Passerelle - Bug #65843

BAN : l'import des rues est trop gourmand en SQL

01 juin 2022 10:55 - Thomas Noël

Statut:	Fermé	**Début:**	01 juin 2022
Priorité:	Normal	**Echéance:**	
Assigné à:	Pierre Ducroquet	**% réalisé:**	0%
Catégorie:		**Temps estimé:**	0:00 heure
Version cible:			
Patch proposed:	Oui	**Planning:**	Non

Description

Actuellement chaque mise à jour provoque un update général de la table, alors qu'en fait ça bouge peu.

Exemple de solution : l'import des rues qui télécharge les fichiers BAN pourrait regarder l'entête HTTP « Last-Modified » pour éviter de mettre à jour une base qui n'a pas changé depuis le dernier import.

Autre solution : que ça soit hebdomadaire, avec
une répartition sur les jours de la semaine, pour que tous les tenants ne se mettent pas à jour en même temps (et l'affichage du jour de la mise à jour dans l'UI)
la possibilité de forcer la mise à jour sur l'UI

Demandes liées:

Lié à Passerelle - Bug #66118: TypeError: can't subtract offset-naive and off...	**Fermé**	**10 juin 2022**
Lié à Passerelle - Development #67375: BAN: reprendre l'optimisation de l'imp...	**Rejeté**	**13 juillet 2022**

Révisions associées

Révision 9b487848 - 03 juin 2022 09:25 - Pierre Ducroquet

ban import: use Last-Modified to reduce load (#65843)

Révision c0154bdd - 07 juin 2022 08:44 - Pierre Ducroquet

ban import: update only when needed (#65843)

Révision 58f482cf - 07 juin 2022 08:44 - Pierre Ducroquet

ban import: update streets only when needed (#65843)

Révision 7479f1d1 - 14 juin 2022 16:56 - Pierre Ducroquet

ban import: fix ban_id field (#65843)

Historique

#2 - 01 juin 2022 11:22 - Pierre Ducroquet

- Fichier 0001-ban-import-use-Last-Modified-to-reduce-load-65843.patch ajouté

- Fichier 0002-ban-import-skip-invalid-lines-65843.patch ajouté

- Statut changé de Nouveau à Solution proposée

- Patch proposed changé de Non à Oui

#3 - 01 juin 2022 11:38 - Thomas Noël

- Statut changé de Solution proposée à Solution validée

0001 : ok mais "30 hours" dans le commentaire alors que c'est hours=36

0002 : je suis pas convaincu, à ma connaissance on n'a jamais eu ce problème, attendons qu'il arrive ? (et on fera un ticket ce jour là)

#4 - 01 juin 2022 16:56 - Pierre Ducroquet

- Assigné à mis à Benjamin Dauvergne

C'est pour ça qu'il ne faut jamais commenter le code, après on se plante ... :)

Je suis en train d'ajouter une réduction du nombre de requêtes SQL, donc j'attends un peu avant de soumettre les nouveaux patchs.

#5 - 02 juin 2022 01:05 - Benjamin Dauvergne

- Assigné à changé de Benjamin Dauvergne à Pierre Ducroquet

#6 - 03 juin 2022 10:10 - Pierre Ducroquet

Du coup, nouveaux patchs (rebase et compagnie, ça passe le jenkins sans gueuler)
Le premier reste identique, ne pas charger inutilement les données de la BAN.
Le second est une grosse optimisation de la partie qui charge les données des régions, départements et villes pour ne plus faire de remplacement systématique en base.
Il faudrait encore optimiser le chargement des rues depuis la BAN, mais je pense qu'on peut garder cette optimisation pour un second temps.

#7 - 03 juin 2022 10:11 - Pierre Ducroquet

- Fichier 0001-ban-import-use-Last-Modified-to-reduce-load-65843.patch ajouté

- Fichier 0002-ban-import-update-only-when-needed-65843.patch ajouté

- Statut changé de Solution validée à Solution proposée

#8 - 03 juin 2022 14:18 - Benjamin Dauvergne

Le code de 0002 pourrait être simplifié en faisant quelque chose de la forme :

```
    keys = {'code': 1}
    attributes = {'a': 1, 'b': 2, 'c': 3}
    try:
        Model.objects.get(**keys, **attributes)
    except Model.DoesNotExist:
        Model.objects.update_or_create(**keys, defaults=attributes)
```

#9 - 03 juin 2022 15:43 - Pierre Ducroquet

update_or_create va refaire un select qui a été fait, avant de déterminer s'il vaut mieux faire un insert ou un update (dommage qu'ils ne fassent pas d'upsert)
Certes on va simplifier le code Python, mais on aura SELECT + SELECT FOR UPDATE + INSERT/UPDATE au lieu de SELECT + INSERT/UPDATE.
Vu le volume de mise à jour c'est pas critique, du coup si vous préférez simplifier le python, allons-y gaiement.
Ci-joint un complément écrit avant le commentaire précédent : la même chose pour les données des rues.

#10 - 03 juin 2022 15:43 - Pierre Ducroquet

- Fichier 0003-ban-import-update-streets-only-when-needed-65843.patch ajouté

#11 - 03 juin 2022 18:12 - Pierre Ducroquet

- Fichier 0004-ban-import-simplify-python-code-65843.patch ajouté

Du coup patch sur le 0002...

Si ça peut simplifier la relecture, je peux fournir un patch 0002 intégrant ce 0004...

#12 - 06 juin 2022 08:06 - Benjamin Dauvergne

- Statut changé de Solution proposée à Solution validée

Pierre Ducroquet a écrit :

> Du coup patch sur le 0002...
>
> Si ça peut simplifier la relecture, je peux fournir un patch 0002 intégrant ce 0004...

Non ça va si tu le fixup dans 0002 avant de commiter. Je vois un souci si un codeDepartement existant devient vide (mais je ne sais pas si le cas existe en vrai), dans ce cas il faudrait faire Model.objects.get(...., department__isnull=True) sinon la suppression de la valeur existante ne se fera pas, mais je valide quand même parce que le cas me semble tordu.

Concernant le select for update, à mon avis il n'y a que 2 scénarios : scénario chargement initiale, 100% des get() échouent, 100% des update_or_create exécutés, effectivement ce sera un poil plus lent, mais sinon 99.99% des get() réussissent et update_or_create() est appelé exceptionnellement (les nombres de mise à jour de ligne par semaine ou de lignes ajoutés ne doit pas dépasser un nombre à 2 chiffres). On est sur du détail.

Si on voulait vraiment aller plus loin ce serait plutôt en batchant les get() et en utilisant bulk_create()/bulk_update().

#13 - 07 juin 2022 08:45 - Pierre Ducroquet

- Statut changé de Solution validée à Résolu (à déployer)

Mergé.

```
commit 58f482cf932e63cc52a6290b094c8858633dbfec (HEAD -> main, origin/wip/65843-ban-weekly, wip/65843-ban-week
ly)
Author: Pierre Ducroquet <pducroquet@entrouvert.com>
Date:   Fri Jun 3 11:58:41 2022 +0200

    ban import: update streets only when needed (#65843)

commit c0154bdd7b8fdc54b68a68822121f3683d9619d4
Author: Pierre Ducroquet <pducroquet@entrouvert.com>
Date:   Wed Jun 1 16:51:11 2022 +0200

    ban import: update only when needed (#65843)

commit 9b4878482e2f7c4e4132ebcebadcea5ee4fc0c69
Author: Pierre Ducroquet <pducroquet@entrouvert.com>
Date:   Wed Jun 1 11:14:34 2022 +0200

    ban import: use Last-Modified to reduce load (#65843)
```

#14 - 10 juin 2022 08:41 - Lauréline Guérin

- Lié à Bug #66118: TypeError: can't subtract offset-naive and offset-aware datetimes ajouté

#15 - 10 juin 2022 11:20 - Pierre Ducroquet

- Fichier 0001-ban-import-fix-ban_id-field-65843.patch ajouté

- Statut changé de Résolu (à déployer) à Solution proposée

Suite à un problème remonté par nroche sur jabber, puis constaté sur des vieux tenants en prod qui ne sont pas maintenus depuis un certain temps...

La colonne ban_id a été introduite en not null avec un default ''. Par habitude, je ne m'attendais pas à ce choix, et du coup la création de la contrainte d'unicité allait échouer.
Donc le patch ci-joint corrige ça :
- introduction d'une migration 27 qui passe la colonne ban_id en nullable et remplace les '' par des null
- passage de la migration précédemment ajoutée en 28 (pour l'ajout de la contrainte)

#16 - 10 juin 2022 11:44 - Thomas Noël

Pierre Ducroquet a écrit :

> constaté sur des vieux tenants en prod qui ne sont pas maintenus depuis un certain temps...

Heu, c'est-à-dire ? C'est géré par du cron, y'a pas trop de raison... (Je conteste pas l'existence d'un soucis mais je veux le comprendre ;))

> - introduction d'une migration 27 qui passe la colonne ban_id en nullable et remplace les '' par des null
> - passage de la migration précédemment ajoutée en 28 (pour l'ajout de la contrainte)

Mais que va-t-il se passer sur les endroits où la migration 27 (actuelle) aurait déjà été jouée ?

#18 - 10 juin 2022 13:00 - Pierre Ducroquet

- Fichier 0001-ban-import-fix-ban_id-field-65843.patch ajouté

Patch corrigé, il passe sur jenkins.

#19 - 10 juin 2022 14:08 - Lauréline Guérin

mon avis: ne pas mélanger migration de schéma et migration de données dans une même migration django

#20 - 10 juin 2022 16:36 - Thomas Noël

Lauréline Guerin a écrit :

> mon avis: ne pas mélanger migration de schéma et migration de données dans une même migration django

Je confirme que c'est systématiquement un nid à problème, il faut décomposer en plusieurs migrations.

#21 - 10 juin 2022 18:09 - Pierre Ducroquet

Hum, et si je la passe en atomic=False, ça règle pas les soucis potentiels ? Si vous avez des détails sur les problèmes, je suis preneur d'ailleurs.

#22 - 10 juin 2022 19:43 - Benjamin Dauvergne

Pierre Ducroquet a écrit :

> Hum, et si je la passe en atomic=False, ça règle pas les soucis potentiels ? Si vous avez des détails sur les problèmes, je suis preneur d'ailleurs.

En atomic=False chaque opération de la migration s'exécute hors d'une transaction, et non c'est pire, si la migration échoue l'état n'est ni avant, ni après. Le souci vient surtout de migration de donnée qui vont générer de la validation de contrainte déférée (je ne doute pas que d'autres cas soit possible, mais j'ai surtout vu celui-là). Ayant eu le problème sur la suppression d'un champ, j'ai juste passé la validation des contraintes en immédiat[1].

#23 - 13 juin 2022 07:59 - Pierre Ducroquet

Benjamin Dauvergne a écrit :

> Pierre Ducroquet a écrit :
>
> > Hum, et si je la passe en atomic=False, ça règle pas les soucis potentiels ? Si vous avez des détails sur les problèmes, je suis preneur d'ailleurs.
>
> En atomic=False chaque opération de la migration s'exécute hors d'une transaction, et non c'est pire, si la migration échoue l'état n'est ni avant, ni après.

Quand chaque étape de la transaction est rejouable ad nauseam, en quoi serait-ce un problème ?

> Le souci vient surtout de migration de donnée qui vont générer de la validation de contrainte déférée (je ne doute pas que d'autres cas soit possible, mais j'ai surtout vu celui-là). Ayant eu le problème sur la suppression d'un champ, j'ai juste passé la validation des contraintes en immédiat.

Je vois très bien le SQL qui sera généré ici, et il n'y a aucune raison qu'un tel problème se pose.

#24 - 14 juin 2022 16:57 - Pierre Ducroquet

- Statut changé de Solution proposée à Résolu (à déployer)

Mergé pour régler le soucis sur l'env de test.

```
commit 7479f1d1430713de4304f1801590c80d5981b408 (HEAD -> main, origin/wip/65843-fix-ban_id, origin/main, origi
n/HEAD, wip/65843-fix-ban_id)
Author: Pierre Ducroquet <pducroquet@entrouvert.com>
Date:   Fri Jun 10 11:06:28 2022 +0200

    ban import: fix ban_id field (#65843)
```

#25 - 14 juin 2022 18:14 - Transition automatique

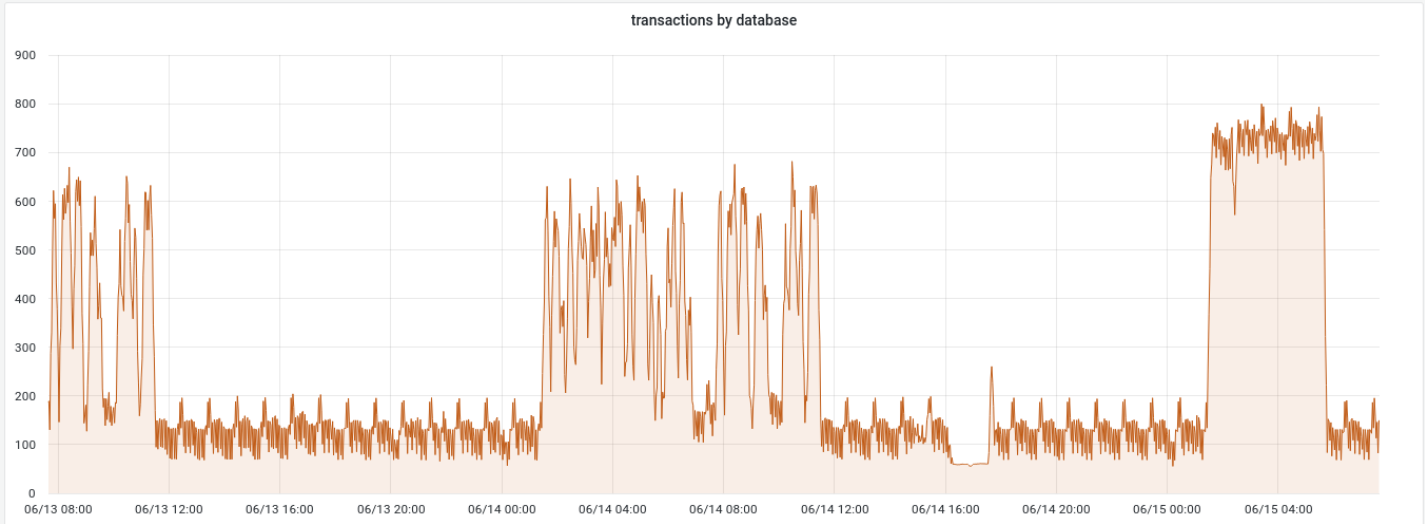
- Statut changé de Résolu (à déployer) à Solution déployée

#26 - 15 juin 2022 08:43 - Pierre Ducroquet

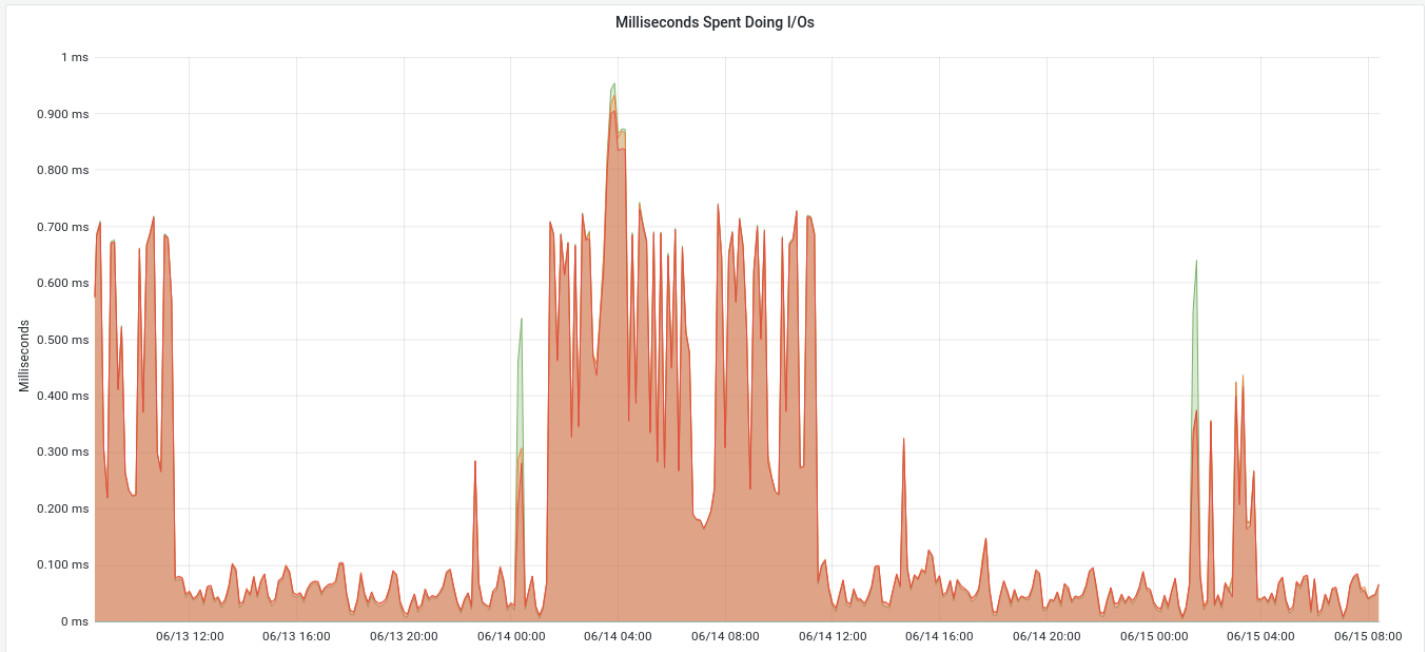
- Fichier transactions.png ajouté

- Fichier ios.png ajouté

Pour information : on est passés de 10H avec un trafic en transactions élevé, à 4H seulement. Je m'attendais à ne plus avoir ce pic, je creuserai.



Par contre, on a une chute nette sur l'utilisation CPU et sur les IOs, ce qui est un très bon signe.



#27 - 13 juillet 2022 21:03 - Lauréline Guérin

- Lié à Development #67375: BAN: reprendre l'optimisation de l'import, qu'on a dû revert ajouté

#28 - 21 août 2022 04:42 - Transition automatique

Automatic expiration

Fichiers

0001-ban-import-use-Last-Modified-to-reduce-load-65843.patch	1,51 ko	01 juin 2022	Pierre Ducroquet
0002-ban-import-skip-invalid-lines-65843.patch	1,14 ko	01 juin 2022	Pierre Ducroquet
0001-ban-import-use-Last-Modified-to-reduce-load-65843.patch	1,51 ko	03 juin 2022	Pierre Ducroquet
0002-ban-import-update-only-when-needed-65843.patch	6 ko	03 juin 2022	Pierre Ducroquet
0003-ban-import-update-streets-only-when-needed-65843.patch	5,28 ko	03 juin 2022	Pierre Ducroquet
0004-ban-import-simplify-python-code-65843.patch	5,49 ko	03 juin 2022	Pierre Ducroquet
0001-ban-import-fix-ban_id-field-65843.patch	2,59 ko	10 juin 2022	Pierre Ducroquet
0001-ban-import-fix-ban_id-field-65843.patch	3,48 ko	10 juin 2022	Pierre Ducroquet
transactions.png	117 ko	15 juin 2022	Pierre Ducroquet
ios.png	111 ko	15 juin 2022	Pierre Ducroquet