w.c.s. - Development #80613

Revoir le stockage et la gestion des "visited_object"

28 août 2023 16:11 - Pierre Ducroquet

Statut:	Fermé	**Début:**	28 août 2023
Priorité:	Haut	**Echéance:**	
Assigné à:	Pierre Ducroquet	**% réalisé:**	0%
Catégorie:		**Temps estimé:**	0:00 heure
Version cible:			
Patch proposed:	Non	**Planning:**	Non

Description

Actuellement, l'information sur les visited_object "squatte" le stockage des sessions.
Cela a conduit ce matin à avoir la requête suivante qui a consommé un temps CPU considérable sur le serveur (d'autant plus que nos logs tronquent à 100ms, j'ignore donc le temps réellement utilisé en permanence par cette requête)

```
select id, session_data from sessions where last_update_time >= ?::timestamp;
```

Ce qui prend du temps n'est pas le filtrage, mais le décodage côté PG des données pour les renvoyer à l'application, l'encodage réseau, le temps réseau, puis le temps de décodage côté applicatif. Par ailleurs, autre facteur à prendre en compte : le stockage dans le session_data, un bytea contenant un objet pickle, est particulièrement inefficace à la mise à jour puisque le code python doit tout reconstruire, et cela fait des enregistrements particulièrement volumineux pour peu de données utiles.

J'ai identifié le code suivant comme fort possiblement responsable de cet appel:

```
pierre@entrouvert-pierred:~/eo/wcs$ git grep select_recent
wcs/sessions.py:        for session in cls.select_recent(ignore_errors=True):
wcs/sessions.py:        for session in self.__class__.select_recent(ignore_errors=True):
wcs/sql.py:    def select_recent(cls, seconds=30 * 60, **kwargs):
```

Cela correspond aux méthodes Session::get_visited_objects et Session::unmark_visited_object.
La seconde prend en paramètre un formdata, elle sait donc quel objet doit être retiré. Un stockage efficace doit prendre en compte ce besoin.
La première renvoie une liste de tous les objets visités, en en excluant éventuellement un utilisateur.

Je suggère de remplacer tout ceci par une table "visited_object" (quelle imagination) avec la définition suivante:

```
CREATE TABLE visited_object (
  object_key character varying not null,
  visit_timestamp timestamp with time zone not null,
  visitor integer
);
```

Je n'ai pas l'impression que l'on puisse imposer une unicité au sein de cette table (et inutile de créer une fausse clé primaire sans intérêt).

Demandes liées:

Lié à w.c.s. - Development #81183: Revoir le stockage et la gestion des "visi...	**En cours**	**28 août 2023**

Révisions associées

Révision 9820b21d - 04 septembre 2023 16:52 - Pierre Ducroquet

session: limit visited objects in db query to sessions with visited objects (#80613)

Instead of fetching all recent sessions, fetch only the sessions with a
visited object, providing a great optimization even without any index.
The time is currently spent in de-toasting, formatting for the PG protocol,
and deparsing and unpickling on Python side, all for nothing for most
sessions.

Révision 0a7eb350 - 05 septembre 2023 16:25 - Pierre Ducroquet

session: limit visited objects in db query to sessions with visited objects (#80613)

Instead of fetching all recent sessions, fetch only the sessions with a
visited object, providing a great optimization even without any index.
The time is currently spent in de-toasting, formatting for the PG protocol,
and deparsing and unpickling on Python side, all for nothing for most
sessions.

Historique

#1 - 28 août 2023 17:00 - Pierre Ducroquet

Option pour l'unicité :

```
CREATE TABLE visited_object (
  object_key character varying not null,
  visit_timestamp timestamp with time zone not null,
  visitors integer[],
  primary key(object_key)
);
```

#3 - 29 août 2023 14:03 - Robot Gitea

- Statut changé de Nouveau à En cours

- Assigné à mis à Pierre Ducroquet

Pierre Ducroquet (pducroquet) a ouvert une pull request sur Gitea concernant cette demande :

- URL : https://git.entrouvert.org/entrouvert/wcs/pulls/624
- Titre : WIP: create a VisitedObjects class and table to reduce Sessions table usage (#80613)
- Modifications : https://git.entrouvert.org/entrouvert/wcs/pulls/624/files

#4 - 30 août 2023 12:35 - Pierre Ducroquet

- Assigné à changé de Pierre Ducroquet à Mikaël Ates

Limite à cette proposition, que je ne voyais pas en lisant simplement le code des sessions : les tests unitaires "s'attendent" à ce que les visites soient au moins partiellement liées à une session au lieu d'un utilisateur. Si c'est bien le cas, il faut alors rajouter un session_id sur la table, mais j'aimerais avoir confirmation de ce fait avant de procéder.

#5 - 30 août 2023 12:37 - Benjamin Dauvergne

- Assigné à changé de Mikaël Ates à Frédéric Péters

#6 - 30 août 2023 12:39 - Robot Gitea

- Assigné à changé de Frédéric Péters à Pierre Ducroquet

Pierre Ducroquet (pducroquet) a ouvert une pull request sur Gitea concernant cette demande :

- URL : https://git.entrouvert.org/entrouvert/wcs/pulls/625
- Titre : WIP: test a simple optimization for visits (#80613)
- Modifications : https://git.entrouvert.org/entrouvert/wcs/pulls/625/files

#7 - 30 août 2023 14:28 - Pierre Ducroquet

La seconde pull request correspond à une optimisation assez basique que j'ai remarqué en faisant la réécriture précédente. Il faudrait a minima intégrer ce changement pour avoir déjà pas mal de bénéfices sur les appels concernés.

#8 - 04 septembre 2023 14:32 - Robot Gitea

- Statut changé de En cours à Solution proposée

#9 - 04 septembre 2023 16:37 - Robot Gitea

- Statut changé de Solution proposée à Solution validée

Frédéric Péters (fpeters) a approuvé une pull request sur Gitea concernant cette demande :

- URL : https://git.entrouvert.org/entrouvert/wcs/pulls/625

#10 - 04 septembre 2023 16:52 - Robot Gitea

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

Pierre Ducroquet (pducroquet) a mergé une pull request sur Gitea concernant cette demande :

- URL : https://git.entrouvert.org/entrouvert/wcs/pulls/625
- Titre : sessions: simple optimization for DB queries (#80613)
- Modifications : https://git.entrouvert.org/entrouvert/wcs/pulls/625/files

#11 - 04 septembre 2023 18:14 - Transition automatique

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

#12 - 04 septembre 2023 21:20 - Pierre Ducroquet

- Statut changé de Solution déployée à En cours

Ce qui a été intégré correspond à une optimisation pour réduire l'importance de ce correctif, mais ne l'évite pas.

#13 - 13 septembre 2023 19:20 - Frédéric Péters

- Lié à Development #81183: Revoir le stockage et la gestion des "visited_object" ajouté

#14 - 13 septembre 2023 19:21 - Frédéric Péters

- Statut changé de En cours à Fermé

> Ce qui a été intégré correspond à une optimisation pour réduire l'importance de ce correctif, mais ne l'évite pas.

Je ferme ici pour garder une association claire entre commits et tickets, j'ai copié vers #81183 pour la suite.

#15 - 13 septembre 2023 19:21 - Robot Gitea

Pierre Ducroquet (pducroquet) a commencé à travailler sur une pull request sur Gitea concernant cette demande :

- URL : https://git.entrouvert.org/entrouvert/wcs/pulls/624
- Titre : WIP: create a VisitedObjects class and table to reduce Sessions table usage (#81183)
- Modifications : https://git.entrouvert.org/entrouvert/wcs/pulls/624/files