Яндекс

Яндекс Почта

“Двухколёсные” бекапы

Евгений Дюков

Бекапы?

- Посмотреть: https://events.yandex.ru/lib/talks/3202/
- Почитать:
  - https://www.postgresql.org/docs/current/static/backup-file.html
  - https://www.postgresql.org/docs/current/static/continuous-archiving.html

Давным-давно*

* В 2013-2015

- Маленькие базы (меньше 1 TB)
- Мало баз/бекапных серверов
- pg_basebackup over NFS
- archive_command = cp ... (NFS)
- Самописная реализация retention policy

Бекапы по 30+ часов

```
root@xivastore01d ~ # fgrep 10404 /var/log/
zkflock.log | fgrep 10609
2015-01-24 02:00:02 INFO    10404 Start
subprocess: /etc/cron.yandex/
pgbackup_database.sh (PID: 10609)

2015-01-25 11:53:43 INFO    10404 Child
returned code: 0 (PID: 10609)
```

Проблемы с pg_basebackup

- Без сжатия упираемся в сеть
- Со сжатием упираемся в одно ядро
- Размеры бекапов
- Отсутствие инкрементов даже на файловом уровне

Barman to the rescue

- File-level increments (reuse_backup=link)
- Built-in retention policy
- Все ещё один поток
- Нет сжатия бекапов

Почтовые метабазы

```
root@pg-backup03i ~ # barman list-backup
xdb2018 | cut -d' ' -f5-12

Feb 24 12:20:07 2016 - Size: 3.3 TiB

Feb 23 18:04:27 2016 - Size: 3.2 TiB

Feb 22 22:05:36 2016 - Size: 3.2 TiB

Feb 21 21:16:17 2016 - Size: 3.1 TiB

Feb 20 17:02:49 2016 - Size: 3.1 TiB

Feb 19 17:07:50 2016 - Size: 3.1 TiB

Feb 18 19:19:20 2016 - Size: 3.0 TiB
```

File-level инкременты нам не подходят

- Изменения размазаны по большому количеству data-файлов
- Нет возможности сделать партиционирование по дате/времени
- 1% изменённых строк выливается в бекап >50% данных

```
Disk usage: 4.9 TB (5.0 TB with WALs)

Incremental size: 2.8 TB (-42.86%)
```

В итоге может получиться одна база - один бекапный сервер

Коллега-DBA

Поищем другие решения?

pg_rman

- Исходники: https://github.com/ossc-db/pg_rman
- Умеет сжимать
- Page-level increments
- Требует интерфейса файловой системы (здравствуй, NFS)
- Однопоточный (https://github.com/ossc-db/pg_rman/issues/5)

pgBackRest

- Исходники: https://github.com/pgbackrest/pgbackrest
- Умеет сжимать
- Многопоточный
- File-level increments :(

Значит, двухколёсные...

Агент на машине с БД (barman-incr)

- Написан на Python
- Многопоточный
- Умеет сжимать данные (gzip/bzip/lzma)
- Может брать только измененные страницы (ориентируется на LSN в page-header’e, PostgreSQL 9.3+)

Изменения в barman

- Запуск агента по ssh
- Зависимости бекапов друг от друга
  - Например, инкрементальные зависят от полных
  - Retention policy не удалит бекап, если он устарел, но зависящий от него нет

Многопоточность

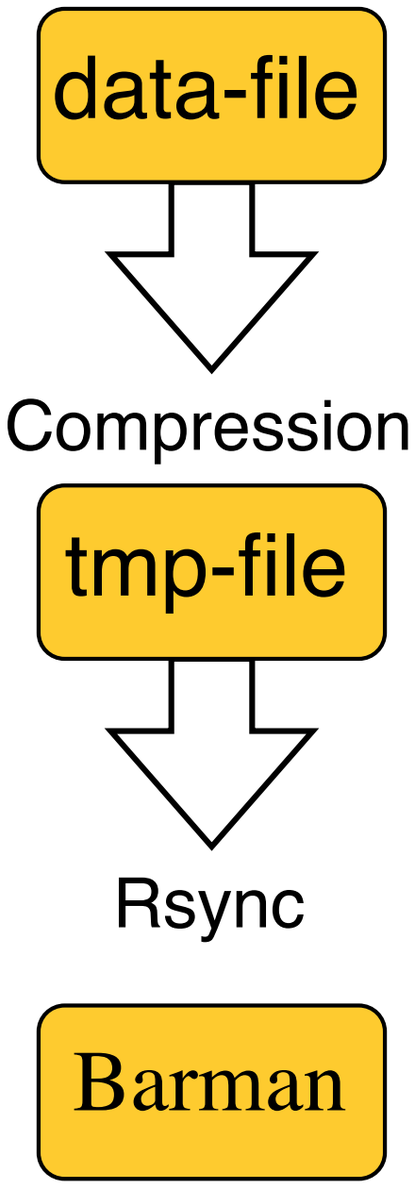
- Multi-threaded
  - Этот вариант отбросили из-за GIL
- Multi-processed
  - Мастер-процесс
  - N рабочих процессов
  - Все необходимые данные передаются небольшой структурой

Сжатие

- Временная директория на машине с БД
- Каждый рабочий процесс работает с одним файлом за раз
- Python is awesome (get_compression_opener < 30 строк!)

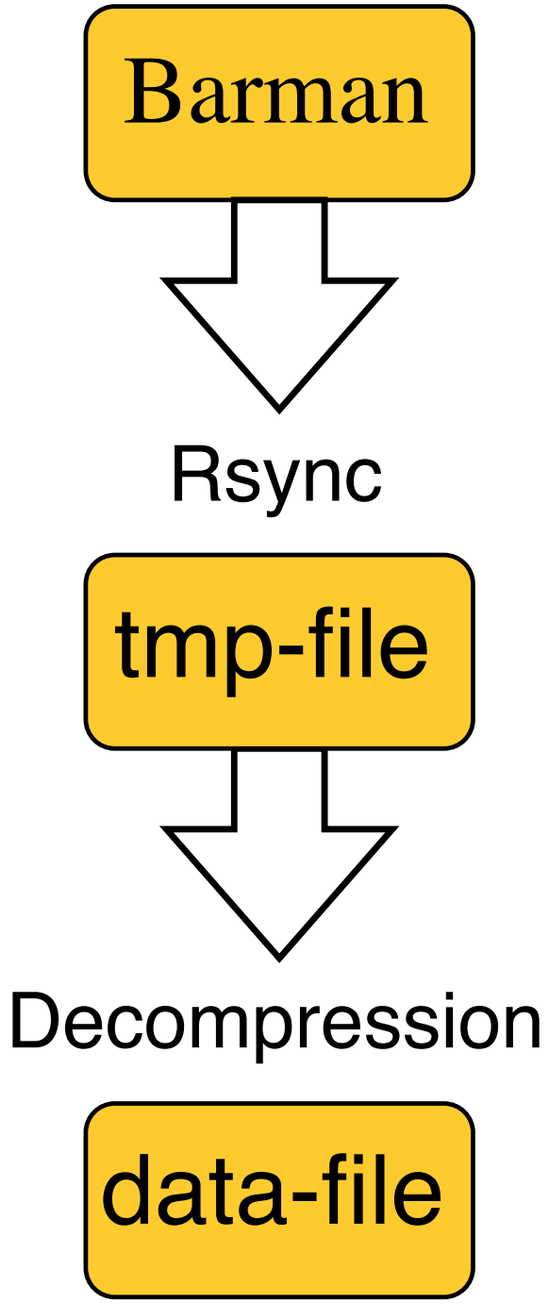
Полный бекап

- Просто читаем файл и пишем во временную директорию, потом передаем на barman-сервер
- Почти rsync, только многопоточный и умеет сжимать данные
- Имя файла и его размер записываем в список, который в конце передадим на barman-сервер



Восстановление из полного бекапа

- Скачиваем список файлов с barman-сервера в tmp-директорию
- Скачиваем файл во временную директорию, потом распаковываем в pgdata



Cumulative / Differential backups

Cumulative

〉 “Предыдущий” бекап - всегда полный

〉 Меньше время восстановления

〉 Занимают больше места

Differential

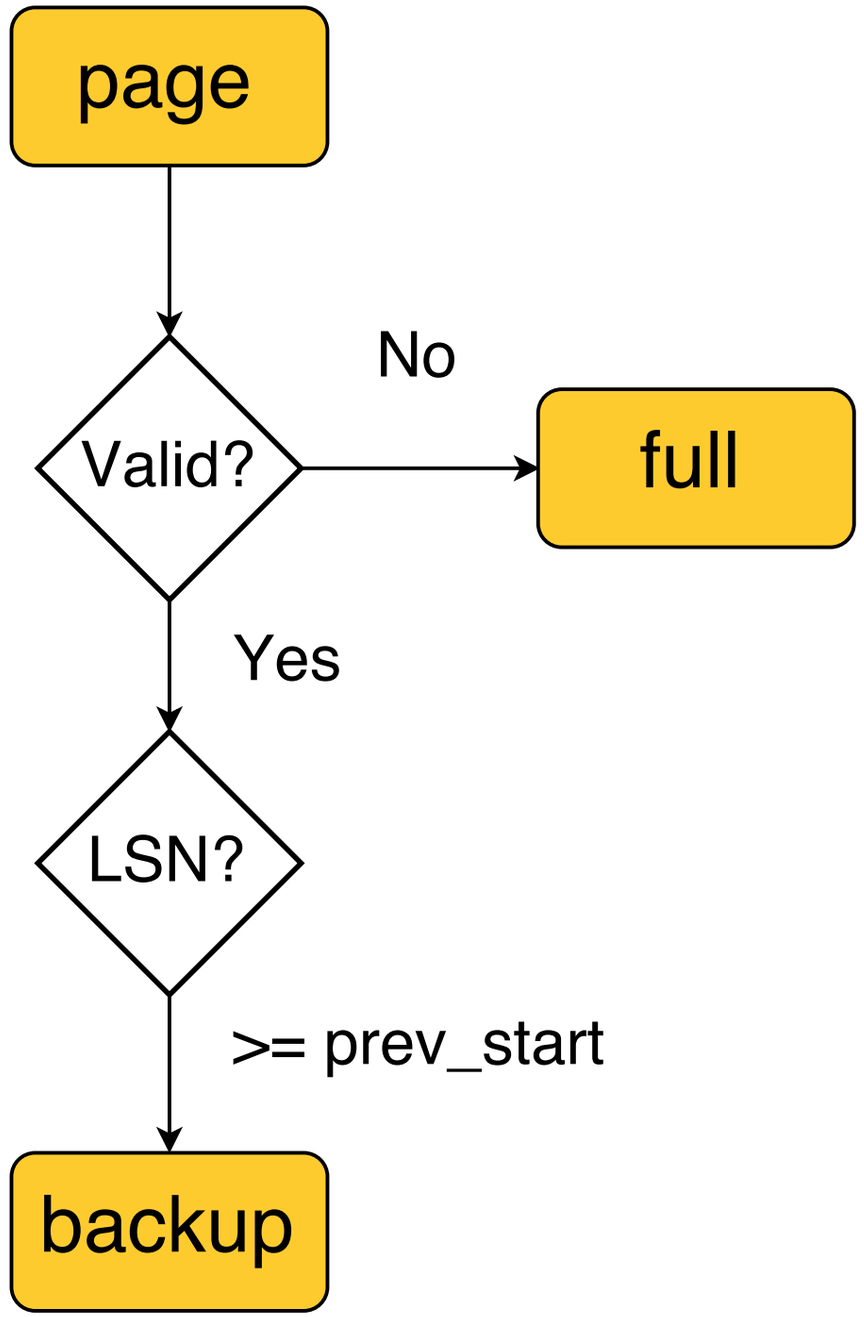
〉 Возможна последовательность инкрементальных бекапов

〉 Восстанавливаться долго

〉 Занимают меньше места

Инкрементальный бекап

- Читаем каждую страницу и проверяем правильность структуры
- Смотрим на pd_lsn в PageHeaderData (меньше lsn последнего бекапа? => эта страница у нас уже есть)
- К файлу дописываем заголовок со списком номеров страниц



Работает ли наш подход?

```
root@pg-backup03i ~ # barman list-backup
xdb2018 | cut -d' ' -f5-12

Jun 11 04:57:09 2016 - Size: 36.0 GiB

Jun 10 02:55:37 2016 - Size: 32.4 GiB

Jun 9 05:18:38 2016 - Size: 39.9 GiB

Jun 8 04:45:25 2016 - Size: 37.3 GiB

Jun 7 05:25:34 2016 - Size: 39.0 GiB

Jun 6 04:33:07 2016 - Size: 28.8 GiB

Jun 5 07:51:02 2016 - Size: 883.9 GiB
```

Консистентность

- barman restore <db> <bkup_id> <dir>
  - Очень долго копируем WAL (в один поток)
- patch config a bit
  - shared_buffers/fsync
- pg_ctl -D ... start
  - Ждем, пока в PostgreSQL достигнет консистентного состояния
- COPY <table> to /dev/null для всех таблиц
  - Без checksums часть случаев не покрывается

Как попробовать?

- Source: https://github.com/secwall/barman
- Quick and dirty install: https://secwall.me/blog/2016/06/18/barman-incremental-backups/
- Upstream issue: https://github.com/2ndquadrant-it/barman/issues/21

Что дальше?

Merge to upstream

- https://github.com/2ndquadrant-it/barman/issues/21
- За полгода ничего не случилось, подождём еще полгода?
- В результате может быть несовместимо с текущим вариантом

Прототипы

- Block change tracking?
  - Меньше чтений - быстрее бекапы
  - Заметный overhead
- Selective restore?
  - Только database
  - Table/page требуют изменений в самом PostgreSQL

Помечтаем?

> “pg_<you-name-it>backup”

Помечтаем?

> “pg_<you-name-it>backup”

〉 Retention policy

Помечтаем?

- “pg_<you-name-it>backup”
- Retention policy
- WAL archive management

Помечтаем?

- “pg_<you-name-it>backup”
- Retention policy
- WAL archive management
- Multithreaded backup/restore

Помечтаем?

- “pg_<you-name-it>backup”
- Retention policy
- WAL archive management
- Multithreaded backup/restore
- Compression

Помечтаем?

- “pg_<you-name-it>backup”
- Retention policy
- WAL archive management
- Multithreaded backup/restore
- Compression
- Page-level increments

Помечтаем?

- “pg_<you-name-it>backup”
- Retention policy
- WAL archive management
- Multithreaded backup/restore
- Compression
- Page-level increments
- Selective restore (database/table/page)

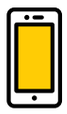
Помечтаем?

- “pg_<you-name-it>backup”
- Retention policy
- WAL archive management
- Multithreaded backup/restore
- Compression
- Page-level increments
- Selective restore (database/table/page)
- Cloud-storage support

Контакты

Евгений Дюков

Системный администратор



+7 (495) 739 70 00, доб. 3508

secwall@yandex-team.ru