Pulp - Issue #723

RPMs with large number of files can exceed mongo document size limit

03/03/2015 04:36 PM - dgregor@redhat.com

Status:	CLOSED - CURRENTRELEASE	**Start date:**	
Priority:	High	**Due date:**	
Assignee:	ttereshc	**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:			
Severity:	3. High	**Groomed:**	No
Version:	2.12.0	**Sprint Candidate:**	No
Platform Release:	2.12.2	**Tags:**	Pulp 2
OS:		**Sprint:**	Sprint 16
Triaged:	Yes	**Quarter:**	

Description

$ rpm -qlp jbossas-javadocs-7.3.0-14.Final_redhat_14.ep6.el5.noarch.rpm | wc -l
38377

Saving this results in:
BSON document too large (18182857 bytes) - the connected server supports BSON document sizes up to 16777216 bytes.

In full disclosure, this is on a customized instance of pulp where we are storing both the sha256 and sha1 copies of the XML metadata. Still, this size limit could hit vanilla upstream pulp with the right RPM.

Perhaps there is a threshold for the number of files in an RPM and anything over that is not stored in the db but generated on the fly during distribution instead. Or the XML fragment could be split out into its own document.

Related issues:

Related to RPM Support - Issue #2747: RPM exceeds mongo document size limit i...	**CLOSED - WONTFIX**
Has duplicate Pulp - Issue #2487: Sync failure - DocumentTooLarge	**CLOSED - DUPLICATE**

Associated revisions

Revision da51b5f3 - 02/28/2017 11:36 PM - ttereshc

Store compressed metadata in DB

This will decrease a probability of hitting maximum document size in MongoDB.

closes #723 https://pulp.plan.io/issues/723

History

#1 - 03/03/2015 08:12 PM - mhrivnak

The community has seen this, I think with chef RPMs, where they basically stuff an entire operating system into one RPM. The "easy" solutions, like putting metadata in its own collection, compressing it, etc. have the unsettling effect of just moving the problem further away.

To fix it robustly, we could either re-generate those snippets at publish time and accept the performance hit, or store them on the filesystem where they can be of any size. In theory we could do fragmentation and still store the pieces in mongo, but that would be painful.

In any case, I suspect we already have an open bug about this.

#2 - 03/10/2015 04:10 PM - dkliban@redhat.com

- Severity set to High

- Triaged changed from No to Yes

#3 - 03/20/2015 08:16 PM - bmbouter

- Severity changed from High to 3. High

#4 - 12/13/2016 05:10 PM - mhrivnak

- Has duplicate Issue #2487: Sync failure - DocumentTooLarge added

#5 - 12/13/2016 05:12 PM - mhrivnak

- Priority changed from Normal to High

Re-adjusting priority based on triage today.

The duplicate issue suggests this can be reproduced by syncing this repo:

https://fedorapeople.org/groups/katello/releases/yum/foreman/1.13/el6/x86_64/

#7 - 12/13/2016 06:53 PM - mhrivnak

- Sprint/Milestone set to 31

#8 - 12/13/2016 10:21 PM - mhrivnak

At one point in the past, I did a PoC where the XML blobs stored on the model were gzip compressed. It worked great. I could not measure any performance degradation in sync or publish. Maybe we should consider doing that.

It does not completely solve the problem, because in theory you could still end up with an XML blob so big that even after compression, it exceeds the mongo document size. But, it would give us a ton more breathing room.

If we do pursue that option, we need to consider whether a migration is warranted. It might be a relatively quick migration, but we'd need testing to be sure. Otherwise we could just compress new entries, and keep track of whether any particular RPM's repodata is compressed.

Ping me if you want to brainstorm any more on this idea.

#9 - 12/16/2016 12:26 AM - ttereshc

- Status changed from NEW to ASSIGNED

- Assignee set to ttereshc

#10 - 12/16/2016 07:01 PM - bmbouter

Based on a quick scan of internet reports of gzip compression we can probably expect at least a 15% reduction worst case. It could be much higher, but it's data dependent. I've not seen any reports where savings were less than 15%. That is pretty good.

+1 to writing the migration which will allow the code to do everything one way.

One other idea to consider. Instead of compressing them, what about spreading the data over more Documents like a singly linked list.

#11 - 01/16/2017 03:06 PM - mhrivnak

- Sprint/Milestone changed from 31 to 32

#12 - 01/20/2017 03:00 AM - ttereshc

- Status changed from ASSIGNED to POST

https://github.com/pulp/pulp_rpm/pull/1018

#13 - 02/06/2017 02:52 PM - mhrivnak

- Sprint/Milestone changed from 32 to 33

#14 - 02/13/2017 01:46 PM - ipanova@redhat.com

- Version set to 2.12.0

#15 - 02/13/2017 01:46 PM - demter@atix.de

just for the record: this seems to also happen with errata sync of
Red_Hat_Software_Collections_RPMs_for_Red_Hat_Enterprise_Linux_7_Server_x86_64_7Server

I get a pulp error like this:

```
pulp_tasks:
- exception:
  task_type: pulp.server.managers.repo.sync.sync
  _href: "/pulp/api/v2/tasks/45c92eda-b5d8-41a7-97d5-124cc28df7d9/"
  task_id: 45c92eda-b5d8-41a7-97d5-124cc28df7d9
  tags:
```

```
  - pulp:repository:ORG-Library-CCV_RHEL7-Red_Hat_Software_Collections__for_RHEL_Server_-Red_Hat_Software_Coll
ections_RPMs_for_Red_Hat_Enterprise_Linux_7_Server_x86_64_7Server
  - pulp:action:sync
  finish_time: '2017-02-13T11:42:22Z'
  _ns: task_status
  start_time: '2017-02-13T11:26:22Z'
  traceback: |
    Traceback (most recent call last):
      File "/usr/lib/python2.7/site-packages/celery/app/trace.py", line 240, in trace_task
        R = retval = fun(*args, **kwargs)
      File "/usr/lib/python2.7/site-packages/pulp/server/async/tasks.py", line 484, in __call__
        return super(Task, self).__call__(*args, **kwargs)
      File "/usr/lib/python2.7/site-packages/pulp/server/async/tasks.py", line 103, in __call__
        return super(PulpTask, self).__call__(*args, **kwargs)
      File "/usr/lib/python2.7/site-packages/celery/app/trace.py", line 437, in __protected_call__
        return self.run(*args, **kwargs)
      File "/usr/lib/python2.7/site-packages/pulp/server/controllers/repository.py", line 810, in sync
        raise pulp_exceptions.PulpExecutionException(_('Importer indicated a failed response'))
    PulpExecutionException: Importer indicated a failed response
  spawned_tasks: []
  progress_report:
    yum_importer:
      content:
        items_total: 0
        state: FINISHED
        error_details: []
        details:
          rpm_total: 0
          rpm_done: 0
          drpm_total: 0
          drpm_done: 0
        size_total: 0
        size_left: 0
        items_left: 0
      comps:
        state: NOT_STARTED
      purge_duplicates:
        state: NOT_STARTED
      distribution:
        items_total: 0
        state: FINISHED
        error_details: []
        items_left: 0
      errata:
        state: FAILED
        error: command document too large
      metadata:
        state: FINISHED
  queue: reserved_resource_worker-0@pulpcapsule2.tld.dq
  state: error
  worker_name: reserved_resource_worker-0@pulpcapsule2.tld
  result:
  error:
    code: PLP0000
    data: {}
    description: Importer indicated a failed response
    sub_errors: []
  _id:
    "$oid": 58a183ed506ac4a71f028e16
  id: 58a183ed506ac4a71f028e16
poll_attempts:
  total: 177
  failed: 1
Error:
```

#16 - 02/13/2017 02:33 PM - ttereshc

@demter, thanks for the report, you hit the issue #2568 where DocumentTooLarge error is happening during sync of errata. In your output you can see that issue is related to errata sync:

```
      errata:
        state: FAILED
        error: command document too large
```

#17 - 02/27/2017 03:55 PM - mhrivnak

- Sprint/Milestone changed from 33 to 34

#18 - 02/28/2017 11:53 PM - ttereshc

- Status changed from POST to MODIFIED

Applied in changeset pulp_rpm:da51b5f326b6faf036cd165365e2dac1a39dd61e.

#19 - 03/06/2017 05:31 PM - semyers

- Platform Release set to 2.12.2

#20 - 03/10/2017 10:16 PM - semyers

- Status changed from MODIFIED to 5

#21 - 04/05/2017 09:12 PM - pthomas@redhat.com

```
[root@hp-dl380pgen8-02-vm-12 ~]# rpm -qa |grep pulp
python-pulp-puppet-common-2.12.2-0.1.beta.el7.noarch
python-pulp-oid_validation-2.12.2-0.1.beta.el7.noarch
pulp-server-2.12.2-0.1.beta.el7.noarch
pulp-docker-plugins-2.3.0-1.el7.noarch
pulp-admin-client-2.12.2-0.1.beta.el7.noarch
python-pulp-ostree-common-1.2.1-0.1.beta.el7.noarch
python-pulp-python-common-1.1.3-1.el7.noarch
python-pulp-bindings-2.12.2-0.1.beta.el7.noarch
python-pulp-common-2.12.2-0.1.beta.el7.noarch
python-kombu-3.0.33-6.pulp.el7.noarch
python-pulp-rpm-common-2.12.2-0.1.beta.el7.noarch
python-pulp-repoauth-2.12.2-0.1.beta.el7.noarch
pulp-rpm-plugins-2.12.2-0.1.beta.el7.noarch
pulp-puppet-plugins-2.12.2-0.1.beta.el7.noarch
python-pulp-client-lib-2.12.2-0.1.beta.el7.noarch
pulp-rpm-admin-extensions-2.12.2-0.1.beta.el7.noarch
pulp-puppet-admin-extensions-2.12.2-0.1.beta.el7.noarch
pulp-ostree-admin-extensions-1.2.1-0.1.beta.el7.noarch
pulp-python-admin-extensions-1.1.3-1.el7.noarch
python-pulp-streamer-2.12.2-0.1.beta.el7.noarch
python-isodate-0.5.0-4.pulp.el7.noarch
python-pulp-docker-common-2.3.0-1.el7.noarch
pulp-selinux-2.12.2-0.1.beta.el7.noarch
pulp-docker-admin-extensions-2.3.0-1.el7.noarch
pulp-ostree-plugins-1.2.1-0.1.beta.el7.noarch
pulp-python-plugins-1.1.3-1.el7.noarch

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin login -u admin -p admin
Successfully logged in. Session certificate will expire at Apr 12 17:23:51 2017
GMT.

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo create --repo-id foreman --feed https://fedorapeople.org/
groups/katello/releases/yum/foreman/1.13/el6/x86_64/
Successfully created repository [foreman]

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo sync run --repo-id foreman
+----------------------------------------------------------------------+
                   Synchronizing Repository [foreman]
+----------------------------------------------------------------------+

This command may be exited via ctrl+c without affecting the request.

Downloading metadata...
[/]
... completed

Downloading repository content...
[-]
[==================================================] 100%
RPMs:       344/344 items
Delta RPMs: 0/0 items

... completed
```

```
Downloading distribution files...
[==================================================] 100%
Distributions: 0/0 items
... completed

Importing errata...
[-]
... completed

Importing package groups/categories...
[-]
... completed

Cleaning duplicate packages...
[-]
... completed

Task Succeeded

Initializing repo metadata
[-]
... completed

Publishing Distribution files
[-]
... completed

Publishing RPMs
[==================================================] 100%
344 of 344 items
... completed

Publishing Delta RPMs
... skipped

Publishing Errata
[\]
... completed

Publishing Comps file
[-]
... completed

Publishing Metadata.
[-]
... completed

Closing repo metadata
[-]
... completed

Generating sqlite files
... skipped

Generating HTML files
... skipped

Publishing files to web
[-]
... completed

Writing Listings File
[-]
... completed

Task Succeeded

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo create --repo-id issue2668 --feed http://download-node-02
.eng.bos.redhat.com/devel/candidates/JBEAP/composing/latest-JBEAP-7.0-RHEL-7/compose/Server/x86_64/os/
Successfully created repository [issue2668]

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo sync run --repo-id issue2668
+----------------------------------------------------------------------+
                   Synchronizing Repository [issue2668]
```

```
+------------------------------------------------------------------------------+

This command may be exited via ctrl+c without affecting the request.

Downloading metadata...
[/]
... completed

Downloading repository content...
[================================================] 100%
RPMs:       296/296 items
Delta RPMs: 0/0 items

... completed

Downloading distribution files...
[================================================] 100%
Distributions: 0/0 items
... completed

Importing errata...
[-]
... completed

Importing package groups/categories...
[-]
... completed

Cleaning duplicate packages...
[-]
... completed

Task Succeeded

Initializing repo metadata
[-]
... completed

Publishing Distribution files
[-]
... completed

Publishing RPMs
[================================================] 100%
296 of 296 items
... completed

Publishing Delta RPMs
... skipped

Publishing Errata
[-]
... completed

Publishing Comps file
[================================================] 100%
2 of 2 items
... completed

Publishing Metadata.
[-]
... completed

Closing repo metadata
[-]
... completed

Generating sqlite files
... skipped

Generating HTML files
... skipped

Publishing files to web
[-]
```

```
... completed

Writing Listings File
[-]
... completed

Task Succeeded

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo create --repo-id rhel7 --feed http://cdn.rcm-internal.red
hat.com/content/dist/rhel/rhui/server/7/7Server/x86_64/os/
Successfully created repository [rhel7]

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo sync run --repo-id rhel7
+----------------------------------------------------------------------+
                    Synchronizing Repository [rhel7]
+----------------------------------------------------------------------+

This command may be exited via ctrl+c without affecting the request.

Downloading metadata...
[\]
... completed

Downloading repository content...
[/]
[==================================================] 100%
RPMs:       11198/11198 items
Delta RPMs: 0/0 items

... completed

Downloading distribution files...
[==================================================] 100%
Distributions: 0/0 items
... completed

Importing errata...
[/]
... completed

Importing package groups/categories...
[\]
... completed

Cleaning duplicate packages...
[\]
... completed

Task Succeeded

Initializing repo metadata
[-]
... completed

Publishing Distribution files
[-]
... completed

Publishing RPMs
[==================================================] 100%
11198 of 11198 items
... completed

Publishing Delta RPMs
... skipped

Publishing Errata
[==================================================] 100%
1260 of 1260 items
... completed

Publishing Comps file
[==================================================] 100%
87 of 87 items
... completed
```

```
Publishing Metadata.
[-]
... completed

Closing repo metadata
[-]
... completed

Generating sqlite files
... skipped

Generating HTML files
... skipped

Publishing files to web
[/]
... completed

Writing Listings File
[-]
... completed

Task Succeeded

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo create --repo-id upload
Successfully created repository [upload]

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo uploads rpm --repo-id upload -f jbossas-javadocs-7.5.14-1
.Final_redhat_1.1.ep6.el6.noarch.rpm -vvv
+----------------------------------------------------------------------+
                            Unit Upload
+----------------------------------------------------------------------+

Extracting necessary metadata for each request...
[==================================================] 100%
Analyzing: jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el6.noarch.rpm
... completed

Files to be uploaded:
  jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el6.noarch.rpm

Creating upload requests on the server...
[==================================================] 100%
Initializing: jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el6.noarch.rpm
... completed

Starting upload of selected units. If this process is stopped through ctrl+c,
the uploads will be paused and may be resumed later using the resume command or
canceled entirely using the cancel command.

Uploading: jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el6.noarch.rpm
[==================================================] 100%
36350612/36350612 bytes
... completed

Importing into the repository...
This command may be exited via ctrl+c without affecting the request.

[|]
Running...

Task Succeeded

Deleting the upload request...
... completed

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo publish run --repo-id upload
+----------------------------------------------------------------------+
                    Publishing Repository [upload]
+----------------------------------------------------------------------+

This command may be exited via ctrl+c without affecting the request.

Initializing repo metadata
```

```
[-]
... completed

Publishing Distribution files
[-]
... completed

Publishing RPMs
[-]
... completed

Publishing Delta RPMs
... skipped

Publishing Errata
[-]
... completed

Publishing Comps file
[-]
... completed

Publishing Metadata.
[-]
... completed

Closing repo metadata
[-]
... completed

Generating sqlite files
... skipped

Generating HTML files
... skipped

Publishing files to web
[-]
... completed

Writing Listings File
[-]
... completed

Task Succeeded

[[root@hp-dl380pgen8-02-vm-12 ~]# ls test/
jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el5.noarch.rpm  jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.
el7.noarch.rpm
jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el6.noarch.rpm
[root@hp-dl380pgen8-02-vm-12 ~]#
[root@hp-dl380pgen8-02-vm-12 ~]#

[root@hp-dl380pgen8-02-vm-12 ~]# pulp-admin rpm repo uploads rpm --repo-id upload --dir test/ -vvv
+----------------------------------------------------------------------+
                            Unit Upload
+----------------------------------------------------------------------+

Extracting necessary metadata for each request...
[==================================================] 100%
Analyzing: jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el7.noarch.rpm
... completed

Files to be uploaded:
  jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el5.noarch.rpm
  jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el6.noarch.rpm
  jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el7.noarch.rpm

Creating upload requests on the server...
[==================================================] 100%
Initializing: jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el7.noarch.rpm
... completed

Starting upload of selected units. If this process is stopped through ctrl+c,
the uploads will be paused and may be resumed later using the resume command or
```

```
canceled entirely using the cancel command.

Uploading: jbossas-javadocs-7.5.14-1.Final_redhat_1.1.ep6.el5.noarch.rpm
[==================================================] 100%
49558304/49558304 bytes
... completed

Importing into the repository...
This command may be exited via ctrl+c without affecting the request.

[/]
Running...

Task Succeeded

Deleting the upload request...
... completed

Uploading: jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el6.noarch.rpm
[==================================================] 100%
36352104/36352104 bytes
... completed

Importing into the repository...
This command may be exited via ctrl+c without affecting the request.

[/]
Running...

Task Succeeded

Deleting the upload request...
... completed

Uploading: jbossas-javadocs-7.5.14-2.Final_redhat_2.1.ep6.el7.noarch.rpm
[==================================================] 100%
31275636/31275636 bytes
... completed

Importing into the repository...
This command may be exited via ctrl+c without affecting the request.

[|]
Running...

Task Succeeded

Deleting the upload request...
... completed

[root@hp-dl380pgen8-02-vm-12 ~]#
```

#22 - 04/12/2017 04:17 PM - bizhang

- Status changed from 5 to CLOSED - CURRENTRELEASE

#23 - 05/05/2017 05:55 PM - ttereshc

- Related to Issue #2747: RPM exceeds mongo document size limit if its filelist > ~15MB added

#24 - 03/08/2018 11:42 PM - bmbouter

- Sprint set to Sprint 16

#25 - 03/09/2018 12:05 AM - bmbouter

- Sprint/Milestone deleted (34)

#26 - 04/15/2019 11:05 PM - bmbouter

- Tags Pulp 2 added