

Streaming Replication avec PostgreSQL



Michel EDWELL

PGsession 7 du 24/09/2015



METEO FRANCE
Toujours un temps d'avance

EPA MeteoFrance

EPA depuis 1993, sous la tutelle du Ministère de l'Écologie, du Développement durable et de l'Énergie

Budget 370,39 millions €

- Subvention de l'état PLF 2015 205 millions €

3152 agents

Grandes lignes du COP 2012->2016

- Intégrer sur un portail de web services une offre de données publiques enrichie
- Créer et alimenter un portail national de web services climatiques
- Doter l'établissement des moyens de calcul lourd adaptés à ses missions
- ...

<http://www.meteofrance.fr/nous-connaître/strategie-et-gouvernance>



Introduction

- Evolutions de la Gestion des Bases de Données à MF
- Haute disponibilité – Pourquoi pas la Replication ?
 - Réplication de la base PACOME
 - Conclusions & perspectives

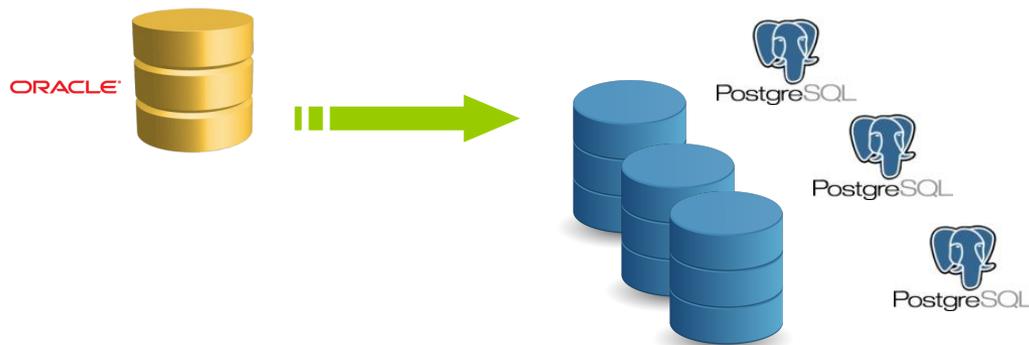


Bases de données de MeteoFrance

100 + serveurs PostgreSQL

10 + serveurs Rac Oracle

80 TB de données



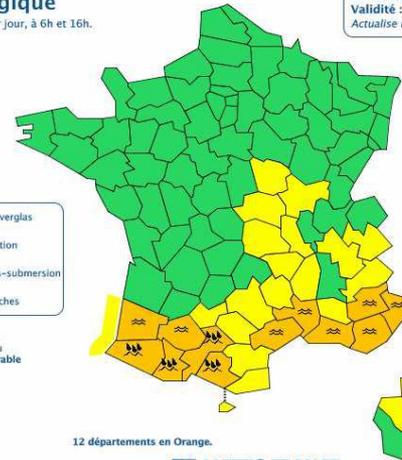
PostgreSQL depuis 2001



Vigilance météorologique

La carte est actualisée au moins 2 fois par jour, à 6h et 16h.

- Une vigilance absolue s'impose des phénomènes dangereux d'intensité exceptionnelle sont prévus...
- Soyez très vigilant, des phénomènes dangereux sont prévus ...



Diffusion : le dimanche 06 novembre 2011 à 10h15
Validité : Jusqu'au lundi 07 novembre 2011 à 06h00
Actualise la carte du dimanche 06 novembre 2011 à 06h01

Consultez le [bulletin national](#)

De l'hérault aux Alpes-maritimes
poursuite des crues.Forts cumuls de
pluie sur les départements pyrénéens.
Crues significatives essentiellement sur
les Pyrénées-Atlantiques.

Cliquez sur la carte pour lire
les [bulletins régionaux](#)

Conseils des pouvoirs publics :
Crues/Orange – Renseignez-vous avant
d'entreprendre vos déplacements et soyez très
prudents. Respectez, en particulier, les
déviations mises en place. – Ne vous engagez
en aucun cas, à pied ou en voiture, sur une
voie immergée. – Dans les zones habituellement
inondables, mettez en sécurité vos biens
susceptibles d'être endommagés et surveillez
la montée des eaux. Précipitations/Orange –
Renseignez-vous avant d'entreprendre un
déplacement ou toute autre activité
extérieure. – Evitez les abords des cours
d'eau. – Renseignez-vous sur les conditions de
circulation.

12 départements en Orange.

METEO FRANCE
Toujours un temps d'avance

Copyright Météo-France

En savoir plus et paramétrer les cookies.

METEO FRANCE Prévisions Météo-France et Vous Données publiques Professionnels et collectivités Boutique particuliers Autres sites A* A*

Données publiques

Vigilance Météo
Phénomènes dangereux
Consultez la carte

Rechercher

Déjà inscrit ?
Veuillez entrer e-mail
Votre mot de passe
Mot de passe oublié ?

Créer un compte
S'inscrire

Accueil Textes officiels et conditions d'accès Catalogue Données libres d'accès Informations sur les stations Référence tarifaire Géoservices FA.Q.

Catalogue

- Observations In situ
- Observations Radar
- Observations Satellite
- Climatologie
- Modèles et données de prévisions
- Prévisions climatiques

Bienvenue sur le portail de données publiques de Météo-France.

Nouveauté
Les sorties des principaux modèles de précision atmosphérique de Météo-France sont désormais disponibles en téléchargement libre sur le site (Rubrique du Catalogue : Modèles et données de prévision/ Modèles en téléchargement direct)

Présentation Générale
Météo-France produit et diffuse quotidiennement un très grand volume d'informations dans le cadre de ses missions de service public.

Un grand nombre d'entre elles peuvent être réutilisées - sous certaines conditions légales, réglementaires et contractuelles - en tant qu'"informations publiques", en application de la loi n°78-753 du 17 juillet 1978, pour des fins différentes de la mission de service public pour laquelle les informations ont été produites ou reçues par Météo-France.

En savoir plus sur la réutilisation des données



METEO FRANCE
Toujours un temps d'avance

Calcul 2013 ...une contrainte externe potentielle



Arrivée des données Calcul 2013
modèles à résolution plus importante

- Risque sur les performances
- Coût estimé en licence Oracle pour doubler la puissance CPU sur les bases opérationnelles (3 clusters) 950 k€



Stockage Clusters Oracle

Groupes de disques : BDP2A2.meteo.fr_BDPA23 (Instance de base de données) - Oracle Enterprise Manager - Mozilla Firefox

Navigation cible

BDPA2.meteo.fr_BDPA23

Base de données Oracle > Performances > Disponibilité > Schéma > Administration

Instance de base de données: BDP2A2.meteo.fr_BDPA23 > Groupes de disques

Groupes de disques

⚠ La base de données peut utiliser les groupes de disques ASM suivants pour le stockage. Vous êtes actuellement connecté à l'instance de base de données en tant qu'utilisateur NORMAL. Par conséquent, vous ne disposez peut-être pas de privilèges permettant d'accéder à certaines informations et opérations liées à ASM. Pour accéder aux ressources ASM associées à la base de données (fichiers, alias, répertoires, modèles, groupes, etc) via l'instance de base de données, vous devez vous connecter à la base de données en tant qu'utilisateur SYSDBA. Pour accéder aux autres opérations liées à ASM, connectez-vous directement à l'instance ASM +ASM3_saxo623

Créer Tout monter Tout démonter

Monter Démonter Rééquilibrer Vérifier Supprimer

Tout sélectionner | Ne rien sélectionner

Sélectionner	Nom	Etat	Redondance	Taille (Go)	Utilisé (Go)	Utilisé (%)	Espace libre utilisable (Go)	Disques membres
<input type="checkbox"/>	DATA	CONNECTED	EXTERN	4 000,00	3 482,50	87,06	517,50	8
<input type="checkbox"/>	DATA2	CONNECTED	EXTERN	5 000,00	1 680,42	33,61	3 319,58	5
<input type="checkbox"/>	FLASH	CONNECTED	EXTERN	1 536,00	3,57	0,23	1 532,43	6
<input type="checkbox"/>	OCRVOT	MOUNTED	EXTERN	30,00	0,43	1,44	29,57	3

✓ CONSEIL L'espace libre utilisable indique la quantité d'espace qui peut être utilisée en toute sécurité pour les données. Une valeur supérieure à zéro signifie que la redondance peut être correctement restaurée après la défaillance d'un disque.

✓ CONSEIL Les opérations Tout monter et Tout démonter monteront et démonteront uniquement les groupes de disques indiqués dans le paramètre Groupes de disques en montage automatique.



Haute disponibilité – Répliquer comment ?

Architecture & Activité du cluster Pacome

Quel est le besoin ? Répliquer comment ?

Mise en oeuvre de l'outil Slony

Mise en oeuvre de 2 serveurs HotStreaming

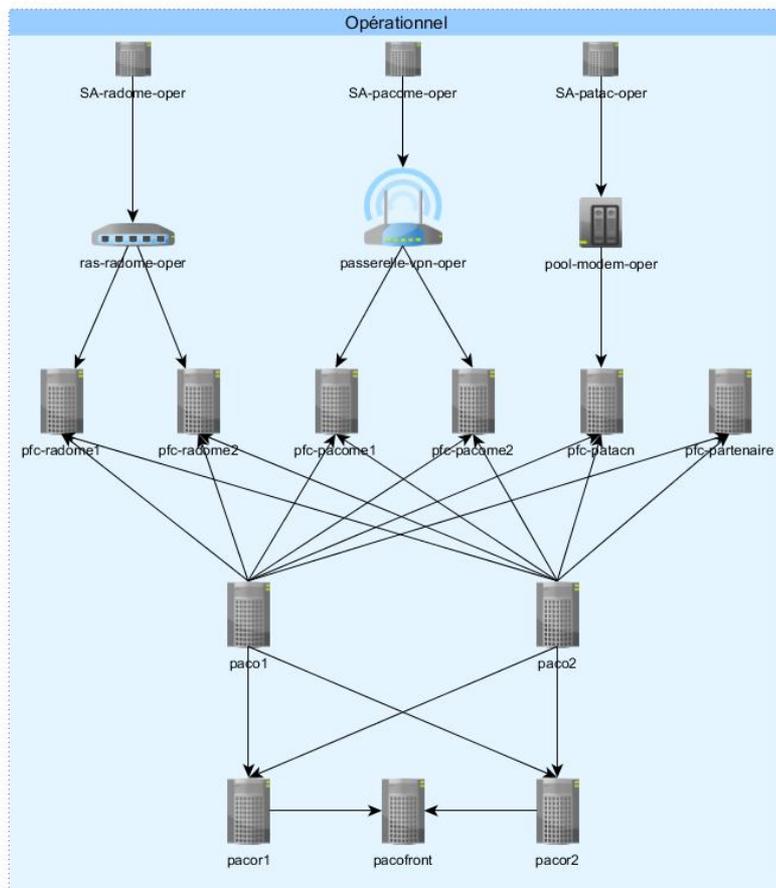
Supervision

Conclusions & perspectives



Architecture PACOME

Centraliser les paramètres observés au sol par les réseaux de stations automatiques



Cluster PACOME

Cluster 4 bases PostgeSql

Version majeure 9.1 (09/2011)

Pgbouncer version 1.4

Base de type DataWareHouse

1,5 To (à terme 4,5 To) de données réparties dans + 800 tables (4636 indexes)

790 toast tables

+ 800 procédures stockées



Ressources

System

paco3int.meteo.fr Hostname
Linux 2.6.18-371.1.2.el5 #1 SMP Kernel
x86_64 GNU/Linux Arch
CentOS release 5.6 (Final) Distribution

vm.dirty_background_bytes 0
vm.dirty_background_ratio 10
vm.dirty_bytes 0
vm.dirty_ratio 40
vm.overcommit_memory 1
vm.overcommit_ratio 50
vm.swappiness 5
vm.zone_reclaim_mode 0

CPU

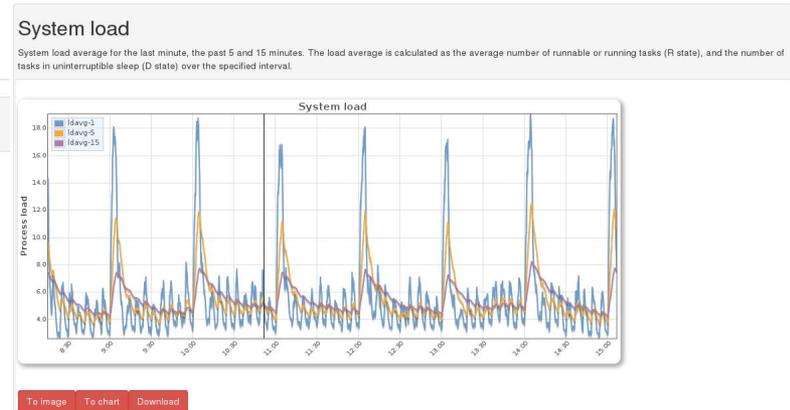
Intel(R) Xeon(R) CPU E5-2640 0
2494.095 Speed
15360 KB Cache
6 Sockets
24 Cores

Memory

61.44 GB Total memory
812.30 MB Free memory
326.46 MB Buffers
21.80 GB Cached
9.54 GB Total swap
9.52 GB Free swap
40.26 GB Commit limit
52.77 GB Committed

Filesystem

Filesystem	Size	Used	Free	Use%	Mount
/dev/cciss/c0d0p3	3.8G	588M	3.1G	16%	/
/dev/cciss/c0d0p8	28G	768M	26G	3%	/opt
/dev/mapper/vg00-pv00-lv01	0.5G	0.5G	0.0G	27%	/var



Activité du cluster

Cluster

Fri Aug 7 08:15:21 2015 to Fri Aug 7 15:06:53 2015

- 1.24 TB** Cluster size
- 4** Databases
- 104865** Connections
- 2391816716** Tuples returned
- 99%** Hit cache ratio
- 2** Extensions (plpgsql, pg_xlog_location_diff)

Databases

Fri Aug 7 08:15:26 2015 to Fri Aug 7 15:06:56 2015

- 1.08 TB** Largest database (SosDatabase)
- 97348** Most connections (SosDatabase)
- 2193250686** Most tuples fetched (SosDatabase)
- 99%** Worst cache utilization (SpsDatabase)

System

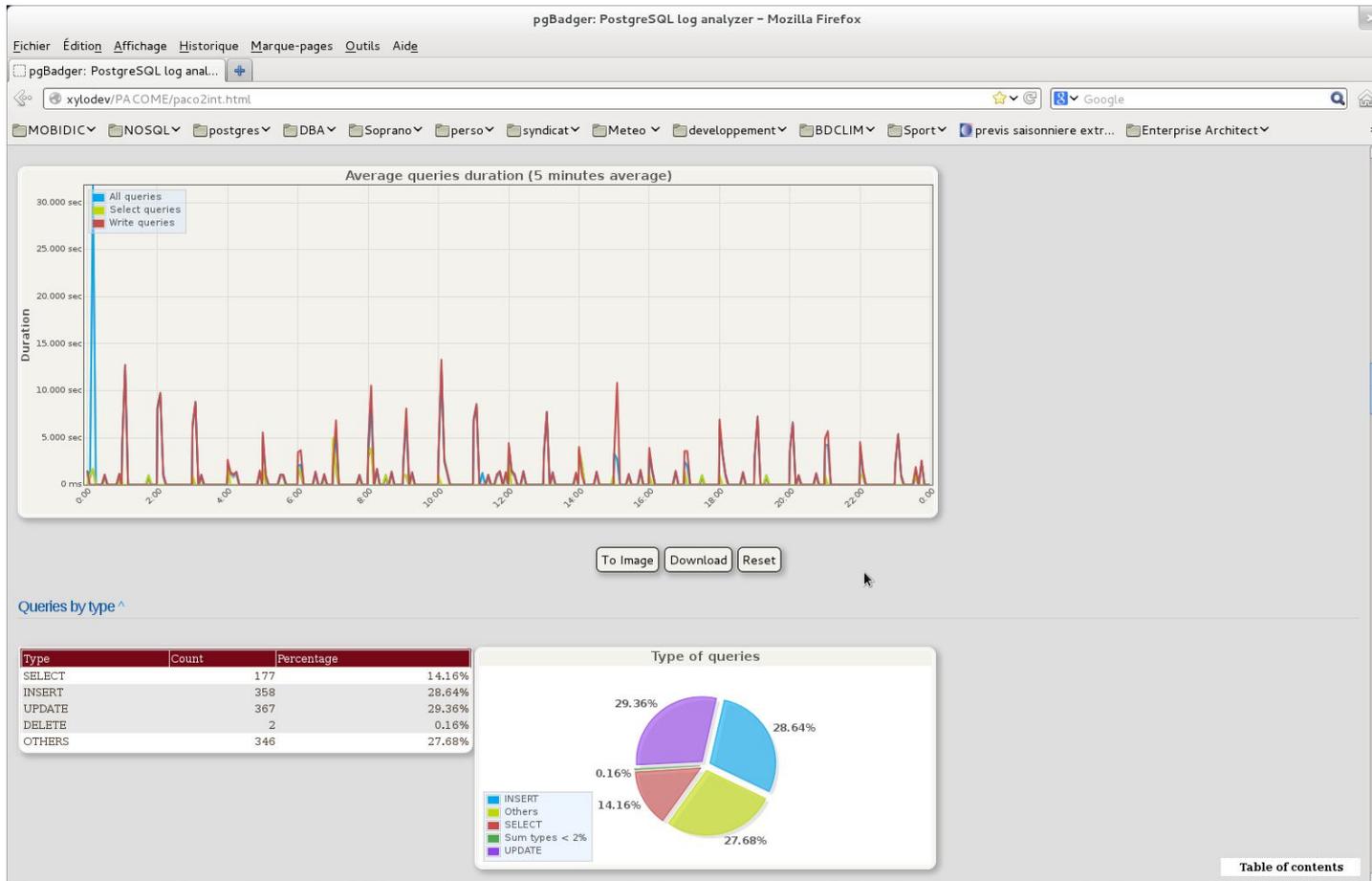
Fri Aug 7 08:15:26 2015 to Fri Aug 7 15:06:56 2015

- 74.81%** Highest CPU utilization (Fri Aug 7 12:02:45 2015)
- 19.06** Highest system load (Fri Aug 7 14:04:26 2015)
- 21.26 GB** Lowest system cache (Fri Aug 7 09:03:23 2015)
- 51.00** Highest device service time (Fri Aug 7 10:02:15 2015)
- nodev** Most read device (5.04 GB)
- nodev** Most written device (20.71 GB)

Database SosDatabase

- 1.08 TB** Total size
- 2** Installed extensions (plpgsql, pg_xlog_location_diff)
- 5** Schemas (metadata, obs_brute, obs_oper, obs_super, public)
 - Last manual vacuum
 - Last manual analyze
- 817** Stored procedures
- 0** Triggers
- 16** sequences
- 1** composite types
- 4636** indexes
- 801** tables
- 790** toast tables
- 2** views

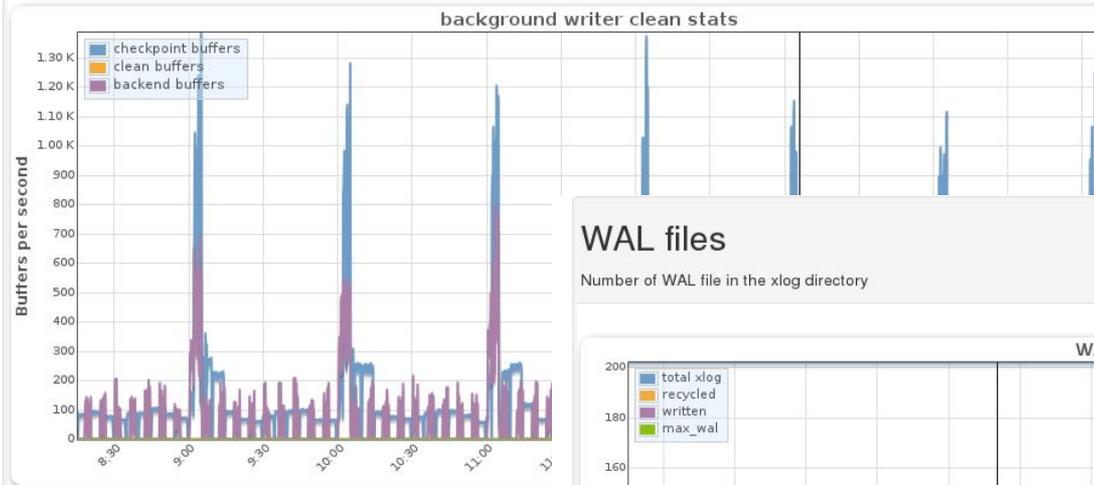
Activité(2) pgBadger



Activité (3) bgwriter & WAL

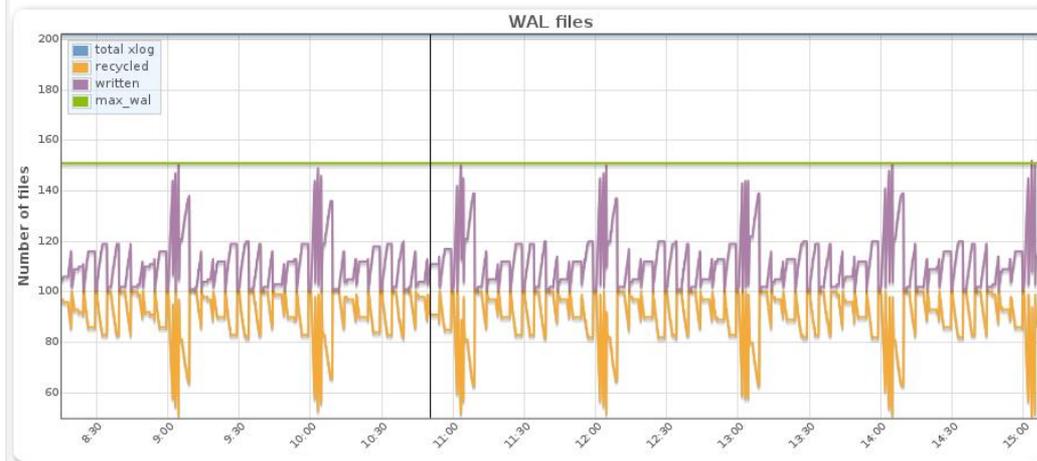
background writer clean stats

Background writer cache cleaning statistics by checkpoints, lru and backends.



WAL files

Number of WAL file in the xlog directory



Quel besoin ?

- Accès + nombreux à la base opérationnelle PACOME
 - Sécuriser la base opérationnelle en terme de requêtes
 - Répartir la charge
- Latence de mise à jour des données n'est pas un facteur critique vu le type d'accès qui y sera réalisé.
- Niveau de service demandé doit être continu avec un mode reprise et supervisé.

Contrainte 9.1 (09/2011)



Réplication PACOME

Le **mode asynchrone asymétrique** répond aux exigences (contrainte version majeure 9.1)

- écriture uniquement sur un seul maitre

Répliquer à priori : les requêtes

Hot-Streaming

Dev 2ndQuadrant intégré 9.0
relecture des journaux de transaction par serveur en standby
pas synchrone... mais très rapide !

Répliquer à posteriori : les changements

Slony

Mise à jour différée des tables sur l'autre serveur



Réplication Asynchrone : slony

Triggers ajoute overhead sur le maitre

Données perdues sur fail-over

Outil de réplication de bases de données puissant et éprouvé

Complexité de mise en œuvre et d'administration

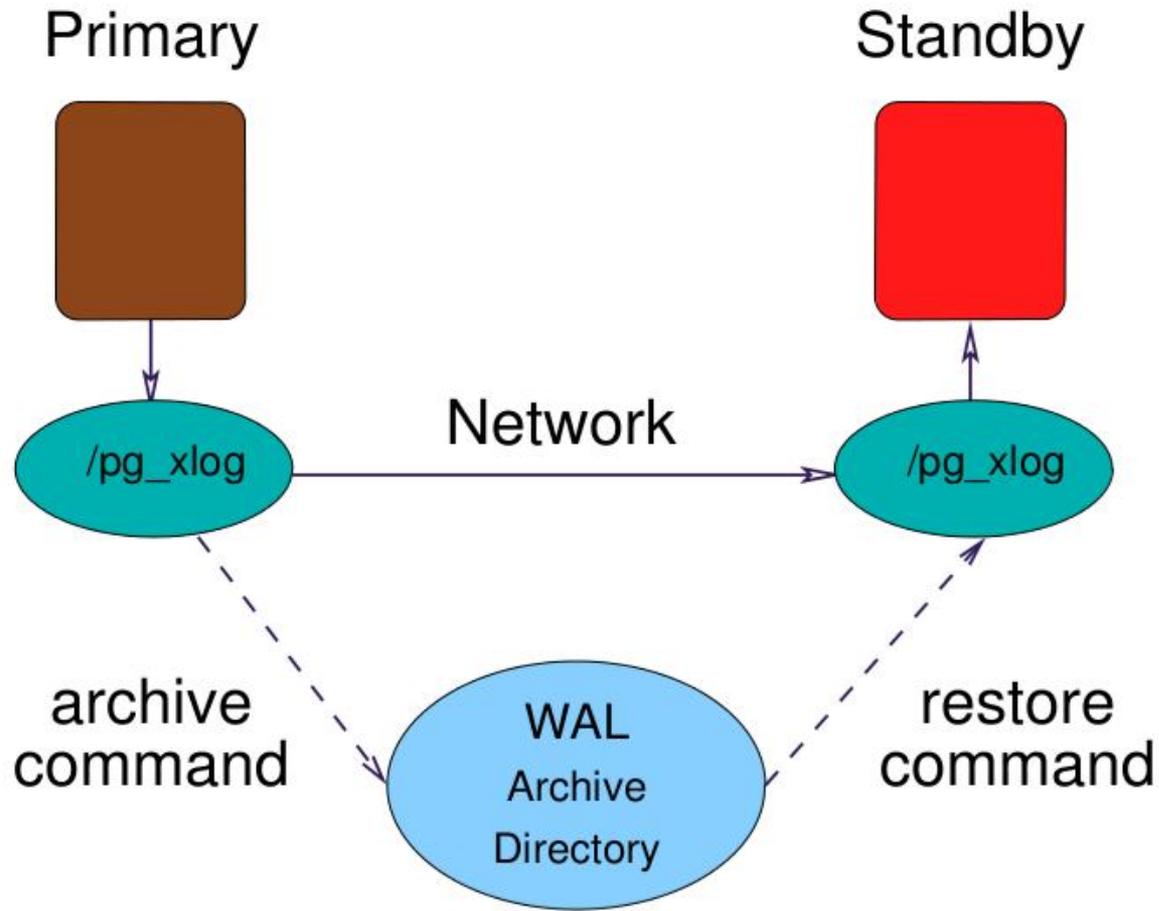
- Granularité niveau table

Extrait d'un dataset de réplication slony :

```
set add table (set id=1, origin=1, id=1, fully qualified name='agro.bilhydri_q_v1',  
comment='bilhydri_q_v1 table');  
set add table (set id=1, origin=1, id=2, fully qualified name='agro.decadagro_v2',  
comment='decadagro_v2 table');  
set add table (set id=1, origin=1, id=3, fully qualified name='alti.alti_v1',  
comment='alti_v1 table');
```



Réplication Hot Streaming



Configuration interchangeable

- Configuration indépendante du rôle dans la réplication
- Déployer des packages rpm
 - postgresql.conf
 - pg_hba.conf
 - fichiers include



Configuration du Standby

Le fichier **recovery.conf** :

- `standby_mode = 'on'`
- `restore_command = 'scp pacobase@paco2int
:/home/pacobase/backup/WALS_MASTER/%f %p 2>
/dev/null'`
- `primary_conninfo = 'host=master port=5432 user=replic
password=xxapplication_name=standby1'`
- `trigger_file =
'/home/pacobase/backup/TRG_FILES/standby1_trigger'`



Synchronisation initiale des données

```
SosDatabase=# select pg_start_backup('Backup SosDatabase 25/03/2015  
18:27',TRUE);
```

```
pg_start_backup
```

```
-----
```

```
17CD/17000020
```

```
(1 row)
```

```
[pacobase@slave:~/log/pgsql]$ time(rsync -avz --bwlimit=20000  
pacobase@master:/home/pacobase/var/pgsql/ --exclude="pg_xlog/*"  
/home/pacobase/var/pgsql)  
real 504m27.557s  
user 363m14.668s  
sys 111m20.843s
```

The copy is inconsistent, but that is okay (WAL replay will correct that).

Signal the backup is complete from psql :

```
SosDatabase=# select pg_stop_backup() ;
```



Synchronisation initiale des données (2)

Le fichier backup_label :

```
[pacobase@paco5int:~/var/pgsql]$ cat backup_label
```

```
START WAL LOCATION: 20EF/3C5AC548 (file 00000001000020EF0000003C)
```

```
CHECKPOINT LOCATION: 20EF/42DD81D0
```

```
BACKUP METHOD: pg_start_backup
```

```
START TIME: 2015-06-22 10:13:08 GMT
```

```
LABEL: synchro init paco5int
```



lowait pg_basebackup

```
[~/var]$time(nice -n 19 ionice -c2 -n7 ../pg_basebackup -h server -U  
replic -D /home/... -vP| pv -q -t --rate-limit 10m)
```

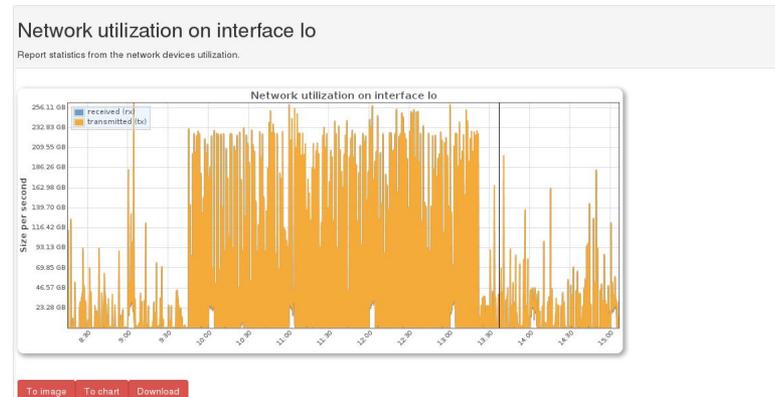
avg-cpu: %user %nice %system %iowait %steal %idle

37,10 0,00 16,73 7,27 0,00 38,91

Device: tps Blk_read/s Blk_wrtn/s Blk_read Blk_wrtn

cciss/c0d0 0,00 0,00

....



Sous le capot

La streaming replication est implémentée avec deux processus

- Walsender sur le maitre
- Walreceiver sur l'esclave

Les processus

- [pacobase@**master**:~/var/pgsql/pg_xlog]\$ pgrep -lf "wal (sender|receiver) process"

2270 postgres: wal sender process replic slave1 (38512) streaming 2B44/18238A70

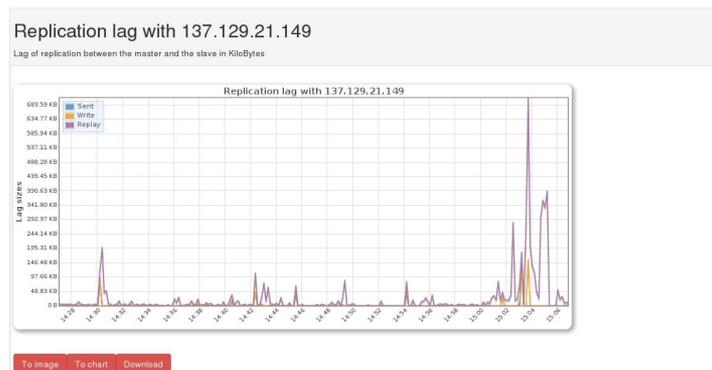
2918 postgres: wal sender process replic slave2 (60529) streaming 2B44/18238A70

- [pacobase@**slave**:~/var/pgsql]\$ pgrep -lf "wal (sender|receiver) process"
- 21053 postgres: wal receiver process streaming 2B44/18238A70



Monitoring

- Système :
 - Espace disque, surtout pour pg_xlog et les espaces d'archivage
 - Les processus de réplication
- Archivage :
 - Nombre de fichiers dans pg_xlog et/ou nombre de fichiers en attente d'archivage
- Replication :
 - Délai de la réplication
 - Disponibilité des serveurs standby



Les outils de monitoring

- L'outil d'administration pgAdmin
- Les procédures stockées
- Le catalogue système pg_stat_replication
- L'extension pg_xlog_location_diff (postgres 9.1)
- PgCluu et PgBadger
- La sonde check_postgres et l'outil de supervision shinken



Catalogue système pg_stat_replication

Version 9.1 catalogue système `pg_stat_replication`.

- > Elle n'est renseignée que sur le maître
- > permet de connaître l'état de tous les esclaves connectés.

Replication du cluster master pacome

23062015-09:50:33 catalogue système pg_stat_replication

```
-[ RECORD 1 ]-----+-----  
pg_current_xlog_location | 2107/6991C818  
-[ RECORD 1 ]-----+-----  
pg_current_xlog_location | 2938/67C0A230  
Procpid      | 2270  
Usesysid    | 474584  
username    | replic  
application_name | standby2  
client_addr  | 137.129.21.149  
client_hostname |  
client_port  | 38512  
backend_start | 2015-08-07 14:26:56.005882+00  
state        | streaming  
sent_location | 2938/67C0A230  
write_location | 2938/67C0A230  
flush_location | 2938/67C0A230  
replay_location | 2938/67C08980  
sync_priority | 0  
sync_state    | async  
-[ RECORD 2 ]-----+-----  
Procpid      | 2918  
Usesysid    | 474584  
username    | replic  
application_name | standby1  
client_addr  | 137.129.47.18  
client_hostname |  
client_port  | 60529  
backend_start | 2015-08-05 13:12:34.473543+00  
...
```

the function calculate the number of Bytes between two given XLOG offsets.

ecart_replication(pg_size_pretty): Compute the difference in xBytes between two WAL locations.

pg_size_pretty

0 bytes

(1 row)



METEO FRANCE
Toujours un temps d'avance

La sonde check-postgres

L'outil shinken et la sonde perl check_postgres
action = connect , hot_standby_delay, etc

```
1. [pacobase@pacor:~/SUPERVISION]$ /usr/bin/check_postgres.pl --action=connection --host=localhost -p 5432 -u  
replic -db SosDatabase
```

POSTGRES_CONNECTION OK: DB "SosDatabase" (host:localhost) version 9.1.14 | time=0.01s

```
2. [pacobase@pacor:~/SUPERVISION]$ /usr/bin/check_postgres.pl --action=hot_standby_delay --dbhost=localhost  
--dbport=5432 --dbname=SosDatabase --dbuser=pacome --dbpass=pacome --dbhost=137.129.21.147  
--dbport=5432 --warning=500 --critical=1000
```

POSTGRES_HOT_STANDBY_DELAY OK: DB "SosDatabase" (host:localhost) 0 | time=0.01s 'délai de
rejeu'=0;500;1000 'délai de réception'=0;500;1000

POSTGRES_HOT_STANDBY_DELAY CRITICAL: DB "SosDatabase" (host:localhost) 893008 | time=0.01s 'délai de
rejeu'=893008;500;1000 'délai de réception'=893008;500;1000



A venir

Supprimer la réplication DRBD (bascule maitre/esclave)

Upgrade du cluster postgres de 9.1 à 9.4

pg_basebackup, etc ...

Outils de monitoring HotStreaming

réplication en cascade



CONCLUSIONS

- Prévoir la réplication au plus tôt dans les projets
- Impact important sur l'architecture
- Exploitation du hot-standby par l'applicatif



Questions ?



METEO FRANCE
Toujours un temps d'avance