




Информационно-аналитические системы



Предпосылки

- ▶ Развитие компьютеров и телекоммуникаций привело к резкому ускорению процессов обмена данными и их накопления.
- ▶ Когда объем хранилища исчисляется миллионами объектов (или даже десятками тысяч) встает вопрос о том, как обеспечить эффективную работу с этими данными.
- ▶ Ручной анализ становится невозможен в полном объеме.
- ▶ Для решения проблемы часть поисково-аналитических задач была переложена на средства автоматизации – информационные системы.

Возможности и проблемы

- ▶ С одной стороны чем большим объемом информации мы располагаем, тем больше у нас возможностей для получения и анализа интересующих нас фактов, событий и явлений.
- ▶ Но с другой стороны это приводит и к увеличению времени поиска нужной информации, анализа результатов поиска, к увеличению объема «шума» в данных.
- ▶ На основании этих предпосылок возник целый класс информационных систем – системы поддержки принятия решений.

Определения и основные задачи

- ▶ **СППР** (системы поддержки принятия решений) – класс систем, автоматизировано выполняющих функции анализа и представления данных.
- ▶ Основные задачи СППР:
 - ▶ ввод данных;
 - ▶ хранение данных;
 - ▶ анализ данных.
- ▶ **СППР** — это системы, обладающие средствами ввода, хранения и анализа данных, относящихся к определенной предметной области, с целью поиска решений.

Еще одно определение

- ▶ **Система поддержки принятия решений** (Decision Support Systems, DSS) – это компьютерная система, которая путем сбора и анализа большого количества информации может влиять на процесс принятия решений организационного плана в бизнесе и предпринимательстве. [TAdviser]

Дополнительные задачи СППР (1)

- ▶ Система также должна обеспечить:
 - ▶ сбор информации (включая интерфейс для ее ввода);
 - ▶ надежное хранение всей поступающей информации;
 - ▶ преобразовать информацию в вид, пригодный для ее анализа человеком.
- ▶ Иными словами: **система обеспечивает аналитикам инструменты** для эффективного выполнения задач анализа данных.

Дополнительные задачи СППР (2)

- ▶ Система сама не генерирует правильные решения, а только предоставляет аналитику данные в соответствующем виде для дальнейшего изучения и анализа.
- ▶ Данные помогают повысить эффективность принимаемых человеком решений, но не ставят задачу сделать всю работу за аналитика.

Классификация СППР по степени «интеллектуальности» решаемых задач

- ▶ **Информационно-поисковые** – поддерживаются возможности поиска.
 - ▶ Нацелены на выполнение заранее определенных запросов.
- ▶ **Оперативно-аналитические** – производится группировка и обобщение данных в произвольном виде, необходимом для анализа в текущий момент.
 - ▶ в данном случае заранее неизвестно, какие запросы будет необходимо выполнить в ходе анализа.
- ▶ **Интеллектуальный** – осуществляется поиск закономерностей в накопленных данных, построение моделей и правил описывающих выявленные закономерности и/или прогнозирующих дальнейшее развитие некоторых процессов.

Как можно их использовать (1)

- ▶ Интерактивные системы позволяют руководителям:
 - ▶ получить полезную информацию из первоисточников;
 - ▶ проанализировать ее;
 - ▶ выявить существующие бизнес-модели для решения определенных задач.

Как можно их использовать (2)

▶ Например:

- ▶ можно проследить за всеми доступными информационными активами;
- ▶ получить сравнительные значения объемов продаж;
- ▶ спрогнозировать доход организации при возможном внедрении новой технологии;
- ▶ и т.п.

Классификация СППР по способу взаимодействия с пользователем

- ▶ **Пассивные системы** – не позволяют выдвинуть конкретное предложение, хотя реализуют средства, в различной степени поддерживающие пользователя при поиске наиболее эффективного решения.
- ▶ **Активные системы** – непосредственно участвуют в поиске и подготовке наиболее оптимального решения.
- ▶ **Кооперативные системы** – предоставляют пользователю возможность доработать найденные ими решение, а затем проверить внесенные пользователем коррективы.

Классификация СППР по способу поддержки

- ▶ **Модельно-ориентированные** – выполняют поиск оптимальных решений основываясь на специально разработанных моделях (статистические, финансовые и т.п.).
- ▶ **Ориентированные на данные** – организуют поддержку пользователя при поиске эффективных решений, агрегируя большие объемы данных из гетерогенных источников.
- ▶ **Ориентированные на знания** – выполняют поиск оптимальных решений основываясь на специально разработанной базе знаний.

Классификация СППР по сфере использования

- ▶ **Настольные** – небольшие системы, ориентированные на использование одним пользователем, работающим на персональном компьютере.
- ▶ **Общесистемные** – используют в своей работе большие хранилища данных и ориентированы на использование многими пользователями.

Состав СППР (1)

- ▶ **Подсистема ввода данных.** В таких подсистемах, называемых OLTP (Online transaction processing), выполняется операционная (транзакционная) обработка данных. Для реализации этих подсистем используют обычные системы управления базами данных (СУБД).
- ▶ **Подсистема хранения.** Для реализации данной подсистемы используют СУБД и концепцию хранилищ данных.

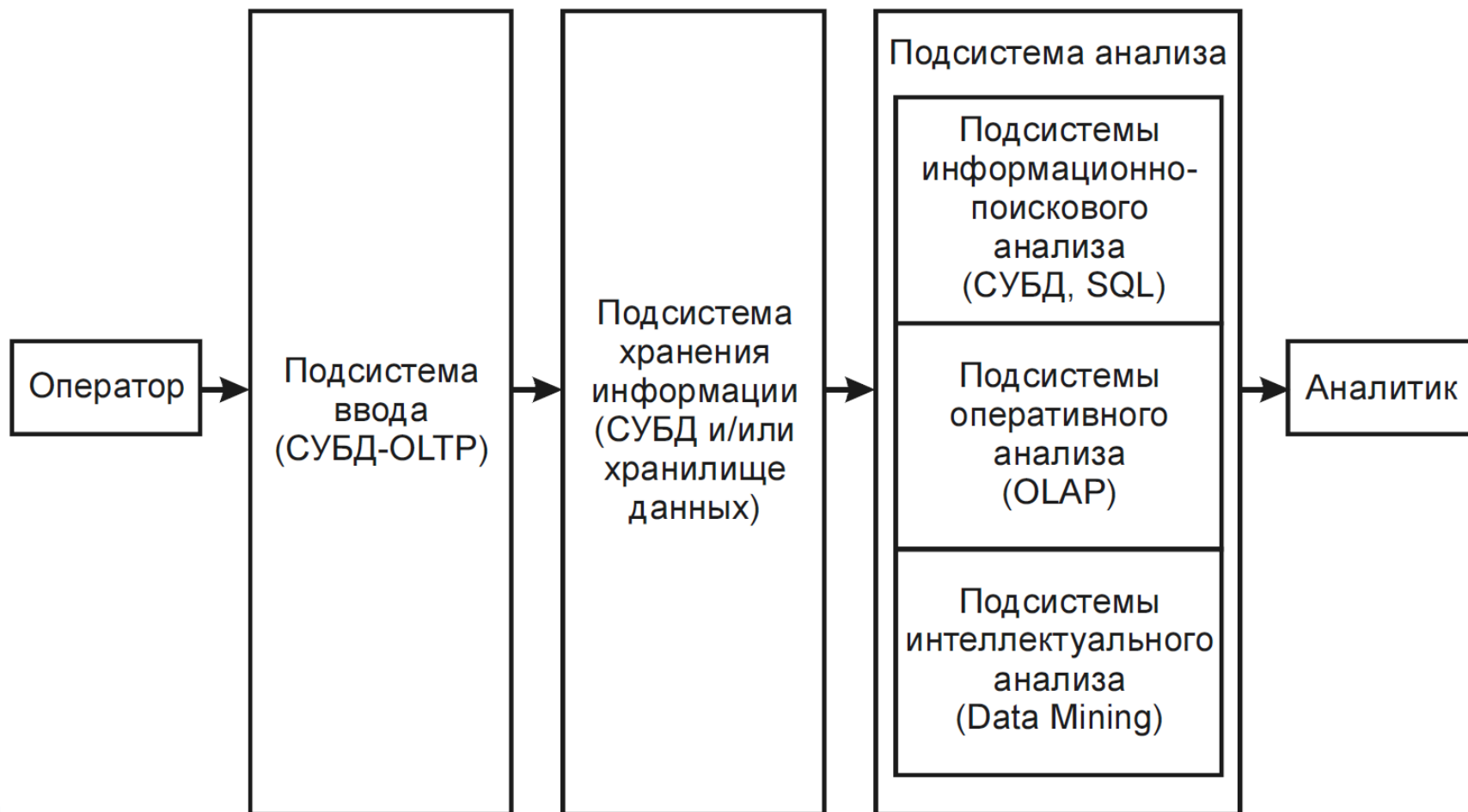
Состав СППР (2)

- ▶ **Подсистема анализа.** Данная подсистема может быть построена на основе:
 - ▶ **подсистемы информационно-поискового анализа** на базе реляционных СУБД и статических запросов с использованием языка структурных запросов SQL (Structured Query Language);

Состав СППР (3)

- ▶ **подсистемы оперативного анализа** – для реализации таких подсистем применяется технология оперативной аналитической обработки данных **OLAP (On-line analytical processing)**, использующая концепцию многомерного представления данных;
- ▶ **подсистемы интеллектуального анализа** – данная подсистема реализует методы и алгоритмы **Data Mining** ("добыча данных").

Обобщенная архитектура СППР



Использование двух СУБД

- ▶ Из схемы явно видно, что **СУБД присутствует в двух элементах**, а следовательно речь идет о двух разных базах данных в пределах одной системы.
- ▶ На первый взгляд назначение второго хранилища данных получается не вполне понятным.
- ▶ Фактически в подсистеме хранения **происходит дублирование информации**, которая уже хранится в подсистеме ввода.
- ▶ Причина этого – противоречие между требованиями, которые предъявляются к хранению и обработке данных в подсистеме сбора и теми, которые необходимы для обеспечения эффективного анализа значительных объемов данных.

- ▶ База данных – основа СППР.

- ▶ **Информационно-аналитическая система (ИАС)** – это комплекс программно-технических средств, информационных ресурсов, методик, которые используются для автоматизации аналитических работ с целью обоснования принятия управленческих решений и других возможных применений.

Разница в требованиях к системам анализа и OLTP (1)

Характеристика	Требования к OLTP	Требования к системе анализа
Степень детализации хранимых данных	Хранение только детализированных данных	Хранение как детализированных, так и обобщенных данных
Качество данных	Допускаются неверные данные из-за ошибок ввода	Не допускаются ошибки в данных
Формат хранения данных	Может содержать данные в разных форматах в зависимости от приложений	Единый согласованный формат хранения данных

Разница в требованиях к системам анализа и OLTP (2)

Характеристика	Требования к OLTP	Требования к системе анализа
Допущение избыточных данных	Должна обеспечиваться максимальная нормализация	Допускается контролируемая денормализация (избыточность) для эффективного извлечения информации
Управление данными	Должна быть возможность в любое время добавлять, удалять и изменять данные	Должна быть возможность периодически добавлять данные
Количество хранимых данных	Должны быть доступны все оперативные данные, требующиеся в данный момент	Должны быть доступны все данные, накопленные в течение продолжительного интервала времени

Разница в требованиях к системам анализа и OLTP (3)

Характеристика	Требования к OLTP	Требования к системе анализа
Характер запросов к данным	Доступ к данным пользователей осуществляется по заранее составленным запросам	Запросы к данным могут быть произвольными и заранее неоформленными.
Время обработки обращений к данным	Время отклика системы измеряется в секундах	Время отклика системы может составлять несколько минут
Характер вычислительной нагрузки на систему	Постоянно средняя загрузка процессора	Загрузка процессора формируется только при выполнении запроса, но на 100%

► Хранилища данных

Основные определения

- ▶ **Хранилище данных (ХД)** — предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.
- ▶ **Оперативные источники данных (ОИД)** – источники импорта данных для ХД.

Предметная ориентированность

- ▶ Это **наиболее фундаментальное отличие ХД от ОИД** (оперативного источника данных).
- ▶ Разные ОИД могут содержать данные, описывающие одну и ту же предметную область с различных точек зрения (например бух. учет, склад, плановый отдел и т.д.).
- ▶ ХД позволяет интегрировать информацию отображающую разные точки зрения на одну предметную область.
- ▶ Также это **позволяет хранить только необходимые для анализа данные**, не дублируя лишнюю информацию.

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 могут **не иметь жесткую привязку** ко времени, поскольку операции над ними происходят в текущий момент.
- ▶ Для полноценного анализа как правило четкая хронология событий является необходимой.
- ▶ ХД решают эту проблему за счет приведения всех дат к единому формату, тем самым обеспечивая возможность хронологического сопоставления событий.

Неизменяемость

- ▶ Во многих ОИД для минимизации объема хранимых данных зачастую устанавливается определенный срок, после которого исторические данные могут быть совсем удалены из системы.
- ▶ В ХД необходимо хранить весь накопленный массив информации.
- ▶ После загрузки данные в ХД только считываются, но не изменяются.
- ▶ Это позволяет в том числе существенно увеличить скорость доступа к данным за счет исключения операций по их модификации.

Организация ХД

- ▶ Хранилища данных разделяются на 3 основные категории:
 - ▶ детальные данные;
 - ▶ агрегированные данные;
 - ▶ метаданные.

Детальные данные

- ▶ **Детальные данные** переносятся непосредственно из ОИД и соответствуют элементарным событиям фиксируемым OLTP-системами.
- ▶ Принято делить все данные на:
 - ▶ **измерения** – наборы данных, необходимые для описания события. Например – площадь квартиры, удаленность от метро и центра города, этаж, кол-во комнат и т.п.;
 - ▶ **факты** – данные отражающие суть события, например, цена продажи квартиры.
- ▶ В процессе эксплуатации если потребность в детальных данных снижается они могут храниться в архивах в сжатом виде на отдельных носителях.

Агрегированные данные

- ▶ На основании детальных данных в ХД могут быть получены агрегированные (обобщенные) данные.
- ▶ В зависимости от возможности агрегации данные делятся на:
 - ▶ **аддитивные** – фактические числовые данные, которые могут просуммированы по всем измерениям;
 - ▶ **полу-аддитивные** – числовые фактические данные, которые могут быть просуммированы только по определенным измерениям;
 - ▶ **неаддитивные** – фактические данные которые не могут быть просуммированы.

Использование агрегированных данных

- ▶ Для обеспечения максимально оперативного доступа к агрегированным данным **часть из них хранится непосредственно в ХД.**
- ▶ Если проводить все расчеты налету, то это очень длительная по времени и накладная с точки зрения загрузки ресурсов операция.
- ▶ Очевидно, что это **ведет к избыточности информации** и увеличению размеров ХД – **важно добиться оптимального соотношения** между вычисляемыми и хранящимися агрегированными данными.
- ▶ **Данные, которые требуются более часто, должны храниться в ХД.**