# Transactions et accès concurrents

Bases de données 2

Thibaut MADELAINE

janvier 2022

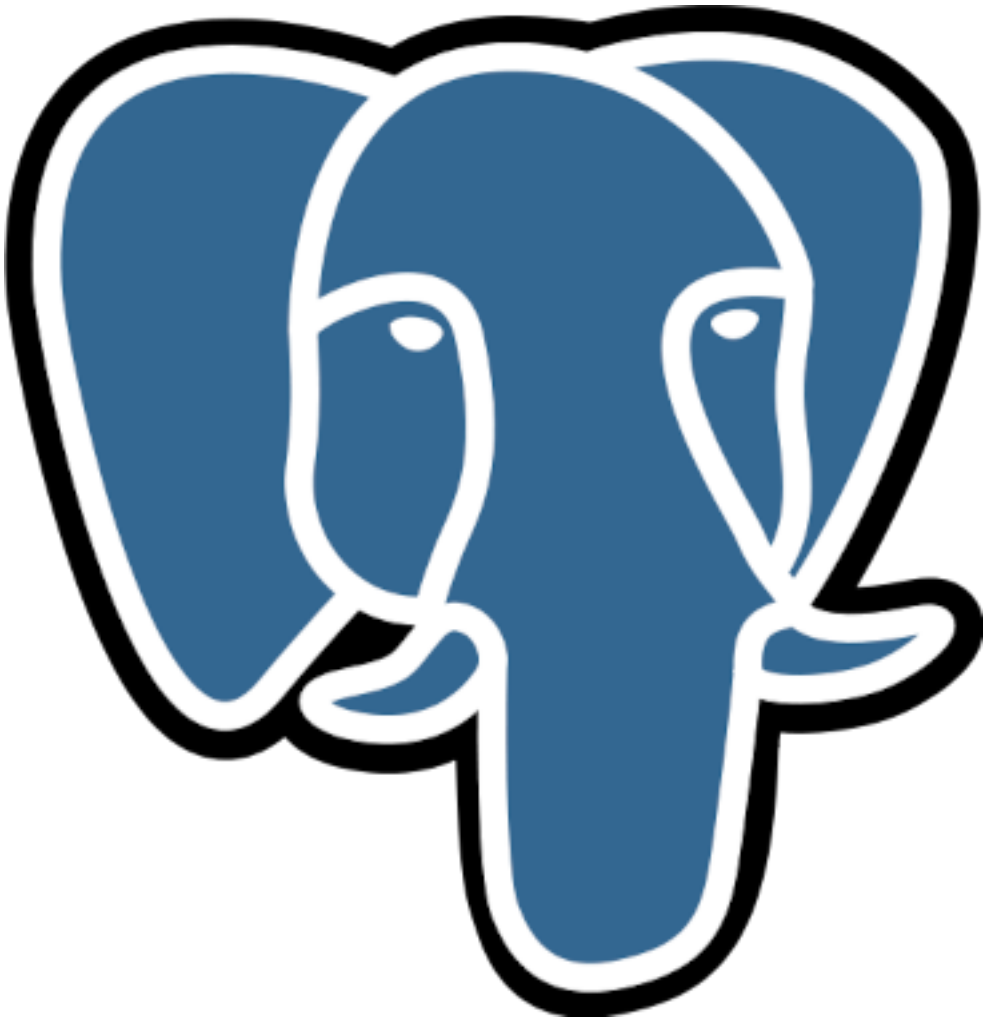
**Chers lectrices & lecteurs,**

Cette formation PostgreSQL est issue des manuels Dalibo. Ils ont été repris par Thibaut MADELAINE pour rentrer dans le format universitaire avec Cours Magistraux, Travaux Dirigés (sans ordinateurs) et Travaux Pratiques (avec ordinateur).

Au-delà du contenu technique en lui-même, l'intention des auteurs est de transmettre les valeurs qui animent et unissent les développeurs de PostgreSQL depuis toujours : partage, ouverture, transparence, créativité, dynamisme… Le but premier de cette formation 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 sur le site gitlab https://gitlab.com/madtibo/cours_dba_pg_universite/-/issues !

# Transactions et accès concurrents



---

## Programme de ce cours

- Semaine 1 : découverte de PostgreSQL
- **Semaine 2** : transactions et accès concurrents
  - MultiVersion Concurrency Control
  - Les transactions dans PostgreSQL
  - Le fonctionnement du *VACUUM*
  - Gestion des verrous

- Semaine 3 : missions du DBA
- Semaine 4 : optimisation et indexation

---

## MultiVersion Concurrency Control (MVCC)

- Le « noyau » de PostgreSQL
- Garantit ACID
- Permet les écritures concurrentes sur la même table

MVCC (Multi Version Concurrency Control) est le mécanisme interne de PostgreSQL utilisé pour garantir la cohérence des données lorsque plusieurs processus accèdent simultanément à la même table.

C'est notamment MVCC qui permet de sauvegarder facilement une base *à chaud* et d'obtenir une sauvegarde cohérente alors même que plusieurs utilisateurs sont potentiellement en train de modifier des données dans la base.

C'est la qualité de l'implémentation de ce système qui fait de PostgreSQL un des meilleurs SGBD au monde : chaque transaction travaille dans son image de la base, cohérent du début à la fin de ses opérations. Par ailleurs les écrivains ne bloquent pas les lecteurs et les lecteurs ne bloquent pas les écrivains, contrairement aux SGBD s'appuyant sur des verrous de lignes. Cela assure de meilleures performances, un fonctionnement plus fluide des outils s'appuyant sur PostgreSQL.

---

### MVCC et les verrous

- Une lecture ne bloque pas une écriture
- Une écriture ne bloque pas une lecture
- Une écriture ne bloque pas les autres écritures…
- …sauf pour la mise à jour de la **même ligne**.

MVCC maintient toutes les versions nécessaires de chaque tuple, ainsi **chaque transaction voit une image figée de la base** (appelée *snapshot*). Cette image correspond à l'état de la base lors du démarrage de la requête ou de la transaction, suivant le niveau d'*isolation* demandé par l'utilisateur à PostgreSQL pour la transaction.

MVCC fluidifie les mises à jour en évitant les blocages trop contraignants (verrous sur UPDATE) entre sessions et par conséquent de meilleures performances en contexte transactionnel.

Voici un exemple concret :

```
# SELECT now();
              now
-------------------------------
 2022-01-23 16:28:13.679663+02
(1 row)

# BEGIN;
BEGIN
# SELECT now();
              now
-------------------------------
 2022-01-23 16:28:34.888728+02
(1 row)

# SELECT pg_sleep(2);
 pg_sleep
----------

(1 row)

# SELECT now();
              now
-------------------------------
 2022-01-23 16:28:34.888728+02
(1 row)
```

---

**Transactions**

- Intimement liées à ACID et MVCC :
  - Une transaction est un ensemble d'opérations atomique
  - Le résultat d'une transaction est « tout ou rien »
- mots clés `BEGIN`, `COMMIT` et `ROLLBACK`

Voici un exemple de transaction:

```
=> BEGIN;
BEGIN
=> CREATE TABLE capitaines (id serial, nom text, age integer);
CREATE TABLE
```

```
=> INSERT INTO capitaines VALUES (1, 'Haddock', 35);
INSERT 0 1
=> SELECT age FROM capitaines;
 age
-----
  35
(1 ligne)

=> ROLLBACK;
ROLLBACK
=> SELECT age FROM capitaines;
ERROR:  relation "capitaines" does not exist
LINE 1: SELECT age FROM capitaines;
                        ^
```

On voit que la table capitaine a existé **à l'intérieur** de la transaction. Mais puisque cette transaction a été annulée (ROLLBACK), la table n'a pas été créée au final. Cela montre aussi le support du DDL transactionnel au sein de PostgreSQL.

```
=> BEGIN;
BEGIN
=> CREATE TABLE capitaines (id serial, nom text, age integer);
CREATE TABLE
=> INSERT INTO capitaines VALUES (1,'Haddock',35);
INSERT 0 1
=> COMMIT;
COMMIT
=> SELECT age FROM capitaines WHERE nom='Haddock';
 age
-----
  35
(1 row)
```

Grâce au mot clé COMMIT, la transaction a été validée. Aucune erreur n'ayant eu lieu, la table est bien créée et l'enregistrement est visible.

```
=> BEGIN;
BEGIN
=> UPDATE capitaines SET age=42;
UPDATE 1
=> SELECT * FROM capitaines;
 id |   nom    | age
```

```
----+---------+-----
  1 | Haddock |  42
(1 ligne)

=> DELETE * FROM capitaines;
ERROR:  syntax error at or near "*"
LIGNE 1 : DELETE * FROM capitaines;
                 ^
=> COMMIT;
ROLLBACK
=> SELECT * FROM capitaines;
 id |   nom   | age
----+---------+-----
  1 | Haddock |  35
(1 ligne)
```

Dans ce dernier cas, une erreur a été détectée durant la transaction. La réponse du serveur à la commande de validation, `COMMIT` est `ROLLBACK`. L'ensemble des modifications effectuées lors de cette transaction sont annulées.

---

**Niveaux d'isolation**

- Chaque transaction (et donc session) est isolée à un certain point :
  - elle ne voit pas les opérations des autres
  - elle s'exécute indépendamment des autres
- On peut spécifier le niveau d'isolation au démarrage d'une transaction:
  - `BEGIN ISOLATION LEVEL xxx;`

Chaque transaction, en plus d'être atomique, s'exécute séparément des autres. Le niveau de séparation demandé sera un compromis entre le besoin applicatif (pouvoir ignorer sans risque ce que font les autres transactions) et les contraintes imposées au niveau de PostgreSQL (performances, risque d'échec d'une transaction).

---

### Niveaux d'isolation dans PostgreSQL

- Niveaux d'isolation supportés
  - `READ COMMITED`
  - `REPEATABLE READ`
  - `SERIALIZABLE`

Le standard SQL spécifie quatre niveaux, mais PostgreSQL n'en supporte que trois.

---

### Niveau `READ UNCOMMITED`

- Autorise la lecture de données modifiées mais non validées par d'autres transactions
- Aussi appelé `DIRTY READS` par d'autres moteurs
- Pas de blocage entre les sessions
- Inutile sous PostgreSQL en raison du MVCC
- Si demandé, la transaction s'exécute en `READ COMMITTED`

Ce niveau d'isolation n'est nécessaire que pour les SGBD non-MVCC. Il est très dangereux : on peut lire des données invalides, ou temporaires, puisqu'on lit tous les enregistrements de la table, quel que soit leur état. Il est utilisé dans certains cas où les performances sont cruciales, au détriment de la justesse des données.

Sous PostgreSQL, ce mode est totalement inutile. Une transaction qui demande le niveau d'isolation `READ UNCOMMITTED` s'exécute en fait en `READ COMMITTED`.

---

### Niveau `READ COMMITTED`

- La transaction ne lit que les données validées en base
- Niveau d'isolation par défaut
- Un ordre SQL s'exécute dans un instantané (les tables semblent figées sur la durée de l'ordre)
- L'ordre suivant s'exécute dans un instantané différent

Ce mode est le mode par défaut, et est suffisant dans de nombreux contextes. PostgreSQL étant MVCC, les écrivains et les lecteurs ne se bloquent pas mutuellement, et chaque ordre s'exécute sur un instantané de la base (ce n'est pas un pré-requis de `READ COMMITTED` dans la norme SQL). On ne souffre

plus des lectures d'enregistrements non valides (`dirty reads`). On peut toutefois avoir deux problèmes majeurs d'isolation dans ce mode :

- Les lectures non-répétables (`non-repeatable reads`) : une transaction peut ne pas voir les mêmes enregistrements d'une requête sur l'autre, si d'autres transaction ont validé des modifications entre temps.
- Les lectures fantômes (`phantom reads`) : des enregistrements peuvent ne plus satisfaire une clause `WHERE` entre deux requêtes d'une même transaction.

---

#### Niveau `REPEATABLE READ`

- Instantané au début de la transaction
- Ne voit donc plus les modifications des autres transactions
- Voit toujours ses propres modifications
- Peut entrer en conflit avec d'autres transactions en cas de modification des mêmes enregistrements

Ce mode, comme son nom l'indique, permet de ne plus avoir de lectures non-répétables. Deux ordres SQL consécutifs dans la même transaction retourneront les mêmes enregistrements, dans la même version. En lecture seule, ces transactions ne peuvent pas échouer (elles sont entre autres utilisées pour réaliser des exports des données, par pg_dump).

En écriture, par contre (ou `SELECT FOR UPDATE`, `FOR SHARE`), si une autre transaction a modifié les enregistrements ciblés entre temps, une transaction en `REPEATABLE READ` va échouer avec l'erreur suivante :

```
ERROR: could not serialize access due to concurrent update
```

Il faut donc que l'application soit capable de la rejouer au besoin.

Dans la norme, ce niveau d'isolation souffre toujours des lectures fantômes, c'est-à-dire de lecture d'enregistrements qui ne satisfont plus la même clause `WHERE` entre deux exécutions de requêtes. Cependant, PostgreSQL est plus strict que la norme et ne permet pas ces lectures fantômes en `REPEATABLE READ`.

---

### Niveau `SERIALIZABLE`

- Niveau d'isolation maximum
- Plus de lectures non répétables
- Plus de lectures fantômes
- Instantané au démarrage de la transaction
- Verrouillage informatif des enregistrements consultés (verrouillage des prédicats)
- Erreurs de sérialisation en cas d'incompatibilité

Le niveau `SERIALIZABLE` permet de développer comme si chaque transaction se déroulait seule sur la base. En cas d'incompatibilité entre les opérations réalisées par plusieurs transactions, PostgreSQL annule celle qui déclenchera le moins de perte de données. Tout comme dans le mode `REPEATABLE READ`, il est essentiel de pouvoir rejouer une transaction si on développe en mode `SERIALIZABLE`. Par contre, on simplifie énormément tous les autres points du développement.

Ce mode empêche les erreurs dues à une transaction effectuant un `SELECT` d'enregistrements, puis d'autres traitements, pendant qu'une autre transaction modifie les enregistrements vus par le `SELECT` : il est probable que le `SELECT` initial de notre transaction utilise les enregistrements récupérés, et que le reste du traitement réalisé par notre transaction dépende de ces enregistrements. Si ces enregistrements sont modifiés par une transaction concurrente, notre transaction ne s'est plus déroulée comme si elle était seule sur la base, et on a donc une violation de sérialisation.

---

## L'implémentation MVCC de PostgreSQL

- Colonnes `ctid`/`xmin`/`xmax`
- Fichiers `clog`
- Avantages/inconvénients
- Opération `VACUUM`
- `Wrap-Around`

---

### `ctid`

- Colonne masquée par défaut
- Codée sur 6 octets

  - 4 octets pour la page
  - 2 octets pour la ligne

- Fournit une adresse physique dans une table

La localisation physique de la version de ligne au sein de sa table. Bien que le `ctid` puisse être utilisé pour trouver la version de ligne très rapidement, le `ctid` d'une ligne change si la ligne est actualisée ou déplacée par un `VACUUM FULL`. Le `ctid` est donc inutilisable comme identifiant de ligne sur le long terme.

---

**xmin et xmax (1/4)**

Table initiale :

| xmin | xmax | Nom | Solde |
|---|---|---|---|
| 100 | | M. Durand | 1500 |
| 100 | | Mme Martin | 2200 |

PostgreSQL stocke des informations de visibilité dans chaque version d'enregistrement.

- `xmin` : l'identifiant de la transaction créant cette version.
- `xmax` : l'identifiant de la transaction invalidant cette version.

Ici, les deux enregistrements ont été créés par la transaction 100. Il s'agit peut-être, par exemple, de la transaction ayant importé tous les soldes à l'initialisation de la base.

---

**xmin et xmax (2/4)**

```
BEGIN;
UPDATE soldes SET solde=solde-200 WHERE nom = 'M. Durand';
```

| xmin | xmax | Nom | Solde |
|---|---|---|---|
| 100 | **150** | M. Durand | 1500 |
| 100 | | Mme Martin | 2200 |
| **150** | | M. Durand | 1300 |

On décide d'enregistrer un virement de 200 € du compte de M. Durand vers celui de Mme Martin. Ceci doit être effectué dans une seule transaction : l'opération doit être atomique, sans quoi de l'argent pourrait apparaître ou disparaître de la table.

Nous allons donc tout d'abord démarrer une transaction (ordre SQL `BEGIN`). PostgreSQL fournit donc à notre session un nouveau numéro de transaction (150 dans notre exemple). Puis nous effectuerons :

```
UPDATE soldes SET solde=solde-200 WHERE nom = 'M. Durand';
```

---

**xmin et xmax (3/4)**

```
UPDATE soldes SET solde=solde+200 WHERE nom = 'Mme Martin';
```

| xmin | xmax | Nom | Solde |
|---|---|---|---|
| 100 | 150 | M. Durand | 1500 |
| 100 | **150** | Mme Martin | 2200 |
| 150 | | M. Durand | 1300 |
| **150** | | Mme Martin | 2400 |

Puis nous effectuerons :

```
UPDATE soldes SET solde=solde+200 WHERE nom = 'Mme Martin';
```

Nous avons maintenant deux versions de chaque enregistrement.

Notre session ne voit bien sûr plus que les nouvelles versions de ces enregistrements, sauf si elle décidait d'annuler la transaction, auquel cas elle reverrait les anciennes données.

Pour une autre session, la version visible de ces enregistrements dépend de plusieurs critères :

- La transaction 150 a-t-elle été validée ? Sinon elle est invisible
- La transaction 150 est-elle *postérieure* à la nôtre (numéro supérieur au notre), et sommes-nous dans un niveau d'isolation (*serializable*) qui nous interdit de voir les modifications faites depuis le début de notre transaction ?
- La transaction 150 a-t-elle été validée après le démarrage de la requête en cours ? Une requête, sous PostgreSQL, voit un instantané cohérent de la base, ce qui implique que toute transaction validée après le démarrage de la requête doit être ignorée.

Dans le cas le plus simple, 150 ayant été validée, une transaction 160 ne verra pas les premières versions : `xmax` valant 150, ces enregistrements ne sont pas visibles. Elle verra les secondes versions, puisque `xmin`=150, et pas de `xmax`.

---

**xmin et xmax (4/4)**

| xmin | xmax | Nom | Solde |
|---|---|---|---|
| 100 | 150 | M. Durand | 1500 |
| 100 | **150** | Mme Martin | 2200 |
| 150 | | M. Durand | 1300 |
| **150** | | Mme Martin | 2400 |

- Comment est effectuée la suppression d'un enregistrement ?
- Comment est effectuée l'annulation de la transaction 150 ?

- La suppression d'un enregistrement s'effectue simplement par l'écriture d'un `xmax` dans la version courante.
- Il n'y a rien à écrire dans les tables pour annuler une transaction. Il suffit de marquer la transaction comme étant annulée dans la CLOG.

---

**CLOG**

- La CLOG (Commit Log) enregistre l'état des transactions.

- Chaque transaction occupe 2 bits de CLOG

La CLOG est stockée dans une série de fichiers de 256 ko, stockés dans le répertoire pg_xact de PGDATA (répertoire racine de l'instance PostgreSQL).

Chaque transaction est créée dans ce fichier dès son démarrage et est encodée sur deux bits puisqu'une transaction peut avoir quatre états.

- TRANSACTION_STATUS_IN_PROGRESS : transaction en cours, c'est l'état initial
- TRANSACTION_STATUS_COMMITTED : la transaction a été validée
- TRANSACTION_STATUS_ABORTED : la transaction a été annulée
- TRANSACTION_STATUS_SUB_COMMITTED : ceci est utilisé dans le cas où la transaction comporte des sous-transactions, afin de valider l'ensemble des sous-transactions de façon atomique.

On a donc un million d'états de transactions par fichier de 256 ko.

Annuler une transaction (ROLLBACK) est quasiment instantané sous PostgreSQL : il suffit d'écrire TRANSACTION_STATUS_ABORTED dans l'entrée de CLOG correspondant à la transaction.

Toute modification dans la CLOG, comme toute modification d'un fichier de données (table, index, séquence), est bien sûr enregistrée tout d'abord dans les journaux de transactions (fichiers WAL dans le répertoire pg_wal).

---

**Avantages du MVCC PostgreSQL**

- Avantages :
  - avantages classiques de MVCC (concurrence d'accès)
  - implémentation simple et performante
  - peu de sources de contention
  - verrouillage simple d'enregistrement
  - rollback instantané
  - données conservées aussi longtemps que nécessaire
- Les lecteurs ne bloquent pas les écrivains, ni les écrivains les lecteurs.
- Le code gérant les instantanés est simple, ce qui est excellent pour la fiabilité, la maintenabilité et les performances.
- Les différentes sessions ne se gênent pas pour l'accès à une ressource commune (l'UNDO).

- Un enregistrement est facilement identifiable comme étant verrouillé en écriture : il suffit qu'il ait une version ayant un `xmax` correspondant à une transaction en cours.
- L'annulation est instantanée : il suffit d'écrire le nouvel état de la transaction dans la `clog`. Pas besoin de restaurer les valeurs précédentes, elles redeviennent automatiquement visibles.
- Les anciennes versions restent en ligne aussi longtemps que nécessaire. Elles ne pourront être effacées de la base qu'une fois qu'aucune transaction ne les considérera comme visibles.

---

**Inconvénients du MVCC PostgreSQL**

- Inconvénients :
  - Nettoyage des enregistrements (VACUUM)
  - Tables plus volumineuses
  - Pas de visibilité dans les index

Comme toute solution complexe, l'implémentation MVCC de PostgreSQL est un compromis. Les avantages cités précédemment sont obtenus au prix de concessions :

- Il faut nettoyer les tables de leurs enregistrements morts. C'est le travail de la commande VACUUM. On peut aussi voir ce point comme un avantage : contrairement à la solution UNDO, ce travail de nettoyage n'est pas effectué par le client faisant des mises à jour (et créant donc des enregistrements morts). Le ressenti est donc meilleur.
- Les tables sont forcément plus volumineuses que dans l'implémentation par UNDO, pour deux raisons :
  - Les informations de visibilité qui y sont stockées. Il y a un surcoût d'une douzaine d'octets par enregistrement.
  - Il y a toujours des enregistrements morts dans une table, une sorte de *fond de roulement*, qui se stabilise quand l'application est en régime stationnaire. Ces enregistrements sont recyclés à chaque passage de VACUUM.
- Les index n'ont pas d'information de visibilité. Il est donc nécessaire d'aller vérifier dans la table associée que l'enregistrement trouvé dans l'index est bien visible. Cela a un impact sur le temps d'exécution de requêtes comme `SELECT count(*)` sur une table : dans le cas le plus défavorable, il est nécessaire d'aller visiter tous les enregistrements pour s'assurer qu'ils sont bien visibles. La *visibility map* permet de limiter cette vérification aux données les plus récentes.
- Le VACUUM ne s'occupe pas de l'espace libéré par des colonnes supprimées (fragmentation verticale).

**Fonctionnement de VACUUM (1/3)**

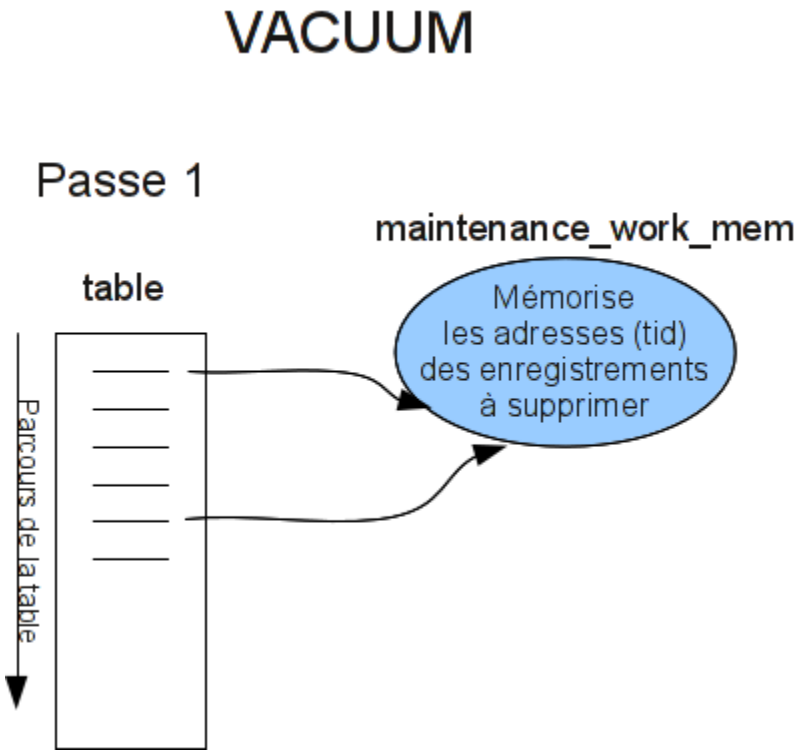



**Figure 1:** Algorithme du vacuum 1/3

Le traitement VACUUM se déroule en trois passes. Cette première passe parcourt la table à nettoyer, à la recherche d'enregistrements morts. Un enregistrement est mort s'il possède un `xmax` qui correspond à une transaction validée, et que cet enregistrement n'est plus visible dans l'instantané d'aucune transaction en cours sur la base.

L'enregistrement mort ne peut pas être supprimé immédiatement : des enregistrements d'index pointent vers lui et doivent aussi être nettoyés. Les adresses (`tid` ou `tuple id`) des enregistrements sont donc mémorisés par la session effectuant le vacuum, dans un espace mémoire dont la taille est à hauteur de `maintenance_work_mem`. Si `maintenance_work_mem` est trop petit pour contenir tous les enregistrements morts en une seule passe, vacuum effectue plusieurs séries de ces trois passes.

Un `tid` est composé du numéro de bloc et du numéro d'enregistrement dans le bloc.

**Fonctionnement de VACUUM (2/3)**

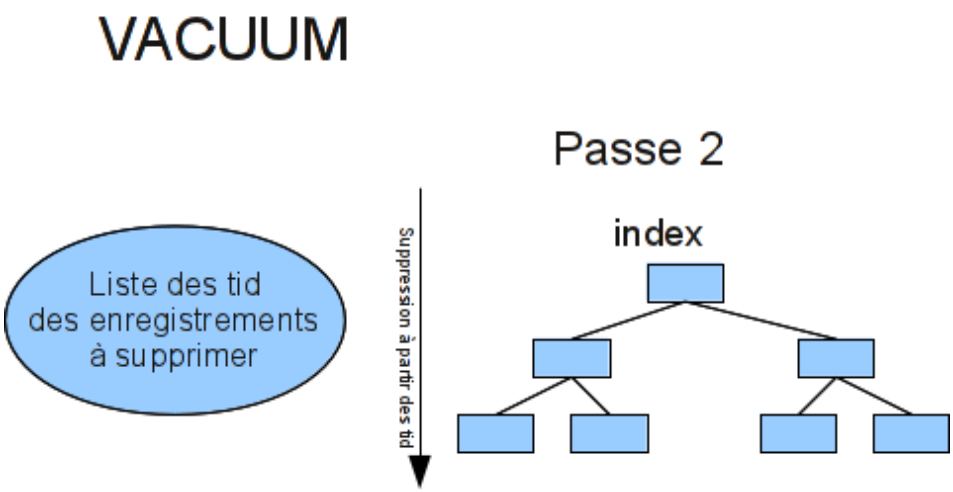



**Figure 2:** Algorithme du vacuum 2/3

La seconde passe se charge de nettoyer les entrées d'index. Vacuum possède une liste de `tid` à invalider. Il parcourt donc tous les index de la table à la recherche de ces tid et les supprime. En effet, les index sont triés afin de mettre en correspondance une valeur de clé (la colonne indexée par exemple) avec un `tid`. Il n'est par contre pas possible de trouver un `tid` directement. Les pages entièrement vides sont supprimées de l'arbre et stockées dans la liste des pages réutilisables, la **Free Space Map** (`FSM`).

---

**Fonctionnement de VACUUM (3/3)**

Maintenant qu'il n'y a plus d'entrée d'index pointant sur les enregistrements identifiés, nous pouvons supprimer les enregistrements de la table elle-même. C'est le rôle de cette passe, qui quant à elle, peut accéder directement aux enregistrements. Quand un enregistrement est supprimé d'un bloc, ce bloc est réorganisé afin de consolider l'espace libre, et cet espace libre est consolidé dans la **Free Space Map** (`FSM`).

Une fois cette passe terminée, si le parcours de la table n'a pas été terminé lors de la passe 1 (la `maintenance_work_mem` était pleine), le travail reprend où il en était du parcours de la table.

---

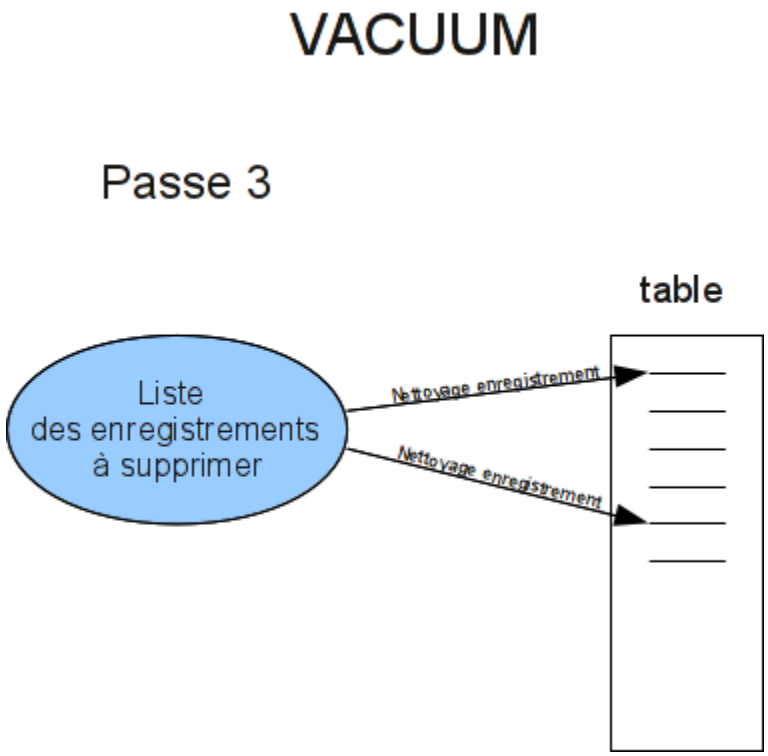



**Figure 3:** Algorithme du vacuum 3/3

### Le problème du Wraparound

*Wraparound* : bouclage d'un compteur

- Le compteur de transactions : 32 bits
- 4 milliards de transactions
- Qu'arrive-t-il si on boucle ?
- Quelles protections ?

Le compteur de transactions de PostgreSQL est stocké sur 32 bits. Il peut donc, en théorie, y avoir un dépassement de ce compteur au bout de 4 milliards de transactions. En fait, le compteur est cyclique, et toute transaction considère que les 2 milliards de transactions supérieures à la sienne sont dans le futur, et les 2 milliards inférieures dans le passé. Le risque de bouclage est donc plus proche des 2 milliards.

En théorie, si on bouclait, de nombreux enregistrements deviendraient invisibles, car validés par des transactions futures. Heureusement PostgreSQL l'empêche. Au fil des versions, la protection est devenue plus efficace.

- Le moteur trace la plus vieille transaction d'une base, et refuse toute nouvelle transaction à partir du moment où le stock de transaction disponible est à 10 millions. Il suffit de lancer un `VACUUM` sur la base incriminée à ce point, qui débloquera la situation, en nettoyant les plus anciens `xmin`.
- Depuis l'arrivée d'`AUTOVACUUM`, celui-ci déclenche automatiquement un `VACUUM` quand le *Wraparound* se rapproche trop. Ceci se produit même si `AUTOVACUUM` est désactivé.

Si vous voulez en savoir plus, la documentation officielle[1] contient un paragraphe sur ce sujet.

Ce qu'il convient d'en retenir, c'est que le système empêchera le wraparound en se bloquant. En cas de blocage, une protection contre ce phénomène se déclenche automatiquement en déclenchant un `VACUUM`, quel que soit le paramétrage d'`AUTOVACUUM`.

---

### commande VACUUM

- Lancement manuel du `VACUUM`
- option `FULL`
  - défragmente la table

[1] https://docs.postgresql.fr/current/maintenance.html

  - verrou exclusif

- option ANALYZE
  - met en plus à jour les statistiques

La commande VACUUM permet de récupérer l'espace utilisé par les lignes non visibles afin d'éviter un accroissement continuel du volume occupé sur le disque.

Cependant, un VACUUM fait rarement gagner de l'espace disque. Il faut utiliser l'option FULL pour cela. La commande VACUUM FULL libère l'espace consommé par les lignes périmées ou les colonnes supprimées, et le rend au système d'exploitation.

Cette variante de la commande VACUUM acquiert un verrou exclusif sur chaque table. Elle peut donc avoir un effet extrêmement négatif sur les performances de la base de données.

Quand faut-il utiliser VACUUM ?

- pour des nettoyages réguliers ;
- il s'agit d'une maintenance de base.

Quand faut-il utiliser VACUUM FULL ?

- après des suppressions massives de données ;
- lorsque la base n'est pas en production ;
- il s'agit d'une maintenance exceptionnelle.

Des VACUUM standards et une fréquence modérée sont une meilleure approche que des VACUUM FULL, même non fréquents, pour maintenir des tables mises à jour fréquemment.

VACUUM FULL est recommandé dans les cas où vous savez que vous avez supprimé ou modifié une grande partie des lignes d'une table, de façon à ce que la taille de la table soit réduite de façon conséquente.
Un REINDEX est exécuté lors d'un VACUUM FULL.

La commande vacuumdb permet d'exécuter facilement un VACUUM sur une ou toutes les bases, elle permet également la parallélisation de VACUUM sur plusieurs tables.

---

## Verrouillage et MVCC

La gestion des verrous est liée à l'implémentation de MVCC.

- Verrouillage d'objets en mémoire

- Verrouillage d'objets sur disque
- Paramètres

---

### Le gestionnaire de verrous

PostgreSQL possède un gestionnaire de verrous

- Verrous d'objet
- Niveaux de verrouillage
- Deadlock
- Vue pg_locks

PostgreSQL dispose d'un gestionnaire de verrous, comme tout SGBD.

Ce gestionnaire de verrous est capable de gérer des verrous sur des tables, sur des enregistrements, sur des ressources virtuelles. De nombreux types de verrous - 8 - sont disponibles, chacun entrant en conflit avec d'autres.

Chaque opération doit tout d'abord prendre un verrou sur les objets à manipuler.

Les noms des verrous peuvent prêter à confusion : `ROW SHARE` par exemple est un verrou de table, pas un verrou d'enregistrement. Il signifie qu'on a pris un verrou sur une table pour y faire des `SELECT FOR UPDATE` par exemple. Ce verrou est en conflit avec les verrous pris pour un `DROP TABLE`, ou pour un `LOCK TABLE`.

Le gestionnaire de verrous détecte tout verrou mortel (`deadlock`) entre deux sessions. Un `deadlock` est la suite de prise de verrous entraînant le blocage mutuel d'au moins deux sessions, chacune étant en attente d'un des verrous acquis par l'autre.

On peut accéder aux verrous actuellement utilisés sur un cluster par la vue `pg_locks`.

---

### Verrous sur enregistrement

- Le gestionnaire de verrous possède des verrous sur enregistrements.
- Ils sont :
  - transitoires
  - pas utilisés pour prendre les verrous définitifs

- Utilisation de verrous sur disque.
- Pas de risque de pénurie de verrous.

Le gestionnaire de verrous fournit des verrous sur enregistrement. Ceux-ci sont utilisés pour verrouiller un enregistrement le temps d'y écrire un `xmax`, puis libérés immédiatement.

Le verrouillage réel est implémenté comme suit :

- Chaque transaction verrouille son objet « identifiant de transaction » de façon exclusive.
- Une transaction voulant mettre à jour un enregistrement consulte le `xmax`. S'il constate que ce `xmax` est celui d'une transaction en cours, il demande un verrou exclusif sur l'objet « identifiant de transaction » de cette transaction. Qui ne lui est naturellement pas accordé. Il est donc placé en attente.
- Quand la transaction possédant le verrou se termine (`COMMIT` ou `ROLLBACK`), son verrou sur l'objet « identifiant de transaction » est libéré, débloquant ainsi l'autre transaction, qui peut reprendre son travail.

Ce mécanisme ne nécessite pas un nombre de verrous mémoire proportionnel au nombre d'enregistrements à verrouiller, et simplifie le travail du gestionnaire de verrous, celui-ci ayant un nombre bien plus faible de verrous à gérer.

Le mécanisme exposé ici est légèrement simplifié. Pour une explication approfondie, n'hésitez pas à consulter l'article suivant[2] issu de la base de connaissance Dalibo.

---

## Conclusion

- PostgreSQL dispose d'une implémentation MVCC complète, permettant :
  - Que les lecteurs ne bloquent pas les écrivains
  - Que les écrivains ne bloquent pas les lecteurs
  - Que les verrous en mémoire soient d'un nombre limité
- Cela impose par contre une mécanique un peu complexe, dont les parties visibles sont la commande VACUUM et le processus d'arrière-plan Autovacuum.

---

[2] https://kb.dalibo.com/verrouillage

## Travaux Dirigés 2

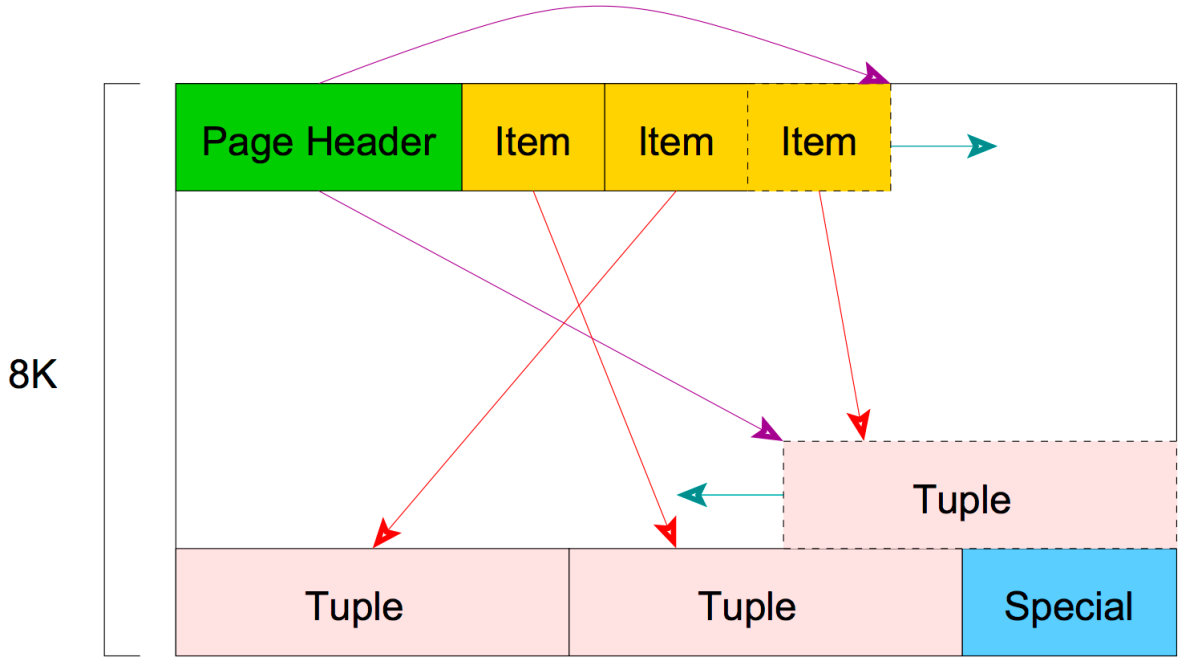
- *MVCC* et Vacuum

### Enoncé

### Question de cours

- Que signifie l'acronyme `ACID` ?
- Combien existe-t-il de niveaux d'isolations dans PostgreSQL ?
- Quels sont les inconvénients du MVCC ?

**Limites dans PostgreSQL** « *block_size (integer): reports the size of a disk block. It is determined by the value of BLCKSZ when building the server. The default value is 8192 bytes. The meaning of some configuration variables (such as shared_buffers) is influenced by block_size.* »




« The `oid` type is currently implemented as an unsigned four-byte integer. Therefore, it is not large enough to provide database-wide uniqueness in large databases, or even in large individual tables. »

« *`ctid`: The physical location of the row version within its table. Note that although the `ctid` can be used to locate the row version very quickly, a row's `ctid` will change if it is updated or moved by*

*VACUUM FULL. Therefore `ctid` is useless as a long-term row identifier. A primary key should be used to identify logical rows.* »

« *`tid`, or tuple identifier (row identifier). This is the data type of the system column `ctid`. A tuple ID is a pair (block number, tuple index within block) that identifies the physical location of the row within its table.* »

Définitions dans le fichier source `src/include/storage/block.h` :

```
typedef uint32 BlockNumber;
#define InvalidBlockNumber      ((BlockNumber) 0xFFFFFFFF)
#define MaxBlockNumber          ((BlockNumber) 0xFFFFFFFE)
```

**Questions**

- Quelle est le nombre maximal de lignes dans une table ?
- Quelle est la taille maximum d'une table sur disque ?

**Exploration des `ctid`** Soit la table t1 :

```
=# CREATE TABLE t1 (c1 int, c2 text);
CREATE TABLE
=# INSERT INTO t1 (c1, c2) SELECT i, md5(i::text) FROM
 ↪  generate_series(1,1000) i;
INSERT 0 1000
=# SELECT ctid, xmin, xmax, c1, c2 FROM t1;
  ctid   |  xmin   | xmax |  c1  |                c2
---------+---------+------+------+----------------------------------
 (0,1)   | 2760651 |    0 |    1 | c4ca4238a0b923820dcc509a6f75849b
 (0,2)   | 2760651 |    0 |    2 | c81e728d9d4c2f636f067f89cc14862c
 (0,3)   | 2760651 |    0 |    3 | eccbc87e4b5ce2fe28308fd9f2a7baf3
(...)
 (0,119) | 2760651 |    0 |  119 | 07e1cd7dca89a1678042477183b7ac3f
 (0,120) | 2760651 |    0 |  120 | da4fb5c6e93e74d3df8527599fa62642
 (1,1)   | 2760651 |    0 |  121 | 4c56ff4ce4aaf9573aa5dff913df997a
 (1,2)   | 2760651 |    0 |  122 | a0a080f42e6f13b3a2df133f073095dd
(...)
 (8,38)  | 2760651 |    0 |  998 | 9ab0d88431732957a618d4a469a0d4c3
 (8,39)  | 2760651 |    0 |  999 | b706835de79a2b4e80506f582af3676a
 (8,40)  | 2760651 |    0 | 1000 | a9b7ba70783b617e9998dc4dd82eb3c5

(1000 lignes)
```

Et la table t2 :

```
=# CREATE TABLE t2 (c1 int);
CREATE TABLE
=# INSERT INTO t2 (c1) SELECT i FROM generate_series(1,1000) i;
INSERT 0 1000
=# SELECT ctid, xmin, xmax, c1 FROM t2;
   ctid   |  xmin   | xmax |  c1
----------+---------+------+-------
 (0,1)    | 2760654 |    0 |     1
 (0,2)    | 2760654 |    0 |     2
 (0,3)    | 2760654 |    0 |     3
 (0,4)    | 2760654 |    0 |     4
 (0,5)    | 2760654 |    0 |     5
 (0,6)    | 2760654 |    0 |     6
(...)
 (0,224)  | 2760654 |    0 |   224
 (0,225)  | 2760654 |    0 |   225
 (0,226)  | 2760654 |    0 |   226
 (1,1)    | 2760654 |    0 |   227
 (1,2)    | 2760654 |    0 |   228
 (1,3)    | 2760654 |    0 |   229
(...)
 (4,94)   | 2760657 |    0 |   998
 (4,95)   | 2760657 |    0 |   999
 (4,96)   | 2760657 |    0 |  1000

(1000 lignes)
```

**Questions**

- Pourquoi le nombre de blocs utilisé est-il différent pour ces 2 tables ?
- Pourquoi la valeur maximum du deuxième champ du `ctid` (l'index du tuple) n'est pas le même pour les 2 tables ?
- Sachant que les informations stockées dans le page header font 24 octets et qu'un item occupe 4 octets, estimer la taille d'un tuple pour ces 2 tables.

**Etude du MVCC** Session 1 :

```
BEGIN ISOLATION LEVEL REPEATABLE READ;
SELECT ctid, xmin, xmax, c1 FROM t2;
```

Session 2 :

```
UPDATE t2 SET c1=c1+1;
```

- Que contient physiquement la table t2 ?

Session 1 :

```
TRUNCATE t2;
```

Session 2 :

```
SELECT ctid, xmin, xmax, c1 FROM t2;
```

- Quel est le résulat de la requête dans la session 2 ?

Session 1 :

```
ROLLBACK;
VACUUM t2;
INSERT INTO t2 (c1) SELECT i FROM generate_series(1002,1500) i;
SELECT ctid, xmin, xmax, c1 FROM t2;
```

- Que contient physiquement la table t2 ?

---

**Solutions du TD 2**

**Question de cours**

- Que signifie l'acronyme ACID ?
  - Atomicité
  - Cohérence
  - Isolation
  - Durabilité
- Combien existe-t-il de niveaux d'isolations dans PostgreSQL ?
  - `READ COMMITTED`.

  - `REPEATABLE READ`.
  - `SERIALIZABLE`

- Quels sont les inconvénients du MVCC ?

  - Besoin de nettoyer les enregistrements grâce au `VACUUM`,
  - Les tables sont plus volumineuses car elles contiennent des enregistrements morts,
  - Pas de visibilité dans les index.

#### Limites dans PostgreSQL

- Quelle est le nombre maximal de lignes dans une table ?

De façon théorique, on peut lister au maximum 65536 (`2^16`) éléments dans une page (le numéro de ligne étant stocké sur 2 octets). Avec un maximum de `2^32` pages, on arrive, de façon théorique, à plus de 280 mille milliards de lignes dans une table.

Ce n'est pas le nombre de lignes qui va poser problème en premier.

- Quelle est la taille maximum d'une table sur disque ?

Au point de vue logique, une table sur disque est découpée en page de 8 Ko. Un numéro de bloc étant stocké sur 4 octets, la taille maximum sur disque d'une table est de 32 To :
`2^32 * 8 * 1024 = 32 * 1024^4`

Dans le cas où cette limite était atteinte, on pourra utiliser les tables partitionnées pour permettre de stocker plus de données.

#### Exploration des `ctid`

- Pourquoi le nombre de blocs utilisé est-il différent pour ces 2 tables ?

Pour la table `t1`, un tuple est composé des colonnes système, d'une colonne entier et d'une colonne texte. Pour la table `t2` un tuple est composé des colonnes système et d'une colonne entier uniquement.
Pour un même nombre de tuple, la table `t1` est plus volumineuse et utilise donc plus de blocs.

- Pourquoi la valeur maximum du deuxième champ du `ctid` (l'index du tuple) n'est pas le même pour les 2 tables ?

L'index du tuple permet de trouver l'adresse d'un tuple au sein d'un bloc.

- Sachant que les informations stockées dans le page header font 24 octets et qu'un item occupe 4 octets, estimer la taille d'un tuple pour ces 2 tables.

Table 1 :

$$24 + 4 \times 120 + t \times 120 = 8192 <=> t = 64$$

Table 2 :

$$24 + 4 \times 226 + t \times 226 = 8192 <=> t = 32$$

On peut vérifier ces valeurs en utilisant l'extension *pageinspect* :

```
=# CREATE EXTENSION pageinspect ;
CREATE EXTENSION

=# SELECT lp_len FROM heap_page_items(get_raw_page('t1', 0)) LIMIT 1;
 lp_len
--------
     61
(1 ligne)

=# SELECT lp_len FROM heap_page_items(get_raw_page('t2', 0)) LIMIT 1;
 lp_len
--------
     28
(1 ligne)
```

La différence entre les 2 calculs tient au fait que notre calcul approximatif ne tenait compte ni du padding, ni de l'espace disponible entre les blocs items et les blocs tuples.

On constate avec ce calcul que le stockage d'une entier sur 4 octets implique un overhead de 85%. le stockage de plus d'information vient réduire ce ratio.

**Etude du MVCC** Jouer les commandes avec *psql*.

- Que contient physiquement la table t2 ?

La table t2 contient toutes les lignes insérée initialement ainsi que les nouvelles lignes issues de notre mise à jour. Le nombre de lignes stocké a donc doublé. La table occupe 9 blocs.

- Quel est le résulat de la requête dans la session 2 ?

La requête est en attente de la fin de la transaction dans la session 1. En effet, un TRUNCATE prend un verrou de très haut niveau : ACCESS EXCLUSIVE. Ce type de verrou bloque toute autre opération.

- Que contient physiquement la table t2 ?

La table t2 contient toujours 9 blocs. Cependant, la première moitié des blocs a été nettoyée par le VACUUM car les lignes n'étaient plus visibles. Cet espace libéré a été réutilisé pour stocker les 500 nouvelles lignes. Il reste des espaces réutilisables pour de nouvelles mises (insertion ou mises à jour).

---

## Travaux Pratiques 2

- *MVCC*, *VACUUM* et verrous

### Rappel

Durant ces travaux pratiques, nous allons utiliser la machine virtuelle du TP 1 pour héberger notre serveur de base de données PostgreSQL.

Effectuez les manipulations nécessaires pour réaliser les actions listées dans la section *Énoncés*.

Vous pouvez vous aider du cours, du dernier TD, des précédents TP, ainsi que de l'aide en ligne ou des pages de manuels (`man`).

### Enoncés

### Effets de MVCC

- Créez une nouvelle base de données et s'y connecter.
- Créez une nouvelle table t1 avec une colonne d'entier et une colonne de texte.
- Ajoutez cinq lignes dans cette table.
- Lisez la table.
- Commencez une transaction et modifiez une ligne.
- Lisez la table.
- Que remarquez-vous ?
- Ouvrez une autre session et lisez la table.
- Qu'observez-vous ?
- Lisez la table ainsi que les informations systèmes *xmin* et *xmax* pour les deux sessions.
- Récupérez maintenant en plus le *ctid*.
- Validez la transaction.
- Installez l'extension *pageinspect*.
- Décodez le bloc 0 de la table t1 à l'aide de cette extension.
- Que remarquez-vous ?

**Traiter la fragmentation** `VACUUM VERBOSE` :

- Exécutez un `VACUUM VERBOSE` sur la table t1.
- Des lignes ont-elles été nettoyées ?

Suppression en début de table :

- Créez une table t2 avec une colonne de type integer.
- Désactivez l'autovacuum pour cette table.
- Insérez un million de lignes dans cette table.
- Récupérez la taille de la table.
- Supprimez les 500000 premières lignes.
- Récupérez la taille de la table. Qu'en déduisez-vous ?
- Exécutez un `VACUUM`.
- Récupérez la taille de la table. Qu'en déduisez-vous ?
- Exécutez un `VACUUM FULL`.
- Récupérez la taille de la table. Qu'en déduisez-vous ?

Suppression en fin de table :

- Créez une table t3 avec une colonne de type integer.
- Désactivez l'autovacuum pour cette table.
- Insérez un million de lignes dans cette table.
- Récupérez la taille de la table.
- Supprimez les 500000 dernières lignes.
- Récupérez la taille de la table. Qu'en déduisez-vous ?
- Exécutez un `VACUUM`.
- Récupérez la taille de la table. Qu'en déduisez-vous ?

**Détecter la fragmentation**

- Installez l'extension *pg_freespacemap*.
- Créez une nouvelle table t4 avec une colonne d'entier et une colonne de texte.
- Désactivez l'autovacuum pour cette table.
- Insérer un million de lignes dans cette table.
- Que rapporte *pg_freespacemap* quant à l'espace libre de la table ?
- Modifier les 200 000 premières lignes.
- Que rapporte *pg_freespacemap* quant à l'espace libre de la table ?
- Exécutez un `VACUUM` sur la table.
- Que rapporte *pg_freespacemap* quant à l'espace libre de la table ? Qu'en déduisez-vous ?

- Récupérez la taille de la table.
- Exécutez un `VACUUM FULL` sur la table.
- Récupérez la taille de la table et l'espace libre rapporté par pg_freespacemap. Qu'en déduisez-vous ?

#### Gestion de l'autovacuum

- Créez une table t5 avec une colonne d'entier.
- Insérer un million de lignes dans cette table.
- Que contient la vue *pg_stat_user_tables* pour cette table ?
- Modifier les 200 000 premières lignes.
- Attendez une minute.
- Que contient la vue _pg_stat_user__tables pour cette table ?
- Modifier 60 lignes supplémentaires de cette table.
- Attendez une minute.
- Que contient la vue pg_stat_user_tables pour cette table ? Qu'en déduisez-vous ?
- Descendez le facteur d'échelle de cette table à 10 % pour le `VACUUM`.
- Modifiez les 200 000 lignes suivantes de cette table ?
- Attendez une minute.
- Que contient la vue *pg_stat_user_tables* pour cette table ? Qu'en déduisez-vous ?

#### Verrous

- Ouvrez une transaction et lisez la table t1.
- Ouvrez une autre transaction, et tentez de supprimer la table t1.
- Listez les processus du serveur PostgreSQL. Que remarquez-vous ?
- Récupérez la liste des sessions en attente d'un verrou avec la vue *pg_stat_activity*.
- Récupérez la liste des verrous en attente pour la requête bloquée.
- Récupérez le nom de l'objet dont on n'arrive pas à récupérer le verrou.
- Récupérez la liste des verrous sur cet objet. Quel processus a verrouillé la table t1 ?
- Retrouvez les informations sur la session bloquante.

---

### Solutions du TP 2

#### Effets de MVCC

- Créez une nouvelle base de données et s'y connecter.

```
postgres=# CREATE DATABASE mvcc;
CREATE DATABASE
postgres=# \c mvcc
```

- Créez une nouvelle table t1 avec une colonne d'entier et une colonne de texte.

```
mvcc=# CREATE TABLE t1 (c1 integer, c2 text);
CREATE TABLE
```

- Ajoutez cinq lignes dans cette table.

```
mvcc=# INSERT INTO t1 VALUES
  (1, 'un'), (2, 'deux'), (3, 'trois'), (4, 'quatre'), (5, 'cinq');
INSERT 0 5
```

- Lisez la table.

```
mvcc=# SELECT * FROM t1;
 c1 |   c2
----+--------
  1 | un
  2 | deux
  3 | trois
  4 | quatre
  5 | cinq
(5 rows)
```

- Commencez une transaction et modifiez une ligne.

```
mvcc=# BEGIN;
BEGIN
mvcc=# UPDATE t1 SET c2=upper(c2) WHERE c1=3;
UPDATE 1
```

- Lisez la table.

```
mvcc=# SELECT * FROM t1;
 c1 |   c2
----+--------
  1 | un
  2 | deux
  4 | quatre
  5 | cinq
  3 | TROIS
(5 rows)
```

- Que remarquez-vous ?

La ligne mise à jour n'apparaît plus, ce qui est normal. Elle apparaît en fin de table. En effet, quand un UPDATE est exécuté, la ligne courante est considérée comme morte et une nouvelle ligne est ajoutée, avec les valeurs modifiées. Comme nous n'avons pas demandé de récupérer les résultats dans un certain ordre, les lignes sont affichées dans leur ordre de stockage dans les blocs de la table.

- Ouvrez une autre session et lisez la table.

```
mvcc=# SELECT * FROM t1;
 c1 |   c2
----+--------
  1 | un
  2 | deux
  3 | trois
  4 | quatre
  5 | cinq
(5 rows)
```

- Qu'observez-vous ?

Les autres sessions voient toujours l'ancienne version de ligne, tant que la transaction n'a pas été validée. Et du coup, l'ordre des lignes en retour n'est pas le même vu que cette version de ligne était introduite avant.

- Lisez la table ainsi que les informations systèmes *xmin* et *xmax* pour les deux sessions.

Voici ce que renvoie la session qui a fait la modification ::

```
mvcc=# SELECT xmin, xmax, * FROM t1;
 xmin | xmax | c1 |   c2
------+------+----+--------
 1930 |    0 |  1 | un
 1930 |    0 |  2 | deux
 1930 |    0 |  4 | quatre
 1930 |    0 |  5 | cinq
 1931 |    0 |  3 | TROIS
(5 rows)
```

Et voici ce que renvoie l'autre session :

```
mvcc=# SELECT xmin, xmax, * FROM t1;
 xmin | xmax | c1 |   c2
------+------+----+--------
 1930 |    0 |  1 | un
 1930 |    0 |  2 | deux
 1930 | 1931 |  3 | trois
```

```
 1930 |    0 |  4 | quatre
 1930 |    0 |  5 | cinq
(5 rows)
```

La transaction 1931 est celle qui a réalisé la modification. La colonne xmin de la nouvelle version de ligne contient ce numéro. De même pour la colonne xmax de l'ancienne version de ligne. PostgreSQL se base sur cette information pour savoir si telle transaction peut lire telle ou telle ligne.

- Récupérez maintenant en plus le *ctid*.

Voici ce que renvoie la session qui a fait la modification :

```
mvcc=# SELECT ctid, xmin, xmax, * FROM t1;
 ctid  | xmin | xmax | c1 |   c2
-------+------+------+----+--------
 (0,1) | 1930 |    0 |  1 | un
 (0,2) | 1930 |    0 |  2 | deux
 (0,4) | 1930 |    0 |  4 | quatre
 (0,5) | 1930 |    0 |  5 | cinq
 (0,6) | 1931 |    0 |  3 | TROIS
(5 rows)
```

Et voici ce que renvoie l'autre session :

```
mvcc=# SELECT ctid, xmin, xmax, * FROM t1;
 ctid  | xmin | xmax | c1 |   c2
-------+------+------+----+--------
 (0,1) | 1930 |    0 |  1 | un
 (0,2) | 1930 |    0 |  2 | deux
 (0,3) | 1930 | 1931 |  3 | trois
 (0,4) | 1930 |    0 |  4 | quatre
 (0,5) | 1930 |    0 |  5 | cinq
(5 rows)
```

La colonne ctid contient une paire d'entiers. Le premier indique le numéro de bloc, le second le numéro de l'enregistrement dans le bloc. Autrement, elle précise la position de l'enregistrement sur le fichier de la table.

En récupérant cette colonne, on voit bien que la première session voit la nouvelle position (enregistrement 6 du bloc 0) et que la deuxième session voit l'ancienne (enregistrement 3 du bloc 0).

- Validez la transaction.

```
mvcc=# COMMIT;
COMMIT
```

- Installez l'extension *pageinspect*.

```
mvcc=# CREATE EXTENSION pageinspect;
CREATE EXTENSION
```

- Décodez le bloc 0 de la table t1 à l'aide de cette extension.

```
mvcc=# SELECT * FROM heap_page_items(get_raw_page('t1',0));
 lp | lp_off | lp_flags | lp_len | t_xmin | t_xmax | t_field3 | t_ctid |
----+--------+----------+--------+--------+--------+----------+--------+-
  1 |   8160 |        1 |     31 |   2169 |      0 |        0 | (0,1)  |
  2 |   8120 |        1 |     33 |   2169 |      0 |        0 | (0,2)  |
  3 |   8080 |        1 |     34 |   2169 |   2170 |        0 | (0,6)  |
  4 |   8040 |        1 |     35 |   2169 |      0 |        0 | (0,4)  |
  5 |   8000 |        1 |     33 |   2169 |      0 |        0 | (0,5)  |
  6 |   7960 |        1 |     34 |   2170 |      0 |        0 | (0,6)  |

 lp | t_infomask2 | t_infomask | t_hoff | t_bits | t_oid
----+-------------+------------+--------+--------+-------
  1 |           2 |       2306 |     24 |        |
  2 |           2 |       2306 |     24 |        |
  3 |       16386 |        258 |     24 |        |
  4 |           2 |       2306 |     24 |        |
  5 |           2 |       2306 |     24 |        |
  6 |       32770 |      10242 |     24 |        |
(6 rows)
```

- Que remarquez-vous ?
  - les six lignes sont bien présentes ;
  - le t_ctid ne contient plus (0,3) mais l'adresse de la nouvelle ligne (ie, (0,6] ;
  - *t_infomask2* est un champ de bits, la valeur 16386 pour l'ancienne version nous indique que le changement a eu lieu en utilisant la technologie HOT.

### Traiter la fragmentation

- Exécutez un VACUUM VERBOSE sur la table t1.

```
mvcc=# VACUUM VERBOSE t1;
INFO:  vacuuming "public.t1"
INFO:  "t1": found 1 removable, 5 nonremovable row versions in 1 out of 1
 ↪  pages
DÉTAIL : 0 dead row versions cannot be removed yet, oldest xmin: 2760700
There were 0 unused item pointers.
```

```
Skipped 0 pages due to buffer pins, 0 frozen pages.
0 pages are entirely empty.
CPU: user: 0.00 s, system: 0.00 s, elapsed: 0.00 s.
INFO:  vacuuming "pg_toast.pg_toast_18739"
INFO:  index "pg_toast_18739_index" now contains 0 row versions in 1 pages
DÉTAIL : 0 index row versions were removed.
0 index pages have been deleted, 0 are currently reusable.
CPU: user: 0.00 s, system: 0.00 s, elapsed: 0.00 s.
INFO:  "pg_toast_18739": found 0 removable, 0 nonremovable row versions in 0
 ↪  out of 0 pages
DÉTAIL : 0 dead row versions cannot be removed yet, oldest xmin: 2760700
There were 0 unused item pointers.
Skipped 0 pages due to buffer pins, 0 frozen pages.
0 pages are entirely empty.
CPU: user: 0.00 s, system: 0.00 s, elapsed: 0.00 s.
VACUUM
```

- Des lignes ont-elles été nettoyées ?

```
INFO:  "t1": found 1 removable, 5 nonremovable row versions in 1 out of 1
 ↪  pages
```

Il y a une ligne obsolète (et récupérable) et cinq lignes vivantes sur le seul bloc de la table.

- Créez une table t2 avec une colonne de type integer.

```
mvcc=# CREATE TABLE t2(c1 integer);
CREATE TABLE
```

- Désactivez l'autovacuum pour cette table.

```
mvcc=# ALTER TABLE t2 SET (autovacuum_enabled=false);
ALTER TABLE
```

- Insérez un million de lignes dans cette table.

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

- Récupérez la taille de la table.

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t2'));
 pg_size_pretty
----------------
 35 MB
(1 row)
```

- Supprimez les 500000 premières lignes.

```
mvcc=# DELETE FROM t2 WHERE id<500000;
DELETE 499999
```

- Récupérez la taille de la table. Qu'en déduisez-vous ?

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t2'));
 pg_size_pretty
----------------
 35 MB
(1 row)
```

Un DELETE ne permet pas de regagner de la place sur le disque. Les lignes supprimées sont uniquement marquées comme étant mortes.

- Exécutez un VACUUM.

```
mvcc=# VACUUM t2;
VACUUM
```

- Récupérez la taille de la table. Qu'en déduisez-vous ?

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t2'));
 pg_size_pretty
----------------
 35 MB
(1 row)
```

- Exécutez un VACUUM FULL.

```
mvcc=# VACUUM FULL t2;
VACUUM
```

- Récupérez la taille de la table. Qu'en déduisez-vous ?

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t2'));
 pg_size_pretty
----------------
 17 MB
(1 row)
```

Dans le cas du VACUUM FULL, on gagne en place disque. L'opération défragmente la table et permet, on de récupérer les espaces morts.

- Créez une table t3 avec une colonne de type integer.

```
mvcc=# CREATE TABLE t3(id integer);
CREATE TABLE
```

- Désactivez l'autovacuum pour cette table.

```
mvcc=# ALTER TABLE t3 SET (autovacuum_enabled=false);
ALTER TABLE
```

- Insérez un million de lignes dans cette table.

```
mvcc=# INSERT INTO t3 SELECT generate_series(1, 1000000);
INSERT 0 1000000
```

- Récupérez la taille de la table.

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t3'));
 pg_size_pretty
----------------
 35 MB
(1 row)
```

- Supprimez les 500000 dernières lignes.

```
mvcc=# DELETE FROM t3 WHERE id>500000;
DELETE 500000
```

- Récupérez la taille de la table. Qu'en déduisez-vous ?

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t3'));
 pg_size_pretty
----------------
 35 MB
(1 row)
```

Un DELETE ne permet pas de regagner de la place sur le disque. Les lignes supprimées sont uniquement marquées comme étant mortes.

- Exécutez un VACUUM.

```
mvcc=# VACUUM t3;
VACUUM
```

- Récupérez la taille de la table. Qu'en déduisez-vous ?

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t3'));
 pg_size_pretty
----------------
 17 MB
(1 row)
```

En fait, il existe un cas où on peut gagner de l'espace disque suite à un VACUUM simple : quand l'espace récupéré se trouve en fin de table et qu'il est possible de prendre rapidement un verrou exclusif sur la table pour la tronquer. C'est assez peu fréquent mais c'est une optimisation intéressante.

### Détecter la fragmentation

- Installez l'extension *pg_freespacemap*.

```
mvcc=# CREATE EXTENSION pg_freespacemap;
CREATE EXTENSION
```

- Créez une nouvelle table t4 avec une colonne d'entier et une colonne de texte.

```
mvcc=# CREATE TABLE t4 (c1 integer, c2 text);
CREATE TABLE
```

- Désactivez l'autovacuum pour cette table.

```
mvcc=# ALTER TABLE t4 SET (autovacuum_enabled=false);
ALTER TABLE
```

- Insérer un million de lignes dans cette table.

```
mvcc=# INSERT INTO t4 SELECT i, 'Ligne '||i FROM generate_series(1, 1000000)
 ↪  AS i;
INSERT 0 1000000
```

- Que rapporte *pg_freespacemap* quant à l'espace libre de la table ?

```
mvcc=# SELECT sum(avail) FROM pg_freespace('t4'::regclass);
 sum
-----
   0
(1 row)
```

- Modifier les 200 000 premières lignes.

```
mvcc=# UPDATE t4 SET c2=upper(c2) WHERE c1<200000;
UPDATE 199999
```

- Que rapporte *pg_freespacemap* quant à l'espace libre de la table ?

```
mvcc=# SELECT sum(avail) FROM pg_freespace('t4'::regclass);
 sum
-----
  32
(1 row)
```

- Exécutez un VACUUM sur la table.

```
mvcc=# VACUUM t4;
VACUUM
```

- Que rapporte *pg_freespacemap* quant à l'espace libre de la table ? Qu'en déduisez-vous ?

```
mvcc=# SELECT sum(avail) FROM pg_freespace('t4'::regclass);
   sum
---------
 8806784
(1 row)
```

Il faut exécuter un VACUUM pour que PostgreSQL renseigne la structure FSM, ce qui nous permet de connaître le taux de fragmentation de la table.

- Récupérez la taille de la table.

```
mvcc=# SELECT pg_size_pretty(pg_table_size('t4'));
 pg_size_pretty
----------------
 58 MB
(1 row)
```

- Exécutez un VACUUM FULL sur la table.

```
mvcc=# VACUUM FULL t4;
VACUUM
```

- Récupérez la taille de la table et l'espace libre rapporté par pg_freespacemap. Qu'en déduisez-vous ?

```
mvcc=# SELECT sum(avail) FROM pg_freespace('t4'::regclass);
 sum
-----
   0
(1 row)

mvcc=# SELECT pg_size_pretty(pg_table_size('t4'));
 pg_size_pretty
----------------
 49 MB
(1 row)
```

VACUUM FULL a supprimé les espaces morts, ce qui nous a fait gagner entre 8 et 9 Mo. La taille de la table maintenant correspond bien à celle de l'ancienne table, moins la place prise par les lignes mortes.

### Gestion de l'autovacuum

- Créez une table t5 avec une colonne d'entier.

```
mvcc=# CREATE TABLE t5 (id integer);
CREATE TABLE
```

- Insérer un million de lignes dans cette table.

```
mvcc=# INSERT INTO t5 SELECT generate_series(1, 1000000);
INSERT 0 1000000
```

- Que contient la vue *pg_stat_user_tables* pour cette table ?

```
mvcc=# \x
Expanded display is on.
mvcc=# SELECT * FROM pg_stat_user_tables WHERE relname='t5';
-[ RECORD 1 ]-------+--------
relid               | 18745
schemaname          | public
relname             | t5
seq_scan            | 0
seq_tup_read        | 0
idx_scan            |
idx_tup_fetch       |
n_tup_ins           | 1000000
n_tup_upd           | 0
n_tup_del           | 0
n_tup_hot_upd       | 0
n_live_tup          | 1000000
n_dead_tup          | 0
n_mod_since_analyze | 1000000
last_vacuum         |
last_autovacuum     |
last_analyze        |
last_autoanalyze    |
vacuum_count        | 0
autovacuum_count    | 0
analyze_count       | 0
autoanalyze_count   | 0
```

- Modifier les 200 000 premières lignes.

```
mvcc=# UPDATE t5 SET id=2000000 WHERE id<200001;
UPDATE 200000
```

- Attendez une minute.

```
mvcc=# SELECT pg_sleep(60);
```

- Que contient la vue _pg_stat_user__tables pour cette table ?

```
mvcc=# SELECT * FROM pg_stat_user_tables WHERE relname='t5';
-[ RECORD 1 ]-------+------------------------------
relid               | 18745
schemaname          | public
relname             | t5
seq_scan            | 1
seq_tup_read        | 1000000
idx_scan            |
idx_tup_fetch       |
n_tup_ins           | 1000000
n_tup_upd           | 200000
n_tup_del           | 0
n_tup_hot_upd       | 0
n_live_tup          | 1000000
n_dead_tup          | 200000
n_mod_since_analyze | 0
last_vacuum         |
last_autovacuum     |
last_analyze        |
last_autoanalyze    | 2021-11-15 22:37:18.039358+01
vacuum_count        | 0
autovacuum_count    | 0
analyze_count       | 0
autoanalyze_count   | 2
```

- Modifier 60 lignes supplémentaires de cette table.

```
mvcc=# UPDATE t5 SET id=2000000 WHERE id<200060;
UPDATE 59
```

- Attendez une minute.

```
mvcc=# SELECT pg_sleep(60);
```

- Que contient la vue pg_stat_user_tables pour cette table ? Qu'en déduisez-vous ?

```
mvcc=# SELECT * FROM pg_stat_user_tables WHERE relname='t5';
-[ RECORD 1 ]-------+------------------------------
relid               | 18745
schemaname          | public
```

```
relname             | t5
seq_scan            | 2
seq_tup_read        | 2000000
idx_scan            |
idx_tup_fetch       |
n_tup_ins           | 1000000
n_tup_upd           | 200059
n_tup_del           | 0
n_tup_hot_upd       | 10
n_live_tup          | 1000000
n_dead_tup          | 0
n_mod_since_analyze | 59
last_vacuum         |
last_autovacuum     | 2020-11-15 22:39:18.597124+01
last_analyze        |
last_autoanalyze    | 2020-11-15 22:37:18.039358+01
vacuum_count        | 0
autovacuum_count    | 1
analyze_count       | 0
autoanalyze_count   | 2
```

Un VACUUM a été automatiquement exécuté sur cette table, suite à la suppression de plus de 200 050 lignes (`threshold + scale factor * #lines`). Il a fallu attendre que l'autovacuum vérifie l'état des tables, d'où l'attente de 60 secondes.

Notez aussi que `n_dead_tup` est revenu à 0 après le VACUUM. C'est le compteur qui est comparé à la limite avant exécution d'un VACUUM.

- Descendez le facteur d'échelle de cette table à 10 % pour le VACUUM.

```
mvcc=# ALTER TABLE t5 SET (autovacuum_vacuum_scale_factor=0.1);
ALTER TABLE
```

- Modifiez les 200 000 lignes suivantes de cette table ?

```
mvcc=# UPDATE t5 SET id=2000000 WHERE id<=400060;
UPDATE 200000
```

- Attendez une minute.

```
mvcc=# SELECT pg_sleep(60);
```

- Que contient la vue *pg_stat_user_tables* pour cette table ? Qu'en déduisez-vous ?

```
mvcc=# SELECT * FROM pg_stat_user_tables WHERE relname='t5';
-[ RECORD 1 ]-------+------------------------------
```

```
relid                | 18745
schemaname           | public
relname              | t5
seq_scan             | 3
seq_tup_read         | 3000000
idx_scan             |
idx_tup_fetch        |
n_tup_ins            | 1000000
n_tup_upd            | 400060
n_tup_del            | 0
n_tup_hot_upd        | 59
n_live_tup           | 1000000
n_dead_tup           | 0
n_mod_since_analyze  | 0
last_vacuum          |
last_autovacuum      | 2021-11-15 22:40:17.799498+01
last_analyze         |
last_autoanalyze     | 2021-11-15 22:40:18.464116+01
vacuum_count         | 0
autovacuum_count     | 2
analyze_count        | 0
autoanalyze_count    | 3
```

Avec un facteur d'échelle à 10%, il ne faut plus attendre que la modification de 100050 lignes.

#### Verrous

- Ouvrez une transaction et lisez la table t1.

```
mvcc=# BEGIN;
BEGIN
mvcc=# SELECT * FROM t1;
 c1 |   c2
----+--------
  1 | un
  2 | deux
  3 | TROIS
  4 | QUATRE
  5 | CINQ
(5 rows)
```

- Ouvrez une autre transaction, et tentez de supprimer la table t1.

```
$ psql mvcc
```

```
psql (11)
Type "help" for help.

mvcc=# DROP TABLE t1;
```

- Listez les processus du serveur PostgreSQL. Que remarquez-vous ?

```
$ ps -o pid,cmd fx
23138 /usr/lib/postgresql/11/bin/postgres -D /var/lib/postgresql/11/main
 ↪  (...)
23140  \_ postgres: 11/main: checkpointer
23141  \_ postgres: 11/main: background writer
23142  \_ postgres: 11/main: walwriter
23143  \_ postgres: 11/main: autovacuum launcher
23145  \_ postgres: 11/main: stats collector
23146  \_ postgres: 11/main: logical replication launcher
 8538  \_ postgres: 11/main: postgres mvcc [local] idle in transaction
 9320  \_ postgres: 11/main: postgres mvcc [local] DROP TABLE waiting
```

La ligne intéressante est la ligne du DROP TABLE. Elle contient le mot clé `waiting`. Ce dernier indique que l'exécution de la requête est en attente d'un verrou sur un objet.

- Récupérez la liste des sessions en attente d'un verrou avec la vue *pg_stat_activity*.

```
-[ RECORD 1 ]----+-----------------------------------------------------------
 ↪  ----------
datid            | 16385
datname          | postgres
pid              | 9320
usesysid         | 16384
usename          | postgres
application_name | psql
client_addr      |
client_hostname  |
client_port      | -1
backend_start    | 2021-11-15 22:41:55.747066+01
xact_start       | 2021-11-15 22:42:00.068183+01
query_start      | 2021-11-15 22:42:00.068183+01
state_change     | 2021-11-15 22:42:00.068185+01
wait_event_type  | Lock
wait_event       | relation
state            | active
backend_xid      | 2760709
backend_xmin     | 2760709
```

```
query            | DROP TABLE t1;
backend_type     | client backend
-[ RECORD 2 ]----+-----------------------------------------------------------
 ↪  ----------
datid            | 16385
datname          | postgres
pid              | 8538
usesysid         | 16384
usename          | postgres
application_name | psql
client_addr      |
client_hostname  |
client_port      | -1
backend_start    | 2021-11-15 22:03:40.831073+01
xact_start       | 2021-11-15 22:41:47.548844+01
query_start      | 2021-11-15 22:41:51.420636+01
state_change     | 2021-11-15 22:41:51.421359+01
wait_event_type  | Client
wait_event       | ClientRead
state            | idle in transaction
backend_xid      |
backend_xmin     |
query            | SELECT * FROM t1;
backend_type     | client backend
```

- Récupérez la liste des verrous en attente pour la requête bloquée.

```
mvcc=# SELECT * FROM pg_locks WHERE pid=9320 AND NOT granted;
-[ RECORD 1 ]------+--------------------
locktype           | relation
database           | 16385
relation           | 18682
page               |
tuple              |
virtualxid         |
transactionid      |
classid            |
objid              |
objsubid           |
virtualtransaction | 3/4804
pid                | 9320
mode               | AccessExclusiveLock
granted            | f
fastpath           | f
```

- Récupérez le nom de l'objet dont on n'arrive pas à récupérer le verrou.

```
mvcc=# SELECT relname FROM pg_class WHERE oid=18682;
-[ RECORD 1 ]
relname | t1
```

- Récupérez la liste des verrous sur cet objet. Quel processus a verrouillé la table t1 ?

```
mvcc=# SELECT * FROM pg_locks WHERE relation=18682;
-[ RECORD 1 ]------+--------------------
locktype           | relation
database           | 16385
relation           | 18682
page               |
tuple              |
virtualxid         |
transactionid      |
classid            |
objid              |
objsubid           |
virtualtransaction | 4/69
pid                | 8538
mode               | AccessShareLock
granted            | t
fastpath           | f
-[ RECORD 2 ]------+--------------------
locktype           | relation
database           | 16385
relation           | 18682
page               |
tuple              |
virtualxid         |
transactionid      |
classid            |
objid              |
objsubid           |
virtualtransaction | 3/4804
pid                | 9320
mode               | AccessExclusiveLock
granted            | f
fastpath           | f
```

Le processus de PID 8538 a un verrou sur t1. Ce processus avait été listé plus haut en attente du client (`idle in transaction`).

- Retrouvez les informations sur la session bloquante.

```
mvcc=# SELECT * FROM pg_stat_activity WHERE pid=8538;
-[ RECORD 1 ]----+------------------------------
datid            | 16385
datname          | postgres
pid              | 8538
usesysid         | 16384
usename          | postgres
application_name | psql
client_addr      |
client_hostname  |
client_port      | -1
backend_start    | 2021-11-15 22:03:40.831073+01
xact_start       | 2021-11-15 22:41:47.548844+01
query_start      | 2021-11-15 22:41:51.420636+01
state_change     | 2021-11-15 22:41:51.421359+01
wait_event_type  | Client
wait_event       | ClientRead
state            | idle in transaction
backend_xid      |
backend_xmin     |
query            | SELECT * FROM t1;
backend_type     | client backend
```

À partir de là, il est possible d'arrêter l'exécution de l'ordre DROP TABLE avec la fonction `pg_cancel_backend()` ou de déconnecter le processus en cours de transaction avec la fonction `pg_terminate_backend()`.