulp

- Sprint/Milestone set to 3.3.0

#8 - 04/03/2020 06:13 PM - rchan

- Sprint changed from Sprint 69 to Sprint 70

#9 - 04/04/2020 12:38 AM - lmjachky

- File test.sh added

- Status changed from ASSIGNED to NEW

- Assignee deleted (lmjachky)

I was able to reproduce the issue. I used the attached test to do so. The output of the test looked like this:

```
2000: 4000
4000: 6000
6000: 8000
8000: 10000
10000: 12000
12000: 14000
14000: 16000
16000: 18000
18000: 20000
20000: 22000
22000: 24000
24000: 25295
26000: 26597
28000: 27908
30000: 29271
32000: 30531
34000: 31868
36000: 33239
38000: 34526
40000: 35833
42000: 37146
44000: 38486
46000: 39804
48000: 41137
50000: 42435
52000: 43745
54000: 45055
56000: 46362
58000: 47703
```

```
60000: 49003
62000: 50342
64000: 51626
66000: 52978
68000: 54312
70000: 54312
Expected: 70000
Actual: 54312
```

The content started to be non-unique after 24,000 fetched units (the number probably depends upon a running system). And it did not matter whether I was retrieving the content units using page_size=2,000 or page_size=200.

I tried to sort synced content by the field "pulp_created" (as dkliban had mentioned) using the built-in Django ordering options (https://docs.djangoproject.com/en/3.0/ref/models/options/#django.db.models.Options.ordering), but with no success. I added "ordering = ['-pulp_created']" to the meta class of the model FileContent, however, the sync failed with the following error message:

```
"traceback": "  File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/rq/worker.py\", line 884, in perform_
job\n
 rv = job.perform()\n  File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/rq/job.py\", line 664, in perf
orm\n
 self._result = self._execute()\n  File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/rq/job.py\", line
670, in _execute\n
   return self.func(*self.args, **self.kwargs)\n  File \"/home/fedora/devel/pulp_file/pulp_file/app/tasks/sync
hronizing.py\", line 45,
in synchronize\n    dv.create()\n  File \"/home/fedora/devel/pulpcore/pulpcore/plugin/stages/declarative_versi
on.py\", line 149,
in create\n    loop.run_until_complete(pipeline)\n  File \"/home/fedora/devel/pulpcore/pulpcore/app/models/rep
ository.py\",
line 753, in __exit__\n    repository.finalize_new_version(self)\n  File \"/home/fedora/devel/pulp_file/pulp_f
ile/app/models.py\",
line 68, in finalize_new_version\n    validate_repo_version(new_version)\n  File
\"/home/fedora/devel/pulpcore/pulpcore/plugin/repo_version_utils.py\", line 138, in validate_repo_version\n

validate_duplicate_content(version)\n  File \"/home/fedora/devel/pulpcore/pulpcore/plugin/repo_version_utils.p
y\",
line 96, in validate_duplicate_content\n    ).distinct(*repo_key_fields).count()\n
File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/django/db/models/query.py\", line 392, in count\n
  return self.query.get_count(using=self.db)\n  File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/djang
o/db/models/sql/query.py\",
 line 504, in get_count\n    number = obj.get_aggregation(using, ['__count'])['__count']\n
 File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/django/db/models/sql/query.py\", line 489,
in get_aggregation\n    result = compiler.execute_sql(SINGLE)\n  File \"/usr/local/lib/pulp/lib64/python3.7/si
te-packages/django/db/models/sql/compiler.py\", line 1133, in execute_sql\n
 cursor.execute(sql, params)\n
 File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/django/db/backends/utils.py\", line 67, in execute\n

  return self._execute_with_wrappers(sql, params, many=False, executor=self._execute)\n
 File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/django/db/backends/utils.py\", line 76, in _execute_
with_wrappers\n
  return executor(sql, params, many, context)\n  File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/djan
go/db/backends/utils.py\",
 line 84, in _execute\n    return self.cursor.execute(sql, params)\n
File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/django/db/utils.py\", line 89, in __exit__\n
   raise dj_exc_value.with_traceback(traceback) from exc_value\n
File \"/usr/local/lib/pulp/lib64/python3.7/site-packages/django/db/backends/utils.py\", line 84, in _execute\n

 return self.cursor.execute(sql, params)\n",

    "description": "SELECT DISTINCT ON expressions must match initial ORDER BY expressions\nLINE 1: SELECT COU
NT(*) FROM (SELECT DISTINCT ON (\"file_filecontent\"...\n
```

I did not experience the above error message during the sync when I added that ordering to the model MasterModel. Still, Pulp did not return unique values. Specifying a default ordering (https://www.django-rest-framework.org/api-guide/filtering/#specifying-a-default-ordering) in FileContentViewSet did not help either way.

The only approach, that was working for me, was to create a custom class that handles the actual ordering, like shown in the commit https://github.com/lubosmj/pulp_file/commit/379ce9298307a4d8886a1cb1f4e2a5bd08906ef7. I just stole the workaround proposed here: https://github.com/encode/django-rest-framework/issues/6886#issuecomment-547120480. What is interesting here is that I did not need to order content by the field "pulp_created" and the test passed as supposed. Yet, It seems like the changes did break some functionality according to Travis

(see the build https://travis-ci.org/github/lubosmj/pulp_file/builds/670764582, only one test failed, but I am not experiencing such a failure in my VM).

Also, there is an option to use a different pagination. Currently, we are using LimitOffsetPagination. There exists CursorPagination (https://www.django-rest-framework.org/api-guide/pagination/#cursorpagination) which requires that there is a unique, unchanging ordering of items in the result set.

#10 - 04/07/2020 08:31 PM - fao89

- Status changed from NEW to ASSIGNED

- Assignee set to fao89

#11 - 04/07/2020 09:28 PM - pulpbot

- Status changed from ASSIGNED to POST

PR: https://github.com/pulp/pulp_file/pull/372

#12 - 04/07/2020 10:57 PM - pulpbot

PR: https://github.com/pulp/pulpcore/pull/634

#13 - 04/07/2020 10:58 PM - fao89

- Assignee changed from fao89 to lmjachky

Lubos did all the work: https://github.com/pulp/pulpcore/pull/634

#14 - 04/08/2020 10:49 PM - Anonymous

- Status changed from POST to MODIFIED

Applied in changeset pulpcore|5ef9ee839bb23d15f0754e44e53358789ce57eb0.

#15 - 04/08/2020 11:17 PM - Anonymous

Applied in changeset pulpcore|937fbddf716fd8e14d0ce65e32d970b77b3955c5.

#16 - 04/13/2020 08:42 PM - pulpbot

PR: https://github.com/pulp/pulpcore/pull/645

#17 - 04/14/2020 04:46 PM - bmbouter

- Status changed from MODIFIED to ASSIGNED

#18 - 04/14/2020 04:46 PM - bmbouter

- Status changed from ASSIGNED to POST

- Assignee changed from lmjachky to fao89

#19 - 04/14/2020 11:33 PM - Anonymous

- Status changed from POST to MODIFIED

Applied in changeset pulpcore|8f6fb1bd5aaa7f9708d9576efb4805426756ffce.

#20 - 04/15/2020 09:55 PM - ttereshc

- Status changed from MODIFIED to CLOSED - CURRENTRELEASE

#21 - 05/08/2020 07:44 PM - ggainey

- Tags Katello added

- Tags deleted (Katello-P1)

Files

test.sh	857 Bytes	04/03/2020	lmjachky