

SYBASE IQ 15.1

Исследования Bloor Research
Автор: Филипп Говард
Ноябрь 2009

Подход, ориентированный на столбцы, которого придерживается Sybase, обеспечит гораздо более высокую производительность при более низкой стоимости по сравнению с традиционными подходами к организации сред анализа, отчетности и хранения данных.

Филипп Ховард

Некоторые факты

Sybase IQ – это поколоночная реляционная база данных, которая была специально спроектирована для аналитических приложений и систем класса Business Intelligence. Ее применение в качестве платформы для хранилища данных может обеспечить ряд весьма существенных преимуществ, в том числе, в области производительности, масштабируемости и совокупной стоимости владения по сравнению с традиционными подходами.

Хотя область применения продукта весьма широка, есть ряд рынков, на которые компания нацеливается в первую очередь. К ним относятся:

- Хранилища данных для компаний-агрегаторов данных, типичным примером которых являются компании, предлагающие услуги доступа к базам данных по большой совокупности клиентов и аналитические сервисы для работы с этой информацией
- Продвинутая аналитика с повышенными требованиями к поддержке сложных и незапланированных запросов, - как в качестве платформы для собственно корпоративного хранилища данных (КХД), так и в качестве аналитического акселератора, когда Sybase IQ служит дополнением для уже существующего КХД.
- В качестве акселератора отчетности для переноса из оперативных систем или централизованных хранилищ данных «тяжелых» отчетов, скорость генерации которых не соответствует требованиям бизнеса.
- Приложения, где хранится большой объем неструктурированных элементов (например, первичные документы и фотоснимки по страховым случаям), а также в которых есть необходимость в течении длительного времени сохранять в он-лайн доступе очень большое количество данных, с возможностью выполнения к ним запросов.

Преимущества Sybase IQ обусловлены используемым подходом к организации БД по столбцам, масштабируемой grid-архитектурой и более высокой производительностью (на порядок выше), чем в базах данных со строковой организацией. При этом для Sybase IQ требуется меньше аппаратных ресурсов. Все эти преимущества становятся особенно очевидными при обработке сложных запросов и сканировании больших таблиц. В последнем случае прямым следствием является отсутствие необходимости в обязательной предварительной агрегации данных, что, по сравнению с традиционными подходами при построении хранилищ данных, означает две вещи. Увеличение производительности и экономию ресурсов на операциях обслуживания базы, и сравнительно меньший размер хранилищ на Sybase IQ (наряду с другими возможностями продукта) это также означает, что Sybase IQ обладает огромным потенциалом с точки зрения производительности при масштабировании системы, т.е. когда число пользователей серьезно возрастает.

Другими словами, Sybase IQ обеспечивает гораздо более высокую производительность при более низкой совокупной стоимости владения. Более того, за исключением поколоночного принципа хранения данных и оптимизированной внутренней архитектуры, в остальном Sybase IQ выглядит аналогично традиционным реляционным БД. Так, например, Вы используете стандартный язык SQL, стандартное аппаратное обеспечение и операционные системы. Схемы БД остаются (или могут оставаться) без изменений, как и приложения. Требования к обучению аналогичны обычным требованиям, и Вы вскоре научитесь добавлять колонку в таблицу так же легко, как и строку и т.п.

Основные выводы

По мнению аналитической компании Bloor Research, потенциальным клиентам следует знать о следующих ключевых моментах.

- В дополнение к поколоночному подходу к хранению данных в Sybase IQ используется ряд специализированных технологий индексирования, которые предназначены для обеспечения еще более высокой производительности при обработке ad-hoc запросов. Эти технологии включают индексы для данных с низкой кардинальностью (что позволяет еще больше сократить требования к хранению и повысить производительность обработки запросов благодаря использованию механизма хранения меток вместо данных), индексы для данных, по которым осуществляется группировка, поиск по диапазону, соединение таблиц, текстового анализа (обеспечение возможности выполнения аналитических запросов к неструктурированным данным, которые могут быть объединены со структурированной информацией), операций сравнения в режиме реального времени для Web-приложений и для анализа данных типа «дата» и «время».
- В Sybase IQ доступны возможности как многопоточной обработки, так и высокой доступности в режиме 24X7 (компания имеет ряд партнерских соглашений с соответствующими поставщиками систем хранения для обеспечения высокой доступности и отказоустойчивости). В частности, выделение отдельных узлов-«читателей» и узлов-«писателей» (где узел - это сервер Sybase IQ в многоузловом кластере IQ Multiplex) позволяют выполнять процедуры чтения и записи параллельно, без влияния друг на друга. В последней версии возможности распараллеливания были существенно расширены, в том числе за счет поддержки нескольких узлов-«писателей», параллельного выполнения запросов и конвейерного распараллеливания (что должно ускорить обработку запросов и загрузку данных). Независимые узлы-«читатели» особенно полезны для компаний-агрегаторов данных, кто предоставляет услуги доступа к своим аналитическим сервисам большому числу клиентов. Узел может быть выделен отдельному клиенту с последующим учетом используемых ресурсов.
- Sybase IQ обеспечивает значительные преимущества в области производительности в сравнении с традиционными базами данных. Помимо уже упомянутых характеристик, она также поддерживает плоские схемы Rcube (Relational Datacube), что может обеспечить серьезные преимущества в сравнении с традиционными схемами «звезда». В частности, плоские схемы Rcube могут значительно ускорить процесс построения хранилища данных, а также улучшить производительность и обеспечить более высокую гибкость. Кроме этого, Sybase IQ позволяет «на лету» вносить изменения в схемы атрибутов (колонки); это означает, что вы можете добавлять/удалять колонки в то время, как сервер Sybase IQ находится в режиме активного использования.
- В Sybase IQ имеются возможности шифрования на уровне столбца и на уровне базы данных. Это особенно важно для агрегаторов данных, предоставляющих услуги большому числу клиентов, когда различные клиентские данные необходимо шифровать по разным алгоритмам. В новую версию Sybase IQ добавлены новые возможности шифрования, в том числе шифрование хранящихся и обрабатываемых данных. Также стоит отметить, что средства администрирования пользователей в Sybase IQ сертифицированы по стандарту Common Criteria (ISO/IEC 15408 EAL3).
- В последнюю версию включена важная функция – поддержка аналитики, встроенной в базу данных, что работает гораздо быстрее, чем в случае традиционных подходов. Это реализуется за счет определяемых пользователем функций, которые обрабатываются, как SQL-функции Sybase IQ. Это означает, что аналитическим процессам доступны все преимущества оптимизатора Sybase IQ и возможности распараллеливания задач. Для использования этих возможностей компания объявила о своем партнерстве с Fuzzy Logix (также ожидаются и другие партнеры).

- Для работы с сервером Sybase IQ используются стандартные ODBC/JDBC/OLE-DB соединения, благодаря чему обеспечивается доступ из любого стандартного BI-инструмента. Система Sybase IQ сертифицирована для работы с ведущими в отрасли средствами, такими как Business Objects, Cognos, MicroStrategy, QlikView, iDashboards, SAS, SPSS, KXEN и другие.
- Кроме того, в Sybase IQ реализованы оптимизированные возможности извлечения, преобразования и загрузки данных (ETL) в форме отдельного дополнительного продукта, который позволяет разработчикам быстро создавать и развертывать наборы данных для анализа в Sybase IQ. Система Sybase IQ также сертифицирована для работы с ведущими в отрасли ETL-инструментами, такими как Informatica, IBM DataStage, Ab Initio, Syncsort и другие. В настоящее время Sybase IQ также поддерживает загрузку данных непосредственно из клиента в обход администратора базы данных, что исключает необходимость его вмешательства в процесс загрузки. Это важно для таких сред, где данные скрываются от администратора из соображений безопасности или конфиденциальности.
- Загрузка в Sybase IQ может также осуществляться в режиме «near real-time» (близко к реальному времени) с использованием инфраструктуры, включающей репликационный сервер от Sybase, стейджинговую базу данных (Sybase ASE или Sybase ASA) и набор скриптов, генерируемых в PowerDesigner. Возможность проводить загрузку данных, не прекращая при этом обработки запросов пользователей, реализуется за счет версионности Sybase IQ – для процесса загрузки создается новая версия данных, в то время как запросы работают на старой версии до момента завершения загрузки.
- Вместе с этой версией Sybase IQ компания представила WorkSpace Data Analytics IDE 2.5 – сопутствующую Sybase IQ среду разработки на базе Eclipse, которая позволяет создавать средства анализа и отчетности для базы данных.
- Управление жизненным циклом информации, в котором поддерживается перемещение данных из хранилища с оперативным доступом в промежуточное хранилище, а затем – в архив, было впервые реализовано в Sybase IQ 15. В последней версии эта функциональность была расширена за счет возможности определения формальных правил хранения данных в рабочей среде WorkSpace.

Резюме

В отличие от стандартных СУБД с построчным принципом хранения данных, которые изначально проектировались для он-лайн обработки транзакций (OLTP), Sybase IQ была спроектирована специально для обработки аналитических запросов и ad-hoc анализа. Таким образом, этот продукт может обеспечить значительные преимущества в области производительности и сокращение совокупной стоимости владения (TCO) по сравнению с традиционными технологиями в средах с большим количеством запросов. Что касается быстродействия, в этой версии данные преимущества были значительно расширены благодаря реализации в базе данных внутренней аналитической обработки.

Очевиднее всего эти преимущества проявятся в средах, где нагрузки по обработке запросов непредсказуемы и где велико количество ad-hoc запросов. Это обусловлено тем, что никакая традиционная база данных не может быть настроена для любых неожиданных запросов. Однако подход хранения данных по колонкам, реализованный в Sybase IQ, обеспечивает эффективную «самонастройку». Кроме этого, сложные запросы, которые требуют выборки из множества таблиц по множеству критериев, а также запросы, требующие сканирования больших таблиц, могут обрабатываться гораздо более эффективно в среде с поколоночной организацией хранения данных. Наряду с этим, Sybase IQ обеспечивает хорошую масштабируемость для больших хранилищ данных, содержащих детальные транзакционные или «субтранзакционные» данные, например, данные о поведении в интернет (clickstream data). Sybase IQ не требует, чтобы данные были преагрегированы для анализа, поэтому пользователи могут эффективно и быстро анализировать данные на атомарном уровне.

Архитектура хранения данных по колонкам, используемая в Sybase IQ, является более эффективной: серверам приходится выполнять гораздо меньше работы для предоставления ответа на конкретный запрос, если данные хранятся по колонкам. Конечно, компания Sybase сделала дополнительные усовершенствования этой архитектуры, но основное отличие заключается именно в этой архитектуре, а так же в сокращении совокупной стоимости владения, которая достигается за счет двух вещей. Во-первых, за счет снижения абсолютной цены, что является результатом улучшенной производительности, и использования технологий сжатия данных, которые Sybase применяет для каждой отдельной колонки. Это приводит к снижению требований к необходимому дисковому пространству и, как следствие, необходимым инвестициям в аппаратное обеспечение. Во-вторых, это сочетается с низкими требованиями к администрированию и настройке, что значительно сокращает необходимые затраты на сопровождение по сравнению с традиционными базами данных других поставщиков.

Однако сегодня на рынке хранилищ данных присутствуют не только поставщики традиционных решений. Высок интерес к поставщикам специализированных программно-аппаратных решений для хранилищ данных (appliances). С одной стороны, это помогло Sybase, поскольку поставило под вопрос однозначное доминирование на этом рынке основных поставщиков баз данных с построчным принципом хранения, поскольку компании получили возможность рассматривать другие доступные опции. Это повысило интерес к Sybase IQ, в частности, как к «акселератору запросов». С другой стороны, интерес к специализированным решениям для хранилищ данных породил новый класс конкурентов, которые заявляют о похожих улучшениях производительности и простоте использования своих решений.

Преимущество Sybase перед этими поставщиками заключается в том, что в Sybase IQ имеются более гибкие возможности настройки. Большинство поставщиков полностью отказываются от индексов или используют их крайне редко. Sybase IQ предлагает широкий спектр возможностей индексирования. Хотя в этом случае увеличивается объем работ по администрированию, продукт становится более адаптируемым к изменениям в объемах и типах данных. Встроенный же «советник по индексам» (index adviser) делает процесс выбора и создания соответствующих индексов более понятным. Более того, в настоящее время не известно ни одной системы хранения данных, которая поддерживала бы текстовый анализ (как это делает Sybase IQ), и возможности управления рабочей нагрузкой смешанных запросов обычно более ограничены, чем в Sybase. Еще одним базовым свойством Sybase IQ является то, что он может легко масштабироваться (причем практически линейно) для обслуживания наборов данных большого объема и большого числа пользователей. Это достигается простым добавлением дисков в общий дисковый пул (в случае роста объема данных) или добавлением новых серверов, «читателей» или «писателей», в кластер IQ Multiplex (в случае роста количества пользователей). Причем такое добавление, дисков или серверов, можно делать постепенно, малыми порциями - по мере необходимости.

В заключении отметим, что Sybase IQ имеет очень хорошую степень конкурентоспособности по отношению к поставщикам как традиционных, так и специализированных технологий. Поскольку Sybase IQ имеет ряд различных преимуществ в различных вариантах использования, мы рекомендуем организациям, планирующим инвестировать в создание хранилищ данных, высокоскоростных аналитических систем и систем класса Business Intelligence, внимательно присмотреться к этому продукту.

ПРОДУКТ

Текущей версией Sybase IQ является версия 15.1. В число дополнительных сопутствующих продуктов, которые были модифицированы в соответствии с функционалом последней версии Sybase IQ, входят система извлечения, преобразования и загрузки данных Sybase ETL, сервер репликации Sybase Replication Server и среда разработки средств анализа WorkSpace Database Analytics IDE. Что касается Sybase ETL, то это решение было оптимизировано специально для использования в сочетании с Sybase IQ и в качестве приемника данных поддерживает только Sybase IQ.

Продукт работает под управлением ОС Windows NT или более поздних версий, Linux (Red Hat и SuSE) и ведущих ОС UNIX от HP, IBM и Sun Microsystems (вскоре будет Oracle).

Логическая архитектура

СУБД Sybase IQ разработана специально для хранилищ данных. Это значит, что она не предназначена для задач обработки транзакций и поэтому не включает возможности, которые могут для этого потребоваться, а оптимизирована именно под задачи хранилищ данных. Это важно, поскольку ведущие коммерческие базы данных предлагают функциональность для задач обоих типов. Таким образом, по сравнению с другими решениями Sybase IQ имеет не такой объемный код и она не так сложна, как эти другие решения.

Еще одно отличие Sybase IQ от других коммерческих баз данных состоит в том, что эта реляционная БД организована по столбцам, а не по строкам. Строковая организация необходима для обработки транзакций, в рамках которых отдельные записи (строки) постоянно добавляются в базу данных и обновляются. Базы данных с ориентацией на столбцы имеют значительное преимущество при обработке запросов, поскольку каждый столбец, фактически, является индексом, но при этом на определение и хранение этих «индексов» дополнительные системные ресурсы не расходуются. Более того, получается, что «проиндексированными» в этом смысле являются абсолютно все столбцы всех таблиц базы, чего просто невозможно добиться при использовании строкового подхода. И все же, Sybase также поддерживает и несколько типов более традиционных индексов, которые могут создаваться дополнительно к базовым индексам-столбцам. Эта возможность рассматривается далее.

Еще одним ключевым преимуществом организации по столбцам является объем данных, который должен быть прочитан при каждом запросе. При обращении к данным в традиционной базе данных происходит полное считывание каждой строки независимо от того, какие поля указаны в запросе. На практике это может привести к тому, что для извлечения всего 20 символов будет считано 3 000 байт. При считывании же данных по столбцам извлекаются только данные, необходимые для выполнения конкретного запроса. Разумеется, различия в производительности будут ничтожны, когда речь идет о чтении одной записи. Однако во многих случаях приходится полностью сканировать таблицы. Если умножить одну операцию чтения на несколько миллионов строк в таблице, различие в производительности будет весьма ощутимым.

Еще одним следствием использования подхода с ориентацией на столбцы является то, что обычно не приходится применять традиционное горизонтальное разбиение на партиции, как в случае строковой организации. Вместо этого в Sybase IQ реализовано вертикальное разбиение по столбцам, а не по строкам. Одно из преимуществ этого подхода заключается в том, что партиции никогда не становятся несбалансированными, поскольку в каждом столбце таблицы всегда будет постоянное число полей. Это значительно снижает требования к управлению партициями и должно устранить необходимость в реорганизации базы данных, которая возникает, когда традиционные партиции становятся несбалансированными, что отрицательно сказывается на производительности.

Кроме того, столбцы легко сжимаются, поскольку для данных разного типа можно использовать разные (и более им подходящие) алгоритмы сжатия. В результате в этом плане Sybase имел историческое преимущество перед другими коммерческими базами данных. Теперь эти различия постепенно стираются, хотя Sybase IQ все еще обладает здесь определенными преимуществами. Например, большинство конкурентов уже тоже предлагают сжатие индексов. Однако только Sybase заявляет, что размер хранилища данных Sybase IQ никогда не превысит размера сырых данных. Такого невозможно добиться ни в одном конкурирующем продукте.

Физическая архитектура

До выхода Sybase IQ версии 15 физическая архитектура продукта строилась вокруг нескольких SMP-узлов (серверов), где один узел назначался администратором и получал права на запись, а несколько узлов чтения занимались обработкой пользовательских запросов. Теперь эта архитектура несколько изменилась, и продукт поддерживает несколько узлов чтения/записи и несколько узлов только чтения. В первом случае узел может быть назначен узлом чтения или узлом записи в зависимости от обстоятельств. Таким образом, в ходе крупного пакетного обновления в ночные часы все узлы чтения/записи могут работать как узлы записи, а в дневное время они могут работать как узлы чтения. Кроме того, систему можно масштабировать по мере необходимости, добавляя в нее новые узлы.

Помимо гибкости узлов чтения/записи и возможности масштабирования за счет простого добавления новых узлов, Sybase также сделал ряд новых функций для загрузки данных. Во-первых, теперь реализовано конвейерное распараллеливание, так что если для индексов имеются самостоятельные структуры данных (это относится к индексам типа "Word" и "High Group" – см. далее), они могут обновляться одновременно с загрузкой данных. Обратите внимание, что распараллеливание поддерживается только на уровне отдельно взятого узла, поэтому вы не получите значительного увеличения скорости загрузки, загружая одну и ту же таблицу с нескольких узлов, за исключением того, что вы снова можете задействовать конвейерное распараллеливание.

Исторически сложилось, что во многих средах данные загружались на сервер, а затем администратор базы данных переносил их в хранилище. Это означает, что администратор видит данные, что во многих средах неприемлемо (например, если работы по сопровождению хранилища передаются сторонней организации), поэтому многие клиенты требуют загружать данные сразу в хранилище. Чтобы удовлетворить это требование, Sybase добавил новую возможность «загрузки с клиента» – данные и крупные объекты (LOB) могут загружаться через ODBC, JDBC, ADO.Net, OLE DB и DBLib.

Еще одной новой функцией является поддержка управления жизненным циклом информации (ILM). Как показано на рис. 1, теперь имеются особые функции, позволяющие работать с разными типами архивных хранилищ: к активным данным добавляется возможность промежуточного и архивного хранения. В целях соблюдения требований регулирования эти хранилища могут быть определены как хранилища с доступом только на чтение, и каждому хранилищу можно назначить разные политики безопасности. Предусмотрены специальные средства хранения в течение установленного периода времени, когда вначале данные помечаются как доступные только для чтения, а затем удаляются.

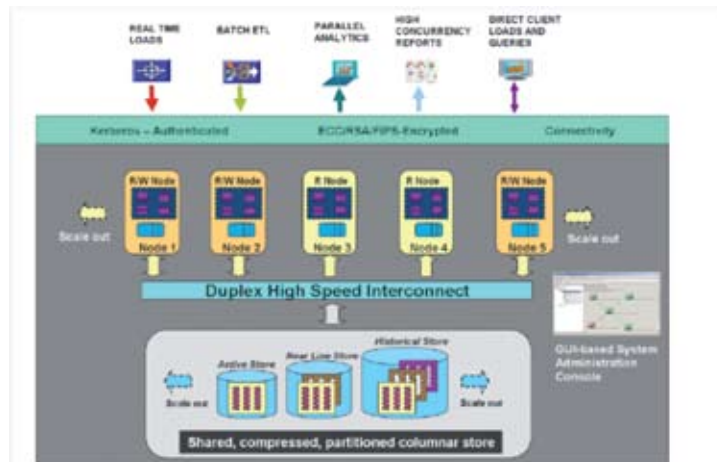


Рисунок 1. Sybase IQ 15 Architecture

Следует отметить, что помимо нескольких узлов чтения и чтения/записи может существовать несколько узлов хранения. Эта архитектура реализует концепцию виртуальных витрин данных, благодаря чему среда масштабируется проще, чем при наличии выделенных витрин данных, взаимодействующих с хранилищем.

Важно подчеркнуть еще одно преимущество этой архитектуры, которое представляет интерес для агрегаторов данных и торговых посредников (целевой рынок Sybase): каждый клиент может обладать собственными узлами чтения и чтения/записи, не связанными ни с какими другими. Это способствует очевидному повышению уровня безопасности и позволяет учитывать расходуемые ресурсы. Также, это позволяет для разных пользователей определить разные уровни обслуживания. Более того, Sybase IQ позволяет также шифровать данные на уровне отдельных столбцов, что еще больше повышает безопасность отдельных клиентов. Фактически, сейчас продукт поддерживает три уровня шифрования: для защиты данных при передаче по сети используется шифрование RSA и шифрование повышенной защищенности ECC, а для шифрования данных внутри базы используется RSA или RSA с поддержкой FIPS 140-2 (протокол повышенной защищенности).

Если какой-либо узел выйдет из строя, его пользователей или его роль («читатель» или «писатель») можно переключить на другой узел. Узлы поддерживают горячую замену, аварийное переключение и балансировку нагрузки. Эти функции не автоматизированы и управляются администратором базы данных, который имеет возможность регулировать динамическое распределение ресурсов в соответствии с потребностями бизнеса. Кроме того, можно воспользоваться приложением балансировки нагрузки OpenSwitch, которое работает на сервере приложений. В Sybase ETL также имеются функции балансировки нагрузки, предназначенные для задач загрузки данных в Sybase IQ. Следует также упомянуть о сотрудничестве компании с рядом производителей систем хранения с целью получения решений с повышенным уровнем высокой доступности (high availability) и восстанавливаемости после аварий (disaster recovery). В этой версии также был переделан механизм внутреннего сетевого взаимодействия между узлами кластера IQ Multiplex, что обеспечило более устойчивый и надежный мониторинг активности узлов, позволяя непрерывно отслеживать состояние узлов и сигнализировать обо всех изменениях. Несколько важных новых возможностей появилось также на уровне хранения данных (SAN). В частности, появился новый уровень абстракции дискового устройства - DBFile, и новое отношение в иерархии хранения - DBSpace/DBFiles, что позволяет размещать отдельные таблицы, индексы, колонки или партиции на отдельных DBSpace – для лучшего управления ресурсами. Появилась возможность создания партиций по диапазону, и вы можете создавать, изменять, объединять переименовывать, разделять и удалять партиции так, как это необходимо.

Также для Sybase IQ существует уникальная HA-DR методология, называемая NonStopIQ, для которой обычно используется два дисковых массива SAN – локальный и удаленный,

связанных между собой синхронным или асинхронным соединением. Большим преимуществом этого решения является не только то, что оно обеспечивает защиту от аварий (disaster recovery), но также то, что оно вообще исключает необходимость выключения системы из работы - даже на время планового обслуживания. Стоит отметить, что все больше компаний внедряют средства оперативного бизнес-анализа и функции обработки запросов в рабочие приложения, вследствие чего хранилище приобретает такое же ключевое значение, что и эти приложения.

Производительность запросов

Хотя каждый столбец, фактически, является индексом, использование специальных индексов во многих случаях обеспечивает ощутимые преимущества. Это как раз та область, где Sybase значительно превосходит других поставщиков. Несомненно, секрет Sybase IQ заключается в возможностях индексирования. По мере возникновения новых потребностей в анализе в Sybase можно просто создавать новые типы индексов. Прелесть этого подхода заключается в том, что новые индексы добавляются в хранилище данных с крайне незначительным влиянием на архитектуру хранилища и аналитические приложения, которые с ним работают. Sybase IQ предлагает целый ряд различных способов индексирования:

- **Low Fast индекс:** Индекс с низкой кардинальностью (обычно используются для полей с менее чем 1 500 уникальных значений). Для его построения используется процесс «токенизации» (создания меток «токенов»), - нецелочисленные данные преобразуются в метки (в случае целочисленных данных число становится собственной меткой), а затем сохраняются метки, а не данные. Это особенно полезно для сокращения объема избыточных данных и экономии дискового пространства. После того, как метки созданы (этот процесс выполняется автоматически) создается битовый (bitmap) индекс, ссылающийся на эти метки.
- **Bit-Wise индекс:** Битовый индекс для полей с высокой кардинальностью, где количество возможных значений превышает 1500 (например, денежные значения). Sybase IQ использует запатентованную технологию, известную как Bit-Wise индексирование. Это удобно, когда нужно выполнить вычисление с набором условий, например, найти общую выручку и количество проданных элементов по цене менее 50 фунтов.
- **High Group индекс:** Фактически это классический B-tree. Однако здесь принципиальное отличие заключается в том, что этот индекс создается, только если ожидается использование столбцов в группировках, - в частности для комбинации механизмов поиска с низкой и высокой кардинальностью. Например, это может быть запрос о количестве проданных единиц продукта и сумме (высокая кардинальность) в группировке по магазинам (низкая кардинальность). Теперь High Group индекс обрабатывается в многопоточном режиме.
- **Fast Projection индекс:** Создается по умолчанию, и представляют собой просто сам столбец. Если пользователь планирует всегда запрашивать столбец целиком, то, поскольку данные хранятся по колонкам, то содержимое столбца может быть прочитано (или, в терминологии IQ - «спроецировано») в отчет или в запрос без необходимости создания каких-либо дополнительных индексов. Это удобно, например, в операторе WHERE. В последней версии возможности индексирования этого типа были расширены: теперь поддерживается размер индекса 3 байта, т.е. можно определить 16 млн уникальных значений, что повышает производительность и эффективность сжатия больших объемов данных.
- **Word Индекс:** Это возможность создания текстового индекса на основе ключевых слов или строки поиска. В последней версии эти индексы сделаны многопоточными.
- **Compare Индекс:** Эта методика индексации позволяет выполнять сравнение столбцов данных (одной и той же таблицы), эквивалентное выражению «если... то... иначе».

Например, «если расходы выше, чем доходы, то...». Этот тип индексов особенно полезен для выполнения сравнений в реальном времени, - в веб-приложениях.

- Join индекс: Как видно из названия, эти индексы предназначены для устранения необходимости объединения таблиц. Как и в случае многих других поддерживаемых индексов, данный тип будет наиболее полезен, когда требования к запросу могут быть определены заранее. В этой версии скорость Join индексов существенно повысилась благодаря параллельному выполнению запросов, что позволяет параллельно сканировать столбцы одновременно и затем также параллельно проводить объединение.
- Time Analytic индекс: предоставляется возможность создавать индексы на основе даты, времени или даты и времени. Следует отметить, что запросы на основе времени особенно сложно обрабатывать в традиционных реляционных базах данных.

Поддерживается ряд расширенных функций, позволяющих использовать эти индексы в самых различных обстоятельствах. К ним относятся сжатие индексов для снижения требований к дисковому пространству (или памяти – битовые массивы могут кэшироваться), совместное использование индексов разных типов, конвейерное распараллеливание для команд GROUP BY и ORDER BY, hash и merge join, и возможность фильтрации битовых массивов с помощью логических операторов, таких как AND и OR. Благодаря этим возможностям индексирования, Sybase IQ позволяет устранить целый ряд традиционных недостатков битового отображения. В частности, то, что оно не подходит для операций объединения таблиц (Join) или операций агрегации данных. Также стоит упомянуть, что хотя Sybase работает достаточно быстро, чтобы устранить необходимость предварительной агрегации данных для обработки на основе OLAP (что является весомым преимуществом с точки зрения администрирования), Sybase IQ поддерживает OLAP-функции такие как, как ранжирование, окна разбиения, взятие процента и вычисление среднего значения. Также необходимо отметить, что в Sybase IQ имеется встроенный Советник по Индексам (Index Adviser), который помогает администратору определять целесообразность добавления нового индекса и его тип.

К другим средствам повышения производительности относятся предикаты с проталкиванием вниз (pushdown) и оптимизация подзапросов (связанные подзапросы, разделение подзапросов и автоматическое упрощение структуры запросов)

Встроенная аналитика (in-Database Analytics)

В традиционных средах функции интеллектуального анализа данных (data mining) выполняются за пределами базы данных: данные извлекаются из хранилища, а затем обрабатываются соответствующим программным обеспечением традиционным образом. Такой подход удваивает количество проблем. Во-первых, снижается производительность, поскольку присутствует процесс извлечения, а сервер приложений не обладает такими же возможностями распараллеливания, что и сервер аналитики. Чтобы нивелировать потерю производительности, обычно аналитики, выполняющие анализ, используют выборку данных. Хотя такой подход позволяет в некоторой степени избежать ухудшения производительности, точность приносится в жертву. В частности, существенная информация, оставшаяся за границами выборки, может быть легко пропущена.

Аналитика, встроенная в базу данных позволяют решить обе эти проблемы. Поскольку алгоритмы интеллектуального анализа выполняются внутри базы данных, нет необходимости ограничиваться выборкой данных, поэтому точность возрастает. По этой же причине можно воспользоваться преимуществами параллельной работы и другими возможностями сервера аналитики, чтобы оптимизировать производительность.

Встроенная аналитика в Sybase реализована с помощью функций, определяемых пользователем (UDF), которые выполняются как функции SQL и создаются в среде Database Analytics IDE (см. далее). Однако компания не делает эту возможность открытой для клиентов – по крайней мере в настоящее время такие функции могут создаваться только

сертифицированными партнерами. Первым из таких партнеров является компания Fuzzy Logic, чья база данных Lytix для Sybase IQ уже выпущена (пока только на платформах Linux и Solaris). Она поддерживает различные алгоритмы, в том числе нейронные сети, кластеризацию k-средних, моделирование по методу Монте-Карло, линейные и логистические регрессии и т. д. Более подробные сведения см. на веб-сайте Sybase: <http://www.sybase.com/detail?id=1065214>.

Рабочая среда WorkSpace

Второй основной особенностью этой версии является выпуск среды Sybase WorkSpace Database Analytic IDE (версии 2.5), поддерживающей Sybase IQ. Разумеется, в целом WorkSpace поддерживает широкий спектр продуктов Sybase, но с точки зрения Sybase IQ существует три особых компонента:

- PowerDesigner 15.1 – средство моделирования данных, обратного проектирования в целях обеспечения поддержки миграции, моделирование перемещения данных (извлечение, преобразование, загрузка и репликация данных) и моделирование жизненного цикла, реализованное для поддержки возможностей управления жизненным циклом информации (ILM), имеющихся теперь в Sybase IQ. Интеграция PowerDesigner с WorkSpace осуществляется через подключаемый модуль Eclipse.
- DB Developer for Sybase IQ 2.5 – настраиваемая среда разработки для Sybase IQ, в том числе уже обсуждавшиеся функции UDF.
- Open Source BIRT 2.2.2 – часть платформы Eclipse, специально предназначенная для разработки отчетов, графиков и т. д.

Управление рабочей нагрузкой

Стратегия Sybase с точки зрения распараллеливания операций противоречит концепциям большинства конкурентов и, вероятно, определяет обстоятельства, в которых Sybase IQ станет оптимальным решением. Большинство поставщиков прилагают основные усилия к повышению скорости параллельного выполнения отдельных запросов. Sybase, в свою очередь, считает, что продукт Sybase IQ сам по себе обеспечивает оптимизацию отдельных запросов. Таким образом, средства параллельной обработки общего назначения (в отличие от ранее обсуждавшегося конвейерного распараллеливания) направлены на поддержку выполнения как можно большего числа одновременных запросов, а не на повышение скорости выполнения отдельных запросов. В этом случае делается попытка обеспечить скорость обработки, постоянную для всех запросов независимо от их типа и нагрузки. Для этого в данной версии Sybase реализован целый ряд новых функций, в том числе возможность определять степень распараллеливания задач и подзапросов, более эффективное использование временного (TEMP) пространства при большем количестве операций, которые могут работать непосредственно со сжатыми данными без необходимости декомпрессии, а также управление текущей рабочей нагрузкой.

Работа с базой данных

В Sybase IQ входит SQL API, обеспечивающий доступ на базе языка SQL. Этот SQL совместим со стандартом SQL-99 и это тот же самый SQL, что используется в Sybase Adaptive Server Anywhere (за несколькими исключениями). Кроме того, он совместим с синтаксисом, используемым в Sybase ASE (то есть T-SQL), поэтому Sybase IQ может естественным образом использовать большую часть хранимых процедур Sybase ASE. В связи с этим стоит упомянуть, что Sybase IQ и Sybase ASE внешне выглядят весьма похоже. В состав продукта Sybase IQ входит графическая утилита SQL Editor.

Sybase IQ также поддерживает интерфейсы ODBC и JDBC (2.0). В качестве альтернативы Sybase IQ обеспечивает функциональность Java 2. Этот язык может использоваться для написания хранимых процедур и определения пользовательских функций. Однако размещение объектов Java в базе данных не поддерживаются.

Также реализована поддержка XML, в том числе возможность сохранять и извлекать XML-документы, а также экспортировать результаты запросов в формат XML (с внедренным определением типа документа). Кроме того, важно принимать во внимание функциональность веб-служб, имеющуюся в Sybase IQ. В базу данных встроен веб-сервер HTTP(S), позволяющий извлекать данные в формате XML, а также по таким стандартам, как SOAP. Также обеспечивается прямая интеграция с Microsoft Visual Studio .NET через провайдера ADO.NET.

Администрирование базы данных

В Sybase IQ поддерживаются традиционные реляционные схемы, в том числе нормализованные схемы, используемые для обработки транзакций, а также схемы «звезда», «снежинка» и «созвездие» (совокупность звезд), используемые для организации хранения данных.

Кроме того, продукт поддерживает плоские схемы (известные как Rcubes), обладающие целым рядом преимуществ, таких как меньшее количество таблиц (в результате чего уменьшается число объединений и повышается производительность), более низкая сложность и более высокая степень управляемости.

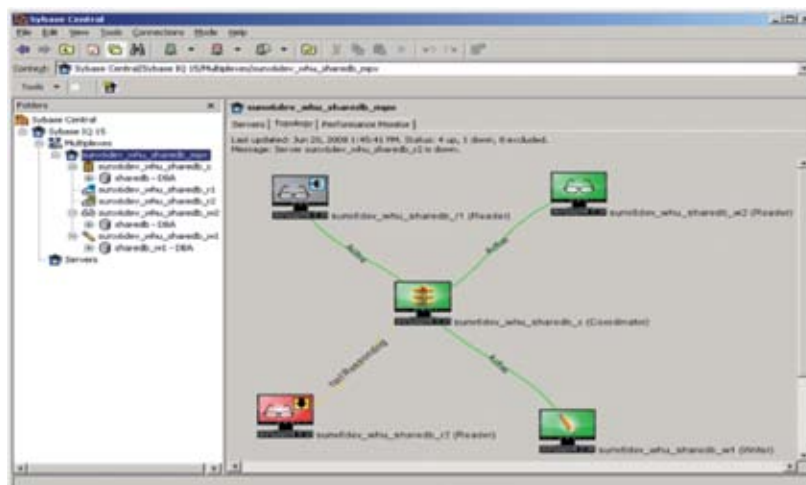


Рисунок 2. Sybase Central

Что касается фактического процесса администрирования, то он выполняется в полностью переработанном средстве Sybase Central, которое представлено на рис. 2, где показана топология среды. Sybase Central поддерживает управление кластерами «одним щелчком мыши», операции обслуживания «на лету» (включая оперативное добавление столбцов), а также графическое представление загрузки процессоров, отдельных потоков, их тайминг, что позволяет легче устранять проблемы. Здесь же определяются параметры безопасности и управление ресурсами, и осуществляется непосредственное управление. При этом все эти возможности в данной версии значительно усовершенствованы.

Как уже упоминалось при рассмотрении управления ресурсами, в этой версии внедрены новые стандарты шифрования. Помимо этого реализована поддержка аутентификации по протоколу Kerberos, которая позволяет использовать общие идентификаторы пользователей и пароли (задаваемые пользователями) в базе данных и операционной системе. Использование LDAP в настоящее время не поддерживается.

КОМПАНИЯ-ПРОИЗВОДИТЕЛЬ

Система Sybase IQ основана на технологии, которую компания Sybase получила в распоряжение после покупки Expressway в 1995 году. В ней всегда делался акцент на масштабирование путем постепенного наращивания мощности - вплоть до очень крупных хранилищ данных (VLDW). В связи с этим необходимо упомянуть о том, что в 2007 году Syb-

ase реализовала VLDW (с использованием Sybase IQ 12.x) объемом один петабайт (1 000 ТБ) совместно с компаниями Sun и BMMsoft. Это было первое хранилище такого размера, которое прошло независимый аудит.

Сейчас Sybase фокусируется на Sybase IQ как на аналитическом сервере в первую очередь для задач предоставления аналитических услуг, продвинутой аналитики и быстрого создания отчетов, о чем уже говорилось выше. Этот подход полностью оправдал себя. Сейчас во всем мире существует свыше 3 000 систем Sybase IQ, развернутых более чем в 1 500 организациях. Более того, доходы от Sybase IQ постоянно растут, так что нет никаких сомнений, что эта стратегия работает. Будет уместно заметить, что значительная доля продаж продукта приходится на организации, в которых не используется Sybase ASE (транзакционная СУБД, основной продукт компании).

Помимо рынков, рассмотренных в начале этого документа, сейчас Sybase активно ведет деятельность в ряде новых интересных областей, где, как ожидает компания, возможности Sybase IQ будут представлять особый интерес:

Управление рисками на рынке ценных бумаг и инвестиций. Sybase RAP The Trading Edition представляет собой продукт, в котором Sybase IQ используется как ядро VLDW, позволяющее соединить в одном хранилище значительные объемы данных реального времени и архивных данных для создания целостного представления о рынке, что необходимо трейдерам, специалистам, отвечающим за управление портфелем продуктов, и ответственным за корпоративные риски.

Соответствие требованиям надзорных органов. Хранение структурированных и неструктурированных данных с созданием отчетности по ним. Зачастую это означает необходимость хранения очень больших, постоянно растущих объемов данных в оперативном режиме (что также может потребоваться для решения аналитических задач). Sybase IQ превосходно для этого подходит. Обратите внимание, что для инвестиционных рынков соответствие требованиям надзорных органов понимается в более широком смысле, и это может быть закрыто с помощью Risk Analytical Platform (RAP). С точки зрения интерфейса и документации система Sybase IQ сама по себе отвечает различным требованиям к безопасности (как уже было показано выше) и доступности, таким как требование 508 (отказ функциональности). Еще одним средством обеспечения соответствия требованиям надзорных органов является возможность зафиксировать выбранные базы IQ в состоянии, в каком они находились на определенный момент времени, разместив их на специальном оборудовании с доступом «только для чтения».

Интеллектуальный анализ данных и текста. В сотрудничестве с партнерами Sybase создает решения для рынков интеллектуального анализа данных и текста. В частности, это позволяет проводить сравнение реляционных данных, хранящихся в таблицах IQ, с нереляционными данными, хранящимися в виде крупных объектов. В связи с этим следует обратить внимание, что Sybase IQ обладает расширенными возможностями сложного анализа, в котором участвуют символьные крупные объекты (CLOB), а также двоичные крупные объекты (BLOB) и XML. В частности, продукт поддерживает возможности индексирования CLOB для текстового поиска.

Компания Sybase заключила ряд договоров о сотрудничестве в отношении Sybase IQ с различными поставщиками, в число которых входят компании, специализирующиеся на аппаратном обеспечении, системах хранения, обеспечении качества данных, управлении интеллектуальными ресурсами предприятия, обработке сложных событий и других областях, а также различные компании VAR и системные интеграторы. Примечательно, что решение одного из этих партнеров (PBS Software), включающее в себя систему Sybase IQ, официально сертифицировано SAP в качестве средства ускорения и организации архивного хранения для SAP BW.

Веб-сайт Sybase IQ: www.sybase.com/bi

ЗАКЛЮЧЕНИЕ

Мы полагали, что Sybase IQ столкнется с двумя препятствиями. Во-первых, его технологии хранения данных по столбцам (что было не очень хорошо понято и воспринято рынком), во-вторых, недоверие в связи с заявлениями в разнице производительности по сравнению с традиционными СУБД. Однако теперь мы не видим этих препятствий. Колонкоориентированная технология хранения данных вообще и Sybase IQ в частности появилась достаточно давно, и сейчас Sybase может приводить достаточное количество реальных примеров использования Sybase IQ клиентами. Во-первых, успех подхода Sybase был подтвержден во многих лидирующих компаниях. Во-вторых, как Sybase, так и другие поставщиками специализированных технологий потратили немало усилий, чтобы сделать очевидным факт, что более быстрая аналитика действительно возможна, разогревая аппетиты компаний, до сих пор страдающих от производительности традиционных СУБД, используя их для аналитических задач.

По мнению аналитической компании Bloor Research, подход, ориентированный на столбцы, которого придерживается Sybase, обеспечит гораздо более высокую производительность при более низкой стоимости по сравнению с традиционными подходами к организации сред анализа, отчетности и хранения данных. Более того, он отличается гораздо большей гибкостью по сравнению с той, что предлагается другими поставщиками. Поэтому мы не видим никаких препятствий дальнейшему успешному продвижению системы. Ранее мы уже рекомендовали рассматривать продукт более широко, и теперь рады отметить, что компания обращает внимание на новые сегменты рынка. При этом мы уверены (а с производительностью и другими возможностями в новой версии эта уверенность лишь укрепилась), что система Sybase IQ должна быть нацелена на весь рынок систем хранения общего назначения. Разумеется, в определенных областях система обладает наиболее значительными преимуществами. Тем не менее, общие возможности должны позволить ей конкурировать на всей площадке.

Дополнительные сведения

Более подробные сведения по этой теме см. по адресу
<http://www.BloorResearch.com/update/1037>

Сведения о компании Bloor Research

Компания Bloor Research является одной из ведущих европейских исследовательских, аналитических и консультационных организаций. Мы объясняем, как обеспечить большую гибкость корпоративных IT-систем за счет эффективного руководства, управления и использования информации. Благодаря независимому, интеллектуальному и хорошо продуманному подходу к предоставлению информации и подготовке публикаций по всем аспектам отрасли информационно-коммуникационных технологий мы заслужили репутацию компании, которая «представляет всю картину». Мы считаем, что предоставление полной информации позволяет:

- описать технологию с точки зрения ее ценности для бизнеса и другие взаимодействующие с ней системы и процессы;
- понять, как новые и инновационные технологии могут быть адаптированы к существующей инфраструктуре информационно-коммуникационных технологий;
- оценить весь рынок и представить все существующие решения, а также способы их более эффективной оценки;
- исключить не относящуюся к вопросу информацию и упростить поиск дополнительных сведений или новостей по инвестициям и реализации;
- обеспечить предоставление всех имеющихся у нас сведений по наиболее подходящему каналу.

Компания основана в 1989 году, и более двадцати лет мы предоставляем результаты исследований и анализа организациям, работающим в сфере IT во всем мире. Мы распространяем информацию через онлайн-подписку, предлагаем специализированные исследовательские услуги, организуем мероприятия и консультационные проекты. Наша цель состоит в том, чтобы превратить наши знания в ценный актив вашего бизнеса.

bi@sybase.ru

Москва, 115114, Дербеневская набережная, д.7, стр.16
телефон/факс: +7 (495) 797-4774

Киев, 01004, ул. Большая Васильковская, 9/2, офис 70
телефон/факс: +380 (44) 490-3498