Лекция №13

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ (ИАД)

ТЕХНОЛОГИИ ХРАНИЛИЩ ДАННЫХ

Хранилище данных (Б. Инмоном) – предметно-ориентированное, привязанное ко времени и неизменяемое собрание данных для поддержки принятия управляющих решений.

Хранилище данных представляет собой репозиторий, содержащий непротиворечивые консолидированные исторические данные корпорации, отражающие ее деятельность за достаточно продолжительный период времени, а также данные о внешней среде ее функционирования.

Объем данных в хранилище как минимум на порядок **превосходит объемы данных в оперативных БД** (так называемых **OLTP-системах**: **On-Line Transaction Processing** – оперативная обработка транзакций).

ТЕХНОЛОГИИ ХРАНИЛИЩ ДАННЫХ

Большей сложностью отличаются **запросы к хранилищу**. Необходима **высокая производительность обработки запросов** и **масштабируемость алгоритмов**. **При загрузке** в хранилище **новых данных** должна выполняться их **верификация**.

Хранилище данных может включать **2** или **3 уровня**.

В первом случае на верхнем уровне располагается обобщенная информация для руководителей всех подразделений предприятия, которым требуются средства анализа данных. Нижний уровень занимают источники данных, в том числе БД оперативной информации.

В трехуровневой архитектуре над двухуровневым хранилищем организуются специализированные хранилища данных для отдельных подразделений.

ТЕХНОЛОГИИ ИАД

Анализ данных в хранилищах базируется на **технологиях ИАД**.

Целью ИАД является извлечение знаний из данных, т.е. обнаружение в исходных данных ранее неизвестных нетривиальных практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных предметных областях.

Наиболее распространенный тип знаний, извлекаемых с помощью технологий ИАД, – это закономерности предметной области.

В зависимости от характера закономерностей предметной области можно разделить на три группы:

1) предметные области с доминированием случайных событий;

2) предметные области, в которых все события причинно обусловлены;

3) предметные области, в которых наблюдаются как причинно обусловленные, так и случайные события.

ТЕХНОЛОГИИ ИАД

Данные в ИАД представляются тремя способами: атрибутивным; структурным; полнотекстовым.

Методы ИАД подразделяют на три класса:

- Алгебраические методы.
- Статистические методы.
- Методы мягких вычислений.

Методы ИАД реализуются в трех технологиях:

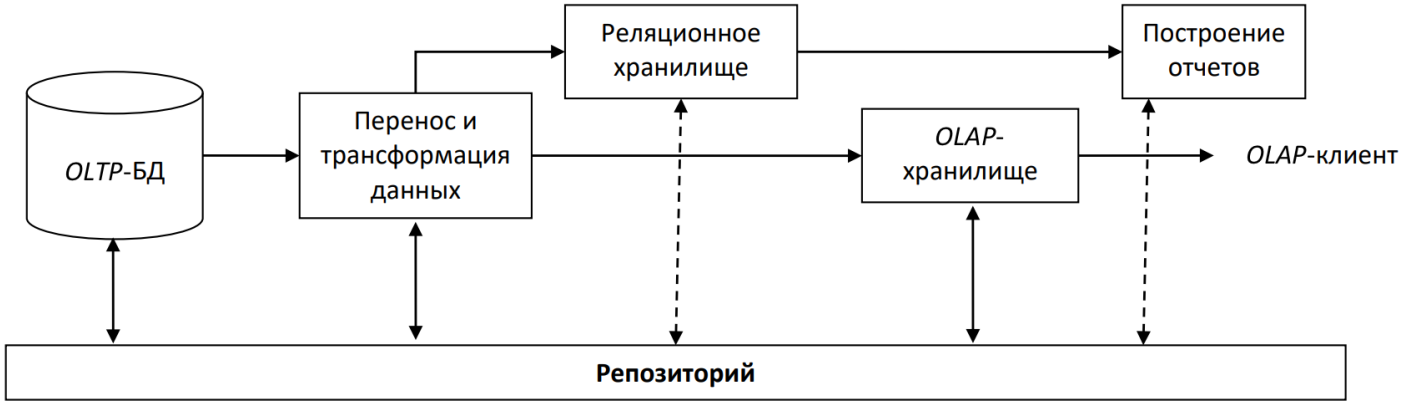
- интерактивной аналитической обработки данных (On-Line Analytical Processing — OLAP);
- глубинного анализа данных (Data Mining — DM);
- визуализации данных.

ТЕХНОЛОГИЯ OLAP И МНОГОМЕРНЫЕ МОДЕЛИ ДАННЫХ

Технология OLAP ориентирована, главным образом, на обработку нерегламентированных запросов к хранилищам данных.

Основной задачей хранилища является представление данных для анализа в одном месте в рамках простой и понятной структуры.

Структура типичного хранилища данных (сплошные стрелки обозначают потоки данных, пунктирные – метаданных).



ТЕХНОЛОГИЯ OLAP И МНОГОМЕРНЫЕ МОДЕЛИ ДАННЫХ

Основная цель анализа данных — качественная и количественная оценка достигнутых результатов и (или) динамики деятельности компании.

Принципы *OLAP* были сформулированы **Э. Коддом**. Центральное место среди них занимает поддержка **многомерного представления данных**.

В многомерной модели данных БД представляется в виде одного или нескольких **кубов данных (гиперкубов)**.

Осями гиперкуба служат основные атрибуты анализируемого бизнес-процесса.

На пересечении осей-измерений **(*dimensions*)**, т.е. в ячейке гиперкуба, содержатся данные, количественно характеризующие анализируемый процесс. Эти данные называются мерами **(*measures*)** или показателями.

В процессе анализа выполняются операции построения сечений (проекций) гиперкуба путем фиксации значений наборов атрибутов-координат.

Многомерность в OLAP-приложениях

Многомерность реализуется в рамках 2-х или 3-х уровневой архитектуры.

Первый уровень поддерживает многомерное представление данных, абстрагированное от их физической структуры. Он содержит средства многомерной визуализации и манипулирования данными.

Второй уровень обеспечивает многомерную обработку. Он включает язык формулирования многомерных запросов (*SQL* для этих целей непригоден) и программный процессор, способный выполнять такие запросы. Он обычно встраивается в ***OLAP*-клиент** или в ***OLAP*-сервер**.

Третий уровень реализует физическую организацию хранения многомерных данных. В рамках него для поддержки многомерных моделей данных используются либо специальные *OLAP*-СУБД, либо обычные реляционные структуры. Обычно *OLAP*-продукты обеспечивают оба эти способа хранения, а также их комбинации:

***MOLAP* (*Multidimensional OLAP*)** — и детальные данные, и агрегаты данных хранятся в многомерной БД;

***ROLAP* (*Relational OLAP*)** — детальные данные хранятся в реляционной БД, агрегаты в специально созданных служебных таблицах;

***HOLAP* (*Hybrid OLAP*)** — детальные данные хранятся в реляционной БД, агрегаты в многомерной БД.

Управление метаданными

В технологии хранилищ данных важную роль играет **управление метаданными.**

Метаданные хранилищ делятся на три группы:

- **Административные** описывают *OLTP*-БД, служащие источниками для *OLAP*, схемы данных хранилища, измерения гиперкубов, физическую организацию данных, формы стандартных отчетов, полномочия пользователей, типовые запросы;
- **Операционные** отражают информацию о текущем состоянии данных, статистике функционирования;
- **Бизнес-метаданные** содержат словарь терминов с их определениями, описания источников и владельцев данных и т.п.