RPM Support - Issue #8893

3rd party repository sync fails with 'InvalidStringData: strings in documents must be valid UTF-8'

06/14/2021 04:14 PM - ggainey

Status:	MODIFIED	**Start date:**	
Priority:	High	**Due date:**	
Assignee:	dalley	**Estimated time:**	0:00 hour
Category:			
Sprint/Milestone:			
Severity:	2. Medium	**Groomed:**	No
Version:		**Sprint Candidate:**	No
Platform Release:		**Tags:**	Katello, Pulp 2
OS:		**Sprint:**	
Triaged:	Yes	**Quarter:**	

Description

ModulemdDefaults are BSON-encoded before being saved into MongoDB because MongoDB (apparently) has restrictions on valid keys which are incompatible with the module data we need to store.

https://github.com/pulp/pulp_rpm/blob/2-master/plugins/pulp_rpm/plugins/importers/yum/repomd/modules.py#L94-L95

...But the serialized BSON strings being created by the encoding function we are using are not always valid UTF-8... Presumably, it works the vast majority of the time, enough so that it wasn't noticed.

Only specific permutations of the input data appear to trigger this. If I delete key from the profiles dictionary, all of a sudden it's UTF-8 compatible. This is presumably why the problem pops into and out of existence.

At some point we attempt to save this string to MongoDB, and Mongo decides to apply a UTF-8 validation to it, and it blows up.

Related issues:

Related to Migration Plugin - Issue #8982: Support migrating any client syste... **CLOSED - CURRENTRELEASE**

Associated revisions

Revision 5c5a7dcc - 07/19/2021 03:28 PM - dalley

Fix saving BSON-encoded data in the database

closes: #8893 https://pulp.plan.io/issues/8893

https://bugzilla.redhat.com/show_bug.cgi?id=1920511

History

#1 - 06/14/2021 04:15 PM - ggainey

From dalley:

```
diff --git a/plugins/pulp_rpm/plugins/importers/yum/repomd/modules.py b/plugins/pulp_rpm/plugins/importers/yum
/repomd/modules.py
index 20aa82eb..4091f858 100644
--- a/plugins/pulp_rpm/plugins/importers/yum/repomd/modules.py
+++ b/plugins/pulp_rpm/plugins/importers/yum/repomd/modules.py
@@ -103,7 +103,7 @@ def _get_profile_defaults(module):
     profile_defaults = {}
     for stream, defaults in module.peek_profile_defaults().items():
         profile_defaults[stream] = defaults.get()
-    return bson.BSON.encode(profile_defaults)
+    return bson.binary.Binary(bson.BSON.encode(profile_defaults))
```

I believe (not 100% sure) this is what we need to do.

And later: ggainey This patch works for me, it fixes the sync. It's based on this line from the documentation.

"Note that, when using Python 2.x, to save binary data it must be wrapped as an instance of bson.binary.Binary. Otherwise it will be saved as a BSON string and retrieved as unicode. Users of Python 3.x can use the Python bytes type."

#2 - 06/14/2021 04:36 PM - dalley

- Triaged changed from No to Yes

This may also require a migration, or potentially not, I'm not sure. We need to make sure that the data coming back out of MongoDB doesn't need to be transformed also.

#3 - 06/29/2021 08:27 PM - dalley

- Status changed from NEW to POST

- Assignee set to dalley

#4 - 07/19/2021 03:29 PM - dalley

- Status changed from POST to MODIFIED

Applied in changeset 5c5a7dcc058b29d89b3a913d29cfcab41db96686.

#5 - 07/19/2021 03:30 PM - dalley

- Related to Issue #8982: Support migrating any client systems that have applied the hotfix for 8982 added