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'etat 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-connaitre/strategie-et-gouvernance

Introduction

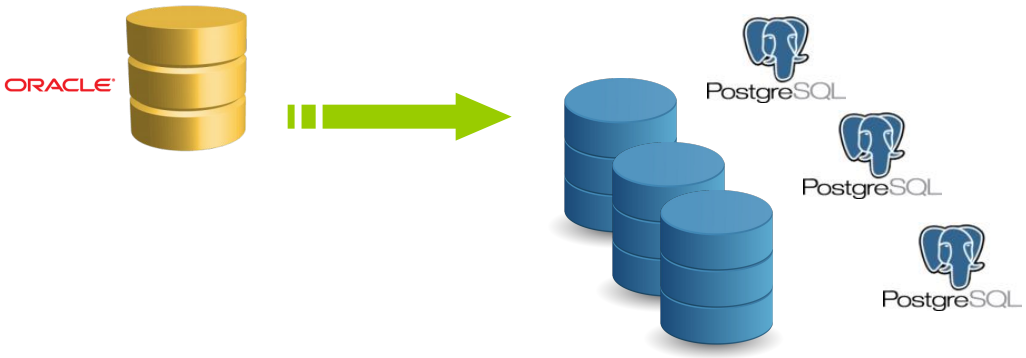
- Evolutions de la Gestion des Bases de Données à MF
- Haute disponibilité – Pourquoi pas la Replication ?
  - Réplication de la base PACOME
  - Conclusions & perspectives

Bases de données de MeteoFrance

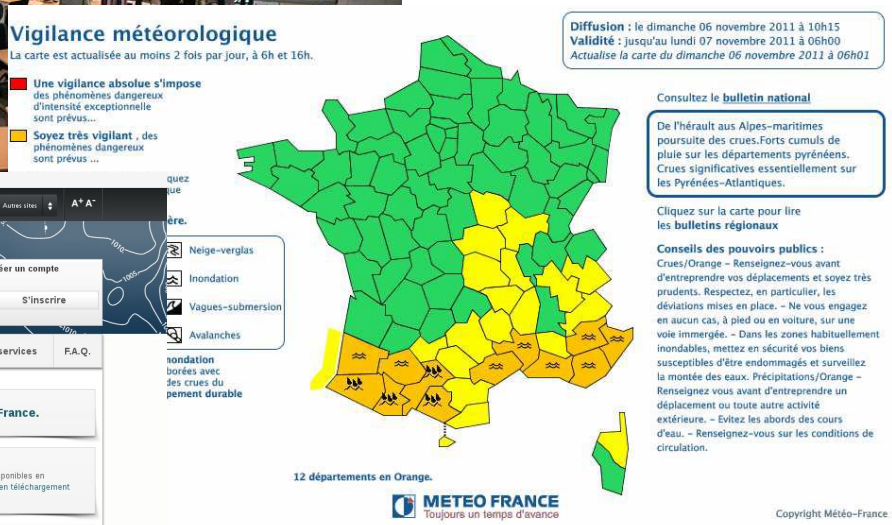
100 + serveurs PostgreSQL

10 + serveurs Rac Oracle

80 TB de données



PostgreSQL depuis 2001



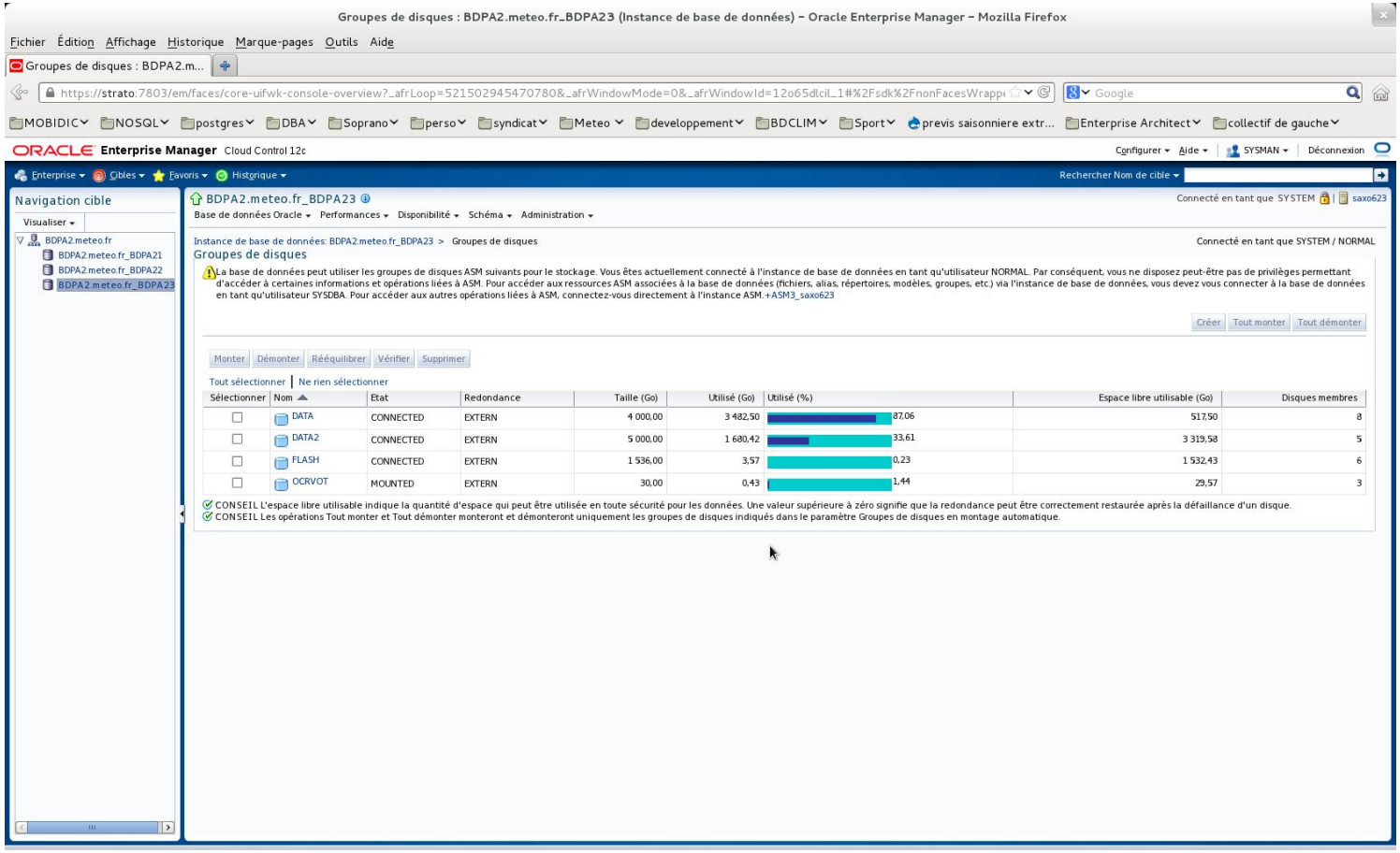
METEO FRANCE
Toujours un temps d'avance

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



Haute disponibilité – Répliquer comment ?

Architecture & Activité du cluster Pacome

Quel est le besoin ? Répliquer comment ?

Mise en oeuvre de l'outil Slony

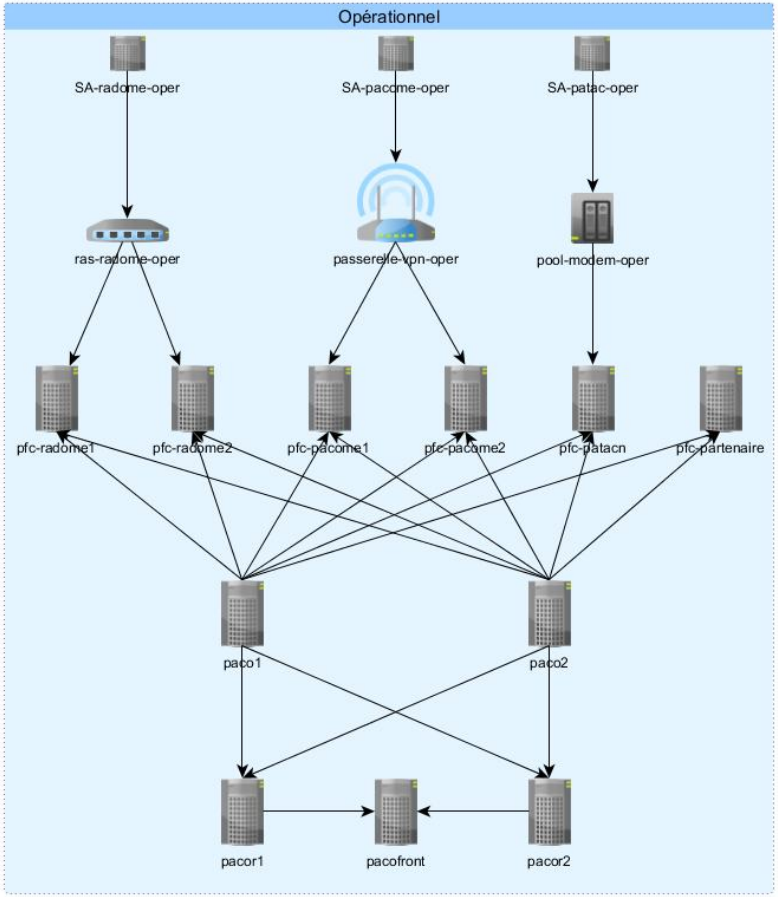
Mise en oeuvre de 2 serveurs HotStreaming

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

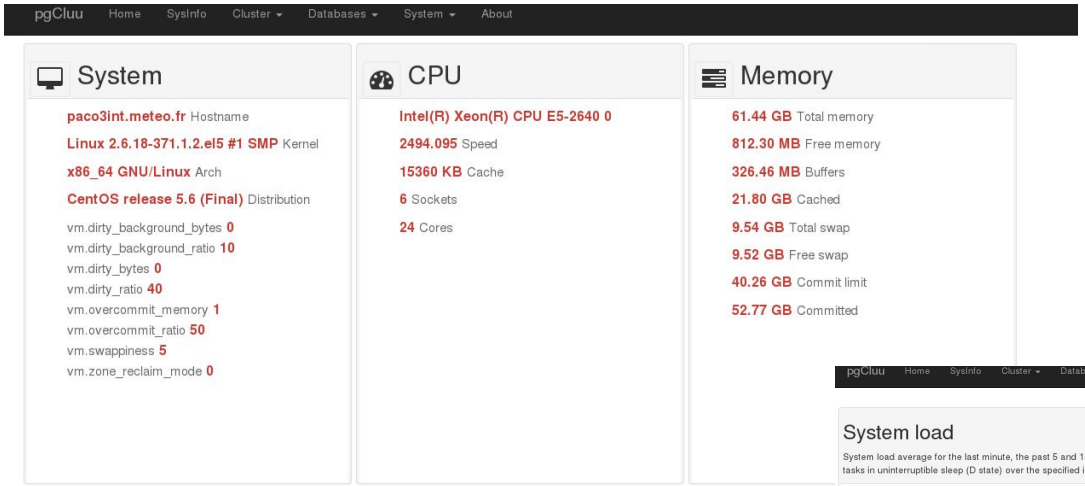
Base de type DataWareHouse

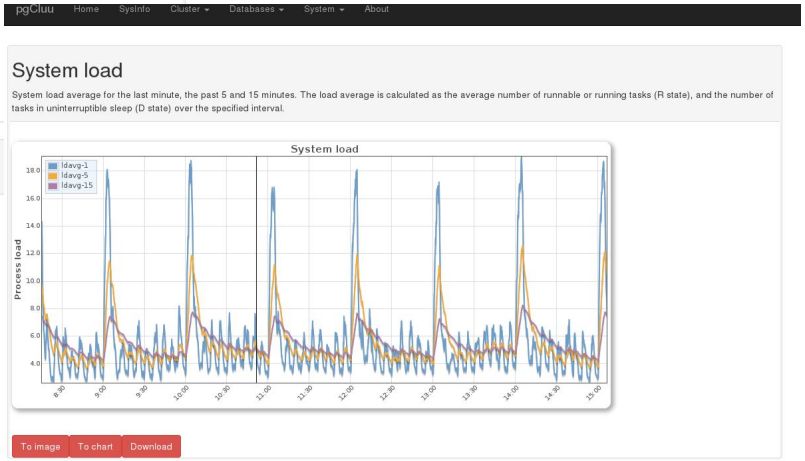
1,5 To (à terme 4,5 To) de données réparties dans + 800 tables (4636 indexes)

790 toast tables

+ 800 procédures stockées

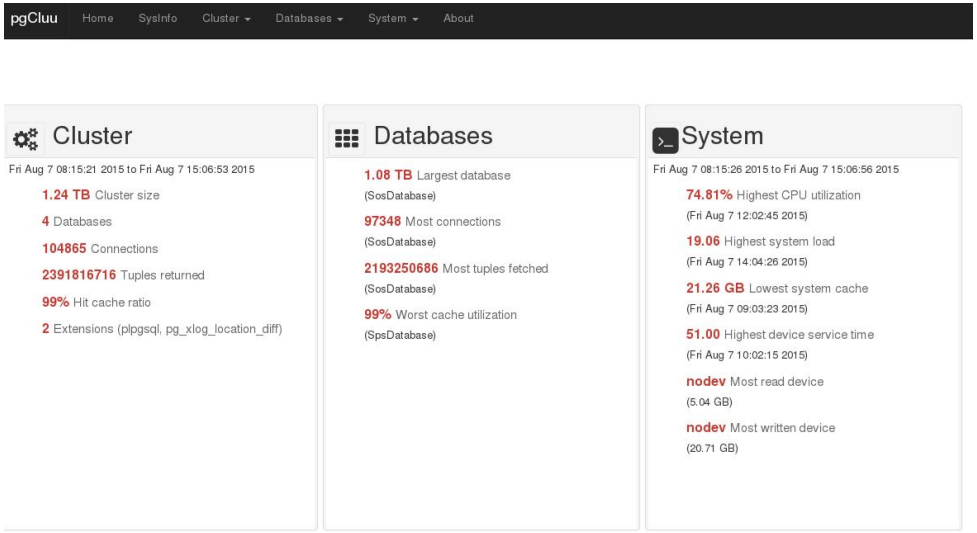
Ressources

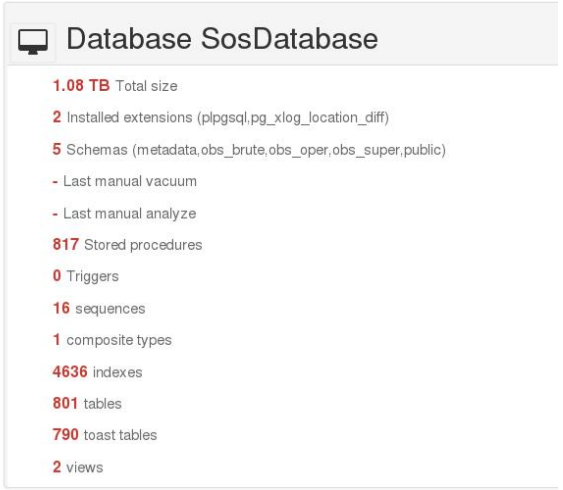




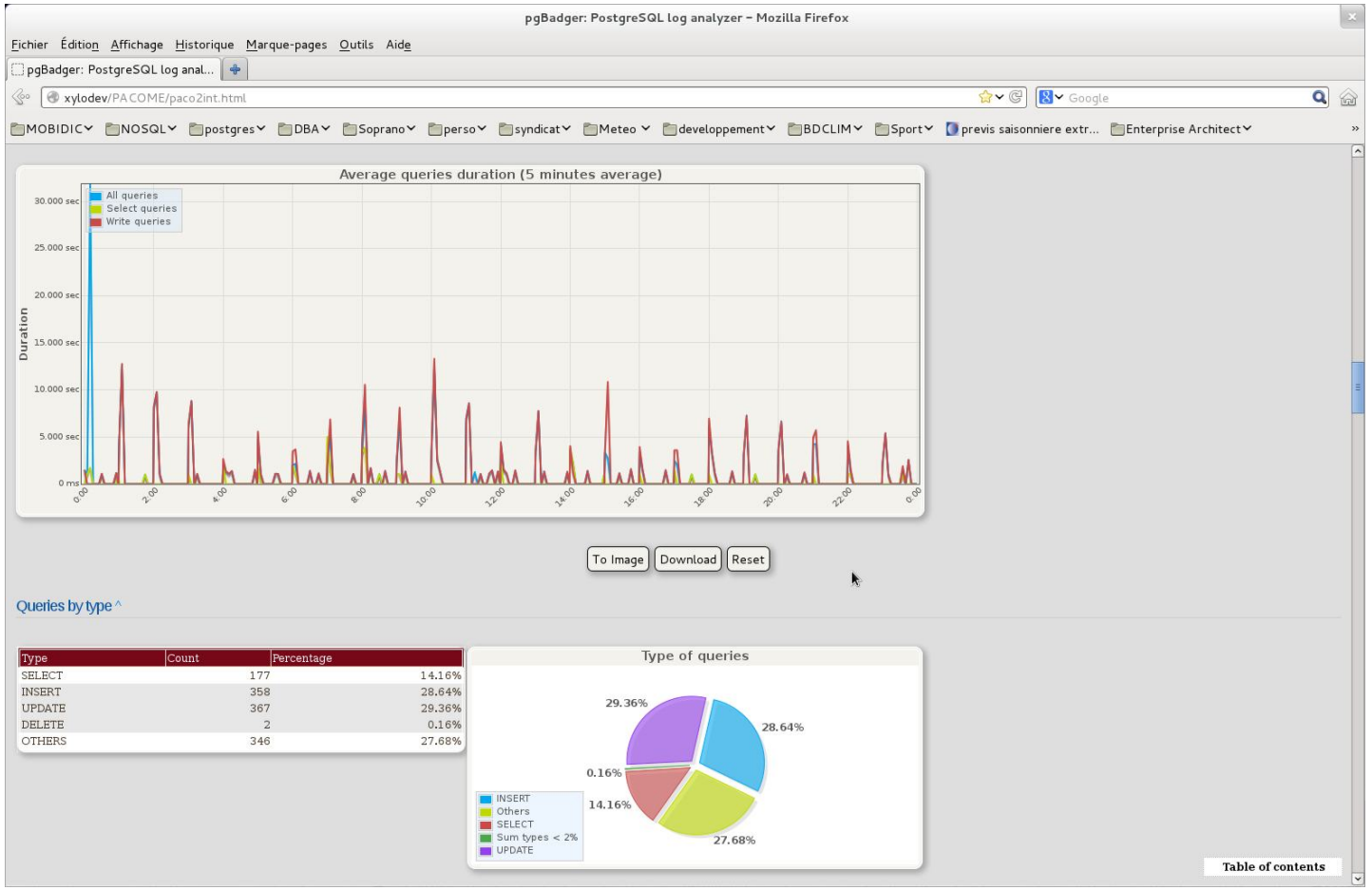
METEO FRANCE
Toujours un temps d'avance

Activité du cluster





Activité(2) pgBadger



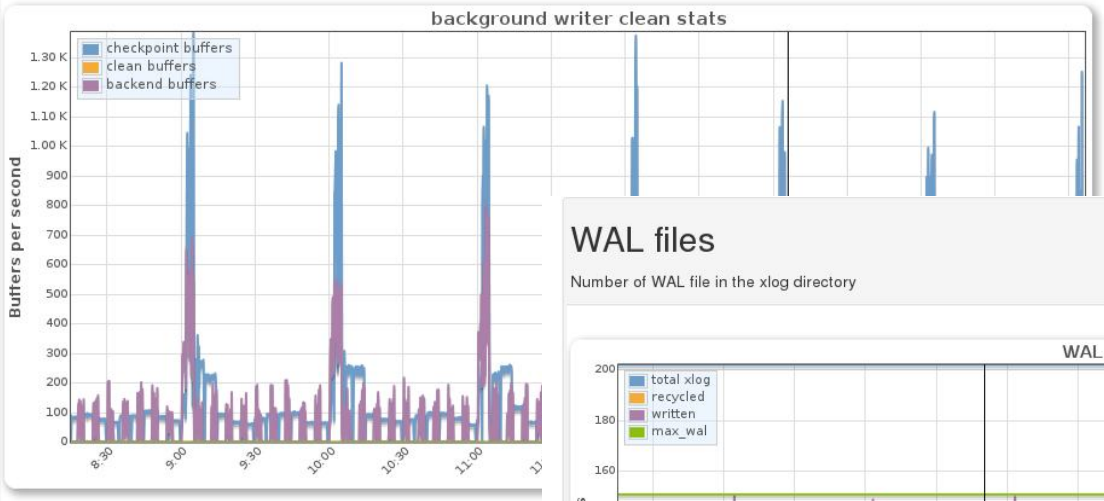
Type	Count	Percentage
SELECT	177	14.16%
INSERT	358	28.64%
UPDATE	367	29.36%
DELETE	2	0.16%
OTHERS	346	27.68%

METEO FRANCE
Toujours un temps d'avance

Activité (3) bgwriter & WAL

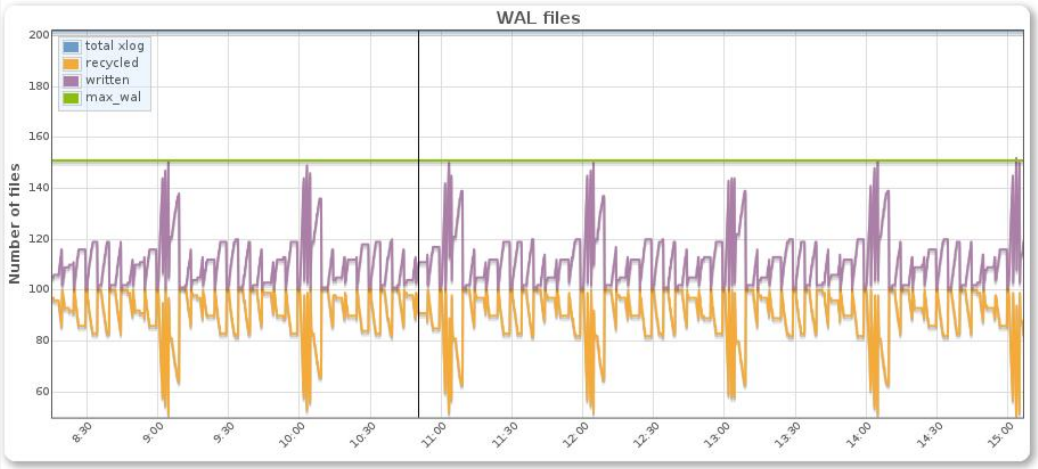
background writer clean stats

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



WAL files

Number of WAL file in the xlog directory



METEO FRANCE
Toujours un temps d'avance

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

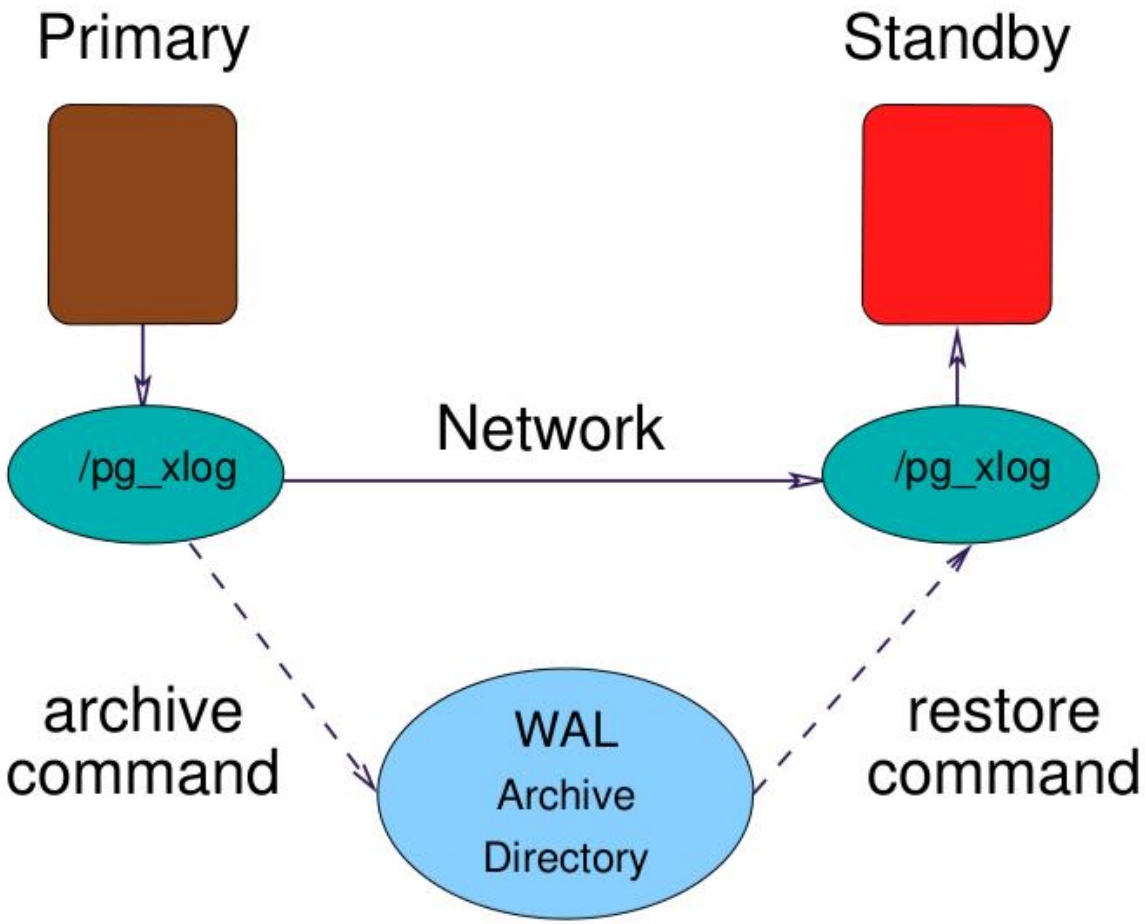
Complexité de mise en œuvre et d'administration

- 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 000000010000020EF0000003C)

CHECKPOINT LOCATION: 20EF/42DD81D0

BACKUP METHOD: pg_start_backup

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

LABEL: synchro init paco5int

Iowait 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)

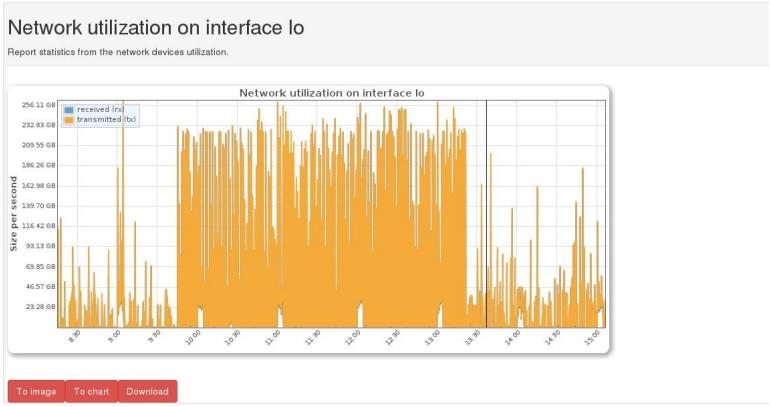
avg-cpu: %user   %nice %system %iowait %steal   %idle

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**

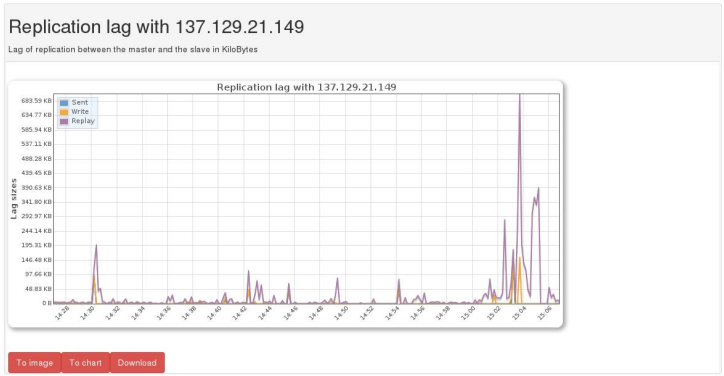
2918 postgres: wal sender process replic slave2 (60529) streaming 2B44/18238A70

- [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 procedures 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
usename          | 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
usename          | 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

METEO FRANCE
Toujours un temps d'avance