КОНЦЕПЦИЯ ХРАНИЛИЩА ДАННЫХ (продолжение)

Данные, поступающие из ОИД в ХД, перемещаемые внутри ХД и поступающие из ХД к аналитикам, образуют следующие информационные потоки (см. рис. 2.5):

- входной поток (Inflow) — образуется данными, копируемыми из ОИД в ХД;
- поток обобщения (Upflow) — образуется агрегированием детальных данных и их сохранением в ХД;
- архивный поток (Downflow) — образуется перемещением детальных данных, количество обращений к которым снизилось;
- поток метаданных (MetaFlow) — образуется потоком информации о данных в репозиторий данных;
- выходной поток (Outflow) — образуется данными, извлекаемыми пользователями;
- обратный поток (Feedback Flow) — образуется очищенными данными, записываемыми обратно в ОИД.

Самый мощный из информационных потоков — входной — связан с переносом данных из ОИД. Обычно информация не просто копируется в ХД, а подвергается обработке: данные очищаются и обогащаются за счет добавления новых атрибутов. Исходные данные из ОИД объединяются с информацией из внешних источников — текстовых файлов, сообщений электронной почты, электронных таблиц и др. При разработке ХД не менее 60 % всех затрат связано с переносом данных.

Процесс переноса, включающий в себя этапы извлечения, преобразования и загрузки, называют ETL-процессом (E — extraction, T — transformation, L — loading: извлечение, преобразование и загрузка, соответственно). Программные средства, обеспечивающие его выполнение, называются ETL-системами. Традиционно ETL-системы использовались для переноса информации из устаревших версий информационных систем в новые. В настоящее время ETL-процесс находит все большее применение для переноса данных из ОИД в ХД и ВД.

2.1. Концепция хранилища данных

Стремление объединить в одной архитектуре СППР возможности OLTP-систем и систем анализа, требования к которым во многом, как следует из табл. 1.1, противоречивы, привело к появлению концепции *хранилищ данных* (ХД).

Концепция ХД так или иначе обсуждалась специалистами в области информационных систем достаточно давно. Первые статьи, посвященные именно ХД, появились в 1988 г., их авторами были Девлин и Мэрфи. В 1992 г. Уильман Г. Инмон подробно описал данную концепцию в своей монографии "Построение хранилищ данных".

В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа. Это позволяет применять структуры данных, которые удовлетворяют требованиям их хранения с учетом использования в OLTP-системах и системах анализа. Такое разделение позволяет оптимизировать как структуры данных оперативного хранения (оперативные БД, файлы, электронные таблицы и т. п.) для выполнения операций ввода, модификации, удаления и поиска, так и структуры данных, используемые для анализа (для выполнения аналитических запросов). В СППР эти два типа данных называются соответственно *оперативными источниками данных* (ОИД) и хранилищем данных.

В своей работе Инмон дал следующее определение ХД.

Внимание! Хранилище данных — предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.

Рассмотрим свойства ХД более подробно.

Предметная ориентация — является фундаментальным отличием ХД от ОИД. Разные ОИД могут содержать данные, описывающие одну и ту же предметную область с разных точек зрения (например, с точки зрения бухгалтерского учета, складского учета, планового отдела и т. п.). Решение, принятое на основе только одной точки зрения, может быть неэффективным или даже неверным. ХД позволяют интегрировать информацию, отражающую разные точки зрения на одну предметную область.

Предметная ориентация позволяет также хранить в ХД только те данные, которые нужны для их анализа (например, для анализа нет необходимости хранить информацию о номерах документов купли-продажи, в то время как их содержимое — количество, цена проданного товара — необходимо). Это существенно сокращает затраты на носители информации и повышает безопасность доступа к данным.

Интеграция — ОИД, как правило, разрабатываются в разное время несколькими коллективами с собственным инструментарием. Это приводит к тому, что данные, отражающие один и тот же объект реального мира в разных системах, описывают его по-разному. Обязательная интеграция данных в ХД позволяет решить эту проблему, приведя данные к единому формату.

Поддержка хронологии — данные в ОИД необходимы для выполнения над ними операций в текущий момент времени. Поэтому они могут не иметь привязки ко времени. Для анализа данных часто важно иметь возможность отслеживать хронологию изменений показателей предметной области. Поэтому все данные, хранящиеся в ХД, должны соответствовать последовательным интервалам времени.

Неизменяемость — требования к ОИД накладывают ограничения на время хранения в них данных. Те данные, которые не нужны для оперативной обработки, как правило, удаляются из ОИД для уменьшения занимаемых ресурсов. Для анализа, наоборот, требуются данные за максимально больший период времени. Поэтому, в отличие от ОИД, данные в ХД после загрузки только читаются. Это позволяет существенно повысить скорость доступа к данным как за счет возможной избыточности хранящейся информации, так и за счет исключения операций модификации. При реализации в СППР концепции ХД данные из разных ОИД копируются в единое хранилище. Собранные данные приводятся к единому формату, согласовываются и обобщаются. Аналитические запросы адресуются к ХД (рис. 2.1).

Такая модель неизбежно приводит к дублированию информации в ОИД и в ХД. Однако Инмон в своей работе утверждает, что избыточность данных,