

Обеспечение качества данных в аналитических системах



Николай ДОРОГОВ,
директор практики
информационно-аналитических
систем Columbus Russia

BI-системы и качество данных

На заре развития бизнес-приложений отчетливо прослеживалось стремление включить в одно приложение и функционал, необходимый для транзакционной обработки данных, и инструменты для анализа и построения сложных отчетов по большим объемам данных. Однако достаточно быстро выяснилось, что аналитические системы существенно отличаются от транзакционных как по функциональным требованиям, так и по технологической реализации. Например, в части функциональных требований для транзакционных систем важны повышение скорости ввода данных по учетным операциям и снижение вероятности ввода ошибочных данных пользователем, что подразумевает необходимость автоматизации регламентированного процесса, а для аналитических систем основным требованием является предоставление удобного инструмента для разностороннего анализа данных, операций «проваливания», детализации данных в уникальной для каждого случая последовательности. В технической

реализации для транзакционной системы ключевая задача – обеспечение надежного механизма добавления записей большим количеством конкурентных пользователей, в то время как для аналитической системы – возможность быстрой выборки больших объемов данных и их агрегации.

Системы класса BI (Business Intelligence) обеспечивают интерфейс для пользователя аналитической отчетности. Они предоставляют инструменты для эффективной интерактивной работы с большими объемами данных, включают в себя средства для работы с OLAP-кубами.

Для полноценной работы BI-системы необходимо, чтобы она функционировала на основе предварительно собранных из разрозненных систем и очищенных данных. Это обусловлено, в частности, тем, что на больших объемах данных обеспечить объединение наборов данных из нескольких систем «на лету» с нормальной производительностью практически невозможно. Кроме того, некоторые транзакционные системы-источники не могут быть постоянно доступны для выполнения на них «тяжелых» запросов вследствие высокой нагрузки со стороны пользователей или из-за узких каналов связи.

Важно, чтобы пользователи доверяли данным из BI-системы, а при достижении этой цели неминуемо возникают задачи, связанные с обеспечением качества данных: пропущенные или неправильно заполненные аналитики, противоречивые данные, нарушения ссылочной целостности, несогласованные между системами справочники и др. Одни проблемы решаются достаточно просто, с другими придется сложнее, главное – осознавать, что задачи обеспечения качества данных нужно будет решать, заранее

готовиться к этому, планировать и осуществлять мониторинг в ходе внедрения и использования системы BI. Так, мы можем годами заносить информацию по клиентам в транзакционную систему, но, только приступив к внедрению BI, в процессе анализа по возрасту обнаружить, что у некоторых клиентов не заполнена графа «дата рождения», у других дата рождения совпадает с датой первого обращения в компанию. Для решения подобных задач зачастую приходится применять комплексное решение, включающее в себя изменения на административном уровне, доработку учетных транзакционных систем и настройку механизмов мониторинга на уровне аналитической системы.

Все эти специфичные моменты привели к тому, что между транзакционными системами и BI выделили уровень сбора, очистки и хранения данных, который можно называть хранилищем данных.

Хранилища данных

Эффективная работа с накопленными за много лет данными позволяет повысить качество анализа, принимать более обоснованные и взвешенные управленческие решения. Но чтобы добраться до этих данных, сделать их доступными для пользователей, необходимо решить ряд серьезных вопросов. Один из них связан с доступом к данным из систем-источников. Хорошо, если мы работаем с документированной системой, хранящей данные в современной СУБД. Еще лучше, если в компании есть специалисты, поддерживающие и развивающие эту систему. В таком случае есть шанс без лишних проблем получить данные из этой системы, преобразовать их по нужным алгоритмам и поместить в область постоянного хранения. Однако иногда в компании

используются системы с недокументированной структурой базы данных, а для хранения данных – файлы собственных форматов или базы данных, для которых не реализованы стандартные интерфейсы доступа. Тогда получение данных может превратиться в нетривиальную задачу с непредсказуемыми сроками решения и трудозатратами. Здесь помогут специализированные инструменты извлечения, преобразования и загрузки данных (Extract-Transform-Load, ETL). Правильный ETL-инструмент должен

средства по обеспечению качества. Такие средства содержат типовые алгоритмы анализа, в них есть возможности настройки рабочей среды для сотрудников, отвечающих за качество данных. Сейчас большинство подобных средств хорошо интегрируются с ETL-инструментами.

При работе над качеством данных важной задачей является выделение и мониторинг его метрик. На первом этапе это могут быть простые показатели, количественно характеризующие долю записей, для которых выполняются условия по

решения, предполагающие необходимость хранения нормализованных данных и построения денормализованных витрин на их основе. Есть подход Ральфа Кимбалла, предлагающий создавать хранилище данных в виде совокупности таблиц фактов, которые соединяются через таблицы согласованных измерений. Решение о выборе подхода должно быть индивидуальным для каждого случая. В качестве общего решения можно рекомендовать подход Кимбалла, который позволяет быстро и эффективно выполнить конкретные бизнес-задачи по анализу и отчетности, и подход с хранением нормализованных данных – при необходимости построить хранилище как элемент ИТ-инфраструктуры компании. При выборе второго пути (с хранением нормализованных данных) имеет смысл задуматься об использовании промышленной отраслевой модели данных. Разработанные по итогам реализации большого количества проектов модели таких компаний, как IBM, Teradata, значительно снижают риски возникновения проблем, связанных с неправильным проектированием модели данных хранилища.

При выборе BI-инструмента следует ориентироваться в первую очередь на потребности бизнес-пользователей, так как именно они будут постоянно использовать это средство в работе.

обеспечивать доступ практически к любому источнику данных, работу на уровне логики приложений при подключении к типовым бизнес-приложениям, высокую производительность при преобразованиях данных и масштабируемость. Хорошие ETL-инструменты не требуют написания кода, позволяя визуально проектировать потоки данных, что повышает скорость разработки, а главное – существенно упрощает поддержку и модификацию решения.

С помощью ETL-инструментов можно успешно решать и часть задач по обеспечению качества данных. Например, требования по непустым значениям полей, по соответствию строк (телефонных номеров, номеров машин, документов) заданным шаблонам, по базовому контролю ссылочной целостности можно контролировать на уровне процессов ETL. Для решения более сложных задач, например, таких как интеллектуальный анализ данных по клиентам, рассчитывающий вероятность того, что под двумя записями о клиентах на самом деле скрывается один человек, целесообразно применять специализированные

заполнению ключевых полей, соответствию строк заданным шаблонам, наличию некорректных ссылок на данные. Наблюдение за изменением метрик дает возможность количественно оценивать эффективность усилий и конкретизировать цели по повышению качества данных.

При подготовке к внедрению всегда решается вопрос о модели данных в хранилище. Существуют

Отраслевая практика

В нашей стране технологии, касающиеся построения хранилищ данных, уже неплохо освоены, причем степень продвинутой применяемых технологий и архитектуры в значительной степени связана с отраслью.

мнение специалиста



Ольга КУЧИНА,
специалист по решениям Business Analytics,
IBM в России и СНГ

Ключевым моментом в обеспечении комфортного пользования системой является та ее часть, которую потребители не видят, – Business Intelligence. Чтобы конечный интерфейс был «дружелюбным» в ежедневном использовании, необходимо, чтобы он был оснащен соответствующими BI-инструментами: надежными базами данных и аналитическим хранилищем, мощными ETL – средствами и инструментами очистки данных. Заказчику логично использовать комплексное решение от единого поставщика, что позволит сократить срок настройки решения, обеспечить техническую поддержку и снизить совокупную стоимость владения решением. Более того, не придется вкладывать дополнительные ресурсы в интеграционный процесс – как правило, вендор предлагает уже интегрированный продукт. В случае с IBM – это портфель программных продуктов Information Management (IBM DB2, IBM Data Stage, InfoSphere, IBM Netezza) и IBM Business Analytics (IBM Cognos TM1, Cognos Controller, Cognos Business Intelligence).

мнение специалиста



Андрей ТИУНОВ,
генеральный директор BI Partner
(ГК «Ай-Теко»)

Вопрос консолидации данных и подготовки информации для использования в BI-приложениях, конечно, является первоочередным. По нашему опыту, все, что связано с самим хранилищем данных, включая обследование источников, проектирование интерфейсов, разработку ETL-компонентов, занимает от 70 до 90% всех трудозатрат на проекте. Это очень объемная задача, которая для своего решения может потребовать от заказчика даже доработки существующих информационных систем. На рынке много отличных инструментов и выбор того или иного средства больше вопрос привычки работать с тем или иным вендором. В силу бюджетных ограничений многие из реализованных на сегодняшний день хранилищ, включая очень масштабные системы, сделаны на встроенных в СУБД инструментах. При грамотном проектировании с точки зрения производительности ETL-процессов этого обычно достаточно, роль специализированных ETL-средств часто переоценивается. Тем не менее для телекоммуникационных компаний, банков, розницы становится критичным использование выделенных ETL-платформ, способных значительно увеличить производительность процессов обновления хранилищ на больших объемах данных. Помимо обеспечения высокой скорости промышленные ETL-инструменты упрощают и ускоряют процессы, связанные с развитием систем, подключением новых источников, изменением правил расчета и пр. В крупных компаниях выделены целые группы, которые постоянно заняты на модификации ETL. Для телекоммуникационных компаний характерна высокая нагрузка на систему хранения со стороны BI, и есть смысл задумываться об использовании специализированных программно-аппаратных платформ. Например, таких как Oracle Exadata. Их применение может увеличить производительность обработки данных в десятки раз.



Евгений КУРИЛОВИЧ,
руководитель проектов, компания «ФОРС»

Систему бизнес-анализа нельзя рассматривать в узком смысле – только как хранилище данных и инструмент для получения аналитической отчетности. Зачастую она является частью более сложной многокомпонентной информационной системы, решающей целый комплекс различных задач. Например, при автоматизации бюджетирования и смежных процессов BI-системы являются обеспечивающими для финансового планирования, оперативного контроля движения денежных средств и управленческой отчетности, что сегодня особенно актуально для большинства предприятий независимо от их отраслевой специфики. Важно, чтобы эти компоненты могли внедряться как независимо друг от друга, так и в составе комплексного решения, обеспечивая интеграцию всех используемых информационных ресурсов и баз данных. Так, архитектура разработанного нами решения предполагает прямую загрузку данных из разрозненных источников в единое аналитическое хранилище. На его базе строится подсистема управленческой отчетности, включающая информационные панели, аналитические отчеты и аналитическую модель данных. Подсистема финансового планирования создана на основе специализированного промышленного продукта Oracle Hyperion Planning. Подсистема контроля платежей разработана как готовое тиражируемое решение, которое поддежит последующей адаптации и настройке в соответствии с конкретными пожеланиями заказчика.

Таким образом, предприятие получает возможность решить сразу несколько ключевых задач финансового управления и создать условия для соответствующего методологического и программного обеспечения. Вся цепочка бизнес-процессов отлажена и правильно выстроена, для того чтобы руководство и заинтересованные сотрудники могли получать актуальную и достоверную финансово-аналитическую информацию в любых необходимых им разрезах.

Например, во многих крупных и средних банках для подготовки управленческой или регуляторной отчетности используются решения на основе технологий хранилищ данных. Большинство таких банков уже прошло этап стихийного и бессистемного роста разрозненных баз данных, каждая из которых содержит данные для решения одной локальной задачи. Как правило, при реализации подобных решений широко использовались кодирование, написание скриптов, а специализированные ETL-средства не применялись. Столкнувшись со сложностями изменения таких решений при возникновении новых требований, банки стали активно изучать промышленные технологии и решения. В настоящее время все больше проектов по хранилищам данных реализуется с использованием промышленных ETL-инструментов, позволяющих отказаться от написания кода, таких как системы от компаний IBM, Informatica. Сейчас в России идет несколько крупных проектов с применением промышленных моделей для банковского хранилища данных. В совокупности со специализированными СУБД и современными BI-инструментами это позволяет банкам строить масштабируемые решения, оперативно реагировать на изменения потребностей бизнеса. В банковском секторе заметна тенденция перехода от стратегии экстенсивного роста к стратегии повышения эффективности работы с существующими клиентами, что невозможно без сбора всех данных о параметрах и действиях каждого клиента. Решения этих задач еще больше повышают требования к качеству данных в хранилище. В таких случаях внедрение хранилища и системы обеспечения качества данных сопровождается хорошо измеримым материальным эффектом за счет сохранения выгодных клиентов, продаж дополнительных банковских продуктов.

Компании из сферы розничной торговли, привыкшие выжимать максимум эффективности из вложений, как правило, имеют менее развитые решения в области аналитических систем и хранилищ данных. Рынок внимательно присматривался к «пионерам», которые внедрили специализированные СУБД для хранилищ и промышленные ETL-инструменты, промышленные модели данных и

современные BI-инструменты. Успешный опыт внедрений у конкурентов помогает ритейлерам принимать решение о запуске новых проектов по хранилищам данных и BI. Как правило, ритейлеры ставят перед собой конкретные задачи, направленные на измеримое повышение эффективности. Видя, как ИТ-инструменты помогают больше заработать или снизить расходы, они принимают взвешенные решения по выбору инструментов и запуску новых проектов.

Тенденции и рекомендации

Основными трендами в развитии аналитических систем и хранилищ данных можно считать персонализацию клиентов, постоянное увеличение объемов используемых данных и ужесточение требований к оперативности поступления информации.

Для того чтобы компания не теряла выгодных клиентов и приобретала новых, необходимо хорошо представлять потребности каждого из них, изучив всю доступную информацию. Многие организации

уже сейчас помимо традиционных источников – собственных транзакционных систем и результатов исследований рынка – используют данные социальных сетей, стремясь еще больше персонализировать свое предложение. Это требует нового качества процессов сбора и очистки данных, приводя к увеличению объемов собираемой, преобразуемой и хранимой информации. С другой стороны, постоянно повышаются требования к оперативности получения результатов анализа. Все больше задач требуют доступа к данным в режиме онлайн. В результате развиваются как программные инструменты, так и подходы к построению архитектуры аналитических систем. Системы, решавшие ранее только задачи ETL, получают возможности обращаться к данным приложений «на лету», используют функционал web-сервисов, становятся системами не просто загрузки, а полноценной интеграции данных. По мере увеличения скорости обработки данных и пропускной способности каналов связи появляется возможность повышения производительности аналитических систем, с постепенным приближением к режиму онлайн.

С целью соответствия возрастающим требованиям бизнеса при планировании проекта по хранилищу данных имеет смысл ориентироваться на продукты глобальных поставщиков. Менять СУБД или ETL-средство после нескольких лет эксплуатации сложно и затратно, поэтому выбирать платформы нужно с учетом увеличения объемов данных и сложности преобразований. При выборе BI-инструмента следует ориентироваться в первую очередь на потребности бизнес-пользователей, так как именно они будут постоянно использовать это средство в работе. Выбор продуктов из линейки одного производителя позволяет получить весомую экономию затрат на программное обеспечение, однако принимать решение можно только после оценки степени эффективности выбираемых средств для решения конкретных задач проекта в реальном окружении ИТ компании. И конечно, нельзя забывать, что выбор даже самого современного программного обеспечения не снижает требований к квалификации команды, которая будет осуществлять внедрение и развитие аналитической системы и хранилища данных. ■