Лекция #12

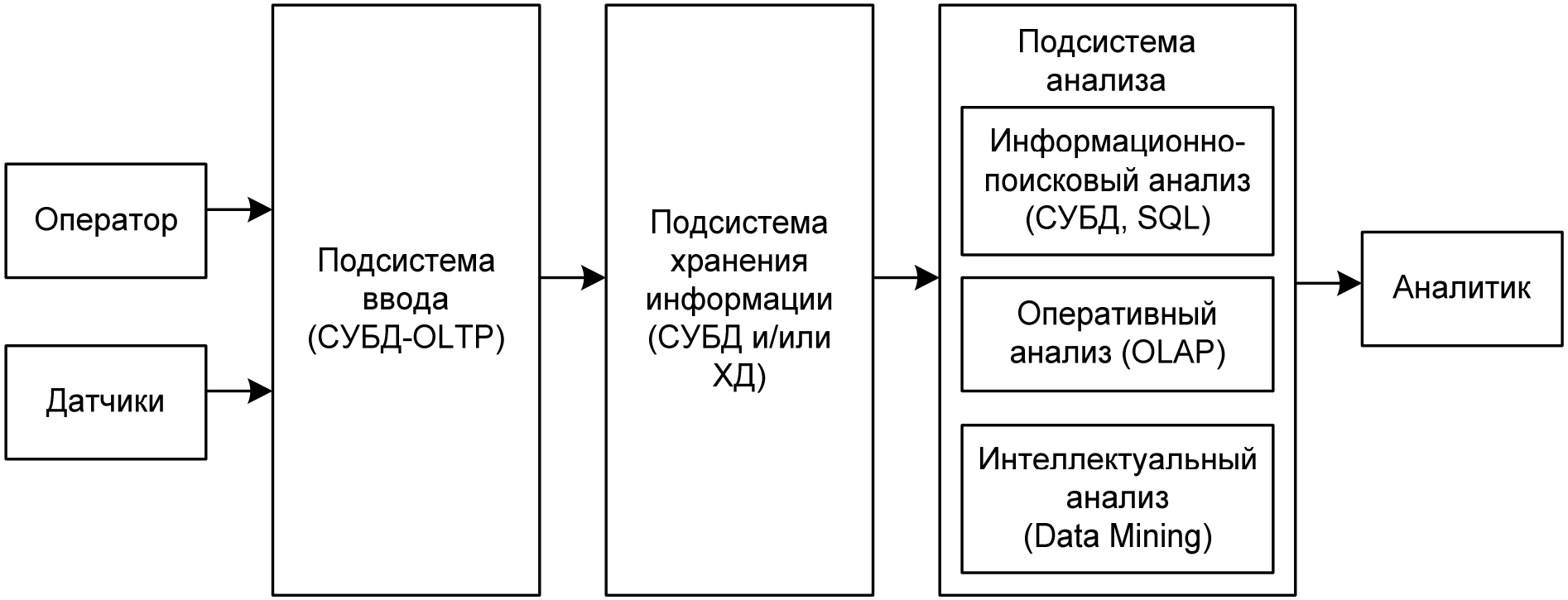
Базы данных

Михаил Моисеев

Технологии анализа данных

СППР

Системы поддержки принятия решений (DSS) – предоставляют собой инструмент для анализа большого объема данных.



Подсистема ввода данных

On-Line Transaction Processing (OLTP) – обеспечивает ручной или автоматический ввод данных в систему.

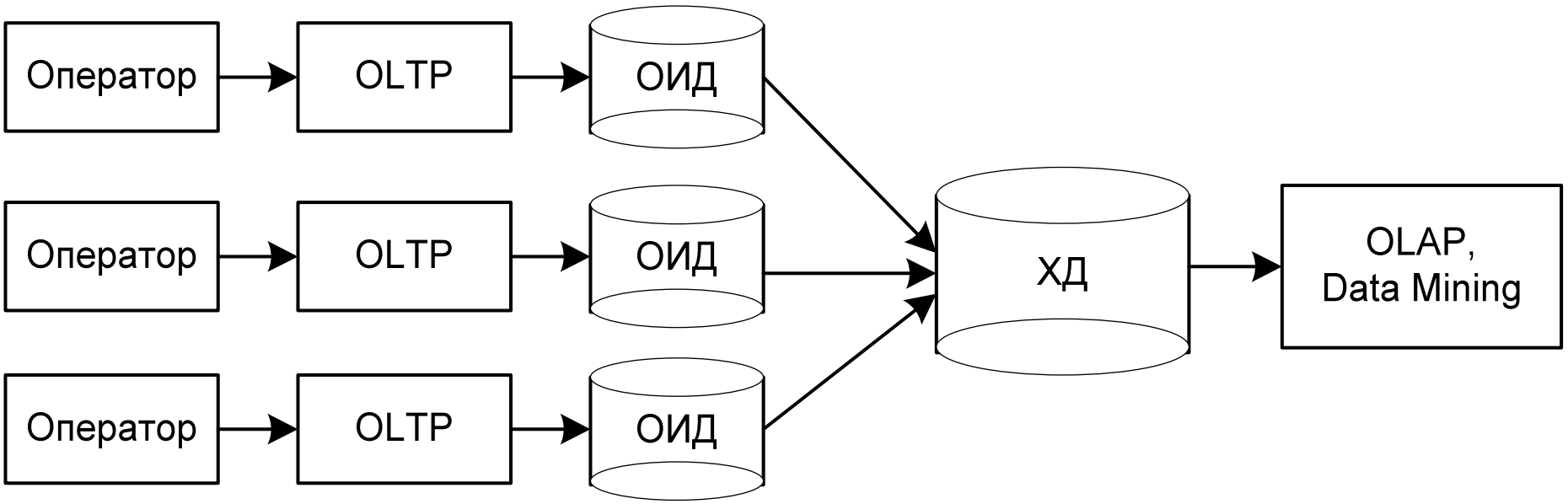
Особенности OLTP:

- Большое число одновременно работающих пользователей;
- Большое число изменений данных;
- Эффективная поддержка механизма транзакций и блокировок;
- Необходимость добавления новых SQL-запросов при изменении требований анализа данных.

OLTP плохо приспособлены для решения задач анализа: возможны ошибки в данных, возможны разные форматы хранения данных, данные нормализированы, хранение оперативных, быстро изменяющихся данных, Фиксированные запросы, средняя загрузка системы (нельзя допустить даже временных перегрузок).

Подсистема хранения информации

Хранилище данных (Data Warehouse) – выделение функций хранения данных для последующего анализа.



Подсистема хранения информации

ОИД – оперативный источник данных (хранение данных OLTP).

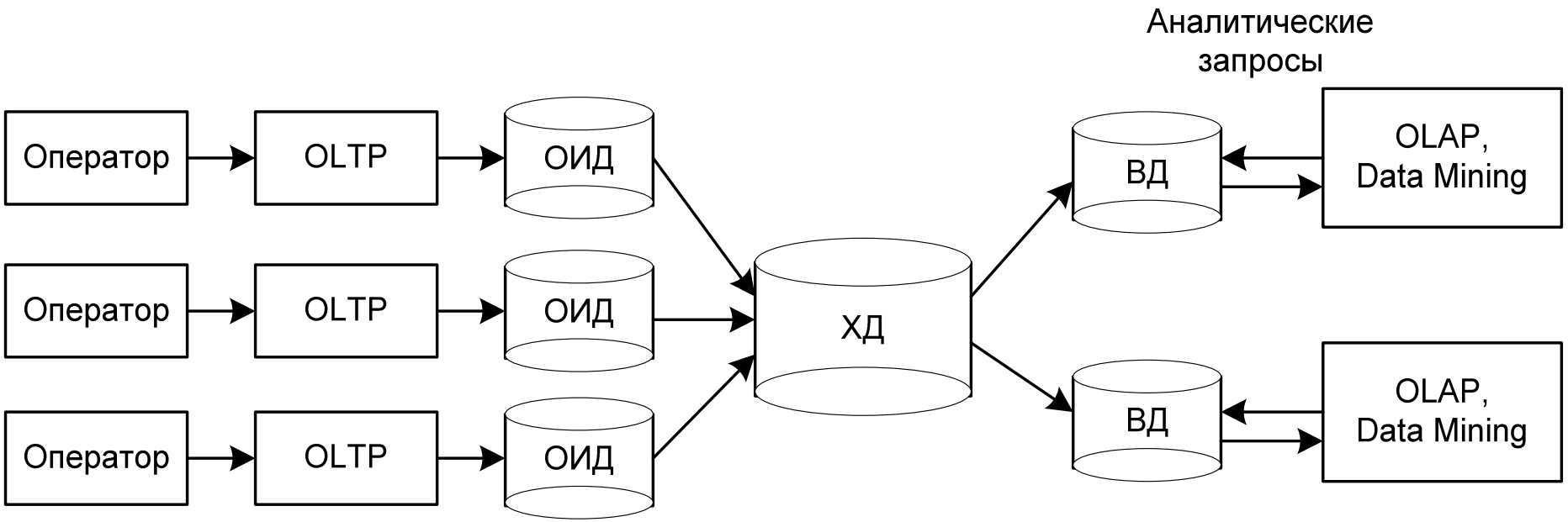
Особенности ХД:

- Предметная ориентация;
- Приведение данных к единому формату;
- Хранение истории изменения данных;
- Постоянство (изменения происходят редко).

Виртуальное ХД – реализуется специальными запросами.

Витрина данных – тематически ориентированное ХД, м.б. несколько ВД для одной системы анализа.

Витрины данных



Перенос данных в ХД

Операция переноса данных из ОИД в ХД (**ETL** -процесс):

- Извлечение данных (**E**xtraction);
  - Выбор ОИД и необходимых данных в них.
- Преобразование данных (**T**ransformation);
  - Обобщение данных (агрегация), уменьшение степени детализации данных;
  - Перевод значений;
  - Создание новых полей (вычисляемых в процессе преобразования);
  - Очистка данных.
- Загрузка данных в ХД (**L**oading).

Очистка данных

Очистка производится на разных уровнях (ячейки, записи, таблицы, …):

- На уровне ячейки – опечатки, пропуски данных, значения вне допустимого диапазона, логически неверные значения, закодированные значения;
- На уровне записи – проверяется непротиворечивость полей друг другу;
- Уровень таблицы – проверка уникальности, проверка на соответствие стандарту записи;
- Уровень БД – проверка целостности данных;
- Уровень нескольких БД – неоднородности структур различных БД.

Процесс очистки данных:

- Выявление проблем в данных;
- Определение правил очистки;
- Проверка правил очистки;
- Непосредственная очистка данных.

OLAP-системы

Online Analytical Processing – технология оперативной аналитической обработки данных, использующая многомерный анализ данных.

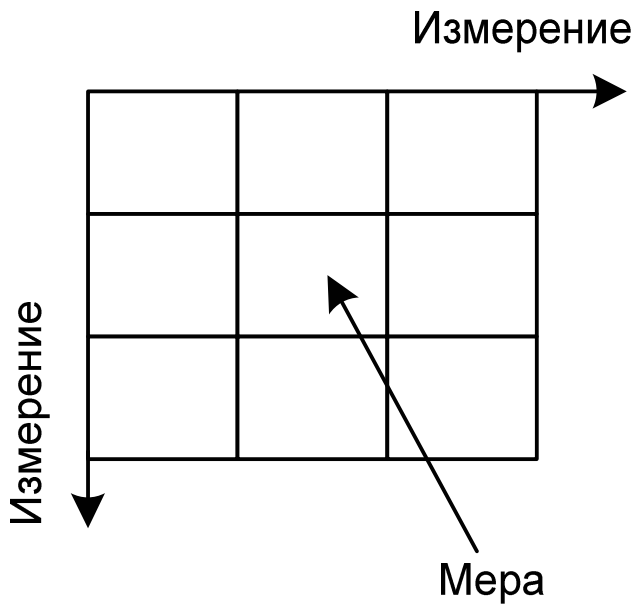
Измерение – последовательность значений анализируемого параметра.
Измерение: **Студенты**, Значения: 4081/1, 4081/2, 4081/3, 4081/4, 4081/5.
Измерение: **Время**, Значения: 02.2008, 03.2008, 04.2008, 05.2008.

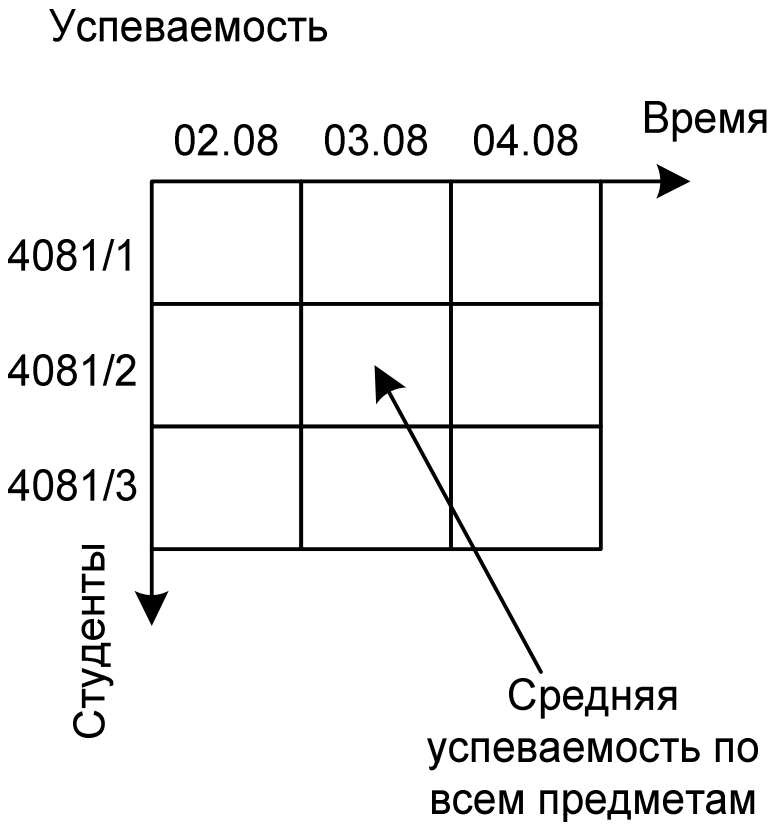
Множественность измерений – многомерный анализ.
Анализ успеваемости по измерениям: Время - Студенты.

Измерение может иметь иерархическую структуру.
Измерение **Студенты**: Институт, курс, группа, студент.
Измерение **Время**: Год, семестр, аттестация, занятие.

OLAP-системы #2

Модель данных – гиперкуб, оси – измерения, на пересечении осей – данные (анализируемые факты, меры).





Операции в OLAP-системах

Срез – формирование подмножества гиперкуба, для фиксированных значений одного или нескольких параметров;

Успеваемость по одному предмету.

Вращение – изменение расположения измерений представляемых пользователю;

Перестановка местами строк и столбцов двумерной таблицы.

Детализация – переход вниз от общего к детальному представлению мер по иерархии измерений;

Переход от рассмотрения успеваемости за все время обучения к успеваемости по семестрам

Консолидация – переход вверх от детального к общему представлению мер по иерархии измерений;

Переход от рассмотрения успеваемости отдельных студентов к успеваемости групп.

Реализация OLAP-систем

OLAP-сервер обеспечивает хранение данных и выполнение над ними необходимых операций.

OLAP-клиент предоставляет интерфейс к многомерной модели данных.

Способы реализации:

- MOLAP – используются многомерные БД;
  - Хранение данных в плоских «файлах»;
  - Быстрый поиск и выборка данных;
  - Объем в 2.5-100 раз больше объема исходных данных;
  - Сильная разреженность гиперкуба;
  - Большие затраты времени при добавлении нового измерения.
- ROLAP - используются реляционные БД;
  - Таблицы измерений, схемы «звезда» и «снежинка»;
  - Таблица фактов, имеющая составной первичный ключ.
- HOLAP – используются и многомерные и реляционные БД.

Файл MOLAP

Time	Subject	Teacher	Student	Department	Faculty	Mark
01.05.2008	БД	Иванов	Сидоров	АиВТ	ФТК	4
01.05.2008	БД	Иванов	Сидоров	АиВТ	ФТК	-
02.05.2008	БД	Иванов	Сидоров	АиВТ	ФТК	-
02.05.2008	ТКС	Петров	Сидоров	АиВТ	ФТК	5
03.05.2008	ТКС	Петров	Алексеев	САУ	ФТК	3
03.05.2008	ТКС	Петров	Алексеев	САУ	ФТК	-
03.05.2008	ТКС	Петров	Алексеев	САУ	ФТК	-
03.05.2008	ТКС	Петров	Алексеев	САУ	ФТК	-

Схема «звезда» ROLAP

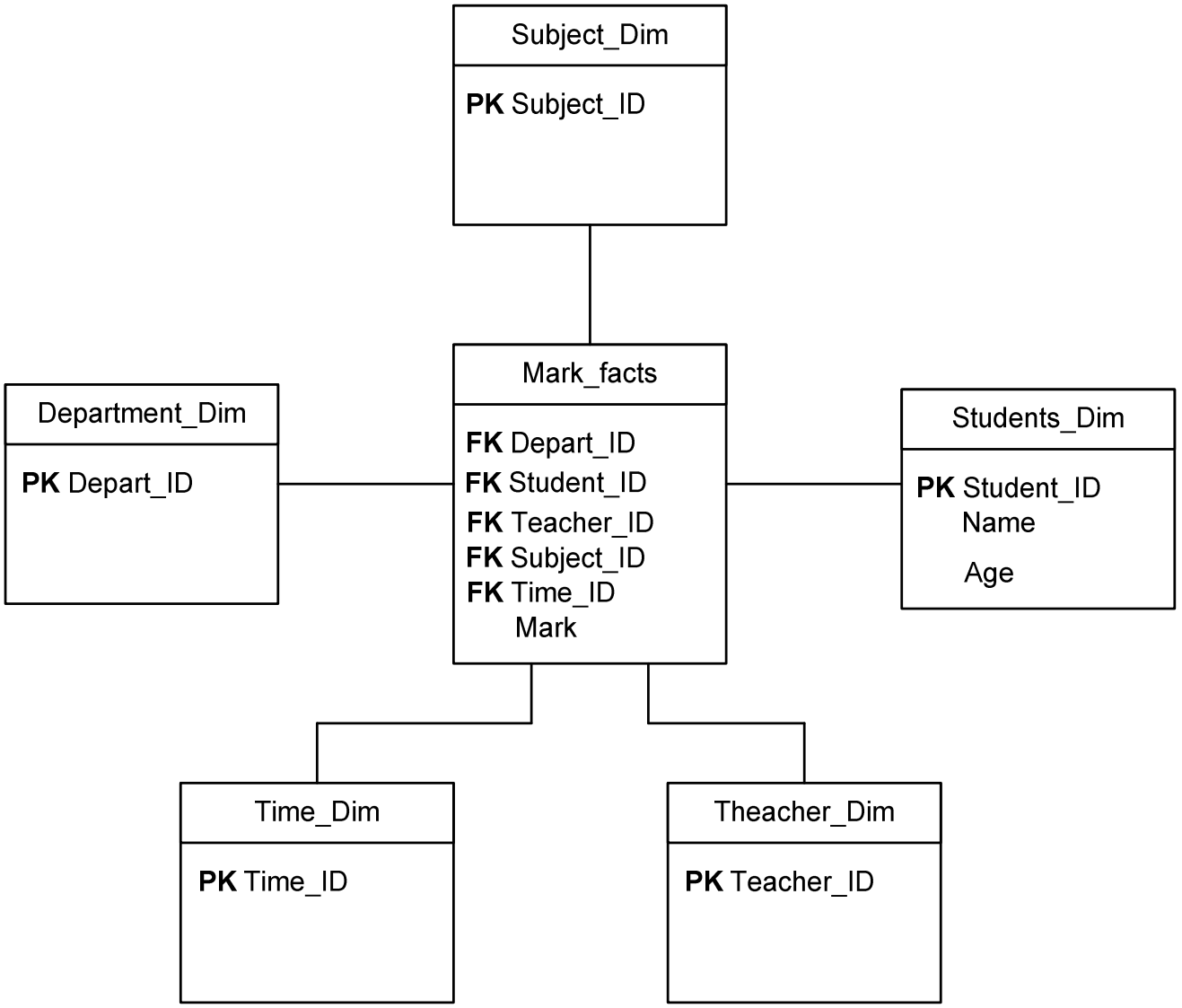
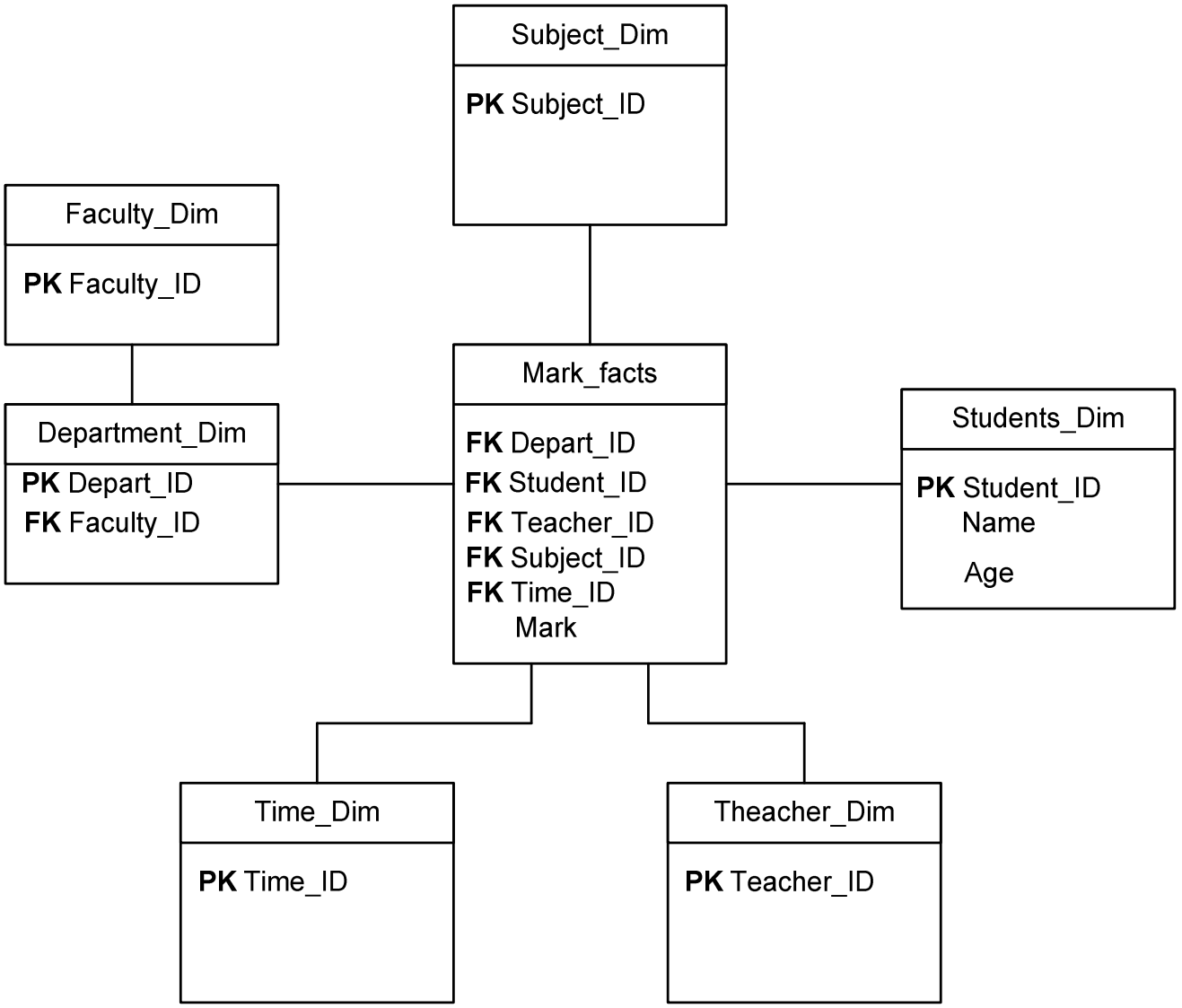


Схема «снежинка» ROLAP



Интеллектуальный анализ данных

Data Mining – «добыча» данных, задача – обнаружение **скрытых** знаний в больших объемах исходных данных.

Свойства обнаруживаемых знаний:

- Знания должны быть новыми, ранее неизвестными;
- Знания должны быть нетривиальны (неочевидные закономерности, которые не могут быть получены другими способами, например OLAP);
- Знания должны быть практически полезны (применимость и достоверность полученных знаний);
- Знания должны быть доступны для понимания человеку (объяснимость, понятный вид).

Задачи Data Mining

Задача классификации - определение класса объекта по его измеренным характеристикам;

Задача регрессии - определение некоторого параметра объекта по его измеренным другим параметрам (определяемый параметр имеет неограниченное множество значений);

Поиск ассоциативных правил – нахождение зависимостей между объектами или событиями (найденные зависимости представляются в виде правил и могут быть использованы для лучшего понимания природы анализируемых данных и для предсказания появления событий).

Задача кластеризации – поиск независимых групп (кластеров) и их характеристик на всем множестве анализируемых данных (помогает лучше понять анализируемые данные).

Основные методы Data Mining

Переборные методы – простота понимания и реализации, отсутствие формальной теории, вычислительные затраты;

Нечеткая логика – отображение на формализованный язык и анализ полученной модели, набор множества правил, которые могут противоречить друг другу.

Генетические алгоритмы – порождение набора моделей, оценка их эффективности, отбраковка неэффективных моделей, порождение нового набора («мутация»).

Нейронные сети – древовидная сеть с фиксированной структурой, процесс обучения на тестовых данных, последующее использование на реальных данных.

Процесс обнаружения знаний

- Понимание и формулировка задачи анализа;
- Подготовка данных для автоматизированного анализа;
- Выбор и применение методов Data Mining, построение моделей;
- Проверка построенных моделей;
- Интерпретация моделей человеком.

Средства Data Mining:

- Oracle;
- MS SQL Server;
- IBM DB2.

Области применения Data Mining

- Интернет-технологии;
- Маркетинг;
- Телекоммуникации;
- Промышленное производство;
- Медицина;
- Банковское дело.

Вопросы

- Для чего предназначены СППР ?
- Перечислите основные подсистемы СППР.
- Каковы особенности OLTP-систем ?
- Для чего предназначены хранилища данных ?
- Как осуществляется перенос данных в ХД ?
- Что такое витрина данных ?
- Как выполняется очистка данных ?
- Что такое OLAP ?
- Какую модель данных использует OLAP ?
- Перечислите операции в OLAP-системах.
- Что такое Data Mining ?
- Перечислите задачи Data Mining.
- Перечислите методы Data Mining.