

Pulp - Issue #4055

Indexes are not created before a migration is run

10/02/2018 09:55 PM - ttereshc

| | | | |
|---|-------------------------|--------------------------|-----------|
| Status: | CLOSED - CURRENTRELEASE | Start date: | |
| Priority: | Normal | Due date: | |
| Assignee: | ttereshc | Estimated time: | 0:00 hour |
| Category: | | | |
| Sprint/Milestone: | | | |
| Severity: | 2. Medium | Groomed: | No |
| Version: | 2.17.0 | Sprint Candidate: | No |
| Platform Release: | 2.17.1 | Tags: | Pulp 2 |
| OS: | | Sprint: | Sprint 43 |
| Triaged: | No | Quarter: | |
| Description | | | |
| <p>The migration for modulemd data introduces a new collection in the DB but doesn't create indexes before migrating a content. So it's possible to have duplicated modulemd units in the DB. Users would run into a problem during upgrade because pulp-manage-db will try to build missing indexes.</p> | | | |
| <pre>E11000 duplicate key error collection: pulp_database.units_modulemd index: name_1_stream_1_version_1_context_1_arch_1 dup key: { : "django", : "1.6", : 20180307130104, : "c2c572ec", : "noarch" } Traceback (most recent call last): File "/usr/lib/python2.7/site-packages/pulp/server/db/manage.py", line 236, in main return _auto_manage_db(options) File "/usr/lib/python2.7/site-packages/pulp/server/db/manage.py", line 303, in _auto_manage_db migrate_database(options) File "/usr/lib/python2.7/site-packages/pulp/server/db/manage.py", line 148, in migrate_database ensure_database_indexes() File "/usr/lib/python2.7/site-packages/pulp/server/db/manage.py", line 186, in ensure_database_indexes model_class.ensure_indexes() File "/usr/lib/python2.7/site-packages/mongoengine/document.py", line 738, in ensure_indexes collection = cls._get_collection() File "/usr/lib/python2.7/site-packages/mongoengine/document.py", line 210, in _get_collection cls.ensure_indexes() File "/usr/lib/python2.7/site-packages/mongoengine/document.py", line 766, in ensure_indexes collection.create_index(fields, background=background, **opts) File "/usr/lib64/python2.7/site-packages/pymongo/collection.py", line 1380, in create_index self.__create_index(keys, kwargs) File "/usr/lib64/python2.7/site-packages/pymongo/collection.py", line 1290, in __create_index sock_info, cmd, read_preference=ReadPreference.PRIMARY) File "/usr/lib64/python2.7/site-packages/pymongo/collection.py", line 205, in _command read_concern=read_concern) File "/usr/lib64/python2.7/site-packages/pymongo/pool.py", line 211, in command read_concern) File "/usr/lib64/python2.7/site-packages/pymongo/network.py", line 100, in command helpers._check_command_response(response_doc, msg, allowable_errors) File "/usr/lib64/python2.7/site-packages/pymongo/helpers.py", line 189, in _check_command_response raise DuplicateKeyError(errmsg, code, response) DuplicateKeyError: E11000 duplicate key error collection: pulp_database.units_modulemd index: name_1_stream_1_version_1_context_1_arch_1 dup key: { : "django", : "1.6", : 20180307130104, : "c2c572ec", : "noarch" }</pre> | | | |
| <p>The migration should build indexes before migrating content. The migration should handle duplicates properly. See example of how a similar case was handled here.</p> | | | |

Associated revisions

Revision 78a2ddd9 - 10/03/2018 05:32 PM - ttereshc

Create indexes before running any migration

closes #4055 <https://pulp.plan.io/issues/4055>

Revision fce66fb8 - 10/03/2018 06:04 PM - ttereshc

Create indexes before running any migration

closes #4055 <https://pulp.plan.io/issues/4055>

(cherry picked from commit 78a2ddd973981c78758d6bd4eca7c22ced8001e5)

Revision a70c5ca2 - 10/03/2018 07:10 PM - ttereshc

Create indexes before running any migration

closes #4055 <https://pulp.plan.io/issues/4055>

(cherry picked from commit 78a2ddd973981c78758d6bd4eca7c22ced8001e5)

Revision 2ebcb7bf - 05/23/2019 02:19 PM - quba42

Add an optional pre index part for DB migrations

This is needed to provide a way to prepare the data base for an index change BEFORE that index is applied. The new mechanism is subject to certain risks and limitations. It should only be used where absolutely necessary.

ref #4055, #4138 <https://pulp.plan.io/issues/4055><https://pulp.plan.io/issues/4138>

closes #4871 <https://pulp.plan.io/issues/4871>

Revision 97b75fd2 - 05/28/2019 05:08 PM - quba42

Add the distribution field to the DB models

Within a standard Debian repository structure, the term "distribution" refers to the unique string given by the path segment between the "dists/" folder, and some "Release" file (without the trailing slash).

Since each "Release" file in the directory structure is associated with exactly one unique distribution string, the terms "distribution" and "release" can be (and often are) used interchangeably.

The distribution string is most commonly (but not always) given by either the "codename" or the "suite". The pulp_deb implementation prior to this commit, has assumed that the distribution string is always equal to codename, and has therefore imposed a uniqueness constraint on the codename for all releases/distributions within a single repository.

Since upstream repository sources make no such assumption and are not necessarily structured using the codename, this has lead to a plathora of unpredictable and buggy behaviour when synchronizing upstream repositories with 'codename != distribution'.

This change fixes these problems by introducing and using a "distribution" field for both the units_deb_release and units_deb_component collections.

revealed #4871 (depends on the fix for this issue) <https://pulp.plan.io/issues/4871>

ref #3464, #4055 <https://pulp.plan.io/issues/3464><https://pulp.plan.io/issues/4055>

fixes #4138, #4705, #4707 <https://pulp.plan.io/issues/4138><https://pulp.plan.io/issues/4705><https://pulp.plan.io/issues/4707>

History

#1 - 10/02/2018 10:30 PM - ttereshc

- Status changed from NEW to POST

https://github.com/pulp/pulp_rpm/pull/1206/

#2 - 10/03/2018 09:42 AM - ttereshc

- Subject changed from Indexes are not created during a migration of modulemd data to Indexes are not created before a migration of modulemd data

- Description updated

#3 - 10/03/2018 11:02 AM - ttereshc

Is there a reason why we create indexes after running migrations and not before?

For performance reasons in case a new index for a large collection is introduced, so it's not rebuilt for every new migrated item?

<https://github.com/pulp/pulp/blob/2-master/server/pulp/server/db/manage.py#L148>

It's a problem for the case when new collections are created during migrations.

If I explicitly create all the indexes in my migration code, then the platform complains that I'm not allowed to due to [this check](#) . We need to create unique indexes before migrating content, otherwise duplicated records can be created in a collection.

Can indexes be created before migrations are run? Any downsides?

#4 - 10/03/2018 11:03 AM - ttereshc

- Status changed from POST to ASSIGNED

- Assignee set to ttereshc

- Sprint set to Sprint 43

#5 - 10/03/2018 02:49 PM - jortel@redhat.com

ttereshc wrote:

Can indexes be created before migrations are run? Any downsides?

The only potential downside is performance if we had migrations doing high volumes of inserts which I doubt we have. In most cases we are updating existing records and it's more likely that creating indexes first would improve performance for migrations doing searches.

Correctness is most important and think we should create indexes first.

#6 - 10/03/2018 05:37 PM - ttereshc

- Project changed from RPM Support to Pulp

- Subject changed from Indexes are not created before a migration of modulemd data to Indexes are not created before a migration is run

- Status changed from ASSIGNED to POST

<https://github.com/pulp/pulp/pull/3691>

#7 - 10/03/2018 05:39 PM - ttereshc

- Platform Release set to 2.17.1

#8 - 10/03/2018 06:01 PM - ttereshc

- Status changed from POST to MODIFIED

Applied in changeset [pulp|78a2ddd973981c78758d6bd4eca7c22ced8001e5](#).

#9 - 10/03/2018 06:04 PM - ttereshc

Applied in changeset [pulp|fce66fb8f0e60e5967516925f160eeae5ae0fb46](#).

#10 - 10/03/2018 07:10 PM - ttereshc

Applied in changeset [pulp|a70c5ca2a87556ad6c0075931fdb1e9c73b75aa1](#).

#11 - 10/12/2018 08:54 AM - ttereshc

- Status changed from MODIFIED to CLOSED - CURRENTRELEASE

#12 - 04/15/2019 10:08 PM - bmbouter

- Tags Pulp 2 added