Pulp - Issue #6463

pulp 3.2.1 duplicate key error when sync

04/07/2020 10:05 PM - binlinf0

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	dalley	**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:	3.7.0		
Severity:	3. High	**Groomed:**	No
Version:		**Sprint Candidate:**	No
Platform Release:		**Tags:**	
OS:		**Sprint:**	Sprint 81
Triaged:	Yes	**Quarter:**	

Description

Noticed we have a few errors when sync repos

```
        "error": {
            "description": "duplicate key value violates unique constraint \"core_repositoryversio
n_repository_id_number_3c54ce50_uniq\"\nDETAIL:  Key (repository_id, number)=(59eb02b1-edab-46e3-a
69b-d69a8b314f20, 2) already exists.\n",
```

Please investigate what could cause this.

Related issues:

Related to Pulp - Backport #7737: Backport request: 6463: duplicate key error...	**CLOSED - CURRENTRELEASE**
Related to RPM Support - Backport #7844: Backport version cleanup fix to 3.6	**CLOSED - CURRENTRELEASE**
Has duplicate Pulp - Issue #7220: When a task crashes, the incomplete repo ve...	**CLOSED - DUPLICATE**

Associated revisions

Revision 1851b70e - 09/18/2020 07:51 PM - dalley

Fix duplicate key error after incomplete sync task

closes: #6463 https://pulp.plan.io/issues/6463

History

#1 - 04/14/2020 05:01 PM - bmbouter

Originally I thought I had seen this error before from another dev, but now I realize it' not the same and this is the first report of it. So to help resolve it we need some more info, but unfortunately I'm not sure exactly what to ask for.

- I believe the error is saying that repository already has a version "2" in it. Is that your read also?
- Can you show us the repsitory versions output for that repository?
- Do have any insight into the various operations that were running when this occurred?
- When's the first time you observed it?
- How many times have you observed it and how frequently?
- How many resource managers are you running?

#2 - 04/21/2020 09:46 PM - binlinf0

Task status

```
# ./get /pulp/api/v3/tasks/d5149ed8-225b-4fbe-831f-2bb4d190d8f2/
HTTP/1.1 200 OK
Allow: GET, PATCH, DELETE, HEAD, OPTIONS
Connection: keep-alive
Content-Length: 3826
Content-Type: application/json
Date: Tue, 21 Apr 2020 19:29:30 GMT
Server: nginx/1.16.1
Vary: Accept, Cookie
X-Frame-Options: SAMEORIGIN

{
    "created_resources": [],
    "error": {
        "description": "duplicate key value violates unique constraint \"core_repositoryversion_repository_id_
number_3c54ce50_uniq\"\nDETAIL:  Key (repository_id, number)=(ec123c49-0900-4eb6-a635-e156d9f1cf67, 2) already
 exists.\n",
        "traceback": "  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/rq/worker.py\", line 8
84, in perform_job\n    rv = job.perform()\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/
rq/job.py\", line 664, in perform\n    self._result = self._execute()\n  File \"/opt/utils/venv/pulp/3.7.3/lib
64/python3.7/site-packages/rq/job.py\", line 670, in _execute\n    return self.func(*self.args, **self.kwargs)
\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/pulp_rpm/app/tasks/synchronizing.py\", lin
e 152, in synchronize\n    dv.create()\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/pulp
core/plugin/stages/declarative_version.py\", line 141, in create\n    with self.repository.new_version() as ne
w_version:\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/pulp_rpm/app/models/repository.p
y\", line 75, in new_version\n    version.save()\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-pac
kages/django/db/models/base.py\", line 741, in save\n    force_update=force_update, update_fields=update_field
s)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/django/db/models/base.py\", line 779, in
 save_base\n    force_update, using, update_fields,\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-
packages/django/db/models/base.py\", line 870, in _save_table\n    result = self._do_insert(cls._base_manager,
 using, fields, update_pk, raw)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/django/db/m
odels/base.py\", line 908, in _do_insert\n    using=using, raw=raw)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64
/python3.7/site-packages/django/db/models/manager.py\", line 82, in manager_method\n    return getattr(self.ge
t_queryset(), name)(*args, **kwargs)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/django
/db/models/query.py\", line 1186, in _insert\n    return query.get_compiler(using=using).execute_sql(return_id
)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/django/db/models/sql/compiler.py\", line
1375, in execute_sql\n    cursor.execute(sql, params)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/sit
e-packages/django/db/backends/utils.py\", line 67, in execute\n    return self._execute_with_wrappers(sql, par
ams, many=False, executor=self._execute)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/dj
ango/db/backends/utils.py\", line 76, in _execute_with_wrappers\n    return executor(sql, params, many, contex
t)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/django/db/backends/utils.py\", line 84,
in _execute\n    return self.cursor.execute(sql, params)\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/
site-packages/django/db/utils.py\", line 89, in __exit__\n    raise dj_exc_value.with_traceback(traceback) fro
m exc_value\n  File \"/opt/utils/venv/pulp/3.7.3/lib64/python3.7/site-packages/django/db/backends/utils.py\",
line 84, in _execute\n    return self.cursor.execute(sql, params)\n"
    },
    "finished_at": "2020-04-21T19:29:13.283924Z",
    "name": "pulp_rpm.app.tasks.synchronizing.synchronize",
    "progress_reports": [],
    "pulp_created": "2020-04-21T19:29:13.117971Z",
    "pulp_href": "/pulp/api/v3/tasks/d5149ed8-225b-4fbe-831f-2bb4d190d8f2/",
    "reserved_resources_record": [
        "/pulp/api/v3/repositories/rpm/rpm/ec123c49-0900-4eb6-a635-e156d9f1cf67/",
        "/pulp/api/v3/remotes/rpm/rpm/e18c1386-3f3a-4edf-93ea-73b314da475c/"
    ],
    "started_at": "2020-04-21T19:29:13.231035Z",
    "state": "failed",
    "worker": "/pulp/api/v3/workers/fdd85b92-d77c-446f-af25-552032266b12/"
}

repo versions
# ./get /pulp/api/v3/repositories/rpm/rpm/ec123c49-0900-4eb6-a635-e156d9f1cf67/versions/
HTTP/1.1 200 OK
Allow: GET, HEAD, OPTIONS
Connection: keep-alive
Content-Length: 867
Content-Type: application/json
Date: Tue, 21 Apr 2020 19:30:40 GMT
Server: nginx/1.16.1
Vary: Accept, Cookie
X-Frame-Options: SAMEORIGIN
```

```
{
    "count": 2,
    "next": null,
    "previous": null,
    "results": [
        {
            "base_version": null,
            "content_summary": {
                "added": {
                    "rpm.package": {
                        "count": 928,
                        "href": "/pulp/api/v3/content/rpm/packages/?repository_version_added=/pulp/api/v3/repo
sitories/rpm/rpm/ec123c49-0900-4eb6-a635-e156d9f1cf67/versions/1/"
                    }
                },
                "present": {
                    "rpm.package": {
                        "count": 928,
                        "href": "/pulp/api/v3/content/rpm/packages/?repository_version=/pulp/api/v3/repositori
es/rpm/rpm/ec123c49-0900-4eb6-a635-e156d9f1cf67/versions/1/"
                    }
                },
                "removed": {}
            },
            "number": 1,
            "pulp_created": "2020-02-06T04:13:45.867491Z",
            "pulp_href": "/pulp/api/v3/repositories/rpm/rpm/ec123c49-0900-4eb6-a635-e156d9f1cf67/versions/1/"
        },
        {
            "base_version": null,
            "content_summary": {
                "added": {},
                "present": {},
                "removed": {}
            },
            "number": 0,
            "pulp_created": "2020-02-04T21:09:59.856262Z",
            "pulp_href": "/pulp/api/v3/repositories/rpm/rpm/ec123c49-0900-4eb6-a635-e156d9f1cf67/versions/0/"
        }
    ]
}
```


This happens when we kick off syncing of all 270 repos. It happens every time we run the sync. We have one resource manager with about 10 workers. Some repos are pointing to upstream repos which have the same content.

#3 - 04/21/2020 09:51 PM - binlinf0

I am removing and recreating most of repos to see if it will happens again. I will will leave one repo with failed sync tasks for troubleshooting purpose.

#4 - 04/26/2020 03:57 PM - dkliban@redhat.com

- Status changed from NEW to CLOSED - NOTABUG

I suspect that the initial repositories were created with an earlier version of Pulp that had a bug related to this. Though I have no been able to find a bug in our issue tracker to point at for sure.

Without specific reproduction steps, we are unable to keep this bug open. Feel free to re-open if you figure out how to reproduce the issue.

#5 - 04/30/2020 01:00 PM - dkliban@redhat.com

- Status changed from CLOSED - NOTABUG to NEW

I was able to reproduce this bug by cancelling a sync of a kickstart repo. My database ended up with a repository version that has complete=False. As a result the next() method is not giving the correct next repository version number on subsequent syncs[0].

[0] https://github.com/pulp/pulpcore/blob/master/pulpcore/app/models/repository.py#L630

#6 - 04/30/2020 01:21 PM - dkliban@redhat.com

The RepositoryVersion context manager needs to cleanup any incomplete versions in the __enter__ method. However, that is probably not safe to do if running outside of a task.

https://github.com/pulp/pulpcore/blob/master/pulpcore/app/models/repository.py#L758

#7 - 04/30/2020 07:22 PM - bmbouter

dkliban@redhat.com wrote:

> The RepositoryVersion context manager needs to cleanup any incomplete versions in the __enter__ method. However, that is probably not safe to do if running outside of a task.
>
> https://github.com/pulp/pulpcore/blob/master/pulpcore/app/models/repository.py#L758

Maybe __enter__ should do that, but the design originally had this cleanup occuring in another place and I believe that is currently broken. Here's the original design to handle this case (RQ OOM, or a power issue abruptly halting a worker causing it's __exit__ not to be called).

1. Worker OOMs while working, leaving the db with it's RepositoryVersion having complete=False
2. Worker restarts thanks to systemd, but it receives a new PID and therefore a new worker name
3. Each Pulp worker checks every few seconds for "missing workers". Those are workers who have stopped heartbeating. That check occurs here which calls check_worker_heartbeat.
4. A worker that is shown as offline triggers mark_worker_offline which should provide all necessary cleanup.

I think the issue is mark_worker_offline is not checking for created_resources with complete=False in the tasks its canceling. For this to occur without race conditions we also have to be sure the creation of a resource becomes associated with a task as a created resource in one database transaction.

#8 - 05/04/2020 09:36 AM - mdellweg

I would not check for complete==False in mark_worker_offline, as this is very special to repository versions. Transferring the cleanup duty to the task (the object persisted in the database) by calling some cleanup provided for certain task methods might be better (more scalable).

But as discussed the matter with repository versions is more complicated due to versions also being created in synchronous calls (by some plugins). With that constraint in mind, __enter__ is the first common location for both code paths.

#9 - 05/05/2020 04:42 PM - fao89

- Triaged changed from No to Yes

- Sprint set to Sprint 72

#10 - 05/15/2020 04:16 PM - rchan

- Sprint changed from Sprint 72 to Sprint 73

#11 - 05/29/2020 02:26 PM - rchan

- Sprint changed from Sprint 73 to Sprint 74

#12 - 06/11/2020 10:27 PM - rchan

- Sprint changed from Sprint 74 to Sprint 75

#13 - 06/26/2020 06:04 PM - rchan

- Sprint changed from Sprint 75 to Sprint 76

#14 - 07/10/2020 08:31 PM - rchan

- Sprint changed from Sprint 76 to Sprint 77

#15 - 07/28/2020 12:05 AM - rchan

- Sprint changed from Sprint 77 to Sprint 78

#16 - 08/07/2020 04:34 PM - rchan

- Sprint changed from Sprint 78 to Sprint 79

#17 - 08/24/2020 01:10 PM - rchan

- Sprint changed from Sprint 79 to Sprint 80

#18 - 09/04/2020 05:24 PM - rchan

- Sprint changed from Sprint 80 to Sprint 81

#19 - 09/11/2020 04:54 AM - dalley

- Status changed from NEW to ASSIGNED

- Assignee set to dalley

#20 - 09/11/2020 06:45 PM - dalley

Here's my reproducer script for future reference:

```
http POST $BASE_ADDR/pulp/api/v3/remotes/rpm/rpm/ name=foo url=http://mirror.linux.duke.edu/pub/centos/8/BaseO
S/x86_64/kickstart/
export REMOTE_HREF=$(http $BASE_ADDR/pulp/api/v3/remotes/rpm/rpm/ | jq -r '.results[0] | .pulp_href')

http POST $BASE_ADDR/pulp/api/v3/repositories/rpm/rpm/ name=foo
export REPO_HREF=$(http $BASE_ADDR/pulp/api/v3/repositories/rpm/rpm/ | jq -r '.results[0] | .pulp_href')
```

```
http POST :${REPO_HREF}sync/ remote=$REMOTE_HREF
export TASK_HREF=$(http $BASE_ADDR/pulp/api/v3/tasks/ | jq -r '.results[0] | .pulp_href')

sleep 10
http PATCH :$TASK_HREF state=canceled
sleep 2
http POST :${REPO_HREF}sync/ remote=$REMOTE_HREF
```

#21 - 09/11/2020 07:55 PM - pulpbot

- Status changed from ASSIGNED to POST

PR: https://github.com/pulp/pulpcore/pull/902

#22 - 09/11/2020 07:55 PM - pulpbot

PR: https://github.com/pulp/pulp_rpm/pull/1849

#23 - 09/15/2020 06:06 PM - bmbouter

- Sprint/Milestone set to 3.7.0

#24 - 09/17/2020 04:05 PM - dalley

- Has duplicate Issue #7220: When a task crashes, the incomplete repo version is not cleaned up and leads to duplicate key error when creating new repo versions added

#25 - 09/18/2020 08:55 PM - dalley

- Status changed from POST to MODIFIED

Applied in changeset pulpcore|1851b70ec76d39ac8a05fa2dbbd96d0fc157253a.

#26 - 09/22/2020 09:21 PM - pulpbot

- Status changed from MODIFIED to CLOSED - CURRENTRELEASE

#27 - 10/27/2020 04:40 PM - ttereshc

- Related to Backport #7737: Backport request: 6463: duplicate key error when sync to pulpcore 3.6/pulp-rpm added

#28 - 11/19/2020 05:46 PM - daviddavis

- Related to Backport #7844: Backport version cleanup fix to 3.6 added