БАЗЫ ДАННЫХ
часть II

Многомерные базы данных

Многомерные БД

Если целью является именно анализ данных, а не выполнение транзакций, используется многомерная модель данных. Технология многомерных баз данных — ключевой фактор интерактивного анализа больших массивов данных с целью поддержки принятия решения. Подобные базы данных трактуют данные как многомерные кубы, что очень удобно именно для их анализа.

Многомерные БД

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

Многомерные БД

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

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

Многомерные БД

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.

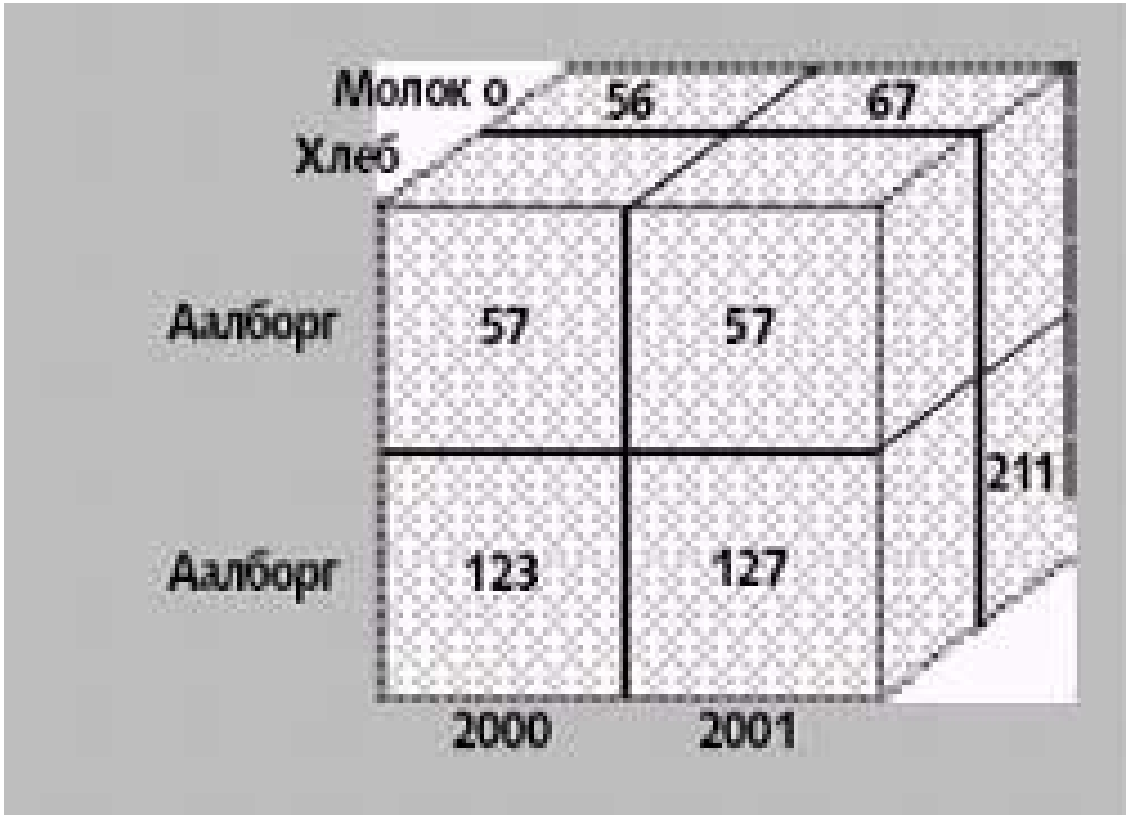
Использование баз данных, поддерживающих SQL, не позволяет сформулировать многие вычисления, такие как совокупные показатели (объем продаж за год к текущему моменту), сочетание итоговых и промежуточных результатов, ранжирование.

Многомерные БД

Многомерные базы данных рассматривают данные как кубы, которые являются обобщением электронных таблиц на любое число измерений. Кубы поддерживают иерархию измерений и формул без дублирования их определений. Набор соответствующих кубов составляет многомерную базу данных (или хранилище данных). Комбинации значений измерений определяют ячейки куба.

Многомерные БД

Пример куба, содержащего данные о продажах. В этом случае куб обобщает электронную таблицу, добавляя к ней третье измерение — время.



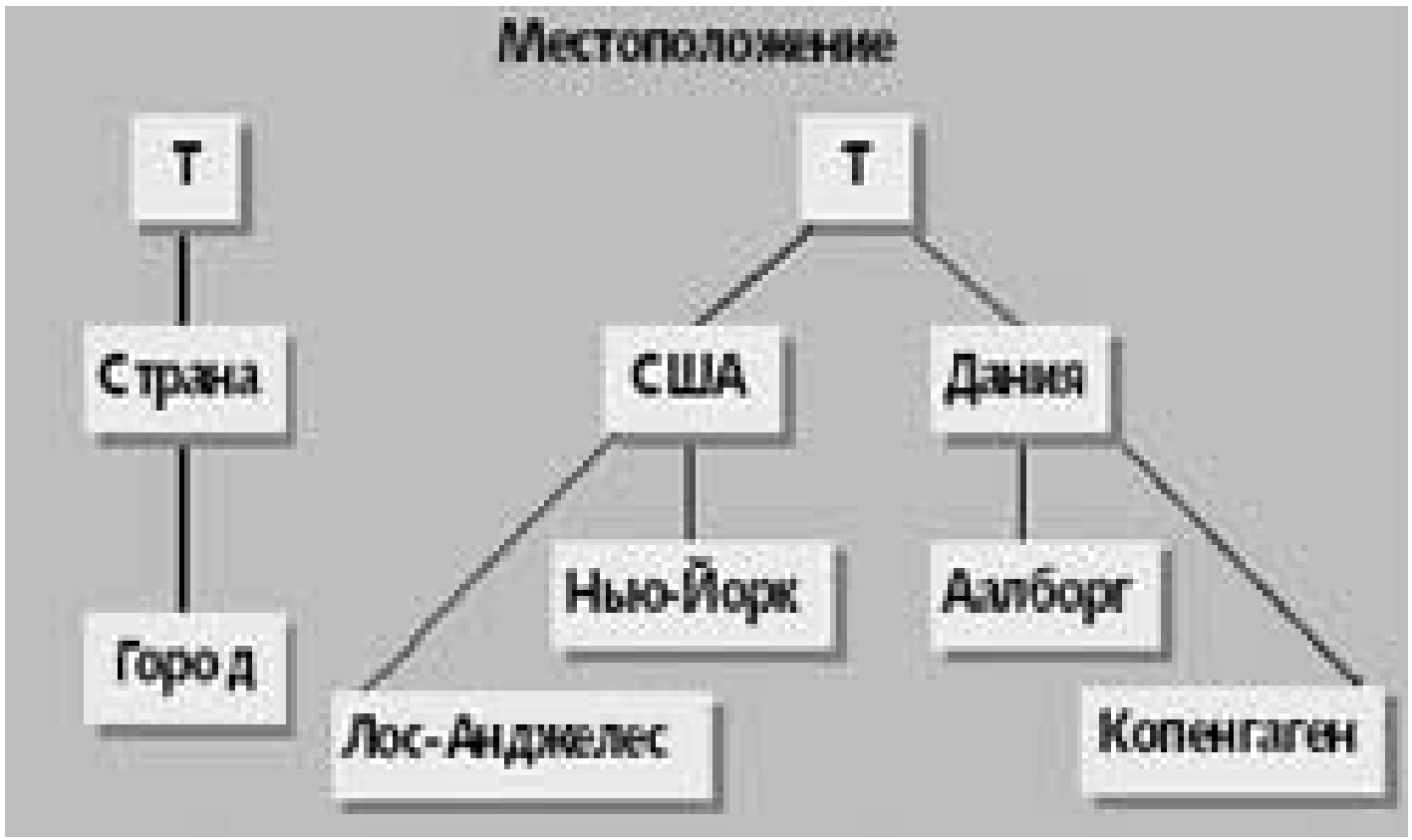
Многомерные БД

Измерения — ключевая концепция многомерных баз данных. Многомерное моделирование предусматривает использование измерений для предоставления максимально возможного контекста для фактов.

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.

Многомерные БД

Пример схемы измерений местоположения. Каждое значение размерности является частью значения Т.



Многомерные БД

В отличие от линейных пространств, с которыми имеет дело алгебра матриц, многомерные модели, как правило, не предусматривают функций упорядочивания или расстояния для значений измерения. Однако для некоторых измерений, таких как время, упорядоченность значений размерности может использоваться для вычисления совокупной информации. Большинство моделей требуют определения иерархии измерений для формирования сбалансированных деревьев — иерархии должны иметь одинаковую высоту по всем ветвям, а каждое значение не корневого уровня — только одного родителя.

Многомерные БД

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.

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

Многомерные БД

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

1. **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.
2. **Мгновенные снимки (snapshot)** моделируют состояние объекта в данный момент времени.
3. **Совокупные мгновенные снимки (cumulative snapshot)** содержат информацию о деятельности организации за определенный отрезок времени.

Многомерные БД

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.

Многомерные БД

1. **Аддитивные параметры** могут содержательным образом комбинироваться в любом измерении.
2. **Полуаддитивные параметры,** которые не могут комбинироваться в одном или нескольких измерениях.
3. **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.

Многомерные БД

Многомерная база данных предназначена для определенных типов запросов:

1. **Запросы вида slice-and-dice** осуществляют выбор, сокращающий куб.
2. **Запросы вида drill-down и roll-up** — взаимообратные операции, которые используют иерархию измерений и параметры для агрегирования. Обобщение до высших значений соответствует исключению размерности.

Многомерные БД

3. **Запросы вида drill-across** комбинируют кубы, которые имеют одно или несколько общих измерений. С точки зрения реляционной алгебры такая операция выполняет слияние (join).
4. **Запросы вида ranking** возвращает только те ячейки, которые появляются в верхней или нижней части упорядоченного определенным образом списка.
5. **Поворот (rotating)** куба дает пользователям возможность увидеть данные, сгруппированные по другим измерениям.

Многомерные БД

Многомерные базы данных реализуют в двух основных формах:

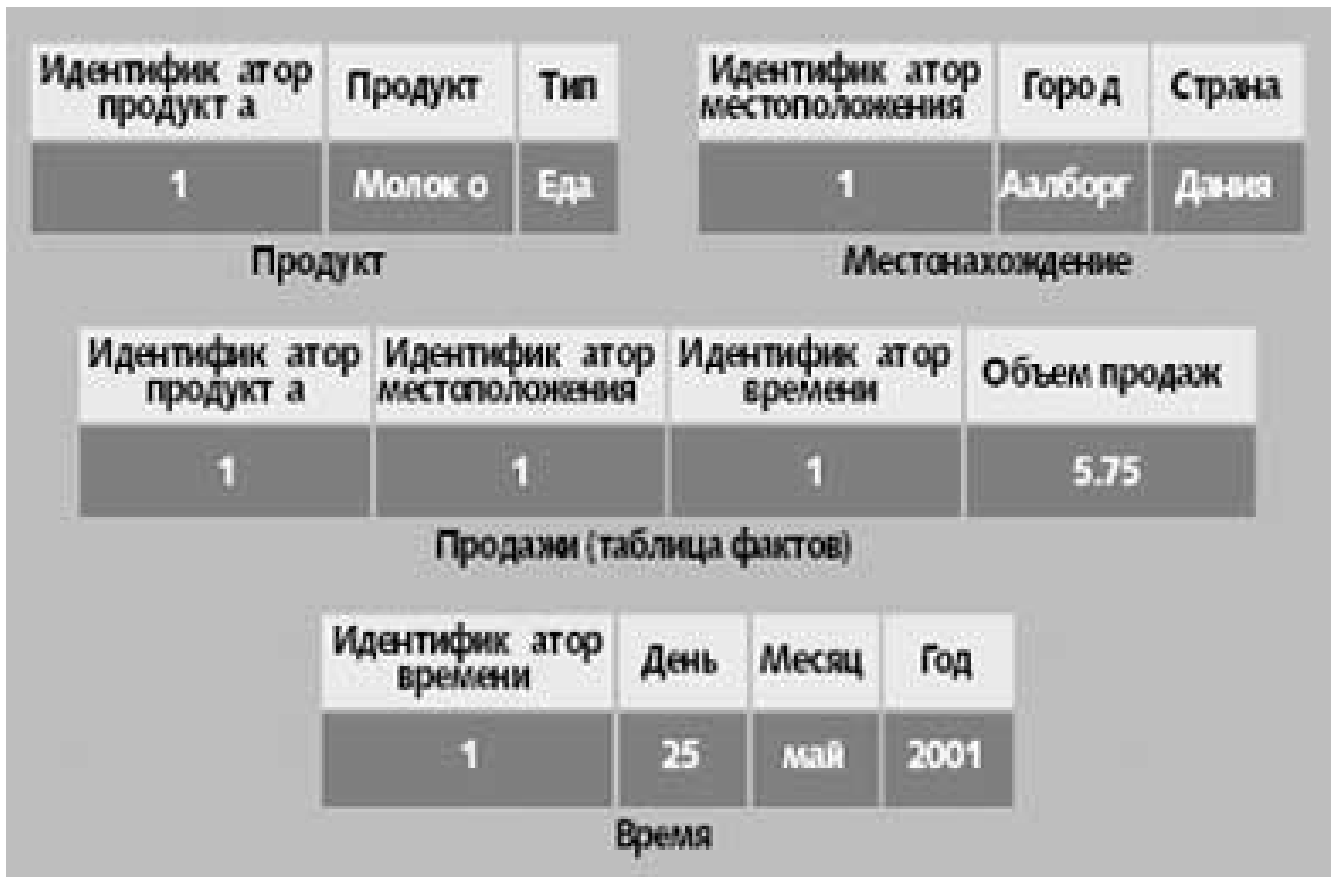
1. Системы многомерной оперативной аналитической обработки (MOLAP) хранят данные в специализированных многомерных структурах.
2. Реляционные системы OLAP (ROLAP) для хранения данных используют реляционные базы данных, а также применяют специализированные индексные структуры, такие как битовые карты, чтобы добиться высокой скорости выполнения запросов.

Многомерные БД

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.

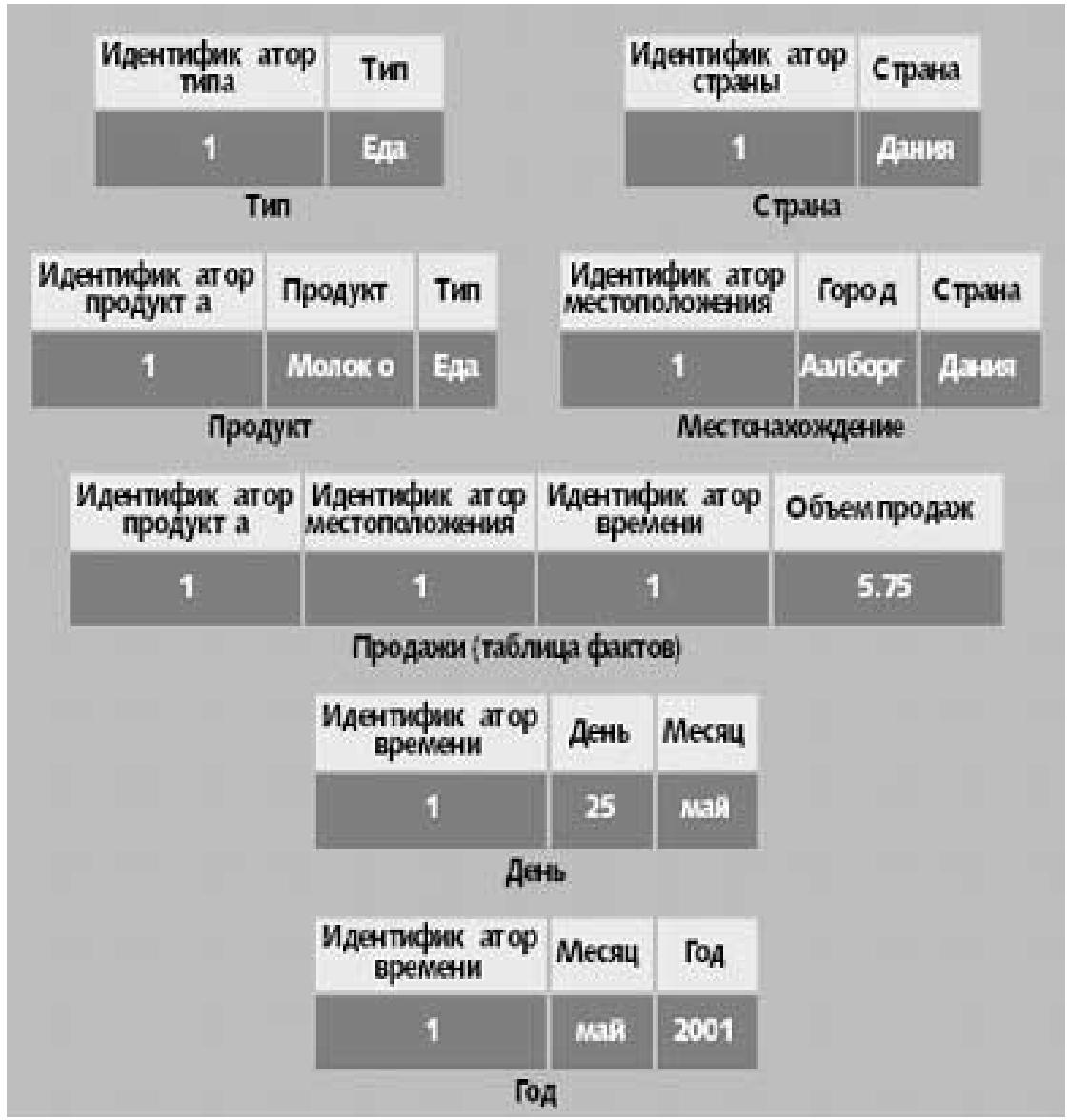
Многомерные БД

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».



Многомерные БД

Схема «снежинка» для куба продаж. Информация из различных уровней в измерении хранится в различных таблицах. Например, названия продуктов и типы продуктов хранятся в таблицах «Продукт» и «Тип» соответственно.



Многомерные БД

За 30 лет с момента своего возникновения технология многомерных баз данных прошла серьезную эволюцию. С недавних пор она стала реализовываться в решениях, предназначенных для массового рынка, а ведущие производители теперь выпускают многомерные ядра вместе со своими реляционными базами данных.

Многомерные БД

Данные, которые необходимо анализировать, становятся все более распределенными. К примеру, это часто необходимо для выполнения анализа, при котором используются данные в формате XML, получаемые с определенных Web-сайтов. Растущая распределенность данных, в свою очередь, требует применения методов, которые позволяют легко добавлять новые данные в многомерные базы данных, тем самым, упрощая задачу создания интегрированного хранилища данных.

Многомерные БД

Технология многомерных баз данных также применяется к новым типам данных, которые современные технологии зачастую не в состоянии адекватно анализировать.

Наконец, технология многомерных баз данных все больше будет применяться там, где результаты анализа напрямую передаются в другие системы, тем самым, исключая участие человека в этом процессе.

ВОПРОСЫ ?

СОВЕТ ДНЯ: