Г.Г. Волков
Филиал БГЭУ (Бобруйск)

ИНФОРМАЦИОННЫЕ ХРАНИЛИЩА ДАННЫХ В КИС

Автоматизированные системы, называемые OLTP-системами, имеют следующие характеристики:

- они максимально оптимизированы для быстрого выполнения транзакций;
- рассчитаны на производство небольшого количества отчетов;
- в них хранятся данные за небольшой временной период;
- они ориентированы на автоматизацию деятельности исполнителей, выполняющих типовые процедуры.

При этом автоматизация почти не затрагивает управленцев – персонал, ответственный за принятие решений. Следовательно, широко распространенные решения на базе ERP-систем задачу быстрого анализа складывающейся на рынке ситуации и возможности прогнозировать ее развитие не решают, анализ – «слабое звено» в цепочке менеджмента.

Попытки построить системы принятия решений, которые обращались бы непосредственно к базам данных OLTP-систем, оказываются в большинстве случаев неудачными.

Во-первых, аналитические запросы «конкурируют» с оперативными транзакциями, блокируя данные и вызывая нехватку ресурсов.

Во-вторых, структура оперативных данных предназначена для эффективной поддержки коротких и частых транзакций и не обеспечивает необходимой скорости выполнения аналитических запросов.

В-третьих, в организации, как правило, функционирует несколько OLTP-систем, каждая со своей базой данных. В этих базах используются различные структуры данных, способы кодирования и т. д. Для аналитика задача построения какого-либо сводного запроса по нескольким подобным базам данных практически неразрешима.

Для обеспечения возможности анализа накопленных данных организации стали создавать хранилища данных. Хранилища данных – это «предметно-ориентированные, интегрированные, стабильные, поддерживающие хронологию наборы данных, организованные для целей поддержки управления, оперативного анализа и принятия решений» (Билл Инмон «Building the Data Warehouse», QED/Wiley, 1991).

Хранилище данных строится на базе клиент-серверной архитектуры, реляционной СУБД и утилит поддержки принятия решений. Данные из промышленной OLTP-системы копируются в хранилище данных таким образом, чтобы построение отчетов и OLAP-анализ не использовал ресурсы промышленной системы и не нарушал ее стабильность. Они загружаются в хранилище с определенной периодичностью, поэтому актуальность данных несколько отстает от OLTP-системы. На сегодняшний день существует два основных подхода к архитектуре хранилищ данных.

Это так называемая корпоративная информационная фабрика (CIF) Билла Инмона и хранилище данных с архитектурой шины (BUS) Ральфа Кимболла. Работа CIF начинается со скоординированного извлечения данных из источников. Затем загружается реляционная база данных с третьей нормальной формой, содержащая атомарные данные. Получившееся нормализованное хранилище используется для того, чтобы наполнить информацией дополнительные репозитории презентационных данных, т. е. данных, подготовленных для анализа. Эти репозитории, в частности, включают специализированные хранилища для изучения и «добычи» данных, а также витрины данных. При таком сценарии конечные витрины данных создаются для обслуживания бизнес-отделов или для реализации бизнес-функций и используют пространственную модель для структурирования суммарных данных. Основная цель пространственной модели – минимизировать время выполнения запроса, поэтому допускается денормализация данных и их группировка вокруг центральной задачи, которую придется выполнять наиболее часто.

В модели BUS первичные данные преобразуются в информацию, пригодную для использования, на этапе подготовки данных. При этом обязательно принимаются во внимание требования к скорости обработки информации и качеству данных. Подготовка данных начинается со скоординированного извлечения данных из источников. Ряд операций совершается централизованно, например, поддержание и хранение общих справочных данных. Другие действия могут быть распределенными.

Область представления пространственно структурирована, при этом она может быть централизованной или распределенной. Пространственная модель хранилища данных содержит ту же атомарную информацию, что и нормализованная модель, но информация структурирована по-другому, чтобы облегчить ее использование и выполнение запросов. Эта модель включает как атомарные данные, так и обобщающую информацию (агрегаты в связанных таблицах или многомерных кубах) в соответствии с требованиями производительности или пространственного распределения данных. Запросы в процессе выполнения обращаются к все более низкому уровню детализации без дополнительного перепрограммирования. В отличие от подхода Билла Инмона, пространственные модели строятся для обслуживания бизнес-процессов, а не бизнес-отделов.

Какие отличия имеют эти два подхода?

Во-первых, архитектуры отличаются способами обращения с атомарными данными: их пространственной организацией у Кимболла и нормализованной – у Инмона.

Во-вторых, если у Инмона хранилище данных – это физически целостный реально существующий объект, то хранилище Кимболла – скорее «виртуальный» объект. Это коллекция витрин данных, которые могут быть пространственно разобщенными.

Вопрос «Чья же модель лучше?» не имеет однозначного ответа. Выбор того или иного технического решения определяется нуждами бизнеса и его конкретными особенностями.