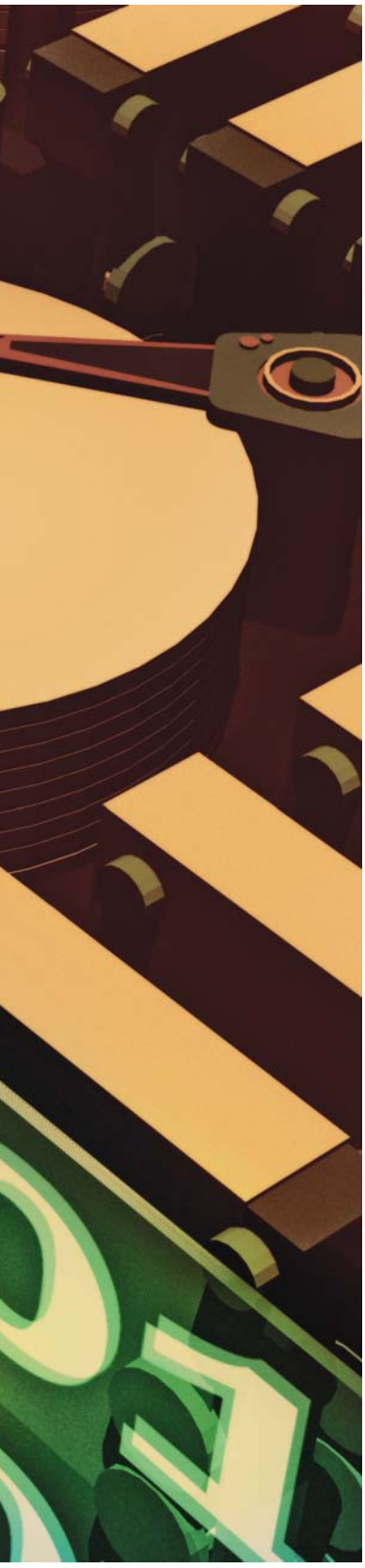
Николай Смирнов



BI без ХД: за и против

Создание корпоративных хранилищ данных зачастую оказывается крайне длительным и дорогостоящим проектом. Обойтись без него при построении аналитической системы — заманчивый вариант. Тем не менее построенные по такой модели системы имеют определенные ограничения

Несколько лет назад на конференции, посвященной теме хранилищ данных, один из слушателей заявил, что аббревиатура ХД на самом деле должна расшифровываться как «худо думали». При этом он подразумевал, что хранилище необходимо только в случае непродуманно выстроенных транзакционных систем, и уверял, что все аналитические задачи можно решить без оного.

Можно долго спорить о правомерности такого утверждения, но практика показывает: построение корпоративных хранилищ данных зачастую оказывается крайне длительным и дорогостоящим проектом, компания может оказаться втянутой в такой проект на год-полтора.

Обойтись без построения хранилища — заманчивый вариант. Например, таким путем пошли в «АльфаСтраховании» (подробнее см. «Цена скорости», ДИС, № 8, 2011).

Тем не менее построенные по такой модели системы имеют определенные ограничения. Для полнофункциональной и всеобъемлющей аналитической системы хранилище данных все же необходимо.

ИСТОЧНИК ИЛИ ЧАСТЬ АРХИТЕКТУРЫ?

«Хранилище зачастую действительно используется именно с целью нивелирования просчетов предыдущих внедрений — это факт. Так происходит, когда выгоднее сделать еще одну «правильную» надстройку, чем убирать все огрехи на нижнем уровне», — отмечает Александр Макеев, руководитель группы аналитической отчетности компании «Северсталь». Тем не менее даже в случае стратегического подхода к построению транзакционных систем и их безукоризненной архитектуры в среднесрочной перспективе обойтись без хранилища все равно не получится.

Во-первых, в компаниях все меняется, в том числе бизнес-процессы. Транзакционная система, первоначально полностью соответствовавшая бизнес-требованиям, рано или поздно устареет. Появление новой «светлой головы», новых бизнес-идей, нового менеджмента рано или поздно поставит ИТ-департамент в тупик с точки зрения настройки системы для получения требуемой отчетности.

Второй важнейший фактор — активность на рынке слияний и поглощений. Приобретаемые компании часто имеют совершенно другую учетную систему. Более того, у них может быть другой характер бизнеса, что не дает возможности тиражировать свою систему на приобретенный актив.

Наконец, технологии. В какой-то момент может оказаться, что очередные требования по предоставлению отчетности невозможно реализовать с сохранением приемлемого быстродействия транзакционных систем и с приемлемыми издержками.

«Есть задачи и обстоятельства, при которых без централизованного хранилища данных не обойтись», — согласен Олег Щербинин, руководитель направления решений BI «Ситроникс ИТ». В основном это связано с наличием разнородных источников данных, необходимостью гармонизации данных и справочников. Также хранилище необходимо в случае больших объемов данных со сложными алгоритмами расчета.

Если же источники данных согласованы друг с другом и приведены к единой нор-

мативно-справочной информации, то с помощью современных инструментов эти данные можно анализировать «на лету».

«Хранилище данных прежде всего необходимо тем аналитическим системам, в которых СУБД является неотъемлемой частью их архитектуры, а не только источником данных», — подчеркивает Петр Травкин, ведущий консультант QlikTech. Но существует ряд систем, для которых создание хранилища просто не требуется в силу отсутствия СУБД в их архитектуре, построенной на принципах обработки вычислений в оперативной памяти. При этом, безусловно, такие системы могут использовать в качестве источника данных и хранилище.

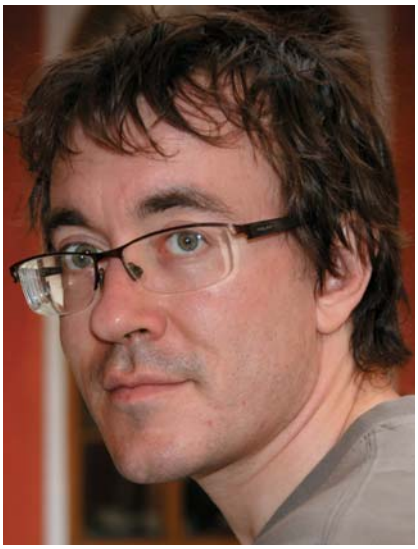
Хранилище данных также решает задачи консолидации данных, их очистки и обогащения. Если эти задачи стоят действительно остро, например в силу большого числа разрозненных систем в организации, то внедрение хранилища необходимо для получения корректной аналитической отчетности.

С одной стороны, сложно представить ситуацию, когда корректно спроектированное и построенное хранилище данных негативно сказывается на эффективности работы с данными. С другой стороны, внедрение хранилища — длительный проект, и у бизнеса всегда есть риск не получить аналитическую отчетность в нужные сроки.

«Конечно, отказаться от построения хранилища можно. И в этом случае понятно, что находится на чашах весов», — резюмирует Макеев. Да, технология обработки в оперативной памяти дает возможность обрабатывать огромный объем информации, но, отказываясь от хранилища, компания в большинстве случаев сильно ограничивает себя в возможностях. Например, она лишается возможности объединять данные из разных источников, создавая сквозную отчетность.

Кроме того, экономя на построении хранилища, организация получает другие издержки — в том числе на доработку транзакционных систем.

Если же говорить о точечных решениях в небольших компаниях, пусть и с огромным объемом информации (например, когда анализируются только продажи), то без хранилища вполне можно обойтись.



«Хранилище зачастую используется именно с целью нивелирования просчетов предыдущих внедрений — это факт», Александр Макеев, руководитель группы аналитической отчетности компании «Северсталь»

ПРОБЛЕМЫ ИНТЕГРАЦИИ

Независимо от того, создается хранилище или нет, как правило, при построении аналитической системы организации часто сталкиваются с вопросами, связанными с качеством данных, а также с управлением НСИ. В использующихся на предприятии системах одни и те же объекты НСИ, например наименования клиентов, поставщиков или товаров, могут быть описаны по-разному. И тогда при создании аналитической системы необходимо вести ключевую НСИ, необходимую для построения отчетности.

Эти две довольно трудоемкие задачи должны включаться в проект построения аналитической системы. Решаются они либо в рамках проекта внедрения хранилища, либо функциональными возможностями самой BI-системы.

Еще один важный вопрос — распределение ответственности за отдельные предметные области, так как должны быть владельцы предметных областей, следящие за алгоритмами расчетов, за способами агрегации по аналитическим измерениям. Выполнение этих двух пунктов позволяет получить качественную отчетность и отдачу от BI-инструмента.

Как правило, аналитические системы, не требующие обязательного наличия хранилища, сами обладают довольно мощным ETL-инструментарием. Они позволяют загрузить данные из разных систем, выполнить необходимые преобразования, например агрегацию, и в итоге корректно выстроить связи.

Облегчить и упростить процесс выверки данных и даже избежать его можно посредством тщательной проверки данных еще на этапе ввода. Однако это не всегда реализуемо по причине отсутствия должного функционала у некоторых транзакционных систем.

Когда требуемого качества данных на этапе их ввода и загрузки в систему не удалось достичь, имеет смысл привлечь к тестированию созданного решения опытного эксперта в предметной области анализируемой информации. При грамотно построенном с точки зрения представления информации аналитическом приложении значительную часть некорректных данных можно будет выявить, исправить в исходных вариантах и загрузить уже корректными в аналитическую систему.

Хранилище данных позволяет решить проблемы, связанные с интеграцией и качеством данных. Однако между появлением данных в транзакционных системах и их загрузкой в хранилище проходит определенное время. Значит ли это, что таким образом компания фактически отказывается от возможностей анализа «на лету»?

«Если архитектура решения позволяет напрямую использовать операционные системы и BI-инструмент обладает возможностью прямых запросов к источникам, то задержка между изменениями данных и их обновлением в отчетах будет минимальной», — говорит Щербинин. Фактически в таких случаях идет речь либо о построении отдельного решения, направленного на оперативную аналитику, либо об охвате основной аналитической системой некоторых транзакционных систем.

«В хранилище данные попадают с задержкой: необходимо не просто пе-

рекачать огромный объем данных, но и обогатить их аналитикой — сделать то, ради чего и существуют хранилища», — объясняет Макеев. В этом случае чаще всего речь идет об анализе, где режим реального времени не нужен.

Однако, во-первых, в любой транзакционной системе есть набор отчетов, позволяющих оперативно оперировать набором показателей, смотря на ситуацию в режиме реального времени. Во-вторых, случаев, когда необходима мгновенная реакция на изменения, немного. Поэтому вполне возможно создание дополнительных решений.

Например, в технологиях SAP есть возможность организации удаленного куба, позволяющего «смотреть» в транзакционную систему и выбирать оттуда то, что еще не пришло в хранилище в виде физической загрузки. Над этими двумя хранилищами — основным и виртуальным — создается объединяющее хранилище, на базе которого строится отчетность. С точки зрения пользователя, это актуальные данные.

Технически задача решаема, но она создает нагрузку на транзакционную систему, что также выливается в издержки, связанные с большими требованиями к вычислительным мощностям.

Наконец, есть и вариант более частой загрузки данных в хранилище.

«Хранилище данных у нас уже было построено, и не использовать его возможности было бы неразумно», Николай Ершов, аналитик отдела бизнес-структуры и аналитики ИТ компании «Л'Этуаль»

«Для каждого требования бизнеса существует своя технология, и связка транзакционных систем с хранилищем способна в любом случае удовлетворить его», — говорит Макеев. Однако важным фактором является сложность такой связки. Все-таки в хранилище данные лежат в существенно более структурированном виде, чем в транзакционной системе. Количество объектов, которыми приходится оперировать в хранилище, в десятки раз меньше, чем число таблиц в ERP. Именно сложность построения связей на уровне ERP будет тормозящим фактором.

НЕ ДУМАТЬ ОБ ОБЪЕМАХ

«Недавно появившиеся технологии оказали и продолжают оказывать влияние на построение систем аналитики», — констатирует Травкин. В частности, технологии резидентных вычислений уже изменили требования бизнес-пользователей к скорости отклика системы. Возможность интерактивного анализа информации в режиме реального времени становится все более востребованной, и поэтому для большинства пользователей время ожидания обработки запроса, превышающее несколько секунд, уже неприемлемо.

Другим примером могут служить мобильные технологии. Аналитические системы должны обеспечивать полноценную работу пользователей с любых устройств, в том числе и с мобильных.

«Сейчас достаточно много технологий, которые можно назвать post-OLAP», — добавляет Щербинин. Часть из них увеличивают быстродействие OLAP-кубов с помощью вычислений в оперативной памяти,

ХРАНИЛИЩЕ УЖЕ БЫЛО...

Сеть магазинов косметики «Л'Этуаль» при построении системы оперативной аналитики применила гибридный подход, позволяющий использовать хранилище данных и интегрированные внешние источники.

В компании существует корпоративное хранилище, над которым построена OLAP-отчетность. Кроме того, используется инструментарий для глубокого изучения данных (data mining) PolyAnalyst — разработка компании Megaputer.

«У всех систем есть свои плюсы и минусы. Когда компания растет, растут и объемы данных, формирование отчетов иногда может быть весьма длительным», — говорит Николай Ершов, аналитик отдела бизнес-структуры и аналитики ИТ компании «Л'Этуаль». Поэтому у бизнеса возникает потребность в инструментах, позволяющих оперативно и с минимальным привлечением ИТ получать достоверную информацию по ключевым показателям деятельности компании. Кроме того, существует много видов отчетности, которые с помощью OLAP-технологий очень сложно реализовать из-за особенностей архитектуры и самих данных.

Фактически компании была нужна система без глубокой аналитики, но с богатыми возможностями визуализации, позволяющая менеджерам оперативно «поиграть» с данными. Большинство существующих на рынке решений просто не справлялись с огромным объемом данных крупного ретейлера.

Все преимущества решения QlikView, которое было выбрано в итоге, укладываются в несколько пунктов. Во-первых, оно «всеядно». Между тем для решения поставленной задачи требовалось объединение данных разнородных источников: хранилища, ERP и прочих транзакционных систем.

Во-вторых, использование резидентных технологий позволяет мгновенно обрабатывать большие объемы данных. Наконец, оно простое в работе. Именно эти факторы и помогают системе быть конкурентоспособной на рынке.

другая часть способна предложить пользователям инструменты анализа, похожие на Excel, но гораздо мощнее, — так называемые self-service BI. Можно даже сказать, что происходит второе рождение концепции оперативной аналитической обработки (Online Analytical Processing, OLAP) для рабочих станций.

«Из таких нововведений серьезно меняет подход к внедрениям только self-service BI», — полагает Щербинин. Данный подход подразумевает, что подготовка источников данных будет сделана традиционными способами, а вот сама аналитика целиком отдана бизнес-пользователю. Оправдается ли такой подход, покажет только время. При этом традиционные подходы — сам OLAP и его реализации (например, Hyperion Essbase) — позволяют с успехом строить BI-системы без хранилищ данных.

«Если мы говорим о резидентных вычислениях применительно к хранилищам, то эта технология способна повлиять на используемые подходы, причем радикально: хранилища данных в этом случае можно резко упростить», — полагает Макеев. Можно будет делать акцент не на объемах, а на взаимосвязях.

Кроме того, появится более прозрачная вертикальная интегрированная структура хранилища. Классическое дисковое хранилище строится слоями, и над каждым из слоев можно строить отчеты.

В случае использования вычислений в оперативной памяти можно будет обойтись одним хранилищем вместо трех-четырех. Это приведет к уменьшению размеров базы, упрощению ее поддержки, отказу от интерфейсов между различными отчетами.

«Резидентные технологии — серьезный прорыв с точки зрения трансформации представлений о хранилищах. Но они вовсе не ведут к отказу от хранилищ», — уверен Макеев.

«ПОЧЕМУ ТАК ДОРОГО?»

Фактически единственным недостатком хранилищ данных являются сроки и бюджеты таких проектов. Так почему же они такие сложные и дорогие?

«Судя по опыту, главная причина заключается в отсутствии четкой постановки со стороны бизнеса при очень дорогих специалистах», — говорит Макеев. Зарплата эксперта по SAP BW, настраивающего хранилище, — от 150 до 400 тыс. руб. в месяц. При этом основное время уходит на понимание организации данных и настроек транзакционных систем. При наличии грамотной технической постановки, четко описывающей пожелания бизнеса, сами технические работы по внедрению хранилища не столь уж и сложны.

И конечно, отдельная проблема — изменение задач в процессе работы, но это особенность всех ИТ-проектов.

В ходе проекта у заказчика часто возникает большой объем административных задач (например, согласование документов). Большие объемы данных тоже добавляют сложности — мало того, что нужно делать очень эффективные структуры хранения и хитрые расчеты, но и отладка таких решений тоже занимает достаточно времени. Значительно влияет на длительность проекта территориальная распределенность заказчика и его способность договариваться с дочерними компаниями.

В целом, если задача очень сложная и у компании нет необходимого времени на ее решение, самым правильным будет проанализировать бизнес-требования и понять, от чего можно отказаться ради достижения результата. Это может серьезно сократить сроки проекта.

Наконец, очень часто — особенно в России — бизнес привык оперировать жесткими формами. Нередко ответить на конкретный бизнес-вопрос проще, чем реализовать требуемую форму отчета. Часто приходится иметь дело с техническими сложностями, не имеющими отношения к решаемой задаче. Такие трудности в работе с заказчиком могут отнимать очень много сил, времени и нервов. CIO.RU

«Для работы системы хранилище данных не требуется. Но оно у нас уже было построено, и не использовать его возможности было бы неразумно», — подчеркивает Ершов. Руководство компании рассматривает хранилище как бесценный кладезь информации, с помощью которой можно оценивать прошлое и прогнозировать будущее.

Что важно, существующие системы не дублируют функциональность, каждая из них выполняет свои задачи. OLAP-отчетность отражает стандартные шаблоны — например, показатели продаж. PolyAnalyst осуществляет глубокие исследования, причем может экспортировать результаты своей работы. В QlikView собраны показатели, которые в OLAP-отчетах не раскрыты детально. Система позволяет «проваливаться» вплоть до конкретной транзакции, причем делать это очень быстро. Данные о продажах, например, могут быть объединены с информацией о чеках, лояльности клиентов, финансовых показателях. Одним из интересных результатов стало построение цифровой карты магазинов с привязкой данных о продажах, остатках, количестве лояльных клиентов и многом другом.

Все приложения были разработаны силами нескольких штатных ИТ-специалистов компании.

Главными потребителями созданной оперативной отчетности являются коммерческий департамент и служба маркетинга.

«Ограничение у системы только одно — ресурсы, которые она потребляет. Все вычисления происходят в оперативной памяти, и если вы хотите получить мощную аналитику, которая будет действительно оперативной, нужны хорошие вычислительные ресурсы — мощное "железо" с достаточным количеством оперативной памяти», — констатирует Ершов.

Объемы данных растут, и уже в ближайшее время планируется переход на более мощные серверы, а также миграция на новую версию системы.