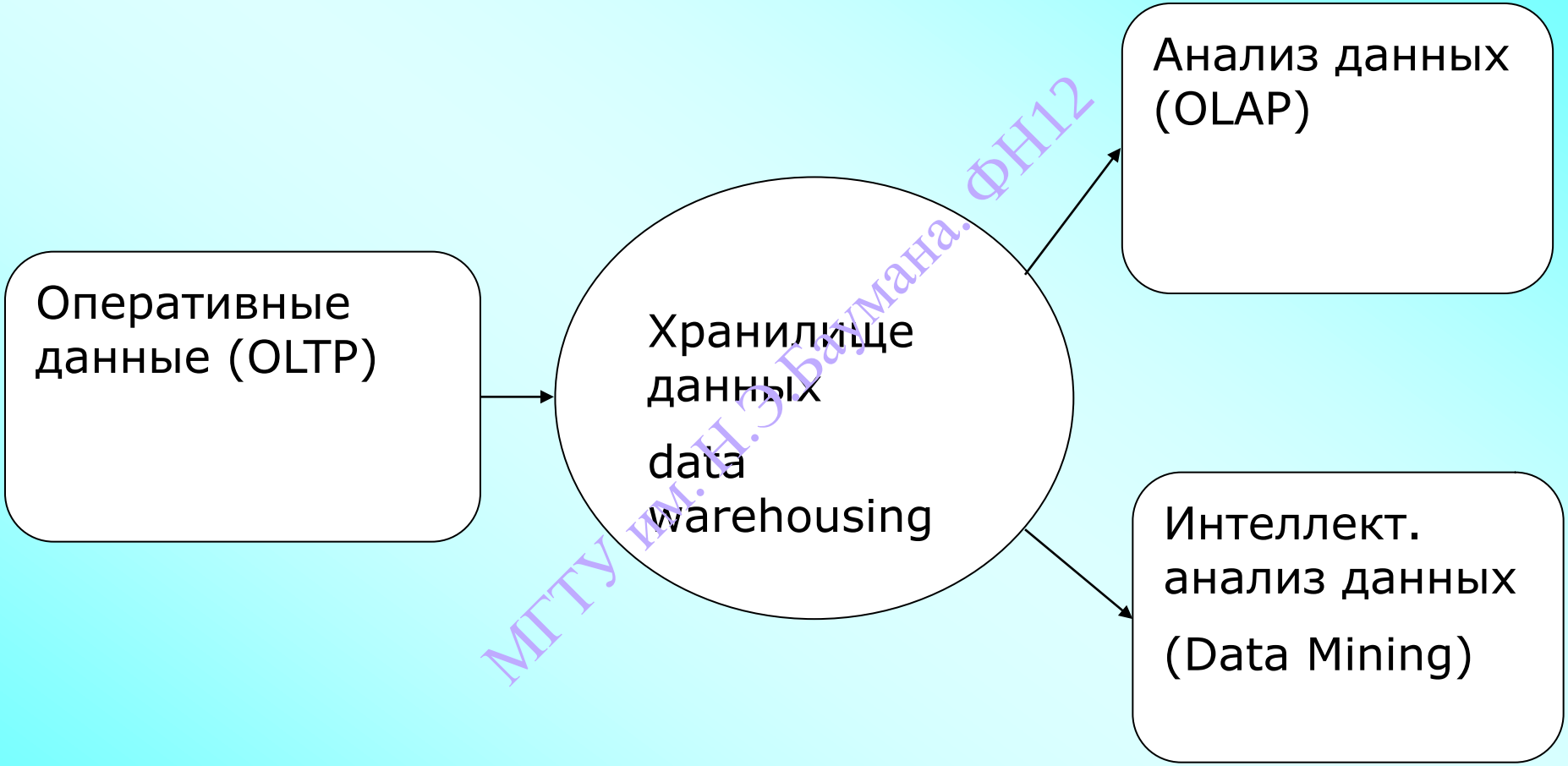
Лекция 17

Концепции хранилищ данных, OLAP и Data Mining

Системы поддержки принятия решений (СППР, Decision Support Systems, DSS).



СППР возникли начале 70-х годов прошлого века. Представляют собой компьютерные системы, которые путем сбора и анализа большого количества информации могут влиять на процесс принятия решений.

Основные задачи СППР

- выбор наилучшего решения из множества возможных (оптимизация),
- упорядочение возможных решений по предпочтительности (ранжирование).

Структура СППР

- информационные хранилища данных;
- средства и методы извлечения, обработки и загрузки данных;
- многомерная база данных и средства анализа OLAP;
- средства Data Mining.

Информационные хранилища данных

Концепция хранения и обработки данных – концепция хранилищ (data warehousing) появилась в IBM .
Окончательно теория была сформулирована У. Г. Инмоном (William H. Inmon) и Р. Кимбаллом (Ralph Kimball) в 90-х годах прошлого столетия.
Название "**метод решения информационно-аналитических задач в области принятия и поддержки решений**".

Хранилище данных – предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации.
Строится на базе систем управления базами данных.
Данные, поступающие в хранилище данных, как правило, доступны только для чтения.

Основные принципы организации хранилища данных
(Inmon W. Building the Data Warehouse. – New York, 1992)

1. **Предметная ориентированность.** Информация в ХД организованы в соответствии с основными аспектами деятельности предприятия (в данные РБД организованы в соответствии с операциями, с приложениями).
2. **Интегрированность.** Исходные данные для хранения поступают из внешних источников (РБД, электронных таблиц,…). Данные приводят к единому формату, агрегируются. Данные накапливаются в хранилище в виде исторических слоев.
3. **Неизменяемость (**Некорректируемость**)**.. Данные в ХД не создаются, не корректируются и не удаляются.
4. **Привязка ко времени (**Зависимость от времени**)**. Данные в ХД всегда должны быть связаны с определенным временным периодом.

Требования к хранилищам данных (Р. Кимбалл)

1. поддержка высокой скорости получения данных из хранилища;
2. поддержка внутренней непротиворечивости данных;
3. возможность получения и сравнения срезов данных (slice and dice);
4. наличие удобных утилит просмотра данных в хранилище;
5. полнота и достоверность хранимых данных;
6. поддержка качественного процесса пополнения данных.

Преимущества использования хранилищ данных

1. ХД содержит информацию за весь требуемый временной интервал в едином информационном пространстве.
2. ХД содержит информацию в едином формате (несоответствия в данных устраняются на этапе сбора информации, организуются единые справочники, все показатели в которых приводятся к одинаковым единицам измерения).
3. Технология ХД обеспечивает построение аналитических отчетов на основе надежных данных и оповещение администратора хранилища об ошибках во входящей информации.
4. Универсальный доступ к данным, ХД предоставляет возможность получать любые отчеты о деятельности предприятия на основе одного источника информации.
5. Ускорение получения аналитических отчетов (работа сервера ХД не мешает работе операторов, в ХД содержится детальная и заранее агрегированные информации, доступна архивная информация.
6. Построение произвольных запросов (технология OLAP).

OLAP

Термин *OLAP* обозначает методы, которые дают возможность пользователям МБД в реальном времени генерировать описательные и сравнительные сводки данных и получать ответы на различные аналитические запросы.

OLAP не подразумевает интерактивную обработку данных (в режиме реального времени), означает процесс анализа многомерных баз данных путем составления эффективных "многомерных" запросов к данным различных типов.

Средства *OLAP* могут быть встроены в корпоративные (масштаба предприятия) системы баз данных, следить за ходом и результативностью своего бизнеса

OLAP-кубы: оси содержат параметры, ячейки — зависящие от них агрегатные данные.

В ячейках OLAP-куба могут содержаться результаты выполнения агрегатных функций SQL (MIN, MAX, AVG, COUNT), дисперсии, СКО и т.д. Значения данных в ячейках – термин summary, исходные данные – термин measure, параметры запросов — термин dimension, значения осях – члены измерений (members).

Требования FASMI

Fast Analysis of Shared Multidimensional Information

- предоставление пользователю результатов анализа за приемлемое время (обычно не более 5 с), пусть даже ценой менее детального анализа;
- возможность осуществления любого логического и статистического анализа, характерного для данного приложения, и его сохранения в доступном для конечного пользователя виде;
- многопользовательский доступ к данным с поддержкой соответствующих механизмов блокировок и средств авторизованного доступа;
- многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий (ключевое требование OLAP);
- возможность обращаться к любой нужной информации независимо от ее объема и места хранения.

Сформулированы в 1995 году на основе 12 правил Кодда

Архитектуры OLAP -серверов

1. MOLAP (Multidimensional OLAP);
2. ROLAP (Relational OLAP);
3. HOLAP (Hybrid OLAP).

MOLAP. Исходные, многомерные данные хранятся в МБД или в многомерном локальном кубе.
Достоинство. **О**беспечивает высокую скорость выполнения OLAP -операций.
Недостаток. МБД избыточна. Куб сильно зависит от числа измерений. При увеличении количества измерений объем куба растет экспоненциально.
ROLAP. Исходные данные хранятся в РБД, в плоских локальных таблицах на файл-сервере. Агрегатные данные – в служебных таблицах РБД.
Достоинство. Нет избыточности (или она минимальная).
Недостаток. Большое время отклика системы (преобразование данных из РБД в многомерные кубы – по запросу OLAP -средства).
HOLAP. Исходные данные остаются в РБД, агрегаты – в МБД. Построение OLAP-куба – по запросу OLAP-средства.

Сравнительные характеристики различных моделей управления данными

Характеристики	Реляционные СУБД OLTP	Реляционные СУБД СППР / ХД	Многомерные СУБД OLAP
Операция	Обновление	Отчет	Анализ
Уровень аналитических требований	Низкий	Средний	Высокий
Экраны	Неизменяемые	Определяемые пользователем	Определяемые пользователем
Объем данных на транзакцию	Небольшой	От малого до большого	Большой
Уровень данных	Детальные	Детальные и суммарные	В основном суммарные
Сроки хранения данных	Только текущие	Исторические и текущие	Исторические, текущие и прогнозируемые
Структурные элементы	Записи		

Data Mining

Термин Data Mining – из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining).

Data Mining – мультидисциплинарная область, развивающаяся на базе теории баз данных, прикладной статистики, искусственного интеллекта, машинного обучения, распознавания образов, алгоритмизации и др.

Статистика – совокупность методов планирования эксперимента, сбора данных, их представления и обобщения, анализа и получения выводов на основании этих данных.

Искусственный интеллект - научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования видов человеческой деятельности, традиционно считающихся интеллектуальными.

Машинное обучение – процесс получения программой новых знаний. «Машинное обучение - это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы» (Митчелл, 1996).

- Статистика
  - Более, чем Data Mining, базируется на теории.
  - Более сосредотачивается на проверке гипотез.
- Машинное обучение
  - Более эвристично.
  - Концентрируется на улучшении работы агентов обучения.
- Data Mining.
  - Интеграция теории и эвристик.

Григорий Пиатецкий-Шапиро (Gregory Piatetsky-Shapiro) - один из основателей направления:
Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.
Технология Data Mining предназначена для поиска неочевидных, объективных и полезных на практике закономерностей в больших объемах данных.

Этапы процесса Data Mining.

- анализ предметной области ;
- постановка задачи;
- подготовка данных;
- построение моделей;
- проверка и оценка моделей;
- выбор модели;
- применение модели;
- коррекция и обновление модели.

Анализ предметной области

Исследование - это процесс познания определенной предметной области, объекта или явления с определенной целью.

Постановка задачи

- формулировка задачи;
- формализация задачи;
- описание статического и динамического поведения исследуемых объектов.

Подготовка данных

•Определение и анализ требований к данным, которые необходимы для осуществления Data Mining. Изучаются вопросы распределения пользователей (географическое, организационное, функциональное); вопросы доступа к данным, которые необходимы для анализа, необходимость во внешних и/или внутренних источниках данных; аналитические характеристики системы (измерения данных, основные виды выходных документов, последовательность преобразования информации и др.).

•Сбор данных. Источником для исходных данных являются хранилища данных, оперативные, справочные и архивные БД, т.е. данные из существующих информационных систем.

•Предварительная обработка данных. Оценивание качества данных. Данные, полученные в результате сбора, должны соответствовать определенным критериям качества.

Построение модели. Для построения моделей используются различные методы и алгоритмы Data Mining, а также используются модели, построенные на основе различных методов. Рабочая группа Data Mining Group предложила стандарт PMML (Predictive Model Markup Language), который позволяет осуществлять обмен моделями, созданными в приложениях различных поставщиков программного обеспечения Data Mining.

Проверка и оценка моделей

Проверка модели подразумевает проверку ее достоверности или адекватности. Эта проверка заключается в определении степени соответствия модели реальности. Адекватность модели проверяется путем тестирования.

Адекватность модели (adequacy of a model) - соответствие модели моделируемому объекту или процессу.

Выбор модели

Если в результате моделирования нами было построено несколько различных моделей, то на основании их оценки мы можем осуществить выбор лучшей из них. В ходе проверки и оценки различных моделей на основании их характеристик, а также с учетом мнения экспертов, следует выбор наилучшей. Достаточно часто это оказывается непростой задачей. Основные характеристики модели, которые определяют ее выбор, - это точность модели и эффективность работы алгоритма

Применение модели

Выбранная модель используется применительно к новым данным с целью решения задач, поставленных в начале процесса Data Mining. Для классификационных и прогнозирующих моделей на этом этапе прогнозируется целевой (выходной) атрибут (target attribute).

Коррекция и обновление модели

Анализ полученных результатов и обновление модели.

Некоторые бизнес-приложения Data Mining

Розничная торговля

- *анализ покупательской корзины*
- *исследование временных шаблонов*
- *создание прогнозирующих моделей*

Банковское дело

- *выявление мошенничества с кредитными карточками*.
- *сегментация клиентов*.
- *прогнозирование изменений клиентуры*

Телекоммуникации

- *анализ записей о подробных характеристиках вызовов.*
- *выявление лояльности клиентов*.

Страхование

- *выявление мошенничества*.
- *анализ риска*.

Медицина
Экспертные системы для постановки медицинских диагнозов. Построены на основе правил, описывающих сочетания различных симптомов различных заболеваний. Технологии Data Mining позволяют обнаруживать в медицинских данных шаблоны, составляющие основу указанных правил.
Молекулярная генетика и генная инженерия
Задача обнаружения закономерностей в экспериментальных данных, определение маркеров (генетических кодов, контролирующих фенотипические признаки живого организма).
Прикладная химия
Выяснение особенностей химического строения тех или иных соединений, определяющих их свойства. Особенно актуальна такая задача при анализе сложных химических соединений, описание которых включает сотни и тысячи структурных элементов и их связей.