Парфенов Ю.П., Девятериков Д.А.

МАСШТАБИРУЕМОЕ И ОТКАЗОУСТОЙЧИВОЕ DBASS ХРАНИЛИЩЕ ИЗ ЛИНЕЙКИ ПРОДУКТОВ POSTGESQL

Аннотация: *Мировая тенденция в разработке прикладного программного обеспечения ориентирована на переход к облачным SaaS-проектам. Для хранения данных в облачных приложениях на смену традиционным серверам баз данных приходят сервисы Database as a Service (DBaaS) – база данных как услуга. Сервис хранения предоставляет прозрачный доступ к системе управления базами данных (СУБД) избавляя пользователя от многих задач администрирования данных. Рынок предлагает немало облачных файловых хранилищ корпоративных данных. Однако в задачах автоматизации бизнеса по-прежнему наиболее востребованы реляционные базы данных. А использование SaaS-приложений в малом и среднем бизнесе выдвигает на первый план сокращение стоимости продукта и приводит к выбору в пользу свободно распространяемых СУБД. Среди таких продуктов, расширенной функциональностью выделяется реляционная СУБД PostgeSQL. В предлагаемой работе рассматриваются способы построения масштабируемого и отказоустойчивого DBaaS-хранилища из линейки продуктов PostgeSQL.*

Ключевые слова: *СУБД, файловые хранилища, корпоративные данные, потоковая репликация баз, оперативный резерв данных, кластер, пользователь, надежность, нагрузочное тестирование, транзакции*

Варианты архитектуры DBaaS-хранилища

С целью ускорения разработки и удешевления хранилища для его реализации предпочтительно использовать проверенные непроприетарные решения. Открытость PostgreSQL привела к большому числу продуктов, расширяющих его функции. В этом же направлении развивается сам PostgreSQL. Так в последней версии PostgreSQL v9.1 реализована потоковая репликация баз с одного ведущего (Master) на несколько ведомых (Slave) серверов, которую можно применить для создания оперативного резерва данных.

Среди современных дополнений PostgreSQL с точки зрения обеспечения отказоустойчивости и доступности данных интерес представляет Pgpool. Подключаемый между клиентами и серверами PostgreSQL, Pgpool-II управляет соединениями в общем пуле, поддерживает репликации данных с ведущего на несколько ведомых серверов, выполняет балансировку нагрузки, перенаправляя Select-запросы на ведомые серверы, переключает запросы на ведомый сервер при отказе ведущего.

DOI: 10.7256/2305-6061.2014.1.10549.

Анализ функциональности и практика использования показали, что требования надежности и доступности хранилища могут быть реализованы штатными средствами PostgreSQL и Pgpool. Реализация масштабируемости требует создания специальной структуры – кластера из отказоустойчивых модулей хранения данных. Каждый модуль обеспечивает надежное хранение размещенных в нем баз данных. Масштабируемость обеспечивается добавлением (и исключением) модулей в кластер. Используя выбранные средства совместимые с PostgreSQL структуру модуля, можно создать в следующих вариантах:

- используя собственную репликацию PostgreSQL-сервера настроить конфигурацию из одного ведущего и двух (для надежности) ведомых серверов;
- использовать Pgpool, настроив его взаимодействие с тремя серверами PostgreSQL с реализацией синхронной схемы репликации данных;
- использовать Pgpool, настроив его на управление тремя серверами PostgreSQL. Один в роли Master и два – Slave. Pgpool обеспечивает репликации на ведомых серверах и адресует им Select-запросы.

Поскольку методы управления масштабированием не зависят от структуры отдельного модуля хранения, в системе могут использоваться разные типы модулей. Архитектура масштабируемого кластерного хранилища, использующего разные модули хранения (МХ) со структурой, определенной в первом и третьем варианте представлена на рис.1.

Сервис хранения обеспечивает прозрачный и защищенный доступ к требуемому модулю хранения. Кроме того, сервис хранения ведет и мониторинг нагрузки каждого модуля для масштабирования хранилища. В функции управления хранилищем входит размещение БД в модулях хранения, хранение статистической и текущей информации о подключениях пользователей и нагрузке модулей, диалог с администратором, реализация заданий по добавлению/освобождению МХ.

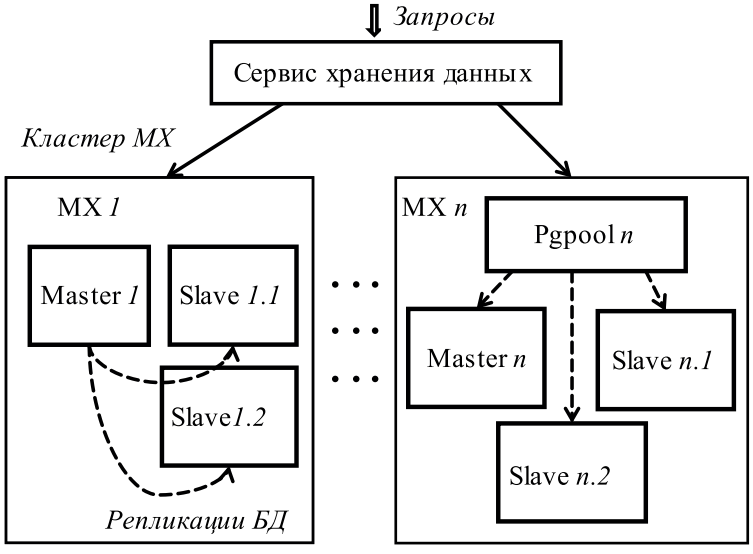

Рис. 1. Кластерное хранилище

DOI: 10.7256/2305-6061.2014.1.10549.

Первоочередной задачей сервиса хранения является слежение за уровнем и балансировка нагрузки отдельных модулей. Решение о подключении или удалении модуля из кластера требует значительных затрат на перемещение и синхронизацию данных и должно основываться на достоверных оценках состояния хранилища и надежного прогноза изменений нагрузки каждого МХ.

Индикатор способности хранилища справляться с потоками транзакций должен:

- быть напрямую связан с критерием качества хранилища для пользователей;
- не требовать хранения большого объема служебных данных и сложной обработки, чтобы не создавать дополнительную нагрузку на серверы;
- допускать надежное прогнозирование для своевременного обнаружения перегрузок.

Естественным индикатором доступности хранилища может служить время ожидания ответа на запрос или исполнения обновляющей транзакции. Времена выполнения отдельных запросов являются непредсказуемыми и интегрально оцениваются средними значениями T(t), вычисляемыми по текущей выборке измеряемых времен ответа на запросы. Пользователем важно, чтобы время ожидания ответа не превышало некоторого заданного порогового значения Тдоп. Поэтому критерий эффективности хранилища для пользователей носит качественный характер и представлен ограничением на среднее значение времени исполнения запроса: $T(t) < Тдоп$. Чтобы убедиться в возможности получения надежной оценки для скользящего среднего времени ответа T(t) на запрос и выбора периода осреднения проведено нагрузочное тестирование PostgreSQL 9.1. Так как было важно проверить возможность получения устойчивых характеристик сервера и характер их зависимости от нагрузки, конкретные параметры используемого оборудования значения не имели. Испытания проводились на стандартном тесте TPC-BX[1], рекомендованном Советом по производительности обработки транзакций для CRM-систем. Тест TPC-B использует базу из четырех таблиц. Наиболее активно используемая таблица содержит 10000000 записей. Генерация смеси TPC-B транзакций выполнялась утилитой pgbench[2] для нагрузочного тестирования PostgreSQL. На сервер PostgreSQL была задана возрастающая нагрузка, которая создавалась разным количеством N подключившихся пользователей N = 40, 80, 120, 160, 200 с продолжительностью непрерывной генерации тестовых транзакций по 600 секунд. Временные параметры каждой транзакции сохранялись в файле и обрабатывались после испытаний программой, разработанной с использованием графических и сглаживающих функций[3]. Средние времена ответа на запрос рассчитывались для разной длины выборки, задаваемой периодами осреднения Θ = 1, 2, 5, 8 и 12 секунд.

Значения средних времен ответа T(t), рассчитанные по запросам, поступившим на Θ = 12 секундном периоде, представлены на рис. 2. Каждая точка графика представляет результат осреднения времен ответа на соответствующем периоде. Каждая ступенька графика соответствует определенной нагрузке N, задаваемой числом от 40 до 200 пользователей.

DOI: 10.7256/2305-6061.2014.1.10549.

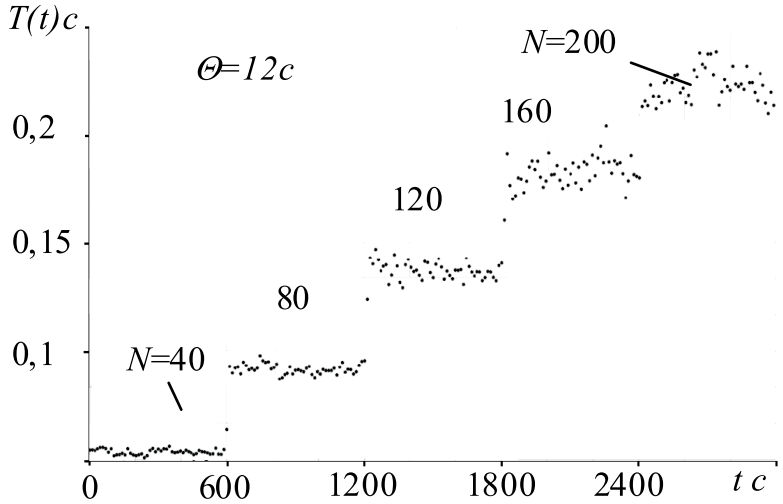

Рис. 2. Оценка влияния нагрузки и интервала осреднения на разброс среднего времени выполнения запроса

Из проведенных испытаний следует, что с ростом нагрузки растет разброс статистического скользящего среднего времени выполнения транзакций T(t). При этом снижение надежности оценки для T(t) можно компенсировать увеличением периода осреднения.

Характер влияния нагрузки на среднее время ответа сервера T(t) представлен на рис. 3. Нагрузка измеряется средним числом – интенсивностью λ(t) транзакций, поступающих на сервер в единицу времени. Точки на рис. 3 соответствуют значениям оценок скользящего среднего времени T(t) для ступенчато возрастающей нагрузки Θ(t). Периоды осреднения Θ = 12 сек. Кривые на рис. 3, представляющие зависимость среднего времени T(t) от нагрузки Θ(t), получены обработкой дискретных значений средних времен методом штрафных регрессионных сплайнов[4].

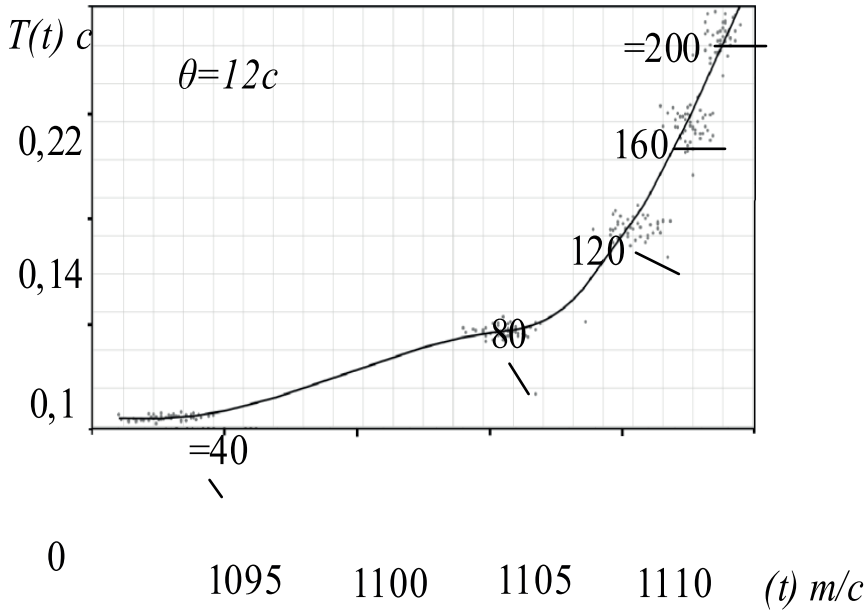

Рис. 3. Влияние нагрузки на среднее время выполнения запросов

DOI: 10.7256/2305-6061.2014.1.10549.

Из проведенного комплекса нагрузочного тестирования следует, что при больших нагрузках средние времена ответа начинают быстро нарастать. Значит, методы оценки и прогнозирования производительности модуля хранения с ростом интенсивности запросов должны одновременно увеличивать как период осреднения, так и частоту контроля среднего времени.

Метод прогнозирования доступности модуля хранения

Фактические значения времен ответа на типовые запросы в CRM-системе при нормальной нагрузке серверов БД должно быть меньше реальных значений допустимых времен ожидания. Это означает, что приближение значения среднего времени ответа сервера к допустимому происходит в области критической нагрузки сервера. Поэтому своевременное обнаружение таких ситуаций, подключение нового модуля хранения и перераспределение нагрузки составляет основную задачу управления кластерным хранилищем. Однако, само по себе приближение нагрузки МХ к критическим значениям недостаточно для принятия решения о необходимости подключения нового модуля, поскольку возрастание нагрузки может носить кратковременный характер. Для принятия правильного решения необходимо учитывать дальнейший прогноз изменения нагрузки. Нагрузка на модуль хранения полностью определяется внешними для хранилища процессами, создаваемыми пользовательскими сервисами. Поэтому прогнозирование тренда нагрузки должно основываться на мониторинге частоты поступления запросов к базам данных, обслуживаемых МХ. Характеристикой нагрузки модуля может служить интенсивность поступающих транзакций $\lambda(t)$, а показателем доступности среднее время ответа $T(t)$. Осреднение должно выполняться на периоде времени с числом запросов, достаточным для получения устойчивой характеристики. При возрастании $\lambda(t)$ растет среднее время ответа $T(t)$, поэтому можно определить такое наибольшее значение $\lambda(t) = \lambda кр(t)$, при котором $T(t) \le Тдоп - Тмасш$, где, Тмасш – запас среднего времени ответа на запросы, необходимый для подготовки и включения нового модуля хранения. Значения $\lambda(t) \ge \lambda кр(t)$ – область критических нагрузок, попадание в которую при дальнейшем росте $\lambda(t)$ требует масштабирования хранилища.

Специфическое влияние на характер нагрузки оказывает ориентация хранилища на корпоративные системы. Бизнес-процессы в таких системах имеют суточные и недельные циклы. Для повышения точности прогнозирования нагрузки и времени обработки транзакций, их мониторинг следует организовать с недельным циклом с сохранением функций $\lambda(t)$ и $T(t)$, $t = 0$ - 24ч. за предыдущие семь дней. Хранение семи функций интенсивности запросов позволит при оценке прогноза изменений нагрузки учитывать динамику ее изменения на данный момент времени в такой же день недели (7 дней назад) с поправкой на тенденцию изменения нагрузки по последним семи дням.

На основе накапливаемых в недельном цикле статистической частоты запросов $\lambda(t)$ и текущей оценки среднего времени ответа $T(t)$ определяется момент наступления критической загрузки, требующей подключение нового МХ.

DOI: 10.7256/2305-6061.2014.1.10549.

Заключение

Высокая надежность, доступность и масштабирование хранилищ данных являются необходимыми условиями эксплуатации SaaS-приложений в корпоративном дата-центре. В работе предложена масштабируемая архитектура, экспериментально обоснованы характеристики, методы оценки и прогнозирования состояния хранилища, использующие PostgreSQL и Pgpool.

В настоящее время продолжаются исследование вариантов архитектур хранилища и ведется разработка прототипа, на котором предполагается отработка методов мониторинга и управления хранилищем. На первом этапе система управления будет предупреждать администратора о приближении к критическим нагрузкам, а реструктуризация выполняться подготовленными сценариями, запускаемыми администратором вручную.

Библиография :

1. The Transaction Processing Performance Council [Электронный ресурс] //TPC-B measures throughput in terms of how many transactions per second a system can perform. [сайт] URL: http://www.tpc.org/tpcb/default.asp (дата обращения: 14.03.2013)
2. PostgreSQL: Documentation: Manuals: [Электронный ресурс] // Appendix F. Additional Supplied Modules F.26. pgbench [сайт] URL: http://www.postgresql.org/docs/devel/static/pgbench.html (дата обращения: 9.04.2013)
3. Dynamic Data Display [Электронный ресурс] URL: http://dynamicdatadisplay.codeplex.com/ (дата обращения: 10.04.2013)
4. Cornell University Operations Research and Industrial Engineering [Электронный ресурс] // Penalized regression splines URL: http://ecommons.library.cornell.edu/bitstream/1813/9131/1/TR001249.pdf (дата обращения: 10.04.2013)

References:

1. The Transaction Processing Performance Council [Elektronnyi resurs] //TPC-B measures throughput in terms of how many transactions per second a system can perform. [sait] URL: http://www.tpc.org/tpcb/default.asp (data obrashcheniya: 14.03.2013)
2. PostgreSQL: Documentation: Manuals: [Elektronnyi resurs] // Appendix F. Additional Supplied Modules F.26. pgbench [sait] URL: http://www.postgresql.org/docs/devel/static/pgbench.html (data obrashcheniya: 9.04.2013)
3. Dynamic Data Display [Elektronnyi resurs] URL: http://dynamicdatadisplay.codeplex.com/ (data obrashcheniya: 10.04.2013)
4. Cornell University Operations Research and Industrial Engineering [Elektronnyi resurs] // Penalized regression splines URL: http://ecommons.library.cornell.edu/bitstream/1813/9131/1/TR001249.pdf (data obrashcheniya: 10.04.2013)

DOI: 10.7256/2305-6061.2014.1.10549.