Хранилище данных

2.1. Концепция хранилища данных

Стремление объединить в одной архитектуре системы поддержки принятия решений (СППР) возможностей OLTP-систем и систем анализа, требования к которым во многом противоречивы, привело к появлению концепции хранилищ дан ных (ХД). Первые статьи, посвященные ХД, появились в 1988 г., их авторами были Б. Девлин и П. Мэрфи. В 1992 г. У. Инмон подробно описал данную концепцию в своей монографии "По строение хранилищ данных".

В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа. Это позволяет приме нять структуры данных, которые удовлетворяют требованиям их хранения с учетом использования в OLTP-системах и системах анализа. Такое разделе ние позволяет оптимизировать как структуры данных оперативного хранения (оперативные БД, файлы, электронные таблицы и т. п.) для выполнения опе раций ввода, модификации, удаления и поиска, так и структуры данных, ис пользуемые для анализа (для выполнения аналитических запросов). В СППР эти два типа данных называются соответственно оперативными источника ми данных (ОИД) и хранилищем данных.

В своей работе Инмон дал следующее определение ХД: *Хранилище данных - предметно-ориентированный, интегрированный, неизменчи вый, поддерживающий хронологию набор данных, организованный для целей под держки принятия решений.*

Рассмотрим свойства ХД более подробно.

Предметная ориентация. Это фундаментальное отличие ХД от ОИД. Разные ОИД могут содержать данные, описывающие одну и ту же пред метную область с разных точек зрения (например, с точки зрения бухгал терского учета, складского учета, планового отдела и т. п.). Решение, при нятое на основе только одной точки зрения, может быть неэффективным или даже неверным. ХД позволяют интегрировать информацию, отра жающую 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 смысла хранить информацию о номерах документов купли-продажи, в то время как их со держимое - 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 не сколькими 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 бывает важно иметь воз можность 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 боль шой период времени. Поэтому, в отличие от ОИД, данные в ХД после за грузки только читаются. Это позволяет существенно повысить скорость доступа к данным, как за счет возможной избыточности хранящейся ин формации, 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, со гласовываются и обобщаются. Аналитические запросы адресуются к ХД (рис. 1). Такая модель неизбежно приводит к дублированию информации в ОИД и в ХД. Однако Инмон в своей работе утверждает, что избыточность данных, хранящихся в СППР, не превышает 1 %! Это можно объяснить следующими причинами:

- При загрузке информации из ОИД в ХД данные фильтруются. Многие из них не попадают в ХД, поскольку лишены смысла с точки зрения использования в процедурах анализа.
- Информация в ОИД носит, как правило, оперативный характер, и данные, потеряв актуальность, удаляются. В ХД, напротив, хранится историческая информация. С этой точки зрения дублирование содержимого ХД данными ОИД оказывается

весьма незначительным. В ХД хранится обобщенная ин формация, которая в ОИД отсутствует.

- Во время загрузки в ХД данные очищаются (удаляется ненужная информа ция), и после такой обработки они занимают гораздо меньший объем.

Избыточность информации можно свести к нулю, используя виртуальное ХД. В данном случае в отличие от классического (физического) ХД данные из ОИД не копируются в единое хранилище. Они извлекаются, преобразуются и интегрируются непосредственно при выполнении аналитических запросов в оперативной памяти компьютера. Фактически такие запросы напрямую адре суются к ОИД (рис. 2). Основными достоинствами виртуального ХД явля ются минимизация объема памяти, занимаемой на носителе информацией и работа с текущими, детализированными данными.

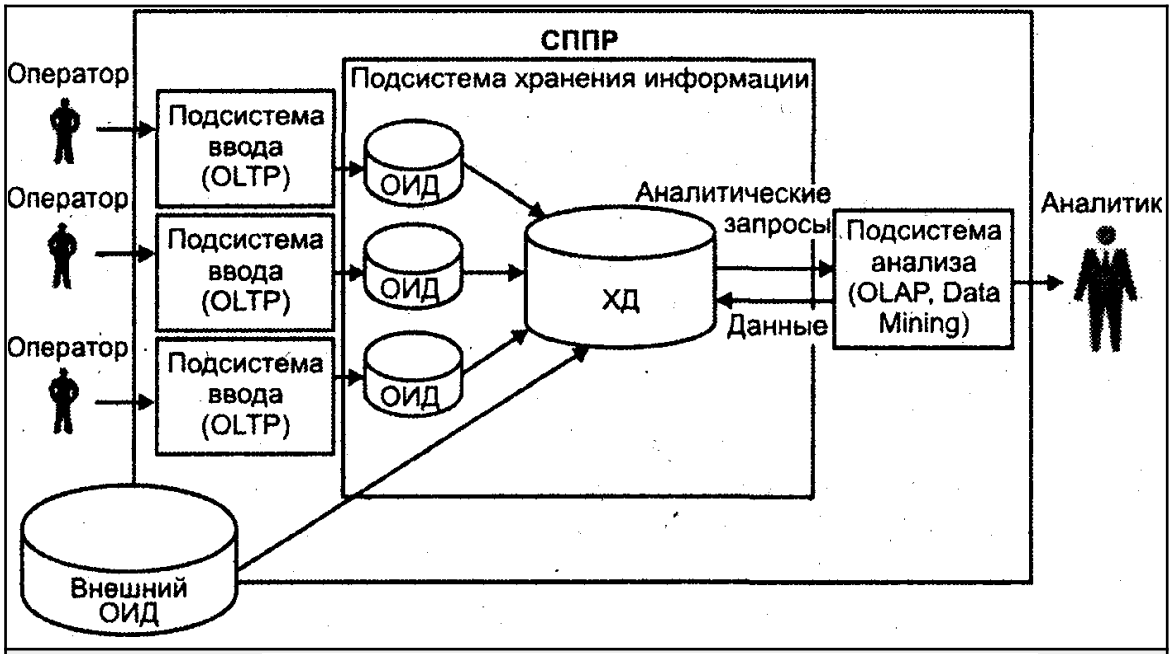

Рис. 1. Структура СППР с физическим ХД

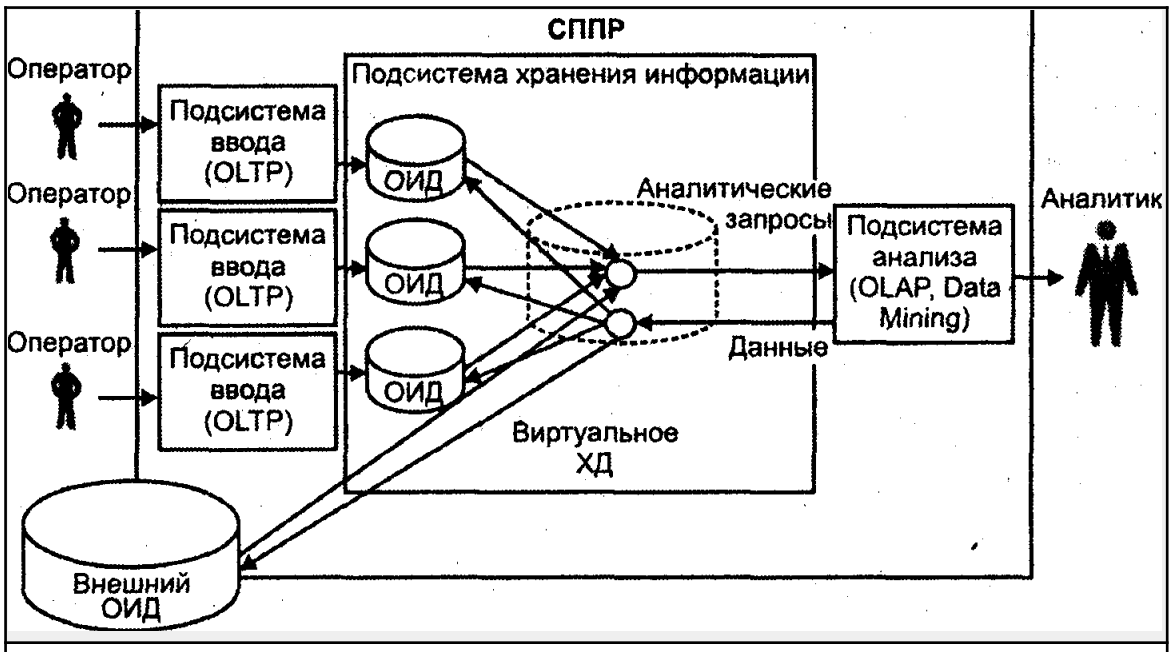

Рис. 2 Структура.СППР с виртуальным ХД

Однако такой подход обладает многими недостатками:

- Время обработки запросов к виртуальному ХД значительно превышает соот ветствующие показатели для физического хранилища. Кроме того, структуры оперативных БД рассчитанные на интенсивное обновление одиночных запи сей, в высокой степени нормализованы. Для выполнения же аналитического запроса требуется объединение большого числа таблиц, что также приводит к снижению быстродействия.
- Интегрированный взгляд на виртуальное хранилище возможен только при выполнении условия постоянной доступности всех ОИД. Таким образом, временная недоступность хотя бы одного из источников может привести ли бо к невыполнению аналитического запроса, либо к неверным результатам. Выполнение сложных аналитических запросов над ОИД требует значитель ных ресурсов компьютеров. Это приводит к снижению быстродействия OLTP-систем, что недопустимо, т. к. время выполнения операций в таких системах часто весьма критично.
- Различные ОИД могут поддерживать разные форматы и кодировки данных. Часто на один и тот же вопрос может быть получено несколько вариантов ответа. Это может быть связано с несинхронностью моментов обновления данных в разных ОИД, отличиями в описании одинаковых объектов и собы тий предметной области, ошибками при вводе, утерей архивов и т. д. В таком случае цель - формирование единого непротиворечивого взгляда на объект управления - может быть не достигнута.

Главным же недостатком виртуального хранилища следует признать практи ческую невозможность получения данных за долгий период времени. При отсутствии физического хранилища доступны только те данные, которые на момент запроса есть в ОИД. Основное назначение OLTP-систем - оператив ная обработка текущих данных, поэтому они не ориентированы на хранение данных за длительный период времени. По мере устаревания данные выгру жаются в архив и удаляются из оперативной БД.

При реализации физического ХД возникают следующие проблемы:

1) Необходимость интеграции данных из неоднородных источников в рас пределенной среде. ХД создаются для интегрирования данных, которые мо гут поступать из разнородных ОИД, физически размещающихся на разных компьютерах: БД, электронных архивов, публичных и коммерческих элек тронных каталогов, справочников, статистических сборников. При создании ХД приходится решать задачу построения системы, согласованно функцио нирующей с неоднородными программными средствами и решениями. При выборе средств реализации ХД приходится учитывать множество факторов, включающих уровень совместимости различных программных компонентов, легкость их освоения и использования, эффективность функционирования и т. д.

2) Потребность в эффективном хранении и обработке очень больших объе мов информации. Свойство неизменности ХД предполагает накопление в нем информации за долгий период времени, что должно поддерживаться по стоянным ростом объемов дисковой памяти. Ориентация на выполнение ана литических запросов и связанная с этим денормализация данных приводят к нелинейному росту объемов памяти, занимаемой ХД при возрастании объема данных. Исследования, проведенные на основе тестового набора TPC-0, по казали, что для баз данных объемом в 100 Гбайт требуется память, в 4,87 раза большая объемом, чем нужно для хранения полезных данных.

3) Необходимость многоуровневых справочников метаданных. Для систем анализа наличие развитых метаданных (данных о данных) и средств их пре доставления конечным пользователям является одним из основных условий успешной реализации ХД. Метаданные необходимы пользователям СППР для понимания структуры информации, на основании которой принимается решение. Например, прежде чем менеджер корпорации задаст системе свой вопрос, он

должен понять, какая информация имеется, насколько она акту альна, можно ли ей доверять, сколько времени может занять формирование ответа и т. д. При создании ХД необходимо решать задачи хранения и удоб ного представления метаданных пользователям,

4) Повышение требований к безопасности данных. Собранная вместе и со гласованная информация об истории развития корпорации, ее успехах и не удачах, о взаимоотношениях с поставщиками и заказчиками, об истории и состоянии рынка дает возможность анализа прошлой и текущей деятельности корпорации и построения прогнозов для будущего. Очевидно, что подобная информация является конфиденциальной и доступ к ней ограничен в преде лах самой компании, не говоря уже о других компаниях. Для обеспечения безопасности данных приходится решать вопросы аутентификации пользова телей, защиты данных при их перемещении в хранилище данных из опера тивных баз данных и внешних источников, защиты данных при их передаче по сети и т. п.

Снизить затраты на создание ХД можно создавая его упрощенный вариант - витрину данных (Data Mart). *Витрина данных (ВД) - это упрощенный вариант ХД, содержащий только тема тически объединенные данные.*

ВД максимально приближена к конечному пользователю и содержит данные, тематически ориентированные на него (например, ВД для работников отдела маркетинга может содержать данные, необходимые для маркетингового ана лиза). ВД существенно меньше по объему, чем ХД, и для ее реализации не требуется больших затрат. Они могут быть реализованы как самостоятельно, так и вместе с ХД. Самостоятельные ВД (рис. 3) часто появляются в организации исторически и встречаются в крупных организациях с большим количеством независимых подразделений, решающих собственные аналитические задачи.

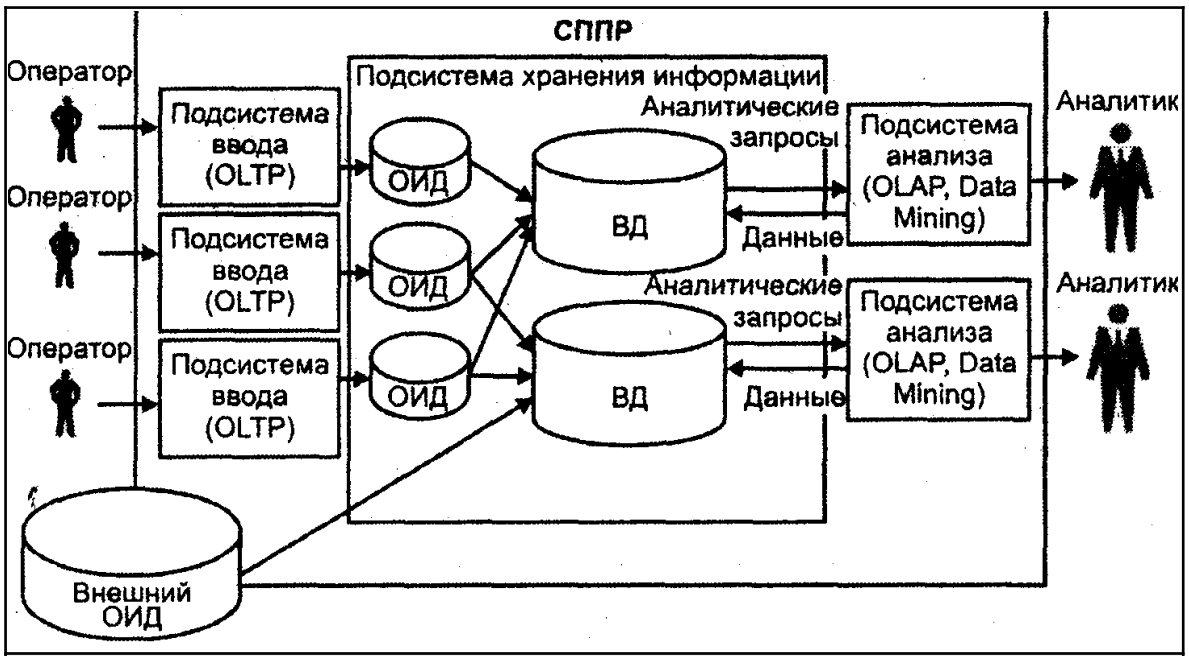

Рис. 3. Структура СППР с самостоятельными ВД

Достоинствами такого подхода являются: быстрое решение определенного круга вопросов, быстрое внедрение автономных ВД и получение отдачи, упрощение процедур заполнения ВД и повышение их производительности..

Недостатками автономных ВД являются: многократное хранение данных в разных ВД, что приводит к увеличению расходов на их хранение и потенциальным проблемам, связанным с необ ходимостью поддержания непротиворечивости данных; отсутствие консолидированности данных на уровне предметной области, а следовательно - отсутствие единой картины.

В последнее время все более популярной становится идея совместить ХД и ВД в одной системе. В этом случае ХД используется в качестве единственно го источника интегрированных данных для всех ВД (рис. 4).

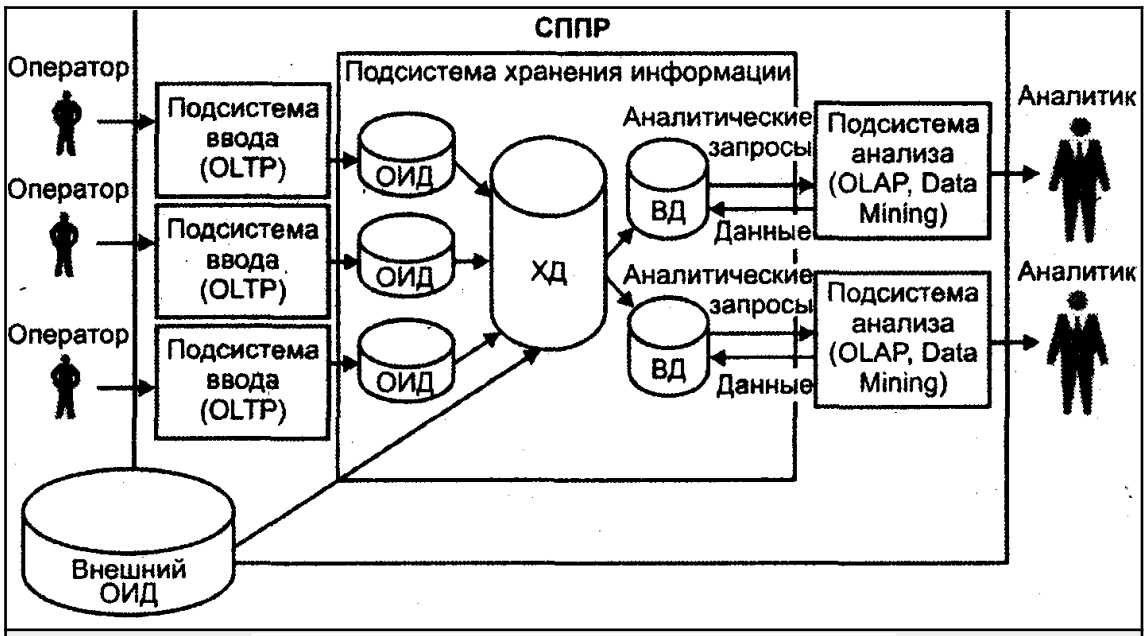

Рис. 4. Структура СППР с ХД и ВД

ХД представляет собой единый централизованный источник информации для всей предметной области, а ВД являются подмножествами данных из храни лища, организованными для представления информации по тематическим разделам данной области. Конечные пользователи имеют возможность дос тупа к детальным данным хранилища, если данных в витрине недостаточно, а также для получения более полной информационной картины. Достоинствами такого подхода являются:

- простота создания и наполнения ВД, поскольку наполнение происходит из единого стандартизованного надежного источника очищенных данных - из ХД;
- простота расширения СППР за счет добавления новых ВД;
- снижение нагрузки на основное ХД.

К недостаткам относятся избыточность (данные хранятся как в ХД, так и в ВД) и дополнительные затраты на разработку СППР с ХД и ВД.

2.2. Организация переноса данных из ОИД в ХД

Самый мощный из информационных потоков, циркулирующих в ХД - входной - связан с перено сом данных из ОИД. Обычно информация не просто копируется в ХД, а под вергается 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 свя зано 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). Программ ные средства, обеспечивающие его выполнение, называются ETL-системами. Рассмотрим более подробно этапы ETL-процесса:

Извлечение данных. Чтобы начать ETL-процесс, необходимо извлечь дан ные из одного или нескольких источников и подготовить их к этапу преобра зования. Можно выделить два способа извлечения данных:

1. Извлечение данных вспомогательными программными средствами непо средственно из структур хранения информации (файлов, электронных таблиц, БД и т. П). Достоинствами такого способа извлечения данных явля ются:

- отсутствие необходимости расширять OLTP-систему (это особенно важно, если ее структура закрыта);
- данные могут извлекаться с учетом потребностей процесса переноса.

2. Выгрузка данных средствами OLTP-систем в промежуточные структуры. Достоинствами такого подхода являются:

- возможность использовать средства OLTP -систем, адаптированные к структурам данных;
- средства выгрузки изменяются вместе с изменениями OLTP-систем;
- возможность выполнения первого шага преобразования данных за счет определенного формата промежуточной структуры хранения данных.

Преобразование данных. После того как сбор данных завершен, необходимо преобразовать их для размещения на новом месте. На этом этапе выполняют ся следующие процедуры:

- обобщение данных (агрегация) - перед загрузкой данные обобщаются. Процедура обобщения заменяет многочисленные детальные данные отно сительно небольшим числом агрегированных данных. Детальные данные разделяют на два класса: измерения и факты. Измере ниями называются наборы данных, необходимые для описания событий. Фактами называются данные, отражающие сущность события. Агрегированные данные получаются из детальных данных путем их сум мирования по измерениям. Для быстрого доступа к наиболее часто запра шиваемым агрегированным данным

они должны сохраняться в ХД, а не вычисляться при выполнении запросов;

- перевод значений (трансформация) - в ОИД данные часто хранятся в за кодированном виде для того, чтобы сократить избыточность данных и па мять для их хранения;
- создание полей - при создании полей для конечных пользователей создается и новая информация. Например, ОИД содержит одно поле для указания количества проданных товаров, а второе - для указания цены одного экземпляра. Для исключения операции вычисления стоимости всех товаров можно создать специальное поле для ее хранения во время преобразования данных;
- очистка данных - направлена на выявление и удаление ошибок и несоответствий в данных с целью улучшения их качества. Очистка данных включает следующие этапы: выявление проблем в дан ных, определение правил очистки, тестирование правил очистки, непо средственная очистка данных.

Загрузка данных. После того как данные преобразованы для размещения в ХД, осуществляется этап их загрузки. При загрузке выполняется запись пре образованных детальных и агрегированных данных. Кроме того, при записи новых детальных данных часть старых данных может переноситься в архив.