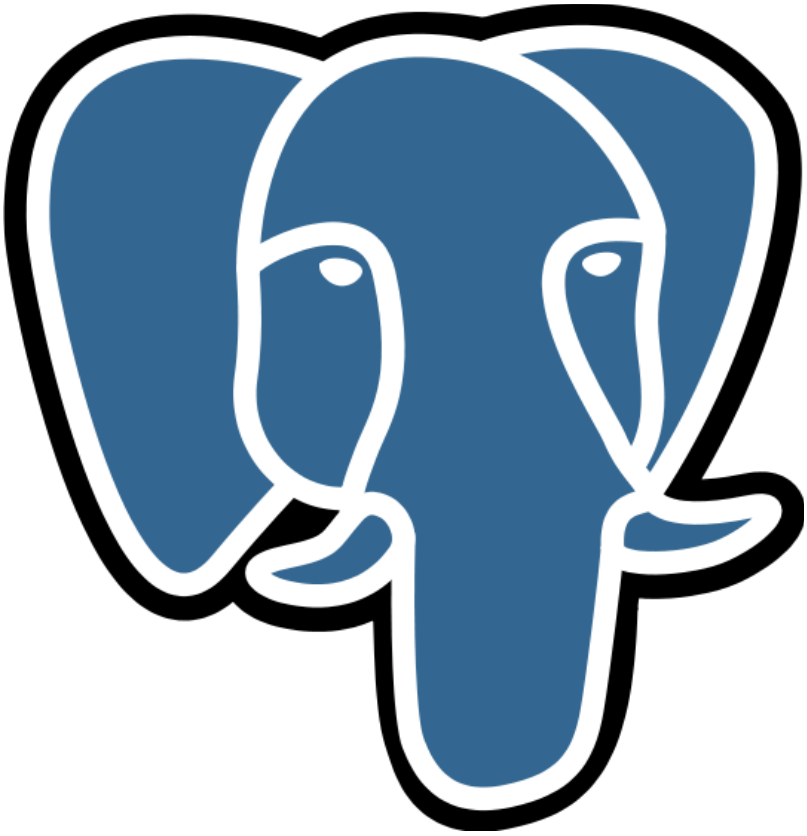
DALIBO

L'expertise PostgreSQL

Gestion d'un sinistre

Table des matières

PostgreSQL : Gestion d'un sinistre



1 Introduction

- une bonne politique de sauvegardes est cruciale
- mais elle n'empêche pas les incidents
- il faut être prêt à y faire face

Ce module se propose de faire une description des bonnes et mauvaises pratiques en cas de coup dur :

- crash de l'instance ;
- suppression / corruption de fichiers ;
- problèmes matériels ;
- sauvegardes corrompues...

Seront également présentées les situations classiques de désastres, ainsi que certaines méthodes et outils dangereux et déconseillés. L'objectif est d'aider à convaincre de l'intérêt qu'il y a à anticiper les problèmes, à mettre en place une politique de sauvegarde pérenne, et à ne pas tenter de manipulation dangereuse sans comprendre précisément à quoi l'on s'expose.

Ce module est en grande partie inspiré de la présentation suivante de Christophe Pettus : http://thebuild.com/presentations/worst-day-fosdem-2014.pdf

1.1 Au menu

- anticiper les désastres
- réagir aux désastres
- rechercher l'origine du problème
- outils utiles
- cas type de désastres

1.2 Licence Creative Commons CC-BY-NC-SA

Vous êtes libres de redistribuer et/ou modifier cette création selon les conditions suivantes :

- Paternité
- Pas d'utilisation commerciale
- Partage des conditions initiales à l'identique

Cette formation (diapositives, manuels et travaux pratiques) est sous licence **CC-BY-NC-SA**.

Vous êtes libres de redistribuer et/ou modifier cette création selon les conditions suivantes :

- Paternité
- Pas d'utilisation commerciale
- Partage des conditions initiales à l'identique

Vous devez citer le nom de l'auteur original de la manière indiquée par l'auteur de l'œuvre ou le titulaire des droits qui vous confère cette autorisation (mais pas d'une manière qui suggérerait qu'ils vous soutiennent ou approuvent votre utilisation de l'œuvre).

Vous n'avez pas le droit d'utiliser cette création à des fins commerciales.

Si vous modifiez, transformez ou adaptez cette création, vous n'avez le droit de distribuer la création qui en résulte que sous un contrat identique à celui-ci.

À chaque réutilisation ou distribution de cette création, vous devez faire apparaître clairement au public les conditions contractuelles de sa mise à disposition. La meilleure manière de les indiquer est un lien vers cette page web.

Chacune de ces conditions peut être levée si vous obtenez l'autorisation du titulaire des droits sur cette œuvre.

Rien dans ce contrat ne diminue ou ne restreint le droit moral de l'auteur ou des auteurs.

Le texte complet de la licence est disponible à cette adresse:
http://creativecommons.org/licenses/by-nc-sa/2.0/fr/legalcode

2 Anticiper les désastres

- un désastre peut toujours survenir
- il faut savoir le détecter le plus tôt possible
- et s'être préparé à y répondre

Il est impossible de parer à tous les cas de désastres imaginables. Le matériel peut subir des pannes, une faille logicielle non connue peut être exploitée, une modification d'infrastructure ou de configuration peut avoir des conséquences imprévues à long terme, une erreur humaine est toujours possible. Les principes de base de la haute disponibilité (redondance, surveillance...) permettent de mitiger le problème, mais jamais de l'éliminer complètement.

Il est donc extrêmement important de se préparer au mieux, de procéder à des simulations, de remettre en question chaque brique de l'infrastructure pour être capable de détecter une défaillance et d'y réagir rapidement.

2.1 Documentation

- documentation complète et à jour
  - emplacement et fréquence des sauvegardes
  - emplacement des traces
  - procédures et scripts d'exploitation
- sauvegarder et versionner la documentation

Par nature, les désastres arrivent de façon inattendue. Il faut donc se préparer à devoir agir en urgence, sans préparation, dans un environnement perturbé et stressant - par exemple, en pleine nuit la veille d'un jour particulièrement critique pour l'activité de la production. Un des premiers points d'importance est donc de s'assurer de la présence d'une documentation claire, précise et à jour, afin de minimiser le risque d'erreurs humaines.

Cette documentation devrait détailler l'architecture dans son ensemble, et particulièrement la politique de sauvegarde choisie, l'emplacement de celles-ci, les procédures de restauration et éventuellement de bascule vers un environnement de secours. Les procédures d'exploitation doivent y être expliquées, de façon détaillée mais claire, afin qu'il n'y ai pas de doute sur les actions à effectuer une fois la cause du problème identifié.

La méthode d'accès aux informations utiles (traces de l'instance, du système, supervision...) devraient également être soigneuseuement documentée afin que le diagnostic du problème soit aussi simple que possible.

Toutes ces informations doivent être organisées de façon claire, afin qu'elles soient immédiatement accessibles et exploitables aux intervenants lors d'un problème. Il est évidemment tout aussi important de penser à versionner et sauvegarder cette documentation, afin que celle-ci soit toujours accessible même en cas de désastre majeur (perte d'un site).

2.2 Procédures et scripts

- procédures détaillées de restauration / PRA
  - préparer des scripts / utiliser des outils
  - minimiser le nombre d'actions manuelles
- tester les procédures régulièrement
  - s'assurer que chacun les maîtrise
- sauvegarder et versionner les scripts

La gestion d'un désastre est une situtation particulièrement stressante, le risque d'erreur humaine est donc accru. Un DBA devant restaurer d'urgence l'instance de production en pleine nuit courera plus de risques de faire une fausse manipulation s'il doit taper une vingtaine de commandes en suivant une procédure dans une autre fenêtre (voire un autre poste) que s'il n'a qu'un script à exécuter.

En conséquence, il est important de minimiser le nombre d'actions manuelles à effectuer dans les procédures, en privilégiant l'usage de scripts d'exploitation ou d'outils dédiés (comme “pitrery” ou “barman” pour restaurer une instance PostgreSQL).

Néanmoins, même cette pratique ne suffit pas à exclure tout risque. L'usage de ces scripts ou de ces outils doit également être correctement documenté, et les procédures régulièrement testées. Dans le cas contraire, l'utilisation d'un script ou d'un outil peut aggraver le problème, parfois de façon dramatique - par exemple, l'écrasement d'un environnement sain lors d'une restauration parce que la procédure ne mentionne pas que le script doit être lancé depuis un serveur particulier.

L'aspect le plus important est de s'assurer par des tests réguliers et manuels que les procédures sont à jour, n'ont pas de comportement inattendu, et sont maîtrisées par toute l'équipe d'exploitation.

Tout comme pour la documentation, les scripts d'exploitation doivent également être sauvegardés et versionnés.

2.3 Supervision et historisation

- tout doit être supervisé
  - réseau, matériel, système, logiciels…
  - les niveaux d'alerte doivent être significatifs
- les métriques importantes doivent être historisées
  - cela permet de retrouver le moment où le problème est apparu
  - quand cela a un sens, faire des graphes

La supervision est un sujet vaste, qui touche plus au domaine de la haute disponibilité.

Un désastre sera d'autant plus difficile à gérer qu'il est détecté tard, la supervision en place doit donc être pensée pour détecter tout type de défaillance (penser également à superviser la supervision !). Attention à bien calibrer les niveaux d'alerte, la présence de trop de messages augmente le risque que l'un d'eux passe inapperçu, et donc que l'incident ne soit détecté que tardivement.

Pour aider la phase de diagnostic de l'origine du problème, il faut prévoir d'historiser un maximum d'informations. La présentation de celles-ci est également importante : il est plus facile de distinguer un pic brutal du nombre de connexions sur un graphique que dans un fichier de traces de plusieurs Go !

2.4 Automatisation

- des outils existent (Pacemaker, repmgr…)
- automatiser une bascule est complexe
- cela peut mener à davantage d'incidents
  - voire à des désastres importants (*split brain*)

Si l'on poursuit jusqu'au bout le raisonnement précédent sur le risque à faire effectuer de nombreuses opérations manuelles lors d'un incident, la conclusion logique est que la solution idéale serait de les éliminer complètement, et d'automatiser complètement le déclenchement et l'exécution de la procédure.

Le problème est que toute solution visant à automatiser une tâche se base sur un nombre limité de paramètres et sur une vision restreinte de l'architecture. Cette vision peut être erronée : typiquement, il est difficile à un outil de bascule automatique de diagnostiquer correctement certains types d'incident,

par exemple une partition réseau. L'outil peut donc détecter à tort à un incident, surtout s'il est réglé de façon à être assez sensible, et ainsi provoquer lui-même une coupure de service inutile. Dans le pire des cas, l'outil peut être amené à prendre une mauvaise décision amenant à une situation de désastre, comme un *split brain* (deux instances PostgreSQL se retrouvent ouvertes en écriture en même temps sur les mêmes données).

Il est donc fortement préférable de laisser un administrateur prendre les décisions potentiellement dangereuses, comme une bascule ou une restauration.

3 Réagir aux désastres

- savoir identifier un problème majeur
- bons réflexes
- mauvais réflexes

En dépit de toutes les précautions que l'on peut être amené à prendre, rien ne peut garantir qu'aucun problème ne surviendra. Il faut donc être capable d'identifier le problème lorsqu'il survient, et être prêt à répondre.

3.1 Symptômes d'un désastre

- crash de l'instance
- résultats de requêtes erronnés
- messages d'erreurs dans les traces
- dégradation importante des temps d'exécution
- processus manquants
  - ou en court d'exécution depuis une durée anormale

De très nombreux éléments peuvent aider à identifier que l'on est en situation d'incident grave. Le plus flagrant est évidemment le crash complet de l'instance PostgreSQL, ou du serveur l'hébergeant, et l'impossibilité pour PostgreSQL de redémarrer.

Les désastres les plus importants sont toutefois pas toujours aussi simples à détecter. Les crash peuvent se produire uniquement de façon ponctuelle, et il existe des cas où l'instance redémarre immédiatement ensuite (typiquement suite au `kill -9` d'un processus backend PostgreSQL). Cas encore plus délicat, il peut également arriver que les résultats de requêtes soient erronés (par exemple en cas de corruption de fichiers d'index) sans qu'aucune erreur apparaisse.

Les symptômes classiques permettant de détecter un problème majeur sont :

- la présence de messages d'erreurs dans les traces de PostgreSQL (notamment des messages `PANIC` ou `FATAL`, mais les messages `ERROR` et `WARNING` sont également très significatifs, particulièrement s'ils apparaissent soudainement en très grand nombre) ;

- la présence de messages d'erreurs dans les traces du système d'exploitation (notamment concernant la mémoire ou le système de stockage) ;
- le constat d'une dégration importante des temps d'exécution des requêtes sur l'instance ;
- l'absence de certains processus critiques de PostgreSQL ;
- la présence de processus présents depuis une durée inhabituelle (plusieurs semaines, mois, ...).

3.2 Bons réflexes 1

- garder la tête froide
- répartir les tâches clairement
- minimiser les canaux de communication
- garder des notes de chaque action entreprise

Une fois que l'incident est repéré, il est important de ne pas foncer tête baissée dans des manipulations. Il faut bien sûr prendre en considération la criticité du problème, notamment pour définir la priorité des actions (par exemple, en cas de perte totale d'un site, quels sont les applications à basculer en priorité ?), mais quelle que soit la criticité ou l'impact il ne faut jamais effectuer une action sans en avoir parfaitement saisi l'impact et s'être assuré qu'elle répondait bien au problème rencontré.

Si le travail s'effectue en équipe, il faut bien faire attention à répartir les tâches clairement, afin d'éviter des manipulations concurrentes ou des oublis qui pourraient aggraver la situation. Il faut également éviter de multiplier les canaux de communication, cela risque de favoriser la perte d'information, ce qui être critique dans une situation de crise.

Surtout, une règle majeure est de prendre le temps de systématiquement noter toutes les actions entreprises. Les commandes passées, les options utilisées, l'heure d'exécution, toutes ces informations sont très importantes, déjà pour pouvoir agir efficacement en cas de fausse manipulation, mais également pour documenter la gestion de l'incident après coup, et ainsi en conserver une trace qui sera précieuse si celui-ci venait à se reproduire.

3.3 Bons réflexes 2

- se prémunir contre une aggravation du problème
  - couper les accès applicatifs
- si une corruption est suspectée :
  - arrêter immédiatement l'instance
  - faire une sauvegarde immédiate
  - travailler sur une copie

S'il y a suspicion de potentielle corruption de données, il est primordial de s'assurer au plus vite de couper tous les accès applicatifs vers l'instance afin de ne pas aggraver la situation. Il est généralement préférable d'avoir une coupure de service plutôt qu'un grand volume de données irrécupérables.

Ensuite, il faut impérativement faire une sauvegarde complète de l'instance avant de procéder à toute manipulation. En fonction de la nature du problème rencontré le type de sauvegarde pouvant être effectué peut varier (un export de données ne sera possible que si l'instance est démarrée et que les fichiers sont lisibles par exemple). En cas de doute, la sauvegarde la plus fiable qu'il est possible d'effectuer est une copie des fichiers à froid (instance arrêtée) - toute autre action (y compris un export de données) pourrait avoir des conséquences indésirables.

Si des manipulations doivent être tentées pour tenter de récupérer des données, il faut impérativement travailler sur une copie de l'instance, restaurée à partir de cette sauvegarde. Ne jamais travailler directement sur une instance de production corrompue, la moindre action (même en lecture) pourrait aggraver le problème !

Voir aussi : https://wiki.postgresql.org/wiki/Corruption

3.4 Bons réflexes 3

- déterminer le moment de démarrage du désastre
- adopter une vision générale plutôt que focalisée sur un détail
- remettre en cause chaque élément de l'architecture
  - aussi stable (et/ou coûteux/complexe) soit-il
- éliminer en priorité les causes possibles côté hardware, système
- isoler le comportement précis du problème
  - identifier les requêtes / tables / index impliqués

La première chose à identifier est l'instant précis où le problème a commencé à se manifester. Cette information est en effet déterminante pour identifier la cause du problème, et le résoudre - notamment pour savoir à quel instant il faut restaurer l'instance si cela est nécessaire.

Il convient pour cela d'utiliser les outils de supervision et de traces (système, applicatif et PostgreSQL) pour remonter au moment d'apparition des premiers symptômes. Attention toutefois à ne pas confondre les symptômes avec le problème lui-même ! Les symptômes les plus visibles ne sont pas forcément apparus les premiers. Par exemple, la charge sur la machine est un symptôme, mais n'est jamais la cause du problème, elle est liée à d'autres phénomènes, comme des problèmes avec les disques ou un grand nombre de connexions, qui peuvent avoir commencé à se manifester bien avant que la charge commence réellement à augmenter.

Si la nature du problème n'est pas évidente à ce stade, il faut examiner l'ensemble de l'architecture en cause, sans en exclure d'office certains composants (baie de stockage, progiciel...), quels que soient leur complexité / coût / stabilité supposés. Si le comportement observé côté PostgreSQL est difficile à expliquer (crash plus ou moins aléatoires, nombreux messages d'erreur sans lien apparent...), il est préférable de commencer par s'assurer qu'il n'y a pas un problème de plus grande ampleur (système de stockage, virtualisation, réseau, système d'exploitation). Un bon indicateur consiste à regarder si d'autres instances / applications / processus rencontrent des problèmes similaires.

Ensuite, une fois que l'ampleur du problème a été cernée, il faut procéder méthodiquement pour en déterminer la cause et les éléments affectés. Pour cela, les informations les plus utiles se trouvent dans les traces, généralement de PostgreSQL ou du système, qui vont permettre d'identifier précisément les éventuels fichiers ou relations corrompus.

3.5 Bons réflexes 4

- en cas de défaillance matérielle, s'assurer de travailler sur du hardware sain et non affecté !!!

Cette recommandation peut paraître aller de soi, mais si les problèmes sont provoqués par une défaillance matérielle, il est impératif de s'assurer que le travail de correction soit effectué sur un environnement non affecté. Cela peut s'avérer problématique dans le cadre d'architecture mutualisant les ressources, comme des environnements virtualisés ou utilisant une baie de stockage.

Prendre également la précaution de vérifier que l'intégrité des sauvegardes n'est pas affectée par le problème.

3.6 Bons réflexes 5

- communiquer, ne pas rester isolé
- demander de l'aide si le problème est trop complexe
  - autres équipes
  - support
  - forums
  - listes

La communication est très importante dans la gestion d'un désastre. Il est préférable de minimiser le nombre de canaux de communication plutôt que de les multiplier (téléphone, e-mail, chat, ticket...) ce qui pourrait amener à une perte d'information, et à des délais indésirables.

Il est primordial de rapidement cerner l'ampleur du problème, et pour cela il est généralement nécessaire de demander l'expertise d'autres administrateurs / équipes (applicatif, système, réseau, virtualisation, SAN...). Il ne faut pas rester isolé et risquer que la vision étroite que l'on a des symptômes (notamment en terme de supervision / accès aux traces) empêche l'identification de la nature réelle du problème.

Si la situation semble échapper à tout contrôle, et dépasser les compétences de l'équipe en cours d'intervention, il faut chercher de l'aide auprès de personnes compétentes, par exemple auprès d'autres équipes, du support. En aucun cas il ne faut se mettre à suivre des recommandations glanées sur Internet, qui ne se rapporteraient que très approximativement au problème rencontré, voire pas du tout. Si nécessaire, on trouve en ligne des forums et des listes de discussions spécialisées sur lesquels il est également possible d'obtenir des conseils - il est néanmoins indispensable de prendre en compte que les personnes intervenant sur ces media le font de manière bénévole. Il est déraisonnable de s'attendre à une réaction immédiate, aussi urgent le problème soit-il, et les suggestions effectuées le sont sans aucune garantie.

3.7 Bons réflexes 6

- dérouler les procédures comme prévu
- en cas de situation non prévue, s'arrêter pour faire le point
  - ne pas hésiter à remettre en cause l'analyse
  - ou la procédure elle-même

Dans l'idéal, des procédures détaillant les actions à effectuer ont été écrites pour le cas de figure rencontré. Dans ce cas, une fois que l'on s'est assuré d'avoir identifié la procédure appropriée, il faut la dérouler méthodiquement, point par point, et valider à chaque étape que tout se déroule comme prévu.

Si une étape de la procédure ne se passe pas comme prévu, il ne faut pas tenter de poursuivre tout de même son exécution sans avoir compris ce qui s'est passé et les conséquences, cela pourrait être dangereux. Il faut au contraire prendre le temps de comprendre le problème en procédant comme décrit précédemment, quitte à remettre en cause toute l'analyse menée aupravant, et la procédure ou les scripts utilisés.

C'est également pour parer à ce type de cas de figure qu'il est important de travailler sur une copie et non sur l'environnement de production directement.

3.8 Bons réflexes 7

- en cas de bug avéré
  - tenter de le cerner et de le reproduire au mieux
  - le signaler à la communauté de préférence en détaillant le moyen de le reproduire

Ce n'est heureusement pas fréquent, mais il est possible que l'origine du problème soit liée à un bug de PostgreSQL lui-même. Dans ce cas, la méthodologie appropriée consiste à essayer de reproduire le problème le plus fidèlement possible et de façon systématique, pour le cerner au mieux.

Il est ensuite très important de le signaler au plus vite à la communauté, généralement sur la liste pgsql-bugs@postgresql.org (nécessite une inscription), en respectant les règles définies dans la documentation : http://www.postgresql.org/docs/current/static/bug-reporting.html Notamment (liste non exhaustive) :

- indiquer la version précise de PostgreSQL installée, et la méthode d'installation utilisée ;
- préciser la plate-forme utilisée, notamment la version du système d'exploitation utilisé et la configuration des ressources du serveur ;
- signaler uniquement les faits observés, éviter les spéculations sur l'origine du problème ;
- joindre le détail des messages d'erreurs observés (augmenter la verbosité des erreurs avec le paramètre `log_error_verbosity`) ;
- joindre un cas complet permettant de reproduire le problème de façon aussi simple que possible.

Pour les problèmes relevant du domaine de la sécurité (découverte d'une

faille), utiliser la liste security@postgresql.org

3.9 Bons réflexes 8

- tester complètement l'intégrité des données
  - pour détecter tous les problèmes
  - pour valider après restauration / correction

Un fois les actions correctives réalisées (restauration, recréation d'objets, mise à jour des données...), il faut tester intensivement pour s'assurer que le problème est bien complètement résolu. Il est donc extrêmement important d'avoir préparé des cas de tests permettant de reproduire le problème de façon certaine, afin de valider la solution appliquée.

En cas de corruption de données, il est également important de tenter de procéder à la lecture de la totalité des données depuis PostgreSQL. La commande suivante, exécutée avec l'utilisateur système propriétaire de l'instance (généralement `postgres`) effectue une lecture complète de toutes les relations :

```
pg_dumpall > /dev/null
```

Cette commande ne devrait renvoyer aucune erreur.

Même si la lecture des données ne renvoie aucune erreur, il est toujours possible que des problèmes subsistent, par exemple des corruptions silencieuses, des index incohérents avec les données... Dans les situations les plus extrêmes (problème de stockage, fichiers corrompus), il est important de tester la validité des données dans une nouvelle instance en effectuant un export/import complet des données. Par exemple, initialiser une nouvelle instance avec `initdb`, sur un autre système de stockage, voire sur un autre serveur, puis lancer la commande suivante (l'application doit-être coupée, ce qui est normalement le cas depuis la détection de l'incident si les conseils précédents ont été suivis) pour exporter et importer à la volée :

```
pg_dumpall -h <serveur_corrompu> -U postgres | psql -h <nouveau_serveur> -U
postgres postgres
vacuumdb -z -h <nouveau_serveur> -U postgres postgres
```

D'éventuels problèmes peuvent être détectés lors de l'import des données, il faut alors procéder au cas par cas. Enfin, même si cette étape s'est déroulée sans erreur, tout risque n'est pas écarté, il reste la possibilité de corruption de données silencieuses. Sauf si la fonctionnalité de checksum de PostgreSQL a été activée sur l'instance, le seul moyen de détecter ce type de problème est

de valider les données fonctionnellement.

Dans tous les cas, en cas de suspicion de corruption de données en profondeur, il est fortement préférable d'accepter une perte de données et de restaurer avant le début de l'incident, plutôt que de continuer à travailler avec des données dont l'intégrité n'est pas assurée.

3.10 Mauvais réflexes 1

- paniquer
- prendre une décision hâtive
  - exemple : supprimer des fichiers du répertoire `pg_xlog`
- lancer une commande sans la comprendre
  - exemple : `pg_resetxlog`

Quelle que soit la criticité du problème rencontré, la panique peut en faire quelque chose de pire. Il faut impérativement garder son calme, et résister au mieux au stress et aux pressions qu'une situation de désastre ne manque pas de provoquer. Il est également préférable d'éviter de sauter immédiatement à la conclusion la plus évidente. Il ne faut pas hésiter à retirer les mains du clavier pour prendre de la distance par rapport aux conséquences du problème, réfléchir aux causes possibles, prendre le temps d'aller chercher de l'information pour réévaluer l'ampleur réelle du problème.

La plus mauvaise décision que l'on peut être amenée à prendre lors de la gestion d'un incident est celle que l'on prend dans la précipitation, sans avoir bien réfléchi et mesuré son impact. Cela peut provoquer des dégâts irrécupérables, et transformer une situation d'incident en situation de crise majeure. Un exemple classique de ce type de comportement est le cas où PostgreSQL est arrêté suite au remplissage du système de fichiers contenant les fichiers WAL, `pg_xlog`. Le réflexe immédiat d'un administrateur non averti pourrait être de supprimer les plus vieux fichiers dans ce répertoire, ce qui répond bien aux symptômes observés mais reste une erreur dramatique qui va rendre le démarrage de l'instance impossible.

Quoi qu'il arrive, ne jamais exécuter une commande sans être certain qu'elle correspond bien à la situation rencontrée, et sans en maîtriser complètement les impacts. Même si cette commande provient d'un document mentionnant les mêmes messages d'erreur que ceux rencontrés (et tout particulièrement si le document a été trouvé via une recherche hâtive sur Internet) ! Là encore, nous disposons comme exemple d'une erreur malheureusement fréquente, l'exécution de la commande `pg_resetxlog` sur une instance rencontrant un problème. Comme l'indique la documentation, *« [cette commande] ne doit être utilisée qu'en dernier ressort quand le serveur ne démarre plus du fait d'une telle corruption »* et *« il ne faut pas perdre de vue que la base de*

données peut contenir des données inconsistantes du fait de transactions partiellement validées » (http://docs.postgresqlfr.org/current/app-pgresetxlog.html). Nous reviendrons ultérieurement sur les (rares) cas d'usage réels de cette commande, mais dans l'immense majorité des cas, l'utiliser va aggraver le problème, en ajoutant des problématiques de corruption logique des données ! Il convient donc de bien s'assurer de comprendre les conséquence de l'exécution de chaque action effectuée.

3.11 Mauvais réflexes 2

- arrêter le diagnostic quand les symptômes disparaissent
- ne pas pousser l'analyse jusqu'au bout

Il est important de pousser la réflexion jusqu'à avoir complètement compris l'origine du problème et ses conséquences.

En premier lieu, même si les symptômes semblent avoir disparus, il est tout à fait possible que le problème soit toujours sous-jacent, ou qu'il ait eu des conséquences moins visibles mais tout aussi graves (par exemple, une corruption logique de données). Ensuite, même si le problème est effectivement corrigé, prendre le temps de comprendre et de documenter l'origine du problème a une valeur inestimable pour prendre les mesures afin d'éviter que le problème ne se reproduise, et retrouver rapidement les informations utiles s'il venait à se reproduire tout de même.

3.12 Mauvais réflexes 3

- ne pas documenter
  - le résultat de l'investigation
  - les actions effectuées

Après s'être assuré d'avoir bien compris le problème recontré, il est tout aussi important de le documenter soigneusement, avec les actions de diagnostic et de correction effectuées. Ne pas le faire, c'est perdre un excellente occasion de gagner un temps précieux si le problème venait à se produire de nouveau. C'est également un risque supplémentaire dans le cas où les actions correctives menées n'auraient pas suffit à complètement corriger le problème ou auraient eu un effet de bord inattendu. Dans ce cas, avoir pris le temps de noter le détail des actions effectuées fera là encore gager un temps précieux.

4 Rechercher l'origine du problème

- quelques pistes de recherche pour cerner le problème
- liste non exhaustive

Les problèmes pouvant survenir sont trop nombreux pour pouvoir tous les lister, chaque élément matériel ou logiciel d'une architecture pouvant subir de nombreux types de défaillances. Cette section liste quelques pistes classiques d'investigation à ne pas négliger pour s'efforcer de cerner au mieux l'étendue du problème, et en déterminer les conséquences.

4.1 Prérequis

- avant de commencer à creuser
  - référencer les symptômes
  - identifier au mieux l'instant de démarrage du problème

La première étape est de déterminer aussi précisément que possible quels sont les symptômes observés, sans en négliger, et à partir de quel moment ils sont apparus. Cela donne des informations précieuses sur l'étendue du problème, et évite de se focaliser sur un symptôme particulier, parce que plus visible (par exemple l'arrêt brutal de l'instance), alors que la cause réelle est très antérieure (par exemple des erreurs IO dans les traces système, ou une montée progressive de la charge sur le serveur).

4.2 Recherche d'historique

- ces symptômes ont-ils déjà été rencontrés dans le passé ?
- ces symptômes ont-ils déjà été rencontrés par d'autres ?
- attention à ne pas prendre les informations trouvées pour argent comptant !

une fois les principaux symptômes identifiés, il est utile de prendre un moment pour déterminer si ce problème est déjà connu. Notamment, identifier dans la

base de connaissances si ces symptômes ont déjà été rencontrés dans le passé (d'où l'importance de documenter bien les problèmes).

Au delà de la documentation interne, il est également possible de rechercher si ces symptômes ont déjà été rencontrés par d'autres. Pour ce type de recherche, il est préférable de privilégier les sources fiables (documentation officielle, listes de discussion, plate-forme de support...) plutôt qu'un quelconque document d'un auteur non identifié.

Dans tous les cas, il faut faire très attention à ne pas prendre les informations trouvées pour argent comptant, et ce même si elles proviennent de la documentation interne ou d'une source fiable ! Il est toujours possible que les symptômes soient similaires mais que la cause soit différente. Il s'agit donc ici de mettre en place une base de travail, qui doit être complétée par une observation directe et une analyse.

4.3 Matériel

- vérifier le système disque (SAN, carte RAID, disques)
- rechercher toute erreur matérielle
- firmwares pas à jour
  - ou récemment mis à jour
- matériel récemment changé

Les défaillances du matériel, et notamment du système de stockage, sont de celles qui peuvent avoir les impacts les plus importants et les plus étendus sur une instance et les données qu'elle contient. Ce type de problème peut également être difficile à diagnostiquer en se contentant d'observer les symptômes les plus visibles, il est facile de sous-estimer l'ampleur des dégâts.

Parmi les bonnes pratiques, il convient de vérifier la configuration et l'état du système disque (SAN, carte RAID, disques). Quelques éléments étant source habituelle de problème :

- le système disque n'honore pas fsync ? (SAN ? virtualisation ?)
- quel est l'état de la batterie du cache en écriture ?

Il faut évidamment rechercher la présence de toute erreur matérielle, au niveau des disques, de la mémoire, des CPU...

Vérifier également la version des firmwares installés. Il est possible qu'une nouvelle version corrige le problème rencontré, ou à l'inverse que le déploiement d'une nouvelle version soit à l'origine du problème.

Dans le même esprit, il faut vérifier si du matériel a récemment été changé, il arrive que de nouveaux éléments soient défaillants.

Il convient de noter que l'investigation à ce niveau peut être grandement

complexifiée par l'utilisation de certaines technologies (virtualisation, baies de stockage), du fait de la mutualisation des ressources, et de la séparation des compétences et des informations de supervision entre différentes équipes.

4.4 Virtualisation

- problèmes de mutualisation des ressources
- configuration du stockage virtualisé
- rechercher toute erreur sur l'hôte / la console d'administration
- mises à jour non appliquées
  - ou appliquées récemment
- modifications de configuration récentes

Tout comme pour les problèmes au niveau du matériel, les problèmes au niveau du système de virtualisation peuvent être complexes à détecter et à diagnostiquer correctement.

Le principal facteur de problème avec la virtualisation est lié à une mutualisation excessive des ressources. Il est ainsi possible d'avoir un total de ressources allouées aux VM supérieur à celles disponibles sur l'hyperviseur, ce qui amène à des comportements de fort ralentissement voire de blocage des systèmes virtualisés. Si ce type d'architecture est couplé à un système de gestion de bascule automatique (Pacemaker, repmgr...), il est possible d'avoir des situations de bascules impromptues, voire des situations de *split brain*, qui peuvent provoquer des pertes de données importantes. Il est donc important de prêter une attention particulière à l'utilisation des ressources de l'hyperviseur, et d'éviter à tout prix la sur-allocation.

Par ailleurs, lorsque l'architecture inclue une brique de virtualisation, il est important de prendre en compte que certains problèmes ne peuvent être observés qu'à partir de l'hyperviseur, et pas à partir du système virtualisé. Par exemple, les erreurs matérielles ou système risquent d'être invisibles depuis une VM, il convient donc d'être vigilant, et de rechercher toute erreur sur l'hôte.

Il faut également vérifier si des modifications ont été effectuées peu avant l'incident, comme des modifications de configuration ou l'application de mises à jour.

Comme indiqué dans la partie traitant du matériel, l'investigation peut être grandement freinée par la séparation des compétences et des informations de supervision entre différentes équipes. Une bonne communication est alors la clé de la résolution rapide du problème.

4.5 Système d'exploitation 1

- erreurs dans les traces
- mises à jour système non appliquées
- modifications de configuration récentes

Après avoir vérifié les couches matérielles et la virtualisation, il faut ensuite s'assurer de l'intégrité du système. La première des vérifications à effecuter est de consulter les traces su système pour en extraire les éventuels messages d'erreur :

- sous Linux, on trouvera ce type d'informations en sortie de la commande `dmesg`, et dans les fichiers traces du système, généralement situés sous `/var/log` ;
- sous Windows, on consultera à cet effet les event logs.

Tout comme pour les autres briques, il faut également voir s'il existe des mises à jour des paquets qui n'auraient pas été appliquées, ou à l'inverse si des mises à jour, installations ou modifications de configuration ont été effectuées récemment.

4.6 Système d'exploitation 2

- opération d'IO impossible
  - FS plein ?
  - FS monté en lecture seule ?
- tester l'écriture sur le PGDATA
- tester la lecture sur le PGDATA

Parmi les problèmes fréquemment rencontrés se trouve l'impossibilité pour PostgreSQL d'accéder en lecture ou en écriture à un ou plusieurs fichiers. La première chose à vérifier est de déterminer si le système de fichiers sous-jacent ne serait pas rempli à 100% (commande `df` sous Linux) ou monté en lecture seule (commande `mount` sous Linux).

Sinon, on peut tester les opérations d'écriture et de lecture sur le système de fichier pour déterminer si le comportement y est global :

- pour tester une écriture dans le répertoire PGDATA, sous Linux :

```
touch $PGDATA/test_write
```

- pour tester une lecture dans le répertoire PGDATA, sous Linux :

```
cat $PGDATA/PGVERSION
```

Pour identifier précisément quels sont les fichiers présentant des problèmes, il est possible de tester la lecture complète des fichiers dans le point de montage :

```
cp -R $PGDATA > /dev/null
```

4.7 Système d'exploitation 3

- consommation excessive des ressources
  - OOM killer
- après un crash, vérifier les processus actifs
  - ne pas tenter de redémarrer si des processus persistent

Sous Linux, l'installation d'outils d'aide au diagnostic sur les serveurs est très important pour mener une analyse efficace, particulièrement le paquet `systat` qui permet d'utiliser la commande `sar`. La lecture des traces système et des traces PostgreSQL permettent également d'avancer dans le diagnostic.

Un problème de consommation excessive des ressources peut généralement être anticipée grâce à une supervision sur l'utilisation des ressources et des seuils d'alerte appropriés. Il arrive néanmoins parfois que la consommation soit très rapide et qu'il ne soit pas possible de réagir suffisamment rapidement.

Dans le cas d'une consommation mémoire d'un serveur Linux qui menacerait de dépasser la quantité totale de mémoire allouable, le comportement par défaut de Linux est d'autoriser par défaut la tentative d'allocation. Si l'allocation dépasse effectivement la mémoire disponible, alors le système va déclencher un processus *Out Of Memory Killer* (OOM killer) qui va se charger de ture les processus les plus consommateurs. Dans le cas d'un serveur dédié à une instance PostgreSQL, il y a de grandes chances que le processus en question appartienne à l'instance. S'il s'agit d'un *OOM killer* effectuant un arrêt brutal (`kill -9`) sur un backend, l'instance PostgreSQL va arrêter immédiatement tous les processus afin de prévenir une corruption de la mémoire et les redémarrer. S'il s'agit du processus principal de l'instance (*postmaster*), les conséquences peuvent être bien plus dramatiques, surtout si une tentative est faite de redémarrer l'instance sans vérifier si des processus actifs existent encore.

Pour un serveur dédié à PostgreSQL, la recommendation est habituellement de

désactiver la sur-allocation de la mémoire, empêchant ainsi le déclenchement de ce phénomène. Voir pour cela les paramètres kernel `vm.overcommit_memory` et `vm.overcommit_ratio`.

4.8 PostgreSQL

- relever les erreurs dans les traces
  - ou messages inhabituels
- vérifier les mises à jour mineures

Tout comme pour l'analyse autour du système d'exploitation, la première chose à faire est rechercher toute erreur ou message inhabituel dans les traces de l'instance. Ces messages sont habituellement assez informatifs, et permettent de cerner la nature du problème (par exemple, si PostgreSQL ne parvient pas à écrire dans un fichier, il indiquera précisément de quel fichier il s'agit).

Si l'instance est arrêtée suite à un crash, et que les tentatives de redémarrage échouent avant qu'un message puisse être écrit dans les traces, il est possible de tenter de démarrer l'instance en exécutant directement le binaire `postgres` afin que les premiers messages soient envoyés vers la sortie standard.

Il convient également de vérifier si des mises à jour qui n'auraient pas été appliquées ne corrigeraient pas un problème similaire à celui rencontré. Identifier les mises à jours appliquées récemment et les modifications de configuration peut également aider à comprendre la nature du problème.

4.9 Paramétrage de PostgreSQL 1

- la désactivation de certains paramètres est dangereuse
- `fsync`
- `full_page_write`

Si des corruptions de données sont relevées suite à un crash de l'instance, il convient particulièrement de vérifier la valeur du paramètre `fsync`. En effect, si celui-ci est désactivé, les écritures dans les journaux de transactions ne sont pas effectuées de façon synchrone, ce qui implique que l'ordre des écritures ne sera pas conservé en cas de crash. Le processus de recovery de PostgreSQL risque alors de provoquer des corruptions si l'instance est malgré tout redémarrée. Ce paramètre ne devrait jamais être positionné à une autre valeur que on, sauf pour répondre à des cas extrêment particuliers (en bref, si

l'on peut se permettre de restaurer intégralement les données en cas de crash, par exemple si les tables ne sont utilisées que pour faire du chargement de données).

Le paramètre `full_page_write` indique à PostgreSQL d'effectuer une écriture complète d'une page chaque fois quelle recevera une nouvelle écriture après un checkpoint, pour éviter un éventuel mélanger entre des anciennes et nouvelles données en cas d'écriture partielle. La désactivation de ce paramètre peut avoir le même type de conséquences que la désactivation de `fsync`.

4.10 Paramétrage de PostgreSQL 2

- option `--data-checksums` de `initdb`
- depuis PostgreSQL 9.3
- détecte les corruptions silencieuses
- au prix d'un impact sur les performances

L'apparition des sommes de contrôles (checksums) permet de se prévenir contre des corruptions silencieuses de données. Pour s'en convaincre, voici un petit exemple.

Tout d'abord, créons un cluster sans utiliser les checksums, et un autre qui les utilisera. Attention, on ne peut modifier un cluster pour changer ce paramètre.

```
$ initdb -D /tmp/sans_checksums/
$ initdb -D /tmp/avec_checksums/ --data-checksums
```

Insérons une valeur de test, sur chacun des deux clusters:

```
postgres=# CREATE TABLE test (name text);
CREATE TABLE

postgres=# INSERT INTO test (name) VALUES ('toto');
INSERT 0 1
```

On récupère le chemin du fichier de la table pour aller le corrompre à la main (seul celui sans checksums est montré en exemple).

```
postgres=# SELECT pg_relation_filepath('test');
 pg_relation_filepath
----------------------
 base/12036/16317
```

```
(1 row)
```

Ensuite, on va s'attacher à corrompre ce fichier, en remplaçant la valeur toto avec un éditeur hexadécimal :

```
$ hexedit /tmp/sans_checksums/base/12036/16317
$ hexedit /tmp/avec_checksums/base/12036/16399
```

Enfin, on peut ensuite exécuter des requêtes sur ces deux clusters.

Sans checksums:

```
postgres=# TABLE test;
 name
------
 qoto
```

Avec checksums:

```
postgres=# TABLE test;
WARNING:  page verification failed, calculated checksum 16321 but expected
21348
ERROR:  invalid page in block 0 of relation base/12036/16387
```

Côté performances, on peut aussi réaliser un benchmark rapide. Celui-ci a été réalisé en utilisant pgbench, avec une échelle de 1000 (base de 16 Go), en utilisant les tests par défaut, avec la configuration de PostgreSQL laissée par défaut. En utilisant la moyenne de trois exécutions, on constate une dégradation de performations très importante, de l'ordre de 20% (226 transactions par seconde avec checksum contre 281 sans).

4.11 Erreur de manipulation

- traces système, traces PostgreSQL
- revue des dernières manipulations effectuées
- historique des commandes

L'erreur humaine fait également partie des principales causes de désastre. Une commande de suppression tapée trop rapidement, un oubli de clause `WHERE` dans une requête de mise à jour, nombreuses sont les opérations qui peuvent provoquer des pertes de données ou un crash de l'instance.

Il convient donc de revoir les dernières opérations effectuées sur le serveur, en commençant par les interventions planifiées, et si possible récupérer l'historique des commandes passées. Des exemples de commandes particulièrement dangereuses :

- `kill -9`
- `rm -rf`
- `rsync`
- `find` → souvent couplé avec des commandes destructices (`rm`, `mv`, `gzip`...)

5 Outils

- quelques outils peuvent aider
  - à diagnostiquer la nature du problème
  - à valider la correction apportée
  - à appliquer un contournement
- ATTENTION :
  - certains de ces outils peuvent corrompre les données !

5.1 Outils - pg_controldata

- fournit des informations de contrôle sur l'instance
- ne nécessite pas que l'instance soit démarrée

L'outil pg_controldata lit les informations du fichier de contrôle d'une instance PostgreSQL. Cet outil ne se connecte pas à l'instance, il a juste besoin d'avoir un accès en lecture sur le répertoire PGDATA de l'instance. Les informations qu'il récupère ne sont donc pas du temps réel, il s'agit d'une vision de l'instance telle qu'elle était la dernière fois que le fichier de contrôle a été mis à jour. L'avantage est qu'elle peut être utilisée même si l'instance est arrêtée.

pg_controldata affiche notamment les informations initialisées lors d'initdb, telles que la version du catalogue, ou la taille des blocs, qui peuvent être cruciales si l'on veut restaurer une instance sur un nouveau serveur à partir d'une copie des fichiers.

Il affiche également de nombreuses informations utiles sur le traitement des journaux de transactions et des checkpoints, par exemple :

- positions du dernier checkpoint et du précédent dans les WAL ;
- nom du WAL correspondant au dernier WAL ;
- timeline sur laquelle se situe le dernier checkpoint ;
- instant précis du dernier checkpoint.

Quelques informations de paramétrage sont également renvoyées, comme la configuration du niveau de WAL, ou le nombre maximal de connexions autorisé.

En complément, le dernier état connu de l'instance est également affiché. Les états potentiels sont :

- `in production` : l'instance est démarrée et est ouverte en écriture ;
- `shut down` : l'instance est arrêtée ;
- `in archive recovery` : l'instance est démarrée et est en mode `recovery` (*Warm* ou *Hot Standby*) ;
- `shut down in recovery` : l'instance est arrêtée et un fichier `recovery.conf` existe ;
- `shutting down` : état transitoire, l'instance est en cours d'arrêt ;
- `in crash recovery` : état transitoire, l'instance est en cours de démarrage suite à un crash ;
- `starting up` : état transitoire, concrètement jamais utilisé.

Bien entendu, comme ces informations ne sont pas mises à jour en temps réel, elles peuvent être erronées. Cet asynchronisme est intéressant pour diagnostiquer un problème, par exemple si pg_controldata renvoie l'état `in production` mais que l'instance est arrêtée, cela signifie que l'arrêt n'a pas été effectué proprement (*crash* de l'instance, qui sera donc suivi d'un *recovery* au démarrage).

Exemple de sortie de la commande :

```
-bash-4.2$ /usr/pgsql-9.4/bin/pg_controldata /var/lib/pgsql/9.4/data
pg_control version number:            942
Catalog version number:               201409291
Database system identifier:           6166553528265460044
Database cluster state:               in production
pg_control last modified:             Fri 30 Oct 2015 08:05:17 PM CET
Latest checkpoint location:           1/7E000028
Prior checkpoint location:            1/7D000028
Latest checkpoint's REDO location:    1/7E000028
Latest checkpoint's REDO WAL file:    0000000A000000010000007E
Latest checkpoint's TimeLineID:       10
Latest checkpoint's PrevTimeLineID:   10
Latest checkpoint's full_page_writes: on
Latest checkpoint's NextXID:          0/2180
Latest checkpoint's NextOID:          32806
Latest checkpoint's NextMultiXactId:  1
Latest checkpoint's NextMultiOffset:  0
Latest checkpoint's oldestXID:        1880
Latest checkpoint's oldestXID's DB:   1
Latest checkpoint's oldestActiveXID:  0
Latest checkpoint's oldestMultiXid:   1
Latest checkpoint's oldestMulti's DB: 1
Time of latest checkpoint:            Thu 29 Oct 2015 10:46:38 PM CET
Fake LSN counter for unlogged rels:   0/1
Minimum recovery ending location:     0/0
Min recovery ending loc's timeline:   0
Backup start location:                0/0
Backup end location:                  0/0
End-of-backup record required:        no
```

```
Current wal_level setting:            hot_standby
Current wal_log_hints setting:        off
Current max_connections setting:      100
Current max_worker_processes setting: 8
Current max_prepared_xacts setting:   0
Current max_locks_per_xact setting:   64
Maximum data alignment:               8
Database block size:                  8192
Blocks per segment of large relation: 131072
WAL block size:                       8192
Bytes per WAL segment:                16777216
Maximum length of identifiers:        64
Maximum columns in an index:          32
Maximum size of a TOAST chunk:        1996
Size of a large-object chunk:         2048
Date/time type storage:               64-bit integers
Float4 argument passing:              by value
Float8 argument passing:              by value
Data page checksum version:           0
```

5.2 Outils - export/import de données

- pg_dump
- pg_dumpall
- COPY
- psql / pg_restore

Les outils pg_dump et pg_dumpall permettent d'exporter des données à partir d'une instance démarrée. Dans le cadre d'un incident grave, il est possible de les utiliser pour :

- extraire le contenu de l'instance ;
- extraire le contenu des bases de données ;
- tester si les données sont lisibles dans un format compréhensibles par PostgreSQL.

Par exemple, un moyen rapide de s'assurer si tous les fichiers des tables de l'instance sont lisibles est de forcer leur lecture complète grâce à la commande suivante :

```
pg_dumpall > /dev/null
```

Attention, les fichiers associés aux index ne sont pas parcourus pendant cette opération. Par ailleurs, ne pas avoir d'erreur ne garantit en aucun cas pas l'intégrité fonctionnelle des données : les corruptions peuvent très bien être

silencieuses.

Si pg_dumpall ou pg_dump rencontrent des messages d'erreur et ne parviennent pas à exporter certaines tables, il est possible de contourner le problème à l'aide de la commande COPY, en sélectionnant exclusivement les données lisibles autour du bloc corrompu.

Il convient ensuite d'utiliser psql ou pg_restore pour importer les données dans une nouvelle instance, probablement sur un nouveau serveur, dans un environnement non affecté par le problème. Pour rappel, même après un export / import de données réalisé avec succès, des corruptions logiques peuvent encore être présentes. Il faut donc être particulièrement vigilant et prendre le temps de valider l'intégrité fonctionnelle des données.

5.3 Outils - pageinspect

- extension
- vision du contenu d'un bloc
- sans le dictionnaire, donc sans décodage des données
- affichage brut
- utilisé surtout en debug, ou dans les cas de corruption
- fonctions de décodage pour heap (table), bt (btree), entête de page, et FSM
- nécessite de connaître le code de PostgreSQL

Voici quelques exemples:

Contenu d'une page d'une table:

```
=# select * from heap_page_items(get_raw_page('dspam_token_data',0)) limit
5;
 lp | lp_off | lp_flags | lp_len |   t_xmin   | t_xmax | t_field3 | t_ctid |
t_infomask2 | t_infomask | t_hoff | t_bits | t_oid
----+--------+----------+--------+------------+--------+----------+--------
+-------------+------------+--------+--------+-------
  1 |    201 |        2 |      0 |            |        |          |        |
|             |            |        |
  2 |   1424 |        1 |     48 | 1439252980 |      0 |        0 | (0,2)  |
5 |       2304 |     24 |        |
  3 |    116 |        2 |      0 |            |        |          |        |
|             |            |        |
  4 |   7376 |        1 |     48 |          2 |      0 |      140 | (0,4)  |
5 |      10496 |     24 |        |
  5 |   3536 |        1 |     48 | 1392499801 |      0 |        0 | (0,5)  |
5 |       2304 |     24 |        |
```

Et son entête:

```
=# select * from page_header(get_raw_page('dspam_token_data',0));
     lsn      | checksum | flags | lower | upper | special | pagesize |
version | prune_xid
--------------+----------+-------+-------+-------+---------+----------
+---------+-----------
 F1A/5A6EAC40 |        0 |     1 |   852 |   896 |    8192 |     8192 |
4 | 1450780148
```

Méta-données d'un index (contenu dans la première page):

```
=# select * from bt_metap('dspam_token_data_uid_key');
 magic  | version | root | level | fastroot | fastlevel
--------+---------+------+-------+----------+-----------
 340322 |       2 |  243 |     2 |      243 |         2
```

La page racine est la 243. Allons la voir:

```
=# select * from bt_page_items('dspam_token_data_uid_key',243) limit 10;
 itemoffset |   ctid     | itemlen | nulls | vars |                 data
------------+------------+---------+-------+------
+-------------------------------------
          1 | (3,1)      |       8 | f     | f    |
          2 | (44565,1)  |      20 | f     | f    | f3 4b 2e 8c 39 a3 cb 80
0f 00 00 00
          3 | (242,1)    |      20 | f     | f    | 77 c6 0d 6f a6 92 db 81
28 00 00 00
          4 | (43569,1)  |      20 | f     | f    | 47 a6 aa be 29 e3 13 83
18 00 00 00
          5 | (481,1)    |      20 | f     | f    | 30 17 dd 8e d9 72 7d 84
0a 00 00 00
          6 | (43077,1)  |      20 | f     | f    | 5c 3c 7b c5 5b 7a 4e 85
0a 00 00 00
          7 | (719,1)    |      20 | f     | f    | 0d 91 d5 78 a9 72 88 86
26 00 00 00
          8 | (41209,1)  |      20 | f     | f    | a7 8a da 17 95 17 cd 87
0a 00 00 00
          9 | (957,1)    |      20 | f     | f    | 78 e9 64 e9 64 a9 52 89
26 00 00 00
         10 | (40849,1)  |      20 | f     | f    | 53 11 e9 64 e9 1b c3 8a
26 00 00 00
```

La première entrée de la page 243, correspondant à la donnée f3 4b 2e 8c 39 a3 cb 80 0f 00 00 00 est stockée dans la page 3 de notre index:

```
=# select * from bt_page_stats('dspam_token_data_uid_key',3);
 blkno | type | live_items | dead_items | avg_item_size | page_size |
free_size | btpo_prev | btpo_next | btpo | btpo_flags
```

```
-------+------+-----------+-----------+--------------+----------
+----------+----------+----------+------+------------
     3 | i    |       202 |         0 |           19 |      8192 |
3312 |        0 |     44565 |    1 |          0

=# select * from bt_page_items('dspam_token_data_uid_key',3) limit 10;
 itemoffset |   ctid    | itemlen | nulls | vars |                 data
------------+-----------+---------+-------+------
+-----------------------------------
          1 | (38065,1) |      20 | f     | f    | f3 4b 2e 8c 39 a3 cb 80
0f 00 00 00
          2 | (1,1)     |       8 | f     | f    |
          3 | (37361,1) |      20 | f     | f    | 30 fd 30 b8 70 c9 01 80
26 00 00 00
          4 | (2,1)     |      20 | f     | f    | 18 2c 37 36 27 03 03 80
27 00 00 00
          5 | (4,1)     |      20 | f     | f    | 36 61 f3 b6 c5 1b 03 80
0f 00 00 00
          6 | (43997,1) |      20 | f     | f    | 30 4a 32 58 c8 44 03 80
27 00 00 00
          7 | (5,1)     |      20 | f     | f    | 88 fe 97 6f 7e 5a 03 80
27 00 00 00
          8 | (51136,1) |      20 | f     | f    | 74 a8 5a 9b 15 5d 03 80
28 00 00 00
          9 | (6,1)     |      20 | f     | f    | 44 41 3c ee c8 fe 03 80
0a 00 00 00
         10 | (45317,1) |      20 | f     | f    | d4 b0 7c fd 5d 8d 05 80
26 00 00 00
```

Le type de la page est `i`, c'est à dire «internal», donc une page interne de l'arbre. Continuons notre descente, allons voir la page 38065:

```
=# select * from bt_page_stats('dspam_token_data_uid_key',38065);
 blkno | type | live_items | dead_items | avg_item_size | page_size |
free_size | btpo_prev | btpo_next | btpo | btpo_flags
-------+------+------------+------------+---------------+-----------
+-----------+----------+----------+------+------------
 38065 | l    |        169 |         21 |            20 |      8192 |
3588 |       118 |       119 |    0 |         65
(1 row)

=# select * from bt_page_items('dspam_token_data_uid_key',38065) limit 10;
 itemoffset |    ctid     | itemlen | nulls | vars |                 data
------------+-------------+---------+-------+------
+-----------------------------------
          1 | (11128,118) |      20 | f     | f    | 33 37 89 95 b9 23 cc 80
0a 00 00 00
          2 | (45713,181) |      20 | f     | f    | f3 4b 2e 8c 39 a3 cb 80
0f 00 00 00
          3 | (45424,97)  |      20 | f     | f    | f3 4b 2e 8c 39 a3 cb 80
26 00 00 00
          4 | (45255,28)  |      20 | f     | f    | f3 4b 2e 8c 39 a3 cb 80
27 00 00 00
          5 | (15672,172) |      20 | f     | f    | f3 4b 2e 8c 39 a3 cb 80
28 00 00 00
          6 | (5456,118)  |      20 | f     | f    | f3 bf 29 a2 39 a3 cb 80
```

```
0f 00 00 00
          7 | (8356,206)  |      20 | f      | f     | f3 bf 29 a2 39 a3 cb 80
28 00 00 00
          8 | (33895,272) |      20 | f      | f     | f3 4b 8e 37 99 a3 cb 80
0a 00 00 00
          9 | (5176,108)  |      20 | f      | f     | f3 4b 8e 37 99 a3 cb 80
0f 00 00 00
         10 | (5466,41)   |      20 | f      | f     | f3 4b 8e 37 99 a3 cb 80
26 00 00 00
```

Nous avons trouvé une feuille (type `l`). Les ctid pointés sont maintenant les adresses dans la table:

```
=# select * from dspam_token_data where ctid = '(11128,118)';
 uid |        token         | spam_hits | innocent_hits |  last_hit
-----+----------------------+-----------+---------------+------------
  40 | -6317261189288392210 |         0 |             3 | 2014-11-10
```

5.4 Outils - pg_resetxlog

- efface les WAL courants
- permet à l'instance de démarrer en cas de corruption d'un WAL
  - comme si elle était dans un état cohérent
  - ce qui n'est pas le cas
- cet outil est dangereux !!!
- utiliser cet outil va corrompre des données

pg_resetxlog est un outil fourni avec PostgreSQL. Il s'utilise manuellement, en ligne de commande. Sa fonctionnalité principale est d'effacer les fichiers WAL courants, et il se charge également de réinitialiser les informations du fichier de contrôle correspondantes. Il est possible de lui spécifier les valeurs à initialiser dans le fichier de contrôle si l'outil ne parvient pas à les déterminer (par exemple, si tous les WAL dans le répertoire `pg_xlog` ont été supprimés).

Attention, pg_resetxlog ne doit **jamais** être utilisé sur une instance démarrée. Avant d'exécuter l'outil, il faut toujours vérifier qu'il ne reste aucun processus de l'instance.

L'objectif de pg_resetxlog est de pouvoir démarrer une instance après un crash si des corruptions de fichiers (typiquement WAL ou fichier de contrôle) empêche ce démarrage. Cette action n'est pas une action de réparation ! La réinitialisation des journaux de transactions implique que des transactions qui

n'étaient que partiellement validées ne seront pas détectées comme telles, et ne seront donc pas annulées lors du *recovery*. La conséquence est que les données de l'instance ne sont plus cohérentes. Il est fort possible d'y trouver des violations de contraintes diverses (notamment clés étrangères), ou d'autres cas d'incohérences plus difficiles à détecter. Après la réinitialisation des WAL, une fois que l'instance a démarré, **il ne faut surtout pas ouvrir les accès à l'application** ! Comme indiqué, les données présentent sans aucun doute des incohérence, et toute action en écriture à ce point ne ferait qu'aggraver le problème. L'étape suivante est donc de faire un export immédiat des données, de les restaurer dans une nouvelle instance initialisée à cet effet (de préférence sur un nouveau serveur, surtout si l'origine de la corruption n'a pa été clairement identifiée), et ensuite de procéder à une validation méthodique des données. Il est probable que certaines données incohérentes puissent être identifier à l'import, lors de la phase de recréation des contraintes : celles-ci échoueront si les données ne les respectent, ce qui permettra de les identifier. En ce qui concerne les incohérences qui passeront au travers de ces tests, il faudra les trouver et les corriger manuellement, en procédant à une validation fonctionnelle des données.

ATTENTION !!!

Il faut donc bien retenir les points suivants :

- pg_resetxlog n'est pas magique ;
- pg_resetxlog rend les données incohérentes (ce qui est souvent pire qu'une simple perte d'une partie des données, comme on aurait en restaurant une sauvegarde).
- n'utiliser pg_resetxlog que s'il n'y a aucun autre moyen de faire autrement pour récupérer les données ;
- ne pas l'utiliser sur l'instance ayant subit le problème, mais sur une copie complète effectuée à froid ;
- après usage, exporter toutes les données et les importer dans une nouvelle instance ;
- valider soigneuseuement les données de la nouvelle instance.

6 Cas type de désastres

- les cas suivants sont assez rares
- ils nécessitent généralement une restauration
- certaines manipulations à haut risque sont possibles
  - mais complètement déconseillées !

Cette section décrit quelques unes des pires situations de corruptions que l'on peut être amené à observer. Dans la quasi-totalité des cas, la seule bonne réponse est la restauration de l'instance à partir d'une sauvegarde fiable.

6.1 Avertissement

- privilégier une solution fiable (restauration, bascule)
- les actions listées ici sont destructives
- la plupart peuvent (et vont) provoquer des incohérences
- travailler sur une copie

La plupart des manipulations mentionnées dans cette partie sont destructives, et peuvent (et vont) provoquer des incohérences dans les données. Tous les experts s'accordent pour dire que l'utilsation de telles méthodes pour récupérer une instance tend à aggraver le problème existant ou à en provoquer de nouveaux, plus graves. S'il est possible de l'éviter, ne pas les tenter (ie: préférer la restauration d'une sauvegarde) !

S'il n'est pas possible de faire autrement (ie: pas de sauvegarde utilisable, données vitales à extraire...), alors TRAVAILLER SUR UNE COPIE.

Il ne faut pas non plus oublier que chaque situation est unique, prendre le temps de bien cerner l'origine du problème, documenter chaque action prise, s'assurer qu'un retour arrière est toujours possible.

6.2 Corruption de blocs dans des index

- messages d'erreur lors des accès par l'index
- données différentes entre un indexscan et un seqscan
- supprimer et recréer l'index (`REINDEX`)

Les index sont des objets de structure complexe, ils sont donc particulièrement vulnérables aux corruptions. Lorsqu'un index est corrompu, on aura généralement des messages d'erreur de ce type :

```
ERROR: invalid page header in block 5869177 of relation base/17291/17420
```

Il peut arriver qu'un bloc corrompu ne renvoie pas de message d'erreur à l'accès, mais que les données elles-mêmes soient altérées. Ce cas est néanmoins très rare dans un block d'index.

Dans la plupart des cas, si les données de la table sous-jacente ne sont pas affectées, il est possible de réparer l'index en le reconstruisant intégrament grâce à la commande `REINDEX`.

6.3 Corruption de blocs dans des tables 1

- cas plus problématique
- restauration probablement nécessaire

Les corruptions de block vont généralement déclencher des erreurs du type suivant :

```
ERROR: invalid page header in block 32570 of relation base/16390/2663
ERROR: could not read block 32570 of relation base/16390/2663: read only 0
of 8192 bytes
```

Si la relation concernée est une table, tout ou partie des données contenues dans ces blocs sont perdues. L'apparition de ce type d'erreur est un signal fort qu'une restauration est certainement nécessaire.

6.4 Corruption de blocs dans des tables 2

- le paramètre `zero_damaged_pages` peut aider
- des données vont certainement être perdues

Néanmoins, s'il est nécessaire de lire le maximum de données possibles de la table, il est possible d'utiliser l'option de PostgreSQL `zero_damaged_pages` pour demander au moteur de réinitialiser les blocs invalides à zéro lorsqu'ils sont lus. Par exemple :

```
> SET zero_damaged_pages = TRUE ;
SET
> VACUUM FULL mycorruptedtable ;
WARNING:  invalid page header IN block 32570 OF relation base/16390/2663;
zeroing OUT page
VACUUM
```

Si cela se termine sans erreur, les blocs invalides ont été réinitialisés. Les données qu'ils contenaient sont évidemment perdues, mais la table peut désormais être accédée dans son intégralité en lecture, permettant ainsi par exemple de réaliser un export des données pour récupérer ce qui peut l'être.

Attention, du fait des données perdues, le résultat peut être incohérent (contraintes non respectées...). Par ailleurs, par défaut PostgreSQL ne détecte pas les corruptions logiques, c'est à dire n'affectant pas la structure des données mais uniquement le contenu. Il ne faut donc pas penser que la procédure d'export complet de données suivi d'un import sans erreur garanti l'absence de corruption.

6.5 Corruption de blocs dans des tables 3

- si la corruption est importante, l'accès au bloc peut faire crasher l'instance
- il est tout de même possible de réinitialiser le bloc
  - identifier le fichier à l'aide de `pg_relation_filepath()`
  - trouver le bloc avec `ctid` / `pageinspect`
  - réinitialiser le bloc avec dd
  - il faut vraiment ne pas avoir d'autre choix

Dans certains cas, il arrive que la corruption soit suffisamment importante pour que le simple accès au bloc fasse crasher l'instance. Dans ce cas, le seul

moyen de réinitialiser le bloc est de le faire manuellement au niveau du fichier, instance arrêtée, par exemple avec la commande dd. Pour identifier le fichier associé à la table corrompue, il est possible d'utiliser la fonction `pg_relation_filepath()` :

```
> SELECT pg_relation_filepath('test_corruptindex') ;
 pg_relation_filepath
----------------------
 base/16390/40995
(1 ROW)
```

Le résultat donne le chemin vers le fichier principal de la table, relatif au PGDATA de l'instance. Attention, une table peut contenir plusieurs fichiers. Par défaut une instance PostgreSQL sépare les fichiers en segments de 1 Go (paramètre `segment_size`). Une table dépassant cette taille aura donc des fichiers supplémentaires (`base/16390/40995.1`, `base/16390/40995.2`...). Pour trouver le fichier contenant le bloc corrompu, il faudra donc prendre en compte le numéro du bloc trouvé dans le champs `ctid`, multiplier ce numéro par la taille du bloc (paramètre `block_size`, 8 Ko par défaut), et diviser le tout par la taille du segment.

Cette manipulation est évidemment extrêment risquée, la moindre erreur pouvant rendre irrécupérables de grandes portions de données. Il est donc fortement déconseillé de se lancer dans ce genre de manipulations à moins d'être absolument certain que c'est indispensable. Encore une fois, ne pas oublier de travailler sur une copie, et pas directement sur l'instance de production.

6.6 Corruption des WAL 1

- situés dans le répertoire `pg_xlog`
- les WAL sont nécessaires au *recovery*
- démarrage impossible s'ils sont corrompus ou manquants
- si les fichiers WAL ont été archivés, les récupérer
- sinon, la restauration est la seule solution viable

Les fichiers WAL sont les journaux de transactions de PostgreSQL. Leur fonction est d'assurer que les transactions qui ont été effectuées depuis le dernier checkpoint ne seront pas perdue en cas de crash de l'instance. Si certains sont corrompus ou manquants (rappel : il ne faut JAMAIS supprimer les fichiers WAL si le système de fichiers est plein !), alors PostgreSQL ne pourra pas redémarrer.

Si l'archivage était activé et que les fichiers WAL affectés ont bien été archivés, alors il est possible de les restaurer avant de tenter un nouveau

démarrage.

Si ce n'est pas possible ou des fichiers WAL archivés ont également été corrompus ou supprimés, l'instance ne pourra pas redémarrer. Dans cette situation, comme dans la plupart des autres évoquées ici, la seule solution permettant de s'assurer que les données le seront pas corrompues est de procéder à une restauration de l'instance.

6.7 Corruption des WAL 2

- `pg_resetxlog` permet de forcer le démarrage
- ATTENTION !!!
- cela va provoquer des pertes de données
- des corruptions de données sont également probables
- ce n'est pas une action corrective !

L'utilitaire `pg_resetxlog` a comme fonction principale de supprimer les fichiers WAL courants et d'en créer un nouveau, avant de mettre à jour le fichier de contrôle pour permettre le redémarrage. Au minimum, cette action va provoquer la perte de toutes les transactions validées effectuées depuis le dernier checkpoint. Il est également probable que des incohérences vont apparaître, certaines relativement simples à détecter via un export / import (icohérences dans les clés étrangères par exemple), certaines complètement invisibles.

L'utilisation de cet utilitaire est extrêmement dangereuse, n'est pas recommandée, et ne peut jamais être considérée comme une action corrective. Il faut toujours privilégier la restauration d'une sauvegarde plutôt que son exécution.

Si l'utilisation de `pg_resetxlog` était néanmoins nécessaire (par exemple pour récupérer quelques données absentes de la sauvegarde), alors il faut travailler sur une copie des fichiers de l'instance, récupérer ce qui peut l'être à l'aide d'un export de données, et les inporter dans une autre instance. Les données récupérées de cette manière devraient également être soigneusement validées avant d'être importée de façon à s'assurer qu'il n'y a pas de corruption silencieuse. Il ne faut en aucun cas remettre une instance en production après une réinitialisation des WAL.

6.8 Corruption du fichier de contrôle

- fichier `global/pg_control`
- contient les informations liées au dernier checkpoint
- sans lui, l'instance ne peut pas démarrer
- restauration nécessaire

Le fichier de contrôle de l'instance contient de nombreuses informations liées à l'activité et au statut de l'instance, notamment l'instant du dernier checkpoint, la position dans les WAL correspondante, le numéro de transaction courant et le prochain à venir… Ce fichier est le premier lu par l'instance. S'il est corrompu ou supprimé, l'instance ne pourra pas démarrer.

Il est possible de forcer la réinitialisation de ce fichier à l'aide de la commande `pg_resetxlog`, qui va se baser par défaut sur les informations contenues dans les fichiers WAL présents pour tenter de “deviner” le contenu du fichier de contrôle. Ces informations seront très certainement erronées, potentiellement à tel point que même l'accès aux bases de données par leur nom ne sera pas possible :

```
-bash-4.2$ pg_isready
/var/run/postgresql:5432 - accepting connections

-bash-4.2$ psql postgres
psql: FATAL:  database "postgres" does not exist
```

Encore une fois, utiliser `pg_resetxlog` n'est aucune une solution, mais doit uniquement être considéré comme un contournement temporaire à une situation désastreuse. Une instance altérée par cet outil ne doit pas être considérée comme saine.

6.9 Corruption du CLOG

- fichier contenu dans `pg_clog`
- statut des différentes transactions
- son altération risque de causer des incohérences

Le fichier CLOG (*Commit Log*) contient le statut des différentes transactions, notamment si celles-ci sont en cours, validées ou annulées. S'il est altéré ou supprimé, il est possible que des transactions qui avaient été marquées

comme annulées soient désormais considérées comme valide, et donc que les modifications de données correspondantes deviennent visibles autres transactions.

C'est évidemment un problème d'incohérence majeur, tout problème avec ce fichier devrait donc être soigneusement analysé. Il est préférable dans le doute de procéder à une restauration et d'accepter une perte de données plutôt que de risquer de maintenir des données incohérentes dans la base.

6.10 Corruption du catalogue système

- le catalogue contient la définition du schéma
- sans lui, les données sont inaccessibles
- situation très délicate

Le catalogue système contient la définition de toutes les relations, les méthodes d'accès, la correspondance entre un objet et un fichier sur disque, les types de données existantes... S'il est incomplet, corrompu ou inaccessible, l'accès au données en SQL risque de ne pas être possible du tout.

Cette situation est très délicate, et appelle là encore une restauration. Si le catalogue était complètement inaccessible, sans sauvegarde la seule solution restante serait de tenter d'extraire les données directement des fichiers data de l'instance, en oubliant toute notion de cohérence, de type de données, de relation... Personne ne veut faire ça.

7 Conclusion

- les désastres peuvent arriver
- il faut s'y être préparé
- faites des sauvegardes !
  - et testez les