

**СППР, хранилища и  
витрины данных,  
интеллектуальный  
анализ данных**

- «Заглядывай вперед или окажешься позади»

Бенджамин Франклин

- «Планировать – это хлопотать по поводу наилучшего метода получения случайного результата»

Амброз Бирс

- « Решить – смириться с перевесом одних внешних влияний над другими»

Амброз Бирс

- « Человека, который преуспел в руководстве, но не искушен в выполнении трех интеллектуальных функций управления (формирование политики, принятие решений и контроль), можно сравнить с циркачом на одноколесном велосипеде – он демонстрирует виртуозные трюки во время представления, но мальчик посыльный на обычном велосипеде движется более устойчиво и перевозит полезный груз»

Стаффорд Бир

- « Планирование – это проектирование желаемого будущего и эффективных путей его достижения.

Это орудие мудрых, но не одних только их.

В руках же мелких людей оно часто превращается в бесполезный ритуал, который порождает кратковременную успокоенность, а не творит будущее, к которому стремятся.

Лучшие образцы планирования являются в такой же степени творениями искусства, как и науки. Здесь, как нигде, важно их гармоническое сочетание.»

Р.Л. Акофф

# Технология Data Mining

(также называемая **Knowledge Discovery in Data**)

изучает процесс нахождения новых, действительных и потенциально полезных знаний в базах данных.

Data Mining лежит на пересечении нескольких наук, главные из которых - это системы баз данных, статистика и искусственный интеллект.

# Системы поддержки принятия решений - СППР

(DSS, Decision Support Systems)

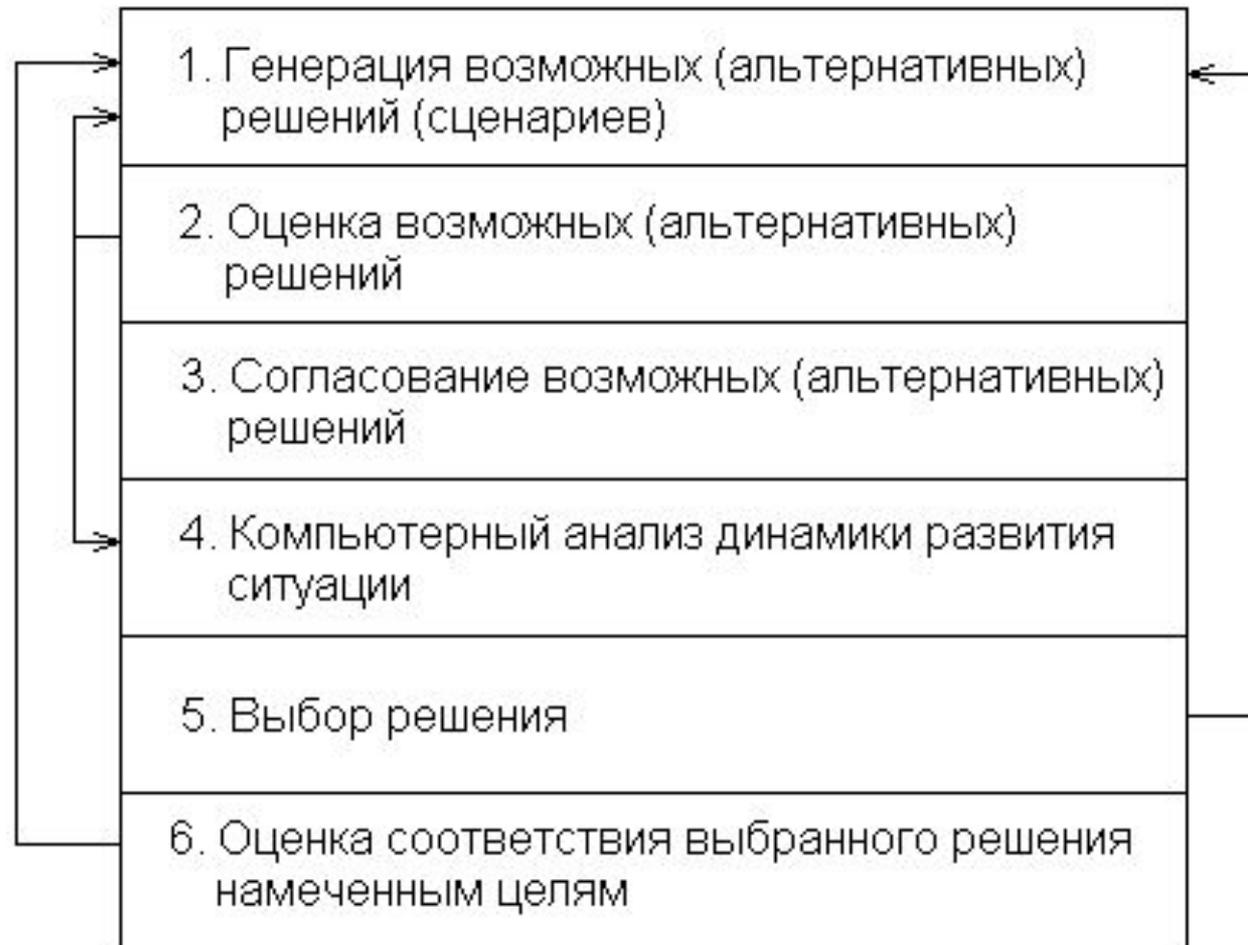
Основная задача СППР - предоставить аналитикам инструмент для выполнения анализа данных.

Необходимо отметить, что для эффективного использования СППР ее пользователь-аналитик должен обладать соответствующей квалификацией.

Система не генерирует правильные решения, а только предоставляет аналитику данные в соответствующем виде (отчеты, таблицы, графики и т. п.) для изучения и анализа.

СППР решают три основные задачи: сбор, хранение и анализ хранимой информации.

# Компьютерный анализ ситуаций, создаваемый СППР

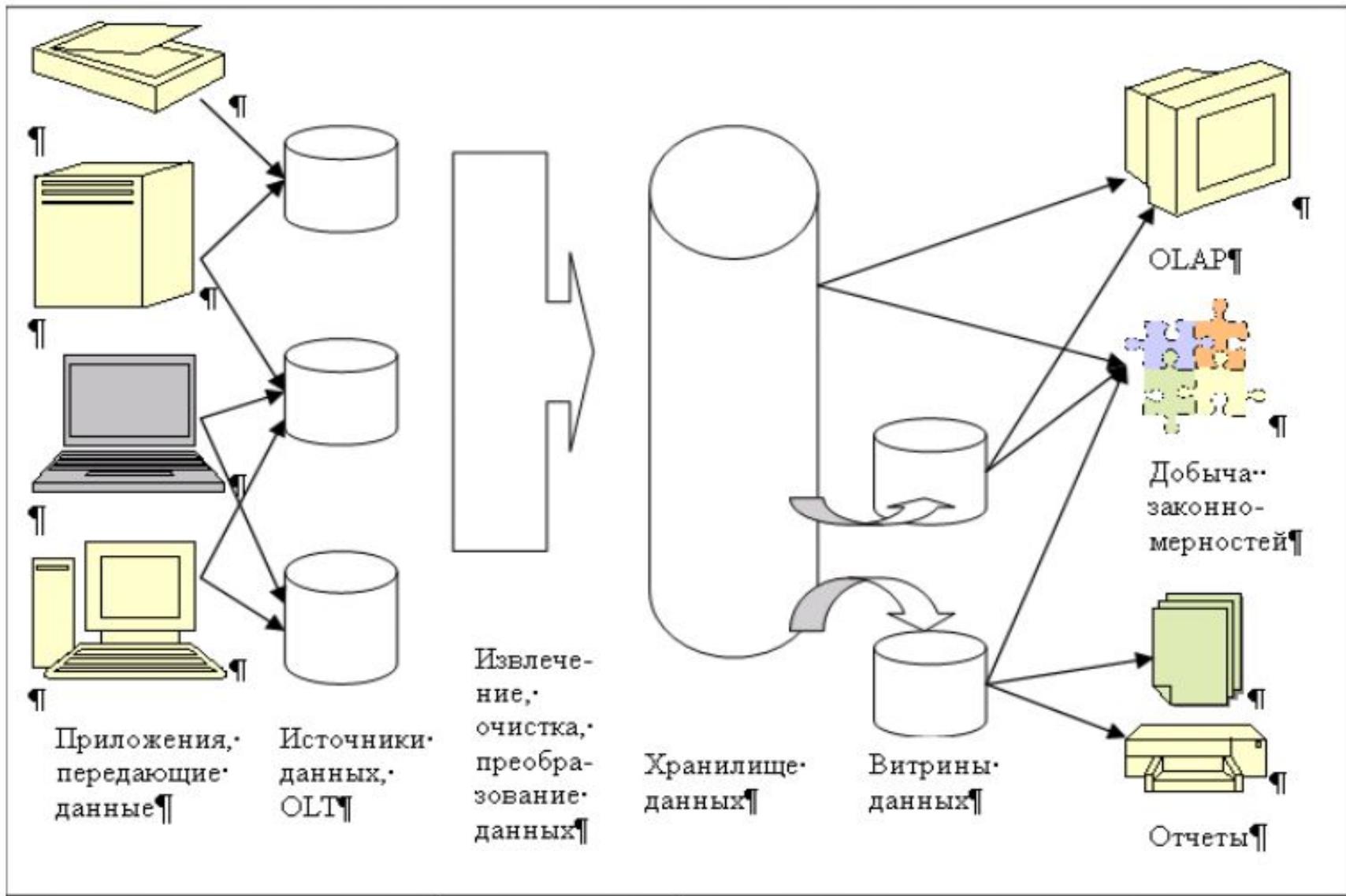


# Классы задач анализа данных

**Информационно-поисковый:** СППР осуществляет поиск необходимых данных. Характерной чертой такого анализа является выполнение заранее определенных запросов.

**Оперативно-аналитический:** СППР производит группирование и обобщение данных в любом виде, необходимом аналитику. В отличие от информационно-поискового анализа в данном случае невозможно заранее предсказать необходимые аналитику запросы. Применяется многомерное представлений данных.

**Интеллектуальный:** СППР осуществляет поиск функциональных и логических закономерностей в накопленных данных, построение моделей и правил, которые объясняют найденные закономерности и/или прогнозируют развитие некоторых процессов (с определенной вероятностью).





Обобщенная архитектура системы поддержки принятия решений

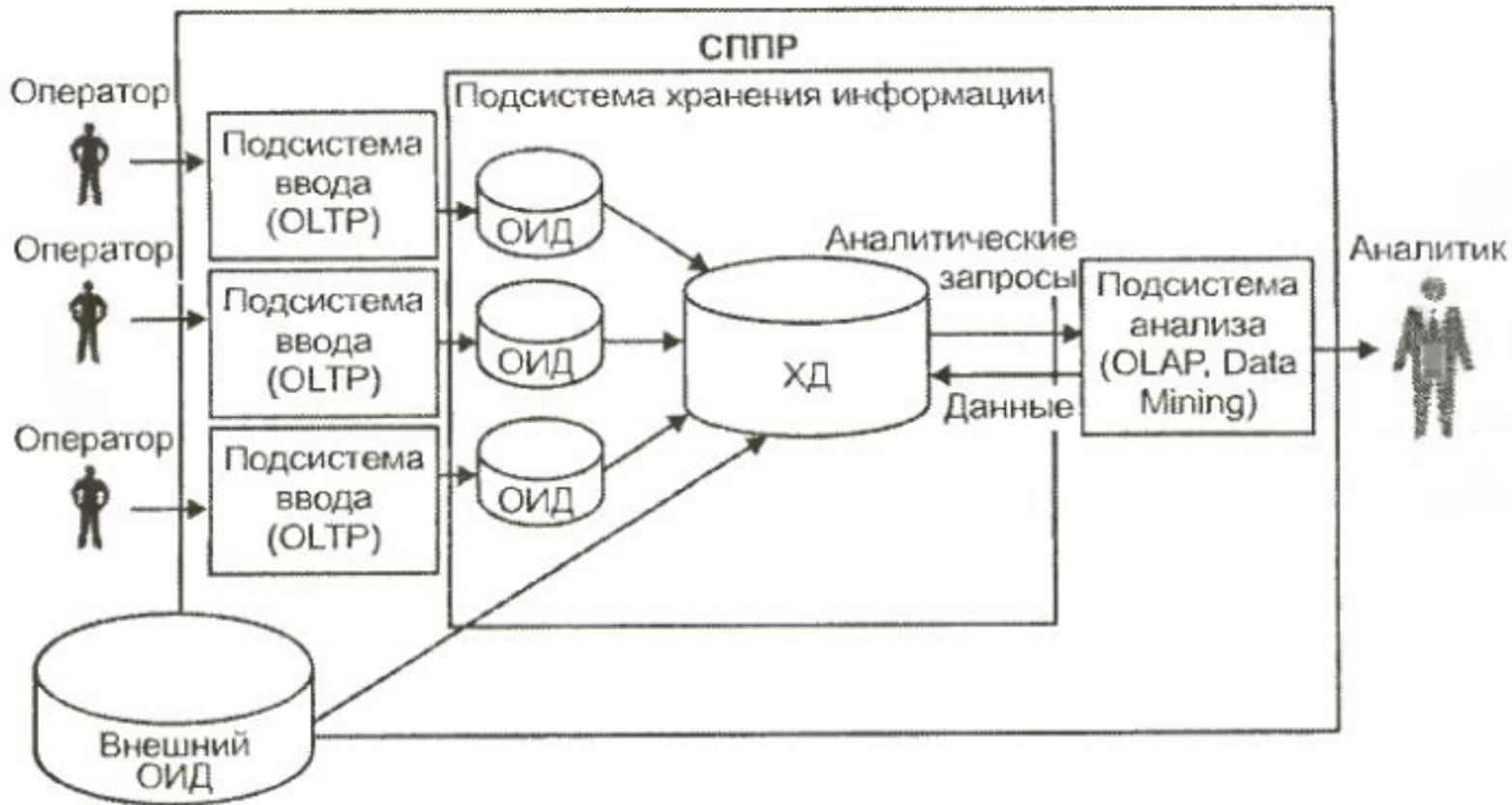
OLTP (Online Transaction Processing), транзакционная система — обработка транзакций в реальном времени. Способ организации БД, при котором система работает с небольшими по размерам транзакциями, но идущими большим потоком, и при этом клиенту требуется от системы минимальное время отклика.

Термин OLTP применяют также к системам (приложениям). OLTP-системы предназначены для ввода, структурированного хранения и обработки информации (операций, документов) в режиме реального времени.

# Хранилища данных

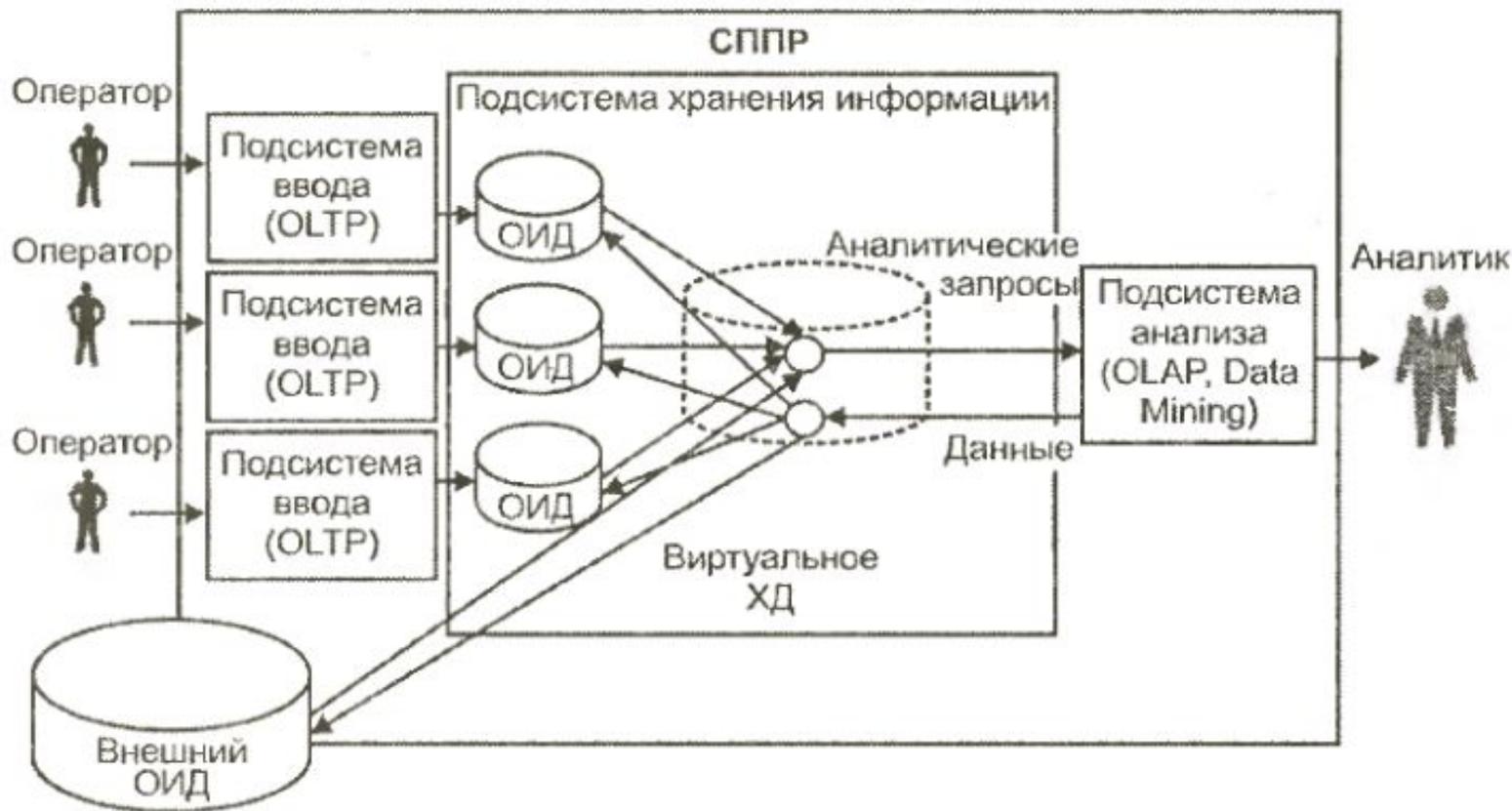
В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа.

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



Структура СДПР с физическим ХД

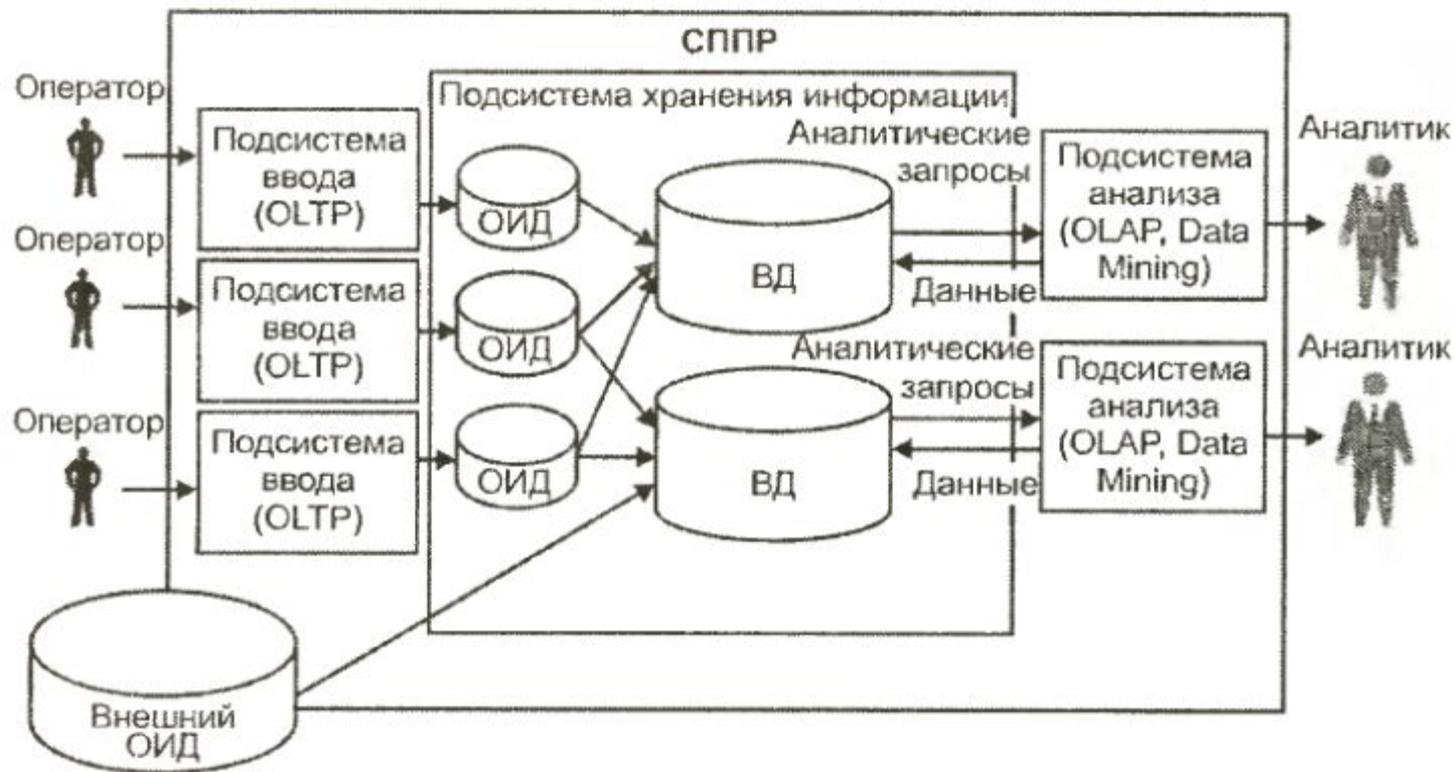
ОИД - оперативные источники данных



Структура СДПР с виртуальным ХД

# Проблемы создания физического ХД:

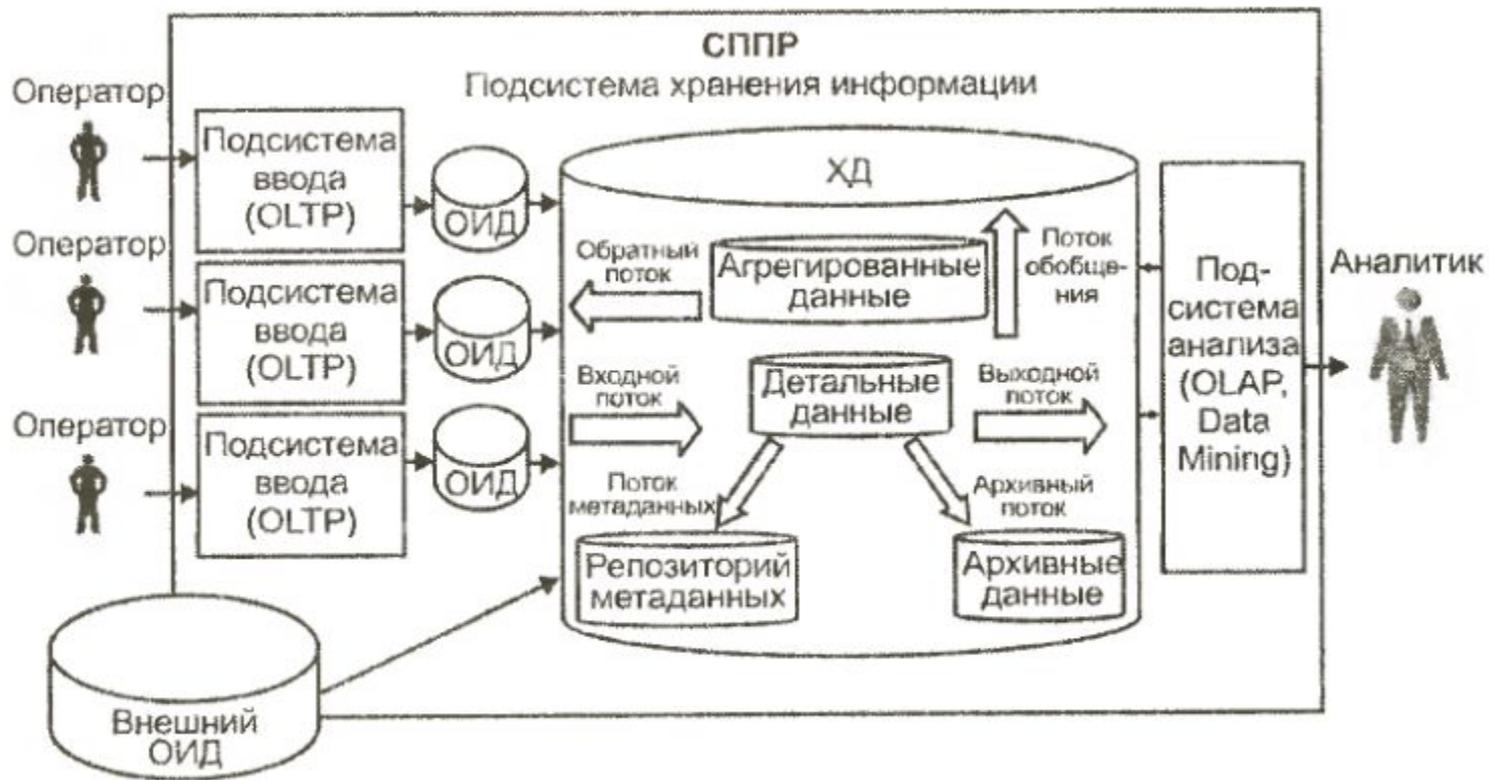
- необходимость интеграции данных из неоднородных источников в распределенной среде;
- потребность в эффективном хранении и обработке очень больших объемов информации;
- необходимость наличия многоуровневых справочников метаданных;
- повышенные требования к безопасности данных.



Структура СППР с самостоятельными ВД

**Витрина данных (ВД)** - это упрощенный вариант ХД, содержащий только тематически объединенные данные.

# Архитектура ХД



Архитектура ХД

# Состав ХД

**Детальными** являются данные, переносимые непосредственно из ОИД. Они соответствуют элементарным событиям, фиксируемым OL TP системами. (Например, продажи, эксперименты и др.). Принято разделять все данные на измерения и факты.

**Измерениями** называются наборы данных, необходимые для описания событий (например, города, товары, люди и т. п.).

**Фактами** называются данные, отражающие сущность события (например, количество проданного товара, результаты экспериментов и т. п.).

На основании детальных данных могут быть получены **агрегированные** (обобщенные) данные.

# Состав ХД

Для удобства работы с ХД необходима информация о содержащихся в нем данных. Такая информация называется метаданными (данные о данных).

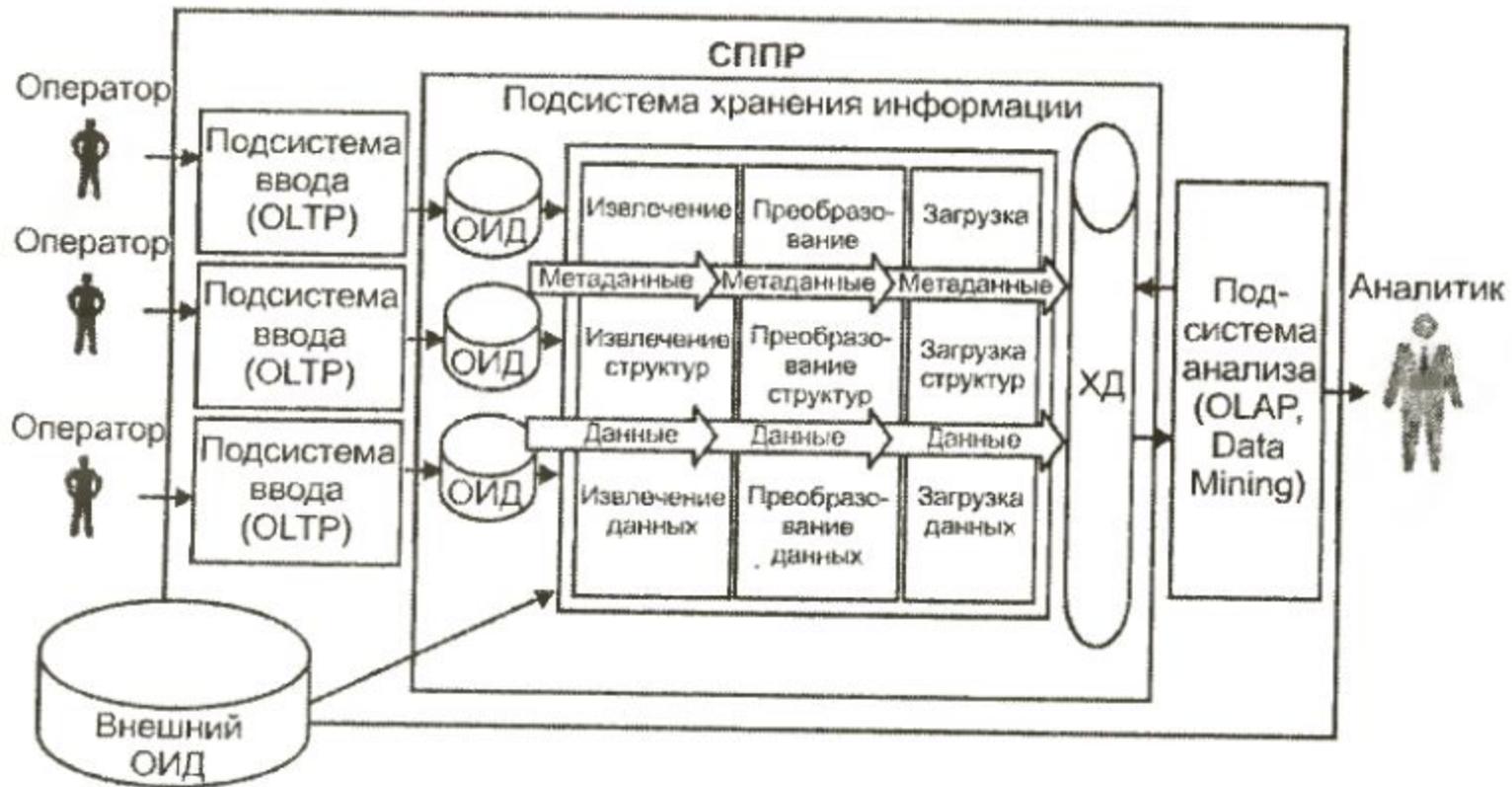
Согласно концепции Дж. Захмана, метаданные должны отвечать на следующие вопросы

- что (описание объектов),
- кто (описание пользователей),
- где (описание места хранения),
- как (описание действий),
- когда (описание времени)
- и почему (описание причин).

# Информационные потоки в ХД

- **Входной поток (Inflow)** образуется данными, копируемыми из оперативных источников данных (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 переносом информации о данных в репозиторий данных;
- **выходной поток (Outflow)** образуется данными, извлекаемыми пользователями;
- **обратный поток (Feedback Flow)** образуется очищенными данными, записываемыми обратно в ОИД.

# ETL- процесс (Extraction, Transformation, Loading)



ETL-процесс

# Очистка данных

Уровень ячейки таблицы:

- Орфографические ошибки (опечатки)
- Отсутствие данных
- Фиктивные значения
- Логически неверные значения
- Закодированные значения
- Составные значения

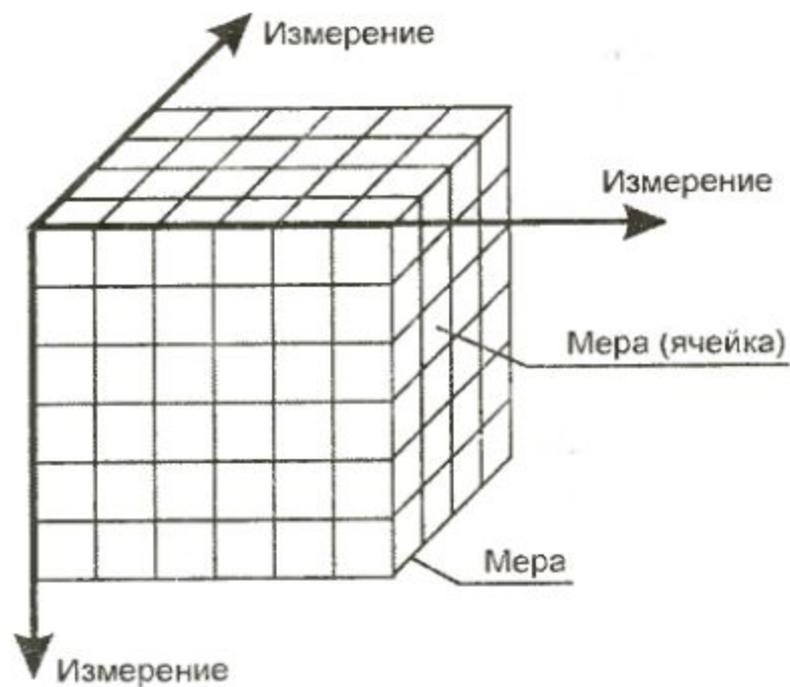
# OLAP-системы

## Многомерная модель данных

**Измерение** - это последовательность значений одного из анализируемых параметров. Например, для параметра "время" это последовательность календарных дней, для параметра "регион" это может быть список городов.

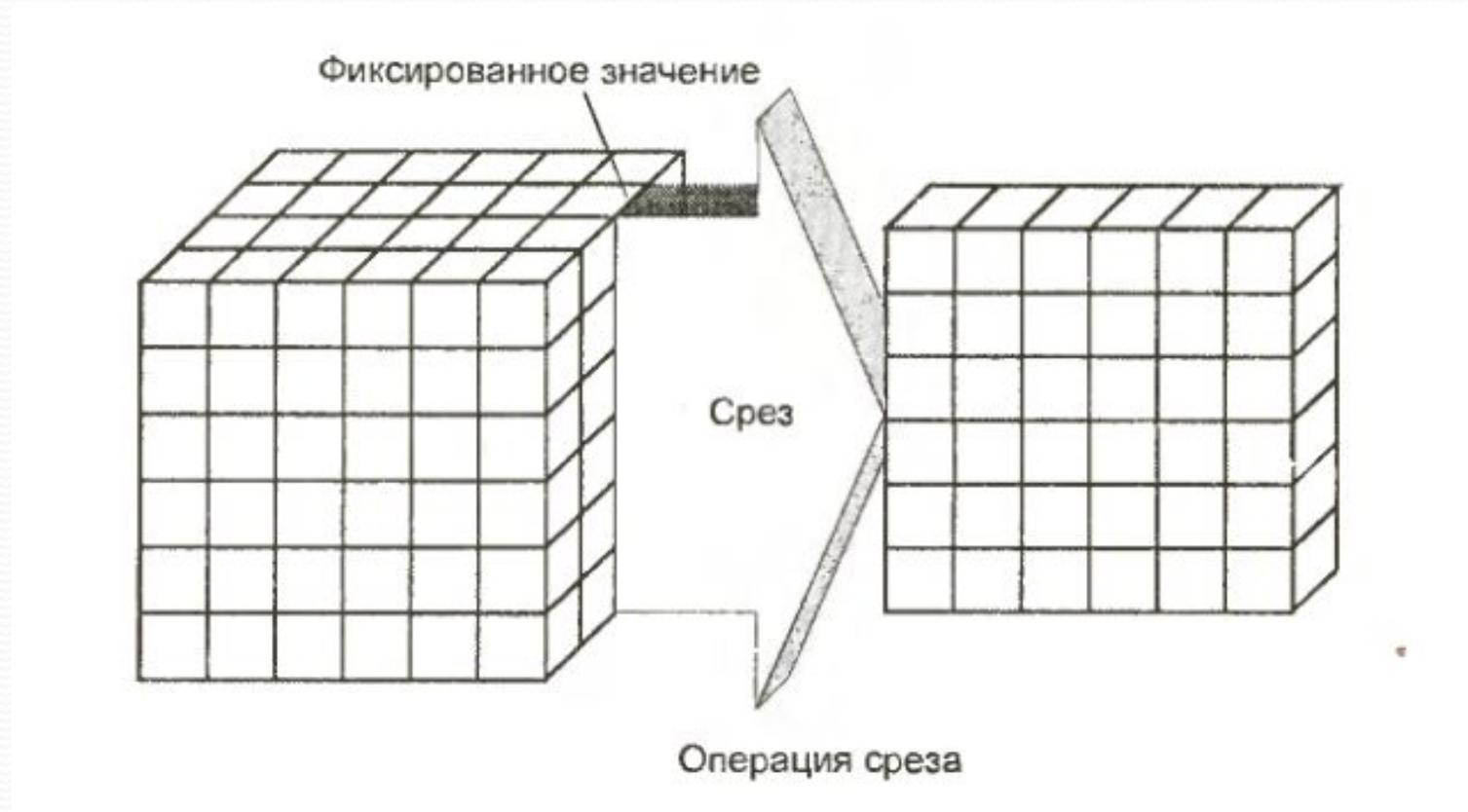
По ученому Кодду, **многомерное концептуальное представление** (multidimensional conceptual view) - это множественная перспектива, состоящая из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Одновременный анализ по нескольким измерениям определяется как многомерный анализ.

# Гиперкуб

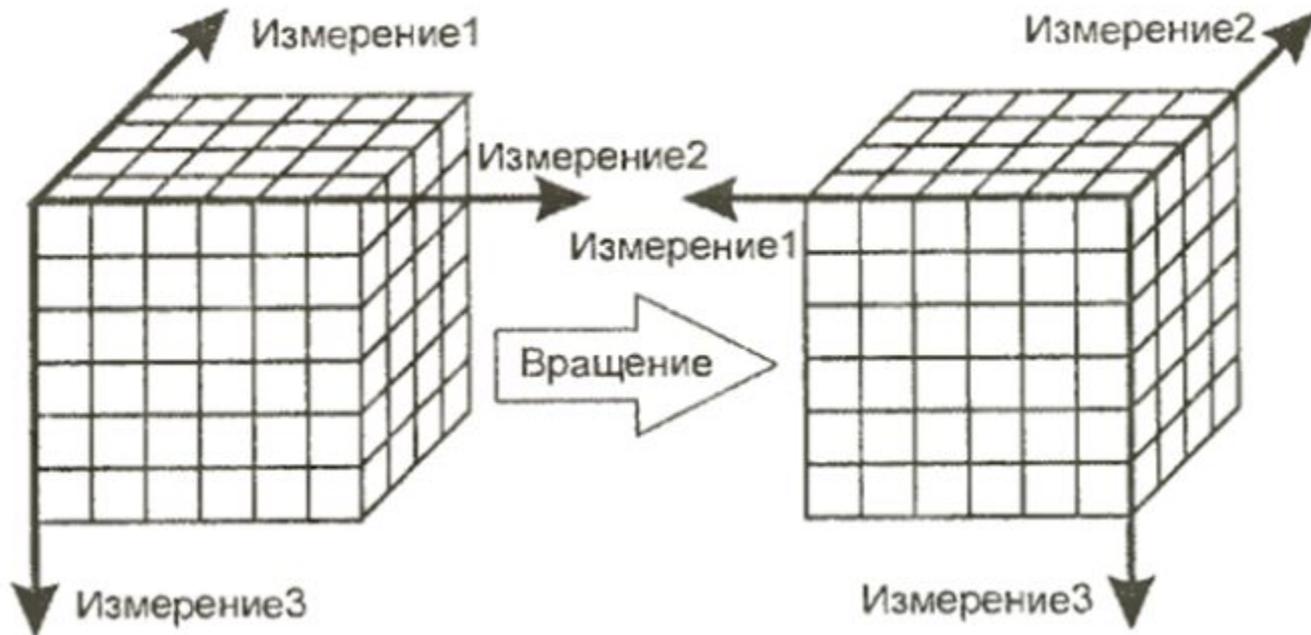


Представление данных в виде гиперкуба

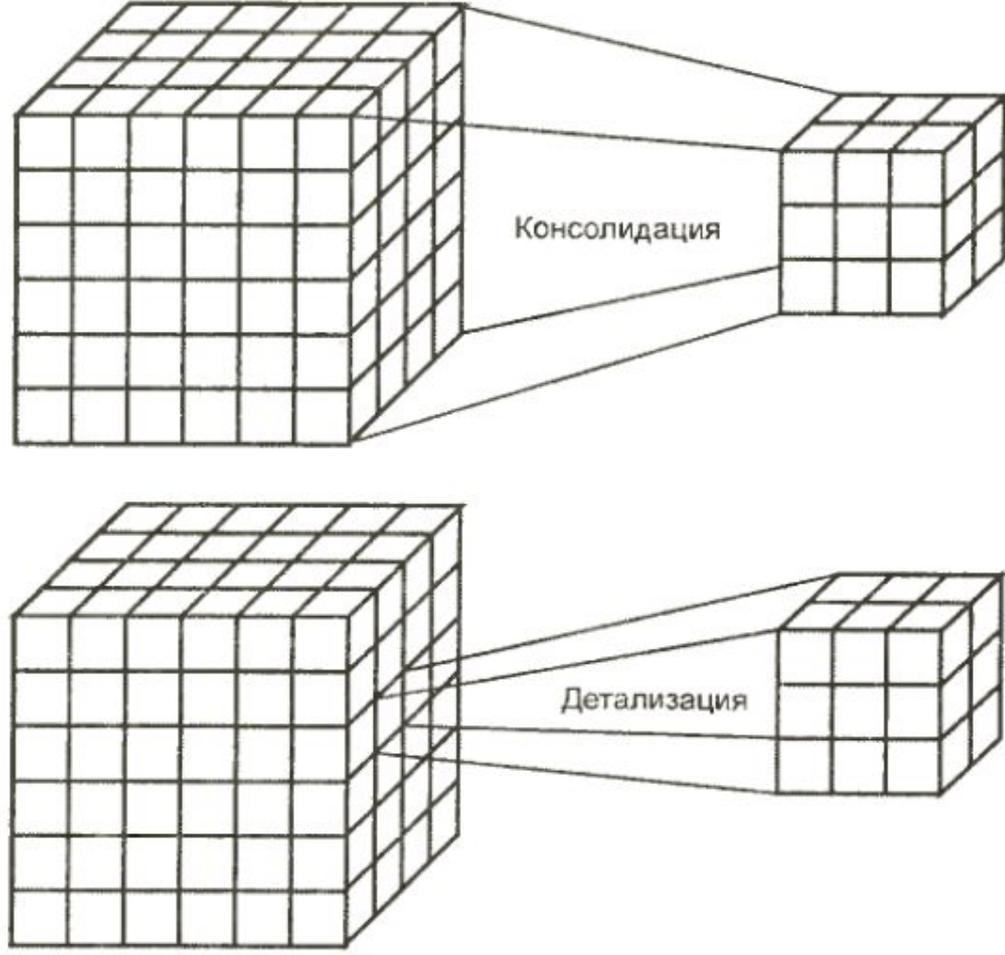
# Операция среза (slice)



# Операция вращения (rotate)



# Консолидация (Drill Up) и детализация (Drill Down)



Операции консолидации и детализации

# Двенадцать правил Кодда

1. Многомерность
2. Прозрачность
3. Доступность
4. Постоянная производительность при разработке отчетов
5. Клиент-серверная архитектура
6. Равноправие измерений
7. Динамическое управление разреженными матрицами.
8. Поддержка многопользовательского режима
9. Неограниченные перекрестные операции
10. Интуитивная манипуляция данными
11. Гибкие возможности получения отчетов
12. Неограниченная размерность и число уровней агрегации

# Дополнительные правила Кодда

1. Пакетное извлечение против интерпретации
2. Поддержка всех моделей OLAP-анализа
3. Обработка ненормализованных данных
4. Сохранение результатов OLAP: хранение их отдельно от исходных данных
5. Исключение отсутствующих значений
6. Обработка отсутствующих значений

# Тест FASMI

F AST (Быстрый)

ANALYSIS (Анализ)

SHARED (Разделяемой)

MULTIDIMENSIONAL (Многомерной)

INFORMATION (Информации)

# OLAP-серверы

MOlap - многомерный (multivariate) OLAP. Для реализации многомерной модели используют многомерные БД;

ROlap - реляционный (relational) OLAP. Для реализации многомерной модели используют реляционные БД;

HOlap - гибридный (hybrid) OLAP. Для реализации многомерной модели используют и многомерные, и реляционные БД.

# MOLAP

Каждый «кубик» преобразуется в отдельную строку таблицы:

Регион	Продукт	Время года	AVG(Продажи)
R1	книги	Весна	9
R1	Еда	Осень	3
R2	книги	Осень	6
R1	книги	ALL	9
R1	ALL	Весна	9
ALL	книги	Весна	9
...	...	...	...
R2	ALL	ALL	6
ALL	Еда	ALL	3
ALL	ALL	Весна	9
ALL	ALL	ALL	6

# MOLAP

## Преимущества:

- поиск и выборка данных осуществляются значительно быстрее,
- легко включить в информационную модель разнообразные встроенные функции.

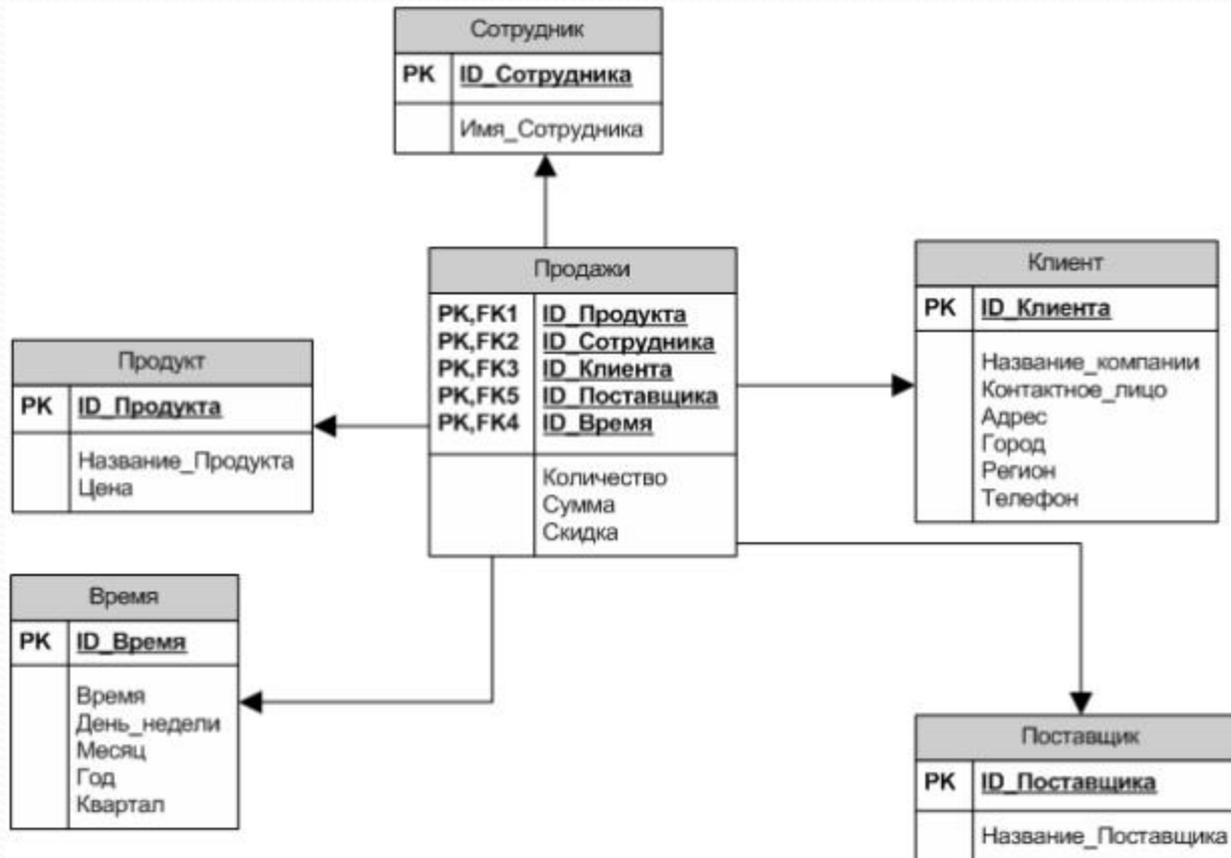
## Недостатки:

- большой объем,
- сложно хранить разреженные данные,
- чувствительны к изменениям структуры многомерной модели.

# MOCLAR – когда ИСПОЛЬЗОВАТЬ?

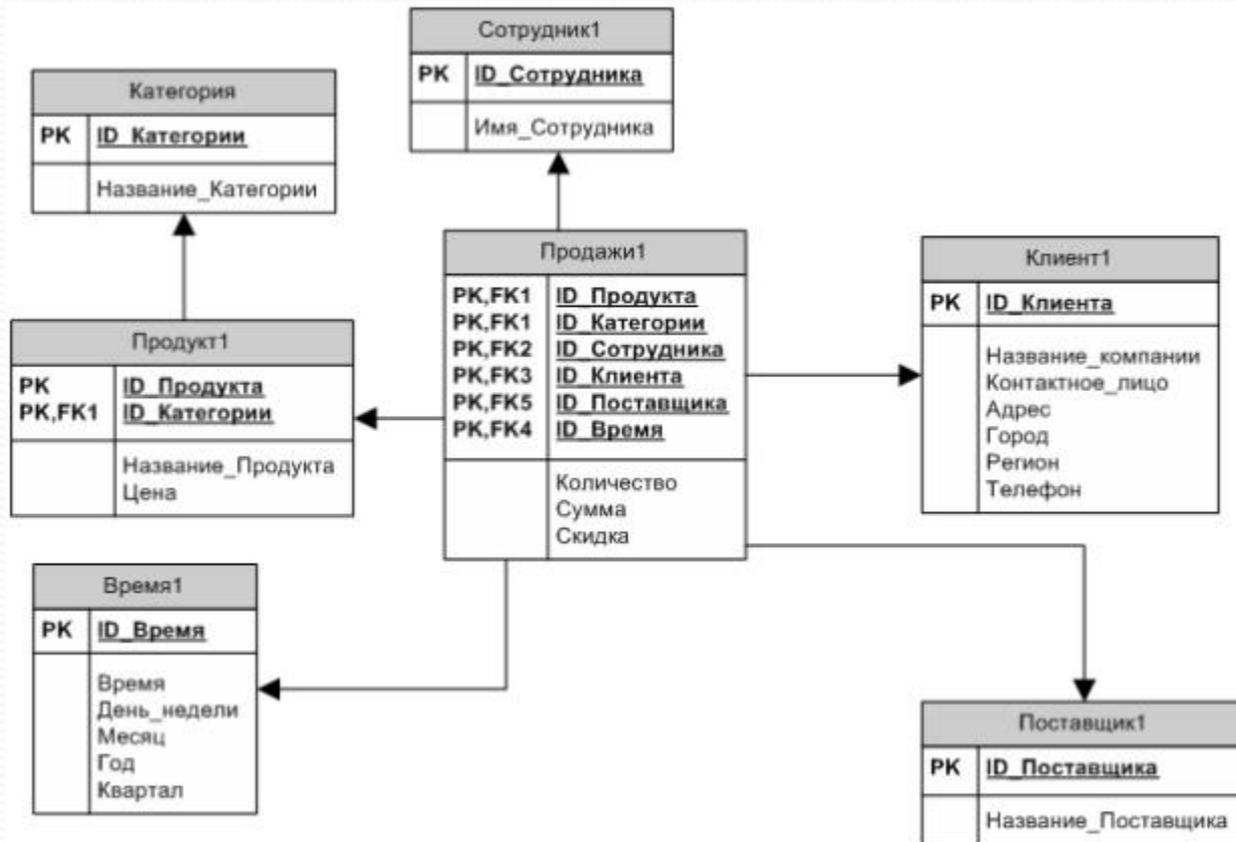
- объем исходных данных для анализа не слишком велик (не более нескольких гигабайт), т. е. уровень агрегации данных достаточно высок;
- набор информационных измерений стабилен;
- время ответа системы на нерегламентированные запросы является наиболее критичным параметром;
- требуется широкое использование сложных встроенных функций.

# ROLAP – схема «звезда»



В центре – таблица фактов, по краям – таблицы измерений

# ROLAP – схема «Снежинка»



# ROLAP

Плюсы:

- в большинстве случаев корпоративные хранилища данных реализуются средствами реляционных СУБД и инструменты ROLAP позволяют производить анализ непосредственно над ними.
- в случае переменной размерности задачи, когда изменения в структуру измерений приходится вносить достаточно часто, ROLAP системы с динамическим представлением размерности являются оптимальным решением, т. к. в них такие модификации не требуют физической реорганизации БД;
- реляционные СУБД обеспечивают значительно более высокий уровень защиты данных и хорошие возможности разграничения прав доступа.

Минусы: низкая скорость работы!