



Настройка приложений баз данных



Настройкой называется совокупность мер, направленных на приведение прикладной системы в соответствие с требованиями по производительности.

Наш журнал второй раз подряд обращается к семинарской деятельности Московской секции ACM SIGMOD (<http://synthesis.ipi.ac.ru/sigmod/rus>). В предыдущем выпуске был опубликован репортаж (http://www.oracle.com/global/ru/oramag/feb2008/feb-08-oracle_rivkin-14-17.pdf) о январском семинаре, на котором с докладом «Тенденции развития коммерческих СУБД» выступил сотрудник Российского представительства корпорации Oracle М.Ривкин. На февральском же собрании секции был заслушан доклад профессора Санкт-Петербургского государственного Университета Бориса Асеновича Новикова (borisnov@acm.org) на тему «Настройка приложений баз данных», в котором автор изложил свои взгляды на это сравнительно новое направление деятельности в области информатики (аннотация доклада со ссылкой на слайды находится по адресу <http://synthesis.ipi.ac.ru/sigmod/seminar/s20080228>). Конечно, теория и практика баз данных возникли совсем недавно - еще на памяти живущего поколения ИТ-шников, но в философском плане понимании - это вечная тема, если под приложениями понимать наши человеческие работы, заботы, стремления, ...

Прочитываем, поскольку никто не сможет сказать лучше автора, основные тезисы доклада Б.А. Новикова, которые опубликованы по указанному

выше адресу на сайте Московской секции ACM SIGMOD. [А.Бачин - я выделил в этих тезисах наиболее яркие, по моему мнению, фразы.]

1. Настройкой называется совокупность мер, направленных на приведение прикладной системы в соответствие с требованиями по производительности.

Вопреки распространенному мнению, требования к системе по характеристикам производительности являются не менее важными, чем функциональные требования.

2. **Аргументы**, основанные на тенденции быстрого роста доступных вычислительных ресурсов, масштабируемости, параллелизме и т.п., **далеко не всегда справедливы**, потому что, во-первых, **плохо спроектированная система может не обладать** необходимыми для масштабируемости свойствами, и, во-вторых, **заказчик далеко не всегда готов** ждать появления более мощного оборудования.

3. Для того чтобы **настройка** была **эффективной**, она должна выполняться **комплексно** и затрагивать **все компоненты** системы и все стадии процесса проектирования, разработки и сопровождения системы. **Настройка только на уровне сервера базы данных** необходима, однако, обычно



оказывается недостаточной вследствие **усилий программистов**, реализующих код приложения.

4. Наиболее важными фазами, определяющими успех настройки, являются проектирование схемы базы данных, разработка кода приложения и разработка запросов.

5. Выбор логической структуры данных может очень существенно повлиять на производительность системы. **Чрезмерно универсальные решения**, на протяжении десятилетий изобретаемые программистами, неизбежно

приводят к **неприемлемым по производительности** результатам. Правильный выбор логической структуры имеет не меньшее значение, чем построение индексов.

6. Современные методы разработки кода приложений, основанные на использовании каркасов, часто провоцируют **слишком большое количество небольших запросов** к базе данных. Для преодоления этой проблемы рекомендуются правила грануляции запросов, которые приводятся ниже.

7. Как правило, оптимизаторы высокопроизводительных СУБД способны построить вполне эффективные планы выполнения запросов, однако, в некоторых случаях необходимо ручное переписывание запросов. **Использование знаний о предметной области позволяет заменять запросы на формально неэквивалентные, но вырабатывающие идентичные результаты в контексте приложения.** Наиболее важными приемами такого типа являются введение избыточных условий и преобразование операций группировки.



В ходе выступления Б.А. Новиков высоко отозвался о книге Dennis Shasha and Philippe Bonnet. Database tuning: principles, experiments and troubleshooting techniques. Morgan-Kaufmann,

San Francisco, CA, 2002, как о ценном источнике сведений по настройке на уровне сервера базы данных, но которая не рассматривает вопросы прикладного уровня. Б.А. Новиков представил собравшимся свою (в соавторстве с Г.Р. Домбровской) книгу «Настройка приложений баз данных» (изд. BHV, СПб, 2006, 240 стр.), по материалам которой во многом был сделан данный доклад. Краткий обзор этой книги «Как возникают и решаются проблемы производительности» в формате принятых в нашем журнале рецензий публикуется в текущем выпуске «Oracle Magazine/Русское Издание».

Теперь несколько впечатлений и зарисовок по ходу мероприятия. Пересказать двухчасовой доклад, прерываемый вопросами заинтересованных и понимающих слушателей, практически невозможно. И очень жаль, что такие семинары привлекают не так много слушателей, как они того заслуживают, потому что творческая дискуссионная атмосфера учебной аудитории была просто переполнена и новой информацией, и вопросами и ответами по уже всем наболевшим темам. Проблему производительности Б.А.Новиков определил таким положением: «Вследствие неприемлемого времени ответа скапливаются большие очереди». Его доклад изобилует примерами из практики, обобщениями и на первый взгляд парадоксальными, но полученными из производственного опыта, утверждениями.

Например:

- бытовавшая всего несколько лет назад сентенция: «Если производительность недостаточна, подождите год, когда будет новая [имеется в виду более мощная] машина», хороша для университетов и неприемлема для практики, поскольку, очевидно, что кое-кто из заказчиков может просто не дожидаться этого радостного события;

- «Performance is not an issue» [Производительность – это не вопрос] – истина в процессах software engineering, но глубоко ошибочно в

реальности. Книг же по настройке производительности на удивление мало, но много широко известных мифов;

- меры по настройке необходимы на всех фазах создания системы, начиная со спецификации требований. Система, не подготовленная во время разработки, малоэффективна;

- когда координатор проекта заявляет: «Заказчики жалуются на медленность системы. Посмотрите, пожалуйста, какие нужно добавить индексы», слегка подрегулировать систему уже не получится, поскольку процесс настройки далеко не ограничивается созданием индексов, заявленная проблема не сформулирована конкретно, а действия по настройке во время разработки, скорее всего, не предпринимались [А.Бачин: эта не лишняя юмора ситуация описана на стр.48 книги Б.А.Новикова];

- всегда имеет место ограниченность ресурсов, в том числе:

- двухнедельные программисты, то есть окончившие двухнедельные курсы по какому-нибудь языку, не предназначены для получения высокой производительности;
- стоимость и время проектов всегда ограничены;
- управление качеством всегда подключается на заключительных этапах проекта, хотя должно начинаться одновременно с ним, ...

- на ранних фазах проектирования часто имеет место явление пессимизации, что, конечно, не означает сознательного выбора худшего решения, но предпринимаемые действия часто ведут к объективному замедлению работы системы. Сюда можно отнести:

- выбор конфигурации на основе рекламных материалов производителей;
- априорные архитектурные решения;
- опережающее проектирование БД;
- эмоциональный выбор программных средств, то есть не

на основе экспериментов, опыта, анализа, а типа «вышла новая версия, давайте попробуем сделать на ней».

- у разработчиков часто имеет место паническая боязнь SQL, ибо в их представлениях:

- декларативный стиль труден для массового программирования [при подготовке репортажа Б.А.Новиков счел нужным расширить и уточнить это положение следующим образом: «Декларативный стиль программирования, в том числе на SQL, труден для человека - не миф, а мое наблюдение. Его в особенности трудно сочетать с объектно - ориентированными методами, которые по существу императивны. То есть это факт, который нужно учитывать при распределении обязанностей в команде разработчиков». Спасибо, Борис Асенович!];

- учебники по Объектно-Ориентированному программированию рекомендуют использование только примитивных запросов;

- функции СУБД дублируются программистами в коде приложения и др.;

- программисты часто не закрывают курсоры, рассчитывая на сборку мусора самой СУБД, которая закрывает курсоры при завершении сеанса, но сеансы никогда не закрываются, а используются для других приложений. Эта ошибка присутствует в большинстве приложений, написанных на java;

- правила гранулярности запросов:

- количество запросов не должно быть пропорционально числу получаемых строк;

- если из таблицы читаются данные, то все необходимые атрибуты должны быть получены в том же запросе;

- условия селекции должны включаться в запрос, а не проверяться в коде приложения;

- остерегайтесь мифов, связанных с

SQL:

- не следует использовать более 5 операций соединения в запросе (может быть оправданно только в недорогих СУБД);

- представления материализуются при выполнении запросов полностью;

- операции соединения выполняются вложенными циклами (только в недорогих СУБД);

- стратегия настройки SQL реализуется, в частности, посредством:

- получения оценок снизу и сопоставления с требованиями времени выполнения запроса;

- анализа планов и выявления операций, вызывающих неудовлетворительную производительность;

- переписывания запросов;

- построения индексов и материализаций;

- использования подсказок оптимизатору;

- и многое, многое другое.

Вместо заключения профессор Б.А.Новиков высказал такое мнение: «Изобретательность программистов предела не имеет, поэтому задачи настройки возникают почти в каждом проекте». Будем надеяться, что это сказано в адрес скороспелых двухнедельных программистов, к которым читатели нашего журнала не относятся.

После завершения основной части и дискуссии по докладу Б.А.Новиков сделал объявление о запланированной на март 2009 г. в нашей стране, в Санкт-Петербурге совместной 12-й конференции EDBT/ICDT. Дополнительная информация находится по адресу <http://www.math.spbu.ru/edbticdt/>. Сегодня етам совсем немного, но до конференции еще год, и сведения об этом мероприятии будут всё прибавляться. Пока же скажем, что:



- **EDBT (Extending Database Technology)** – международная конференция по расширенным технологиям баз данных) – постоянный престижный форум для обмена новейшими достижениями и исследованиями в организации данных. Проводимая каждые два года в самых привлекательных местах Европы (Венеция, Вена, Кембридж, Авиньон, Валенсия и др.) конференция предоставляет исследователям баз данных, практикам, разработчикам и пользователям уникальные возможности познакомиться с новыми идеями, методами и инструментальными средствами, а также обменяться опытом.

- **ICDT (International Conference on Database Theory** - международная конференция по теории баз данных) - научная конференция по исследованию фундаментальных основ систем баз данных. Она обычно проводится в красивых исторических местах Европы (Рим, Брюгге, Париж, Берлин, Прага, Иерусалим, Лондон и др.).

Следуя успешному опыту SIGMOD/PODS, начиная с 2009г., то есть со встречи в Петербурге, EDBT и ICDT будут проводить ежегодные совместные конференции.

Краткую информацию о EDBT

и ICDT можно прочитать в «Энциклопедии технологий баз данных» (автор М.Р. Когаловский, - М., изд. «Финансы и Статистика», 2002, 800 стр.).

Заканчивая репортаж я хочу выразить надежду и твердую уверенность в том, что плодотворное сотрудничество журнала и Московской секции ACM SIGMOD будет успешно развиваться.

Анатолий Бачин,
главный редактор "Oracle Magazine/
Русское Издание"

Московская Секция ACM SIGMOD (ACM Special Interest Group on Management of Data

– Группа ACM по управлению данными) была учреждена ACM в сентябре 1992 года. Наряду с членами ACM и ACM SIGMOD в работе Секции принимают участие многие представители научно-исследовательских институтов Академии наук и других ведомств, ВУЗов, различных государственных и коммерческих организаций.

В соответствии со статусом ACM, деятельность Секции

преследует исключительно научные и образовательные цели и осуществляется на некоммерческой основе, участие в работе Секции является добровольным и безвозмездным. Тематика работы Секции включает широкий спектр перспективных проблем технологии баз данных и информационных систем.

ACM – Association for Computing Machinery – Международная ассоциация по вычислительному делу

Автоматизация государственного языка

Е.И. Коржов,
директор компании "Текон"

Источник:
журнал "Computerworld Украина" №10 от 11 марта 2008г.,
<http://www.tecon.com.ua/product135>

Одно из наиболее требовательных к ресурсам направлений в автоматизации человеческой деятельности – обработка текстовой информации. Необходимость такой обработки вызвала бурный рост рынка приложений, использующих полнотекстовый поиск в хранилищах электронных документов и документоориентированных базах данных. Способствовало этому росту и развитие Internet/Intranet-технологий, систем документооборота.

Проблемы обработки текста

Необходимо отметить, что хранение и обработка текстовой информации порождает множество проблем, как «организационных» (например, массовое преобразование информации из одного вида в другой: запись стенограмм, распознавание текста и др.), так и «технологических». Последние во многом связаны с тем, что компьютеры изначально были ориентированы на обработку числовой информации, проведение инженерных расчетов для решения задач физики, химии, обороны,

космоса. Соответственно, долгие годы большинство разрабатываемых алгоритмов имели «числовую» направленность. Позже компьютеры стали применять для решения «учетных» задач, что вызвало рост интереса к разработке несложных «текстовых» алгоритмов (сортировка, поиск в таблицах).

Современная деятельность человека становится все более информационно-ориентированной. Все чаще ему требуется найти не только конкретную, но и «похожую» информацию, найти связи (явные и скрытые) между различными фактами, определить правдивость или ложность сведений, выявить «качественные» тенденции, не выразимые в традиционных графиках и диаграммах. Такого рода задачи возникают в самых разных сферах человеческой деятельности, причем в ряде случаев (в здравоохранении, деятельности правоохранительных органов и органов государственного управления) на их решение может отводиться несколько минут и даже меньше.

Одними из самых распространенных инструментов для хранения и обработки информации являются решения компании Oracle, обладающие широкими «текстовыми» возможностями.

Текстовые возможности Oracle

Oracle Text позволяет хранить в СУБД Oracle наряду с «обычными» табличными данными документы и строить запросы к этим документам, так и к информации, хранимой в файлах ОС или на Web-страницах.

Для реализации текстовых возможностей в СУБД Oracle используются индексы специального вида. В Oracle Text используются три вида «текстовых» индексов:

- для полнотекстового поиска по текстовым документам;
- для упрощенного и ускоренного поиска по «каталогам» (1-2 строчным текстовым описаниям);
- для построения «классификаций» документов.





Oracle в мире | Настройка приложений баз данных

Среди возможностей Oracle Text следует выделить:

- работу с документами, хранящимися в СУБД, в файловой системе и в Web;
- поддержку более 150 различных форматов документов;
- богатый язык поисковых запросов, включающий логические связи, поиск с учетом расстояний между словами, использование метасимволов и т.д.;
- мощные лингвистические опции (поиск с учетом словоформ, нечеткий поиск, расширение запросов при помощи тезауруса, тематический анализ, реферирование, классификация, кластерный анализ).

Текстовые возможности Oracle находят применение как в различных прикладных разработках, так и в ряде других продуктов корпорации:

- **Oracle Ultra Search** — средство для организации поиска по разнородным информационным ресурсам;
- **Oracle Internet Application Server** — средство создания Internet/Intranet-приложений (содержит портлет для поиска по документам и формам портала);
- **Oracle Collaboration Suite** — средство организации коллективной работы, включающее электронную и голосовую почту, средства публикации, календарь;
- **Oracle Content Management SDK** — средство управления документами различной природы с возможностью поддержания структуры каталогов, разграничением прав доступа, ведения версий документов.

Новейшая технология Oracle Secure Enterprise Search обеспечивает доступ пользователей к корпоративной и внешней информации, осуществляемый в соответствии с правами, которые установлены администратором по безопасности. При этом все источники информации защищены, а результаты поиска фильтруются таким образом, что пользователю становятся видны только ссылки на разрешенные

ему для просмотра ресурсы.

Все вроде бы хорошо. Но есть одна проблема — многие из перечисленных возможностей работают только для текстов на английском и нескольких других иностранных языках. Для украинского и русского языков Oracle Text реализует только базовые поисковые возможности. Соответственно, ограничены и возможности других продуктов Oracle, поскольку они не задействуют лингвистические технологии, которые зависят от лексики, грамматики и семантики языка.

Решение проблем

Для преодоления указанного недостатка фирмой «Текон» совместно с ее партнерами — «Гарант-Парк-Интернет» (Россия) и «Трайидент Софтвэр» (Украина) — был создан продукт Ukrainian Context Optimizer (UCO). В нем использованы технологии и алгоритмы, которые прошли проверку на практике и успешно работают в самых разных отраслях.

UCO for Oracle позволяет значительно расширить возможности Oracle в украиноязычных базах данных и документов. Этот продукт решает проблему полноценной реализации для украинского языка всех существующих функций Oracle Text:

- поиск документов с учетом всех грамматических форм слов запроса;
- поиск слов, близких по написанию; нечеткий поиск;
- расширение запроса синонимами, более общими и более частными понятиями (на основе тезауруса общей лексики украинского языка);
- поиск документов по ключевым темам, которые автоматически выделяются при индексировании;
- выявление смысловых связей между темами документов, что позволяет расширять или уточнять запрос близкими по смыслу темами;
- составление общих и тематических рефератов документов;
- автоматическое построение иерархических рубрикаторов, классификация документов по

рубрикам;

- поддержка всех кодовых страниц украинского языка.

Интеграция с технологиями и продуктами Oracle.

В UCO есть возможность интеграции с существующими приложениями, использующими Oracle Text, что позволяет расширить их функциональные возможности. Учет особенностей украинского языка, реализованный в UCO, значительно упрощает поиск и обработку документов при помощи продуктов Oracle (Ultra Search, Collaboration Suite, Internet Application Server, Content Management SDK). Разработчики различных решений на базе технологий Oracle, связанных с обработкой текста (например, электронных архивов, систем управления знаниями и систем управления документооборотом), смогут при помощи UCO более полно реализовать возможности поиска и обработки информации.

Интеграция корпоративных архивов.

По оценкам экспертов, до 90% объема корпоративной информации представлено в виде текста. Oracle Ultra Search совместно с UCO позволяет объединить все корпоративные ресурсы в единое информационное пространство.

Аналитические возможности.

Алгоритмы анализа текста и лингвистическое обеспечение, входящие в UCO, позволяют не только искать документы, но и проводить их аналитическую обработку: маршрутизацию и классификацию потока документов (например, ленты новостей, электронной почты), представление тематической структуры исследуемого массива информации, представление динамики изменения тематического поля вокруг интересующего объекта, выявление ассоциативных связей между темами/объектами.