Redmine - Feature #31921

Changes to properly support 4 byte characters (emoji) when database is MySQL

2019-08-17 18:10 - Marius BĂLTEANU

Status:	Closed		**Start date:**	
Priority:	Normal		**Due date:**	
Assignee:	Go MAEDA		**% Done:**	0%
Category:	Database		**Estimated time:**	0.00 hour
Target version:	4.1.0			
Resolution:	Fixed			

Description

Currently, Redmine with MySQL as database doesn't support very well 4 byte characters (for ex: emojis) and you need some experience with MySQL in order to properly configure MySQL (and Redmine) to support those characters. Because of this, there are reported plenty of tickets that I'm going to relate to this issue later.

I would like to work on this in order to improve the default Redmine installation and also, to provide some documentation regarding how you can migrate an existing Redmine installation to support 4 byte characters.

For now, I've created two patches:
1. 0001-Update-default-database-config-for-MySQL.patch
- Adds to the database.yml.example the required encoding and collation, including a note to inform users that are safe only for new installations.
- Adds a test to ensure that an issue can be created using a emoji in description
- Updates some old instructions
- These settings will be default in Rails 6: https://github.com/rails/rails/pull/33608

2. attachment:0002-Task-to-check-mysql-support-for-utf8mb4.patch
This task checks the MySQL configuration (innodb_file_per_table, innodb_large_prefix, innodb_file_format) and ENGINE, ROW_FORMAT and TABLE_COLLATION for Redmine tables. Any feedback or ideas to improve this task are welcome!

I'll try to create later a Wiki page regarding the steps required to migrate and existing installation or at least some useful links.

Related issues:

Related to Redmine - Defect #24242: Adding comments to ticket returns interna...	**Closed**
Related to Redmine - Defect #28774: Internal Error when Submit the Descriptio...	**Closed**
Related to Redmine - Defect #27406: Internal Server Error while posting smile...	**Closed**
Related to Redmine - Defect #25959: "Smile of the death" problem	**Closed**
Related to Redmine - Defect #24992: MailHandler: an unexpected error occurred...	**Needs feedback**
Related to Redmine - Defect #23586: Create index on mysql exceed limits	**Closed**
Related to Redmine - Defect #22119: Error: Data too long for column 'notes' w...	**Closed**
Related to Redmine - Defect #21398: Mysql: 500 server error when submitting 4...	**Closed**
Related to Redmine - Defect #20143: Mailhandler cannot handle 4-byte characters	**New**
Related to Redmine - Patch #19742: RedmineInstall: MySQL: collation_database	**Closed**
Related to Redmine - Defect #18866: MySQL: disappear after 4-Byte UTF-8	**Closed**
Related to Redmine - Defect #30848: Error when creating issue with emoji in d...	**Closed**
Related to Redmine - Defect #27984: Arabic Support Issues	**Closed**
Related to Redmine - Defect #27803: Can't post a smiley face in issues descri...	**Closed**
Related to Redmine - Defect #27361: Failed when using Emoji	**Closed**
Related to Redmine - Defect #27238: Mysql Error after upgrading Redmine from ...	**Closed**
Related to Redmine - Defect #26386: Mysql: Unable to update ticket with Emoji	**Closed**
Related to Redmine - Defect #24030: When SVN or Git repository has a commit c...	**Closed**
Related to Redmine - Defect #23557: Special (micro) character in message fiel...	**Closed**
Related to Redmine - Defect #22618: subject utf-8 char vs mysql2	**Closed**
Related to Redmine - Defect #19334: Error 500 when uploading file with umlaut...	**Closed**
Related to Redmine - Defect #10772: 4-byte utf-8 characters	**Closed**
Related to Redmine - Patch #32054: Add test for 4 byte characters (emoji) sup...	**Closed**

Related to Redmine - Defect #34235: Losing data due to unsupported characters	**Closed**
Has duplicate Redmine - Defect #33649: Ascii code in the 'Description' causes...	**Closed**
Has duplicate Redmine - Defect #34574: Email Fetcher catch no Mails with smiley	**Closed**

Associated revisions

Revision 18454 - 2019-09-12 12:02 - Go MAEDA

Update default database config for MySQL to support 4 byte characters (emoji) (#31921).

Patch by Marius BALTEANU and Go MAEDA.

History

#1 - 2019-08-17 18:11 - Marius BĂLTEANU

- Related to Defect #24242: Adding comments to ticket returns internal server error added

#2 - 2019-08-17 18:11 - Marius BĂLTEANU

- Related to Defect #28774: Internal Error when Submit the Description with Vietnamese in Unicode fonts added

#3 - 2019-08-17 18:12 - Marius BĂLTEANU

- Related to Defect #27406: Internal Server Error while posting smile in issue's description added

#4 - 2019-08-17 18:13 - Marius BĂLTEANU

- Related to Defect #25959: "Smile of the death" problem added

#5 - 2019-08-17 18:13 - Marius BĂLTEANU

- Related to Defect #24992: MailHandler: an unexpected error occurred when receiving email: invalid byte sequence in UTF-8 added

#6 - 2019-08-17 18:14 - Marius BĂLTEANU

- Related to Defect #23586: Create index on mysql exceed limits added

#7 - 2019-08-17 18:15 - Marius BĂLTEANU

- Related to Defect #22119: Error: Data too long for column 'notes' when copy paste pictures added

#8 - 2019-08-17 18:15 - Marius BĂLTEANU

- Related to Defect #21398: Mysql: 500 server error when submitting 4 bytes utf8 (to be saved in the 'notes' field) added

#9 - 2019-08-17 18:15 - Marius BĂLTEANU

- Related to Defect #20143: Mailhandler cannot handle 4-byte characters added

#10 - 2019-08-17 18:18 - Marius BĂLTEANU

- Description updated

#11 - 2019-08-17 18:20 - Marius BĂLTEANU

- Related to Patch #19742: RedmineInstall: MySQL: collation_database added

#12 - 2019-08-17 18:20 - Marius BĂLTEANU

- Related to Defect #18866: MySQL: disappear after 4-Byte UTF-8 added

#13 - 2019-08-17 18:22 - Marius BĂLTEANU

- Related to Defect #30848: Error when creating issue with emoji in description added

#14 - 2019-08-17 18:29 - Marius BĂLTEANU

- Related to Defect #27984: Arabic Support Issues added

#15 - 2019-08-17 18:29 - Marius BĂLTEANU

- Related to Defect #27803: Can't post a smiley face in issues description added

#16 - 2019-08-17 18:29 - Marius BĂLTEANU

- Related to Defect #27361: Failed when using Emoji added

#17 - 2019-08-17 18:29 - Marius BĂLTEANU

- Related to Defect #27238: Mysql Error after upgrading Redmine from 2.5 to 3. added*

#18 - 2019-08-17 18:30 - Marius BĂLTEANU

- Related to Defect #26386: Mysql: Unable to update ticket with Emoji added

#19 - 2019-08-17 18:30 - Marius BĂLTEANU

- Related to Defect #24030: When SVN or Git repository has a commit comment include an emoji (4 bytes charactor), error occurs added

#20 - 2019-08-17 18:30 - Marius BĂLTEANU

- Related to Defect #23557: Special (micro) character in message field causes internal server error added

#21 - 2019-08-17 18:31 - Marius BĂLTEANU

- Related to Defect #22618: subject utf-8 char vs mysql2 added

#22 - 2019-08-17 18:31 - Marius BĂLTEANU

- Related to Defect #19334: Error 500 when uploading file with umlaut in filename from Mac added

#23 - 2019-08-17 18:32 - Marius BĂLTEANU

- Related to Defect #10772: 4-byte utf-8 characters added

#24 - 2019-08-18 04:19 - Go MAEDA

As mentioned in [#19742#note-3](#), the collation utf8mb4_unicode_ci is really problematic for Japanese users. I think it is better to leave it to users without setting a default value. At least, we should not set utf8mb4_unicode_ci as the default value. Or, setting utf8mb4_general_ci or utf8mb4_bin is better.

The problem of utf8mb4_unicode_ci is that it treats many different kinds of characters as the same when comparing. For example, the following combinations of Japanese words are treated as the same words.

- "□□" (means mother, pronounced as "haha") and "□□" (means dad, pronounced as "papa")
- "□□□" (means a crow, pronounced as "karasu") and "□□□" (means glass, pronounced as "garasu")
- "□□□" (means a camera, pronounced as "kamera") and "□□□" (means [Gamera](#), pronounced as "gamera")
- "□□" (menas Paris, pronounced as "pari") and "□□" (means Bali in Indonesia, pronounced as "bari")

Imagine that searching the word "□□□" hits not only "□□□" (a crow) but also "□□□" (glass). Setting the collation to utf8mb4_unicode_ci makes issues filter and full-text search unusable in some languages.

#25 - 2019-08-18 15:33 - Marius BĂLTEANU

- File 0001-Update-default-database-config-for-MySQL.patch added

- File 0002-Task-to-check-mysql-support-for-utf8mb4.patch added

Thanks for your detailed response, I totally missed that note and now I understand the problem.

I've removed collation from default settings.

#26 - 2019-08-19 13:37 - Marius BĂLTEANU

- File deleted (0001-Update-default-database-config-for-MySQL.patch)

#27 - 2019-08-19 13:37 - Marius BĂLTEANU

- File deleted (0002-Task-to-check-mysql-support-for-utf8mb4.patch)

#28 - 2019-08-19 13:37 - Marius BĂLTEANU

- File deleted (0001-Update-default-database-config-for-MySQL.patch)

#29 - 2019-08-19 13:37 - Marius BĂLTEANU

- File deleted (0002-Task-to-check-mysql-support-for-utf8mb4.patch)

#30 - 2019-09-03 09:17 - Marius BĂLTEANU

- File 0001-Update-default-database-config-for-MySQL.patch added

Go Maeda, it is ok from you point of view to set only the encoding to encoding: utf8mb4?

#31 - 2019-09-03 10:03 - Go MAEDA

Marius BALTEANU wrote:

> Go Maeda, it is ok from you point of view to set only the encoding to encoding: utf8mb4?

Yes, I think it is OK.

But I am not sure whether the comment "# Remove encoding for existing installation on MySQL" is necessary.

It would be even better if there is a comment like "Use "utf8" instead of "utf8mb4" if you use MySQL earlier than #{VERSION}".

#32 - 2019-09-04 02:19 - Go MAEDA

I think the change against test/unit/issue_test.rb should not be merged because it may break tests on the official CI server if the encoding of the MySQL database on the CI server is not utf8mb4.
http://www.redmine.org/builds/

No one except Jean-Philippe Lang can update configurations of the CI server.

#33 - 2019-09-05 16:58 - Go MAEDA

I updated Marius's patch:

- Removed "with ruby1.9". The version of Ruby is unnecessary because Redmine no longer supports Ruby 1.8 that require 'mysql' adapter instead of 'mysql2' adapter. Please see the example file in Redmine 2.6 for reference: source:tags/2.6.9/config/database.yml.example
- Removed the test not to break the CI server
- Added the comment "Use "utf8" instead of "utfmb4" for MySQL prior to 5.7.7"
- Removed "# Remove encoding for existing installation on MySQL" because an existing installation may require encoding setting

```diff
diff --git a/config/database.yml.example b/config/database.yml.example
index 57bc51605..727b4d89b 100644
--- a/config/database.yml.example
+++ b/config/database.yml.example
@@ -1,4 +1,4 @@
-# Default setup is given for MySQL with ruby1.9.
+# Default setup is given for MySQL 5.7.7 or later.
 # Examples for PostgreSQL, SQLite3 and SQL Server can be found at the end.
 # Line indentation must be 2 spaces (no tabs).

@@ -8,7 +8,8 @@ production:
   host: localhost
   username: root
   password: ""
-  encoding: utf8
+  # Use "utf8" instead of "utfmb4" for MySQL prior to 5.7.7
+  encoding: utf8mb4

 development:
   adapter: mysql2
@@ -16,7 +17,8 @@ development:
   host: localhost
   username: root
   password: ""
-  encoding: utf8
+  # Use "utf8" instead of "utfmb4" for MySQL prior to 5.7.7
+  encoding: utf8mb4

 # Warning: The database defined as "test" will be erased and
 # re-generated from your development database when you run "rake".
@@ -27,7 +29,8 @@ test:
   host: localhost
   username: root
   password: ""
-  encoding: utf8
+  # Use "utf8" instead of "utfmb4" for MySQL prior to 5.7.7
+  encoding: utf8mb4

 # PostgreSQL configuration example
 #production:
```

#34 - 2019-09-09 19:44 - Marius BĂLTEANU

- Assignee deleted (Marius BĂLTEANU)

- Target version changed from Candidate for next major release to 4.1.0

Go MAEDA wrote:

> I updated Marius's patch:
>
> - Removed "with ruby1.9". The version of Ruby is unnecessary because Redmine no longer supports Ruby 1.8 that require 'mysql' adapter instead of 'mysql2' adapter. Please see the example file in Redmine 2.6 for reference: source:tags/2.6.9/config/database.yml.example
> - Removed the test not to break the CI server
> - Added the comment "Use "utf8" instead of "utfmb4" for MySQL prior to 5.7.7"
> - Removed "# Remove encoding for existing installation on MySQL" because an existing installation may require encoding setting

Thanks for updating my patch, I agree with your changes. Regarding the test, I'll create a new ticket assigned to Jean-Philippe to update the CI server. Until then, I propose to deliver this in 4.1.0.

#35 - 2019-09-12 12:02 - Go MAEDA

- Status changed from New to Closed

- Assignee set to Go MAEDA

- Resolution set to Fixed

Committed the updated patch in #31921#note-33. Thanks.

#36 - 2019-09-12 23:54 - Marius BĂLTEANU

- Related to Patch #32054: Add test for 4 byte characters (emoji) support added

#37 - 2019-12-20 15:38 - Deoren Moor

Marius BALTEANU wrote:

> I would like to work on this in order to improve the default Redmine installation and also, to provide some documentation regarding how you can migrate an existing Redmine installation to support 4 byte characters.
>
> ...
>
> I'll try to create later a Wiki page regarding the steps required to migrate and existing installation or at least some useful links.

I understand that this feature has only *just* released today in v4.1.0, but is there a new page somewhere covering the conversion steps? I looked for it, but didn't spot it.

If the page doesn't exist yet, would you be willing to stub out such a page so that others could opt-in for (Watch) notifications when the page is updated?

Many thanks for taking the time to push this forward.

#38 - 2020-02-05 15:09 - Gérard L

Deoren Moor wrote:

> Marius BALTEANU wrote:
>
> > I would like to work on this in order to improve the default Redmine installation and also, to provide some documentation regarding how you can migrate an existing Redmine installation to support 4 byte characters.
> >
> > ...
> >
> > I'll try to create later a Wiki page regarding the steps required to migrate and existing installation or at least some useful links.
>
> I understand that this feature has only *just* released today in v4.1.0, but is there a new page somewhere covering the conversion steps? I looked for it, but didn't spot it.
>
> If the page doesn't exist yet, would you be willing to stub out such a page so that others could opt-in for (Watch) notifications when the page is updated?
>
> Many thanks for taking the time to push this forward.

+1

#39 - 2020-02-29 01:40 - Martin Rys

With the help of this article I've managed to convert from utf8 to utf8mb4 on a 4.1.0 bitnami install
http://blog.andriylesyuk.com/4-byte-utf-8-problem-in-redmine-and-how-to-fix-it/

Both database.yml and my.cnf were using utf8mb4, so there was nothing to do there for me.

I've already had row_format set to DYNAMIC (meaning InnoDB was already Barracuda and not Antelope, verified via SHOW TABLE STATUS), so I only listed all tables in the database

```
SELECT table_name, CCSA.character_set_name, CCSA.collation_name FROM information_schema.`TABLES` T, informatio
n_schema.`COLLATION_CHARACTER_SET_APPLICABILITY` CCSA WHERE CCSA.collation_name = T.table_coll
ation AND T.table_schema = "bitnami_redmineplusagile";
```

And converted every single table from utf8 to utf8mb4 with the appropriate collation (Czech here since this is a Czech install)

```
...
ALTER TABLE wikis CONVERT TO CHARACTER SET utf8mb4 COLLATE utf8mb4_czech_ci;
ALTER TABLE workflows CONVERT TO CHARACTER SET utf8mb4 COLLATE utf8mb4_czech_ci;
```

After that I verified that everything works by posting a comment in an issue with an emoji

EDIT: Ironically it seems that posting an emoji in here will cut the text off.

#40 - 2020-06-23 12:08 - Holger Just

- Has duplicate Defect #33649: Ascii code in the 'Description' causes 500 error added

#41 - 2021-01-25 00:30 - Marius BĂLTEANU

- Related to Defect #34235: Losing data due to unsupported characters added

#42 - 2021-02-28 17:55 - Marius BĂLTEANU

- Has duplicate Defect #34574: Email Fetcher catch no Mails with smiley added

Files

0001-Update-default-database-config-for-MySQL.patch	2.32 KB	2019-09-03	Marius BĂLTEANU