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