Module M3

Mémoire & journalisation

24.04

Table des matières

Sur ce document

Formation	Module M3
Titre	Mémoire & journalisation
Révision	24.04
PDF	https://dali.bo/m3_pdf
EPUB	https://dali.bo/m3_epub
HTML	https://dali.bo/m3_html
Slides	https://dali.bo/m3_slides
TP	https://dali.bo/m3_tp
TP (solutions)	https://dali.bo/m3_solutions

Vous trouverez en ligne les différentes versions complètes de ce document.

Chers lectrices & lecteurs,

Nos formations PostgreSQL sont issues de nombreuses années d'études, d'expérience de terrain et de passion pour les logiciels libres. Pour Dalibo, l'utilisation de PostgreSQL n'est pas une marque d'opportunisme commercial, mais l'expression d'un engagement de longue date. Le choix de l'Open Source est aussi le choix de l'implication dans la communauté du logiciel.

Au-delà du contenu technique en lui-même, notre intention est de transmettre les valeurs qui animent et unissent les développeurs de PostgreSQL depuis toujours : partage, ouverture, transparence, créativité, dynamisme… Le but premier de nos formations est de vous aider à mieux exploiter toute la puissance de PostgreSQL mais nous espérons également qu'elles vous inciteront à devenir un membre actif de la communauté en partageant à votre tour le savoir-faire que vous aurez acquis avec nous.

Nous mettons un point d'honneur à maintenir nos manuels à jour, avec des informations précises et des exemples détaillés. Toutefois malgré nos efforts et nos multiples relectures, il est probable que ce document contienne des oublis, des coquilles, des imprécisions ou des erreurs. Si vous constatez un souci, n'hésitez pas à le signaler via l'adresse formation@dalibo.com[1] !

À propos de DALIBO

DALIBO est le spécialiste français de PostgreSQL. Nous proposons du support, de la formation et du conseil depuis 2005.

Retrouvez toutes nos formations sur https://dalibo.com/formations

[1] mailto:formation@dalibo.com

Remerciements

Ce manuel de formation est une aventure collective qui se transmet au sein de notre société depuis des années. Nous remercions chaleureusement ici toutes les personnes qui ont contribué directement ou indirectement à cet ouvrage, notamment :

Jean-Paul Argudo, Alexandre Anriot, Carole Arnaud, Alexandre Baron, David Bidoc, Sharon Bonan, Franck Boudehen, Arnaud Bruniquel, Pierrick Chovelon, Damien Clochard, Christophe Courtois, Marc Cousin, Gilles Darold, Jehan-Guillaume de Rorthais, Ronan Dunklau, Vik Fearing, Stefan Fercot, Pierre Giraud, Nicolas Gollet, Dimitri Fontaine, Florent Jardin, Virginie Jourdan, Luc Lamarle, Denis Laxalde, Guillaume Lelarge, Alain Lesage, Benoit Lobréau, Jean-Louis Louër, Thibaut Madelaine, Adrien Nayrat, Alexandre Pereira, Flavie Perette, Robin Portigliatti, Thomas Reiss, Maël Rimbault, Julien Rouhaud, Stéphane Schildknecht, Julien Tachoires, Nicolas Thauvin, Be Hai Tran, Christophe Truffier, Cédric Villemain, Thibaud Walkowiak, Frédéric Yhuel.

Forme de ce manuel

Les versions PDF, EPUB ou HTML de ce document sont structurées autour des slides de nos formations. Le texte suivant chaque slide contient le cours et de nombreux détails qui ne peuvent être données à l'oral.

Licence Creative Commons CC-BY-NC-SA

Cela inclut les diapositives, les manuels eux-mêmes et les travaux pratiques. Cette formation peut également contenir quelques images et schémas dont la redistribution est soumise à des licences différentes qui sont alors précisées.

Marques déposées

PostgreSQL® Postgres® et le logo Slonik sont des marques déposées[3] par PostgreSQL Community Association of Canada.

Versions de PostgreSQL couvertes

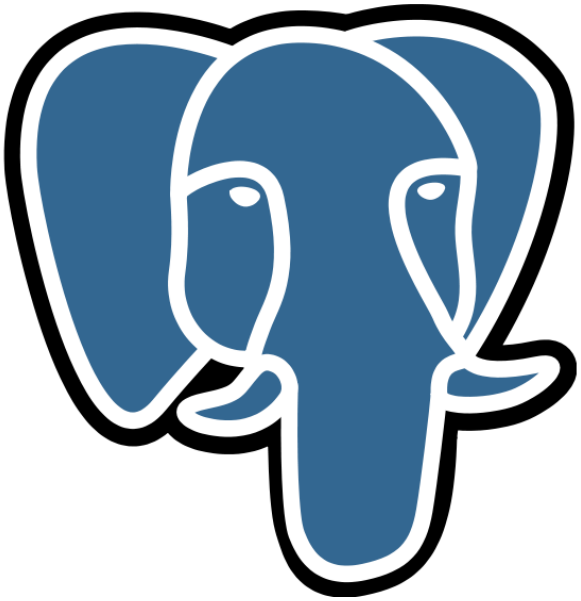
Ce document ne couvre que les versions supportées de PostgreSQL au moment de sa rédaction, soit les versions 12 à 16.

Sur les versions précédentes susceptibles d'être encore rencontrées en production, seuls quelques points très importants sont évoqués, en plus éventuellement de quelques éléments historiques.

Sauf précision contraire, le système d'exploitation utilisé est Linux.

[3] https://www.postgresql.org/about/policies/trademarks/

1/ Mémoire et journalisation dans PostgreSQL

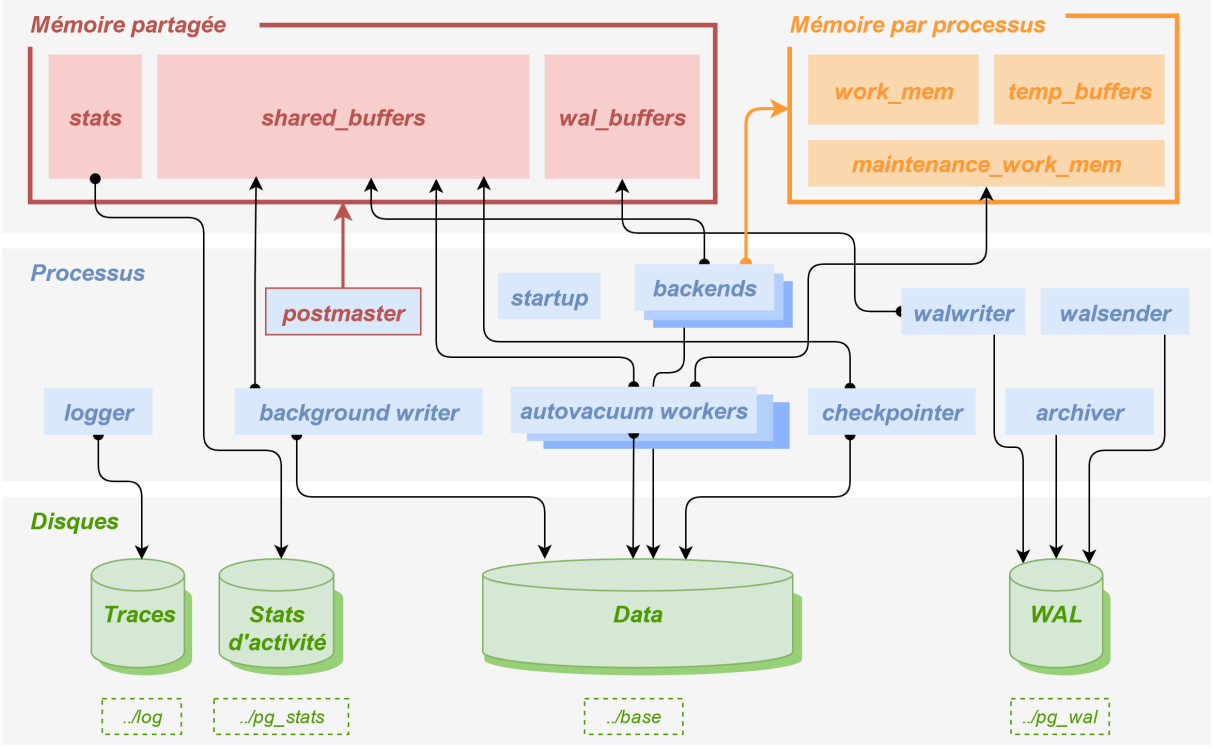


1.1 AU MENU

La mémoire & PostgreSQL :

- Mémoire partagée
- Mémoire des processus
- Les *shared buffers* & la gestion du cache
- La journalisation

1.2 RAPPEL DE L'ARCHITECTURE DE POSTGRESQL



1.3 MÉMOIRE PARTAGÉE

- L'implémentation dépend de l'OS
- Quelles zones ?
- Quelle taille ?

La zone de mémoire partagée statique est allouée au démarrage de l'instance. Le type de mémoire partagée est configuré avec le paramètre `shared_memory_type`. Sous Linux, il s'agit par défaut de `mmap`, sachant qu'une très petite partie utilise toujours `sysv` (System V). Il est en principe possible de basculer uniquement en `sysv` mais ceci n'est pas recommandé et nécessite généralement un paramétrage du noyau Linux. Sous Windows, le type est `windows`.

1.3.1 Zones de la mémoire partagée

- `shared_buffers`
  - cache disque des fichiers de données
- `wal_buffers`
  - cache disque des journaux de transactions
- `max_connections`
  - 100… ou plus ?
- `track_activity_query_size`
  - à monter
- verrous
  - `max_connections`, `max_locks_per_transaction`
- etc
- Modification → redémarrage

Les principales zones de mémoire partagées décrites ici sont fixes, et les tailles calculées en fonction de paramètres. Nous verrons en détail l'utilité de certaines de ces zones dans les chapitres suivants.

Shared buffers :

Les *shared buffers* sont le cache des fichiers de données présents sur le disque. Ils représentent de loin la volumétrie la plus importante.

Paramètre associé : `shared_buffers` (à adapter systématiquement)

Wal buffers :

Les *wal buffers* sont le cache des journaux de transaction.

Paramètre associé : `wal_buffers` (rarement modifié)

Données liées aux sessions :

Cet espace mémoire sert à gérer les sessions ouvertes, celles des utilisateurs, mais aussi celles ouvertes par de nombreux processus internes.

Principaux paramètres associés :

- `max_connections`, soit le nombre de connexions simultanées possibles (défaut : 100, souvent suffisant mais à adapter) ;
- `track_activity_query_size` (défaut : 1024 ; souvent monté à 10 000 s'il y a des requêtes de grande taille) ;

Verrous :

La gestion des verrous est décrite dans un autre module[1]. Lors des mises à jour ou suppressions dans les tables, les verrous sur les lignes sont généralement stockés dans les lignes mêmes ; mais de nombreux autres verrous sont stockés en mémoire partagée. Les paramètres associés pour le dimensionnement sont surtout :

- `max_connections` à nouveau ;
- `max_locks_per_transaction`, soit le nombre de verrous possible pour une transaction (défaut : 64, généralement suffisant) ;
- `max_pred_locks_per_relation`, nombre de verrous possible pour une table si le niveau d'isolation « sérialisation » est choisi ;
- mais encore les paramètres liés aux nombres de divers processus internes, ou de transactions préparées.

Modification :

Toute modification des paramètres régissant la mémoire partagée imposent un redémarrage de l'instance.

[1] https://dali.bo/m4_html#verrouillage-et-mvcc

1.3.2 Taille de la mémoire partagée

```
-- v15+
SHOW shared_memory_size ;
SHOW shared_memory_size_in_huge_pages ;
```

À partir de la version 15, le paramètre `shared_memory_size` permet de connaître la taille complète de mémoire partagée allouée (un peu supérieure à `shared_buffers` en pratique). Dans le cadre de l'utilisation des *Huge Pages*, il est possible de consulter le paramètre `shared_memory_size_in_huge_pages` pour connaître le nombre de pages mémoires nécessaires (mais on ne peut savoir ici si elles sont utilisées) :

```
postgres=# \dconfig shared*
              Liste des paramètres de configuration
            Paramètre             |             Valeur
----------------------------------+---------------------------------
 shared_buffers                   | 12GB
 shared_memory_size               | 12835MB
 shared_memory_size_in_huge_pages | 6418
 ...
```

Des zones de mémoire partagée non statiques peuvent exister : par exemple, à l'exécution d'une requête parallélisée, les processus impliqués utilisent de la mémoire partagée dynamique. Depuis PostgreSQL 14, une partie peut être pré-allouée avec le paramètre `min_dynamic_shared_memory` (0 par défaut).

1.4 MÉMOIRE PAR PROCESSUS

- `work_mem`
  - × `hash_mem_multiplier` (v 13)
- `maintenance_work_mem`
  - `autovacuum_work_mem`
- `temp_buffers`
- Pas de limite stricte à la consommation mémoire d'une session !
  - ni à la consommation totale
- Augmenter prudemment & superviser

Les processus de PostgreSQL ont accès à la mémoire partagée, définie principalement par `shared_buffers`, mais ils ont aussi leur mémoire propre. Cette mémoire n'est utilisable que par le processus l'ayant allouée.

Le paramètre le plus important est `work_mem`, qui définit la taille de la mémoire de travail d'un processus lors d'une requête, principalement lors d'opérations de tri : `ORDER BY`, certaines jointures, déduplication… Autre paramètre capital, `maintenance_work_mem` est la mémoire pour les opérations de maintenance lourdes : `VACUUM`, `CREATE INDEX`, ajouts de clé étrangère…

Cette mémoire est rendue immédiatement après la fin de l'ordre concerné.

Opérations de maintenance & maintenance_work_mem :

`maintenance_work_mem` peut être monté à 256 Mo à 1 Go sur les machines récentes, car il concerne des opérations lourdes rarement exécutées plusieurs fois simultanément. Monter au-delà est rare, mais peut avoir un intérêt dans les créations de très gros index.

Paramétrage de work_mem :

Pour `work_mem`, c'est beaucoup plus compliqué.

Si `work_mem` est trop bas, beaucoup d'opérations de tri, y compris nombre de jointures, ne s'effectueront pas en RAM. Par exemple, si une jointure par hachage impose d'utiliser 100 Mo en mémoire, mais que `work_mem` vaut 10 Mo, PostgreSQL écrira des dizaines de Mo sur disque à chaque appel de la jointure. Si, par contre, le paramètre `work_mem` vaut 120 Mo, aucune écriture n'aura lieu sur disque, ce qui accélérera forcément la requête.

Trop de fichiers temporaires peuvent ralentir les opérations, voire saturer le disque. Un `work_mem` trop bas peut aussi contraindre le planificateur à choisir des plans d'exécution moins optimaux.

Par contre, si `work_mem` est trop haut, et que trop de requêtes le consomment simultanément, le danger est de saturer la RAM. Il n'existe en effet pas de limite à la consommation des sessions de PostgreSQL, ni globalement ni par session !
Or le paramétrage de l'overcommit sous Linux est par défaut très permissif, le noyau ne bloquera rien. La première conséquence de la saturation de mémoire est l'assèchement du cache système (complémentaire de celui de PostgreSQL), et la dégradation des performances. Puis le système va se mettre à swapper, avec à la clé un ralentissement général et durable. Enfin le noyau, à court de mémoire, peut être amené à tuer un processus de PostgreSQL. Cela mène à l'arrêt de l'instance, ou plus fréquemment à son redémarrage brutal avec coupure de toutes les connexions et requêtes en cours.

Toutefois, si l'administrateur paramètre correctement l'overcommit[2], Linux refusera d'allouer la RAM et la requête tombera en erreur, mais le cache système sera préservé, et PostgreSQL ne tombera pas.

Suivant la complexité des requêtes, il est possible qu'un processus utilise plusieurs fois `work_mem` (par exemple si une requête fait une jointure et un tri, ou qu'un nœud est parallélisé). À l'inverse, beaucoup de requêtes ne nécessitent aucune mémoire de travail.

La valeur de `work_mem` dépend donc beaucoup de la mémoire disponible, des requêtes et du nombre de connexions actives.

Si le nombre de requêtes simultanées est important, `work_mem` devra être faible. Avec peu de requêtes simultanées, `work_mem` pourra être augmenté sans risque.

Il n'y a pas de formule de calcul miracle. Une première estimation courante, bien que très conservatrice, peut être :

> `work_mem` = mémoire / `max_connections`

On obtient alors, sur un serveur dédié avec 16 Go de RAM et 200 connexions autorisées :

```
work_mem = 80MB
```

Mais `max_connections` est fréquemment surdimensionné, et beaucoup de sessions sont inactives. `work_mem` est alors sous-dimensionné.

Plus finement, Christophe Pettus propose en première intention[3] :

> `work_mem` = 4 × mémoire libre / `max_connections`

Soit, pour une machine dédiée avec 16 Go de RAM, donc 4 Go de *shared buffers*, et 200 connections :

```
work_mem = 240MB
```

[2] https://dali.bo/j1_html#configuration-du-oom
[3] https://thebuild.com/blog/2023/03/13/everything-you-know-about-setting-work_mem-is-wrong/

Dans l'idéal, si l'on a le temps pour une étude, on montera `work_mem` jusqu'à voir disparaître l'essentiel des fichiers temporaires dans les traces, tout en restant loin de saturer la RAM lors des pics de charge.

En pratique, le défaut de 4 Mo est très conservateur, souvent insuffisant. Généralement, la valeur varie entre 10 et 100 Mo. Au-delà de 100 Mo, il y a souvent un problème ailleurs : des tris sur de trop gros volumes de données, une mémoire insuffisante, un manque d'index (utilisés pour les tris), etc. Des valeurs vraiment grandes ne sont valables que sur des systèmes d'infocentre.

Augmenter globalement la valeur du `work_mem` peut parfois mener à une consommation excessive de mémoire. Il est possible de ne la modifier que le temps d'une session pour les besoins d'une requête ou d'un traitement particulier :

```
SET work_mem TO '30MB' ;
```

hash_mem_multiplier :

À partir de PostgreSQL 13, un paramètre multiplicateur peut s'appliquer à certaines opérations particulières (le hachage, lors de jointures ou agrégations). Nommé `hash_mem_multiplier`, il vaut 1 par défaut en versions 13 et 14, et 2 à partir de la 15. `hash_mem_multiplier` permet de donner plus de RAM à ces opérations sans augmenter globalement `work_mem`.

Tables temporaires

Les tables temporaires (et leurs index) sont locales à chaque session, et disparaîtront avec elle. Elles sont tout de même écrites sur disque dans le répertoire de la base.

Le cache dédié à ces tables pour minimiser les accès est séparé des *shared buffers*, parce qu'il est propre à la session. Sa taille dépend du paramètre `temp_buffers`. La valeur par défaut (8 Mo) peut être insuffisante dans certains cas pour éviter les accès aux fichiers de la table. Elle doit être augmentée avant la création de la table temporaire.

1.5 SHARED BUFFERS

- *Shared buffers* ou blocs de mémoire partagée
  - partage les blocs entre les processus
  - cache en lecture ET écriture
  - double emploi partiel avec le cache du système (voir `effective_cache_size`)
  - importants pour les performances !
- Dimensionnement en première intention & avant tests :
  - `shared_buffers` = ¼ RAM
  - si > 8 Go : *Huge Pages*, `max_wal_size`…

PostgreSQL dispose de son propre mécanisme de cache. Toute donnée lue l'est de ce cache. Si la donnée n'est pas dans le cache, le processus devant effectuer cette lecture l'y recopie avant d'y accéder dans le cache.

L'unité de travail du cache est le bloc (de 8 ko par défaut) de données. C'est-à-dire qu'un processus charge toujours un bloc dans son entier quand il veut lire un enregistrement. Chaque bloc du cache correspond donc exactement à un bloc d'un fichier d'un objet. Cette information est d'ailleurs, bien sûr, stockée en en-tête du bloc de cache.

Tous les processus accèdent à ce cache unique. C'est la zone la plus importante, par la taille, de la mémoire partagée. Toute modification de données est tracée dans le journal de transaction, **puis** modifiée dans ce cache. Elle n'est donc pas écrite sur le disque par le processus effectuant la modification, sauf en dernière extrémité (voir *Synchronisation en arrière plan*).

Tout accès à un bloc nécessite la prise de verrous. Un *pin lock*, qui est un simple compteur, indique qu'un processus se sert du buffer, et qu'il n'est donc pas réutilisable. C'est un verrou potentiellement de longue durée. Il existe de nombreux autres verrous, de plus courte durée, pour obtenir le droit de modifier le contenu d'un buffer, d'un enregistrement dans un buffer, le droit de recycler un buffer… mais tous ces verrous n'apparaissent pas dans la table `pg_locks`, car ils sont soit de très courte durée, soit partagés (comme le *spin lock*). Il est donc très rare qu'ils soient sources de contention, mais le diagnostic d'une contention à ce niveau est difficile.

Les lectures et écritures de PostgreSQL passent toutefois toujours par le cache du système. Les deux caches risquent donc de stocker les mêmes informations. Les algorithmes d'éviction sont différents entre le système et PostgreSQL, PostgreSQL disposant de davantage d'informations sur l'utilisation des données, et le type d'accès qui y est fait. La redondance est donc habituellement limitée.

Dimensionner correctement ce cache est important pour de nombreuses raisons.

Un cache trop petit :

- ralentit l'accès aux données, car des données importantes risquent de ne plus s'y trouver ;
- force l'écriture de données sur le disque, ralentissant les sessions qui auraient pu effectuer uniquement des opérations en mémoire ;
- limite le regroupement d'écritures, dans le cas où un bloc viendrait à être modifié plusieurs fois.

Un cache trop grand :

- limite l'efficacité du cache système en augmentant la redondance de données entre les deux caches ;
- peut ralentir PostgreSQL, car la gestion des `shared_buffers` a un coût de traitement ;
- réduit la mémoire disponible pour d'autres opérations (tris en mémoire notamment).

Ce paramétrage du cache est malgré tout moins critique que sur de nombreux autres SGBD : le cache système limite la plupart du temps l'impact d'un mauvais paramétrage de `shared_buffers`, et il est donc préférable de sous-dimensionner `shared_buffers` que de le sur-dimensionner.

Pour dimensionner `shared_buffers` sur un serveur dédié à PostgreSQL, la documentation officielle[4] donne 25 % de la mémoire vive totale comme un bon point de départ et déconseille de dépasser 40 %, car le cache du système d'exploitation est aussi utilisé.

Sur une machine dédiée de 32 Go de RAM, cela donne donc :

```
shared_buffers = 8GB
```

Le défaut de 128 Mo n'est donc pas adapté à un serveur sur une machine récente.

Suivant les cas, une valeur inférieure ou supérieure à 25 % sera encore meilleure pour les performances, mais il faudra tester avec votre charge (en lecture, en écriture, et avec le bon nombre de clients).

Modifier `shared_buffers` impose de redémarrer l'instance.

Attention : une valeur élevée de `shared_buffers` (au-delà de 8 Go) nécessite de paramétrer finement le système d'exploitation (*Huge Pages* notamment) et d'autres paramètres comme `max_wal_size`, et de s'assurer qu'il restera de la mémoire pour le reste des opérations (tri…).

Un cache supplémentaire est disponible pour PostgreSQL : celui du système d'exploitation. Il est donc intéressant de préciser à PostgreSQL la taille approximative du cache, ou du moins de la part du cache qu'occupera PostgreSQL. Le paramètre `effective_cache_size` n'a pas besoin d'être très précis, mais il permet une meilleure estimation des coûts par le moteur. Il est paramétré habituellement aux alentours des ⅔ de la taille de la mémoire vive du système d'exploitation, pour un serveur dédié.

Par exemple pour une machine avec 32 Go de RAM, on peut paramétrer en première intention dans `postgresql.conf` :

```
shared_buffers = '8GB'
effective_cache_size = '21GB'
```

Cela sera à ajuster en fonction du comportement observé de l'application.

1.5.1 Notions essentielles de gestion du cache

- Buffer pin
- Buffer dirty/clean
- Compteur d'utilisation
- Clocksweep

Les principales notions à connaître pour comprendre le mécanisme de gestion du cache de PostgreSQL sont :

Buffer pin

Un processus voulant accéder à un bloc du cache (*buffer*) doit d'abord épingler (*to pin*) ce bloc pour forcer son maintien en cache. Pour ce faire, il incrémente le compteur *buffer pin*, puis le décrémente quand il a fini. Un buffer dont le pin est différent de 0 est donc utilisé et ne peut être recyclé.

Buffer dirty/clean

Un buffer est *dirty* (sale) si son contenu a été modifié en mémoire mais pas encore sur disque. Le fichier de données n'est plus à jour (bien que la modification ait été pérennisée sur disque dans les journaux de transactions).

Un buffer est *clean* (propre) s'il n'a pas été modifié. Le bloc peut être supprimé du cache sans nécessiter le coût d'une écriture sur disque.

Compteur d'utilisation

Cette technique vise à garder dans le cache les blocs les plus utilisés.

À chaque fois qu'un processus a fini de se servir d'un buffer (quand il enlève son pin), ce compteur est incrémenté (à hauteur de 5 dans l'implémentation actuelle). Il est décrémenté par le *clocksweep* évoqué plus bas.

Seul un buffer dont le compteur est à zéro peut voir son contenu remplacé par un nouveau bloc.

Clocksweep (ou algorithme de balayage)

Un processus ayant besoin de charger un bloc de données dans le cache doit trouver un buffer disponible. Soit il y a encore des buffers vides (cela arrive principalement au démarrage d'une instance), soit il faut libérer un buffer.

L'algorithme *clocksweep* parcourt la liste des buffers de façon cyclique à la recherche d'un buffer *unpinned* dont le compteur d'utilisation est à zéro. Tout buffer visité voit son compteur décrémenté de 1. Le système effectue autant de passes que nécessaire sur tous les blocs jusqu'à trouver un buffer à 0. Ce *clocksweep* est effectué par chaque processus, au moment où ce dernier a besoin d'un nouveau buffer.

1.5.2 Ring buffer

But : ne pas purger le cache à cause :

- des grandes tables
- de certaines opérations
  - *Seq Scan* (lecture)
  - `VACUUM` (écritures)
  - `COPY`, `CREATE TABLE AS SELECT…`
  - …

Une table peut être plus grosse que les *shared buffers*. Sa lecture intégrale (lors d'un parcours complet ou d'une opération de maintenance) ne doit pas mener à l'éviction de tous les blocs du cache.

PostgreSQL utilise donc plutôt un *ring buffer* quand la taille de la relation dépasse 1/4 de `shared_buffers`. Un *ring buffer* est une zone de mémoire gérée à l'écart des autres blocs du cache. Pour un parcours complet d'une table, cette zone est de 256 ko (taille choisie pour tenir dans un cache L2). Si un bloc y est modifié (`UPDATE`…), il est traité hors du *ring buffer* comme un bloc sale normal. Pour un `VACUUM`, la même technique est utilisée, mais les écritures se font dans le *ring buffer*. Pour les écritures en masse (notamment `COPY` ou `CREATE TABLE AS SELECT`), une technique similaire utilise un *ring buffer* de 16 Mo.

Le site *The Internals of PostgreSQL*[5] et un README[6] dans le code de PostgreSQL entrent plus en détail sur tous ces sujets tout en restant lisibles.

1.5.3 Contenu du cache

2 extensions en « contrib » :

- `pg_buffercache`
- `pg_prewarm`

[5] https://www.interdb.jp/pg/pgsql08.html
[6] https://github.com/postgres/postgres/blob/master/src/backend/storage/buffer/README

Deux extensions sont livrées dans les *contribs* de PostgreSQL qui impactent le cache.

`pg_buffercache` permet de consulter le contenu du cache (à utiliser de manière très ponctuelle). La requête suivante indique les objets non système de la base en cours, présents dans le cache et s'ils sont *dirty* ou pas :

```
pgbench=# CREATE EXTENSION pg_buffercache ;

pgbench=# SELECT
    relname,
    isdirty,
    count(bufferid) AS blocs,
    pg_size_pretty(count(bufferid) * current_setting ('block_size')::int) AS taille
FROM pg_buffercache b
INNER JOIN pg_class c ON c.relfilenode = b.relfilenode
WHERE relname NOT LIKE 'pg\_%'
GROUP BY
        relname,
        isdirty
ORDER BY 1, 2 ;
```

```
        relname         | isdirty | blocs | taille
------------------------+---------+-------+---------
 pgbench_accounts       | f       |  8398 | 66 MB
 pgbench_accounts       | t       |  4622 | 36 MB
 pgbench_accounts_pkey  | f       |  2744 | 21 MB
 pgbench_branches       | f       |    14 | 112 kB
 pgbench_branches       | t       |     2 | 16 kB
 pgbench_branches_pkey  | f       |     2 | 16 kB
 pgbench_history        | f       |   267 | 2136 kB
 pgbench_history        | t       |   102 | 816 kB
 pgbench_tellers        | f       |    13 | 104 kB
 pgbench_tellers_pkey   | f       |     2 | 16 kB
```

L'extension `pg_prewarm` permet de précharger un objet dans le cache de PostgreSQL (si le cache est assez gros, bien sûr) :

```
=# CREATE EXTENSION pg_prewarm ;
=# SELECT pg_prewarm ('nom_table_ou_index', 'buffer') ;
```

Il permet même de recharger dès le démarrage le contenu du cache lors d'un arrêt (voir la documentation[7]).

Ces deux outils sont décrits dans le module de formation X2[8].

[7] https://docs.postgresql.fr/current/pgprewarm.html
[8] https://dali.bo/x2_html

1.5.4 Synchronisation en arrière plan

- Le *Background Writer* synchronise les buffers
  - de façon anticipée
  - une portion des pages à synchroniser
  - paramètres : `bgwriter_delay`, `bgwriter_lru_maxpages`, `bgwriter_lru_multiplier` et `bgwriter_flush_after`
- Le *checkpointer* synchronise les buffers
  - lors des checkpoints
  - synchronise toutes les dirty pages
- Écriture directe par les *backends*
  - en dernière extrémité

Afin de limiter les attentes des sessions interactives, PostgreSQL dispose de deux processus, le *Background Writer* et le *Checkpointer*, tous deux essayant d'effectuer de façon asynchrone les écritures des buffers sur le disque. Le but est que les temps de traitement ressentis par les utilisateurs soient les plus courts possibles, et que les écritures soient lissées sur de plus grandes plages de temps (pour ne pas saturer les disques).

Le *Background Writer* anticipe les besoins de buffers des sessions. À intervalle régulier, il se réveille et synchronise un nombre de buffers proportionnel à l'activité sur l'intervalle précédent, dans ceux qui seront examinés par les sessions pour les prochaines allocations. Quatre paramètres régissent son comportement :

- `bgwriter_delay` (défaut : 200 ms) : la fréquence à laquelle se réveille le *Background Writer* ;
- `bgwriter_lru_maxpages` (défaut : 100) : le nombre maximum de pages pouvant être écrites sur chaque tour d'activité. Ce paramètre permet d'éviter que le *Background Writer* ne veuille synchroniser trop de pages si l'activité des sessions est trop intense : dans ce cas, autant les laisser effectuer elles-mêmes les synchronisations, étant donné que la charge est forte ;
- `bgwriter_lru_multiplier` (defaut : 2) : le coefficient multiplicateur utilisé pour calculer le nombre de buffers à libérer par rapport aux demandes d'allocation sur la période précédente ;
- `bgwriter_flush_after` (défaut : 512 ko sous Linux, 0 ou désactivé ailleurs) : à partir de quelle quantité de données écrites une synchronisation sur disque est demandée.

Pour les paramètres `bgwriter_lru_maxpages` et `bgwriter_lru_multiplier`, *lru* signifie *Least Recently Used* que l'on pourrait traduire par « moins récemment utilisé ». Ainsi, pour ce mécanisme, le *Background Writer* synchronisera les pages du cache qui ont été utilisées le moins récemment.

Le *checkpointer* est responsable d'un autre mécanisme : il synchronise tous les blocs modifiés lors

des checkpoints. Son rôle est d'effectuer cette synchronisation, en évitant de saturer les disques en lissant la charge (voir plus loin).

Lors d'écritures intenses, il est possible que ces deux mécanismes soient débordés. Les processus *backend* peuvent alors écrire eux-mêmes dans les fichiers de données (après les journaux de transaction, bien sûr). Cette situation est évidemment à éviter, ce qui implique généralement de rendre le *bgwriter* plus agressif.

1.6 JOURNALISATION

- Garantir la durabilité des données
- Base encore cohérente après :
  - arrêt brutal des processus
  - crash machine
  - …
- Écriture des modifications dans un journal **avant** les fichiers de données
- WAL : *Write Ahead Logging*

La journalisation, sous PostgreSQL, permet de garantir l'intégrité des fichiers, et la durabilité des opérations :

- L'intégrité : quoi qu'il arrive, exceptée la perte des disques de stockage bien sûr, la base reste cohérente. Un arrêt d'urgence ne corrompra pas la base.
- Toute donnée validée (`COMMIT`) est écrite. Un arrêt d'urgence ne va pas la faire disparaître.

Pour cela, le mécanisme est relativement simple : toute modification affectant un fichier sera d'abord écrite dans le journal. Les modifications affectant les vrais fichiers de données ne sont écrites qu'en mémoire, dans les *shared buffers*. Elles seront écrites de façon asynchrone, soit par un processus recherchant un buffer libre, soit par le *Background Writer*, soit par le *Checkpointer*.

Les écritures dans le journal, bien que synchrones, sont relativement performantes, car elles sont séquentielles (moins de déplacement de têtes pour les disques).

1.6.1 Journaux de transaction (rappels)

Essentiellement :

- `pg_wal/` : journaux de transactions
  - sous-répertoire `archive_status`
  - nom : *timeline*, journal, segment
  - ex : `00000002 00000142 000000FF`
- `pg_xact/` : état des transactions
- **Ces fichiers sont vitaux !**

Rappelons que les journaux de transaction sont des fichiers de 16 Mo par défaut, stockés dans `PGDATA/pg_wal` (`pg_xlog` avant la version 10), dont les noms comportent le numéro de *timeline*, un numéro de journal de 4 Go et un numéro de segment, en hexadécimal.

```
$ ls -l
total 2359320
...
-rw------- 1 postgres postgres 33554432 Mar 26 16:28 00000002000001420000007C
-rw------- 1 postgres postgres 33554432 Mar 26 16:28 00000002000001420000007D
...
-rw------- 1 postgres postgres 33554432 Mar 26 16:25 000000020000014300000023
-rw------- 1 postgres postgres 33554432 Mar 26 16:25 000000020000014300000024
drwx------ 2 postgres postgres    16384 Mar 26 16:28 archive_status
```

Le sous-répertoire `archive_status` est lié à l'archivage.

D'autres plus petits répertoires comme `pg_xact`, qui contient les statuts des transactions passées, ou `pg_commit_ts`, `pg_multixact`, `pg_serial`, `pg_snapshots`, `pg_subtrans` ou encore `pg_twophase` sont également impliqués.

Tous ces répertoires sont critiques, gérés par PostgreSQL, et ne doivent pas être modifiés !

1.6.2 Checkpoint

- « Point de reprise »
- À partir d'où rejouer les journaux ?
- Données écrites au moins au niveau du checkpoint
  - il peut durer
- Processus `checkpointer`

PostgreSQL trace les modifications de données dans les journaux WAL. Ceux-ci sont générés au fur et à mesure des écritures.

Si le système ou l'instance sont arrêtés brutalement, il faut que PostgreSQL puisse appliquer le contenu des journaux non traités sur les fichiers de données. Il a donc besoin de savoir à partir d'où rejouer ces données. Ce point est ce qu'on appelle un *checkpoint*, ou « point de reprise ».

Les principes sont les suivants :

Toute entrée dans les journaux est idempotente, c'est-à-dire qu'elle peut être appliquée plusieurs fois, sans que le résultat final ne soit changé. C'est nécessaire, au cas où la récupération serait interrompue, ou si un fichier sur lequel la reprise est effectuée était plus récent que l'entrée qu'on souhaite appliquer.

Tout fichier de journal antérieur au dernier point de reprise valide **peut être supprimé** ou recyclé, car il n'est plus nécessaire à la récupération.

PostgreSQL a besoin des fichiers de données qui contiennent toutes les données jusqu'au point de reprise. Ils peuvent être plus récents et contenir des informations supplémentaires, ce n'est pas un problème.

Un checkpoint n'est pas un « instantané » cohérent de l'ensemble des fichiers. C'est simplement l'endroit à partir duquel les journaux doivent être rejoués. Il faut donc pouvoir garantir que tous les blocs modifiés dans le cache *au démarrage du checkpoint* auront été synchronisés sur le disque quand le checkpoint sera terminé, et marqué comme dernier checkpoint valide. Un checkpoint peut donc durer plusieurs minutes, sans que cela ne bloque l'activité.

C'est le processus `checkpointer` qui est responsable de l'écriture des buffers devant être synchronisés durant un checkpoint.

1.6.3 Déclenchement & comportement des checkpoints - 1

- Déclenchement périodique (idéal)
  - `checkpoint_timeout`
- ou : Quantité de journaux
  - `max_wal_size` (pas un plafond !)
- ou : `CHECKPOINT`
- À la fin :
  - sync
  - recyclage des journaux
- Espacer les checkpoints peut réduire leur volumétrie

Plusieurs paramètres influencent le comportement des checkpoints.

Dans l'idéal les checkpoints sont périodiques. Le temps maximum entre deux checkpoints est fixé par `checkpoint_timeout` (par défaut 300 secondes). C'est parfois un peu court pour les instances actives.

Le checkpoint intervient aussi quand il y a beaucoup d'écritures et que le volume des journaux dépasse le seuil défini par le paramètre `max_wal_size` (1 Go par défaut). Un checkpoint est alors déclenché.

L'ordre `CHECKPOINT` déclenche aussi un *checkpoint* sans attendre. En fait, il sert surtout à des utilitaires.

Une fois le checkpoint terminé, les journaux sont à priori inutiles. Ils peuvent être effacés pour redescendre en-dessous de la quantité définie par `max_wal_size`. Ils sont généralement « recyclés », c'est-à-dire renommés, et prêt à être réécris.

Cependant, les journaux peuvent encore être retenus dans `pg_wal/` si l'archivage a été activé et que certains n'ont pas été sauvegardés, ou si l'on garde des journaux pour des serveurs secondaires.

À cause de cela, le volume de l'ensemble des fichiers WAL peut largement dépasser la taille fixée par `max_wal_size`. Ce n'est **pas** une valeur plafond !

Il existe un paramètre `min_wal_size` (défaut : 80 Mo) qui fixe la quantité minimale de journaux à tout moment, même sans activité en écriture. Ils seront donc vides et prêts à être remplis en cas d'écriture imprévue. Bien sûr, s'il y a des grosses écritures, PostgreSQL créera au besoin des journaux supplémentaires, jusque `max_wal_size`, voire au-delà. Mais il lui faudra les créer et les remplir intégralement de zéros avant utilisation.

Après un gros pic d'activité suivi d'un checkpoint et d'une période calme, la quantité de journaux va très progressivement redescendre de `max_wal_size` à `min_wal_size`.

Le dimensionnement de ces paramètres est très dépendant du contexte, de l'activité habituelle, et de la régularité des écritures. Le but est d'éviter des gros pics d'écriture, et donc d'avoir des checkpoints essentiellement périodiques, même si des opérations ponctuelles peuvent y échapper (gros chargements, grosse maintenance…).

Des checkpoints espacés ont aussi pour effet de réduire la quantité totale de journaux écrits. En effet, par défaut, un bloc modifié est intégralement écrit dans les journaux à sa première modification après un checkpoint, mais par la suite seules les modifications de ce bloc sont journalisées. Espacer les checkpoints peut économiser beaucoup de place disque quand les journaux sont archivés, et du réseau s'ils sont répliqués. Par contre, un écart plus grand entre checkpoints peut allonger la restauration après un arrêt brutal, car il y aura plus de journaux à rejouer.

En pratique, une petite instance se contentera du paramétrage de base ; une plus grosse montera `max_wal_size` à plusieurs Go.

Si l'on monte `max_wal_size`, par cohérence, il faudra penser à augmenter aussi `checkpoint_timeout`, et vice-versa.

Pour `min_wal_size`, rien n'interdit de prendre une valeur élevée pour mieux absorber les montées d'activité brusques.

Enfin, le checkpoint comprend un *sync* sur disque final. Toujours pour éviter des à-coups d'écriture, PostgreSQL demande au système d'exploitation de forcer un vidage du cache quand `checkpoint_flush_after` a déjà été écrit (par défaut 256 ko). Avant PostgreSQL 9.6, ceci se paramétrait au niveau de Linux en abaissant les valeurs des *sysctl* `vm.dirty_*`. Il y a un intérêt à continuer de le faire, car PostgreSQL n'est pas seul à écrire de gros fichiers (exports `pg_dump`, copie de fichiers…).

1.6.4 Déclenchement & comportement des checkpoints - 2

- Dilution des écritures
  - `checkpoint_completion_target` × durée moy. entre 2 checkpoints
- Surveillance :
  - `checkpoint_warning`
  - `log_checkpoints`
  - Gardez de la place ! sinon crash…

Quand le checkpoint démarre, il vise à lisser au maximum le débit en écriture. La durée d'écriture des données se calcule à partir d'une fraction de la durée d'exécution des précédents checkpoints, fraction fixée par le paramètre `checkpoint_completion_target`. Sa valeur par défaut est celle préconisée par la documentation pour un lissage maximum, soit 0,9 (depuis la version 14, et auparavant le défaut de 0,5 était fréquemment corrigé). Par défaut, PostgreSQL prévoit donc une durée maximale de 300 × 0,9 = 270 secondes pour opérer son checkpoint, mais cette valeur pourra évoluer ensuite suivant la durée réelle des checkpoints précédents.

Il est possible de suivre le déroulé des checkpoints dans les traces si `log_checkpoints` est à `on`. De plus, si deux checkpoints sont rapprochés d'un intervalle de temps inférieur à `checkpoint_warning` (défaut : 30 secondes), un message d'avertissement sera tracé. Une répétition fréquente indique que `max_wal_size` est bien trop petit.

Enfin, répétons que `max_wal_size` n'est pas une limite en dur de la taille de `pg_wal/`.

La partition de `pg_wal/` doit être taillée généreusement. Sa saturation entraîne l'arrêt immédiat de l'instance !

1.6.5 WAL buffers : journalisation en mémoire

- Mutualiser les écritures entre transactions
- Un processus d'arrière plan : `walwriter`
- Paramètres notables :
  - `wal_buffers`
  - `wal_writer_flush_after`
- Fiabilité :
  - `fsync` = `on`
  - `full_page_writes` = `on`
  - sinon **corruption** !

La journalisation s'effectue par écriture dans les journaux de transactions. Toutefois, afin de ne pas effectuer des écritures synchrones pour chaque opération dans les fichiers de journaux, les écritures sont préparées dans des tampons (*buffers*) en mémoire. Les processus écrivent donc leur travail de journalisation dans des *buffers*, ou *WAL buffers*. Ceux-ci sont vidés quand une session demande validation de son travail (`COMMIT`), qu'il n'y a plus de *buffer* disponible, ou que le *walwriter* se réveille (`wal_writer_delay`).

Écrire un ou plusieurs blocs séquentiels de façon synchrone sur un disque a le même coût à peu de chose près. Ce mécanisme permet donc de réduire fortement les demandes d'écriture synchrone sur le journal, et augmente donc les performances.

Afin d'éviter qu'un processus n'ait tous les buffers à écrire à l'appel de `COMMIT` , et que cette opération ne dure trop longtemps, un processus d'arrière-plan appelé *walwriter* écrit à intervalle régulier tous les buffers à synchroniser.

Ce mécanisme est géré par ces paramètres, rarement modifiés :

- `wal_buffers` : taille des *WAL buffers*, soit par défaut 1/32e de `shared_buffers` avec un maximum de 16 Mo (la taille d'un segment), des valeurs supérieures (par exemple 128 Mo[9]) pouvant être intéressantes pour les très grosses charges ;
- `wal_writer_delay` (défaut : 200 ms) : intervalle auquel le *walwriter* se réveille pour écrire les buffers non synchronisés ;
- `wal_writer_flush_after` (défaut : 1 Mo) : au-delà de cette valeur, les journaux écrits sont synchronisés sur disque pour éviter l'accumulation dans le cache de l'OS.

Pour la fiabilité, on ne touchera pas à ceux-ci :

[9] https://thebuild.com/blog/2023/02/08/xtreme-postgresql/

- `wal_sync_method` : appel système à utiliser pour demander l'écriture synchrone (sauf très rare exception, PostgreSQL détecte tout seul le bon appel système à utiliser) ;
- `full_page_writes` : doit-on réécrire une image complète d'une page suite à sa première modification après un checkpoint ? Sauf cas très particulier, comme un système de fichiers *Copy On Write* comme ZFS ou btrfs, ce paramètre doit rester à `on` pour éviter des corruptions de données (et il est alors conseillé d'espacer les checkpoints pour réduire la volumétrie des journaux) ;
- `fsync` : doit-on réellement effectuer les écritures synchrones ? Le défaut est `on` et **il est très fortement conseillé de le laisser ainsi en production**. Avec `off`, les performances en écritures sont certes très accélérées, mais en cas d'arrêt d'urgence de l'instance, les données seront totalement corrompues ! Ce peut être intéressant pendant le chargement initial d'une nouvelle instance par exemple, sans oublier de revenir à `on` après ce chargement initial. (D'autres paramètres et techniques existent pour accélérer les écritures et sans corrompre votre instance, si vous êtes prêt à perdre certaines données non critiques : `synchronous_commit` à `off`, les tables *unlogged*…)

1.6.6 Compression des journaux

- `wal_compression`
  - compression des enregistrements
  - moins de journaux
  - un peu de CPU

`wal_compression` compresse les blocs complets enregistrés dans les journaux de transactions, réduisant le volume des WAL, la charge en écriture sur les disques, la volumétrie des journaux archivés des sauvegardes PITR.

Comme il y a moins de journaux, leur rejeu est aussi plus rapide, ce qui accélère la réplication et la reprise après un crash. Le prix est une augmentation de la consommation en CPU.

Les détails et un exemple figurent dans ce billet du blog Dalibo[10].

[10] https://blog.dalibo.com/2024/01/05/cambouis.html

1.6.7 Limiter le coût de la journalisation

- `synchronous_commit`
  - perte potentielle de données validées
- `commit_delay` / `commit_siblings`
- Par session

Le coût d'un `fsync` est parfois rédhibitoire. Avec certains sacrifices, il est parfois possible d'améliorer les performances sur ce point.

Le paramètre `synchronous_commit` (défaut : `on`) indique si la validation de la transaction en cours doit déclencher une écriture synchrone dans le journal. Le défaut permet de garantir la pérennité des données dès la fin du `COMMIT`.

Mais ce paramètre peut être modifié dans chaque session par une commande `SET`, et passé à `off` **s'il est possible d'accepter une petite perte de données** pourtant committées. La perte peut monter à `3 × wal_writer_delay` (600 ms) ou `wal_writer_flush_after` (1 Mo) octets écrits. On accélère ainsi notablement les flux des petites transactions. Les transactions où le paramètre reste à `on` continuent de profiter de la sécurité maximale. La base restera, quoi qu'il arrive, cohérente. (Ce paramètre permet aussi de régler le niveau des transactions synchrones avec des secondaires.)

Il existe aussi `commit_delay` (défaut : `0`) et `commit_siblings` (défaut : `5`) comme mécanisme de regroupement de transactions[11]. S'il y au moins `commit_siblings` transactions en cours, PostgreSQL attendra jusqu'à `commit_delay` (en microsecondes) avant de valider une transaction pour permettre à d'autres transactions de s'y rattacher. Ce mécanisme, désactivé par défaut, accroît la latence de certaines transactions afin que plusieurs soient écrites ensembles, et n'apporte un gain de performance global qu'avec de nombreuses petites transactions en parallèle, et des disques classiques un peu lents. (En cas d'arrêt brutal, il n'y a pas à proprement parler de perte de données puisque les transactions délibérément retardées n'ont pas été signalées comme validées.)

[11] https://docs.postgresql.fr/current/wal-configuration.html

1.7 AU-DELÀ DE LA JOURNALISATION

- Sauvegarde PITR
- Réplication physique
  - par *log shipping*
  - par *streaming*

Le système de journalisation de PostgreSQL étant très fiable, des fonctionnalités très intéressantes ont été bâties dessus.

1.7.1 L'archivage des journaux

- Repartir à partir :
  - d'une vieille sauvegarde
  - les journaux archivés
- Sauvegarde à chaud
- Sauvegarde en continu
- Paramètres
  - `wal_level`, `archive_mode`
  - `archive_command` ou `archive_library`

Les journaux permettent de rejouer, suite à un arrêt brutal de la base, toutes les modifications depuis le dernier checkpoint. Les journaux devenus obsolète depuis le dernier *checkpoint* (l'avant-dernier avant la version 11) sont à terme recyclés ou supprimés, car ils ne sont plus nécessaires à la réparation de la base.

Le but de l'archivage est de stocker ces journaux, afin de pouvoir rejouer leur contenu, non plus depuis le dernier checkpoint, mais **depuis une sauvegarde**. Le mécanisme d'archivage permet de repartir d'une sauvegarde binaire de la base (c'est-à-dire des fichiers, pas un `pg_dump`), et de réappliquer le contenu des journaux archivés.

Il suffit de rejouer tous les journaux depuis le checkpoint précédent la sauvegarde jusqu'à la fin de la sauvegarde, ou même à un point précis dans le temps. L'application de ces journaux permet de rendre

à nouveau cohérents les fichiers de données, même si ils ont été sauvegardés en cours de modification.

Ce mécanisme permet aussi de fournir une sauvegarde continue de la base, alors même que celle-ci travaille.

Tout ceci est vu dans le module *Point In Time Recovery*[12].

Même si l'archivage n'est pas en place, il faut connaître les principaux paramètres impliqués :

wal_level :

Il vaut `replica` par défaut depuis la version 10. Les journaux contiennent les informations nécessaires pour une sauvegarde PITR ou une réplication vers une instance secondaire.

Si l'on descend à `minimal` (défaut jusqu'en version 9.6 incluse), les journaux ne contiennent plus que ce qui est nécessaire à une reprise après arrêt brutal sur le serveur en cours. Ce peut être intéressant pour réduire, parfois énormément, le volume des journaux générés, si l'on a bien une sauvegarde non PITR par ailleurs.

Le niveau `logical` est destiné à la réplication logique[13].

(Avant la version 9.6 existaient les niveaux intermédiaires `archive` et `hot_standby`, respectivement pour l'archivage et pour un serveur secondaire en lecture seule. Ils sont toujours acceptés, et assimilés à `replica`.)

archive_mode & **archive_command**/**archive_library** :

Il faut qu'`archive_mode` soit à `on` pour activer l'archivage. Les journaux sont alors copiés grâce à une commande shell à fournir dans `archive_command` ou grâce à une bibliothèque partagée indiquée dans `archive_library` (version 15 ou postérieure). En général on y indiquera ce qu'exige un outil de sauvegarde dédié (par exemple pgBackRest ou barman) dans sa documentation.

1.7.2 Réplication

- *Log shipping* : fichier par fichier
- *Streaming* : entrée par entrée (en flux continu)
- Serveurs secondaires très proches de la production, en lecture

La restauration d'une sauvegarde peut se faire en continu sur un autre serveur, qui peut même être actif (bien que forcément en lecture seule). Les journaux peuvent être :

- envoyés régulièrement vers le secondaire, qui les rejouera : c'est le principe de la réplication par *log shipping* ;

[12] https://dali.bo/i2_html
[13] https://dali.bo/w5_html

- envoyés par fragments vers cet autre serveur : c'est la réplication par *streaming*.

Ces thèmes ne seront pas développés ici. Signalons juste que la réplication par *log shipping* implique un archivage actif sur le primaire, et l'utilisation de `restore_command` (et d'autres pour affiner) sur le secondaire. Le *streaming* permet de se passer d'archivage, même si coupler *streaming* et sauvegarde PITR est une bonne idée. Sur un PostgreSQL récent, le primaire a par défaut le nécessaire activé pour se voir doté d'un secondaire : `wal_level` est à `replica` ; `max_wal_senders` permet d'ouvrir des processus dédiés à la réplication ; et l'on peut garder des journaux en paramétrant `wal_keep_size` (ou `wal_keep_segments` avant la version 13) pour limiter les risques de décrochage du secondaire.

Une configuration supplémentaire doit se faire sur le serveur secondaire, indiquant comment récupérer les fichiers de l'archive, et comment se connecter au primaire pour récupérer des journaux. Elle a lieu dans les fichiers `recovery.conf` (jusqu'à la version 11 comprise), ou (à partir de la version 12) `postgresql.conf` dans les sections évoquées plus haut, ou `postgresql.auto.conf`.

1.8 CONCLUSION

Mémoire et journalisation :

- complexe
- critique
- mais fiable
- et le socle de nombreuses fonctionnalités évoluées

1.8.1 Questions

N'hésitez pas, c'est le moment !

1.9 QUIZ

https://dali.bo/m3_quiz

1.10 TRAVAUX PRATIQUES

1.10.1 Mémoire partagée

But : constater l'effet du cache sur les accès.

Se connecter à la base de données **b0** et créer une table `t2` avec une colonne `id` de type `integer`.

Insérer 500 lignes dans la table `t2` avec `generate_series`.

Pour réinitialiser les statistiques de `t2` :

- utiliser la fonction `pg_stat_reset_single_table_counters`
- l'OID en paramètre est dans la table des relations `pg_class`, ou peut être trouvé avec `'t2'::regclass`

Afin de vider le cache, redémarrer l'instance PostgreSQL.

Se connecter à la base de données **b0** et lire les données de la table `t2`.

Récupérer les statistiques IO pour la table `t2` dans la vue système `pg_statio_user_tables`. Qu'observe-t-on ?

Lire de nouveau les données de la table `t2` et consulter ses statistiques. Qu'observe-t-on ?

Lire de nouveau les données de la table `t2` et consulter ses statistiques. Qu'observe-t-on ?

1.10.2 Mémoire de tri

But : constater l'influence de la mémoire de tri

Ouvrir un premier terminal et laisser défiler le fichier de traces.

Dans un second terminal, activer la trace des fichiers temporaires ainsi que l'affichage du niveau LOG pour le client (il est possible de le faire sur la session uniquement).

Insérer un million de lignes dans la table `t2` avec `generate_series`.

Activer le chronométrage dans la session (`\timing on`). Lire les données de la table `t2` en triant par la colonne `id` Qu'observe-t-on ?

Configurer la valeur du paramètre `work_mem` à `100MB` (il est possible de le faire sur la session uniquement).

Lire de nouveau les données de la table `t2` en triant par la colonne `id`. Qu'observe-t-on ?

1.10.3 Cache disque de PostgreSQL

But : constater l'effet du cache de PostgreSQL

Se connecter à la base de données `b1`. Installer l'extension `pg_buffercache`.

Créer une table `t2` avec une colonne `id` de type `integer`.

Insérer un million de lignes dans la table `t2` avec `generate_series`.

Pour vider le cache de PostgreSQL, redémarrer l'instance.

Pour vider le cache du système d'exploitation, sous **root** :

```
# sync && echo 3 > /proc/sys/vm/drop_caches
```

Se connecter à la base de données **b1**. En utilisant l'extension `pg_buffercache`, que contient le cache de PostgreSQL ? (Compter les blocs pour chaque table ; au besoin s'inspirer de la requête du cours.)

Activer l'affichage de la durée des requêtes. Lire les données de la table `t2`, en notant la durée d'exécution de la requête. Que contient le cache de PostgreSQL ?

Lire de nouveau les données de la table `t2`. Que contient le cache de PostgreSQL ?

Configurer la valeur du paramètre `shared_buffers` à un quart de la RAM.

Redémarrer l'instance PostgreSQL.

Se connecter à la base de données **b1** et extraire de nouveau toutes les données de la table `t2`. Que contient le cache de PostgreSQL ?

Modifier le contenu de la table `t2`, par exemple avec :

```
UPDATE t2 SET id = 0 WHERE id < 1000 ;
```

Que contient le cache de PostgreSQL ?

Exécuter un checkpoint. Que contient le cache de PostgreSQL ?

1.10.4 Journaux

But : Observer la génération de journaux

Insérer 10 millions de lignes dans la table `t2` avec `generate_series`. Que se passe-t-il au niveau du répertoire `pg_wal` ?

Exécuter un checkpoint. Que se passe-t-il au niveau du répertoire `pg_wal` ?

1.11 TRAVAUX PRATIQUES (SOLUTIONS)

1.11.1 Mémoire partagée

> Se connecter à la base de données **b0** et créer une table `t2` avec une colonne `id` de type `integer`.

```
$ psql b0
```

```
b0=# CREATE TABLE t2 (id integer);
CREATE TABLE
```

> Insérer 500 lignes dans la table `t2` avec `generate_series`.

```
b0=# INSERT INTO t2 SELECT generate_series(1, 500);
INSERT 0 500
```

> Pour réinitialiser les statistiques de `t2` :
>
> - utiliser la fonction `pg_stat_reset_single_table_counters`
> - l'OID en paramètre est dans la table des relations `pg_class`, ou peut être trouvé avec `'t2'::regclass` Cette fonction attend un OID comme paramètre :

```
b0=# \df pg_stat_reset_single_table_counters
```

```
List of functions
-[ RECORD 1 ]-------+--------------------
Schema              | pg_catalog
Name                | pg_relation_filepath
Result data type    | text
Argument data types | regclass
Type                | func
```

L'OID est une colonne présente dans la table `pg_class` :

```
b0=# SELECT relname, pg_stat_reset_single_table_counters(oid)
     FROM pg_class WHERE relname = 't2';
```

```
 relname | pg_stat_reset_single_table_counters
---------+-------------------------------------
 t2      |
```

Il y a cependant un raccourci à connaître :

```
SELECT pg_stat_reset_single_table_counters('t2'::regclass) ;
```

> Afin de vider le cache, redémarrer l'instance PostgreSQL.

```
# systemctl restart postgresql-15
```

> Se connecter à la base de données **b0** et lire les données de la table `t2`.

```
b0=# SELECT * FROM t2;
[...]
```

> Récupérer les statistiques IO pour la table `t2` dans la vue système `pg_statio_user_tables`. Qu'observe-t-on ?

```
b0=# \x
Expanded display is on.

b0=# SELECT * FROM pg_statio_user_tables WHERE relname = 't2' ;

-[ RECORD 1 ]---+-------
relid           | 24576
schemaname      | public
relname         | t2
heap_blks_read  | 3
heap_blks_hit   | 0
idx_blks_read   |
idx_blks_hit    |
toast_blks_read |
toast_blks_hit  |
tidx_blks_read  |
tidx_blks_hit   |
```

3 blocs ont été lus en dehors du cache de PostgreSQL (colonne `heap_blks_read`).

> Lire de nouveau les données de la table `t2` et consulter ses statistiques. Qu'observe-t-on ?

```
b0=# SELECT * FROM t2;
[...]
b0=# SELECT * FROM pg_statio_user_tables WHERE relname = 't2';

-[ RECORD 1 ]---+-------
relid           | 24576
schemaname      | public
relname         | t2
heap_blks_read  | 3
heap_blks_hit   | 3
…
```

Les 3 blocs sont maintenant lus à partir du cache de PostgreSQL (colonne `heap_blks_hit`).

> Lire de nouveau les données de la table `t2` et consulter ses statistiques. Qu'observe-t-on ?

```
b0=# SELECT * FROM t2;
[...]
b0=# SELECT * FROM pg_statio_user_tables WHERE relname = 't2';

-[ RECORD 1 ]---+-------
relid           | 24576
schemaname      | public
```

```
relname         | t2
heap_blks_read  | 3
heap_blks_hit   | 6
…
```

Quelle que soit la session, le cache étant partagé, tout le monde profite des données en cache.

1.11.2 Mémoire de tri

> Ouvrir un premier terminal et laisser défiler le fichier de traces.

Le nom du fichier dépend de l'installation et du moment. Pour suivre tout ce qui se passe dans le fichier de traces, utiliser `tail -f` :

```
$ tail -f /var/lib/pgsql/15/data/log/postgresql-Tue.log
```

> Dans un second terminal, activer la trace des fichiers temporaires ainsi que l'affichage du niveau LOG pour le client (il est possible de le faire sur la session uniquement).

Dans la session :

```
postgres=# SET client_min_messages TO log;
SET
postgres=# SET log_temp_files TO 0;
SET
```

Les paramètres `log_temp_files` et `client_min_messages` peuvent aussi être mis en place une fois pour toutes dans `postgresql.conf` (recharger la configuration). En fait, c'est généralement conseillé.

> Insérer un million de lignes dans la table `t2` avec `generate_series`.

```
b0=# INSERT INTO t2 SELECT generate_series(1, 1000000);
```

```
INSERT 0 1000000
```

> Activer le chronométrage dans la session (`\timing on`). Lire les données de la table `t2` en triant par la colonne `id` Qu'observe-t-on ?

```
b0=# \timing on
b0=# SELECT * FROM t2 ORDER BY id;
```

```
LOG:  temporary file: path "base/pgsql_tmp/pgsql_tmp1197.0", size 14032896
   id
---------
       1
       1
       2
       2
       3
[...]
Time: 436.308 ms
```

Le message `LOG` apparaît aussi dans la trace, et en général il se trouvera là.

PostgreSQL a dû créer un fichier temporaire pour stocker le résultat temporaire du tri. Ce fichier s'appelle `base/pgsql_tmp/pgsql_tmp1197.0`. Il est spécifique à la session et sera détruit dès qu'il ne sera plus utile. Il fait 14 Mo.

Écrire un fichier de tri sur disque prend évidemment un certain temps, c'est généralement à éviter si le tri peut se faire en mémoire.

> Configurer la valeur du paramètre `work_mem` à `100MB` (il est possible de le faire sur la session uniquement).

```
b0=# SET work_mem TO '100MB';
SET
```

> Lire de nouveau les données de la table `t2` en triant par la colonne `id`. Qu'observe-t-on ?

```
b0=# SELECT * FROM t2 ORDER BY id;

   id
---------
       1
       1
       2
       2
[...]
Time: 240.565 ms
```

Il n'y a plus de fichier temporaire généré. La durée d'exécution est bien moindre.

1.11.3 Cache disque de PostgreSQL

> Se connecter à la base de données `b1`. Installer l'extension `pg_buffercache`.

```
b1=# CREATE EXTENSION pg_buffercache;
CREATE EXTENSION
```

> Créer une table `t2` avec une colonne `id` de type `integer`.

```
b1=# CREATE TABLE t2 (id integer);
CREATE TABLE
```

> Insérer un million de lignes dans la table `t2` avec `generate_series`.

```
b1=# INSERT INTO t2 SELECT generate_series(1, 1000000);
INSERT 0 1000000
```

> Pour vider le cache de PostgreSQL, redémarrer l'instance.

```
 # systemctl restart postgresql-15
```

> Pour vider le cache du système d'exploitation, sous **root** :
>
> `# sync && echo 3 > /proc/sys/vm/drop_caches`

> Se connecter à la base de données **b1**. En utilisant l'extension `pg_buffercache`, que contient le cache de PostgreSQL ? (Compter les blocs pour chaque table ; au besoin s'inspirer de la requête du cours.)

```
b1=# SELECT relfilenode, count(*)
  FROM pg_buffercache
  GROUP BY 1
  ORDER BY 2 DESC
  LIMIT 10;

 relfilenode | count
-------------+-------
             | 16181
        1249 |    57
        1259 |    26
        2659 |    15
[...]
```

Les valeurs exactes peuvent varier. La colonne `relfilenode` correspond à l'identifiant système de la table. La deuxième colonne indique le nombre de blocs. Il y a ici 16 181 blocs non utilisés pour l'instant dans le cache (126 Mo), ce qui est logique vu que PostgreSQL vient de redémarrer. Il y a quelques blocs utilisés par des tables systèmes, mais aucune table utilisateur (on les repère par leur OID supérieur à 16384).

> Activer l'affichage de la durée des requêtes. Lire les données de la table `t2`, en notant la durée d'exécution de la requête. Que contient le cache de PostgreSQL ?

```
b1=# \timing on
Timing is on.

b1=# SELECT * FROM t2;

   id
---------
       1
       2
       3
       4
       5
[...]
Time: 277.800 ms

b1=# SELECT relfilenode, count(*) FROM pg_buffercache
     GROUP BY 1 ORDER BY 2 DESC LIMIT 10 ;
```

```
 relfilenode | count
-------------+-------
             | 16220
       16410 |    32
        1249 |    29
        1259 |     9
        2659 |     8
[...]
Time: 30.694 ms
```

32 blocs ont été alloués pour la lecture de la table `t2` (*filenode* 16410). Cela représente 256 ko alors que la table fait 35 Mo :

```
b1=# SELECT pg_size_pretty(pg_table_size('t2'));

 pg_size_pretty
----------------
 35 MB
(1 row)

Time: 1.913 ms
```

Un simple `SELECT *` ne suffit donc pas à maintenir la table dans le cache. Par contre, ce deuxième accès était déjà beaucoup rapide, ce qui suggère que le système d'exploitation, lui, a probablement gardé les fichiers de la table dans son propre cache.

> Lire de nouveau les données de la table `t2`. Que contient le cache de PostgreSQL ?

```
b1=# SELECT * FROM t2;

   id
---------
[...]
Time: 184.529 ms
```

```
b1=#  SELECT relfilenode, count(*) FROM pg_buffercache
      GROUP BY 1 ORDER BY 2 DESC LIMIT 10 ;

 relfilenode | count
-------------+-------
             | 16039
        1249 |    85
       16410 |    64
        1259 |    39
        2659 |    22
[...]
```

Il y en en a un peu plus dans le cache (en fait, 2 fois 32 ko). Plus vous exécuterez la requête, et plus le nombre de blocs présents en cache augmentera. Sur le long terme, les 4425 blocs de la table `t2` peuvent se retrouver dans le cache.

> Configurer la valeur du paramètre `shared_buffers` à un quart de la RAM.

Pour cela, il faut ouvrir le fichier de configuration `postgresql.conf` et modifier la valeur du paramètre `shared_buffers` à un quart de la mémoire. Par exemple :

```
shared_buffers = 2GB
```

> Redémarrer l'instance PostgreSQL.

```
# systemctl restart postgresql-15
```

> Se connecter à la base de données **b1** et extraire de nouveau toutes les données de la table `t2`.
> Que contient le cache de PostgreSQL ?

```
b1=# \timing on
b1=# SELECT * FROM t2;

   id
---------
       1
[...]
Time: 340.444 ms
```

```
b1=# SELECT relfilenode, count(*) FROM pg_buffercache
     GROUP BY 1 ORDER BY 2 DESC LIMIT 10 ;

 relfilenode | count
-------------+--------
             | 257581
       16410 |   4425
        1249 |     29
[...]
```

PostgreSQL se retrouve avec toute la table directement dans son cache, et ce dès la première exécution.

PostgreSQL est optimisé principalement pour du multi-utilisateurs. Dans ce cadre, il faut pouvoir exécuter plusieurs requêtes en même temps et donc chaque requête ne peut pas monopoliser tout le cache. De ce fait, chaque requête ne peut prendre qu'une partie réduite du cache. Mais plus le cache est gros, plus la partie octroyée est grosse.

> Modifier le contenu de la table `t2`, par exemple avec :
>
> ```
> UPDATE t2 SET id = 0 WHERE id < 1000 ;
> ```
>
> Que contient le cache de PostgreSQL ?

```
b1=# UPDATE t2 SET id=0 WHERE id < 1000;
UPDATE 999
```

```
b1=# SELECT
    relname,
    isdirty,
    count(bufferid) AS blocs,
    pg_size_pretty(count(bufferid) * current_setting ('block_size')::int) AS taille
FROM pg_buffercache b
INNER JOIN pg_class c ON c.relfilenode = b.relfilenode
WHERE relname NOT LIKE 'pg\_%'
GROUP BY
```

```
        relname,
        isdirty
ORDER BY 1, 2 ;
```

```
 relname | isdirty | blocs | taille
---------+---------+-------+--------
 t2      | f       |  4419 | 35 MB
 t2      | t       |    15 | 120 kB
```

15 blocs ont été modifiés (`isdirty` est à `true`), le reste n'a pas bougé.

> Exécuter un checkpoint. Que contient le cache de PostgreSQL ?

```
b1=# CHECKPOINT;
CHECKPOINT
```

```
b1=# SELECT
    relname,
    isdirty,
    count(bufferid) AS blocs,
    pg_size_pretty(count(bufferid) * current_setting ('block_size')::int) AS taille
FROM pg_buffercache b
INNER JOIN pg_class c ON c.relfilenode = b.relfilenode
WHERE relname NOT LIKE 'pg\_%'
GROUP BY
        relname,
        isdirty
ORDER BY 1, 2 ;
```

```
 relname | isdirty | blocs | taille
---------+---------+-------+--------
 t2      | f       |  4434 | 35 MB
```

Les blocs *dirty* ont tous été écrits sur le disque et sont devenus « propres ».

1.11.4 Journaux

> Insérer 10 millions de lignes dans la table `t2` avec `generate_series`. Que se passe-t-il au niveau du répertoire `pg_wal` ?

```
b1=# INSERT INTO t2 SELECT generate_series(1, 10000000);
INSERT 0 10000000
```

```
$ ls -al $PGDATA/pg_wal
total 131076
$ ls -al $PGDATA/pg_wal
total 638984
drwx------  3 postgres postgres     4096 Apr 16 17:55 .
drwx------ 20 postgres postgres     4096 Apr 16 17:48 ..
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000033
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000034
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000035
…
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000054
```

```
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000055
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000056
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000057
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000058
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000059
drwx------  2 postgres postgres        6 Apr 16 15:01 archive_status
```

Des journaux de transactions sont écrits lors des écritures dans la base. Leur nombre varie avec l'activité récente.

> Exécuter un checkpoint. Que se passe-t-il au niveau du répertoire `pg_wal` ?

```
b1=# CHECKPOINT;
CHECKPOINT

$ ls -al $PGDATA/pg_wal
total 131076
total 638984
drwx------  3 postgres postgres     4096 Apr 16 17:56 .
drwx------ 20 postgres postgres     4096 Apr 16 17:48 ..
-rw-------  1 postgres postgres 16777216 Apr 16 17:56 000000010000000000000059
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000005A
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000005B
…
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 000000010000000000000079
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000007A
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000007B
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000007C
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000007D
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000007E
-rw-------  1 postgres postgres 16777216 Apr 16 17:55 00000001000000000000007F
drwx------  2 postgres postgres        6 Apr 16 15:01 archive_status
```

Le nombre de journaux n'a pas forcément décru, mais le dernier journal d'avant le checkpoint est à présent le plus ancien (selon l'ordre des noms des journaux).

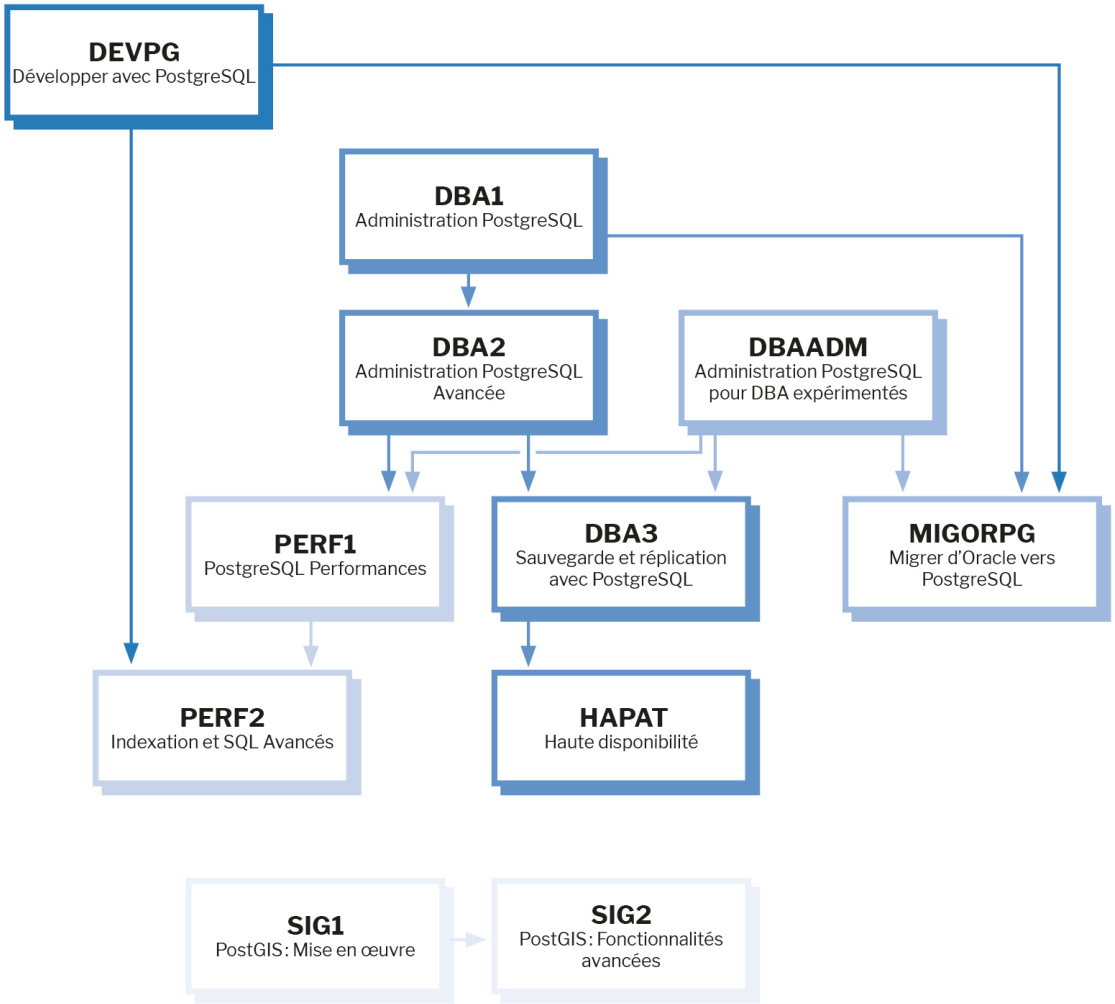
Ici, il n'y a ni PITR ni archivage. Les anciens journaux sont donc totalement inutiles et sont donc recyclés : renommés, il sont prêts à être remplis à nouveau. Noter que leur date de création n'a pas été mise à jour !

Les formations Dalibo

Retrouvez nos formations et le calendrier sur https://dali.bo/formation

Pour toute information ou question, n'hésitez pas à nous écrire sur contact@dalibo.com.

Cursus des formations



Retrouvez nos formations dans leur dernière version :

- DBA1 : Administration PostgreSQL
  https://dali.bo/dba1
- DBA2 : Administration PostgreSQL avancé
  https://dali.bo/dba2
- DBA3 : Sauvegarde et réplication avec PostgreSQL
  https://dali.bo/dba3
- DEVPG : Développer avec PostgreSQL
  https://dali.bo/devpg
- PERF1 : PostgreSQL Performances
  https://dali.bo/perf1
- PERF2 : Indexation et SQL avancés
  https://dali.bo/perf2
- MIGORPG : Migrer d'Oracle à PostgreSQL
  https://dali.bo/migorpg
- HAPAT : Haute disponibilité avec PostgreSQL
  https://dali.bo/hapat

Les livres blancs

- Migrer d'Oracle à PostgreSQL
  https://dali.bo/dlb01
- Industrialiser PostgreSQL
  https://dali.bo/dlb02
- Bonnes pratiques de modélisation avec PostgreSQL
  https://dali.bo/dlb04
- Bonnes pratiques de développement avec PostgreSQL
  https://dali.bo/dlb05

Téléchargement gratuit

Les versions électroniques de nos publications sont disponibles gratuitement sous licence open source ou sous licence Creative Commons.