Тема №7 Технология многомерных баз данных

1. Многомерные модели
2. Электронные таблицы и отношения
3. Кубы
4. Измерения
5. Факты
6. Параметры
7. Запросы
8. Реализация

Многомерные модели

Реляционная модель данных, которая была предложена Э.Ф. Коддом в 1970 году, и за которую десятилетие спустя он получил премию Тьюринга, служит основой современной многомиллиардной отрасли баз данных.

За последние десять лет сложилась многомерная модель данных, которая используется, когда целью является именно ***анализ данных***, а ***не выполнение транзакций***.

Подобные базы данных трактуют данные как многомерные кубы, что очень удобно именно для их анализа.

Многомерные модели

Многомерные модели рассматривают данные либо как факты с соответствующими численными параметрами, либо как текстовые измерения, которые характеризуют эти факты.

В розничной торговле, к примеру, покупка — это факт, объем покупки и стоимость — параметры, а тип приобретенного продукта, время и место покупки — измерения.

Запросы агрегируют значения параметров по всему диапазону измерения, и в итоге получают такие величины, как общий месячный объем продаж данного продукта.

Многомерные модели

Многомерные модели данных имеют три важных области применения, связанных с проблематикой анализа данных:

- Хранилища данных интегрируют для анализа информации из нескольких источников на предприятии.
- Системы оперативной аналитической обработки (online analytical processing — OLAP) позволяют оперативно получить ответы на запросы, охватывающие большие объемы данных в поисках общих тенденций.
- Приложения добычи данных служат для выявления знаний за счет полуавтоматического поиска ранее неизвестных шаблонов и связей в базах данных.

Электронные таблицы и отношения

Электронные таблицы не подходят для управления и хранения многомерных данных, поскольку они слишком жестко связывают данные с их внешним видом, не отделяя структурную информацию от желаемого представления информации.

Например, добавление третьего измерения, такого как время, или группировка данных по обобщенным типам продуктов требует значительно более сложной настройки.

Очевидное решение состоит в использовании отдельной электронной таблицы для каждого измерения. Но это оправдано только в ограниченной степени, поскольку анализ подобных наборов таблиц быстро становится громоздким.

Электронные таблицы и отношения

Использование баз данных, поддерживающих SQL, значительно увеличивает гибкость обработки структурированных данных.

Однако сформулировать многие вычисления, такие как совокупные показатели (объем продаж за год к текущему моменту), сочетание итоговых и промежуточных результатов, ранжирование, например, определение десяти самых продаваемых продуктов, посредством стандартного варианта SQL весьма сложно.

При перестановке строк и столбцов необходимо вручную специфицировать и комбинировать различные представления.

Электронные таблицы и отношения

Электронные таблицы и реляционные базы данных адекватно обрабатывают массивы данных, которые имеют незначительное число измерений, но они ***не полностью отвечают требованиям углубленного анализа данных.***

Решение состоит в том, чтобы использовать технологию, которая предусматривает поддержку полного спектра средств многомерного моделирования данных.

Кубы

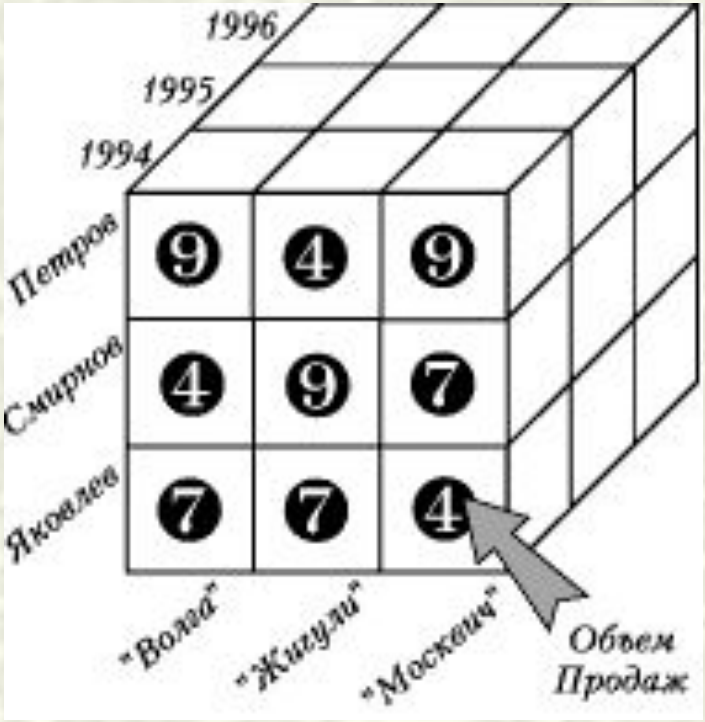
Многомерные базы данных рассматривают данные как кубы, которые являются обобщением электронных таблиц на любое число измерений.

Кроме того, кубы поддерживают иерархию измерений и формул без дублирования их определений.

Набор соответствующих кубов составляет многомерную базу данных (или хранилище данных).

Кубы

Куб с тремя измерениями:



Кубы

Кубами легко управлять, добавляя новые значения измерений.

В обычном обиходе этим термином обозначают фигуру с тремя измерениями, однако теоретически куб может иметь любое число измерений.

На практике чаще всего кубы данных имеют от 4 до 12 измерений.

Современный инструментарий часто сталкивается с нехваткой производительности, когда так называемый гиперкуб имеет свыше 10-15 измерений.

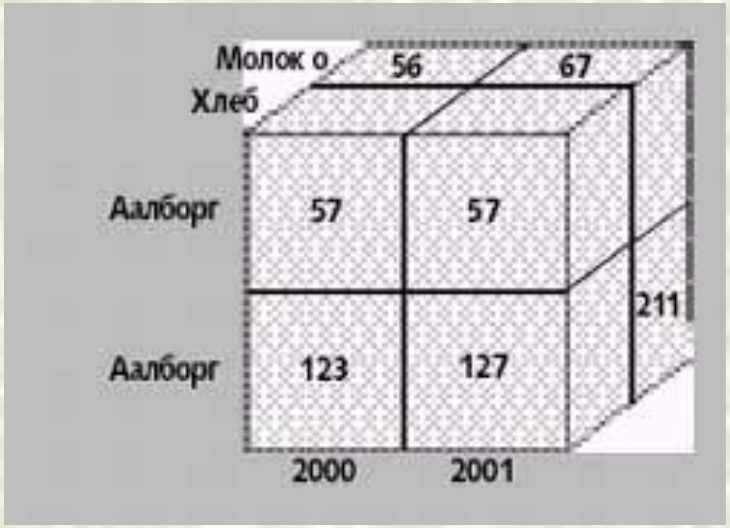
Кубы

Комбинации значений измерений определяют ячейки куба. В зависимости от конкретного приложения ячейки в кубе могут располагаться как разрозненно, так и плотно.

Кубы, как правило, становятся разрозненными по мере увеличения числа размерностей и степени детализации значений измерений.

Кубы

Куб, содержащий данные по продажам в двух датских городах:



В соответствующих ячейках хранятся данные об объеме продаж. В примере можно обнаружить «факт» — непустую ячейку, содержащую соответствующие числовые параметры — для каждой комбинации время, продукт и город, где была совершена, по крайней мере, одна продажа.

Кубы

В общем случае куб позволяет представить только два или три измерения одновременно, но можно показывать и больше за счет вложения одного измерения в другое.

Таким образом, путем проецирования куба на двух- или трехмерное пространство можно уменьшить размерность куба, агрегировав некоторые размерности, что ведет к работе с более комплексными значениями параметров.

Измерения

Измерения — ключевая концепция многомерных баз данных.

Многомерное моделирование предусматривает использование измерений для предоставления максимально возможного контекста для фактов.

В отличие от реляционных баз данных, контролируемая избыточность в многомерных базах данных, в общем, считается оправданной, если она увеличивает информационную ценность.

Поскольку данные в многомерный куб часто собираются из других источников, например, из транзакционной системы, проблемы избыточности, связанные с обновлениями, могут решаться намного проще.

Измерения

Измерения используются для выбора и агрегирования данных на требуемом уровне детализации.

Измерения организуются в иерархию, состоящую из нескольких уровней, каждый из которых представляет уровень детализации, требуемый для соответствующего анализа.

Иногда бывает полезно определять несколько иерархий для измерения. Например, модель может определять время как в финансовых годах, так и в календарных

Измерения

Схема «Местоположение» для данных продаж:

Из трех уровней измерений местоположения самый низкий — «Город». Значения уровня «Город» группируются в значения на уровне «Страна».

Уровень Т представляет все измерения.

Измерения

В отличие от линейных пространств, с которыми имеет дело алгебра матриц, многомерные модели, как правило, не предусматривают функций упорядочивания или расстояния для значений измерения.

Единственное «упорядочивание» состоит в том, что значения более высокого уровня содержат значения более низких уровней.

Для некоторых измерений, таких как время, упорядоченность значений размерности может использоваться для вычисления совокупной информации, такой как общий объем продаж за определенный период.

Факты

Факты представляют субъект — некий шаблон или событие, которые необходимо проанализировать.

В большинстве многомерных моделей данных факты однозначно определяются комбинацией значений измерений; факт существует только тогда, когда ячейка для конкретной комбинации значений не пуста.

Большинство многомерных моделей требуют, чтобы каждому факту соответствовало одно значение на более низком уровне каждого измерения, но в некоторых моделях это не является обязательным требованием

Факты

Каждый факт обладает некоторой ***гранулярностью***, определенной уровнями, из которых создается их комбинация значений измерений.

Например, гранулярность факта в кубе — это (Год x Продукт x Город).

(Год x Тип x Город) и (День x Продукт x Город) — соответственно более грубая и более тонкая гранулярности.

Факты

Хранилища данных, как правило, содержат следующие три типа фактов:

- **События (event)**
- **Мгновенные снимки (snapshot)**
- **Совокупные мгновенные снимки (cumulative snapshot)**

Факты

События (event), по крайней мере, на уровне самой большой гранулярности, как правило, моделируют события реального мира, при этом каждый факт представляет определенный экземпляр изучаемого явления.

Примерами могут служить продажи, щелчки мышью на Web-странице или движение товаров на складе.

Факты

Мгновенные снимки (snapshot) моделируют состояние объекта в данный момент времени, такие как уровни наличия товаров в магазине или на складе и число пользователей Web-сайта.

Один и тот же экземпляр явления реального мира, например, конкретная банка бобов, может возникать в нескольких фактах.

Факты

Совокупные мгновенные снимки (cumulative snapshot) содержат информацию о деятельности организации за определенный отрезок времени.

Например, совокупный объем продаж за предыдущий период, включая текущий месяц, можно легко сравнить с показателями за соответствующие месяцы прошлого года.

Факты

Хранилище данных часто содержит все три типа фактов.

Одни и те же исходные данные, например, движение товаров на складе, могут содержаться в трех различных типах кубов:

поток товаров на складе,

список товаров и

поток за год к текущей дате.

Параметры

Параметры состоят из двух компонентов:

- численная характеристика факта, например, цена или доход от продаж;

- формула, обычно простая агрегативная функция, скажем, сумма, которая может объединять несколько значений параметров в одно.

В многомерной базе данных параметры, как правило, представляют свойства факта, который пользователь хочет изучить.

Параметры

При вычислениях три различных класса параметров ведут себя совершенно по-разному.

Аддитивные параметры могут содержательным образом комбинироваться в любом измерении.

Например, имеет смысл суммировать общий объем продаж для продукта, местоположения и времени, поскольку это не вызывает наложения среди явлений реального мира, которые генерируют каждое из этих значений.

Параметры

Полуаддитивные параметры, которые не могут комбинироваться в одном или нескольких измерениях.

Например, суммирование запасов по разным товарам и складам имеет смысл, но суммирование запасов товаров в разное время бессмысленно, поскольку одно и то же физическое явление может учитываться несколько раз.

Параметры

Неаддитивные параметры не комбинируются в любом измерении, обычно потому, что выбранная формула не позволяет объединить средние значения низкого уровня в среднем значении более высокого уровня.

Аддитивные и неаддитивные параметры могут описывать факты любого рода, в то время как полуаддитивные параметры, как правило, используются с мгновенными снимками или совокупными мгновенными снимками.

Запросы

Многомерная база данных естественным образом предназначена для определенных типов запросов:

- **Запросы вида slice-and-dice**
- **Запросы вида drill-down и roll-up**
- **Запросы вида drill-across**
- **Запросы вида ranking**
- **Поворот (rotating)**

Запросы

Запросы вида slice-and-dice осуществляют выбор, сокращающий куб.

К примеру, можно рассмотреть сечение куба, приняв во внимание только те ячейки, которые касаются хлеба, а затем еще больше сократить его, оставив ячейки, относящиеся только к 2000 году.

Фиксация значения измерения сокращает размерность куба, но при этом возможны и более общие операции выбора.

Запросы

Запросы вида drill-down и roll-up — взаимообратные операции, которые используют иерархию измерений и параметры для агрегирования.

Обобщение до высших значений соответствует исключению размерности.

Например, свертка от уровня «Город» до уровня «Страна» агрегирует значения для Аалборга и Копенгагена в одно значение — Дания.

Запросы

Запросы вида drill-across комбинируют кубы, которые имеют одно или несколько общих измерений. С точки зрения реляционной алгебры такая операция выполняет слияние (join).

Запросы вида ranking возвращает только те ячейки, которые появляются в верхней или нижней части упорядоченного определенным образом списка, например, 10 самых продаваемых продуктов в Копенгагене в 2000 году.

Поворот (rotating) куба дает пользователям возможность увидеть данные, сгруппированные по другим измерениям.

Реализация

Многомерные базы данных реализуют в двух основных формах:

1. Системы многомерной оперативной аналитической обработки (MOLAP) хранят данные в специализированных многомерных структурах.

   Системы MOLAP, как правило, содержат средства для обработки разреженных массивов и применяют усовершенствованную индексацию и хеширование для поиска данных при выполнении запросов.

Реализация

2. Реляционные системы OLAP (ROLAP) для хранения данных используют реляционные базы данных, а также применяют специализированные индексные структуры, такие как битовые карты, чтобы добиться высокой скорости выполнения запросов.

Реализация

Системы MOLAP, как правило, позволяют добиться более эффективного использования дискового пространства, а также меньшего времени ответов при обработке запросов.

Системы ROLAP, как правило, лучше масштабируются с ростом числа фактов, которые они могут хранить, более гибкими в том, что касается переопределения кубов, и лучше поддерживают частые обновления.

Достоинства двух подходов объединены в гибридном решении, при котором для хранения сводных данных более высокого уровня используется технология MOLAP, а в системах ROLAP размещаются детальные данные.

Реализация

В ROLAP, как правило, используются схемы «звезда» и «снежинка», при которых данные хранятся в таблицах фактов и таблицах измерений.

Таблица фактов содержит одну строку для каждого факта в кубе. Для каждого измерения отводится отдельный столбец, содержащий значение параметра для конкретного факта, а также столбец для каждого измерения, которое содержит внешний ключ, ссылающийся на таблицу измерений для конкретного измерения.

Реализация

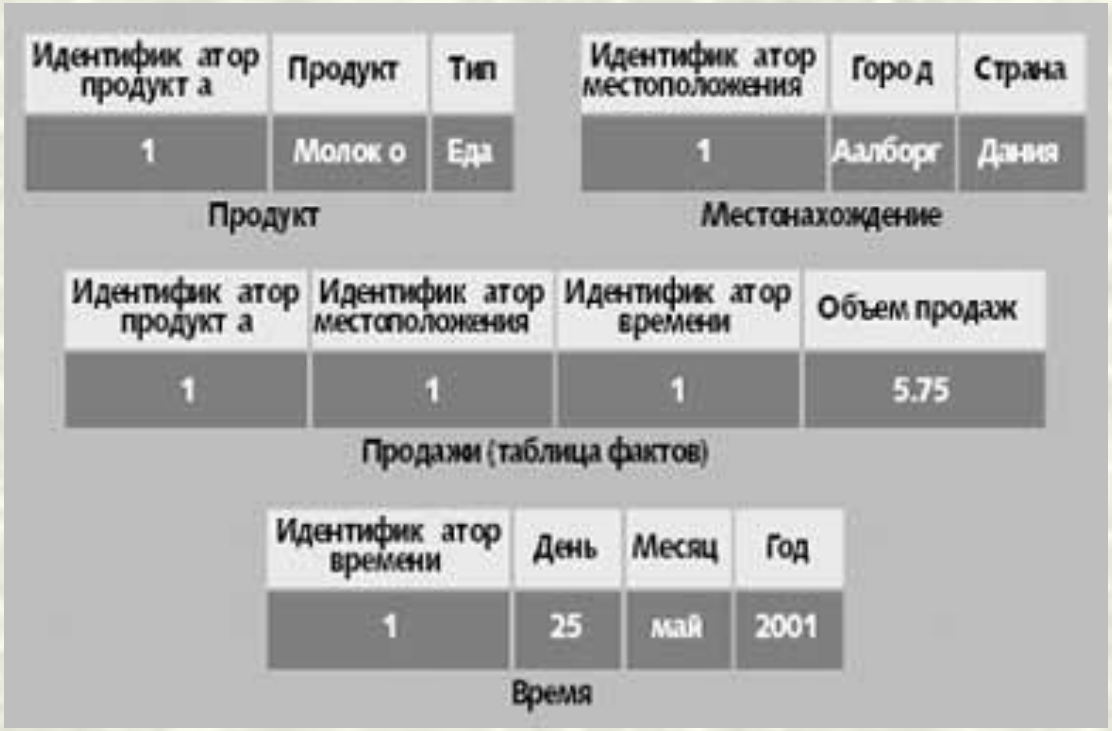
Схемы «звезда» и «снежинка» отличаются в том, как они поддерживают измерения, и выбор между ними, в основном, зависит от того, какими свойствами должна обладать разрабатываемая система.

В схеме «звезда» на каждое измерение отводится одна таблица.

Таблица измерений содержит ключевой столбец, по одному столбцу для каждого уровня измерений с текстовыми описаниями значений этого уровня, и по одному столбцу для каждого свойства уровня в измерении.

Реализация

Схема «звезда» для куба продаж:



Информация со всех уровней в измерении хранится в одной таблице измерений, например, названия продуктов и типы продуктов хранятся в таблице «Продукт»

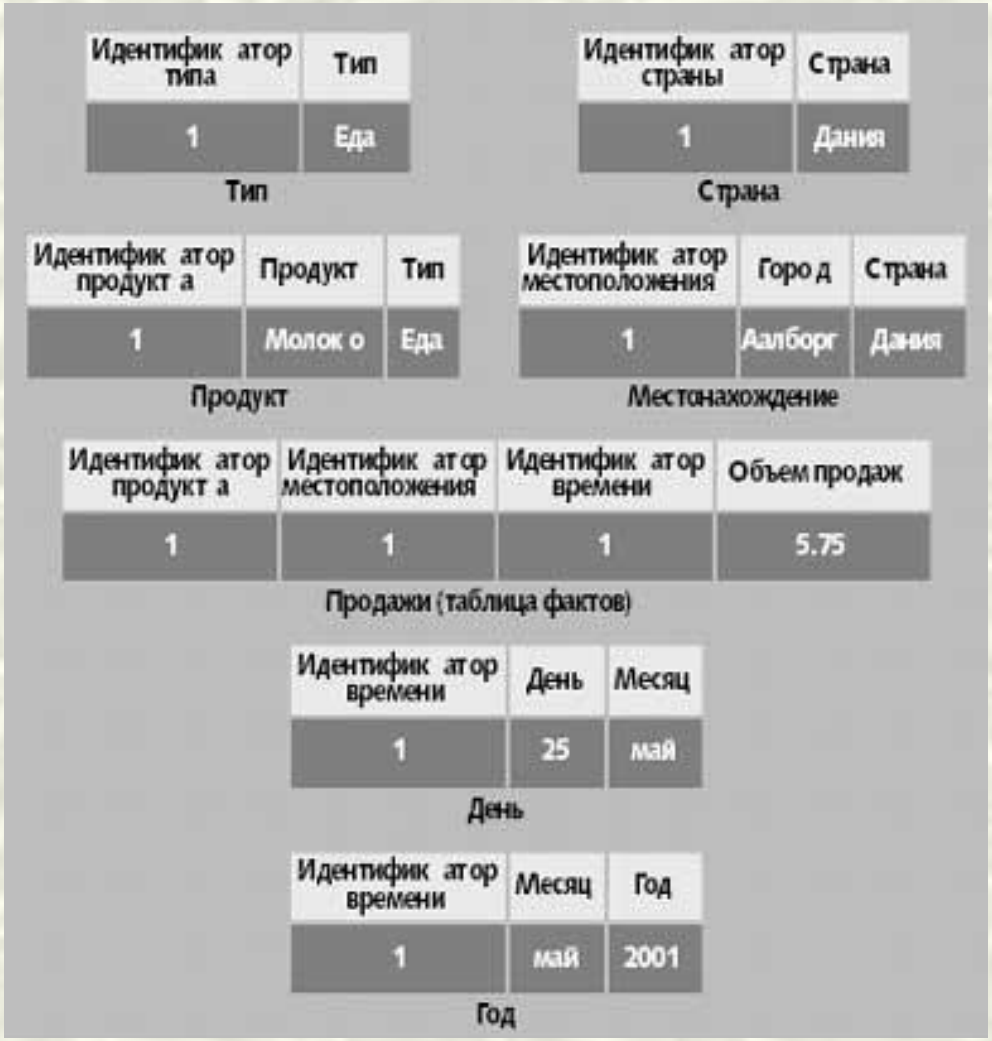
Реализация

Схема «снежинка» содержит по одной таблице для каждого уровня измерений, избегая избыточности, что может оказаться весьма полезным в некоторых ситуациях.

Каждая из таблиц измерений содержит ключ, столбец с текстовыми описаниями значений уровней, возможно, столбцы для свойств уровней.

Таблицы более низких уровней могут также содержать внешний ключ для доступа к более высокому уровню.

Реализация



Тип

Идентификатор типа	Тип
1	Еда

Страна

Идентификатор страны	Страна
1	Дания

Продукт

Идентификатор продукта	Продукт	Тип
1	Молоко	Еда

Местонахождение

Идентификатор местоположения	Город	Страна
1	Аалборг	Дания

Продажи (таблица фактов)

Идентификатор продукта	Идентификатор местоположения	Идентификатор времени	Объем продаж
1	1	1	5.75

День

Идентификатор времени	День	Месяц
1	25	май

Год

Идентификатор времени	Месяц	Год
1	май	2001

Информация из различных уровней в измерении хранится в различных таблицах.

Например, названия продуктов и типы продуктов хранятся в таблицах «Продукт» и «Тип» соответственно