

# ОБРАБОТКА И ХРАНЕНИЕ ИНФОРМАЦИИ

---

К.с.н., доцент А.Н. Тарасова

# Формы представления данных

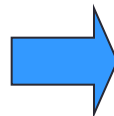
➔ **Неструктурированные данные**

➔ **Слабоструктурированные данные**

➔ **Структурированные данные**

\* Данные – сведения, которые характеризуют систему, явление, процесс или объект, представленные в определенной форме и предназначенные для дальнейшего использования.

625003, г. Тюмень,  
ул. Ленина, д.16, к.105



Поле	Значение
Индекс	625003
Город	Тюмень
Улица	Ленина
Дом	16
Квартира	105

**Столбцы** (переменные, атрибуты, признаки)

**Строки**  
(записи,  
прецеденты,  
примеры,  
операции,  
наблюдения)

					...	
					...	
					...	
					...	
					...	
...	...	...	...	...	...	...

# Представления наборов данных

→ Упорядоченные наборы данных

1	Дата	Количество	Сумма
2	01.01.2007	4	283,31
3	01.01.2007	1	173,32
4	01.01.2007	1	72,48
5	02.01.2007	12	405,76
6	02.01.2007	6	303,13
7	03.01.2007	6	521,16
8	03.01.2007	3	156,96
9			

→ Неупорядоченные наборы данных

1	Оператор	Машина	Дефект
2	Иванов	5	Сбой нарезки
3	Сидоров	3	Неточность выреза
4	Иванов	1	Дефект балансира
5	Петров	5	Неточность выреза
6			

→ Транзакционные данные


Одна транзакция

1	Код транзакции	Товар
2	10200	Йогурт "Чудо"
3	10200	Сахарный песок
4	10201	Батон "Рязанский"
5	10201	Сок "Добрый"
6		

# Виды данных

- Непрерывные данные – данные, значения которых могут принимать какое угодно значение в некотором интервале.
- Дискретные данные – значения признака, общее число которых конечно, может быть подсчитано при помощи натуральных чисел.

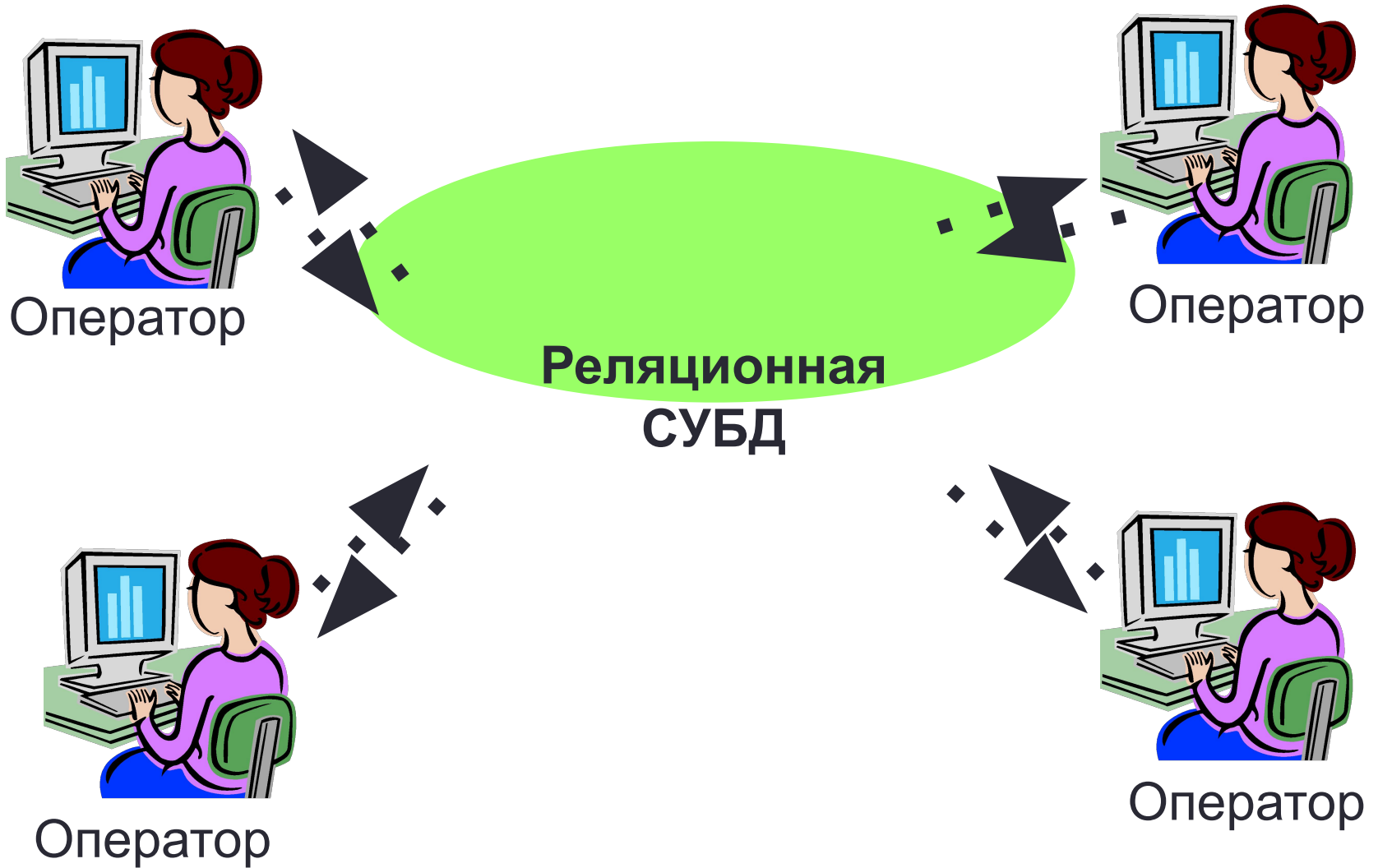
# Основные форматы данных

- Целый (количество товара, код товара и т.п.)
  - Вещественный (цена, скидка и т.п.)
  - Строковый (Фамилия, наименование, адрес и т.п.)
  - Логический
  - Дата/время
- 
- **Категориальные**  
(номинальные)
  - **Упорядоченные**  
(ординальные, порядковые)

- **Данные транзакций** – это сведения, хранимые с целью отслеживания взаимодействий, или бизнес-транзакций, выполняемых организацией.
- **Системы оперативной обработки транзакций (OLTP)** - служат для хранения данных о выполняемых бизнес-транзакциях и призваны поддерживать повседневную деятельность организации.



# OLTP-система



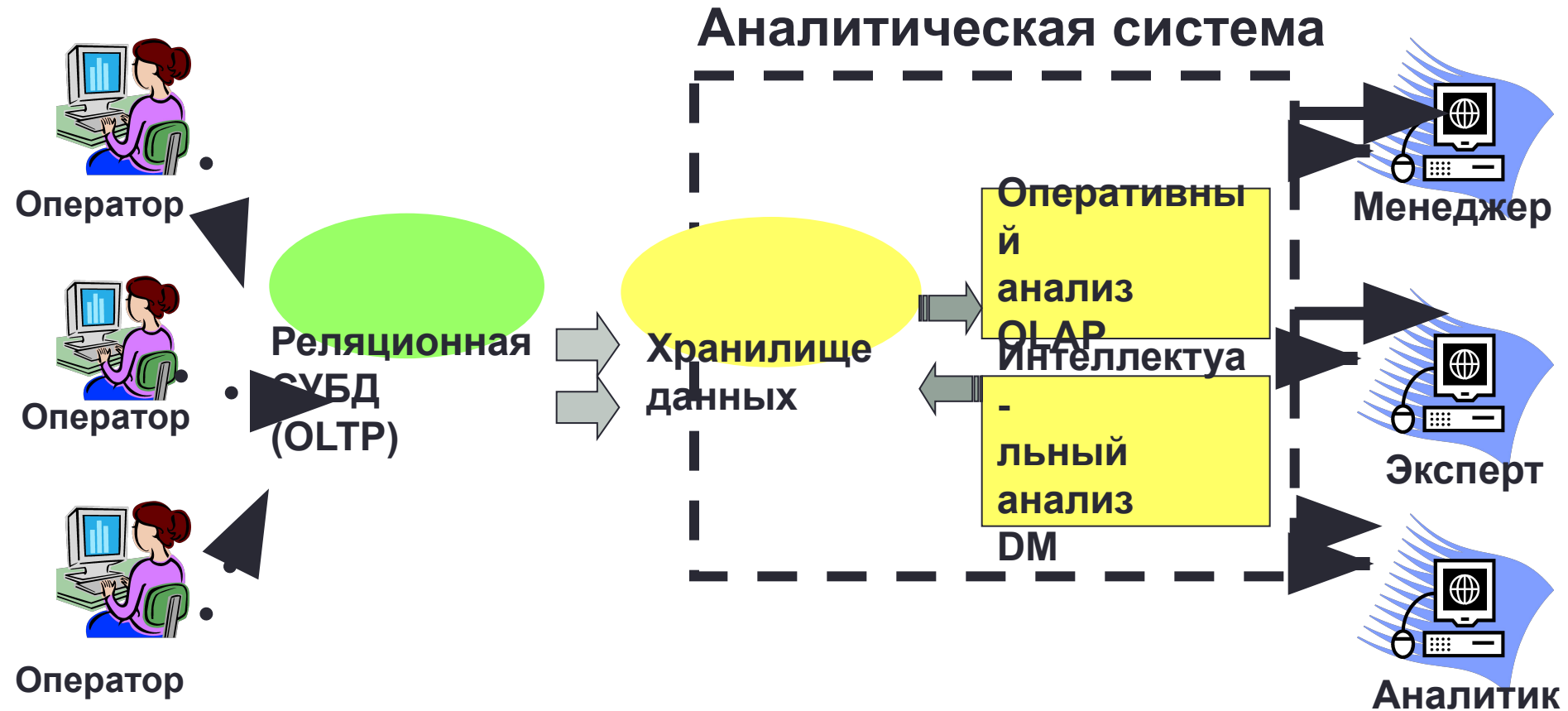
## Проблемы использования данных из OLTP-систем для бизнес-анализа

- Гигантский объем обрабатываемых транзакций.
- Вмешательство в бизнес-операции.
- Архивирование.
- Использование различных систем оперативной обработки транзакций

# Системы поддержки принятия решений (СППР)

- Информационные системы, ориентированные на аналитическую обработку данных с целью получения знаний, необходимых для разработки решений в области управления.

# Структура СППР



# Отличия СППР и OLTP-систем

Свойство	OLTP-система	СППР
Цели использования данных	Быстрый поиск, простейшие алгоритмы обработки	Аналитическая обработка с целью поиска скрытых закономерностей, построения прогнозов, моделей и т.д.
Уровень обобщения (детализации) данных	Детализированные	Как детализированные, так и обобщенные (агрегированные)
Формат хранения данных	Данные могут храниться в различных форматах в зависимости от приложения, в котором они были созданы	Данные хранятся и обрабатываются в едином формате
Время хранения данных	В пределах отчетного периода	Годы, десятилетия

# Отличия СППР и OLTP-систем

Свойство	OLTP-система	СППР
Требования к качеству данных	Возможны некорректные данные (ошибки регистрации, ввода и др)	Ошибки в данных не допускаются, т.к. могут привести к некорректной работе аналитических алгоритмов
Изменение данных	Данные могут добавляться, изменяться и удаляться	Допускается только пополнение; ранее добавленные данные изменяться не должны, что позволяет обеспечить их хронологию.
Доступ к данным	Должен быть обеспечен доступ ко всем текущим (оперативным) данным	Должен быть обеспечен доступ к историческим (т.е. накопленным за достаточно длительный период времени) данным с соблюдением их хронологии

# Отличия СППР и OLTP-систем

Свойство	OLTP-система	СППР
Периодичность обновления	Часто, но в небольших объемах	Редко, но в больших объемах
Характер выполнения запросов	Стандартные, настроенные заранее	Нерегламентированные, формируемые аналитиком «на лету» в зависимости от требуемого анализа
Время выполнения запроса	Несколько секунд	До нескольких минут и даже более

# Консолидация данных - это

Комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.



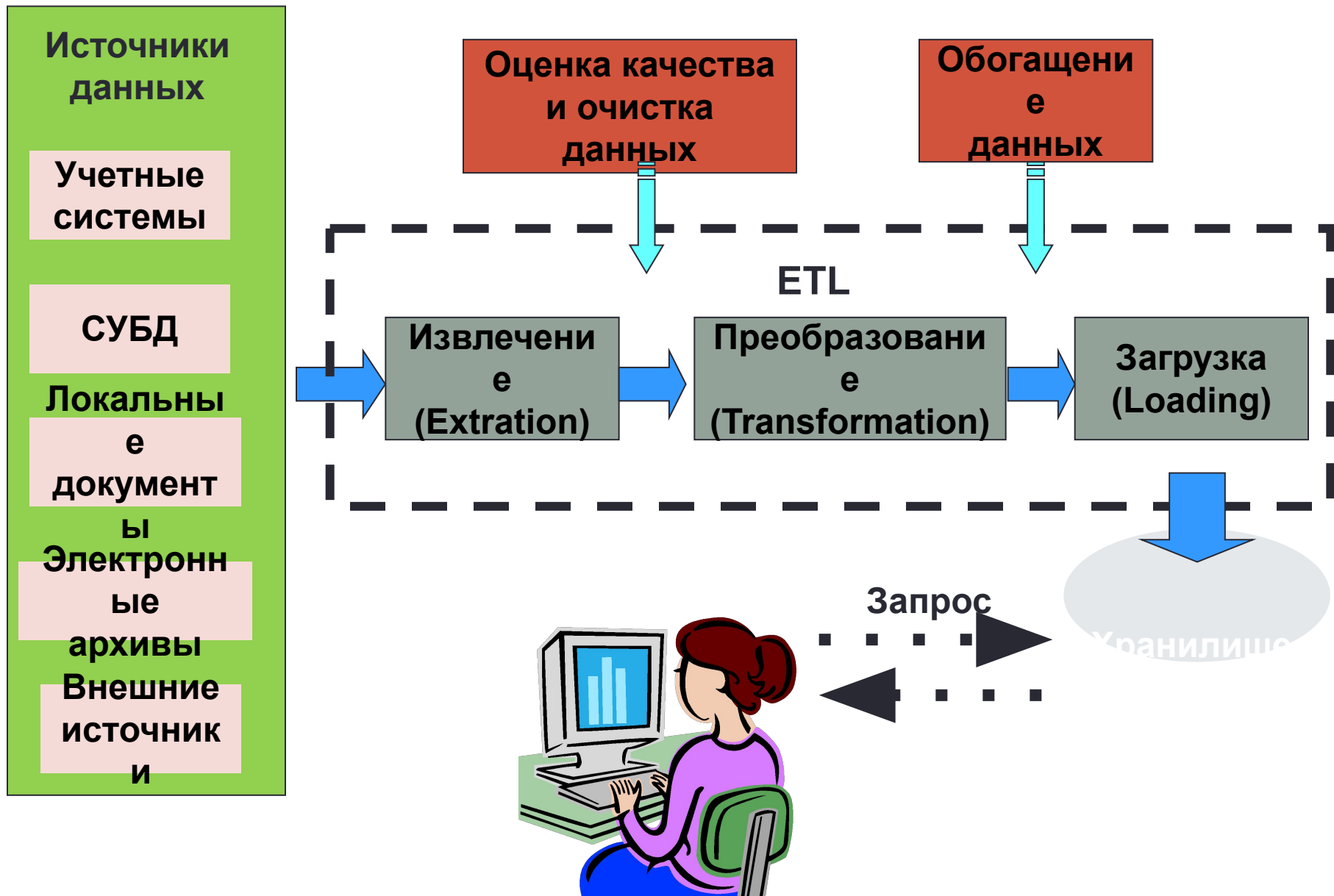
## Критерии оптимальности данных:

- Обеспечение высокой скорости доступа к данным.
- Компактность хранения.
- Автоматическая поддержка целостности структуры данных.
- Контроль непротиворечивости данных.

# Задачи консолидации данных:

- **Выбор источников данных:** данные, хранящиеся в отдельных (локальных) файлах; базы данных различных СУБД; специализированные хранилища данных (ХД).
- **Разработка стратегии консолидации.**
- **Оценка качества данных.**
- **Обогащение.**
- **Очистка.**
- **Перенос в хранилище данных.**

# Обобщенная схема процесса консолидации



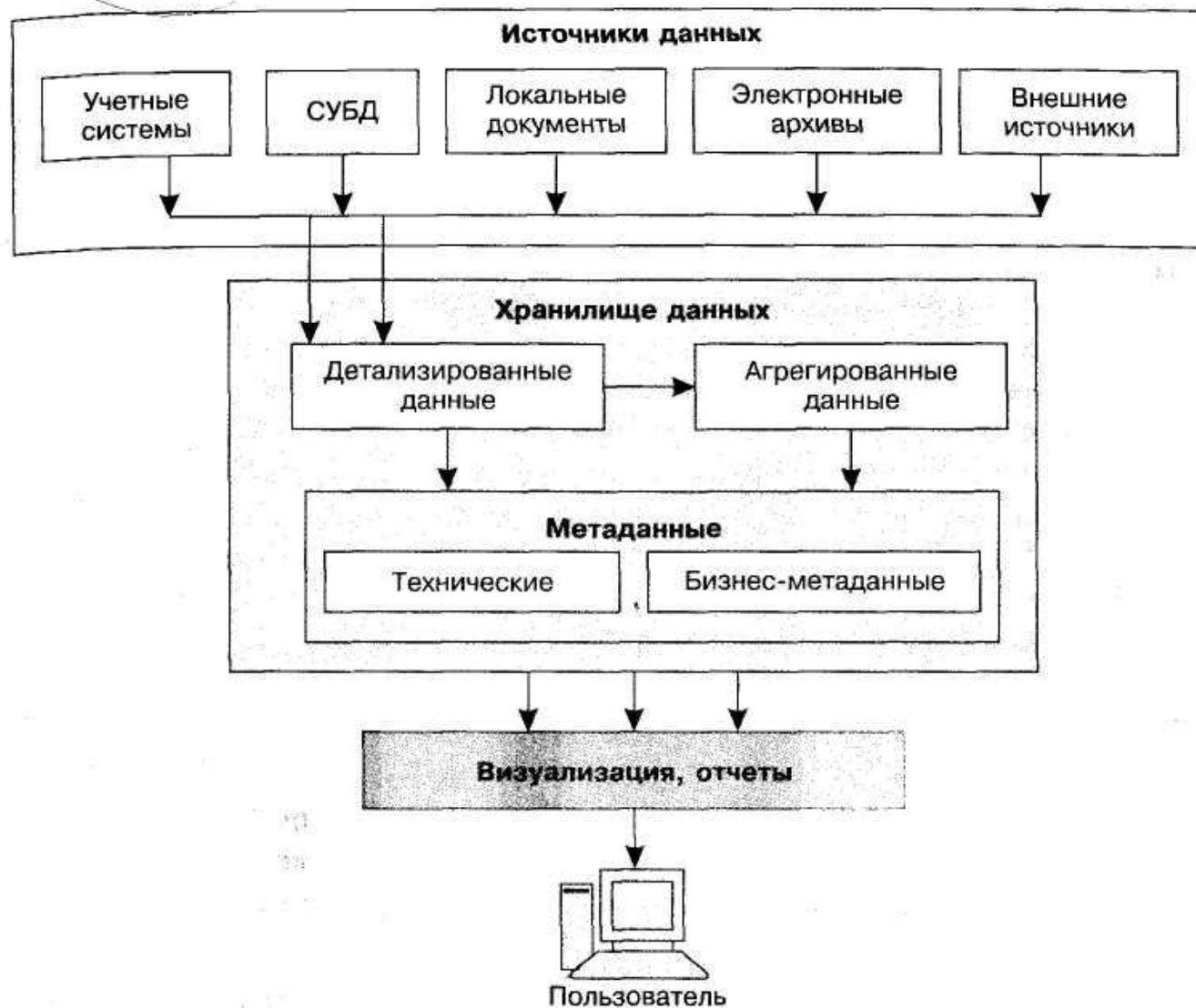
# Хранилище данных -

- Разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов.
- Предметно-ориентированный, интегрированный, неизменяемый и поддерживающий хронологию набор данных, предназначенный для обеспечения принятия управленческих решений.

# Основные требования к ХД

- Высокая скорость получения данных из хранилища;
- Автоматическая поддержка внутренней непротиворечивости данных;
- Возможность получения и сравнения срезов данных;
- Наличие удобных средств для просмотра данных в хранилище;
- Обеспечение целостности и достоверности хранящихся данных.

# Концептуальная схема ХД



# Использование ХД:

- Регулярные отчеты;
- Нерегламентированные отчеты;
- Интеллектуальный анализ данных.

# Архитектуры ХД

## ROLAP

Структура куба  
(многомерное  
хранилище)

Предобработанные  
агрегаты  
(реляционное  
хранилище)

Детализированные  
значения  
(реляционное  
хранилище данных)

## MOLAP

Структура куба  
(многомерное  
хранилище)

Предобработанные  
агрегаты  
(многомерное  
хранилище)

Детализированные  
значения  
(многомерное  
хранилище данных)

## HOLAP

Структура куба  
(многомерное  
хранилище)

Предобработанные  
агрегаты  
(многомерное  
хранилище)

Детализированные  
значения  
(реляционное  
хранилище данных)



**Спасибо за внимание!**