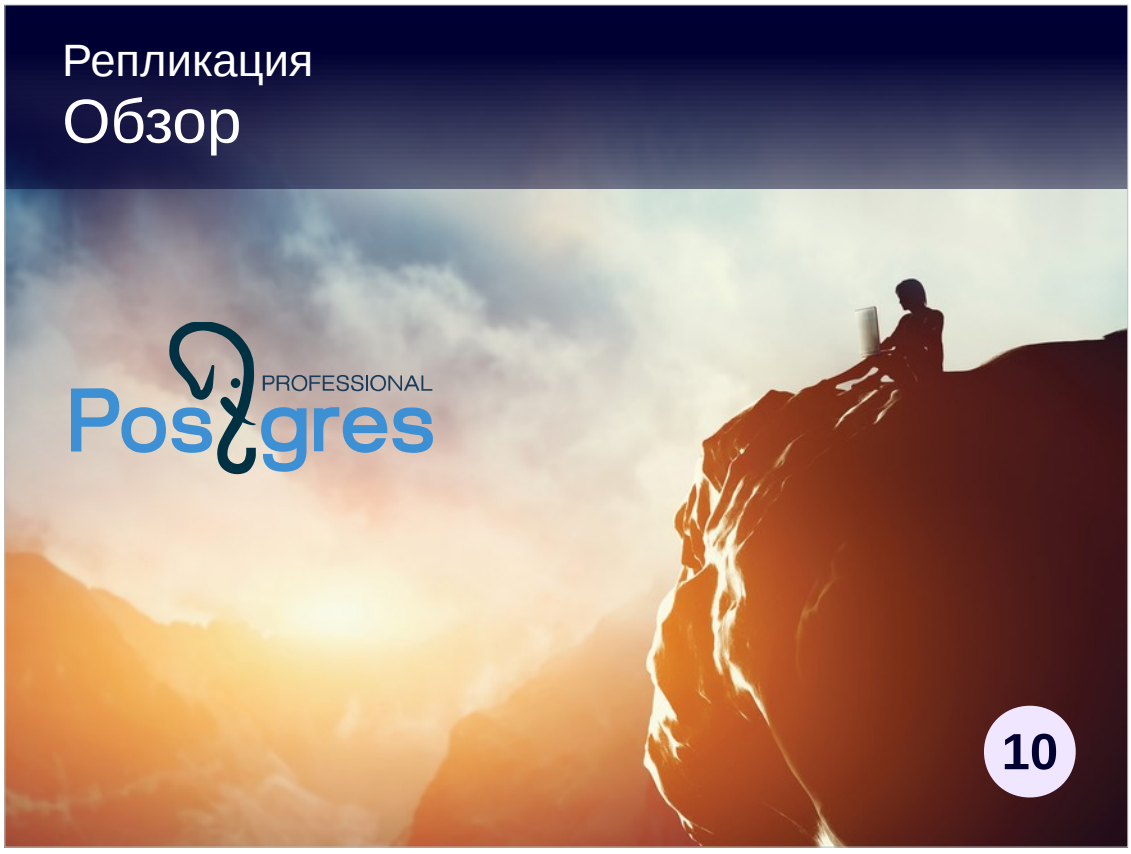


Авторы: Егор Рогов, Павел Лузанов

Обратная связь
Отзывы, замечания и предложения направляйте по адресу:
edu@postgrespro.ru

Темы

Задачи и виды репликации

Физическая репликация

Логическая репликация

Варианты использования механизма репликации

Задачи репликации

Репликация

процесс синхронизации нескольких копий кластера баз данных на разных серверах

Задачи

отказоустойчивость	при выходе из строя одного из серверов система должна сохранить доступность (возможна деградация производительности)
масштабируемость	распределение нагрузки между серверами

Одиночный сервер, управляющий базами данных, может не удовлетворять предъявляемым требованиям.

Во-первых, отказоустойчивость: один физический сервер — это возможная точка отказа. Если сервер выходит из строя, система становится недоступной.

Во-вторых, производительность. Один сервер может не справляться с нагрузкой. Зачастую апгрейду сервера предпочтительна возможность масштабирования — распределения нагрузки на несколько серверов.

Таким образом, речь идет о том, чтобы иметь несколько серверов, работающих над одними и теми же базами данных. Под репликацией понимается процесс синхронизации этих серверов.

Виды репликации

Физическая

- мастер-реплика: поток данных только в одну сторону
- трансляция потока журнальных записей или файлов журнала
- требуется двоичная совместимость серверов
- возможна репликация только всего кластера

Логическая

- поставщик-подписчик: поток данных возможен в обе стороны
- информация о строках (уровень журнала logical)
- требуется совместимость на уровне протокола
- возможна выборочная репликация отдельных таблиц

Взаимодействие нескольких серверов можно организовать по-разному. Два основных подхода, доступных в PostgreSQL — физическая и логическая репликация.

При физической репликации серверы имеют назначенные роли: мастер и реплика. Мастер передает на реплику журнальные записи (в виде файлов или потока записей); реплика применяет эти записи к своим файлам данных. Применение происходит чисто механически, без «понимания смысла» изменений, поэтому важна двоичная совместимость между серверами (одинаковые платформы и основные версии PostgreSQL). Поскольку журнал общий для всего кластера, то и реплицировать можно только кластер целиком.

При логической репликации в журнал добавляется информация более высокого уровня, позволяющая реплике разобраться в изменениях на уровне строк таблиц (параметр wal_level = logical). Для такой репликации не нужна двоичная совместимость, реплика должна лишь уметь декодировать содержащуюся в журнале логическую информацию. Такая репликация позволяет при необходимости проигрывать не все изменения, а только касающиеся отдельных таблиц.

Логическая репликация доступна, начиная с версии 10; более ранние версии должны были использовать расширение pg_logical, либо организовывать репликацию с помощью триггеров.

Физическая репликация

Схема работы репликации

Способы доставки журналов

Использование реплики

Переключение на реплику (и обратно)

Варианты настройки и использования репликации

Сначала познакомимся с физической репликацией.

Механизм ее работы — передача на реплику изменений в виде записей журнала предзаписи. Это очень эффективный механизм, но он требует, чтобы между серверами была двоичная совместимость (основная версия сервера, операционная система, аппаратная платформа).

Физическая репликация всегда однонаправлена: в ней может существовать только один мастер (и произвольное число реплик).

Репликация

Резервная копия

- базовая резервная копия — pg_basebackup
- журналы упреждающей записи — непрерывное архивирование

Непрерывное восстановление

- разворачиваем резервную копию,
  создаем управляющий файл recovery.conf (standby_mode = on)
  и запускаем сервер
- сервер восстанавливает согласованность
  и продолжает применять поступающие журналы
- доставка — поток по протоколу репликации или архив WAL
- подключения (только для чтения) разрешаются
  сразу после восстановления согласованности

Настройка репликации выполняется очень похоже на настройку физического резервного копирования. Отличие в том, что резервная копия поднимается сразу, не дожидаясь поломки основного сервера, и работает в режиме постоянного восстановления (standby_mode = on): все время читает и применят новые сегменты WAL, приходящие с мастера. Таким образом, реплика постоянно поддерживается в почти актуальном состоянии и в случае сбоя мы имеем сервер, готовый подхватить работу.

Если реплика не допускает подключений, она называется «теплым резервом». Однако можно сделать и «горячий резерв» — в процессе восстановления реплика будет допускать подключения для чтения данных (как только будет восстановлена согласованность данных).

В отличие от резервной копии репликация не позволяет восстановиться на произвольный момент в прошлом. Иными словами, репликацию невозможно использовать, чтобы исправить допущенную ошибку (хотя есть возможность настроить репликацию так, чтобы она отставала от мастера на определенное время).

Использование реплики

Допускаются

- запросы на чтение данных (select, copy to, курсоры)
- установка параметров сервера (set, reset)
- управление транзакциями (begin, commit, rollback...)
- создание резервной копии (pg_basebackup)

Не допускаются

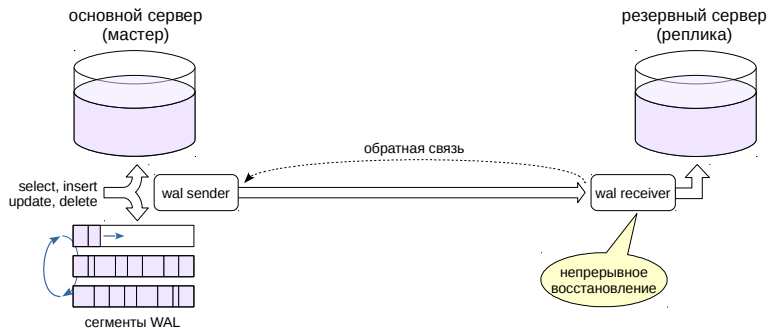
- любые изменения (insert, update, delete, truncate, nextval...)
- блокировки, предполагающие изменение (select for update...)
- команды DDL (create, drop...), в том числе создание временных таблиц
- команды сопровождения (vacuum, analyze, reindex...)
- управление доступом (grant, revoke...)
- не срабатывают триггеры и пользовательские (advisory) блокировки

В режиме горячего резерва на реплике не допускаются любые изменения данных (включая последовательности), блокировки, команды DDL, такие команды, как vacuum и analyze, команды управления доступом — словом, все, что так или иначе изменяет данные.

При этом реплика может выполнять запросы на чтение данных. Также будет работать установка параметров сервера и команды управления транзакциями — например, можно начать (читающую) транзакцию с нужным уровнем изоляции.

Кроме того, реплику можно использовать и для изготовления резервных копий (конечно, принимая во внимание возможное отставание от мастера).

Потоковая репликация

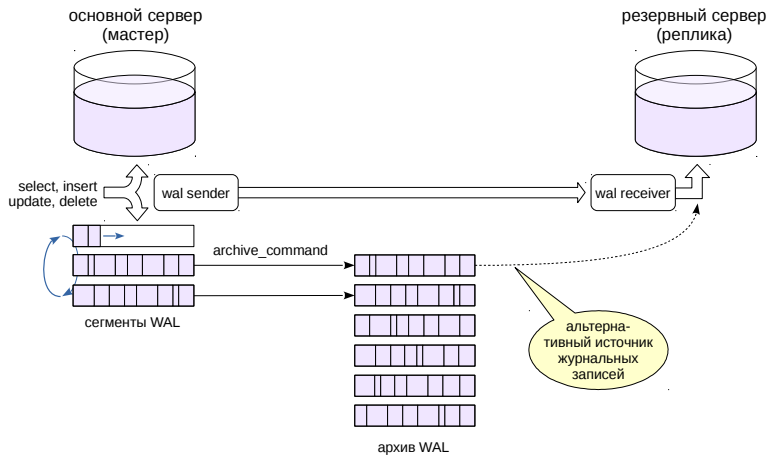


Есть два способа доставки журналов от мастера к реплике. Основной, который используется на практике — потоковая репликация.

В этом случае реплика подключается к мастеру по протоколу репликации и читает поток записей WAL. За счет этого при потоковой репликации отставание реплики сведено к минимуму, и даже к нулю при синхронном режиме.

Тонкий момент. Очистка на основном сервере может удалить версии строк, которые нужны для снимка запроса на реплике. Пострадавший запрос на реплике в таком случае отменяется. Эта проблема решается в случае потоковой репликации с помощью специального механизма *обратной связи:* при этом мастер будет знать, какой номер транзакции нужен для снимка данных на реплике и отложит очистку.

Репликация с архивом



При использовании потоковой репликации есть опасность, что мастер удалит сегмент WAL, данные из которого еще не переданы на реплику. Для уверенности надо либо использовать слот репликации, либо применять потоковую репликацию вместе с архивом WAL (который все равно необходим для организации резервного копирования).

При использовании архива специальный процесс archiver на мастере записывает заполненные сегменты журнала с помощью команды archive_command (этот механизм рассматривается в модуле «Резервное копирование»).

Если реплика не сможет получить очередную журнальную запись по протоколу репликации, она попробует прочитать ее из архива с помощью команды restore_command из recovery.conf.

В принципе, репликация может работать и с одним только архивом, без потоковой репликации. Но в этом случае:

- реплика вынужденно отстает от мастера на время заполнения сегмента;

- мастер ничего не знает о существовании реплики, поэтому очистка может удалить версии строк, нужные для снимка на реплике (можно настроить задержку применения конфликтующих записей, но далеко не всегда понятно, на какое время).

Переключение на реплику

Плановое переключение

останов основного сервера для технических работ без прерывания обслуживания

ручной режим

Аварийное переключение

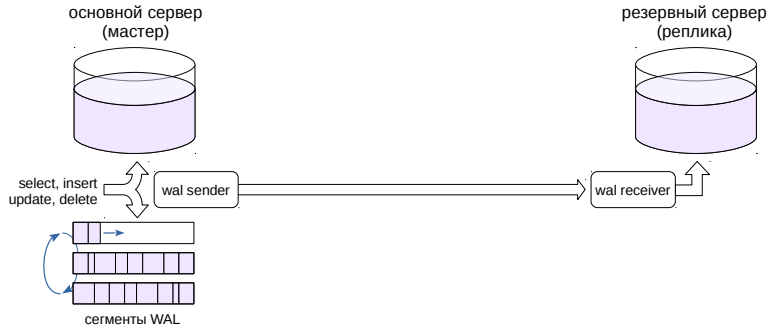
переход на реплику из-за сбоя основного сервера

ручной режим,
но в принципе можно автоматизировать с помощью дополнительного кластерного ПО

Причины перехода на резервный сервер бывают разные. Это может быть необходимость проведения технических работ на основном сервере — тогда переход выполняется в удобное время в штатном режиме. А может быть сбой основного сервера, и в таком случае переходить на резервный сервер нужно как можно быстрее, чтобы не прерывать обслуживание пользователей.

Даже в случае сбоя переход осуществляется вручную, так как PostgreSQL не имеет встроенного кластерного программного обеспечения (которое должно следить за состоянием серверов и инициировать переход).

Переключение на реплику



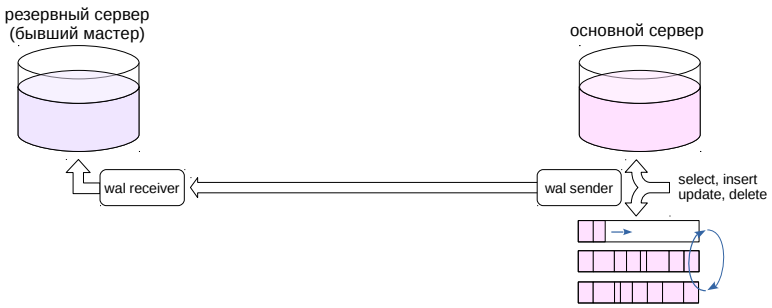
Переход на реплику в картинках: вначале имеются два сервера (мастер и реплика), между ними настроена репликация.

Переключение на реплику

В случае выхода основного сервера из строя или при штатной необходимости работ, реплике дается команда прекратить восстановление и стать самостоятельным сервером, а бывший мастер отключается.

Конечно, требуется способ перенаправить пользователей на новый сервер; но это выполняется внешними средствами вне PostgreSQL.

Восстановление мастера



После восстановления бывшего основного сервера или окончания других работ на нем, он подключается в качестве реплики к новому работающему мастеру.

Восстановление мастера

Простое подключение бывшего мастера — не работает

- проблема потери записей WAL, не попавших на реплику из-за задержки

Восстановление «с нуля» из резервной копии

- на месте бывшего мастера разворачивается абсолютно новая реплика
- процесс занимает много времени (отчасти можно ускорить rsync)

Утилита pg_rewind

- «откатывает» потерянные записи WAL, заменяя соответствующие страницы на диске страницами с нового мастера
- есть ряд ограничений, ограничивающий применение

Если переход на реплику произошел из-за аппаратуры (необходима замена дисков или сервера) или операционной системы (необходима переустановка ОС), то единственный вариант состоит в изготовлении абсолютно новой реплики на новом сервере.

Если же переход был штатным, полезен способ быстро вернуть старый мастер в строй (теперь уже в качестве реплики).

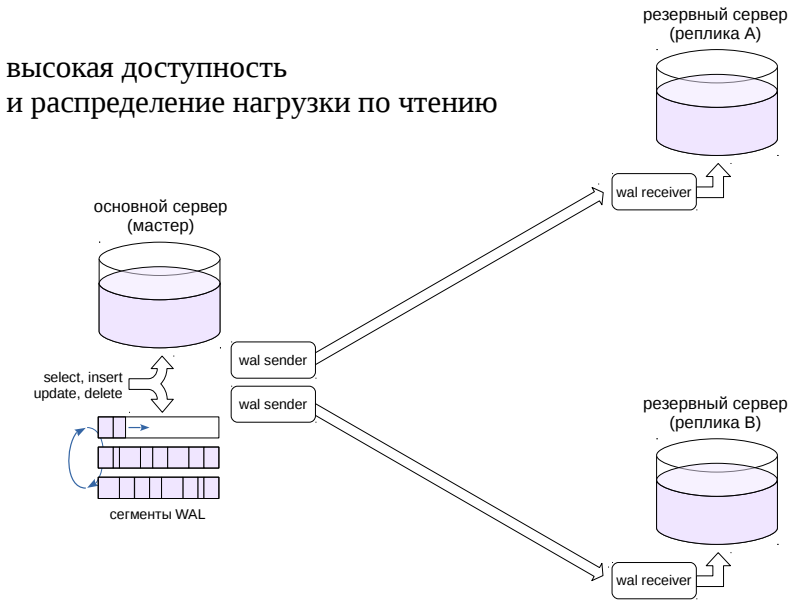
К сожалению, простой способ — просто подключить старый мастер к новому по репликационному протоколу — не работает. Причина в том, что из-за задержек в репликации часть записей WAL могла не дойти до реплики. Если на старом мастере остались такие записи, то применение записей с нового мастера приведет к повреждению базы.

Всегда есть вариант создать абсолютно новую реплику путем изготовления базовой резервной копии. Однако для больших баз данных этот процесс может занимать много времени. Его можно укорить с помощью rsync.

Еще более быстрый вариант состоит в использовании утилиты pg_rewind https://postgrespro.ru/docs/postgresql/10/app-pgrewind (доступна с версии 9.5, для 9.3 и 9.4 есть расширение).

Утилита определяет записи WAL, которые не дошли до реплики (вплоть до ближайшей контрольной точки), и находит страницы, затронутые этими записями. Найденные страницы (которых должно быть немного) заменяются страницами с нового мастера. Кроме того, утилита копирует с сервера-источника (нового мастера) все служебные файлы. Дальше работает обычный процесс восстановления.

1. Несколько реплик



Механизм репликации позволяет сконструировать систему так, чтобы она отвечала предъявляемым к ней требованиям. Рассмотрим несколько типичных задач и средства их решения.

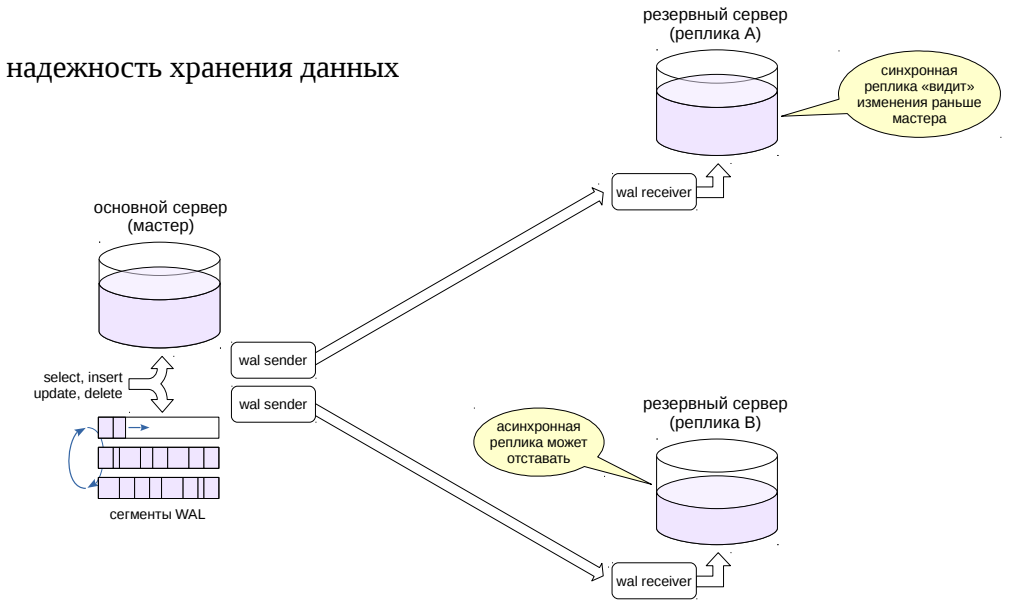
Задача: обеспечить высокую доступность и распределение нагрузки по чтению.

Для решения необходимо иметь мастер-сервер и несколько реплик. Реплики можно использовать для выполнения запросов только на чтение; при сбое основного сервера можно перейти на реплику с минимальным временем простоя.

Надо учитывать, что каждой реплике будет соответствовать отдельный процесс wal writer и, при необходимости, отдельный слот репликации.

Распределение нагрузки по чтению между репликами должно решаться внешними средствами.

2. Синхронная репликация



Задача: в случае сбоя основного сервера, не потерять никакие данные при переходе на реплику.

Решение состоит в использовании синхронной репликации. В случае одного сервера синхронная запись журнала упреждающей записи гарантирует, что зафиксированные данные не будут потеряны при сбое. Аналогичный механизм работает и для репликации: фиксация изменений на мастере не завершается до тех пор, пока не получает подтверждение от реплики. При необходимости синхронностью можно управлять на уровне транзакций.

Синхронная репликация не обеспечивает идеальной согласованности данных между серверами: изменения могут становятся видимыми на мастере и на реплике в разные моменты времени.

Начиная с версии 9.6 синхронизация может происходить с несколькими репликами; в версии 10 доступна синхронизация с учетом кворума.

На иллюстраций реплика Б асинхронная и может отставать; реплика А синхронная. При фиксации изменений мастер выполняет следующее:

- делает запись в журнал упреждающей записи (таким образом, изменение не потеряется при сбое);

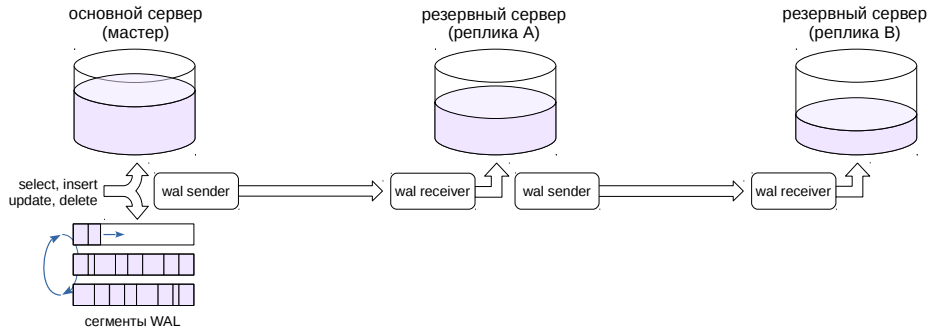
- дожидается подтверждения от реплики о получении записи WAL;

- изменяет состояние транзакции в буфере xact.

Таким образом, запрос к синхронной реплике может увидеть изменения даже раньше, чем запрос к мастеру.

3. Каскадная репликация

несколько реплик
без дополнительной нагрузки на мастер



Задача: иметь несколько реплик, не создавая дополнительной нагрузки на основной сервер.

Задача решается с помощью каскадной репликации, при которой одна реплика передает записи WAL другой реплике и так далее.

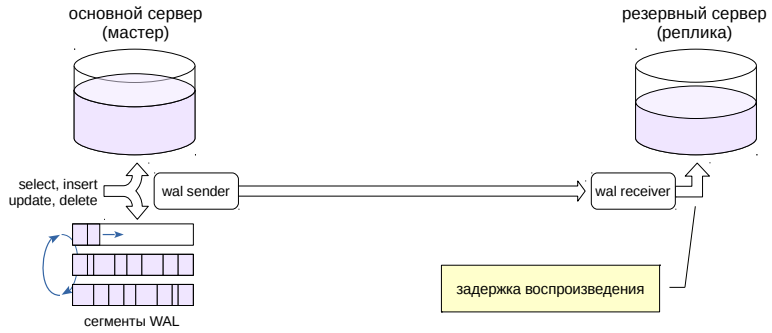
При каскадной репликации не поддерживается синхронизация. Однако обратная связь поступает основному серверу от всех реплик, так что этот функционал работает в полном объеме.

При необходимости переключения следует выбирать ближайшую к мастеру реплику, как наименее запаздывающую.

На иллюстрации: на основном сервере только один процесс wal sender; реплики передают записи WAL друг другу по цепочке. Чем дальше от мастера, тем большее может накопиться запаздывание. Схема мониторинга усложняется: процесс надо контролировать на нескольких серверах.

4. Отложенная репликация

«машина времени»
и возможность восстановления на определенный момент без архива



Задача: иметь возможность просмотреть данные на некоторый момент в прошлом и, при необходимости, восстановить сервер на этот момент.

Проблема в том, что обычный механизм восстановления из архива на момент времени (point-in-time recovery) в принципе позволяет решить задачу, но требует большой подготовительной работы и занимает много времени. А способа построить снимок данных по состоянию на произвольный момент в прошлом в PostgreSQL нет.

Задача решается созданием реплики, которая применяет записи WAL не сразу, а через установленный интервал времени.

Чтобы задержка работала правильно, необходима синхронизация часов между серверами.

Если вывести реплику из режима непрерывного восстановления, остаток записей будет применен без каких-либо задержек.

При использовании обратной связи надо проявлять осторожность, поскольку большая задержка вызовет разрастание таблиц на мастере из-за того, что очистка не будет удалять старые версии строк, которые могут быть нужны реплике.

Логическая репликация

Поставщики и подписчики

Обнаружение и разрешение конфликтов

Варианты настройки и использования репликации

Встроенная логическая репликация доступна в версиях PostgreSQL, начиная с 10. Для более ранних версий аналогичный функционал доступен в расширении pg_logical.

Механизм передачи — включение в физический журнал WAL логической информации об измененных строках. Для этого требуется установка уровня журнала logical.

Другой способ организации логической репликации состоит в использовании триггеров для перехвата изменений и передача этой информации с помощью очереди на реплику. Такой способ, однако, менее эффективен.

При логической репликации у сервера нет выделенной роли мастера или реплики, что позволяет организовать в том числе и двунаправленную репликацию «мастер-мастер».

Логическая репликация

Поставщик

- выдает изменения данных построчно в порядке их фиксации (реплицируются команды INSERT, UPDATE, DELETE)
- возможна начальная синхронизация
- всегда используется слот логической репликации
- параметр wal_level = logical

Подписчик

- получает и применяет изменения
- без разбора, трансформаций и планирования — сразу выполнение
- возможны конфликты с локальными данными

Логическая репликация использует модель «поставщик-подписчик». На одном сервере создается *публикация,* которая может включать ряд таблиц одной базы данных. Другой сервер может создать *подписку* на эту публикацию и получать и применять изменения.

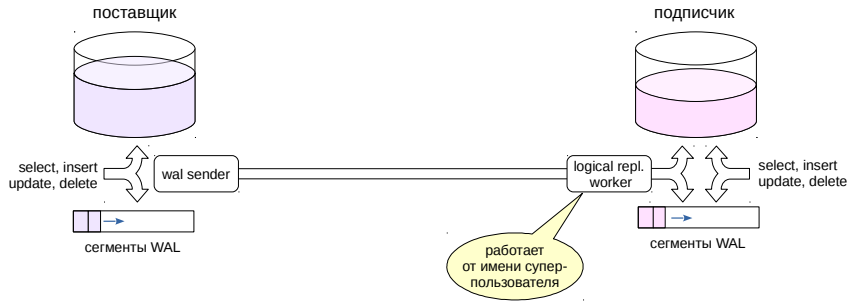
Реплицируются только измененные строки. DDL не передается, то есть таблицы-приемники на реплике надо создавать вручную. Но есть возможность автоматической начальной синхронизации содержимого таблиц при создании подписки.

Технически информация об измененных строках записывается в журнал на мастере; процесс wal sender отсылает подписчику необходимую информацию. Процесс logical replication worker на реплике принимает логическую информацию и применяет изменения. Чтобы гарантировать надежность передачи (отсутствие потерь и повторов), в обязательном порядке используется *слот логической репликации* (похожий на обычный слот репликации).

Применение изменений происходит без выполнения команд SQL и связанных с этим накладных расходов на разбор и планирование, поэтому нагрузка на реплику будет ниже, чем на мастер.

https://postgrespro.ru/docs/postgresql/10/logical-replication

Логическая репликация



На рисунке: фоновый процесс logical replication worker на сервере-подписчике получает информацию от поставщика и применяет ее. В это же время сервер работает обычным образом и принимает запросы и на чтение, и на запись.

Конфликты

Режимы идентификации для изменения и удаления

- столбцы первичного ключа (по умолчанию)
- столбцы указанного уникального индекса с ограничением NOT NULL
- все столбцы
- без идентификации (по умолчанию для системного каталога)

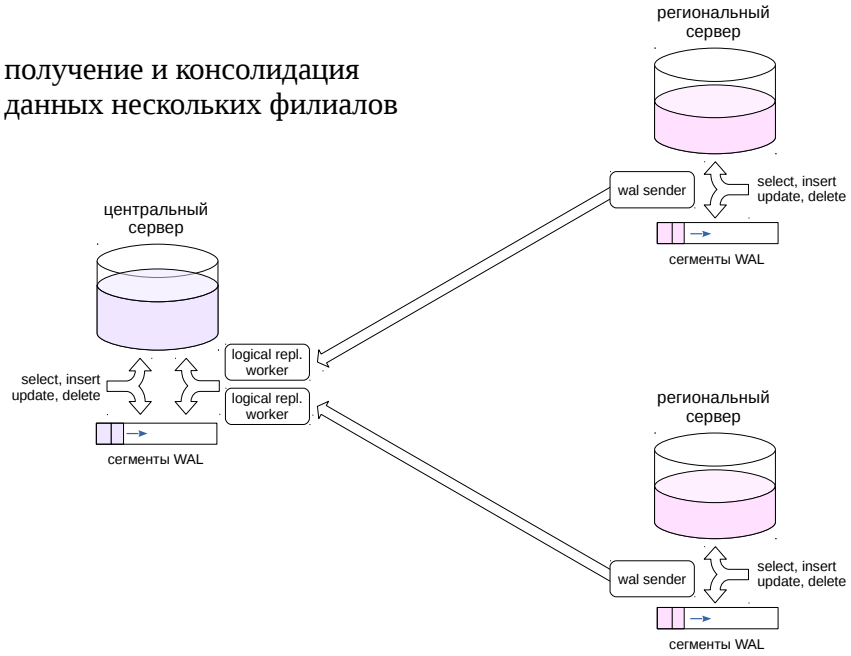
Конфликты — нарушение ограничений целостности

- репликация приостанавливается до устранения конфликта вручную
- либо исправление данных,
- либо пропуск конфликтующей транзакции

Вставка новых строк происходит достаточно просто. Интереснее обстоит дело при изменениях и удалениях — в этом случае надо как-то идентифицировать старую версию строки. По умолчанию для этого используются столбцы первичного ключа, но при определении таблицы можно указать и другие способы (replica identity): использовать уникальный индекс или использовать все столбцы. Можно вообще отказаться от поддержки репликации для некоторых таблиц (по умолчанию — таблицы системного каталога).

Поскольку таблицы на поставщике и на подписчике могут изменяться независимо друг от друга, при вставке новых версий строк возможно возникновение конфликта — нарушение ограничения целостности.
В этом случае процесс применения записей приостанавливается до тех пор, пока конфликт не будет разрешен вручную. Можно либо исправить данные на подписчике так, чтобы конфликт не происходил, либо отменить применение части записей.

1. Консолидация



Рассмотрим несколько задач, которые можно решить с помощью логической репликации.

Пусть имеются несколько региональных филиалов, каждый из которых работает на собственном сервере PostgreSQL. Задача состоит в консолидации части данных на центральном сервере.

Для решения задачи на региональных серверах создаются публикации необходимых данных. Центральный сервер подписывается на эти публикации. Полученные данные можно обрабатывать с помощью триггеров на стороне центрального сервера (например, приводя данные к единому виду).

Такая же схема, развернутая наоборот, позволяет, например, передавать справочную информацию от центрального сервера региональным.

Технический аспект: поскольку репликация основана на передаче журнала через слот, между серверами необходимо постоянное соединение, так как во время разрыва соединения центральный сервер вынужден сохранять файлы WAL.

С точки зрения бизнес-логики также есть множество особенностей, требующих всестороннего изучения. В каких-то случаях может оказаться проще передавать данные пакетно раз в определенный интервал времени.

На иллюстрации: на центральном сервере работают два процесса приема журналов, по одному на каждую подписку.

2. Обновление серверов

обновление основной версии
без прерывания обслуживания

24

Задача: обеспечить обновление основной версии сервера без прерывания обслуживания клиентов.

Поскольку между разными основными версиями нет двоичной совместимости, физическая репликация не помогает. Однако логическая репликация дает возможности для решения этой задачи.

Как обычно, требуются внешние средства для переключения пользователей между серверами.

Вначале создается новый сервер с желаемой версией PostgreSQL.

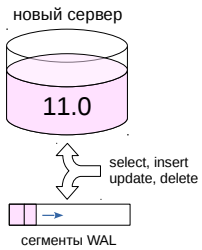
2. Обновление серверов

обновление основной версии
без прерывания обслуживания

Затем между серверами настраивается логическая репликация всех необходимых баз и данные синхронизируются. Это возможно благодаря тому, что для логической репликации не требуется двоичной совместимости между серверами.

2. Обновление серверов

обновление основной версии
без прерывания обслуживания

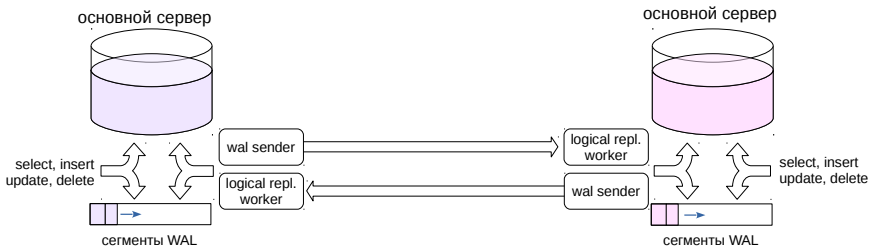


После этого клиенты переключаются на новый сервер, а старый выключается.

На самом деле процесс обновления с помощью логической репликации гораздо более сложен и сопряжен со значительными трудностями. Несколько подробнее он рассматривается в курсе DBA2, тема «Обновление сервера».

3. Мастер-мастер

кластер, в котором
данные могут изменять несколько серверов



27

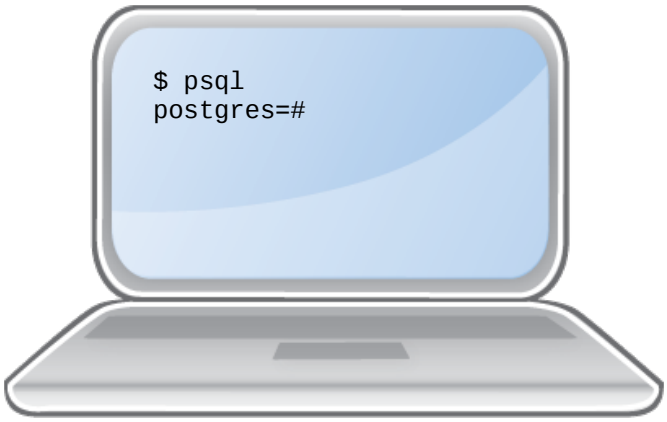
Задача: обеспечить надежное хранение данных на нескольких серверах с масштабированием нагрузки на запись.

Обычная физическая репликация позволяет масштабировать нагрузку на чтение. Логическая репликация позволяет изменять данные одновременно на нескольких серверах. При этом прикладная система должна быть построена таким образом, чтобы избегать конфликтов при изменении данных в одних и тех же таблицах. Например, гарантировать, что разные серверы работают с разными диапазонами ключей.

Надо учитывать, что система мастер-мастер, построенная на логической репликации, не обеспечивает выполнение глобальных распределенных транзакций. При использовании синхронной репликации можно гарантировать надежность, но не согласованность данных между серверами. Кроме того, никаких средств для автоматизации обработки сбоев, подключения или удаления узлов из кластера и т. п. в PostgreSQL не предусмотрено — эти задачи должны решаться внешним средствами.

На иллюстрации: в конфигурации мастер-мастер каждый из серверов создает публикацию и подписку. Между ними происходит двусторонний обмен журнальными записями. Заметим, что в настоящее время средства, доступные в PostgreSQL 10, не позволяют реализовать такую двунаправленную репликацию, но рано или поздно эта возможность должна будет появиться в ядре (см. расширения pg_logical https://www.2ndquadrant.com/en/resources/pglogical/ и BDR https://www.2ndquadrant.com/en/resources/bdr/).

Демонстрация



Итоги

Механизм репликации основан на передаче журнальных записей на реплику и их применении

- трансляция потока записей или файлов WAL

Физическая репликация создает точную копию всего кластера

- однонаправленная
- требует двоичной совместимости

Логическая репликация передает изменения строк отдельных таблиц

- разнонаправленная
- совместимость на уровне протокола

Практика

1. Настройте физическую потоковую репликацию между двумя серверами в синхронном режиме.
2. Проверьте работу репликации. Убедитесь, что при остановленной реплике фиксация не завершается.
3. Выведите реплику из режима восстановления.
4. Создайте две таблицы на обоих серверах.
5. Настройте логическую репликацию первой таблицы от одного сервера к другому, а второй — в обратную сторону.
6. Проверьте работу репликации.

1. Для этого на мастере установите параметры:

- synchronous_commit = on,

- synchronous_standby_names = 'replica',

а на реплике в файле recovery.conf в параметр primary_conninfo добавьте «application_name=replica».