Хранилища Данных: вчера, сегодня, завтра

Казалось бы, только вчера я делал доклад на тему «Хранилища данных в банках» на одной очень уважаемой ИТ-тусовке, после которого несколько ИТ-директоров окружили меня и начали задавать вопросы «Почем шкафы (с серверами, разумеется) продаете?». На самом деле, эта история произошла в начале 2000-х. С тех пор «шкафы» принято называть системами хранения данных (СХД), а представление о «хранилище данных» (ХД) у большинства ИТ-специалистов совпадает с его определением в Википедии - очень большая предметно-ориентированная информационная корпоративная база данных, специально разработанная и предназначенная для подготовки отчетов, анализа бизнес-процессов с целью поддержки принятия решений в организации.

Если «копнуть» еще глубже в историю, то сама концепция хранилища данных (с окончанием «… о бизнесе») появился с подачи компании IBM в конце 80-х. В то время эта концепция представляла собой модель архитектуры для организации потока данных из оперативных (учетных) систем к лицам, принимающим решения. С тех пор хранилище данных как система получило массу имен: система поддержки принятия решений, система бизнес-анализа, корпоративная информационно-аналитическая система. Концептуальная архитектура же подобной системы остается неизменной и по сей день, включая в себя следующие «слои»:

- **Слой систем-источников данных**: ими служат любого рода учетные системы, использующиеся в организации – ERP, CRM, MRP, автоматизированные банковские системы, биллинговые системы (в телекоммуникационных компаниях), разрозненные Excel-файлы в конце концов;
- **Слой доступа к данным** – фактически, интерфейс хранилища данных к системам-источникам, отвечающий за извлечение, преобразование и загрузку (ETL - extract, transform, load) данных в хранилище;
- **Слой хранения данных и метаданных** – собственно, сердце хранилища, та самая «очень большая … база данных»;
- **Слой доступа к информации** – интерфейс конечных пользователей системы (потребителей информации полученной из данных) для работы с хранилищем данных.

Возникает вопрос – если все так просто и прозрачно, в какой из учетных систем нет модуля отчетности? Зачем было придумывать что-то новое? Ответ достаточно тривиален.

ОТВЕТЫ НА ВОПРОСЫ

Начнем с того, что учетных систем в любой организации всегда несколько и в каждой из них одни и те же данные учитываются по-разному: «М» и «Ж» для определения пола клиента в одной, и «0» и «1» для той же цели – в другой. Поэтому ответ на простейший вопрос «каково соотношение мужчин и женщин в части приносимых организации доходов?» требует как минимум одной «лишней» операции – приведения значений такого измерения

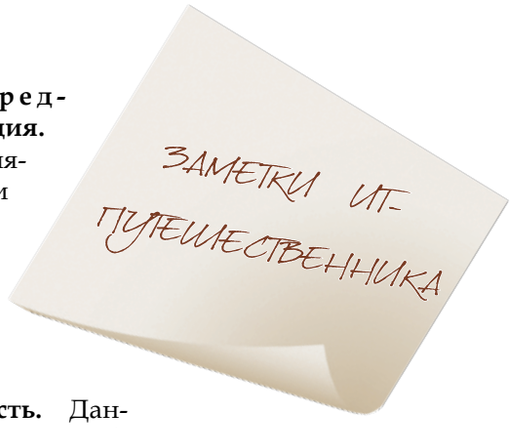
анализа как «пол» к единому знаменателю. А это вполне распространенный вопрос в любом клиент-ориентированном бизнесе.

Возьмем все учетные системы от одного поставщика (фантастика, но допустим), которые настолько интегрированы друг с другом, что подобные казусы и разночтения исключены. Вопрос решен? Не все так просто – заранее сложно «просчитать» какого рода информация потребуется тому или иному лицу, принимающему решения, а плодить отчеты сотнями на все случаи жизни – такого ни одному ИТ-директору не пожелаешь.

Можем взять удобное средство бизнес-анализа (этот термин тоже появился в конце 80-х) и «прикрутить» его прямо к базам данных учетных систем: теперь каждый пользователь может сам себя обеспечить информацией для построения отчета, «перетаскивая» показатели и измерения на панель формирования запроса (никакого знания SQL – ура!). Вопрос закрыт? К сожалению, нет - оказывается наиболее «одаренные» аналитики своими запросами рано или поздно просто кладут на лопатки любую учетную систему. Вместо того чтобы учитывать (для чего, собственно, такие системы и предназначены), она начинает часами пыхтеть, чтобы вернуть пользователю выборку типа «все операции всех клиентов по дням за последние три года». Это, опять же, вполне типовой запрос в любом банке.

ОК, удваиваем аппаратные мощности, строим кластер, делаем реплику баз данных учетных систем, внедряем сложные инструменты балансировки нагрузки, устраиваем систему интеллектуального резервирования и/или кэширования данных (нужное подчеркнуть) и забываем про проблемы производительности. Все счастливы? Не тут-то было – оказывается, злобным аналитикам нужна историчность: сколько нам был должен клиент позавчера, в каком регионе он находился месяц назад, к какому сегменту относился в прошлом году? В учетных системах историчности нет априори. Но даже если мы решим эту проблему путем хранения ежедневных «срезов» данных, с которыми практически невозможно работать, да и проблему с корректировками задним числом это не снимает, остаются проблемы одинаковой трактовки информации. К примеру, «клиент» - это тот с кем у нас сейчас есть действующий договор или тот, с кем у нас были хоть какие-то договорные отношения на протяжении последнего года? Проверки, очистки и стандартизации данных, восполнение недостающих данных, таких как определение связанных лиц… И это далеко не полный перечень.

Хранилища данных избавлены от таких проблем, так как данные в них физически перемещенные из учетных систем и организованы по следующим принципам:



- **Проблемно-предметная ориентация.** Данные объединяются в категории и хранятся в соответствии с описываемыми областями, а не с используемыми приложениями.
- **Интегрированность.** Данные объединены таким образом, чтобы удовлетворять всем требованиям предприятия в целом, а не единственной функции бизнеса.
- **Некорректируемость.** Данные в хранилище не создаются, т.е. поступают из внешних источников, не корректируются и не удаляются.
- **Зависимость от времени.** Данные в хранилище точны и корректны только в том случае, когда они привязаны к некоторому промежутку или моменту времени.

Данный перечень можно дополнить рядом требований к информации, которые удовлетворяет хранилище данных: правильность (за счет проверки и очистки данных), своевременность (за счет предрассчитанных показателей), доступность, релевантность, краткость и защищенность (за счет применения инструментов бизнес-анализа). Становится очевидным, что подобные системы появились для эффективной поддержки следующей важной цепочки: данные > информация > знания > решения. Хотя сами хранилища данных решений не принимают, как бы этого не хотелось.

БОРЬБА ПОДХОДОВ

Почти одновременно с появлением концепции хранилища данных возникли и две противоборствующие школы подходов к созданию сердца хранилища (слоя хранения данных).

Ральф Кимбалл предлагал создавать хранилища как набор отдельных витрин, направленных на решение конкретных бизнес-задач и согласованных между собой на уровне размерностей (измерений - справочников).

Билл Инмон, напротив, настаивал на создании хранилища основанного на нормализованной модели данных, включающей в себя сущности с атрибутами, отражающие суть деятельности организации.

Каждый из этих подходов имел свои преимущества и недостатки. Первый обладал быстротой появления первых результатов, но имел место рост сложности согласования каждой последующей витрины с уже имеющимися. Второй – устойчивость всего хранилища во времени, но необходимость длительного начального проектирования

всего хранилища была явным недостатком. Со временем оба этих подхода были объединены в один гибридный: витрины с показателями для анализа строятся поверх детального нормализованного слоя атомарных данных, который в свою очередь проектируется и создается итерационным путем.

АРХИТЕКТУРА ХРАНИЛИЩА ДАННЫХ

Сегодня о хранилищах данных не говорит только ленивый. У каждой уважающей себя компании за рубежом, хранилище уже есть (и порой - не одно). Самое крупное находится в распоряжении у eBay и содержит более 6 петабайт данных (1ПБ=1024ТБ). В украинских компаниях хранилища тоже есть, правда не у всех – первыми их установили себе операторы мобильной связи. У одного из украинских телекомов хранилище появилось первым в СНГ почти десять лет назад. Потом к ним подключились банки, и если бы не кризис, эту эстафету активно переняли бы и страховые компании, розничные сети, производители и дистрибьюторы товаров народного потребления. Именно компании в этих сферах накопили огромные объемы данных о своих клиентах или продуктах, ценность которых сложно переоценить, ведь без их анализа эффективно делать бизнес уже не представляется возможным.

Концептуальная архитектура хранилища данных, предложенная более 20 лет назад, уже давно оформилась в стандартизированную:

- Слой доступа к данным представлен разнообразными платформами интеграции данных, включающими не только ETL-инструменты, но и программные продукты для очистки и обогащения данных, а также специфические механизмы доступа к данным в системах-источниках (используя журналы операций последних);
- Слой хранения данных логичным образом реализуется за счет обычных (реляционных) и необычных (колоночно-ориентированных) баз данных с огромным количеством опций для повышения производительности и снижения стоимости хранения. К примеру, за счет партиционирования, сжатия, хранения разно востребованных данных на носителях с различной производительностью и стоимостью;
- Слой доступа к данным уже давно строится с использованием инструментов бизнес-анализа (BI - business intelligence), которые не только поддерживают все мыслимые способы работы с данными (OLAP-анализ, ad-hoc-запросы), но и предоставляют обширные возможности визуализации, такие как таблицы, графики, дэшборды. Но самое главное, что они позволяют не-ИТ пользователям самим получать то, что им нужно. Это возможно осуществить по расписанию, или моментально, за счет наличия данных для анализа прямо в оперативной памяти, а также в своем привычном интранет-портале или прямо на мобильном терминале.

Помимо упомянутых типов программных продуктов, хранилища обросли большим количеством вспомогательных инструментов. Нужно знать, какой бизнес-смысл скрывается за тем или иным показателем: есть глоссарий, и средства управления метаданными, не говоря уже о различных средствах аудита всей системы, профилирования данных, управления ETL-процессами и т.д.

Также появились и оформились в отдельные продукты правила извлечения данных из распространенных систем-источников (так называемые коннекторы), логические модели данных хранилища (чаще всего их называют индустри-

Что касается поставщиков (вендоров) соответствующего программного обеспечения для хранилищ данных, то выбор сейчас более чем широк:

- Все мега-вендоры (IBM, Microsoft, Oracle, SAP, SAS) предлагают практически полный стек программных технологий и, зачастую, методологий. Не в последнюю очередь благодаря удачным поглощениям других поставщиков;
- Много нишевых игроков гордятся лучшими в своем классе инструментами, к примеру, Informatica со своей интеграционной платформой;
- Некоторые поставщики предоставляют сразу целый программно-аппаратный комплекс - хранилище «из коробки», к примеру - Teradata.

Но, как показывает практика, появление успешного хранилища данных в компании, редко зависит от выбора того или иного подхода его построения и используемых при этом технологий и методологий: его создают и потом используют живые люди. Именно их компетенции, опыт и вовлеченность влияют на результат гораздо больше.

альными), наборы показателей и размерностей для анализа тех или иных областей бизнеса (аналитические решения с преднастроенными отчетами и дэшбордами).

Кроме того, аналитики и другие потребители информации давно уже не являются единственными прямыми клиентами хранилищ данных. Эти системы стали источником эталонных данных о бизнесе для других приложений, как оперативных (полное досье клиента со всей его историей теперь может быть доступно оператору контакт-центра), так и аналитических (выявление закономерностей требует больших массивов данных за длительное время).

БУДУЩЕЕ ЗА SAAS

Из всего изложенного выше может сложиться впечатление, что хранилища данных достигли своего технологического совершенства, и их развитие будет лежать лишь в плоскости увеличения объемов данных, скорости их обработки, снижения стоимости владения. На самом деле это не так. Уже завтра может измениться месторасположение хранилищ - они переедут в «облака» и работа с ними будет строиться по принципу «ПО как услуга» (SaaS - software as a service). Но это далеко не самая главная перспектива их развития. Значительно более важным является расширение области применения хранилищ со структурированных данных на неструктурированные - по содержанию SMS своих абонентов операторы мобильной связи смогут определять вероятность их оттока, не говоря уже об аналитике изображений и видео материалов.

Последним прогнозом на завтра, и, возможно, самым важным есть превращение хранилища данных из систем помощи в принятии решений в системы, подсказывающие правильные решения - какую новую услугу нужно предложить клиенту, позвонившему в банковский контакт-центр, какие товарные позиции и в каком объеме нужно закупить супермаркету. Хотя принятие окончательного решения всегда останется за живым человеком.

Максим БОДАЕВ
Директор по развитию бизнеса,
Citia Business & Technology Consulting