

# **Использование хранилищ данных и технологии OLAP**

**Хранилища данных (Data Ware House) и технологии  
OLAP (On Line Analytical Processing)**

# Общие сведения

- Технология комплексного многомерного анализа данных получила название OLAP (On-Line Analytical Processing). OLAP — это ключевой компонент организации хранилищ данных. Концепция OLAP была описана в 1993 году Эдгаром Коддом, известным исследователем баз данных и автором реляционной модели данных

# Общие сведения

- OLAP (On-Line Analytical Processing) является ключевым компонентом построения и применения хранилищ данных. Эта технология основана на построении многомерных наборов данных — OLAP-кубов, оси которого содержат параметры, а ячейки — зависящие от них агрегатные данные.

# Требования к приложениям для многомерного анализа:

- 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

# Требования к приложениям для многомерного анализа: Продолжение

- многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий (это — ключевое требование OLAP);
- возможность обращаться к любой нужной информации независимо от ее объема и места хранения.

# Хранилище данных (data warehouse).

- Хранилище данных – это интегрированный накопитель информации, собранной из других систем, на основе которого строятся процессы принятия решений и анализа данных. Все хранилища данных имеют некоторые общие признаки:

# Признаки хранилища данных

- Информация в хранилище данных организовывается вокруг базовых понятий, используемых в деятельности предприятия;
- "Сырые" данные собираются из неинтегрированных оперативных и унаследованных приложений, очищаются от ошибок, затем агрегируются и представляются в виде, понятном бизнес-пользователям;
- Процесс создания хранилища является итеративным, требует регулярного перепроектирования в течение всего жизненного цикла приложения.

# Типичная структура хранилищ данных

- Основная идея OLAP заключается в построении многомерных кубов, которые будут доступны для пользовательских запросов
- Однако исходные данные для построения OLAP-кубов обычно хранятся в реляционных базах данных
- Нередко это специализированные реляционные базы данных, называемые также хранилищами данных (Data Warehouse)



# Типичная структура хранилищ данных

## Продолжение

- Типичная структура хранилища данных существенно отличается от структуры обычной реляционной СУБД. Как правило, эта структура денормализована (это позволяет повысить скорость выполнения запросов), поэтому может допускать избыточность данных.

# Типичная структура хранилищ данных

Продолжение

■ Основными составляющими структуры хранилищ данных являются таблица фактов (fact table) и таблицы измерений (dimension tables).

# Таблица фактов

- Таблица фактов является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно говорят о четырех наиболее часто встречающихся типах фактов. К ним относятся:

# Таблица фактов

- факты, связанные с транзакциями (Transaction facts). Они основаны на отдельных событиях (типичными примерами которых являются телефонный звонок или снятие денег со счета с помощью банкомата);
- факты, связанные с «моментальными снимками» (Snapshot facts). Основаны на состоянии объекта (например, банковского счета) в определенные моменты времени, например на конец дня или месяца. Типичными примерами таких фактов являются объем продаж за день или дневная выручка;

# Таблица фактов

- факты, связанные с элементами документа (Line-item facts). Основаны на том или ином документе (например, счете за товар или услуги) и содержат подробную информацию об элементах этого документа (например, количестве, цене, проценте скидки);
- факты, связанные с событиями или состоянием объекта (Event or state facts). Представляют возникновение события без подробностей о нем (например, просто факт продажи или факт отсутствия таковой без иных подробностей).

# Таблица фактов

- Содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Чаще всего это целочисленные значения либо значения типа «дата/время» — ведь таблица фактов может содержать сотни тысяч или даже миллионы записей, и хранить в ней повторяющиеся текстовые описания, как правило, невыгодно — лучше поместить их в меньшие по объему таблицы измерений. При этом как ключевые, так и некоторые неключевые поля должны соответствовать будущим измерениям OLAP-куба. Помимо этого таблица фактов содержит одно или несколько числовых полей, на основании которых в дальнейшем будут получены агрегатные данные.

# Таблицы измерений

- Таблицы измерений содержат неизменяемые либо редко изменяемые данные
- В подавляющем большинстве случаев эти данные представляют собой по одной записи для каждого члена нижнего уровня иерархии в измерении.
- Таблицы измерений также содержат как минимум одно описательное поле (обычно с именем члена измерения) и, как правило, целочисленное ключевое поле (обычно это суррогатный ключ) для однозначной идентификации члена измерения

# Таблицы измерений

- Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов
- скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов; например, добавление новой записи в таблицу измерений, характеризующую товары, производится только при появлении нового товара, не продававшегося ранее.



# Хранилище данных “Звезда”

- Одно измерение куба может содержаться как в одной таблице (в том числе и при наличии нескольких уровней иерархии), так и в нескольких связанных таблицах, соответствующих различным уровням иерархии в измерении. Если каждое измерение содержится в одной таблице, такая схема хранилища данных носит название «звезда» (star schema).

# Хранилище данных “Снежинка”

- Если же хотя бы одно измерение содержится в нескольких связанных таблицах, такая схема хранилища данных носит название «снежинка» (snowflake schema). Дополнительные таблицы измерений в такой схеме, обычно соответствующие верхним уровням иерархии измерения и находящиеся в соотношении «один ко многим» в главной таблице измерений, соответствующей нижнему уровню иерархии, иногда называют консольными таблицами (outrigger table).

# Есть два основных вида хранилищ данных:

общекорпоративные хранилища  
данных (enterprise data  
warehouse)

- киоски (или витрины) данных  
(data mart).

# Корпоративное хранилище данных

- Корпоративное хранилище данных содержит информацию о всех сторонах деятельности организации. Обычно оно формируется на основании данных, касающихся нескольких различных аспектов - например, клиентов, продуктов и продаж - и служит для поддержки принятия как тактических, так и стратегических решений.

# Киоски данных

- Киоски данных содержат некоторое подмножество всех данных корпорации, которое создается для использования его отдельными подразделениями или отделами организации. В отличие от корпоративных хранилищ, киоски данных часто строятся снизу вверх на основе информационных ресурсов подразделения, используемых конкретным приложением поддержки принятия решений или группой пользователей.

# Применяются три способа хранения данных:

- MOLAP (Multidimensional OLAP) — исходные и агрегатные данные хранятся в многомерной базе данных. Хранение данных в многомерных структурах позволяет манипулировать данными как многомерным массивом, благодаря чему скорость вычисления агрегатных значений одинакова для любого из измерений. Однако в этом случае многомерная база данных оказывается избыточной, так как многомерные данные полностью содержат исходные реляционные данные.

# Способы хранения данных

- ROLAP (Relational OLAP) — исходные данные остаются в той же реляционной базе данных, где они изначально и находились. Агрегатные же данные помещают в специально созданные для их хранения служебные таблицы в той же базе данных.
- HОLAP (Hybrid OLAP) — исходные данные остаются в той же реляционной базе данных, где они изначально находились, а агрегатные данные хранятся в многомерной базе данных.

# Недостатки OLAP

- Повышенные требования к аппаратному обеспечению (в большей части к объему оперативной памяти)
- (Не всегда) Достаточная нестабильная работа при сложном разрезе многомерного куба с большим объемом информации и одновременной перетасовкой измерений



# Примеры программного обеспечения

- OLAP-функциональность реализована в средствах статистической обработки данных (продукты компаний StatSoft и SPSS);
- электронные таблицы (неплохими средствами многомерного анализа обладает Microsoft Excel 2000) С его помощью можно создать и сохранить в виде файла небольшой локальный многомерный OLAP-куб и отобразить его двух- или трехмерные сечения.
- средства разработки содержат библиотеки классов или компонентов, (такие, например, как компоненты DecisionCube в Borland Delphi и Borland C++Builder)
- многие компании предлагают элементы управления ActiveX и другие библиотеки, реализующие подобную функциональность

# Вопросы

1. Расшифруйте и поясните значение терминов OLAP и OLTP.
2. Что представляют собой хранилища данных и для чего они предназначены?
3. Как соотносятся между собой хранилища данных и OLAP?
4. Приведите примеры программных средств, реализующих OLAP-функциональность.
5. Что такое таблицы фактов и таблицы измерений?
6. Почему таблицы данных в OLAP должны быть денормализованы?

# Вопросы

7. Чем отличаются хранилища данных типа “Звезда” и хранилища данных типа “Снежинка”? Какому типу отдается предпочтение при построении ХД?
8. Для чего предназначены корпоративные хранилища данных и киоски данных?
9. Охарактеризуйте хранилища данных по способу хранения данных (MOLAP, ROLAP, HOLAP).
10. Что означает тест FASMI? Расшифруйте значение каждой литеры.